

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN



MASTER THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF ARTS IN GENERAL LINGUISTICS

A theoretical approach to automatic loanword detection

Author:
Marisa DELZ

1st Supervisor:
Prof. Dr. Gerhard JÄGER
2nd Supervisor:
Dr. des. Johann-Mattis LIST

SEMINAR FÜR SPRACHWISSENSCHAFT
EBERHARD-KARLS-UNIVERSITÄT TÜBINGEN

September 2013

Ich versichere, dass ich die Arbeit ohne Benutzung anderer als der angegebenen Quellen angefertigt habe und dass die Arbeit in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegen hat und von dieser als Teil einer Prüfungsleistung angenommen wurde. Alle Stellen und Personen, welche mich bei der Vorbereitung und Anfertigung der Abhandlung unterstützten, wurden genannt und Ausführungen, die wörtlich oder sinngemäss übernommen wurden, sind als solche gekennzeichnet.

Tübingen, den September 9, 2013

Marisa Delz

Zusammenfassung

Vor einigen Jahren haben automatische Methoden und Computeranalysen ihren Weg in die Geisteswissenschaften gefunden. Vor allem die Computerlinguistik untersucht und entwickelt neue Methoden. Es ist daher nicht überraschend, dass das Interesse an unterschiedlichen Computeranalysen im Bereich der historischen Linguistik an Interesse gewonnen hat. Neue Ansätze haben die Sicht auf die Untersuchungsmethoden innerhalb der Sprachevolution verändert. Biologische Evolution und Sprachevolution weisen verschiedene Gemeinsamkeiten auf. Die Ähnlichkeiten zwischen Phylogenetik und Linguistik haben zu einer Kombination dieser Bereiche geführt. Die Phylogenetik stellt eine große Anzahl von mathematischen und auch implementierten Methoden zur Verfügung, um unterschiedliche Prozesse zu analysieren. Einige dieser Methoden können auf Grund der Gemeinsamkeiten dieser Bereiche in die historische Linguistik übernommen werden. In der historischen Linguistik ist die Entlehnung ein bekannter evolutionärer Prozess, bei welchem Wörter der einen Sprache in eine andere entlehnt werden. Der Prozess der Entlehnung weist große Ähnlichkeiten mit dem aus der Phylogenetik bekannten Prozess des Horizontalen Gentransfers auf. Horizontaler Gentransfer beschreibt die Übertragung von Genen von einem Organismus in einen anderen. Die Gemeinsamkeit von Entlehnung und Horizontalem Gentransfer ist die Übertragung von Genen oder Wörtern, wobei der Organismus oder die Sprache nicht verwandt sein müssen. Die Phylogenetik stellt mehrere mathematische Methoden und Analysen zur Verfügung, um Horizontalen Gentransfer zu erkennen. Diese könnten in die Linguistik übernommen werden. In dieser Arbeit werden die Hintergründe von Entlehnung und die Grundlagen der Phylogenetik erklärt. Des Weiteren wird die Kombination der beiden Bereiche erläutert. Der neue baumbasierte Ansatz soll zeigen, ob die Methoden aus der Phylogenetik in die Linguistik aufgenommen werden können und ob diese Entlehnungen erkennen können.

Abstract

For several years, computational methods found their way into humanities. Especially in the field of computational linguistics several analysis and methods are studied. It is not surprising that computational analysis arouse interest in the field of historical linguistics. Due to such methods, language evolution can be studied from another point of view. Biological and linguistic evolution show certain parallels. Especially the parallels between phylogenetics and linguistics arouse the interest of combining both fields. Phylogenetics provide a great number of mathematical and computational methods for computing different tasks. Based on the parallels, the methods can be adapted into historical linguistics. In historical linguistics, the process of borrowing is a well-known evolutionary process where words are borrowed from one language and adapted into another. Borrowing has its corresponding parallel within phylogenetics, namely horizontal gene transfer. Horizontal gene transfer is the process of transferring genes from one organism to another. The similarity between borrowing and horizontal gene transfer is the transfer of genes or words whereas the organisms or languages are not related. Phylogenetics provides several computational methods and analysis to detect horizontal gene transfer. The methods might be adapted into linguistics to detect borrowing. This paper introduces the background of borrowing and phylogenetics as well as the combination of both fields. The new tree-based approach should indicate if provided methods of phylogenetics can be adapted into linguistics for the detection of borrowing.

Acknowledgements

I would like to express my greatest gratitude to the people who have helped and supported me throughout my project.

I am truly and indebtedly grateful to my first supervisor Prof. Dr. Gerhard Jäger and my second supervisor Johann-Mattis List for their valuable guidance and support throughout my project and my theses. Without their hints during my research and writing phase it would have been more difficult to finish this project. Their support and knowledge helped me to understand the topic of phylogenetics in a clearer way and helped me to establish the approach. I would also like to thank Prof. Dr. Gerhard Jäger for providing and preparing the language data used in my approach. Additionally, I would like to thank Johann-Mattis List for explanations and insights on LingPy and its implementation, as well as for providing a case study and additional material. Besides, I would like to thank Prof. Dr. Daniel Huson for his helpful suggestions to my idea. Thanks go also to Heike Cardoso for corrections and suggestions on my thesis. Johannes Dellert for helpful insights on LingPy and discussion and correction of this topic. Philip Schulz for thoughts and suggestions. Johannes Wahle for insightful discussions and suggestions and moral support.

Contents

List of Figures	viii
List of Tables	ix
1 Introduction	1
2 Linguistic background of loanwords	3
2.1 Loanwords	4
2.2 Processes in loanword adaption	10
2.3 Theoretical approaches to loanword adaption	13
2.4 Automatic loanword detection	17
3 Phylogenetics	19
3.1 Theoretical Background on Phylogenetics	20
3.2 Phylogenetic Trees	23
3.2.1 Reconstruction of phylogenetic trees	28
3.2.2 Working with phylogenetic trees	32
3.3 Phylogenetic Networks	34
3.3.1 Different Types of Networks	36
4 Phylogenetics in Historical Linguistics	41
4.1 LingPy	44
4.1.1 The Python Library for Historical Linguistics	45
4.1.2 Borrowing Detection with LingPy	47
5 Borrowing Detection with Horizontal Transfer	56
5.1 Species trees and gene trees	57
5.2 Horizontal Gene Transfer	63
5.3 T-REX: a web server	67
5.4 Horizontal Language Transfer and LingPy	69
6 Automatic and Manually Approaches: A Comparison	72
6.1 The World Loanword Database (WOLD)	72
6.1.1 Background and content of WOLD	72
6.1.2 Representation and Findings in WOLD	77
6.1.3 Leipzig-Jakarta List	80
6.2 Automatic versus Manually Approaches	81
7 Conclusion	85

References	87
A Some Information on the WOLD database	I
B The Swadesh 100-word list	IV
C The Leipzig-Jakarta 100-word list	V
D Expert Tree of the Indo-European languages	VI
E Concept Tree “Mountain” of the Indo-European languages	VIII

List of Figures

3.1	Darwin's first sketch of a tree	21
3.2	Darwin's tree in <i>On the Origin of Species</i>	22
3.3	Haeckel's famous pedigree of man	24
3.4	The main steps towards the reconstruction of a phylogenetic tree .	29
3.5	A representation from a tree to a network	35
3.6	An illustration of a split network.	37
3.7	An illustration of a rooted hybridization network.	39
4.1	Heackel's illustration of a pedigree for the Indo-European language family.	42
4.2	Schleicher's illustration of a pedigree for the Indo-European language family.	42
4.3	Workflow through the LingPy program.	46
4.4	Gain-loss mappings	48
4.5	The reference tree for Info-European languages based on Ringe (2002)	50
4.6	The reference tree for Info-European languages based on Southworth (1964)	50
4.7	The MLN for IELex using the reference tree of Ringe (2002) . . .	51
4.8	The MLN for IELex using the reference tree of Southworth (1964)	51
5.1	The expert tree of Germanic and Romance Languages	59
5.2	The concept tree for mountain of Germanic and Romance languages	60
5.3	The concept tree with the missing entries	61
5.4	The concept tree without the missing entries	62
5.5	Horizontal gene transfer	63
5.6	Horizontal transfer between the expert tree and the concept tree .	64
5.7	Horizontal transfer within the expert tree	65
5.8	The HGT network for Germanic and Romance languages	68
6.1	A map of the languages in WOLD	74
6.2	A part of the webside representing the vocabulary list	77

List of Tables

4.1	Conceptual parallels between biological and linguistic evolution	43
4.2	The top links for the minimal lateral network in figure 4.7	52
4.3	The top links for the minimal lateral network in figure 4.8	53
6.1	The main differences between an automatic and a manual approach	81
A.1	The LWT project languages	I
A.2	The semantic fields	II
A.3	The lexical borrowing rates	III

1 Introduction

Historical linguistics is a well-studied field within linguistics where the interest of using computational methods and analysis continuously increased. Computational studies also found their way into historical linguistics and since then scientists are interested in studying language evolution from another point of view. The study of language classification arouse interest. Within the study of language classification, the usage of phylogenetic method's increased in recent years. Phylogenetics is a field of biology that studies evolutionary relationships between organisms. The basic idea goes back to Darwin, who constructed a tree of life for classifying organisms according to their evolution and relationship.

In linguistics, August Schleicher was one of the first who introduced a pedigree for the Indo-European languages. In an open letter exchange with his friend Ernst Haeckel, the discussion about the similarities between biological evolution and linguistic evolution arose. Darwin also indicates in one of his papers that there are parallels between biological and linguistic evolution.

The usage of a tree for representing classifications is not new in linguistics. In the field of Syntax, trees are used for representing word and phrase relations within a sentence. Therefore, the usage of trees for representing other kinds of relations can easily be adapted into linguistics. Schleicher's pedigree for the Indo-European languages represent the relation and evolution of the languages and can therefore be compared to Darwin's tree of life. The representation of relationships within a tree is not the only parallel between biology and linguistics.

With the parallels between biological and linguistic evolution, the usage of phylogenetic methods within linguistics becomes increasingly more interesting. The parallels are the basis for adapting phylogenetic methods in linguistics and for developing new approaches. With the adaption of phylogenetic methods, linguists also adapt the computational methods. The approaches are used for the detection and explanation of linguistic evolution. The reconstruction of a tree for representing a relationships is one parallel. Phylogenetic methods can be used for creating a language tree automatically showing the classification between the languages. This method became widely used and is currently a popular representation for the classification. Other parallels between biology and linguistics can be drawn. One parallel is horizontal gene transfer and borrowing.

Words can be borrowed into other languages. Such words are loanwords. The words undergo an adaption and are then fully integrated into the borrowing languages. The borrowing takes place between two languages who are in contact with each other.

Horizontal gene transfer is the transfer of a gene from one species to another.

The gene is transferred between two organisms without sexual reproduction. The process of borrowing and the horizontal gene transfer take place between two languages or organisms. Both undergo the same process of transformation. Therefore, the processes can be compared. The various approaches for detecting horizontal gene transfer might also be used for the detection of borrowing.

In this paper, a tree-based approach for detecting horizontal gene transfer is introduced and applied to linguistics. The methods are not yet implemented in linguistics and the approach will analyse if the phylogenetic methods are an adequate method of detecting borrowing within languages.

The second chapter is an introduction to loanwords and their linguistic background. The process of the adaption is explained and theories for the adaption are introduced. A small statement on automatic loanword detection is made to show the requirement of such an approach. The third chapter is an introduction to phylogenetics. The theoretical background and the fundamental ideas of phylogenetics are introduced. Afterwards, the two main concepts are explained, namely phylogenetic trees and phylogenetic networks. The fourth chapter states the usage of phylogenetics in historical linguistics. A python library called LingPy is introduced. It is the only software package where phylogenetic methods are implemented for the usage in linguistics and it enables the detection of borrowing. In the fifth chapter the theoretical approach is introduced by explaining some background on phylogenetic methods and the methods themselves. The theoretical approach is also compared to that of the software package LingPy. In the last chapter, the automatic approaches are compared to a manually constructed database. The database is explained and a comparison is made.

2 Linguistic background of loanwords

Language is one of the biggest and fastest changing system in humanity. Due to the contact of languages, the languages change rapidly and develop further over years. In first place, people contribute to language change. Those changes are studied in historical linguistics. Language can change in different ways: lexical, morphological, phonological and semantically. The changes occur over time and affect the language. Most of the time, the language will adapt such changes and after a while they are assimilated (Bußmann, 2008).

One form of language change could happen through a change within one language, while people adapt easier word forms or change complex forms to simpler forms. This can be illustrated by looking at verbs in German and English. Verbs are grouped into strong and weak verbs. Strong verbs are inflected differently, whereas weak verbs always have the same inflection. In the process of simplification, the weak inflection is adapted for some strong verbs and the old form of the verb is replaced by the new form (Bußmann, 2008; Delz, Layer, Schulz, & Wahle, 2012).

Another type of language change is the adaption of new words. Language contact or the change of the living conditions is the reason for the adaption. This adaption is reflected to as borrowing. A language borrows linguistic expressions from distinct languages. In most cases, the borrowing language does not possess a word for a specific description or concept and therefore needs to borrow it from a language which already has a word for this description or concept. Political, cultural, social and economic developments can the reason for this need to surge. The import of new products, forms of sport, technology or economic strategies can be named. The borrowed words are classified as either loanwords or foreign words (Bußmann, 2008).

This paper will focus on loanwords. It is important to distinguish a loanword from a foreign word. Therefore, I will firstly introduce a definition of loanwords which will be used throughout this paper. Afterwards, I will compare loanwords to foreign words and even to cognates and provide some examples on loanwords and the reasons for their adaption. Different processes are applied to adapted words. Those processes are described in section 2.2. Afterwards, theories are explained which can be used for analysing loanword adaption. This all leads to the question if loanwords should be detected automatically. Several reasons are listed and explained in the last section of this chapter.

2.1 Loanwords

As stated above, loanwords are words adapted by one language from another one. Bußmann (2008) uses this definition:

Definition 2.1 *Entlehnungen einer Sprache A aus einer Sprache B, die sich in Lautung, Schriftbild und Flexion vollständig an die Sprache A angeglichen haben.*

In other words, a loanword is borrowed from language B by language A, whereas the word is phonological, lexical and inflectional fully adapted in language B. This can also be illustrated:

(1) *language A* \implies *adaption* \implies *language B*

The illustration in (1) shows that borrowing is a process of integration of a foreign word. Mostly, this process takes place between languages of the same time period. The loanword is a widely used word of its source language. Some speakers of another language do not have a word with the same meaning and are therefore borrowing it from the source language. Within the borrowing language, the word is adapted and integrated into the language. The speaker does so in the most comfortable way and will pronounce the word as he would if it were of his mother tongue. The original word changes phonologically, lexically and also in inflection. After these steps, the word becomes a loanword.

This is distinct from a foreign word of language A which is used in language B without adaption. Cognates are etymologically related words from different languages that are derived from a single common ancestor. The following graphs illustrate the differences between the processes. All three illustrations should give a clearer process of the different processes.

(2) Foreign words:
language A \implies *language B*

(3) Cognates (the source language indicates the same language):
source language \implies *language A*
source language \implies *language B*

Cognates derive from one single form present in an ancestral language. Two cognates can occur in the same language or in different ones, but they always have one single common ancestor. This process develops over time and the establishment in the languages must be during the same time period. For example, Latin words can be found in German and English.

- (4) a. Latin: *discus* (meaning: disc, a circular plate) \implies German: *Tisch* (table, a plate with legs)
- b. Latin: *discus* (meaning: disc, a circular plate) \implies English: *Dish* (is still a plate)

Both words have the same ancestor which is the Latin word *discus*. Most of the time, the relatedness is obvious, but the meaning might have changed. This is the case for German, where the meaning disc has changed into the meaning of table.

This is not the case for loanwords and foreign words. Loanwords and foreign words have a direction, they always start in language A and end up in language B. The major difference is that a foreign word is adopted, whereas a loanword is integrated into language B. By adoption, the word will not change and it will stay similar to its origin word. The meaning of the word, being a loanword or a foreign word, stays the same. By integration, an adaption takes place and the word is integrated into the language via customizing its phonology, inflection and typeface. The words are adapted or adopted because the language needs a word to describe a particular meaning or for other reasons.

Haugen (1950) defines borrowing as a result of language mixture, where the reproduction of linguistic patterns in language A are previously found in language B. If a word is borrowed, it is modified to fit in the borrowed language. After this modification, a native speaker of the source language may not recognize the borrowed word at all. This modification happens due to linguistic patterns. The linguistic patterns of the source language might not be represented in the borrowing language, therefore the word changes. The change is done in small steps until the word fits into the language. It depends on the borrowing language how much a word will change. Therefore, borrowing is a process not a state (Haugen, 1950).

The process of borrowing is also a historical one, because the words are integrated in a language and adapted over time (Haugen, 1950). The whole process of borrowing consists of three parts:

1. The borrowing part, where the word is chosen from a foreign language and used in the borrowing language.
2. The process of adaption and integration, where the word is adapted into the borrowing language.
3. The end result being the loanword.

The borrowing itself, is a short process which takes place between two languages being located in the same historical time period, as stated above. The process of

adaption and integration is a longer process which takes place over a certain time period. The loanword itself is the final result of both processes (Haugen, 1950). After a while, these words are no longer seen as loanwords. They belong to the language like every other native word and are perceived as such by the language community. Along with the process of borrowing, the phonological change takes place. Haugen (1950) claims that the native speakers imitate the foreign sound sequences while modifying the sound sequences according to the patterns of their native language. This is the “process [...] in which the speaker substitutes ‘the most nearly related sounds’ of his native tongue for those of the other language” (Haugen, 1950, p. 215). Next to the phonological, there is also a grammatical process. The native speaker modifies the word according to the grammar of his native language. The words need to fit into a category. For example, if a verb is borrowed it may be integrated in one verbal category of the language and all borrowed verbs may end up in this category. The same happens for nouns and their gender. The borrowed nouns are integrated in the gender system of the borrowing language (Haugen, 1950).

Most of the time, the speakers who borrow words are bilingual speakers. They take words from their second language and use them in their mother tongue. Some of those words are possible candidates for substained integrations in the language, others are not powerful enough to be integrated. When a word is useful, monolingual speakers start to use it and the word is adapted into the language (Yip, 2006; Haugen, 1950; Peperkamp & Dupoux, 2003).

People who know a second language mostly live close or on the border to another country, have relatives in other countries, or business partners. They came in contact with other languages because of those circumstances and without knowing the language before.

Others learn a second language during their education or because of travelling and meeting other people. Schools and universities open up possibilities for studying and living abroad and as do other exchanging programs.

In former times, this was more difficult than nowadays. It was not naturally to learn a language in school or in other institutions for educational reasons. They came in contact with other languages and cultures because of the above-mentioned circumstances, like moving, living close to another country or having relatives abroad. Most of them live as nomads until they settle down. During this time, the people meet up while moving and could have exchange experiences, utensils or other things. In this time, the oldest loanwords are adapted. After they settle down, the people did not stop moving and travelling. They get their inspiration from other countries, their cultures and most important their religion.

Religious terms are one of the most adapted words. If the population integrates the religion, they also integrate the corresponding terms. This is an easy way to keep the meaning of the words and more important their religious function. As time goes on, the living conditions changes and time opens up more possibilities. The contact between countries and their languages becomes easier with every step in time. Ships, locomotive, cars and other vehicles made it possible to manage longer distances and easier to import new things for working, living, eating and so on. The invention of technology, like the telephone, radio, television, the internet and much more are leading to even more contact between countries and their languages. People are adapting new words instead of creating a new one for their own language. The process of borrowing is easier and more efficient than the creation of a new word.

The adaption is not only due to the moving of the people, but also to the history of language. Most loanwords were adapted from old languages like Latin, Greek and others (Joseph & Janda, 2003). Latin was the language used in the church and therefore, the language of educated people. Later on, a tendency of loanwords coming from neighbouring states could be made out. For example, German has a long list of loanwords from French (Volland, 1986). The same holds for English, it also incorporate many loanwords from French (Baugh, 1935). People who knew French or came into contact with the French language adopted words into their native language. Additionally, French became the language of the upper class and educated people. Most of these words came from a cultural, religious or economic background.

Nowadays, most words come from the technical, economical and scientific fields. They are transferred through different types of media like newspaper, radio, television or the internet. Most of these words arise in the English language because innovations are made in the United States or in companies where English is the common language. This is due to the fact that English became the world language, spoken by many people all over the world and thought in school as the first foreign language. English became an international language and because new developments receive an English name and description to be sold all over the world. These are adopted or adapted in other languages. In those cases, it is hard to differentiate between loanwords and foreign words. Here the historical process can be used for the identification of the loanwords. One example is the word *Google*. It is a proper English name for the best-known search engine. Then this name was used for identifying every search engine in the internet and it became a fixed term in other languages. Everyone automatically links *Google* to an internet search engine. In German, the name has been integrated into the language over years. The noun *Google* is officially included in the lexicon of the

German language. Further, *Google* is also a verb: to google sth. meaning to search something with the help of the search engine Google. The verb is also inserted in the German verb system with its corresponding inflection.

- (5) a. German: Er googelt das Wort.
English: He googles the word.
- b. German: Er googelte das Wort.
English: He googled the word.
- c. German: Er hat das Wort gegoogelt.
English: He has googled the word.

The different inflections of the verb indicates that to google is a weak verb in the German language. The *t* in *googelt* indicates the inflection for the third person singular present tense which is similar the *s* of *googles* in English. In example (5-b), *te* is the inflection for the past tense similar to *ed* in English and *hat* plus *gegoogelt* is the pendant to *has googled*.

It is obvious that speakers borrow words from other languages because they do not have words which carry the same meaning in their native languages. This is due to language contact.

Language contact is caused by speakers of one language which come into contact with speakers of a different language generally due to moving. The circumstances and reasons under which a word can be borrowed induce speakers to adapt a new word for expressing a specific meaning in their native language. For example, in earlier times Germans adapted words from the high class in the French society. The words sounded classy and they used them to establish a gap between themselves and the lower class of the German society. The French words are adapted in the German language, but they do not replace the German words.

- (6) a. French: Chaiselongue (meaning: a specific couch) - German: Sofa (meaning: specific kind of couch)
- b. French: Trottoir (meaning: sidewalk) - German: Bürgersteig (meaning: sidewalk)

Other words are adapted from French into German because the object has no words in German with this meaning (Volland, 1986).

- (7) a. old French: raisin, rosin - German: Rosine - English: raisin (meaning in all languages: dried grape)
- b. old French: pastee - German: Pastete - English: pie (meaning in all languages: a special kind of pie)

Nowadays, most loanwords come from economy and sports. The words have the same meaning in the borrowing language and are adapted for representing the object. Most of the words change according to the borrowing language. The confusion between loanwords and foreign words increase for words which are borrowed in younger times. Other words are so much integrated in a language that the origin of the words is almost forgotten. Here are some examples from French loanwords in English (Kemmer, n.d.):

- (8) a. Old French: parlement - English: parliament (meaning: comes from parler-to speak and is now an institution)
b. Old French: saumon - (Middle) English: salmon (meaning: the fish and the food)
c. Old French: mireor - Middle English: mirour - English: mirror (meaning: a surface that reflects the image)

As one can see in these examples, most borrowed words are nouns. It is also common that mainly nouns are borrowed from the cultural background of other languages. For example, gender is less likely to be borrowed into another language (Joseph & Janda, 2003). This means that a language like English which only has one gender will not borrow the three gender system which is present in the German language. The same can be said about affixes, articles, inflections and even particular sounds (Haugen, 1950). It is also less likely to borrow words from the basic vocabulary (Joseph & Janda, 2003). Swadesh (1955) made a list of words which are non-cultural and universal. Most of these words are present in each language. The first list contained about 100 items and was later on modified by Swadesh (1955). The 100-words swadesh list can be found in the appendix. He inserted words which according to him should be contained in the list. Those words are cultural concepts like mother and father, numerals, natural objects and animals (Swadesh, 1955). The words in this list are said to be resistant against language evolution, especially borrowing, and are contained in most of the languages. Sometimes, words from the basic vocabulary can be borrowed. This is the case for the English word *mountain*. This word is a loan from the French (Joseph & Janda, 2003).

- (9) Old French: montaigne - English: mountain

Other Germanic languages use a word of a different stem for *mountain*. The German word is *berg*, the Swedish word is also *berg*, in Dutch it is *bjerg* and the Afrikaans word for *mountain* is also *berg*. There are more words from the basic vocabulary in English which are borrowed from French (Joseph & Janda, 2003):

- (10) a. Old French: face - English: face (meaning: the front part of the head)
- b. Old French: estomac - English: stomach (meaning: an organ that stores food)
- c. Old French: riviere - English: river (meaning: a stream of water)

The words in old French originate in the Latin language. The detection of loanwords in the basic vocabulary is challenging. Mostly it is not clear which words are loanwords and which are not. As said before, one can identify these loanwords by means of historical processes. If the historical process is known, loanwords can be detected.

2.2 Processes in loanword adaption

Words which are adapted in a language undergo a process during the adaption. The words change with respect to the system of the borrowing language. As stated above, most of the time bilingual speakers introduce the word in their native language. Other speakers of the language pick up the word. During this process, the word is adapted in the language and different processes of change take place during the adaption.

The major changes take place in the phonology of the word. Peperkamp and Dupoux (2003) called this change which is applied to words which are adapted in a language, transformation. Speakers use these transformations to convert sounds which are not present in their native language into well known sounds. “Words from a source language that are ill-formed in the borrowing language are thus transformed into well-formed words” (Peperkamp & Dupoux, 2003, p. 367). This can be seen as transformation or as a type of repair strategy by the speakers. This repair strategy can take on the form of changing the sound, deleting the sound or adding a sound. The most common strategy is the change of a sound. Most speakers choose the sound closest to in their native language (Haugen, 1950). The phonological distance between two sounds plays a crucial role, whereas the sound in the native language with the smallest distance to the sound in the source language is chosen. There are several examples for the change of sounds (Peperkamp & Dupoux, 2003; Yip, 2006):

- (11) a. Korean listeners: [li:d - ri:d] - English: to lead
- b. Cantonese listeners: [rejz - lej si:] - English: raze

In the example (11-a), the discrimination between [l] and [r] is shown. Korean listeners are sensible of this difficulty and are therefore changing the sounds from

[l] to [r] (Peperkamp & Dupoux, 2003). In the other example (11-b), Yip (2006) shows the change of the initial [r] and the final sound [z]. Both of these sounds are not present in Cantonese. The special thing about Cantonese is the property of having a so called interlanguage, namely Hong Kong English. In Hong Kong English the word is [ɹeɪs]. The initial sound [r] changed to [w] and [l], both being alveolar approximants. The final sound [z] changed to an [s] (Yip, 2006). According to Yip (2006), the devoicing of the final sound [z] is an influence of the native language Cantonese, because Cantonese does not have voiced fricatives. Therefore, all English [z] sounds are replaced in Cantonese. The change from Hong Kong English to Cantonese is smaller than from English to Cantonese. Having this interlanguage, the change is not as big as without such an intermediate step. There are also some examples of the reduction of sounds (Peperkamp & Dupoux, 2003; Yip, 2006):

- (12) a. White Humong: [pe.si] from the English word *pepsi*
 b. Cantonese: [sipin] from the English word *spleen*
 c. Cantonese: [kip] from the English word *creep*

In the example (12-a) shown by Peperkamp and Dupoux (2003), the [p] is lost during adapting *pepsi* from the English language into White Humong. In both Cantonese examples shown by Yip (2006), the central sounds [l] and the [r] are lost. As one can see, if an [l] is in the initial position of the loanword, the sound changes and if it is present in the middle of the word, the sound is lost.

The opposite of reduction is addition. This might also happen during the process of adaption.

- (13) a. Japanese listeners: [kurimu] from the English word *cream*

The Japanese listeners break up the consonant clusters by adding another vowel, in this case it is a [u] (Peperkamp & Dupoux, 2003; Olah, 2007).

There are also other processes or transformations which occur during the adaption of words. One of those is the shift of sounds or accents.

- (14) a. French listeners: *télévision* from the English word *television*

French listeners have a contrast in stress compared to other languages. The English word is stressed on the syllable *vi*, whereas the French word is stressed on the syllable *sion*. Mostly, they adapt the word and instead of changing a sound, they shift the stress (Peperkamp & Dupoux, 2003). This transformation is less common than changing or reducing sounds and called shift. Haugen (1950) defines another kind of shift, namely loanshift. He suggests a shift as change in the usage of native words. This might happen for synonyms. For example a language

A has two words a_1 and a_2 with the same meaning and both words overlap with the word b_1 from language B. This overlap can lead to the adaption of word b_1 and the displacement of one of the words a_1 or a_2 .

Another transformation in the adaption process is the process of insertion of the words to the grammatical system of the borrowing language. This happens parallel to the phonological transformation. Haugen (1950) claims that the borrowed words also need to fit in the grammar of the borrowing language. This is a kind of process which also needs to be taken into account while talking about loanword adaption. As I said before, nouns are the most common words to be borrowed. The gender of nouns can be divided into three groups, known as feminine, neuter and masculine. While adapting a noun, one needs to assign one of the three gender to the loanword. In a language like German, where all three genders have a particular article, all loanwords are most of the time inserted into the same category (Haugen, 1950). Only in certain cases this strategy will change. This depends upon the gender system within the source language and the borrowing language. For example, if a German word is borrowed into the English language, the choice of the article is obvious, but the noun still needs a gender for assigning pronouns to it. English has an easier and clearer system than German and therefore words coming from German into English might end up in the neuter gender class in English. On the other hand, if we look at the other way around, it might not be as easy. The German gender system is richer than the English and therefore the gender for adapted words might be chosen more cautious. This process is distinct in most languages, some have a so called default gender or article and others don't.

Another case is the adaption of orthographical forms. There are cases where the plural *-s* in English is borrowed with the stem of the word into the other language (Haugen, 1950).

(15) a. English: *car* - Norwegian: *kars*

The English word *cars* is borrowed into Norwegian with the plural *-s*. The loanword is *kars*. The plural of the Norwegian word is *karser*. This is a phenomenon which might also appear in other languages and can be seen as a kind of word plus grammar adaption.

The last change described which can occur parallel to the phonological change, is the change in orthography. Spelling has an influence on the adaption of words. A study by Vendelin and Peperkamp (2006) shows influences of orthography in loanword adaption. Also Haugen (1950) claims that the process of borrowing

has an influence on the spelling of the word. This influence can be considered from two perspectives. The first is the situation where the word is adapted in the language via phonological contact. In this case, the pronunciation is known but no written form is present. The speakers write the word as they speak it. Therefore, the word is written as the speakers would write it according to their native language. The original form can still be identified via pronunciation but not necessarily from its orthography. In the second case a written form of a word is adapted into a language. Here, the speakers pronounce the word as they would according to their native language. The pronunciation has no relation to that of the original word. The original word can be identified because of the written form and less from its phonological form.

2.3 Theoretical approaches to loanword adaption

Approaches to loanword adaption are used in different studies. Most use a well-known theory as a framework for describing the adaption of words and the correct transformation of these words into the borrowing language. As commonly the case in science, different scientists have different opinions and therefore different approaches for the adaption. The most common theories are rule-based or constraint-based systems. Whereas the constraint-based system mostly ends up in a framework of Optimality Theory. A less frequent theory would be the one of speech perception.

A rule-based system is as the name suggests a model with a set of rules which can be applied to the word. The rules describe how the adaption take place. The rules are fixed according to representations of similar words in the borrowing language. The rules are applied to each word which a language wants to borrow. This makes it difficult to expand the system or the model. Silverman (1992) gives an example of a rule-based system in his study. The rule-based system contains two levels, the *Perceptual Level* and the *Operative Level*. On the first level, the word or the input is parsed and interpreted as segments in the borrowing language. This process is based on constraints from the native phonological system and is acting as a filter for the input. If the native phonological constraints hold in the first principle, the second principle is applied. On the Operative Level, rules which I will elucidate after, are applied to the segments (Silverman, 1992; Jacobs & Gussenhoven, 2000). The segments, which are the output from the first principle, undergo phonological processes and are realized “in conformity with native prosodic constraints on syllable and metrical structure” (Silverman, 1992, p. 290). Silverman (1992) shows in his study examples of English loanwords in

Cantonese. The following is an example of the English word *shaft* borrowed into Cantonese. In the Perceptual Level the English input word is parsed and its segments interpreted in Cantonese. The *Perceptual Uniformity Hypothesis* serves like a filter to the native language (Silverman, 1992).

(16) *Perceptual Uniformity Hypothesis*

At the Perceptual Level, the native segment inventory constrains segmental representation in a uniform fashion, regardless of string position. The English word *shaft* is therefore parsed as [sɛf]. In Cantonese, fricatives and affricates may only appear in the onset position and not in the coda position, while in English they can appear in both positions. A process of occlusivisation is applied to fricatives and affricates in coda positions (Silverman, 1992). The process will formally look like this:

$$(17) C \rightarrow [-cont]/-]_{\sigma}$$

This rule is applied at the second level, namely the Operative Level. The rule will change the output of the first principle to [sɛp]. The adaption will look like this:

$$(18) \text{ original word} \longrightarrow \text{Perceptual Level} \longrightarrow \text{Operative Level } shaft \longrightarrow [sɛf] \\ \longrightarrow [sɛp].$$

As one can see, at the Perceptual Level the segments are parsed and it is a segment-by-segment representation. At the Operative Level, the rule comes into play and the phonological process triggers the change from f to p (Silverman, 1992).


The problem with a rule-based system is that the rules can lead to an incorrect output. This is due to the fact that rules are hard to change or to be added additionally. The rules are established according to the specific loanword phonology of a language. Therefore, rules need to be added for every specific loanword phonology (Jacobs & Gussenhoven, 2000). Additionally, the rule-based model only includes language specific rules. Therefore, every language needs its own rule-based system for the adaption of words.

The constraint-based system is the counterpart to the rule-based one. Mostly, the constraints are embedded in a framework of Optimality Theory (OT). Several studies are based on this system, like the ones from Rose (2012), Paradis and LaCharité (1997), Vendelin and Peperkamp (2004) and Moira (1993). In a constraint-based system, several constraints are defined and ranked. The input of the model, is the original word with its pronunciation in the source language. Moira (1993) argues that the constraint-based model only needs a set of ranked

constraints which are either universal constraints or motivated by the native language. The adaption or the transformation of the loanwords is made by applying the constraints to the possible representations of the word in the native language. The “set of ranked constraints examines the set of all possible output representations for a given input, and assigns degrees of well-formedness to these” (Moira, 1993, p. 263). Each borrowing language has such a set of ranked constraints depending on its phonology. The highest ranked constraint must be satisfied while going through the set. The set of ranked constraints can be seen as a list of transformations which need to be applied to each possible representation of the word step by step. The representation which fulfils the most constraints is the optimal representation. An optimal representation is relative which means that an optimal representation in one language can be suboptimal in another one. It can also happen that two constraints are violated by the same representation. In this case, the representation which violates less constraints is chosen (Moira, 1993).

An account of Optimality Theory describes the grammaticality of a word or a representation and is represented with the help of a tableaux.

(19)

Input: //	constraint 1	constraint 2	constraint 3
a.  representation 1	*	*	* *
b. representation 2	*	* *!	*

The columns represent the constraints and the rows the different representations of the loanword. The first constraint is the highest one in the ranking, followed by the second constraint and the third constraint (constraint 1 » constraint 2 » constraint 3). It is also said that constraint 1 dominates constraint 2 and constraint 2 dominates constraint 3. The representations of the loanwords can fulfil the constraint or violate the constraint. There can also be more than one violation. The stars or asteriks represent the number of violation for the representation and the corresponding constraint. If a representation does worse than another representation on the same constraint and this constraint distinguishes the representations, an exclamation mark indicates the worse one. Once a representation gets an exclamation mark, it will stop being a candidate for the optimal representation. The grey colouring visualizes the suboptimal representations. The optimal candidate is shown via the pointing finger.

Moira (1993) explains a constrained-based system within an OT framework. First of all, Moira (1993) defines some constraints which represent and define well-

formed words in Cantonese. The process is as follows: the input is perceived from the English language, “this perceived input is then checked by a group of ranked constraints that are independently motivated for native Cantonese, and minimal adjustments are made to produce an output that is optimal with respect to the constraints. Prominent among the constraints are (i) a set of syllable-structure conditions, (ii) a strong preference for matching the input as closely as possible, and (iii) a tendency towards bi-syllabic Minimal Words” (Moira, 1993, p.261). There is also a set of possible candidates which consists of different representations of the word in Cantonese. For example, the English word is *cut* and for the set of candidates Moira (1993) chooses $k^h a t.$, $k^h a. t^\square.$, $k^h a.(t)$. Additional candidates can be added infinitely. The \square indicates an empty node which is realized as an epenthetic segment. The parentheses show an unparsed segment which is deleted in the representation. The ranked constraints are checking each candidate and rejecting non optimal candidates. The constraints and the tableaux for the OT analysis are stated in Moira (1993). The result of Moira (1993) is that for the English word *cut* the optimal Cantonese pronunciation would be $k^h a t.$. The word English word is adapted without a change into Cantonese as Moira (1993) states in her paper.

In this framework and in the other constrained-based systems of Rose (2012), Paradis and LaCharité (1997), and Vendelin and Peperkamp (2004), the Optimality Theory distinguishes between the representation of the words during the application of the constraints. This leads to the most optimal representation which is adapted into the borrowing language. With the ranking of the constraints, the optimal transformation of a loanword can be found. The optimal representation does not mean that it is also the right one. Yip (2006) compares the optimal representation with data taken from the Cantonese language. The comparison shows that the optimal representation is mostly the right one and the one which was actually adapted in the language.

The field of speech perceptions differs in its point of view on the adaption of loanwords. The two frameworks explained above are developed with respect to constraints or rules representing the phonology of the native language or the loanwords. A framework of speech perception claims that the adaption happens during perception. In the perception, “the phonetic form of the source words is faithfully copied onto an abstract underlying form, and [...] adaptations are produced by the standard phonological processes in production” (Peperkamp & Dupoux, 2003, p. 368). Peperkamp and Dupoux (2003) claim in their study that non-native sounds can be decoded in the perceptual process and the words can be repaired. Repaired in the sense that the input word is ill-formed in the

borrowing language and gets adapted via repairs of the sounds to a well-formed loanword. “The process of decoding [...] maps the non-native sound patterns onto the closest native ones” (Peperkamp & Dupoux, 2003, p. 369). Compared to the other frameworks, in the framework of speech perception adaption takes place in perception not in the production process. Peperkamp and Dupoux (2003) claim that this explains a phenomenon in Cantonese. Cantonese lacks the voiced fricative [v], but in loanwords the sound changes into the sound [w] and not into an [f]. In the framework of speech perception, the change from [v] to [w] is made because in the underlying form [w] is the closest sound to [v].

There are different approaches to loanword adaption. Each of them having their own advantages and disadvantages and different points of view. The adaption of loanwords is a broad field and can be represented in more than one framework. This is the case for most language phenomena.

2.4 Automatic loanword detection

Loanword adaption is one side on the field of loanword studies, the other is loanword detection. In loanword adaption, the process of loanword transformations and different theories are described which can help to adapt loanwords. But what about loanword detection? How can we find loanwords which are already adapted by a language. I argued before that the original words undergo a phonological transformation process to be adapted in the borrowing language. Sometimes, the loanwords are so much integrated in a language that the speakers do not know that the word is a loanword at all. But how do we know which words are loanwords? In historical linguistics, the history of languages and the origin of words are studied. With the help of the historical process, loanwords can be detected. The reconstruction of the history of a word is time-consuming and needs to be done for each word individually. If each word needs to be reconstructed, one will find loanwords, foreign words, cognates and native words. This costs time and is not effective. Nowadays, databases are constructed which represent historical processes or loanwords.

Another advantage nowadays, are computers and algorithms. Although there are less algorithms for loanword detection, it will be a big help for the detection of loanwords in languages. The detection can be made with the help of language databases. These databases include the same words for several languages and give a great background on the vocabulary of the different languages. Such databases can be found widely over the internet. With the help of algorithms, loanwords can be detected automatically. This will be an efficient method for detecting loanwords because the algorithm can find a great amount on loanwords in a small

period of time. If one would search each loanword manually in the vocabulary of the languages, this would take much longer. Additionally, one needs to be an expert in the language to read and understand the words. Compared to the manually search, more loanwords can be found automatically. The algorithms may work more precisely than humans and make less mistakes. A more significant connection can be drawn between the languages and the loanwords, and between the source languages and the borrowing languages. Similarities can be found and even the language contact can be reconstructed. The previous presented studies would gain significance and the processes and theories represented can be applied to more data. With more data, the theories would reach more precise results and the results would strengthen the theories.

Automatic detection of loanwords would bring the studies on loanword detection and on historical linguistics one step further in language evolution. It might only be a small part, but an important one. Language contact can be explained in more detail, language evolution will reach another level in the explanation of language change and language contact and evolutionary events like borrowing can be detected more easily.

3 Phylogenetics

Phylogenetics is a field of study and analysis of evolutionary relationships between different groups in biology and bioinformatics. In biology, these groups can be different classificational units like organism families, genera or species, but also individuals within a species. In linguistics, phylogenetics can be used for detecting evolutionary relationships between languages or language families and different concepts of words.

The fundamental idea goes back to Darwin, who constructed a *tree of life* for representing the relation between organisms. He was one of the first who classified organisms according to their genealogical development and relationship. This genealogical development is now known as *phylogeny*. Darwin illustrated the phylogenetic order by using the symbol of a tree with one trunk that branches out into different directions. The idea of reconstruction phylogeny as a tree is still present in phylogenetic systematics (Lecointre, 2006). Darwin reconstructed his tree of life through his knowledge and intuition. The idea of using a tree for the representation was developed further and explicit methods and ideas for the reconstruction evolved.

The basic ideas of phylogenetic systematics were introduced by Willi Hennig. A German entomologist who began developing *phylogenetic systematics* before World War II (Wiley & Lieberman, 2011). During the development of phylogenetic systematics, “some of these ideas remain basic to the discipline [...], while others have to be discarded [...]” (Wiley & Lieberman, 2011, p.2). Those basic ideas are the so called foundation for the systematics. Additional studies on phylogenetics inspired Hennig’s ideas to further developments. These ideas contribute to a bigger theory of phylogenetics and to formally described algorithms and models. Wiley and Lieberman (2011) stated that “phylogenetics is a dynamic discipline” (Wiley & Lieberman, 2011, p.2), the development is not completed and phylogenetic systematics are still studied, also in different fields.

Both, Darwin and Hennig, reconstructed their trees to represent evolutionary developments and relationships. A tree can not only be used for the representation of evolutionary events in biology, but also as for example language evolution. Phylogenetic methods can be used to describe different evolutionary phenomena of language history.

I will first give an overview on phylogenetics and the theoretical background with its representations of trees. Afterwards, I will introduce some methods within the field of phylogenetics as well as their technique to detect different phenomena within trees. In the last section, I will compare phenomena in biology and historical linguistics, showing similarities and differences.

3.1 Theoretical Background on Phylogenetics

The basic ideas in evolution go back to Darwin. Darwin did not know anything about genes, the structure of DNA or other organisms which are responsible for inheritance, but he did know that inheritance is present. He knows that “organisms resemble their parents; that the variation in the appearance of organisms within a single species is heritable; and that more organisms are produced each generation than can possibly all survive and themselves reproduce” (Eldredge, 2005, p.69). His grandfather, Erasmus Darwin, had published his work *Zoonomia* in 1801. This work already showed basic approaches towards evolution. In one of his notebooks, Darwin quoted phrases and passages from the *Zoonomia* and began to write his own thoughts next to them (Eldredge, 2005). The idea of evolution and *natural selection* was established. *Natural selection* is a biological mechanism in genetics. The process selects for adaptive genes while maladaptive genes are selected against. Therefore, it regulates the transmission of adaptive genes to the next generation. To Darwin the environment appeared to play a crucial role in natural selection. He studied the evolution of animals and plants, also taking their environment into account. Although, Darwin never used the term *evolution*, his thoughts on it has already arisen. Darwin uses the terms *transmutation* or as later on, in his work *Origin of Species, descent with modifications* instead of evolution. The term evolution came later in his life into vogue (Eldredge, 2005).

While his thoughts on evolution arose, Darwin thought about populations and individuals. While thinking about forming new individuals via inheritance and the dying of other individuals, he came to a point where he thought about the occurrence and death of populations. He asked himself: What would it look like, if evolution were true? The answer to this question can be seen as the metaphor for a *tree of life* where the population is represented by the branches all going back into one trunk (Eldredge, 2005). New populations would evolve on older and thicker branches. The order depends on the parents and the transmutation of the populations. The term *Tree of Life* goes way further back and is actually a bible phrase (Penny, 2011). Therefore, Darwin did not introduce this term but rather taking it to represent his concept of a tree of life.

The first sketch of Darwin’s *tree of life* is shown in figure 3.1. Darwin himself stated that the tree looks more like a coral and should therefore be called *coral of life* (Eldredge, 2005). Nowadays, one would refer to this tree as an unrooted tree or network but not a rooted tree. Those terms are described in more detail below. Nevertheless, it is the first sketch of a hierarchical system and a visualisation for evolution which found its way to the field of phylogenetic systematics. The tree

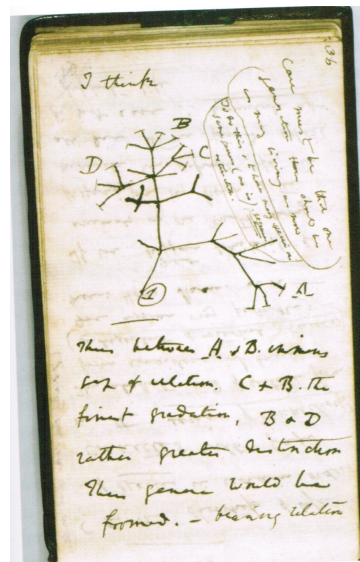


Figure 3.1: Darwin's first sketch of a tree

visualizes not only the evolution of populations but also gives a classification for them and their relationships to each other. Eldredge (2005) stated that the process of establishing evolution was turned around by the development of the tree. "Now we can see if evolution is true by generating evolutionary trees - and then checking if they hold up over time with the generation of new data" (Eldredge, 2005, p.105). This idea is a great scientific discovery and a cornerstone in the theory of evolution. It became a method for detecting all sorts of evolution between different organisms.

All of Darwin's thoughts on evolution are published in his work *Origin of Species*. The tree in figure 3.2 is the only illustration in the book. He used this figure several times for illustrating an expected outcome of evolution or as he calls it *descent of modification*.

Darwin's approach was the basis of the work of Willi Hennig. He introduced his approach on *phylogenetic systematics* which is nowadays called *cladistics*. In his work *Grundzüge einer Theorie der Phylogenetischen Systematic* in 1950 and later in his English work *Phylogenetic systematics* in 1966, he stated his basic ideas. Wiley and Lieberman (2011, p.2) summarized them in the following way:

1. The relationship that provide the cohesion of living and extinct organisms are genealogical ("descent") relationships.
2. Such relationships exist for individuals within populations, populations within species, and between species themselves.
3. All other types of relationships (e.g.: similarity, ecology) have maximum relevance when understood within the context of genealogical descent.

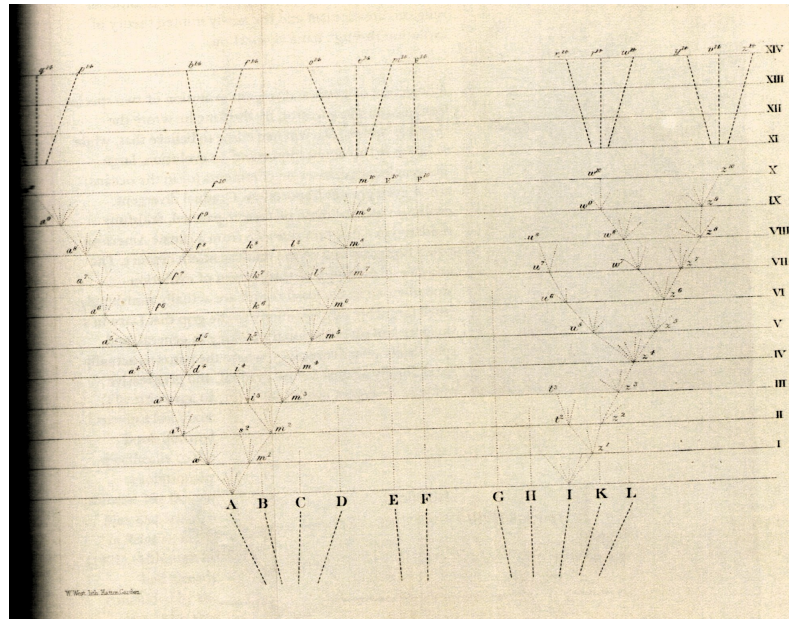


Figure 3.2: Darwin's tree in *On the Origin of Species*

4. The genealogical descent among species may be recovered by searching for particular characters (evolutionary innovations, synapomorphies) that document these relationships. Further, not all of the similarities that arise through descent are equally applicable to discovering particular relationships; some are applicable at one level of inquiry while others are applicable at different levels of inquiry.
5. Of the many possible ways of classifying organisms, the best general reference system is one that exactly reflects the genealogical relationships of the species classified.

Those basic ideas are a major part in the evolution of systematics. Hennig's theory on phylogenetic systematics is a modification of Darwin's theory and his tree of life. In cladistics, the organisms are ordered according to their common ancestor. Therefore, all organisms with a common ancestor are grouped together via the use of Darwin's descent of modification concept (Lecointre, 2006). This group is also called taxon. The taxon is associated with a proper scientific name according to the group of organisms. If there is no scientific name, the taxon receives another name describing the group. The plural form of taxon is taxa. The theory and practice comprising this describing, naming an grouping of organisms is called *Taxonomy* (Wiley & Lieberman, 2011). Organisms are chosen according to their relationship with each other and the tree is build by their diversity. The diversity is relevant for the evolution of the organism and the taxon. Each organism has a set of characters which is an observable attribute. The state of

a character is used for discriminating it within a group of organisms. For each character it is assumed to have similar states and that those states are *homologous*. Homologous meaning similar, where the states can be identical or differ slightly. Homologous can also have another explanation which is stated below. In cladistics, not all character states are homologous but certain resemblances might be convergent. Those cannot be detected immediately and can even contradict with other similarities (Lecointre, 2006). A data matrix is used for coding the characters and their assumptions which are that characters have similar states. With the help of the data matrix, all possible trees are build. The trees integrate the smallest number of evolutionary events needed by the data matrix for building the tree. “We keep only the most parsimonious tree - the one with the fewest number of evolutionary steps.” (Lecointre, 2006, p.16-17)

This is a more detailed description of the basic ideas stated above. All in all, these ideas and the technique behind them are studied and established further in the field of phylogeny. Phylogenetic systematics, as stated by Hennig, are focusing on trees and methods for building them. Darwin however already stated that his first sketch looked more like a coral. With this statement, he referred to what’s nowadays called networks. In phylogenetics, both trees and networks can be found. They are used for representing different evolutionary events and different techniques for describing evolutionary phenomena in biology.

3.2 Phylogenetic Trees

Hennig uses Darwin’s ideas of developing methods to reconstruct trees. Haeckel (1874) uses Darwin’s idea to create the first pedigree. He built a pedigree for different organisms, like plants, animals, bacteria, and even humans. The pedigree for humans is one of the most famous illustrations of Haeckel (1874).

In this illustration, the pedigree shows more similarities to a tree than Darwin’s illustrations in figure 3.1 and figure 3.2. Trees can be used to represent different relationships, while they keep the hierarchical structure of the represented organism. The relatedness between the organisms can be illustrated in a clear and intuitive way. Therefore, the concept of trees became famous for representing relatedness and dependence of different organisms.

As one can see, a tree consists of nodes and branches. Haeckel (1874) illustrated this in a pictorial way. Nowadays, the representation of trees is illustrated as in the example (20). The tree is a top-down tree, with the root on top and the nodes and leaves below.

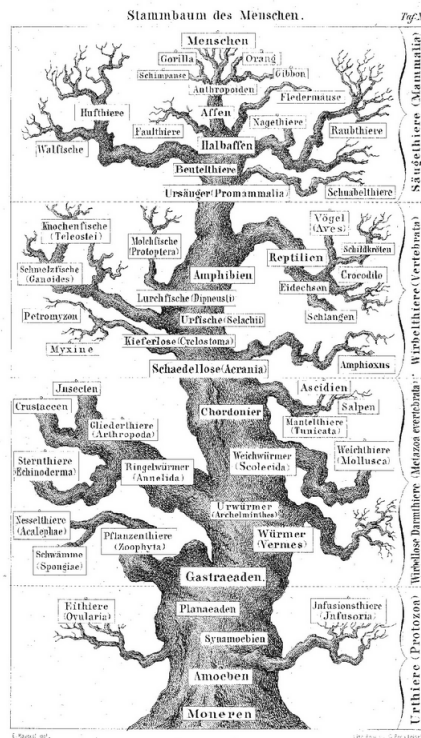
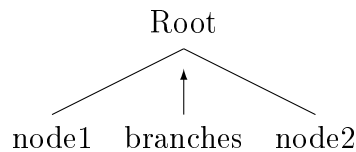


Figure 3.3: Haeckel's famous pedigree of man

(20)



Trees can be illustrated in different ways, the root can change its place. It can be found at the bottom as Haeckel (1874) and Darwin illustrated it or at the top as illustrated in example (20), but it can also appear on the left or the right side. Trees can be found in different fields of science where each one has its own main representation.

There is also a mathematics definition of trees which is stated by Lecointre (2006, p.21):

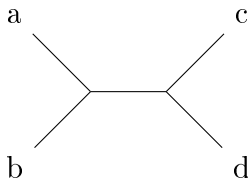
Definition 3.1 *A tree is a noncyclic, connected graph.*

All nodes are connected with their ancestor. The tree has to be noncyclic and all branches are at least binary branched. Binary branching refers to the fact that each node has two branches each pointing to one child and the fact that two nodes are only linked by one branch is called noncyclic. This definition of a graph is the basis of the phylogenetic trees.

There are two different kinds of trees, *unrooted trees* and *rooted trees*. The examples below are taken from Lecointre (2006, p.22) where for simplicity reasons only four taxa are used.

The advantage of unrooted trees is that “they are consistent with a limited number of rooted trees” (Wiley & Lieberman, 2011, p.101). In other words, there can be different rooted trees built out of one unrooted tree. The illustration in (21) represents an unrooted tree which consists of four taxa.

(21)



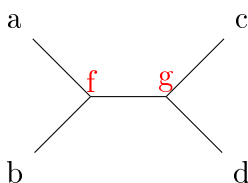
This is one of four possible representations of the tree. For the purpose in this section, only one is needed.

Huson, Rupp, and Scornavacca (2010, p.25) come up with a formal definition.

Definition 3.2 *Given a set of taxa χ , a phylogenetic tree T on χ consists of a tree $T = (V, E)$, in which all nodes have degree $\neq 2$, together with a taxon labeling $\lambda: \chi \rightarrow V$ that assigns exactly one taxon to every leaf and non to any internal node.*

In the definition, V indicates the set of nodes, E indicates the set of edges or branches and the phylogenetic tree indicates an unrooted tree. In example (21), the set of taxa would be $\chi = \{a, b, c, d\}$ and the tree would be the same graph without nodes shown in example (21). Per definition, each taxa in the set would be assigned to one node by change. One of the results would be the tree in (22), but the unrooted tree could also have another labeling. Taking the mathematical definition in 3.1 into account, the tree needs to be a noncyclic graph and connected. All nodes in (21) are connected with each other, but what about being noncyclic? I stated above that noncyclic in the sense of Lecointre (2006) means two nodes are linked by one path. In example (22), I labeled the inner nodes in red which do not assign a label according to definition 3.2.

(22)



If the inner nodes are labeled, the illustration of a binary tree gets clearer. In a binary tree, each node is connected with two children. Both inner nodes, f and g , are connected with their corresponding children. The children of f are a and b and the children of g are c and d . Additionally, both inner nodes are connected

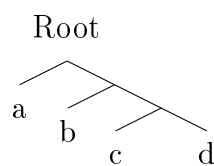
with each other. This is the reason, why the unrooted tree is not binary branched, but it is noncyclic in terms of linking two nodes with one path. Therefore, the unrooted tree in (21) might not be binary branched but it fulfils both definitions.

An unrooted tree can be transformed into a rooted tree, whereas each taxon in the set can be the root. Huson et al. (2010) also come up with a definition for rooted trees:

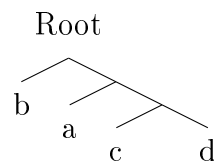
Definition 3.3 *Given a set of taxa χ , a rooted phylogenetic tree consists of a rooted tree $T = (V, E, \rho)$ and the taxon labeling $\lambda : \chi \rightarrow V$ that assigns exactly one taxon to every leaf and non to an internal node. All nodes, except ρ , must have degree $\neq 2$.*

It is the same than for the unrooted tree. The taxa set $\chi = \{a, b, c, d\}$ includes all taxa and they are assigned to the nodes of the raw tree. The trees are different depending on which taxon is the root. Here are all possible rooted trees stated, resulting from the one represented in (21).

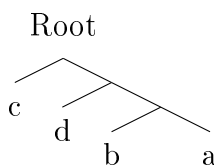
(23) a. The tree is rooted on a:



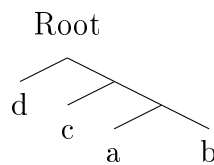
b. The tree is rooted on b:



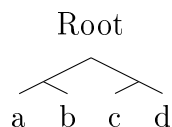
c. The tree is rooted on c:



d. The tree is rooted on d:



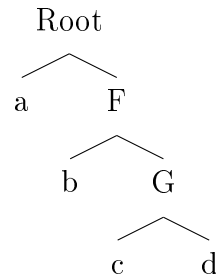
e. The tree has a midpoint root



Depending on the root, the trees change. Lecointre (2006) states the idea from Hennig, where the tree should be rooted on the outgroup. Depending on the outgroup, the tree is built differently. There are methods for constructing trees and choosing the optimal one. Those are described later on. Again, we take the

definition in 3.1 into account. The rooted trees are connected graphs and they are noncyclic. But what about binary branching? Here we can see that each node is connected to two children. The root and the inner nodes (capital letters) are taken into account, too:

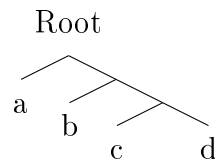
(24) a. The tree is rooted on a:



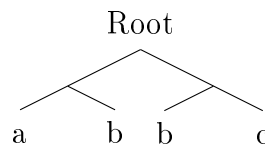
The representation in (24) gives us a clear picture on a binary branched tree. Each node, namely the root and the inner nodes B and C, are connected with two children. Therefore, the tree is a classical example of a binary tree.

Most rooted trees are used to represent a *species tree* or a *gene tree*. Those are two specific terms in phylogenetics. The species tree represents the evolutionary history of an organism, whereas a gene tree represents the evolutionary history of its genes.

(25) a. species tree:



b. gene tree:



The species tree of the organism is different from the gene tree. This indicates that the evolutionary history of an organism might differ from the one of its genes. Within the gene tree, different evolutionary events can happen which cause the gene tree and its species trees to be distinct. Those evolutionary events can be the duplication of genes, the loss of genes or the transfer of genes. The gene tree can be mapped and compared to a species tree for indicating the difference in their history. A gene tree can be displayed within a species tree (Huson et al., 2010).

Another reason for the usage of gene and species trees is the relation of two or more organisms to their ancestor organism. The ancestor organism and its corresponding evolutionary history would be represented within a species tree. The inner nodes of the tree represent the speciation of the descendant organism. Each

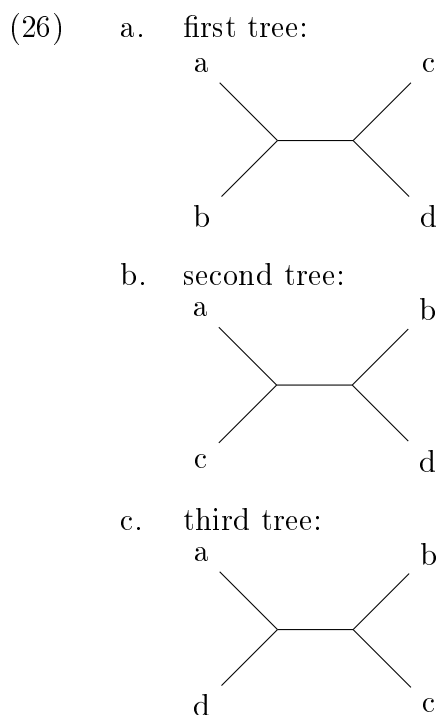
descendant organism has its own tree to represent its genes. The two or more gene trees can be mapped to each other to a bigger gene tree for representing their common history. This bigger gene tree can then be displayed within the species tree of the ancestor organism to compare the history. This method is used for comparing the speciation events and time of the speciation of the descendant organisms (Wiley & Lieberman, 2011).

Multiple gene trees can also be mapped to each other to form a single tree representing the species tree. This is done if no species tree is currently present or cannot be computed.

3.2.1 Reconstruction of phylogenetic trees

After the introduction of trees, their different representations and their different types, we want to focus on the computation of unrooted phylogenetic trees. There are two main methods which are used to compute unrooted trees. These results of the applying methods will be an optimal unrooted tree. Afterwards, this optimal tree can be rooted.

The problem with phylogenetic trees is that they can be represented in one way as it is also the case for an unrooted tree. The unrooted tree in example (20) is built on a set of four taxa which allow the construction of three different unrooted trees (Lecointre, 2006).



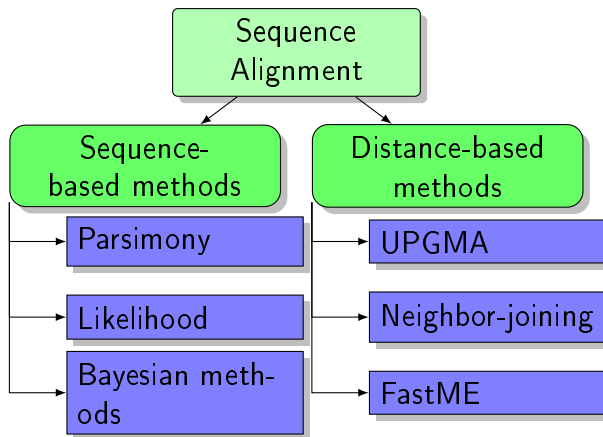


Figure 3.4: The main steps towards the reconstruction of a phylogenetic tree

By using computational methods to construct an unrooted tree, it is made possible to compute and represent an *optimal tree*. In phylogenetics, this is also called *the reconstruction problem*. The problem of reconstructing the tree can be solved in different manners. The main goal of the reconstruction is to find the optimal and so called *true tree* for a given set of species.

The illustration in figure 3.4 shows the main steps towards the reconstruction of an optimal tree. The process of computing is called *phylogenetic inference* and there are different ways to achieve the goal of constructing a true tree.

The first step in the process of phylogenetic inference is alignment. *Sequence alignment* is the comparison between two or more sequences. A sequence is a chain or string containing elements like genes. The elements of one string are assigned to the elements of the other string or to a gap. Mostly this is done while writing the sequences among each other whereas the order of the elements stays the same. The number of identical or similar elements indicates the *homology* between two sequences. In this case, homologous sequences are evolutionary related and share a common ancestor. Evolutionary events can be indicated by the alignment of different elements or of an element to a gap. Mutation correspond to the alignment of different elements and duplication or loss to the alignment of an element to a gap. If two sequences are aligned, it is called *pairwise sequence alignment* and if more sequences are aligned, it is called *multiple sequence alignment*. The comparison done via sequence alignment is the basis of the reconstruction methods (Huson et al., 2010).

The first group to look at are the sequence-based methods. As Huson et al. (2010, p.33) stated, “*Sequence-based methods* usually search for a phylogenetic tree T that optimally explains a given multiple sequence alignment M .” The input for all methods are alignments, mostly multiple sequence alignments, on a set of

taxa. The tree is reconstructed via the alignment and with the help of specified methods. The three main methods are as stated above in figure 3.4, namely maximum Parsimony, maximum Likelihood and Bayesian method. Those are broadly explained below.

The maximum parsimony method is the most widespread and famous method for sequence-based reconstruction. The basic idea of a parsimony method is to find the phylogenetic tree which represents the minimum number of evolutionary events. The detection of the events is done via multiple sequence alignment. As I stated above, the number of similar alignments indicate the relatedness of the sequences. The difference between elements indicate evolutionary changes. A phylogenetic tree reflects the relatedness of the sequences and the number of evolutionary events. Depending on the tree and the root, the placing of the evolutionary events might differ. The parsimony method detects the tree which can explain the relation of the aligned sequences while using the minimum number of evolutionary events. According to Huson et al. (2010), the parsimony method can be divided further into a small parsimony problem and a large parsimony problem. Both problems can be solved. For solving the small parsimony problem, different algorithms are provided. The large parsimony methods can be solved using different methods and their corresponding algorithms.

The second method is the maximum likelihood estimation. The basic idea of the maximum likelihood method is to reconstruct a phylogenetic tree with branch lengths using multiple sequence alignment and an underlying model of sequence evolution. The evolutionary events are computed by a model. Huson et al. (2010) and Felsenstein (2004) give examples of different models of sequence evolution. The models are used to compute the probabilities of evolutionary changes along a given tree. Additionally, the model describes the selection of the root and specifies the evolution of the sequences along the branches of the tree. The tree with the optimal and highest likelihood of the branch lengths is the maximum likelihood tree. Maximum likelihood can also be computed by using an algorithm. The most famous algorithm is the one from Felsenstein. The algorithm efficiently computes the maximum likelihood score and the tree with the best score is considered to be the optimal one (Huson et al., 2010).

The last methods to elaborate are the bayesian ones. Bayesian inference is a method used on phylogenetic trees while estimating the posterior probability. “Generally speaking, the *posterior probability* of a result is the conditional probability of the result being observed, computed *after* seeing a given input dataset” (Huson et al., 2010, p.45). Again, a given evolutionary model is assumed. Multiple sequence alignment is established which makes the computing of a phylogenetic tree via calculating the posterior probability possible. This posterior

probability is obtained from the prior probability with the help of the likelihood while using Bayes' Theorem. The main goal of bayesian inference is not one single optimal tree, but rather a sample of optimal trees according to their posterior probability. Such a sample of trees is used for further processes, where more than one tree will be needed. The method uses the *Markov chain Monte Carlo* approach to avoid the problem of normalization over all computed trees. The idea of the markov chain is to sample the results of the posterior probability distribution using a chain. The chain contains the phylogenetic trees computed with the posterior probability method. While going through the chain, at each step a new tree is proposed and the decision of replacing or keeping the old one is done via a probabilistic decision. The result should be a chain of binary branched trees. The distribution of the trees within the chain should approximate the posterior probability distribution of the phylogenetic trees (Huson et al., 2010).

The second group are the distance-based methods. "*Distance-based methods* usually construct a phylogenetic tree T from a given distance matrix D " (Huson et al., 2010, p.33). The input for creating a distance matrix are aligned sequences. The distance matrix is created by using different methods. One of which being the *Hamming distances* which takes the aligned sequences as input and calculates the positions where the sequences differ. The result is a distance matrix which is the basis for the distance-based methods (Huson et al., 2010). The three main methods are displayed in figure 3.4, namely UPGMA, Neighbor-joining, and FastME.

The first and oldest method is UPGMA (unweighted pair group method using arithmetic averages). UPGMA produces a rooted tree with the help of a distance matrix. The method is based on clustering. At each state in the given data, two clusters are merged and at the same time a new node is created in the tree. The tree is built bottom-up and has the root at the top. First, the leafs are created, then the inner nodes and last but not least the root. Each node refers a height which depends on the cluster. For example, if a cluster contains only one node the height of the node is 0. The length of the edge is computed via the difference of the heights representing at the corresponding nodes. Any tree, which is computed by this method, has the property that all leaves have the same distance to the root (Huson et al., 2010).

The neighbor-joining method is the successor of the UPGMA method. The neighbor-joining method computes an unrooted phylogenetic tree with edge lengths given a distance matrix. The method decides which two clusters are joined so that their nodes become neighbors or siblings in the tree. The average distance of each cluster according to all other clusters is calculated to balance the effect of

large distances. This avoids the problem of the need of an ultrametric tree. In an ultrametric tree, all nodes have the same distances to the root. A new neighbor-joining matrix is created to compute a new pair of neighbors. The clusters with the minimum entry in the new matrix are paired. In this way, new pairs of clusters are created where a cluster represents a node on the tree (Huson et al., 2010).

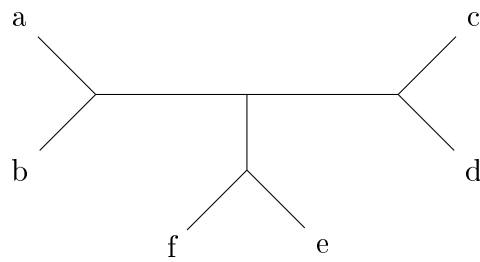
The third method is FastME which is developed within a framework called *balanced minimum evolution (BME)*. Given a distance matrix, the method computes a binary branched tree. To every edge in the tree, a *balanced edge length* is assigned. This length is calculated via the *balanced average distances* between both taxa represented by the nodes. Finding the optimal tree with this method is an NP-hard task. Therefore, heuristics for computing an BME tree need to be taken into account. The heuristics for FastME is based on two phases within the algorithm. First, an initial tree is created and second, the tree is improved in an iterative way using *nearest neighbour interchange (NNI)* operations. In an NNI operation, subtrees which are attached to the same edge are swapped in all possible ways. The NNI operation finds the minimum entry in the neighbor-joining matrix through iteration. The FastME algorithm is faster than the neighbor-joining method. This is due to the fact that the edge lengths are balanced. The NNI moves can be made constantly, as long as all balanced averages are calculated. This is the advantage of the FastME algorithm (Huson et al., 2010).

3.2.2 Working with phylogenetic trees

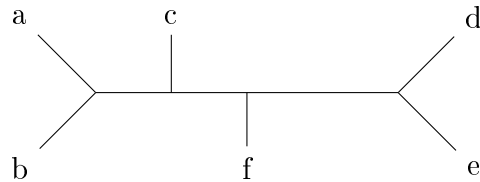
The next step after having reconstruct and compute an optimal tree is working with it. There are different methods which can be applied to trees. Two main methods are introduced here, namely the comparison of trees and the creation of consensus trees.

Two trees can be compared for measuring their similarity. This is mostly done with the help of two measures, the *Robinson-Foulds distance* and the *quartet distance*. Given two unrooted phylogenetic trees, the distance is computed by the number of transformations needed to transform one tree into the other one. Each node can be seen as a split of the tree. In the example (27) two unrooted trees are displayed. We want to transform the tree in (27-b) into tree (27-a). The transformations which are required are displayed in (28).

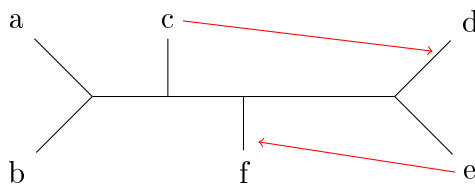
(27) a. Tree one:



b. Tree two:



(28) Transformations:



If two nodes are the same, they are contracted during the transformation and the split is removed. If a node is present in one tree but not in the other a node or a split is added. The Robinson-Foulds distance computes the symmetric difference of all splits within the two trees. For the quartet distance, two unrooted phylogenetic trees need to be given. For each tree, a set of so called *quartet trees* is created. Each quartet tree is caused by a set of four taxa. The set of four taxa is a subset to the set of all taxa represented in the tree. A quartet tree can be seen as a restrictive tree for the given unrooted phylogenetic tree induced by the subset of four taxa. In a restriction a new phylogenetic tree is received from the subset by suppressing all taxa not present in the subset. Taking the set of quartet trees for each given tree, the quartet distance can be computed (Huson et al., 2010). Both distances take two unrooted phylogenetic trees as input for the comparison. This is due to the fact that the distances of unrooted trees are more precise than the distances for rooted trees. The position of the root has a large effect on the distance between two trees. Therefore, the unrooted trees are compared and can be rooted thereafter (Huson et al., 2010).

Two or more unrooted phylogenetic trees need to be given as input to the consensus method. The trees are more like a collection of different trees computed from the same set of taxa. The trees within the collection could be gene trees. As stated above, gene trees represent the evolutionary history of the organism's

genes. On the other hand, distinct trees can be reconstructed using different reconstruction methods. Although given the same alignment, different methods compute different trees. Those trees are then added into a single collection for the consensus method. Even taking different reconstructive methods into account, only the bayesian method produces a set of possible trees. All of the trees, resulting from the bayesian method, can also be contained in a single collection. We can make the assumption that the trees contained in one collection have the same evolutionary tree. To confirm this assumption, a consensus tree is constructed by the consensus method. Within a consensus tree, “those parts of the evolutionary history on which the different phylogenetic trees agree” can be represented (Huson et al., 2010, p.63). There are different consensus methods. Huson et al. (2010) discusses three different methods, two for unrooted trees and one for rooted trees. The *strict consensus* method and the *majority consensus* method are the two most popular and important methods for unrooted phylogenetic trees. While the *Adams consensus* method is applied to rooted trees. The idea of a consensus tree is mostly used with unrooted trees. This is the case, because most reconstruction methods produce unrooted trees and the root would affect the construction of a consensus tree (Huson et al., 2010).

3.3 Phylogenetic Networks

In a broad sense one can say that if a tree is cyclic, it is a network. Therefore, Wiley and Lieberman (2011) used the term *cyclic graph* for introducing networks. The introduction of trees given above is the basis for networks. Darwin describes his first sketch of the tree of life in figure 3.1 as coral of life (Eldredge, 2005). The picture can therefore be seen as a network, more precisely as an unrooted network. Networks and trees do not differ that much as also Huson et al. (2010, p.68) stated that “[phylogenetic] networks provide an alternative to phylogenetic trees”. Networks are better suited for representing evolutionary events and reticular evolutionary events, like horizontal gene transfer.

In literature, different definitions of phylogenetic networks can be found each focusing on a specific type of network. The specific networks are not named according to their specification, but are still addressed as a phylogenetic network. Huson et al. (2010) give a general definition of a network:

Definition 3.4 *A phylogenetic network is any graph used to represent evolutionary relationships (either abstractly or explicitly) between a set of taxa that labels some of its nodes (usually the leaves)*

Explicit evolutionary relationships are represented in explicit networks which are a kind of rooted phylogenetic network. The events and the kind of network are

described in the sections below (Huson et al., 2010).

“The envisioned role of rooted phylogenetic networks in biology is to describe the evolution of life in a way that explicitly includes reticulate events ” (Huson et al., 2010, p.70). This would not be possible within a tree. This is explained in more detail below.

Doolittle (1999) introduced a so called *network of life*.

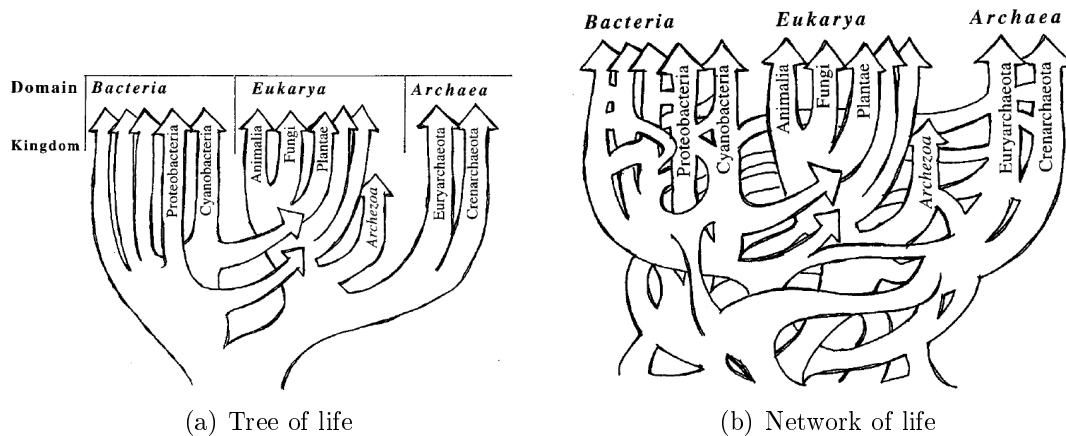


Figure 3.5: A representation from a tree to a network

The tree in figure 5(a) represents a tree which can be reconstructed with the help of the above mentioned methods. The problem Doolittle (1999) stated is that evolutionary events cannot be displayed within a tree. The evolutionary events may come from multiple trees or more than one event can be represented by one taxon. Therefore, he uses the network shown in 5(b) for representing evolutionary history. Networks, which represent evolutionary reticular events, are also called *explicit networks*. Other networks mostly visualize incompatible taxasetes and are called *abstract networks* (Huson et al., 2010).

Networks can also be divided in two groups, unrooted networks and rooted networks, and are defined analogously to unrooted and rooted trees. Unrooted networks do not have a root and are similar to an unrooted tree, where the edges can be spread to all sides. Huson et al. (2010) defines an unrooted network as follows:

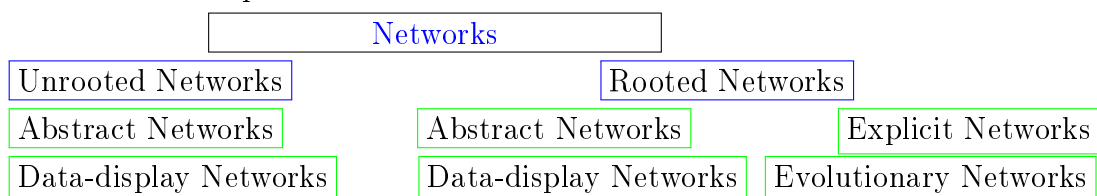
Definition 3.5 *An unrooted phylogenetic network N on χ is any unrooted graph whose leaves are bijectively labeled by the taxa in χ .*

Rooted networks on the other hand, look more like a tree. Their branches emerge from one root and are built up to a tree-like network. It is similar to the one shown in figure 5(b). The nodes can also be connected through reticular branches, representing evolutionary events. Huson et al. (2010) defines a rooted network as follows:

Definition 3.6 A rooted phylogenetic network N on χ is a rooted DAG [(direct acyclic graph)] whose set of leaves is bijective labeled by the taxa in χ . Any node of indegree ≥ 2 is called reticulate node and all others are called tree nodes. Any edge leading to a reticulate node is called a reticulate edge and all others are called tree edges.

Unrooted and rooted networks are alternatives to unrooted and rooted trees. The networks can represent more data, incompatible datasets, evolutionary history and evolutionary events. If we are talking about unrooted networks, we can also refer to them as abstract networks. Mostly, unrooted or abstract networks are used for representing and visualizing incompatible datasets. Rooted networks, on the other hand, can be refer to as abstract and explicit networks. This division depends on the type of rooted network. If the network contains and represents evolutionary events, it is an explicit network. Otherwise, it is an abstract network (Huson et al., 2010).

Networks can also be divided into data-display and evolutionary networks. Morrison (2011) makes this division in his book. In this case, we can also draw the connection to unrooted and rooted networks and to abstract and explicit networks. Data-display networks are unrooted and abstract networks. Given some different and incompatible taxaset, the data-display network indicates the relationships between the samples. It is more or less a diagram visualizing the possible relationships among the taxa without making any assumption on evolutionary change. Evolutionary change is represented in evolutionary networks. Those are explicit networks and therefore rooted. The root represents the ancestor of all species analysed within the taxaset. The branches demonstrate the path to the corresponding descendants. Along that path, the evolutionary change takes place. This change happens through evolutionary events and indicates the evolutionary history. This can all be represented in a evolutionary network (Morrison, 2011). All of those representations of networks can be quite confusing. Here is a short overview of all representations.



3.3.1 Different Types of Networks

The illustration above states the main representations of networks. Both representations can further be represented by different types of networks. Before, we start talking about different types, we should concentrate on a single group

of representations. In line with Huson et al. (2010) I take unrooted and rooted networks as the main representations. Within these representations, more types of networks can be classified. The unrooted networks can be divided into *split networks* and *quasi-median networks*. The rooted networks can be divided into four types of networks, namely *cluter networks*, *hybridization networks*, *recombination networks* and *Duplication-Loss-Transfer networks*. I will first introduce the unrooted networks and then the rooted networks.

Split networks are one type of unrooted networks which depend on a set of splits. As stated above, the splits can be represented by nodes. We have a taxa set χ which includes a number of splits S . The splits may be weighted indicating character changes, distance or other representations. The set of splits S can be used for creating an unrooted phylogenetic network where each split indicates one edge in it (Huson et al., 2010). An illustration of a split network is given in figure 3.6 which is taken from Huson et al. (2010).

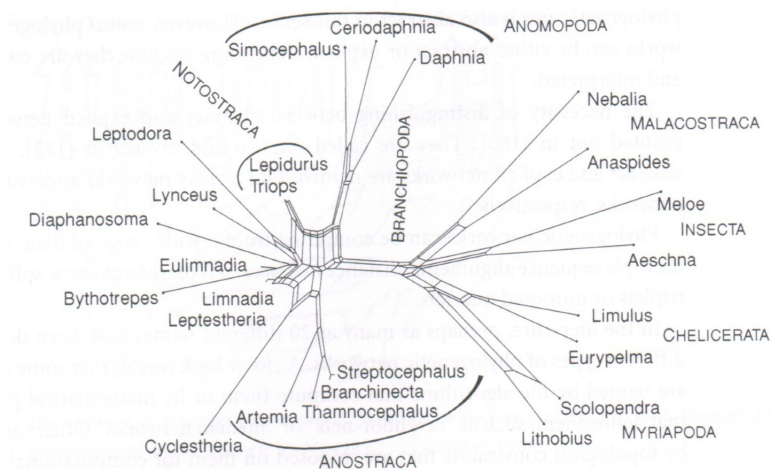


Figure 3.6: An illustration of a split network.

A split network can contain a different number and different types of data. The splits of the data are represented by the network. The network is computed by using an algorithm, as for example the *convex hull algorithm* or the *circular network algorithm*. Split networks can be computed from different inputs, namely from distances, trees and sequences. When computing a network from a distance, the input is a distance matrix. The distance matrix is used for creating the set of weighted splits. The two most popular methods for doing this are: *split decomposition method* and *neighbor-net method*.

The input is always a distance matrix. The decomposition network creates a set of weighted splits that is *weakly compatible*. This property ensures that the network is not too complicated.

The neighbor-net method also takes a distance matrix as input and creates a set of weighted splits. This set is circular and can be used as input for the circular algorithm. The hereby created networks receive their corresponding name: *decomposition network* and *neighbor-net network* (Huson et al., 2010).

A split network can be computed from a set of unrooted phylogenetic trees. As stated above, the trees might differ because they are either gene trees, computed with different methods, or multiple trees from a bayesian analysis. This method is similar to the method of building a consensus tree except that here a consensus network is built. The networks is called *consensus split network* or *super split network*. It can “visualize conflicting signals in a set of trees” (Huson et al., 2010, p.73).

The third network is computed on the basis of sequences. The input is a multiple sequence alignment where every character pair indicates a split. Using this set of splits, a split network can be computed using the convex hull algorithm. The columns in the alignment are the labels for the edges present in the corresponding split. This split network is called *median network* (Huson et al., 2010).

The other unrooted phylogenetic network is the quasi-median network. This network was constructed to representing multi-state characters. The input of the network is a multiple sequence alignment. The quasi-median network is a generalisation of the split network. The network is rarely used in practice, because the resulting network of a multiple sequence alignment is too large and complicated. An alternative is the computation of a subnetwork with the *median-joining algorithm*. The network would be a *median-joining network* (Huson et al., 2010).

The other main group of networks are the rooted networks. The four types of rooted networks discussed here are: cluster networks, hybridization networks, recombination networks and DLT networks.

Cluster networks are an abstract type of network, also called data-display network. The network represents a set of clusters. Each cluster is a group which provides assumptions of evolutionary relatedness within the taxa. The network can represent a cluster in two different ways, either *hardwired* or *softwired*. The cluster network does so in the hardwired sense. This means that there is a tree edge in the network such that the set of labels on the nodes of the edge are equal to the cluster (Huson et al., 2010). This can easily be calculated by the *cluster-popping algorithm*. The cluster network is an abstract rooted phylogenetic network and it can be used for visualizing sets of rooted trees (Huson et al., 2010). The other three networks are all explicit networks, representing evolutionary events and history.

The hybridization network is computed out of a set of taxa which was developed with the help of a model of evolution. This model indicates evolutionary events, like speciation, descent-with-modification and hybridization events. All of them can be visualized in a rooted network. The speciation events are displayed at the corresponding tree node and the hybridization events are represented by the reticular nodes in the network (Huson et al., 2010). In theory, a hybridization network can also be built out of two or more gene trees. The topology of the trees differ and the assumption is made that this is due to hybridization. Computationally, this can only be implemented with two rooted trees (Huson et al., 2010). The figure in 3.7 is an illustration of a rooted hybridization network, taken from (Huson et al., 2010).

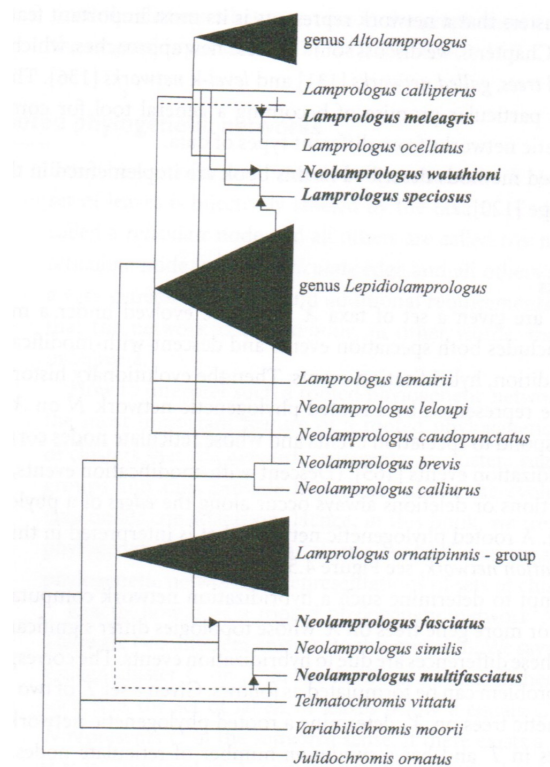


Figure 3.7: An illustration of a rooted hybridization network.

The next explicit network is the recombination network. The input is a set of taxa which was developed by an evolutionary model. Therefore, it includes evolutionary events like, speciation, descent-with-modification and recombination events. The evolutionary history is represented in a recombination network. Again, the tree nodes represent the speciation events and the reticular nodes the recombination events. According to Huson et al. (2010), the following labels are given:

- a labeling of all nodes by sequences, and

- a labeling of all tree edges by positions in the sequences at which mutations occur.

“These labellings must be compatible in the sense that the sequences assigned to the tree nodes of the network differ exactly by the indicated mutations, while the sequences assigned to reticular nodes must be obtainable from the sequences assigned to the parents nodes by suitable recombinations” (Huson et al., 2010, p.78).

The third explicit network is the DLT network, where D stand for duplication, L for losses and T for transfers. The input is again a set of taxa developed by a model of evolution. It concludes speciation events, descent-with-modification, gene duplication, gene loss and horizontal-gene-transfer events. This model is used for mapping a gene tree to its species tree. By applying a duplication-loss-transfer scenario, the gene tree can be mapped to its species tree and the differences between the trees can be shown via evolutionary events (Huson et al., 2010).

All types of networks represented in this section can either be computed by an algorithm or explained mathematically. Huson et al. (2010) provides further explanations and algorithms in his book. Some of the algorithms are implemented in programs and can be tested. Huson et al. (2010) list some of these software programs, additionally Morrison (2011) provides a list of software for data-display networks and evolutionary networks in his book.

4 Phylogenetics in Historical Linguistics

While talking a lot about linguistics and phylogenetics in this paper, the two fields can be compared to show their parallels. Before phylogenetics come into play, I will compare biology and linguistics.

Darwin (1871) was one of the first biologist who stated that processes in language and biology show parallels.

“The formation of different languages and of distinct species, and the proofs that both have been developed through a gradual process, are curiously the same. But we can trace the origin of many words further back than in the case of species, for we can perceive that they have arisen from the imitation of various sounds, as in alliterative poetry. We find in distinct languages striking homologies due to the community of descent, and analogies due to a similar process of formation.”
(p.57-58)

This statement of Darwin leads to the discovery of similarities between biology and linguistics, as the pedigrees of Haeckel (1874) and Schleicher (1873) show. Haeckel was a biologist, while Schleicher was a linguist. Atkinson and Gray (2005) state in their article that Haeckel introduced Schleicher to the theory of Darwin. Schleicher had already used pedigrees for representing language history and so did Haeckel, but with the theory of Darwin the similarities between both are revealed. Both trees are famous for representing one of the first pedigrees, each is popular in the corresponding field of its author. The first contact between biology and linguistics was established by using the same method for representing evolution and relationships.

Atkinson and Gray (2005) summarized some general parallels between biology and linguistics which are displayed in table 4.1.

One famous comparison are cognates and homologies. In section 2, I introduced not only loanwords, but also cognates. Cognates are set of words which are etymologically related having the same ancestor. In biology, homology can have different meanings. Lecointre (2006) states that homology can have two different meanings:

1. Two homologous structures are inherited from a common ancestor.
2. By comparing organisms, a structure of characters is homologous if another structure has the same characters.

Zehnte Tabelle.
Stammbaum der indogermanischen Sprachen.

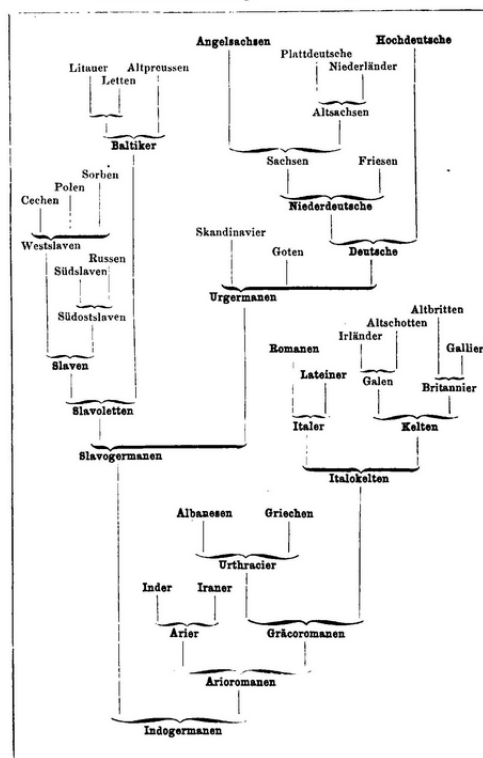


Figure 4.1: Heackel's illustration of a pedigree for the Indo-European language family.

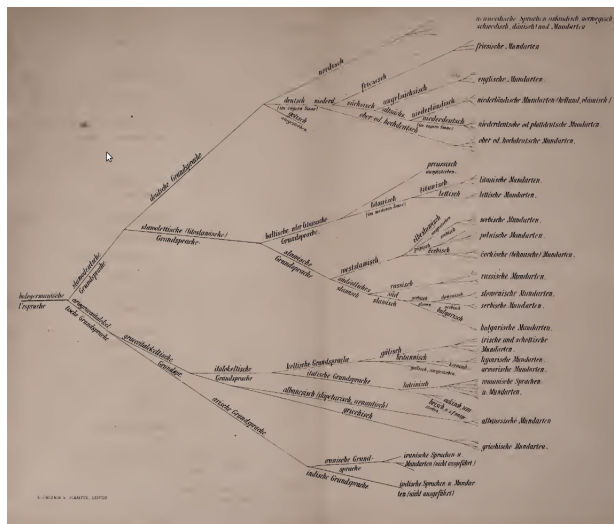


Figure 4.2: Schleicher's illustration of a pedigree for the Indo-European language family.

List (n.d.-c) states that homology in the sense of sharing a common ancestor can be divided into three specific relations: *orthology*, *paralogy* and *xenology*. According to List (n.d.-c), orthology refers to different genes which are related via

Biological evolution	Linguistic evolution
Discrete characters	Lexicon, syntax, and phonology
Homologies (Orthology, Paralogy)	Cognates
Mutation	Innovation
Drift	Drift
Natural selection	social selection
Cladogenesis	Lineage splits
Horizontal gene transfer (Xenology)	Borrowing
Play hybrids	Language Creoles
Geographic clines	Dialects/dialect chains
Fossils	Ancient texts
Extinction	Language death

Table 4.1: Conceptual parallels between biological and linguistic evolution

speciation, paralogy refers to different genes related via duplication and xenology refers to genes related via transfer.

This classification of homology into three parts affects the relation between homology and cognation. Within this classification, sharing a common ancestor can be due to three different evolutionary events, namely speciation, duplication and lateral transfer. In linguistic, there is a strict distinction between cognates and loanwords. Cognates are descendants of a common ancestor, whereas loanwords share a common ancestor because of borrowing. Therefore, the parallel can only be drawn between cognation and orthology and paralogy (List, n.d.-c). Xeneology is the same than gene transfer which is related to borrowing. The original table presented by Atkinson and Gray (2005) uses the overall term for homology and relates it to cognation. I modified the table in 4.1 according to the classification of List (n.d.-c).

The parallel between horizontal gene transfer and borrowing is the most interesting one for this paper. In general, horizontal gene transfer is a method in biology for describing the inheritance of a gene between two unrelated organisms. Morrison (2011) describes it in this way:

“HGT (horizontal gene transfer) occurs when a small piece of a genome (usually a whole gene) is transferred between unrelated organisms by means other than sexual reproduction.” (p.112)

The counterpart is *Inheritance*, where the gene is inherited from the parent(s) to their children. In linguistics, this would be something like word or language transmission from one generation to the next generation, taking language change into account.

In linguistics, borrowing is the process of a word being transferred and adapted

into another language. The result of this process is the loanword. Borrowing might happen between languages or language families which are not related to each other (Haugen, 1950). Atkinson and Gray (2005) uses the same example for describing the process of borrowing as I used in section 2, namely the one of the English word *mountain*. English is a Germanic language and for all Germanic languages the same word for *mountain* might be expected. But English borrowed the word from French.

(29) Old French: *montaigne* - English: *mountain*

Other Germanic languages have distinct words for *mountain*, for example German has the word *berg* and Dutch has the word *bjerg*. Romance Languages have words similar to *mountain*, for example French has the word *montagne* and Spanish the word *montaña*. The English word is borrowed from the Romance language family into the Germanic language family. The word does not have the same ancestor before the borrowing and is therefore horizontally transferred.

Huson et al. (2010) stated that horizontal gene transfer can be represented within an explicit rooted phylogenetic network, namely DLT. Within this network additional events, namely duplication and loss events, are also represented. Therefore, the question arises, whether the process of borrowing can also be visualized in a similar way. One approach to borrowing detection is proposed by Minett and Wang (2003). Their goal is to detect borrowing of lexical items among “a family of genetically related languages” (Minett & Wang, 2003, p.3). Their methods for detecting lexical borrowing are distance-based and character-based.

However, as I stated in section 2, borrowing mostly depends on phonology and sounds. I will therefore focus on approaches which are based on phonological and sound borrowings. Firstly, I will introduce *LingPy*. It is a python package including different modules for automatic sequence analysis in historical linguistics. The package includes basic cluster algorithms from phylogenetics which can be used for reconstructing the borrowing process. Additionally, the fragment builds on phonological data and is therefore closer to an analysis which I am to achieve. Secondly, I will introduce my own theoretical approach to borrowing detection. This approach is based on the detection of horizontal gene transfer. A gene tree is mapped to a species tree for detecting transfer events.

4.1 *LingPy*

Computational methods became quite popular in scientific fields like linguistics. In computational linguistics, corpora and databases are created automatically by using different tools. Automatically created corpora are obviously bigger than the

ones created manually. The same holds for searching different patterns through the usage of corpora. Tools and methods are created for searching big corpora efficiently. With this development, various linguistic questions can be answered using a great amount of data provided by different corpora. On the one hand, this makes the linguistic theory more reliable and on the other hand, rare phenomena might be revealed. In computational linguistics, tools are created for different approaches as for example in natural language processing and machine translation. The field of technology grows faster and faster and with it, the demand of developments to computational linguistics. Why not also use computational methods in historical linguistics?

In biology and phylogenetics, computational methods of detecting different phenomena, are already present. Huson et al. (2010) stated most of them in a mathematical way and also provided algorithms for detecting specific phenomena.

If Atkinson and Gray (2005) can compare phylogenetics and historical linguistics in a theoretical way, why not use the computational methods of phylogenetics in historical linguistics?

LingPy is a python package which contains all sorts of different methods for quantitative analysis in historical linguistics (List, n.d.-b) and it can be included easily in every python script. It includes several methods for analysing linguistic data. Most of the program is based biological and phylogenetic methods. This is explained in more detail on the homepage www.lingpy.org. LingPy is a great development for analysing linguistic data in an automatic way. It is way more efficient than collecting and analysing data manually. The existence of large databases is a good basis for using computational tools and programs to detect linguistic phenomena. More data can be processed and the analysis gains significance. Nevertheless, there are also problems while handling different types of data. The next section gives an introduction to LingPy and its main methods and states advantages and disadvantages of working with different databases. Afterwards, I will explain the detection of borrowing in more detail, using different studies.

4.1.1 The Python Library for Historical Linguistics

LingPy is a program for data analysis in historical linguistics. Computational methods in historical linguistics develop during the last years and became quite popular for analysing linguistic data. Most of the method used in historical linguistics came a biological or phylogenetical background. As we saw above, there are quite a lot of similarities between the two fields, therefore some methods

can be modified and used in historical linguistic.

List and Moran (2013) showed a workflow of the LingPy package. I will use that illustration and go step by step through it for explaining the main steps of the program.

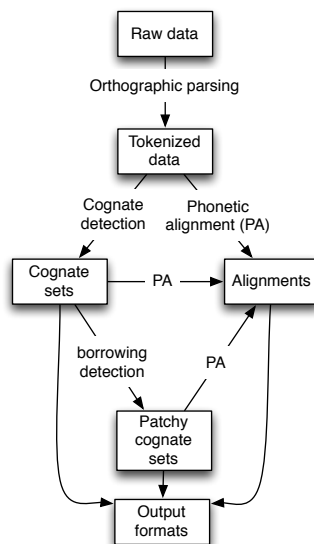


Figure 4.3: Workflow through the LingPy program.

The input data is a simple format, easy to create and it can be edited by the user. It is parsed using a parser and the words are tokenized. An orthographic parser is implemented in LingPy. The orthographic parser does not only tokenize the words, but represents the tokens in IPA format. IPA stands for *International Phonetic Alphabet* which aims to include a symbol for each sound of any language. For the representation of IPA tokens one should add an orthography profile. This profile represents the letters and their corresponding IPA sound (List & Moran, 2013). The result is the tokenized data which is needed for phonetic alignment or cognate detection.

Phonetic alignment can be compared to sequence alignment. The words contained in the tokenized data are compared to each other. Each word contains a sequence of IPA symbols which is aligned to the sequence of other words. The alignment indicates the similarity between the words. LingPy contains algorithms, like the *Needleman-Wunsch algorithm* and the *Smith-Waterman algorithm* for automatic sequence alignment. The algorithms are implemented with slight modifications (List & Moran, 2013).

For the detection of cognates a phonological basis is needed. This is provided by the tokenized data. LingPy contains four different methods to detect cognates. The main task of all methods is the grouping of the words into clusters. These

clusters are also called cognates clusters because after the grouping each cluster contains a set of cognates. The four methods differ in their computational techniques which leads to the grouping of the words (List & Moran, 2013). The results of the cognate detection can be saved in a file or plotted in a tree. The workflow on the webpage <http://www.lingpy.org/tutorial/workflow.html> shows the plotting of a phylogenetic tree which is calculated using the neighbor-joining algorithm.

Having the clusters of cognates, one can detect borrowing with LingPy. This is the case, because “[incompatible] (patchy) cognate sets often point to either borrowing or wrong cognate assessments in the data” (List & Moran, 2013, p.16). The main requirements of the borrowing detection are the cognate sets of the given data and a reference tree of the languages contained in the data. The reference tree can be provided by the user or computed by LingPy. There are three different methods implemented to detect borrowing. The main task of the methods is the computation of evolutionary events. These events are represented in a *minimal lateral network (MLN)* (List & Moran, 2013). The methods only differ in their algorithms for the detection of the evolutionary events. The output can be saved in a file and the network can be saved in its corresponding data format (List & Moran, 2013).

4.1.2 Borrowing Detection with LingPy

The detection of borrowing is the most interesting part of LingPy for this paper. Therefore, I will introduce two studies on borrowing detection below. Both studies are based on the same Indo-European languages contained in a dataset named IELex (Dunn, n.d.), but differ in their reference trees. The most interesting part of this section will be the outcome of the borrowing detection visualized in a MLN network.

The two main processes to detect borrowing within LingPy are the computation of gain-loss events and the visualization of these evolutionary events within a minimal lateral network (MLN).

The first process within the borrowing detection is *gain-loss mapping*. The gain-loss mapping is the underlying idea of detecting evolutionary events. In each method, such a *gain-loss scenario*, how List, Nelson-Sathi, Martin, and Geisler (n.d.) calls it, is created. This scenario indicates the evolution of a character along the reference tree (List et al., n.d., p. 10). The development of a character is indicated with a presence (1) or an absence (0). “A *gain event* (also called *origin*) is defined as the change from state 0 to state 1, and a *loss event* is defined as the change from state 1 to state 0, respectively” (List et al., n.d., p.10). This

changes are represented within the nodes.

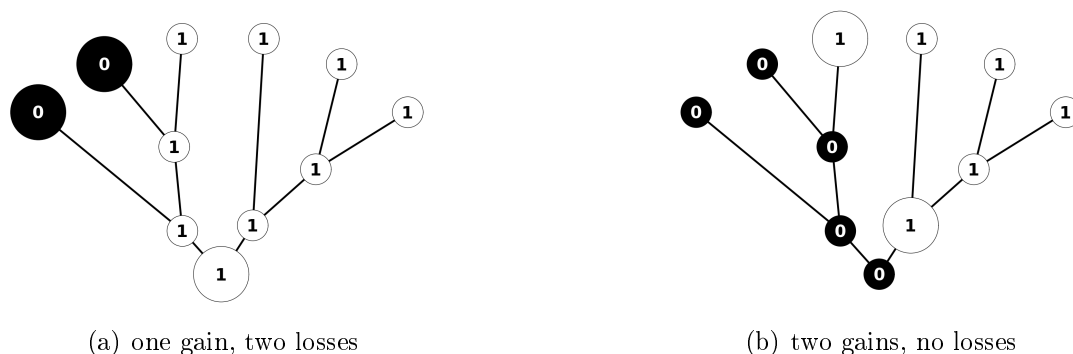


Figure 4.4: Gain-loss mappings

The two figures of gain-loss mappings, are taken from List (n.d.-a). They represent a gain-loss mapping analysis for different words having the same meaning as the Latin word *computare*. The Spanish word *contar*, the French word *compter* and the Italian word *contare* are cognates to the Latin word *computare*. The three Germanic languages have distinct words referring to the same meaning. The English word *count* is a cognate to the Latin word *computare*, whereas the German word *zählen* and the Danish word *tælle* are different words from a Proto-Germanic ancestor **taljan-* (List, n.d.-a). According to this difference between English and German and Dutch, the gain-loss mapping should indicate the change of the English word. The trees in figure 4.4 refer to two different gain-loss scenarios.

In the first scenario, in figure 4(a), there are one gain and two losses. Those indicate that the English word has the same ancestor as the Romance languages and the corresponding words in German and Dutch are lost. In the second scenario, in figure 4(b), the two gains indicate that all Germanic languages have the same ancestor. Two gains and no losses are present in this scenario where one gain is the occurrence of the (loan)word in English. The second scenario is the historically correct one, because the English word *count* was borrowed from the French word *conter* (List, n.d.-a). The question is, how do we find the right scenario automatically?

The gain-loss scenario is used for the detection of evolutionary events. LingPy provides three different methods to detect evolutionary events, but only one method uses gain-loss scenarios for selecting the optimal one. This is the parsimony-based approach. “In order to find a consistent way of selecting the most parsimonious scenario, we test different *models* that assign different penalties for the scenario, depending on the number of gains and loss events proposed by them. A model is defined as the ratio between penalties for gain and loss events” (List et al., n.d., p.10). The figures in 4.4 can be seen as two different models, according to the explanation of List et al. (n.d.). All possible scenarios are computed using

bottom-up approaches which means from the nodes to the root. The method is a bottom-up approach which computes all possible gain-loss scenarios. The trees given in the article are the other way around as the ones in figure 4.4. “The most parsimonious scenario for a given model is the one which minimizes the overall penalty” List et al. (n.d.). If we computed an optimal model for a given dataset, the results can be displayed.

This is done within a minimal lateral network (MLN) which brings us to the second process. There are two things needed for creating a MLN, a reference tree and gain-loss scenarios. The reference tree is the basis of the network, representing the relationship between the languages in the dataset. The gain-loss scenario or the optimal model is used for drawing the lateral events between the different languages. “Borrowing events are assumed for all patterns for which more than one origin was inferred by a given gain-loss model, and links are drawn between the nodes in which the characters originate” (List et al., n.d., p.12). The edges of the MLN are weighted, whereas the weights reflect the number of patterns. The MLN is represented for each of the two datasets in the next section.

Borrowing in Indo-European Languages

The dataset of the Indo-European languages is the same than List (n.d.-a) uses in his study. The dataset is a subset of the Indo-European Lexical Cognacy Database (IELex, (Dunn, n.d.)) which contains 40 Indo-European languages with 7 518 words clustered into 1 194 cognate sets (List, n.d.-a). The borrowings within the data are already known. Therefore, the correctness and accuracy of the methods to detect borrowings can be tested. The dataset was modified by List (n.d.-a). He corrected errors in the cognate set and added some unobserved cases of borrowing.

I will compare two different results displayed in a MLN. Both studies use the same method to detect the borrowing, but the different reference tree leads to different results in the computation of the MLN.

In the study of List (n.d.-a), the reference tree is a binary branched family tree based on the article of Ringe, Warnow, and Taylor (2002).

The reference tree in figure 4.5 is taken out of the Supplemental Material I of List (n.d.-a).

In the second study, a small case study of Johann-Mattis List, the reference tree is created according to Southworth (1964). Southworth (1964) creates a family tree based on phonological data. Additionally, the tree is not binary branching but on some nodes multi branching.

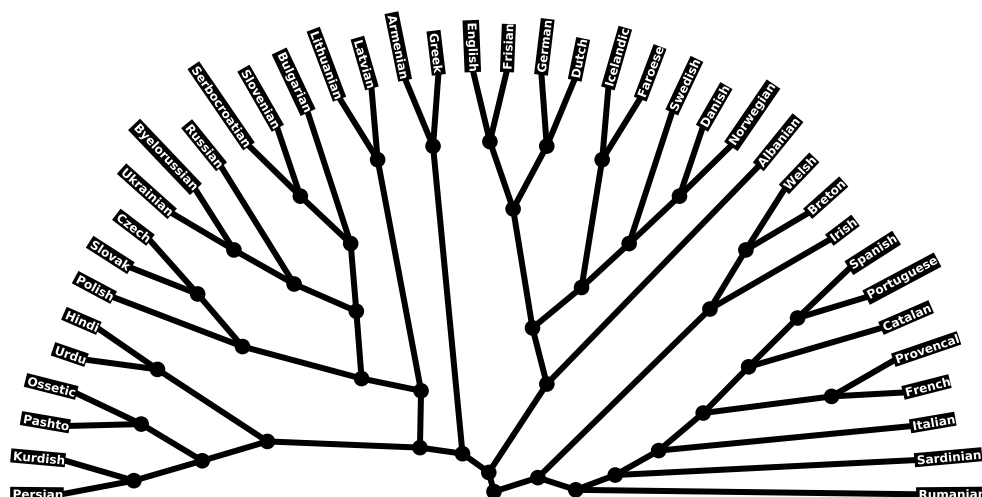


Figure 4.5: The reference tree for Info-European languages based on Ringe (2002)

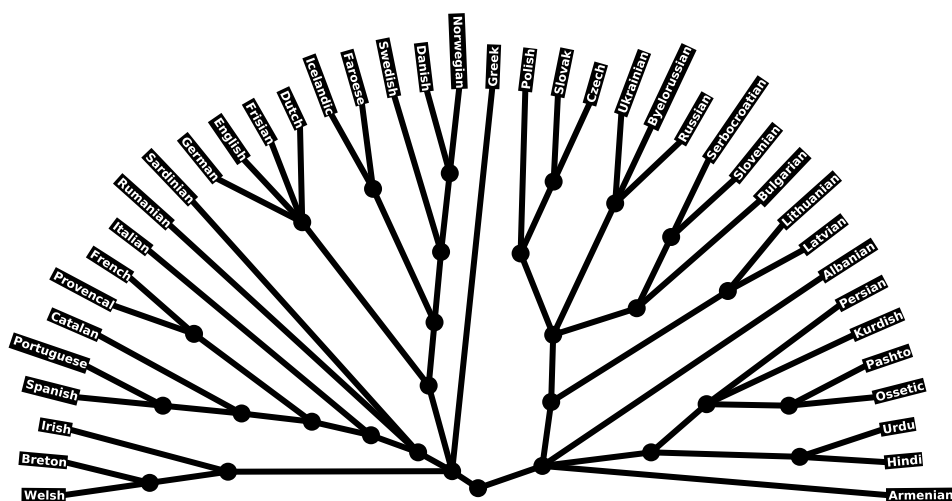


Figure 4.6: The reference tree for Info-European languages based on Southworth (1964)

The parsimony-based approach was tested with both datasets and their corresponding cognate sets. Nine different models were tested and “the model that yielded the highest p-value in the Wilcoxon rank-sum test of contemporary and ancestrals VSDs [(vocabulary size distributions)] was selected as the best one” (List, n.d.-a, p.8). The VSD is a restriction from Nelson-Sathi et al. (2011) to determine an optimal model. The vocabulary size distribution is defined as the number of words a language needs to express a given cognate set. The number of words from one language should not differ greatly from the number of words

in another language. The greater the VSD number, the more different are the cognate sets and the less optimal is the model. The optimal model is represented in a MLN.

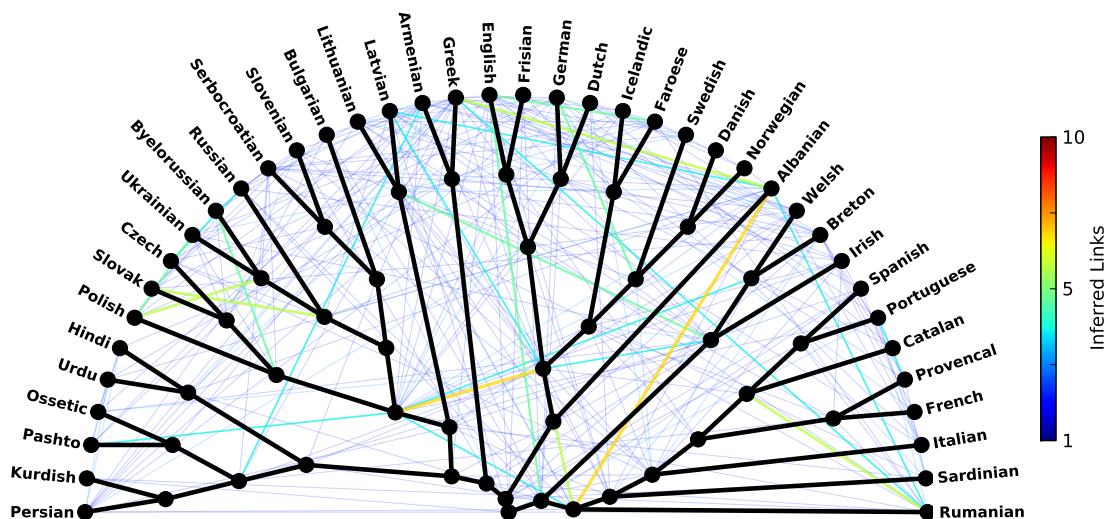


Figure 4.7: The MLN for IELex using the reference tree of Ringe (2002)

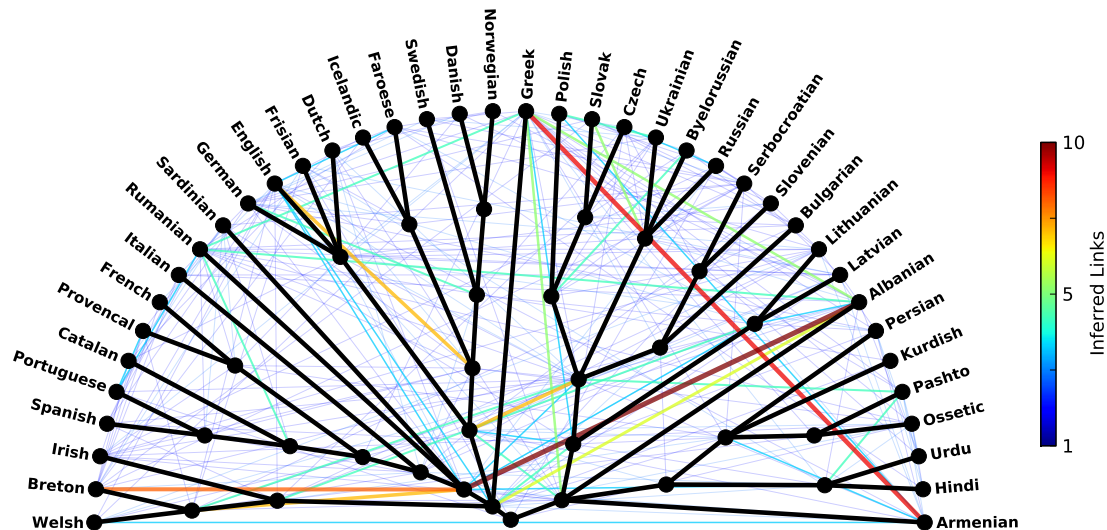


Figure 4.8: The MLN for IELex using the reference tree of Southworth (1964)

It is obvious that the two MLNs are different. According to their corresponding reference tree, the grouping of the languages differ in the MLNs. In the MLN, based on the reference tree taken out of List (n.d.-a), the grouping is based on one root. This is due to the fact that the reference tree is binary branched. The main groups are the following (starting from the right):

Romance languages, Celtic languages, Albanian + Germanic languages, Greek + Armenian, Balto-Slavic languages, and Indo-Iranian languages.

These groups can additionally be divided into two main groups. The tree is binary branched therefore only two languages or language families can share a node. The branching is responsible for the grouping of the language. If the tree would be multi-branched, the grouping may look different.

This is the case for the MLN based on the multi-branching tree of Southworth (1964). There are two main groups based on one root. The language groups contained in their corresponding main groups are the following (starting from the right):

First Group: Armenian, Indo-Iranian languages, Albanian, and Balto-Slavic languages.

Second Group: Greek, Germanic languages, Romance languages, and Celtic languages.

The underlying structure from the reference tree and the MLN is the same. Therefore, the two main groups are each connected to a node which is connected to the root. The node is multi branching, therefore such a hierarchy as in the network in 4.7 is not needed.

The grouping depends on the reference tree, but one would assume that the detected borrowings should not differ within the same dataset. The first obvious difference between the MLNs are the different weights for the edges. The weights are represented by the number of cognate sets or words. The more cognate sets or words, the greater the weight. The question arises if there are some inferred events which may not be displayed due to the different branching of the reference trees? I will look at the inferred links with the highest weight which means all links with a weight ≥ 5 . The main links between two languages are listed with their corresponding weight, started from the one with the highest weight. Additionally, I will look if a link between English and the Romance languages is present.

Node	Node	weight
Germanic languages	Slavic languages	6
Albanian	Romance languages	6
Polish	Byelorussian + Ukraine	5
Albanian	Greek	5
Romance languages	Germanic languages	5
Rumanian	Iberian Romance languages	5
East Slavic languages	Slovak	5
English	Romance languages	4

Table 4.2: The top links for the minimal lateral network in figure 4.7

Node	Node	weight
Albanian	Romance languages	10
Greek	Armenian	9
Breton	Romance languages	8
Germanic languages	Slavic languages	7
Scandinavian languages	English	7
Celtic languages	Romance languages	7
Albanian	Germanic + Romance + Celtic languages	6
Slavic + Albanian + Indo-Iranian + Armenian languages	Greek	5
Greek	Albanian	5
Slovak	East Slavic languages	5
English	Romance languages	3

Table 4.3: The top links for the minimal lateral network in figure 4.8

Comparing the results, the difference is obvious. The MLNs show a clear difference and so does the table with the weights. Four links are present in both networks and two of them have different weight. The other links are different. I take all different links of each network into account. If the link is present in one list but absent in the other, I will check if there is a link with a smaller weight or if the link is absent. I start with the first list in table 4.2.

- The first different link is the one between Polish and Belyorussian+Ukraine. This link is not present in the other network. This can be due to the fact that Belyorussian and Ukraine share the same node with Russian. The link in the other network excludes Russian. Therefore, no link between Polish and Belyorussian and Ukraine can be drawn. But Polish is linked to each language with a small weight. Therefore, not the exact link is present but a derivation of the link.
- The link between the Romance languages and the Germanic languages cannot be present in the other network, because the two language families share the same node and are therefore already connected.
- The link between Rumanian and the Iberian Romance languages is present in the other network, but with a smaller weight and therefore not listed in the table 4.3 above.

Having a look at the links present in table 4.3 and not in table 4.2.

- The link between Greek and Armenian cannot be present in the other network, because they share the same node and are already connected.
- The exact link between Breton and the Romance languages is not present in the other network. Therefore, a derivation of the link is present, namely the link between Breton and the Western Romance languages which exclude Rumanian and Sardinian.
- The link between English and the Scandinavian languages is not present. My first thought was that the link does not need to be present because they share a node. But they also share a node in the network 4.8. The reason is the multi-branching tree. To avoid confusion, English is linked to the Scandinavian languages.
- The link between the Celtic languages and the Romance language is also not present in the other network. In this case, it is due to a common node. In the network 4.7, they have a common node and need no link. In the network 4.8, they also share a node. This is also due to the multi-branching network in 4.8. The common node between the languages is also shared by Greek and the Germanic languages. Therefore, they need to be linked to avoid confusion.
- The two last links between Albanian and Germanic + Romance + Celtic languages and between Greek and Albanian + Armenian + Slavic + Indo-Iranian languages are not present. This is due to the structure of the reference trees. If one would link Albanian and Greek the corresponding group of languages, they would all be linked to the root. The root is the only node where the all of the languages, to which Greek and Albanian are linked, are present.

Additionally, I listed the link between English and the Romance languages. As I stated throughout the whole paper, one traditional example of borrowing is the one of the English word *mountain*. The word is borrowed from the Romance languages. This link can be found in both networks which is a nice proof for the loanword in English.

The difference between the links is due to the difference of the reference tree. The trees prohibit certain links to be drawn because of the representation of the grouping. The close relation between languages can be explained with cognates. If this is not the case, the relation between languages is due to borrowing. The two

MLNs is a great visualisation of the borrowings computed with the parsimony-based method.

There are still things which might be important to detect within borrowing. One such thing, is the direction of borrowing. The MLNs cannot show any direction. For example, the link between English and the Romance languages indicates borrowing but one cannot say if English borrows from the Romance languages or if the Romance languages borrow from English. This is an important issue in the process of borrowing. Although, there is no solution provided at the moment, the direction is something which should be taken into account. Nevertheless, the process of automatic borrowing detection implemented in LingPy is efficiently and reliable. The results are visualized in an descriptively way and can clearly be interpreted.

5 Borrowing Detection with Horizontal Transfer

In section 4, I presented the parallels between biological and linguistic evolution. I stated that the most interesting part is the parallel between *horizontal gene transfer (HGT)* and borrowing. This idea is based on the detection of the events of horizontal gene transfer within languages and the idea to use this transfer to represent borrowing events. This chapter introduces an approach of horizontal gene transfer and its usage in linguistics. Nelson-Sathi et al. (2011) draw also the parallel between HGT and borrowing.

There are several methods and approaches for detecting horizontal gene transfer events. Nelson-Sathi et al. (2011) implemented a method based on borrowing models. This method is also contained in LingPy.

I will focus on the detection of horizontal gene transfers by mapping a gene tree to a species tree. This tree-based method is common in phylogenetics and has been used for several years. For constructing a gene tree and a species tree within linguistics, the phylogenetic methods can be adapted. For the construction of such a gene tree and species tree in linguistics, language data is needed. The reconstruction is done according to a phylogenetic distance-based approach and a phylogenetic reconstruction method.

The approaches of detecting horizontal gene transfer can also be adapted into linguistics to detect evolutionary events, like borrowing. The transfer of gene events can be visualized by representing the transfer within the structure of the species tree. Additionally, it is tested if this representation can also be useful within linguistics.

Firstly, I will introduce species trees and gene trees in more detail. Especially, the ones based on the languages which are used for this approach. I will also explain the underlying data and the computational methods used for the reconstruction of the trees. Afterwards, I will introduce approaches used for the detection of horizontal gene transfer. The focus lies on different tree-based methods and computing transfer events. In the next subsection, I will explain T-REX, a web server containing applications for working with phylogenetic trees and networks. The detection of horizontal gene transfer is done automatically and the result is visualized. The interesting part is the reconstruction of the language trees and if borrowing can be detected with the same way as horizontal gene transfer. In the last section, I will compare this theoretical tree-based approach with LingPy, illustrating similarities and differences.

The approach should give a clearer insight into the usage of tree-based methods and their detection of evolutionary events in linguistics.

5.1 Species trees and gene trees

Before I explain horizontal gene transfer, I will introduce linguistic species trees and gene trees.

Species trees represent the evolutionary history of an organism, whereas gene trees represent its genes. Within the evolution of genes, different evolutionary events can take place. These evolutionary events cannot be seen within a species tree. Therefore, gene trees are reconstructed and compared to their corresponding species tree to detect such events and explain the evolution of the genes. Gene and species trees can also be used within other computational methods. Multiple gene trees can be used to reconstruct a species tree of the ancestor species. If the species tree is reconstructed, the gene tree and the species tree can be compared to get a better insight on the speciation of the different species.

Within linguistics, both ways of using a species tree and a gene tree can be integrated.

I will refer to the species tree as *expert tree* and to the gene tree as *concept tree*. Languages can be seen as linguistic organisms and words or concepts can be seen as linguistic genes, because languages contains words as an organism contains genes. The expert tree is formed by a set of languages, whereas the concept tree represents a word contained in this set of languages.

In the second scenario, where multiple gene trees are mapped, the genes are synonymous to the concepts. The evolutionary history of each concept is represented by a concept tree. By mapping all concept trees, an expert tree representing the ancestor language can be reconstructed. The greater concept tree, including all concept trees, can also be compared to an existing species tree of the language family. With the comparison similarities and differences of the speciation of the languages can be illustrated. The first scenario can also be integrated into linguistics. This is also the scenario which forms the basis of this approach.

The concept tree is mapped to an expert tree. The mapping of the concept tree to an expert tree can be used to discover evolutionary events. As already stated in section 4, biological evolution and linguistic evolution are parallels in various ways. Horizontal gene transfer and borrowing is one of these parallels. Horizontal gene transfer is an evolutionary event which can only be detected within genes. The same holds for the process of borrowing which can only be detected within words. The mapping of the concept tree to the expert tree allows us to detect such events. Horizontal gene transfer and its detection via the mapping of two trees onto each other is explained in the next section. First, I want to explain the computation of the expert and concept trees, the underlying data and the appearance.

The data used in this approach is from the *Automated Similarity Judgement Program (ASJP)* (Wichmann et al., 2012). The main goal of the ASJP is the automated classification of languages through comparison of words. The ASJP compares pairs of languages to find lexical similarities. For each of these pairs a *Lexical Similarity Percentage (LSP)* is computed. A list of common meanings of two languages is created and the LSP presents the number of items on this list. There might be factors which are irrelevant to the meanings represented by the list and the LSP is corrected respectively. Lexical similarity might not be enough for classifying languages, because some languages can also have phonological resemblance. To compensate this, a *Phonological Similarity Percentage (PSP)* is calculated. The PSP is subtracted from the LSP and results in a *Subtracted Similarity Percentage (SSP)*. The SSPs serves as a database for the generation of branching structures for languages or phylogenetic trees which represent the classification. By comparing the branching structures of a language to family trees from historical linguistics, the automated classifications are close to the ones of historical linguists (Brown, Holman, Wichmann, & Velupillai, 2008).

ASJP provides a database containing the languages and their corresponding lexical and phonologically transcribed words. Originally, the database was based on the 100-words list of Swadesh (1955) which can be found in the appendix. Currently, the ASJP database includes a list with 40 concepts and 6,139 languages. The database consists of a file which includes all the information needed. The main part is a list with all 40 concepts, all sounds needed for the phonological description, and for each language concept its phonological representation. The data is used for computing a distance matrix and the identified distances are used for computing the trees.

The distance between the languages or concepts are represented in a distance matrix. The distances are computed with an alignment called *Needelman-Wunsch algorithm* (Huson et al., 2010). This algorithm is a global alignment which is applied to two sequences or in this case phonetic representations of two words. This is done for each word pair and a distance matrix is created. The distance matrix is the basis for the reconstruction of a tree. As already stated in section 3, there are different distance-based reconstruction methods. Jäger (2013) compares different distance-based algorithms in his article and discovered that the FastME algorithm is one of the algorithms leading to the best results. The trees in this project are also reconstructed using the FastME algorithm. This computation is done for the reconstruction of the expert tree, as well as for the reconstruction of all concept trees. The output trees of FastME are all binary branched and unrooted. The trees are rooted with respect to an outgroup. The outgroup is

a language or phonological representation which lays outside of a main group to which it is closely related. An example is given below.

The expert tree includes all languages of a specific sample with their phonological representations of the 40 concepts. A sample can contain different languages, as for example only the Germanic languages or all Indo-European languages. A concept tree is reconstructed for each of the 40 concepts contained in each language. There are 40 different distance matrices and 40 different concept trees. The expert tree of all Indo-European languages is displayed in the appendix. Please note that the tree contains 292 languages and is therefore split up in the middle to make it readable. Additionally, I created an expert tree including a sample set of Germanic and Romance languages.

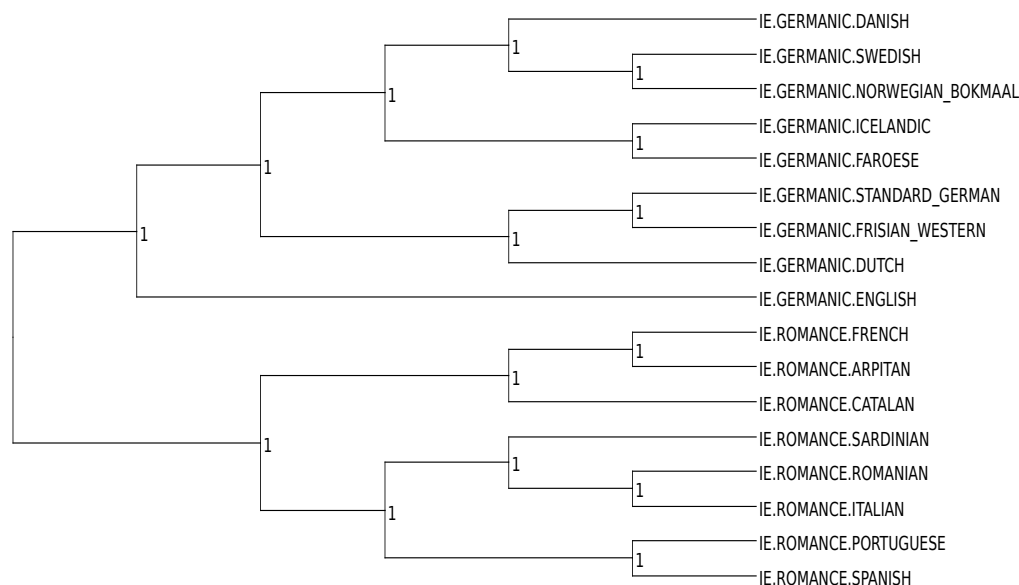


Figure 5.1: The expert tree of Germanic and Romance Languages

The languages are all clustered as expected. All Germanic languages are in one cluster and all Romance languages in another. Within the Germanic languages, the Scandinavian languages (Swedish, Danish, Norwegian) represent one cluster, Icelandic and Faroese are closely related, the West Germanic languages are grouped and English functions as an outgroup. Within the Romance languages, French and its dialect Arpitian are closely related with Catalan. This is not surprising, because a part of Catalonia is now France. Italian, Romanian and Sardinian, as well as Spanish and Portuguese are related as expected. Normally, the tree is rooted on the outgroup. In this case the tree is rooted on two groups and no outgroup is chosen as both group lead back to the root.

is grouped within the Romance languages. This is due to an evolutionary change in its history and this change mostly indicates an evolutionary event.

Before we talk about evolutionary events, there is one thing one needs to be aware of: missing entries. Not every language in the ASJP database contains a phonological transcription for every word. In some languages entries are missing. This is due to the restrictions of the ASJP database or to the missing transcription of languages. For the concept tree above this is not the case, all languages contain an entry with the meaning of mountain. I would like to illustrate the missing entries by using all Germanic languages and their the concept of *thou* or *you*.

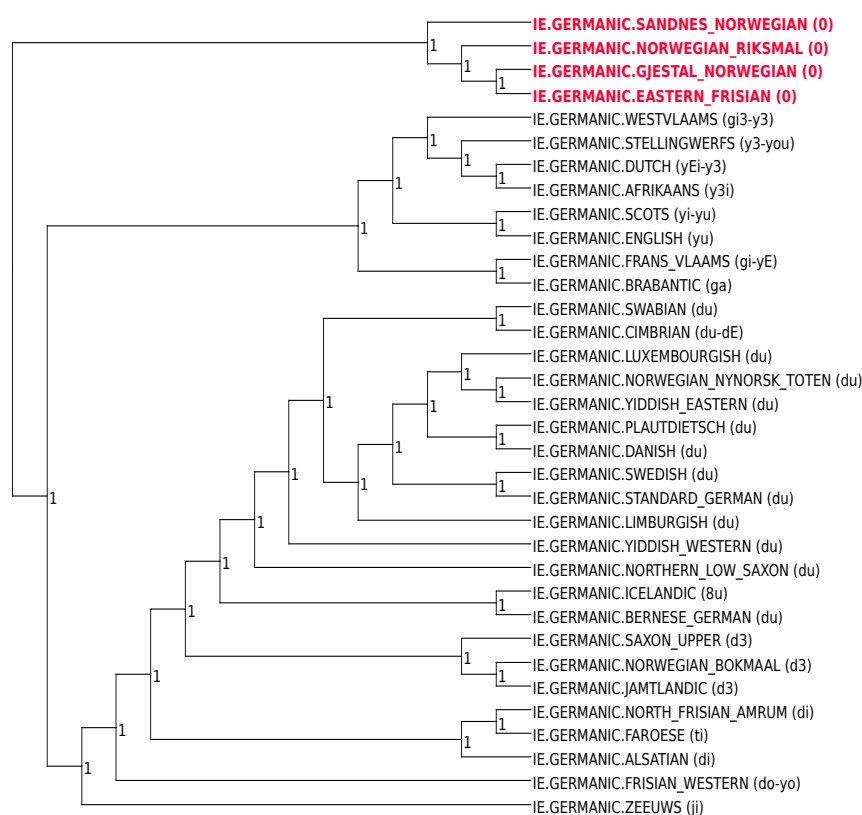


Figure 5.3: The concept tree with the missing entries

What happens is that all languages which do not have an entry for this meaning are related. Those are the first four languages: Sandnes_Norwegian, Norwegian_Riksmal, Gjestal_Norwegian, and Eastern_Frisian which are marked in red. The missing entries are indicated by a 0. Those languages are the outgroup the tree is rooted on. If we want to detect horizontal gene transfer or other evolutionary events within this tree, we need to sort out this group. If we map the concept tree to the species tree, the group would be treated as any other group. The algorithm would detect events and transfers, because the language are moved

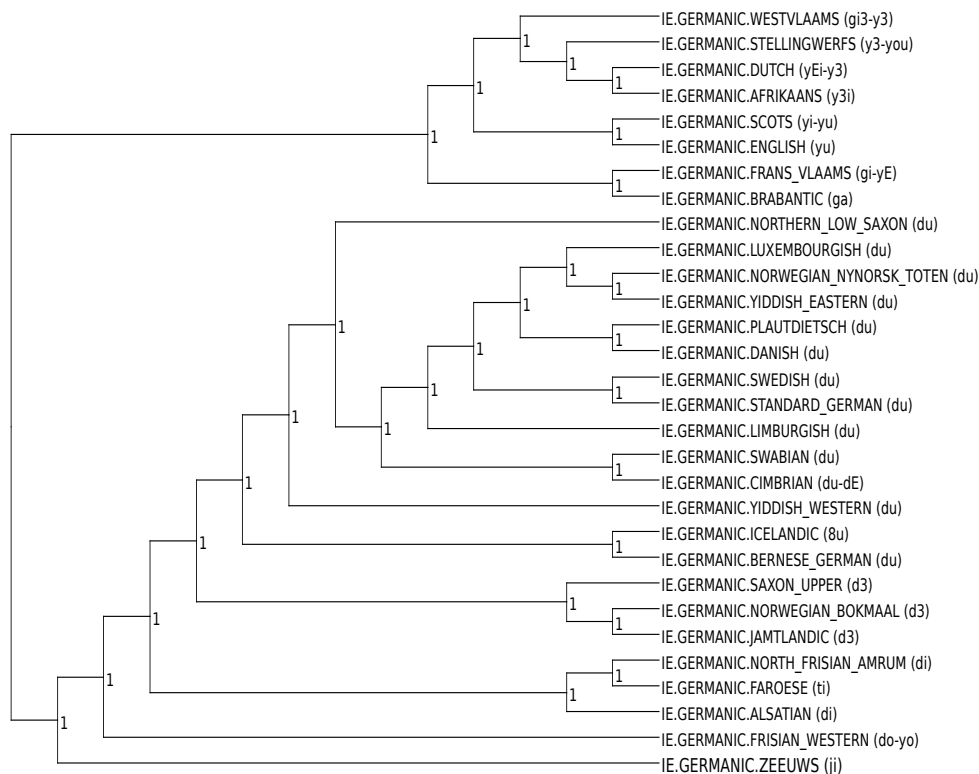


Figure 5.4: The concept tree without the missing entries

out of their original position. This can lead of an unwanted detection of evolutionary events. Therefore, we want to sort out all languages which do not have an entry within a concept. This was done with a implementation taking the list with the original names and checks the entries in the database. All missing entries are indicated with a 0 which makes it easy to sort out the corresponding languages. A new list with all language names present in the language sample is created for each concept. Afterwards, the new list of names is used for sorting out the corresponding lines within the distance matrix and creates a new distance matrix. The new matrix can be used for computing a new tree without the languages with missing entries. This tree is represented in figure 5.4. As one can see, the four languages with the missing entries are no longer present. The only thing done was to remove the outgroup. The other groups are still grouped together according to their phonological representation which can be seen in figure 5.4. Behind the language names, the representations are displayed in brackets. This might not be relevant right know, but for computing horizontal gene transfer events and for reconstructing a network, this step is need for gaining better results.

5.2 Horizontal Gene Transfer

Atkinson and Gray (2005) stated the parallels between biological and linguistic evolution, one of them is horizontal gene transfer and borrowing. This connection gets clearer if we have a look at a description of horizontal gene transfer which Morrison (2011) gave in his book. I already quoted the description in section 4, but I want to repeat it here:

“HGT (horizontal gene transfer) occurs when a small piece of a genome (usually a whole gene) is transferred between unrelated organisms by means other than sexual reproduction.” (p.112)

Morrison (2011) describes horizontal gene transfer with the illustration displayed in figure 5.5.

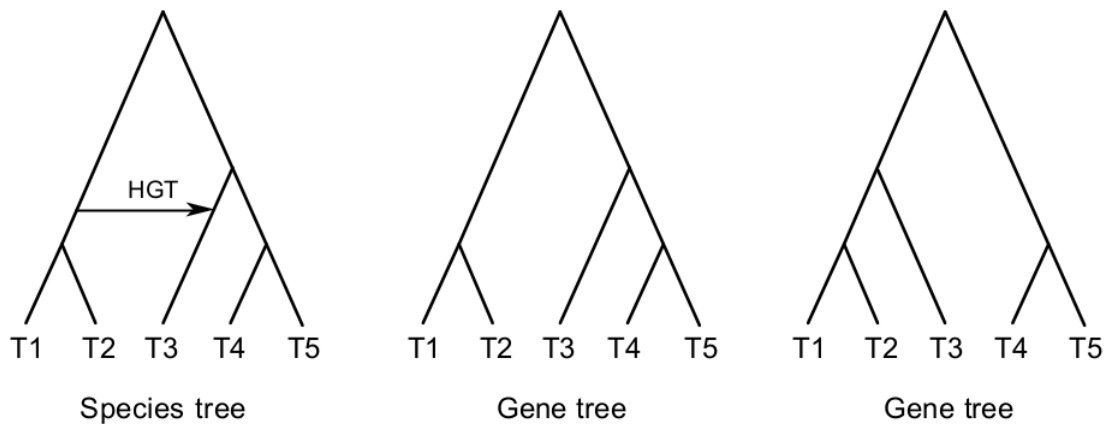


Figure 5.5: Horizontal gene transfer

The illustration displays a species tree representing an organism and two possible gene trees. The nodes of the species tree and the gene tree are labeled with the same taxa. Two of the taxa are involved in the transfer, namely horizontal gene transfer. By a comparison of the right gene tree to the species tree, it can be seen that no transfer is involved. The gene tree represents the same history as the species tree. A comparison between the left gene tree and the species tree indicates different histories. This difference was caused by a transfer. This horizontal gene transfer is marked within the species tree (Morrison, 2011).

In biology, horizontal gene transfer indicates for example exogenous DNA transfer between individual bacteria. Bacteria can acquire genes from other bacteria or from their environment. This acquirement is a horizontal gene transfer and it can lead to significant consequences like the transfer of antibiotic resistance.

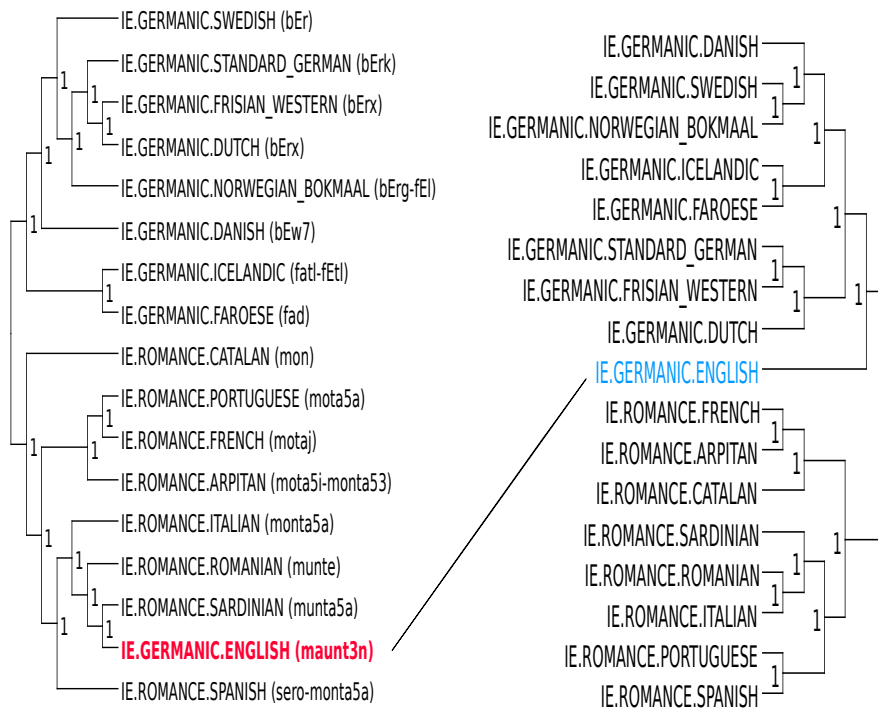


Figure 5.6: Horizontal transfer between the expert tree and the concept tree

The linguistic counterpart to horizontal transfer is borrowing. Borrowing is a process taking place between two individual languages.

Figure 5.6 represents the comparison between the expert tree of the Germanic and Romance languages and the concept tree of mountain. This comparison clearly indicates the transfer of the English language into the group of Romance languages.

Figure 5.7 indicates the result of the comparison. The red arrow indicates the transfer from the Romance languages to the English language. The transfer indicates the borrowing of the word and brings along the adaptation of the word to the English language.

Horizontal gene transfer can be detected using different techniques and methods. Auch (2010) stated three different models in his dissertation. One computational, one similarity and one phylogenetic model. The basic idea of a computational model is a character or sequence based method for detecting genes which deviate from the average composition. The similarity model uses an algorithm to seek similarities between a gene and a group of genes. If the taxonomic distance is larger than expected, it is supposed that the gene derived via transfer. The phylogenetic model uses the mapping of a gene tree to a species tree for the detection of horizontal gene transfer.

The comparison of a species and a gene tree indicates the difference between the

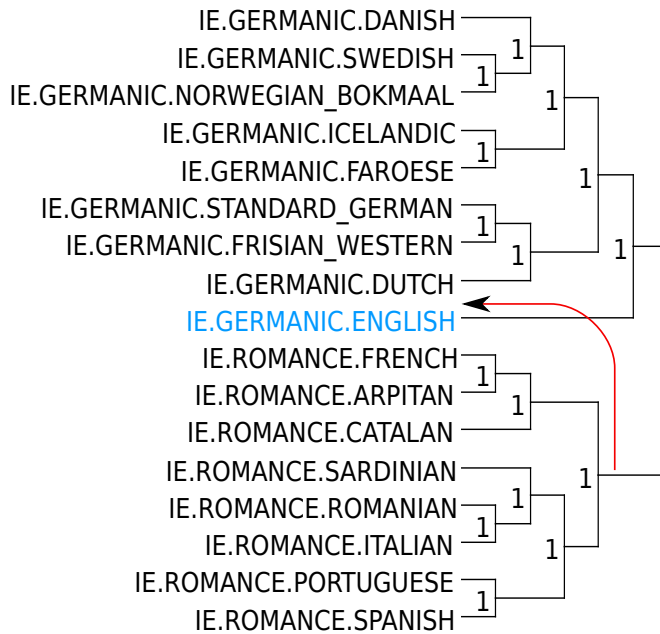


Figure 5.7: Horizontal transfer within the expert tree

two. Those differences can be reconciled by assuming a specific number and a specific type of evolutionary event. Therefore, the mapping of a gene tree into a species tree is also called *reconciliation* or *gene-tree reconciliation* (Morrison, 2011). The difference(s) between the two trees can be computed in many different ways. I will shortly introduce a widely used approach for the detection of horizontal gene transfers. The basic idea is the detection of the number of transfer via mapping a species tree and a gene tree.

Hallett and Lagergren (2001) introduce an approach in which a set of gene trees is mapped to the species tree and the mapping derives a possible reconciliation. The reconciliation explains the evolutionary event. The model is called *subtree transfer model* and comes close to the *SPR (subtree prune and regraft) method*. Within the SPR method, a subtree from the phylogenetic tree is pruned and re-grafted at a different position in the tree (Huson et al., 2010). The number of transformations until it is possible to map the gene tree to the species tree are counted. In figure 5.5, one transformation is needed until the gene tree has the same evolutionary history. The number of transformations indicate the number of transfers. The transformation of all gene trees into the species tree with the smallest number is sought. This number indicate all transfers between the two trees. Hallett and Lagergren (2001) explain the method in more detail and mathematically. The method is implemented and called *LatTrans algorithm* and is widely used within phylogenetics.

An adaption of the method where one gene tree is transformed to fit into one species tree is also common. Boc, Philippe, and Makarenkov (2010) introduce this approach and implement it in an application of the online web server T-REX (Alix, Vladimir, et al., 2012). Pairs of branches of the species tree are tested against the hypothesis that a HGT event has occurred. Then the gene tree is gradually transformed into the species tree using the SPR method also used in the approach of Hallett and Lagergren (2001). Additionally, the direction can be computed using an optimization criteria. There are four possible criteria: the least-square distance described in Boc and Makarenkov (2003), the Robinson-Foulds distance and the Quartet distance described in Huson et al. (2010) and the bipartition dissimilarity described in ? (?). According to Boc et al. (2010), the bipartition dissimilarity criteria has advantages over the other three criteria. The bipartition dissimilarity is defined over a bipartition vector, where the vector indicates the direction of the transfer. For a closer look on the advantages of the bipartition dissimilarity and for a mathematical description of the method, please have a look at Boc et al. (2010). Indicating the direction is a new and interesting outcome within the algorithm. The other algorithms can only compute the HGT events, but not their direction. This algorithm is implemented in the application available on the web server T-REX. The direction of the HGTs are visualized using arrows. The outcome and visualization of the approach are explained in the next section.

There are also other approaches for the detection of horizontal gene transfer using different mathematical methods for the computation. The approach of Boc and Makarenkov (2003) is also implemented in the application used by the web server T-REX. Additionally, there is also another program which is worth to mention here RIATA-HGT. The algorithm represents another approach on the detection of transfer events and can visualize the results. The algorithm is implemented by Nakhleh, Ruths, and Wang (2005).

Each algorithm has its advantages and disadvantages. The algorithm of Boc et al. (2010) is faster than the one introduced by Nakhleh et al. (2005). The algorithm of Nakhleh et al. (2005) and the program RIATA-HGT are included in the software package PhyloNet. The algorithm of Hallett and Lagergren (2001) is implemented in its own software package called *LatTrans*. Both algorithms implemented in T-REX can be used freely and online on the web server. Nevertheless, each algorithm can be used for detecting HGT events and each one results in a good visualization of the transfer events, either in a list or in a network.

5.3 T-REX: a web server

Tree and reticulogram Reconstruction (T-REX) is a web sever including different applications for reconstructing phylogenetic trees and networks and for detecting horizontal gene transfer (HGT) events. It is the only online server which includes the reconstruction of a reticulogram and a network displaying horizontal transfer events. A reticulogram is a special kind of unrooted network and is described in more detail in Alix et al. (2012) and Huson et al. (2010). The web server includes different applications for drawing, computing and validating phylogenetic trees and networks (Alix et al., 2012):

1. Visualizing trees by loading up a phylogenetic tree in a corresponding format
2. Drawing and Modifying trees and saving them in a corresponding format
3. Inferring trees using different distance-based methods
4. Reconstructing trees using a distance matrix with missing values
5. Inferring reticulograms from a distance matrix
6. Detection of horizontal gene transfer events
7. Multiple sequence alignment using two widely used algorithms
8. Transforming sequences into distances
9. Computation of the Robinson-Foulds distance
10. Conversion of a distance matrix into the newick format (for representing trees) and the other way around
11. Generating random phylogenetic trees

The most interesting application is the detection of HGT events. The program uses a gradual reconciliation of a species tree (or expert tree) and a gene tree (or concept tree) to determine an optimal HGT scenario. Within a network, the gene transfers are indicated by an arrow pointing from one gene to another. The arrows are ordered according to their inference.

The program can also be used to detect horizontal transfers between languages. The inputs of the program are an expert tree and a concept tree. The HGT events are computed by using the bipartition dissimilarity described by Boc et al. (2010), the Robinson-Foulds distance described in Huson et al. (2010) and

least-squares coefficient described by Boc and Makarenkov (2003). These computational methods indicate the proximity between a language in the expert tree and a language in the concept tree. The values of the computational methods and the HGT events are all listed in an output file (Alix et al., 2012).

As I stated above, horizontal gene transfer events can be used for the detection of borrowing events between languages. T-REX is applied to an expert tree and a concept tree and the horizontal gene transfer events are computed. The question is whether the application computes the expected results for language borrowing. To make sure the results are correct and can be interpreted in the right way, I use the common example of the English word *mountain*. I chose the expert tree displayed in figure 5.1 and the concept tree of the concept mountain displayed in figure 5.2. Those trees are the inputs for the program on T-REX. The expert tree is the underlying tree structure and the transfer events are indicated by the red arrows.

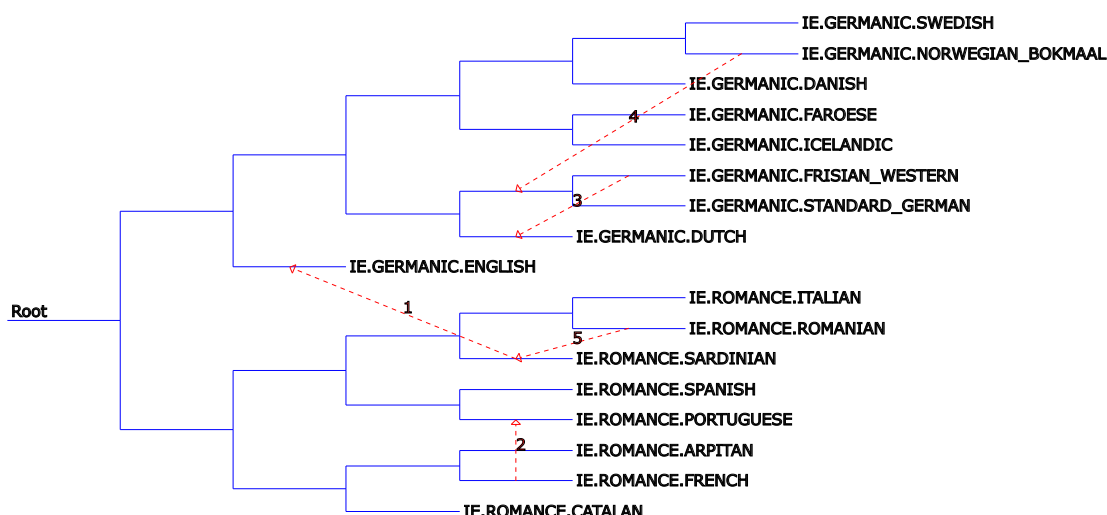


Figure 5.8: The HGT network for Germanic and Romance languages

The arrow from the Romance languages to English indicates the borrowing of the word *mountain*. As one can see, there are two arrows, one from Romanian to the Sardinian and then to English. This is due to the fact that in the concept tree English shares a node with Sardinian and both languages are closely related to Romanian. In the expert tree two arrows indicate this relation between the languages. Interpreting these arrows one might think that English has borrowed *mountain* from Sardinian and Romanian. This is not what we would expect. Figure 5.7 displays the expected transfer where English borrows the word *moun-*

tain from the Romance languages. Actually, English borrowed the word *mountain* from Old French *montaigne*, as already stated in section 2, and not from Sardinian and Romanian. The problem might come from the data. The ASJP database only contains words from presently spoken French and not from older variants. This is the case for all languages present in the ASJP database. Therefore, no connection can be drawn between Old French and English. A concept tree is computed using distance-based methods. Therefore, the languages with the smallest distances are the most closely related. English is related to Sardinian and both are related to Romanian due to the smallest distance. The horizontal transfer is correct with respect to the input data, but it does not represent the historically correct borrowing process of the English word.

If the arrows between Sardinian, Romanian and English indicate borrowing, the other arrows should also indicate borrowings between the languages. This assumption is questionable. The arrows do not indicate borrowing, but the relation of the languages. Within a language family it is not surprising to find cognates. The difficulty is, to distinguish between cognates and loanwords. This cannot be done within the application of T-REX. The program simply links every movement or difference between the expert tree and the concept tree and cannot distinguish between different language phenomena. We also need to keep in mind that in biology homologies are detected differently. For detecting horizontal events between genes no such distinction is needed. For the detection of borrowing between languages, cognates need to be recognizable and taken into account. With the detection of cognates, the program would come to another result and would detect the correct borrowing. This can be done with an adaption of the algorithm, but not and within the application of T-REX.

5.4 Horizontal Language Transfer and LingPy

The tree-based approach represented here and the methods implemented in LingPy have differences and similarities. I would not say that one approach is better than the other, but rather compare the two approaches and see whether they can be combined or not.

The differences between this approach and LingPy starts with the input data. LingPy detects borrowing with gain-loss scenarios or a corresponding model and a reference tree. There are two different methods to compute and analyse borrowings: a parsimony method and a topdown method (List, n.d.-b). In this approach, the input is an expert tree and a concept tree. The expert tree and the reference tree are basically the same. Both represent the relation between different languages. The concept tree can be mapped to an expert tree using

different methods of horizontal gene transfer. Most common is the use of the SPR method and transform the concept tree into the expert tree. The steps are counted, because they can indicate evolutionary events like horizontal transfer. The best transfer and the direction can be computed using different methods like least-square distance, Robinson-Foulds distance, Quartet distance or bipartition dissimilarity. The methods introduced in the tree-based approach and the ones implemented in LingPy are all suited for the detection of borrowing.

The visualization of the borrowings also differ within both methods. LingPy uses the minimal lateral network (MLN) for the representation of borrowings (List & Moran, 2013). The reference tree is the underlying structure of the MLN and the gain-loss scenarios are used for linking the languages. In section 4, two MLN are displayed. The close relation between languages can be explained with cognates. If this is not the case, the relation between languages is due to borrowing. Therefore, the links get their weight from summing over the cognates within a cluster. Within this tree-based approach, there is no fixed resulting network. One possible representation would be a network like the resulting HGT network of T-REX (Alix et al., 2012). The expert tree would also represent the underlying structure and the HGT events are drawn using arrows. The advantage over the MLN would be that the arrows can represent a direction. If we have a look at figure 5.8, the arrow points from the Romance language Romanian to English. Although, there is an intermediate step the direction of the borrowing would be the right one. The expected network is represented in 5.7. The question arise if the expected network can be a result of the algorithm. An implementation of the algorithm within linguistics is needed for answering this question.

The similarities and differences between the two approaches show that non is better than the other. Both can detect horizontal transfer events and display them within a network.

The advantage of LingPy is the detection of cognate sets. The detection is already implemented and the methods for detection borrowing events are based on the cognate detection (List, n.d.-b). This is an important task which need to be integrated in this tree-based approach. If the cognates are detected, links between languages which are not due to borrowing would disappear. The resulting network including cognate detection would differ from the ones displayed above. The advantage of the tree-based approach might be the direction of the borrowing. The arrows in figure 5.8 indicate the right direction of the borrowing. It need to be checked and tested whether this holds also for linguistic data. The scenario we would expect is displayed in figure 5.7. It is questionable if this expected result can be achieved. This is not due to the algorithm but due to the data. If the concept tree is mapped to the expert tree, English is directly related to Sardinian

and Romanian. The algorithm computes the transfer from the position in the concept tree to the position in the expert tree. The English word can not be transfer from the node containing all Romance languages. This is not the fault of the algorithm. For a clearer insight, an implementation of the algorithm within linguistics is needed. The results gives us a better explanation.

The difference between the methods is the abstraction. In LingPy, cognate sets are used to detect evolutionary events. Close related language, where the relation is not due to cognates, are considered to be related because of borrowing. The characters within the cognate clusters need to be known to detect single loanwords. In the tree-based approach, concepts are used to detect different evolutionary events instead of direct cognate sets. The concept refers directly to a word which can be detect as loanword. The method is the more automatic one and is efficient in the detection of single loanwords and the relation between the languages due to borrowing.

Both approaches have an advantage over the other. Nevertheless, they are pretty similar and might work hand in hand. It might not make sense to implement the tree-based approach from scratch. The missing cognate detection would always lead back to use LingPy for this part. So why not use LingPy as a basis for implementing the tree-based approach? The cognate detection can be done with LingPy. Each concept tree contains the different phonological representation of a concept. The cognates could be marked and not be considered within the detection of borrowing. A method with the corresponding tree-based algorithm can be implemented. It is already possible to build a reference or expert tree within LingPy. The same can be done for the concept tree with the marked cognates. The result would be a network where the expert tree is the underlying structure and the horizontal transfer events are indicated by arrows. This is one idea for an implementation of the theoretical approach introduced above.

6 Automatic and Manually Approaches: A Comparison

As we saw in the last sections, automatic approaches for the detection of loanwords are rare but in progress. The usage of computational methods to detect language phenomena are widely studied in the field of linguistics. The idea of an automatic process within linguistics becomes more and more popular, as the field of computational linguistics shows. Nevertheless, automatic approaches for the detection of borrowing are few and until now not widely used within linguistics. But the demand of such processes increases.

The counterpart to an automatic approach is a manual approach. A manual approach is nothing less than for example creating a database from scratch. This is what Haspelmath and Tadmor (2009) did. They created a database of loanwords. The database contains 41 different languages all representing a vocabulary list containing similar words. The database was built manually. For each language an expert translated or transcribed the words in their corresponding language, marking the loanword and even adding additional information.

This chapter should point out the difference between an automatic and a manual approach and it should function as a motivation for the usage of automatic processes within linguistics.

I will first introduce the World Loanword Database, its content, representation and findings. Additionally, I will introduce the Leipzig-Jakarta list which is an alternative for the swadesh list. Afterwards, I will compare the manual approach to an automatic one.

6.1 The World Loanword Database (WOLD)

The World Loanword Database (WOLD) is a database edited by Haspelmath and Tadmor (2009). The database is an example for a collection of languages and their corresponding vocabularies. They marked inherited words and loanwords within different languages. WOLD is an example of a database edited manually and by several authors. It has not yet been done automatically, but it is a great source to look up loanwords.

6.1.1 Background and content of WOLD

The WOLD database is an empirical study of borrowability of words. Haspelmath and Tadmor (2009) started a project called *Loanword Typology (LWT) Project* for representing languages and a part of their corresponding vocabulary where

inherited words and loanwords are marked. There is no comparable project like this and its' therefore unique in his representation.

The goals of the project were to identify lexically borrowed words. For the project, Haspelmath and Tadmor (2009) chose to base their empirical study on classical methods of linguistic typology: (the list is taken from Haspelmath and Tadmor (2009, p. 1))

- (30) a. establishing a world wide sample of languages
- b. surveying the types of loanwords found in these languages, on the basis of a fixed list of lexical meanings
- c. attempting generalizations across the languages of the sample

Those are the main parts which need to be fulfilled for an empirical basis of a study. Before we want to focus on the project, Haspelmath and Tadmor (2009) rises and answers the question why it is important to distinguish between borrowed and inherited words.

- (31) a. It is important to separate inherited words from loanwords, to assess genealogical relatedness between languages. Loanwords confirm the historical contact between languages, although the languages do not belong to the same family.
- b. The lexical borrowing depends of the type of contact. As stated in the second chapter, cultural, political or another situation can lead to borrow words.
- c. The borrowing patterns might be influenced by linguistic factors, like phonology or grammar.

These reasons and the classical methods are the basis and the guidelines of the LWT project.

The LWT project is a collaborative project between different authors. The result was a publication and a database. The different authors are specialists of different languages and their history. Every author worked on his own small project and all small projects ended up into a single grater project, namely the WOLD database. The LWT project ended up in one fixed list of 1,460 items which is called the *LWT meaning list* (Haspelmath & Tadmor, 2009). The authors were asked to provide counterparts for each item on the LWT meaning list and add additional information about the historical circumstances of the borrowing. They could also add additional loanwords to the meaning list which are special or well known in the corresponding language. The WOLD database includes 41 subdatabases of which each representing one language. Each subdatabase contains the words of the language which are the counterparts of the meaning list. It could be the case

that the number of words in the subdatabases varies. This is the case because some authors add additional words, others have to leave out words which are not represented in their language. Each word contains information about orthography, analyzability, loanword status, age of the words, morpheme-by-morpheme gloss and optional information added by the author. Each loanword contains information about the source word and its corresponding language, as well as information about the borrowing circumstances (Haspelmath & Tadmor, 2009). The languages are selected due to “the world’s genealogical, geographical, typological and sociolinguistic diversity” (Haspelmath & Tadmor, 2009). For each language a specialist is needed who would be willing to invest the time and effort of collecting words and information, complete the database and write an article on the work. Haspelmath and Tadmor (2009) stated that their language sample is not ideal and that some language families are over- or under-represented. This is due to the fact that it is hard to find a specialist on each language and who will also support the project. The language sample is not “fully representative of world’s diversity [but] it is much better than anything that existed before [the] project” (Haspelmath & Tadmor, 2009, p. 3).

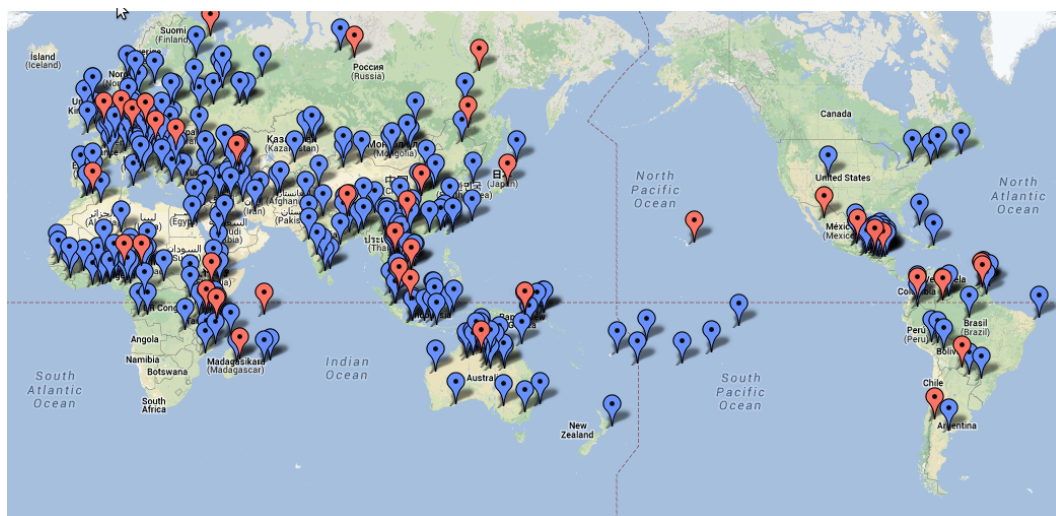


Figure 6.1: A map of the languages in WOLD

The map of the languages in Figure 6.1 is taken from the WOLD webpage (Haspelmath & Tadmor, n.d.). The red symbol indicates all languages included in the database and the blue symbol indicates source languages of loanwords. As one can see, the database includes languages from all over the world and a great distribution over language families. All languages are also listed in Table A.1 in the appendix which is taken from Haspelmath and Tadmor (2009, p. 4).

The LWT meaning list contains 1,460 lexical meanings which have counterparts in any language. It could be the case that there are languages which lack a certain lexical meaning or in other words, do not have a counterpart which represents

this meaning. The lack of a meaning can lead back to cultural or biographical variations. For example, an Amazonian language has no word for *snowshoe* because without snow they do not need snowshoes and have therefore no word representing such a meaning in their language (Haspelmath & Tadmor, 2009). Therefore, it is possible that the number of words in a language varies from the other languages. In the case of additional loanwords in a language, the number of words also varies compared to the other languages. All in all, it is not said that each language contains the same words or lexical meanings. Most of the meanings overlap, but there could also be missing ones and additional ones (Haspelmath & Tadmor, 2009). One should also be aware of the difference between a word and lexical meaning. If this would be a list of words, one would assume that each language contains translations of the words. In a list of meanings, one assumes a transliteration or a transcription of the word. Therefore, the words in the other languages are called counterparts and not translations (Haspelmath & Tadmor, 2009).

The LWT meaning list contains three pieces of information, namely a label, a description of the meaning, and a typical context. For languages, which are originally written in non-Latin scripts, the spelling in the original script can be added additionally. If a language contains two slightly different words for the same meaning, the words are added as one entry in the list. Otherwise, if the words differ greatly, two entries are added to the list representing the same meaning. As said above, the counterpart is more a transcription or transliteration of a meaning, but it need to be a fixed expression in the language. It cannot be a kind of description or explanation of the meaning.

The list of meanings is divided into 24 semantic fields. “Of these, 22 were semantic fields retained from Buck’s (1949) list and Key’s IDS list (slightly renamed in some cases), and two fields were added” (Haspelmath & Tadmor, 2009, p. 6). A list of the semantic fields can be found in Table A.2 in the appendix which is taken from Haspelmath and Tadmor (2009, p. 7).

The words are allocated into their corresponding fields. For most of the words the grouping “is fairly obvious (e.g. animal names in field [Animals], body parts in field [The body]), but in many other cases the grouping of the words is somewhat arbitrary, and alternative groupings are possible but might preferred by other scholars” (Haspelmath & Tadmor, 2009, p. 6). However, the semantic fields are a good way to group the words and to give a first overview of the content. The words receive a *LWT meaning code* to map the word to its corresponding field. Additionally, Haspelmath and Tadmor (2009) assign a word class to each meaning, represented by part-of-speech labels. There are five labels (noun, verb, adjective, adverbs, functional words) representing *things and entities, ac-*

tions and processes, properties, manner and location, and grammatical meanings (Haspelmath & Tadmor, 2009).

The most important part in this project is the information about the borrowed status of a word. The authors identified the loanwords and added a degree of certainty to it. There are five degrees of certainty (Haspelmath & Tadmor, 2009):

- (32)
- a. 0 - no evidence for borrowing
 - b. 1 - very little evidence for borrowing
 - c. 2 - perhaps borrowed
 - d. 3 - probably borrowed
 - e. 4 - clearly borrowed

There is no such degree like “clearly inherited”, because one cannot be sure if the word was borrowed at some earlier time (Haspelmath & Tadmor, 2009). The degree 0 also adds the information that the word might be an inherited word. Therefore, a label like “clearly inherited” is not needed. The information about the age of a word gives information about the time up to which a language can be reconstructed. For the loanwords the age gives information about the time when the word might be borrowed. With this information, the history of the loanwords can be reconstructed. Older and more recent loanwords can be established which gives information for which kind of words are more likely to be borrowed in a specific point in time. This information can be used to reconstruct language contact.

The authors could also add additional information to the loanwords. This information contains the source word and the donor language of the loanword. This information is important and helpful for the reconstruction of language contact, for historical linguistics and the search of the original word. Another additional information is the effect on the lexicon in the borrowing language. It contains the modification of the word in the borrowing language, whether it replaced a word, coexists with a word having the same meaning, or is inserted in the lexicon of the language. The last additional information contains the contact situation of the languages. The authors provided names for the specific situation which led to lexical borrowing (Haspelmath & Tadmor, 2009).

With all this information, the LWT project and the WOLD database provides all needed information for the represented words. Although, there are “only” 41 languages contained in the project, the information leaves nobody’s wishes unfulfilled. There is a lot of information which can be extracted from the database for further studies and projects.

6.1.2 Representation and Findings in WOLD

The LWT project is represented in the World Loanword Database (WOLD) (Haspelmath & Tadmor, n.d.). The WOLD is an online database and can be reached under <http://wold.livingsources.org/>. The website provides all the information from the LWT project in a visualized way.

It is divided into different partitions or categories, representing different information of the languages. All categories are constructed in a similar way, therefore I will give a more detailed explanation of the first category and describe the others in less detail. The first category is the vocabulary.

Home Vocabularies Languages Meanings Authors Newsblog Contact					
RDF+XML					
Vocabularies					
The World Loanword Database consists of the 41 vocabularies listed below. Each vocabulary is a separately citable publication that should be cited as in the following example:					
Schadeberg, Thilo. 2009. "Swahili vocabulary." In: Haspelmath, Martin & Tadmor, Uri (eds.) World Loanword Database. Munich: Max Planck Digital Library, 1625 entries. http://wold.livingsources.org/vocabulary/1					
For the users' convenience the complete reference can be found under "citation".					
ID ↓ [help]	Vocabulary [help]	Authors [help]	Number of words [help]	Percentage of loanwords [help]	
1	Swahili	by Thilo Schadeberg	1625	30	cite
2	Iraqw	by Maarten Mous	1117	15	cite
3	Gawwada	by Mauro Tosco	982	15	cite
4	Hausa	by Ari Awagana & H. Ekkehard Wolff with Doris Löhr	1452	24	cite
5	Kanuri	by Doris Löhr & H. Ekkehard Wolff with Ari Awagana	1427	21	cite
6	Tarifiyt Berber	by Maarten Kossmann	1533	53	cite
7	Seychelles Creole	by Susanne Michaelis with Marcel Rosalie & Katrin Muhme	1880	13	cite

Figure 6.2: A part of the website representing the vocabulary list

The vocabulary contains a list of all 41 languages, their id or count, their corresponding author, the number of words listed for the language, the percentage of loanwords in the language, and a hyperlink for citing the source. The small [help] hyperlinks under each categories give information and an explanation of the category. This hyperlink can be found in each table represented on the webpage. The author's names are hyperlinks too which are linked to a list of all authors and their contact information. By clicking on the languages, another table appears. This table represents all words listed for this language. The words are represented in conjunction with their additional information, like their LWT code, their meaning, their borrowed status (above it is called degrees of certainty), and their source word/language if available. The LWT code can be mapped to the corresponding id of the semantic field and the id of the word. The meaning of the word represents the semantic category to which the word belongs. Again, the meaning is a hyperlink leading to the semantic field and the hyperlink of the

word leads to a description of it.

This is more or less the overall representation of the website. Each category contains a list which represents the corresponding information, while hyperlinks represent the underlying information. Therefore, almost all the information can be found under one category. It can be seen as a many layer database. Firstly, only the most important information for the corresponding category is shown and the hyperlinks lead to the layers directly below this information, the next hyperlinks lead to the next layer containing more detailed information and so on. The next category is the Languages. The map in figure 6.1 shows the languages with their language family and vocabulary are listed, with hyperlinks leading to more information.

The third category is Meaning, referring to the semantic fields. It contains a list of all 24 semantic fields, their id, the number of meanings, the borrowed score, the age score, and the simplicity score. All of the semantic fields function as hyperlinks leading to their subcategories. There is also a complete list of all meanings, containing the LWT code of the words, the semantic category (part-of-speech labels), the semantic field, the borrowed score, the age score, the simplicity score, and the representation.

The website contains more additional information about the authors, a newsblog, a glossary and contact information. On each side on the webpage a rdf file can be downloaded containing the source information as XML (Haspelmath & Tadmor, n.d.).

Haspelmath and Tadmor (2009) stated some results and findings while establishing the database of which one is concerned with the lexical borrowing across the languages. The borrowing rates are different between the languages. This can be due to the fact that some languages have been studied longer and in more detail than others. Therefore, the longer studied languages might contain a more precise representation and classification of loanwords than shorter studies ones. The terms longer/shorter do not only refer to the timespan of the study, but also the history of a language. The more about a languages history is known, the more words can be classified. This is important for loanwords. Loanwords can be integrated at any time in a language. Here again, the more history is revealed about, the more might have been known about language contact and the more loanwords might have been classified.

Another important point for the borrowing rate is the age of the languages. Not all languages are of the same age. For example, *Old High German* is an older language and developed around the year 600 A.D., whereas *Saramaccan*, one of the creole languages developed around 1651, is a much younger language and might have had less time to borrow words (Haspelmath & Tadmor, 2009). “Lex-

ical borrowing is universal” (Haspelmath & Tadmor, 2009, p. 55), as one can see no language in the database which contains only inherited words and no loanwords. Therefore, Haspelmath and Tadmor (2009, p.55) claims that “the average borrowing rate, at 24.4%, is substantial and higher than expected. ” The question arises, if there is a type of language which has a greater tendency to borrow words than others. There is no clear answer to this question. While looking at Table A.3 in the appendix, taken from Haspelmath and Tadmor (2009, p. 56), it is clear that the languages with the highest borrowing rate are all different. They are very distinct in their typological as well as sociolinguistic type. The borrowing rate of each language has to be explained in a specific way rather than in a general explanation.

Another interesting finding in the semantic word classes is the difference between content words and function words. Empirically, it is said that content words are more likely to be borrowed compared to function words. Most of the languages comply with this theory, but three languages do not fulfil the statement. In those three languages, the percentage of borrowing is higher for function words as for content words. Haspelmath and Tadmor (2009) also compared the borrowing rate of nouns and verbs.

Empirically, nouns are presumably more likely to be borrowed compared to verbs. This cannot be said for all languages in the WOLD. Some languages have more borrowed nouns while others have more borrowed verbs. Haspelmath and Tadmor (2009) claim that it has something to do with isolated and synthetic languages. “The more synthetic the language [is], the more adaption is required” (Haspelmath & Tadmor, 2009, p. 63). Most synthetic languages have a complex verb system which makes it more complicated to integrate a new verb in the system. A lot of modifications have to be made to the morphosyntactic system. Therefore, they are less likely to borrow verbs. For isolated languages, it is the other way around. Most of the languages have a simple verbal system and therefore verbs can easily be integrated in the language. Whereas, it cannot be said that isolated languages borrow less nouns. The borrowing rate for nouns is more or less the same over all languages.

Grammatical categories do not play such a significant role here, it is more the reason that names of things and concepts can easily be borrowed and integrated in a language. Nouns can easily be integrated in a system, because most of the languages have a simple noun system. The changes and modifications on the loanwords are less and therefore the nouns are more likely to be borrowed by synthetic languages (Haspelmath & Tadmor, 2009).

Talking about things and concepts, the loanword frequency says a lot about the most borrowed semantic field in the database. The three semantic fields with the

highest loanword frequency and the three semantic fields with the lowest loanword frequency are (Haspelmath & Tadmor, 2009):

- (33)
- a. Religion and Beliefs
 - b. Clothing and grooming
 - c. The house
 - d. The body
 - e. Spatial relations
 - f. Sense perception

The first three semantic fields in (33-a-c) are the fields with the highest loanword frequency. It is intuitive that words from religious context are borrowed into other languages. Religious terminology has been present since the early days and religion plays a crucial role in the history of almost every language. Religion is widely spread over the world and people all over the world who adapt a religion into their culture they adapt also the terminology of the religion. On the other hand, it is also intuitive that the terminology describing parts of the body are less borrowed (Haspelmath & Tadmor, 2009). This goes hand in hand with Swadesh (1955). His list of basic vocabulary contains also body parts and he claims that those parts are present in every language and therefore resistant against borrowing.

6.1.3 Leipzig-Jakarta List

One major result of WOLD is the Leipzig-Jakarta List. The list is named after the location where it was established and created. It represents the 100 words contained in the basic vocabulary list of the database and can be found in the appendix. The list takes all the factors of the project into account, like unborrowed score, the representation score, the simplicity score, and the age score. Those are multiplied to produce a composite score. This score is used to rank the words on the list. Therefore, “it is a full-fledged basic vocabulary ranking” (Haspelmath & Tadmor, 2009, p. 68).

The list introduced by Swadesh (1955) is in some points different to the one of Haspelmath and Tadmor (2009). The Swadesh list is established and edited by Morris Swadesh. He created this list manually and with nothing less than his knowledge. It is claimed that the list is only based on his intuition, but he didn’t get the chance of using modern tools for creating such a list. Haspelmath and Tadmor (2009) however, used the tools of computational linguistics and the internet for creating an “empirically-based basic vocabulary list” (Haspelmath & Tadmor, 2009, p. 72). Both lists contain 100 words, where 62 words overlap in

the lists. This shows that Swadesh (1955) established a good list just with his knowledge, whereas the Leipzig-Jakarta list “has a strong empirical foundation and is thus a more reliable for scientific purposes” (Haspelmath & Tadmor, 2009, p. 73).

6.2 Automatic versus Manually Approaches

Databases are a common tool within linguistics. It is a tool which is widely used and proved to be successful in linguistics. Databases are used to store data in a specific format and visualize the data, so others can use the it. There are many different databases online which can be used for different tasks. For example, the ASJP database, introduced in section 5.5, was used to construct different language trees. The WOLD database is the only one comprising loanwords and their borrowing process.

On the other hand, automatic processes for the detection of loanwords and their borrowings are rare. LingPy is to my knowledge the only software package already implemented. The phylogenetic methods introduced above within the new approach are partially implemented for phylogenetics, but not for linguistics.

The manual and the automatic approach both have advantages and disadvantages over the other. Each approach covers something which is not present in the other approach. For the sake of simplicity, I made a table with the main differences.

Automatic Approach	Manual Approach
fast detection	time costly detection
less precise	more precise
computational methods	human mind
network	vocabulary list
great amount of data	less data

Table 6.1: The main differences between an automatic and a manual approach

The first main difference consists of the time cost of the detection. The automatic approach is quite fast in detecting borrowings. LingPy creates the minimal lateral network within seconds. The algorithm detects the cognates, clusters them, computes a gain-loss scenario, analyses it and computes the MLN. This is all done in a short time span thanks to algorithms. On the contrary, within a database the detection of loanwords is time costly. For each language a specialist goes through the list of more than 1 000 words and checks each word to see if it is a loanword. This is very time consuming. It took years for building the database and detecting all loanwords. The database contains 41 languages and each languages around 1 000 words which makes around 41 000 words within the whole

database. The small database used in the case studies in section 4 for LingPy contain 40 words and originally 9 413 words. We need to keep in mind that this is only a part of the used database IELex. The IELex database contains 152 languages and 32 588 words. If we would use the methods of LingPy on a bigger dataset, it would be slower, but not that much. A software package like LingPy does not need years to detect the loanwords. Therefore, it can be concluded that the automatic approach is much quicker in detecting borrowings than the manual approach.

On the other hand, the point of accuracy also plays an important role. The automatic approach can detect more borrowing, but is it as precise as the manual approach? Within a database the detection of loanwords is precise. With precise I mean not the detection but the information needed to detect borrowing. If we would talk about the accuracy within the detection, the automatic approach might make less mistakes than the manual detection of loanwords. The mistakes of an automatic approach can be corrected by changing or working on the implementation. The mistakes of a human need to be found and corrected manually. It is not said that a human does not make a mistake twice. For avoiding mistakes one has to check the whole work twice and even than it is not said that there are no more mistakes.

The specialist of each corresponding language follows the constructions given by the person responsible for the creation of the database. The detection of loanwords is based on knowledge and research. The specialist knows the history and the evolution of the language in detail and can use his experience for the decision if a word is a loanword. He can clearly describe how and when the loanword surged and developed and trough which processes it went during the adaption. . He can even explain what caused the adaption and the language contact between languages. This detailed knowledge is not present in an automatic approach. The automatic approach depends on its input data and the algorithms for computing the detection. An implemented algorithm cannot have additional thoughts or experience. An automatic approach can be trained on a dataset and this data can be seen as learned words of the algorithm. Neither LingPy not the new tree-based approach takes this into account. The manual approach has no problem with the detection of cognates. They can easily be sorted out or are not even taken into account. In an automatic approach everything needs to be implemented. The cognate detection is only one task which needs to be faced in an automatic approach. The way more difficult task is the direction of borrowing. Each loanword has a source word and a source language. The specialists know this through experience and research. Within an automatic approach, this causes serious problems. The algorithms for detecting HGT events might be a solution

for it, but until now, no one knows how a direction can be assigned to the borrowing. LingPy can clearly detect borrowings and can link the languages, but the direction cannot be shown. Therefore, the source language and the borrowing language cannot clearly be indicated. The languages are linked, but it is not clear which language is the source and which the borrowing language.

This leads me to the next point: computational methods versus human mind. This is a standard argument while comparing computational and manually methods. The human mind always knows more than a computer. The computational methods are only as good as their programmer and computational methods only do the whatever was implemented. Computers will never be as intelligent as humans. The database contains more accurate information than the output of the program. Most would expect this to be the case. The best example is the identification of the source word and the source language. If the computer does not have an input including this information or if the computational method is not able to compute this, the information is simply ignored.

The last point is about visualization. The automatic approaches represent their results within a network. For the construction of a network, a program is necessary. One can create such a network manually but this would again cost time. Within the database, all words contained in the vocabulary of a language including their additional information is visualized by list. A list can easily be expanded if new entries are recorded. The list is alphabetically ordered and the words are easily to find. The disadvantage and at the same time the advantage is that every word has to be looked up to see whether it is a loanword or not. This can be done with the information of the language or of the semantic field. If single loanwords are sought, this representation is adequate while if the set of all borrowings between two languages are sought, this representation is inadequate. The network, on the other hand, arranges this information well. The links between two languages are obvious, but the network misses the information of the single loanwords. LingPy has a method which lists all links between languages. With this list the related languages can be found. The network is only a good visualization for an automatic approach but not for the manual approach. Drawing and creating a network by hand would again be time consuming and would not contribute to the online database. One can parse the WOLD database and get all the information needed for creating a network automatically. However, the representation of a list is adequate for a database.

The advantage of the simple representation of the database makes it user-friendly. The visualization of a network through the automatic approach is quite user-friendly but a computational background is need for the usage of the software package. The network can be interpreted intuitively, but getting to the results

might be challenging for some people. Whereas the database can be used after merely introduction.

The database is not very large with its 41 languages and around 41 000 words. Much larger databases exist, but it is the only database containing loanwords which makes it an acceptable database. It contains a lot of information about loanwords which cannot yet be detected automatically. The databases advantages do not lie in its size, but in the information available to enhance the automatic approaches.

Here once again I would like to emphasize that both approaches have their advantages and disadvantages over the other. The automatic approach is of more interest within the field of linguistics. Since computational methods are pervasive, they are also integrated in linguistics. Nevertheless, scientists are thankful for databases which can be taken as a gold standard to ensure the correctness of their calculated results and the additional information they provide. Without databases the input for the automatic approaches might not be that large. It is even questionable if the automatic approaches would have evolved in the way they did without the presence of digital data. Both approaches are important for the detection of borrowing.

7 Conclusion

The paper shows the connection between biology and linguistics and the usage of phylogenetic methods within historical linguistics. These methods can be modified and integrated in the linguistic field. Not only language classification is an interesting example, but also the detection of borrowing.

The automatic approaches are a step in the right direction within borrowing detection. The phylogenetic methods can be used with some modification in linguistics. The comparison between the manually constructed database and the automatic approach shows that the automatic approach is more efficient within the detection of borrowing. There are some cases which cannot be represented within an automatic approach. The automatic approach cannot make use of all available information. A serious issue is the detection of the direction in which a borrowing event took place. The methods implemented in LingPy cannot differentiate between the source and the borrowing language, while this information is crucial for the borrowing process.

The theoretical approach introduced in this paper shows that the tree-based methods for detecting horizontal gene transfer can be an efficient method to detect borrowings. Some of the methods are represented as online application on a web server. This application uses language trees and is, in contrast to the others, able to represent the direction of the borrowing. Its embedded algorithm can detect the direction of the horizontal gene transfer. If it were modified it might also detect the direction of the borrowing. If this were to be realized and functioning, a great step within the automatic borrowing detection would be taken.

Nevertheless, the approach is only a theoretical one. This paper showed that the methods used in the tree-based approach are adequate for detecting borrowings between languages. A next step would be the implementation. An idea would be an implementation integrated into LingPy. It would be useful to integrate the approach in the software package. The phylogenetic methods cannot be used from scratch, they need to be modified for the usage in linguistics. One problem stated was the detection of cognates. Such methods are already implemented in LingPy. Therefore, an implementation of the approach would always lead back to the usage of the cognate detection methods. The tree-based approach would fit into the package and a new network could easily be integrated. Another idea would be to modify the existing phylogenetic methods.

A further thought would include a usage of the DLT network. The computation of DLT scenarios can be used to detect additional events, namely duplication and loss events. These duplication and loss events can indicate the duplication or the loss of a word. A word is also borrowed within another language, if it denotes the

absence of a word with such meaning. If the concept trees are modified by sorting out all languages with a missing entry of a word, the loss can also be due to a missing entry. This network and its scenarios are not considered in this study, but would definitely be worth a thought. It could be another advantage within the detection of borrowing.

The tree-based approach shows the adequacy of the methods generally within linguistics and in specific for the detection of borrowing. A theoretical explanation is not sufficient. The implementation is crucial for the approach. It is the only way to test the methods with linguistic data and check if the results are as expected.

References

- Alix, B., Vladimir, M., et al. (2012). T-rex: a web server for inferring, validating and visualizing phylogenetic trees and networks. *Nucleic acids research*, 40(W1), W573–W579. Available from <http://www.trex.uqam.ca/>
- Atkinson, Q. D., & Gray, R. D. (2005). Curious parallels and curious connections—phylogenetic thinking in biology and historical linguistics. *Systematic Biology*, 54(4), 513–526.
- Auch, A. (2010). *A phylogenetic potpourri: computational methods for analysing genome-scale data*. Hochschulschrift.
- Baugh, A. C. (1935). The chronology of french loan-words in english. *Modern Language Notes*, 50(2), 90–93.
- Boc, A., & Makarenkov, V. (2003). New efficient algorithm for detection of horizontal gene transfer events. In *Algorithms in bioinformatics* (pp. 190–201). Springer.
- Boc, A., Philippe, H., & Makarenkov, V. (2010). Inferring and validating horizontal gene transfer events using bipartition dissimilarity. *Systematic biology*, 59(2), 195–211.
- Brown, C. H., Holman, E. W., Wichmann, S., & Velupillai, V. (2008). Automated classification of the world's languages: a description of the method and preliminary results. *STUF-Language Typology and Universals*, 61(4), 285–308.
- Bußmann, H. (Ed.). (2008). *Lexikon der sprachwissenschaft: mit ...14 tabellen ...*. Stuttgart: Kröner.
- Dagan, T., Artzy-Randrup, Y., & Martin, W. (2008). Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. *Proceedings of the National Academy of Sciences*, 105(29), 10039–10044.
- Darwin, C. (1871). *The descent of man*. D. Appleton and Company.
- Delz, M., Layer, B., Schulz, S., & Wahle, J. (2012, March). Overgeneralisation of verbs - the change of the german verb system. In *Proceedings of the 9th international conference on the evolution of language* (p. 96-103). Kyoto, Japan.
- Doolittle, W. F. (1999). Phylogenetic classification and the universal tree. *Science*, 284(5423), 2124–2128.
-

- Dunn, M. (n.d.). *Indo-european lexical cognacy database*. Available from <http://ielex.mpi.nl/>
- Eldredge, N. (2005). *Darwin: discovering the tree of life*. New York [u.a.]: Norton.
- Felsenstein, J. (2004). *Inferring phylogenies*. Sunderland, Mass.: Sinauer. Available from <http://www.ulb.tu-darmstadt.de/tocs/103801863.pdf>
- Geisler, H., & List, J.-M. (n.d.). *Beautiful trees on unstable grounds: Notes on the data problem in lexicostatistics*. Wiesbaden.
- Haeckel, E. H. P. A. (1874). *Anthropogenie oder entwicklungsgeschichte des menschen*. Leipzig: Verlag von Wilhelm Engelmann.
- Hall, B. G. (2005). *Phylogenetic trees made easy: a how-to manual* (2. ed., 2. print. ed.). Sunderland, Mass.: Sinauer.
- Hallett, M. T., & Lagergren, J. (2001). Efficient algorithms for lateral gene transfer problems. In *Proceedings of the fifth annual international conference on computational biology* (pp. 149–156).
- Harper, D. (n.d.). *Online etymology dictionary*. Available from <http://www.etymonline.com/index.php>
- Haspelmath, M., & Tadmor, U. (n.d.). *World loanword database*. Available from <http://wold.livingsources.org/>
- Haspelmath, M., & Tadmor, U. (2009). *Loanwords in the world's languages: a comparative handbook*. Walter de Gruyter.
- Haugen, E. (1950). The analysis of linguistic borrowing. *Language*, 26(2), 210–231.
- Huson, D. H., Rupp, R., & Scornavacca, C. (2010). *Phylogenetic networks: concepts, algorithms and applications*. Cambridge University Press.
- Hyman, L. M. (2010). The role of borrowing in the justification of phonological grammars. *Studies in African linguistics*, 1(1).
- Jacobs, H., & Gussenhoven, C. (2000). Loan phonology: perception, salience, the lexicon and ot. *Optimality Theory: Phonology, syntax, and acquisition*, 193–209.
- Jäger, G. (2013). *Evaluating distance-based phylogenetic algorithms for automated language classification*.
-

- Jin, G., Nakhleh, L., Snir, S., & Tuller, T. (2007). Inferring phylogenetic networks by the maximum parsimony criterion: a case study. *Molecular Biology and Evolution*, 24(1), 324–337.
- Joseph, B. D., & Janda, R. D. (2003). *The handbook of historical linguistics*. Wiley Online Library.
- Kemmer, S. (n.d.). *Loanwords: Major periods of borrowing in the history of english*. Available from <http://www.ruf.rice.edu/kemmer/Words/loanwords.html>
- Lecointre, G. (2006). *The tree of life: a phylogenetic classification* (H. Le Guyader, Ed.). Cambridge, MA: Belknap Press of Harvard Univ. Pr.
- List, J.-M. (n.d.-a). *Improving phylogeny-based network approaches to investigate the history of the chinese dialects*.
- List, J.-M. (n.d.-b). *Lingpy documentation*. Available from www.pypi.python.org/pypi/lingpy/2.0
- List, J.-M. (n.d.-c). *Sequence comparison in historical linguistics*.
- List, J.-M., & Moran, S. (2013, August). An open source toolkit for quantitative historical linguistics. In *Proceedings of the 51st annual meeting of the association for computational linguistics: System demonstrations* (pp. 13–18). Sofia, Bulgaria: Association for Computational Linguistics. Available from <http://www.aclweb.org/anthology/P13-4003>
- List, J.-M., Nelson-Sathi, S., Martin, W., & Geisler, H. (n.d.). *Language dynamics and change: Using phylogenetic networks to model chinese dialect history*.
- Minett, J. W., & Wang, W. S.-Y. (2003). On detecting borrowing: distance-based and character-based approaches. *Diachronica*, 20(2), 289–331.
- Moira, Y. (1993). Cantonese loanword phonology and optimality theory. *Journal of East Asian Linguistics*, 2(3), 261–291.
- Morrison, D. A. (2011). *Introduction to phylogenetic networks*. Uppsala, Sweden: RJR Productions.
- Nakhleh, L., Ruths, D., & Wang, L.-S. (2005). Riata-hgt: a fast and accurate heuristic for reconstructing horizontal gene transfer. In *Computing and combinatorics* (pp. 84–93). Springer.
-

- Nelson-Sathi, S., List, J.-M., Geisler, H., Fangerau, H., Gray, R. D., Martin, W., et al. (2011). Networks uncover hidden lexical borrowing in indo-european language evolution. *Proceedings of the Royal Society B*, 278(1713), 1794–1803.
- Nichols, R. (2001). Gene trees and species trees are not the same. *Trends in Ecology & Evolution*, 16(7), 358–364.
- Olah, B. (2007). English loanwords in japanese: Effects, attitudes and usage as a means of improving spoken english ability. *Bunkyo Gakuin Daigaku Ningengakubu Kenkyuu Kiyu*, 9(1), 177–188.
- Paradis, C., & LaCharité, D. (1997). Preservation and minimality in loanword adaptation. *Journal of Linguistics*, 33(02), 379–430.
- Penny, D. (2011). Darwin’s theory of descent with modification, versus the biblical tree of life. *PLoS biology*, 9(7), e1001096.
- Peperkamp, S., & Dupoux, E. (2003). Reinterpreting loanword adaptations: the role of perception. In *Proceedings of the 15th international congress of phonetic sciences* (Vol. 367, p. 370).
- Ringe, D., Warnow, T., & Taylor, A. (2002). Indo-european and computational cladistics. *Transactions of the philological society*, 100(1), 59–129.
- Rose, Y. (2012). Perception, representation, and correspondence relations in loanword phonology. In *Proceedings of the annual meeting of the berkeley linguistics society* (Vol. 25).
- Schleicher, A. (1873). *Die darwinsche theorie und die sprachwissenschaft* (2. ed. ed.). Weimar: Hermann Böhlau.
- Silverman, D. (1992). Multiple scansions in loanword phonology: evidence from cantonese. *Phonology*, 9(2), 298–328.
- Southworth, F. C. (1964). Family-tree diagrams. *Language*, 40(4), 557–565.
- Swadesh, M. (1955). Towards greater accuracy in lexicostatistic dating. *International journal of American linguistics*, 21(2), 121–137.
- Vendelin, I., & Peperkamp, S. (2004). Evidence for phonetic adaptation of loanwords: an experimental study. *Actes des Journées d’Etudes Linguistique*, 129–131.
- Vendelin, I., & Peperkamp, S. (2006). The influence of orthography on loanword adaptations. *Lingua*, 116(7), 996–1007.
-

- Volland, B. (1986). *Französische entlehnungen im deutschen: Transferenz und integration auf phonologischer, graphematischer, morphologischer und lexikalisch-semantischer ebene* (Vol. 163). Walter de Gruyter.
- Wardhaugh, R. (2009). *An introduction to sociolinguistics*. Wiley-Blackwell.
- Wichmann, S., Müller, A., Velupillai, V., Wett, A., Brown, C. H., Molochieva, Z., et al. (2012). *The asjp database*. Available from <http://wwwstaff.eva.mpg.de/wichmann/ASJPHomePage.htm>
- Wiley, E. O., & Lieberman, B. S. (2011). *Phylogenetics: theory and practice of phylogenetic systematics*. John Wiley & Sons.
- Yip, M. (2006). The symbiosis between perception and grammar in loanword phonology. *Lingua*, 116(7), 950–975.
-

A Some Information on the WOLD database

Table A.1: The LWT project languages

Language	Affiliation	Main location(s)
Archi	Lezgian, Nakh-Daghestanian	Daghestan, Russian Federation
Bezhta	Tsezic, Nakh-Daghestanian	Daghestan, Russian Federation
Ceq Wong	Aslian, Austro-Asiatic	West Malaysia
Dutch	Germanic, Indo-European	Netherlands
English	Germanic, Indo-European	Britain, USA, Canada, Australia
Gawwada	Cushitic, Afro-Asiatic	Ethiopia
Gurindji	Pama-Nyungan	Australia
Hausa	Chadic, Afro-Asiatic	Nigeria, Niger
Hawaiian	Polynesian, Austronesian	Hawai'i
Hup	Nadahup	Brazil, Colombia
Imbabura Quechuan	Quechuan	Ecuador
Indonesian	Malayic, Austronesian	Indonesian
Iraqw	Cushitic, Afro-Asiatic	Tanzania
Japanese	Japanese-Ryukyuan	Japan
Kali'na	Cariban	Venezuela
Kanuri	Saharan	Nigeria, Niger
Ket	Yeniseian	Russia
Kildin Saami	Uralic	Russia
Lower Sorbian	Slavic, Indo-European	Germany
Malagasy	Southeast Barito, Austronesian	Madagascar
Manange	Bodish, Sino-Tibetan	Nepal
Mandarin Chinese	Sinitic, Sino-Tibetan	China
Mapudungun	(isolate)	Chile, Argentina
Old High German	Germanic, Indo-European	Northern Germany
Oroqen	Tungusic	China
Otomi	Otomanguean	Mexico
Q'eqchi'	Mayan	Guatemala, El Salvador, Belize
Romanian	Romance, Indo-European	Romania
Sakha	Turkic	Siberia
Saramaccan	English-based creole	Surinam
Selice Romani	Indo-Iranian, Indo-European	Slovakia
Seychelles Creole	French-based creole	Seychelles
Swahili	Banut, Niger-Congo	Tanzania, Kenya, Uganda, D. R. Congo
Takia	Oceania, Austronesian	Papua New Guinea
Thai	Tai-Kadai	Thailand
Tarifit Berber	Afro-Asiatic	Morocco
Vietnamese	Viet-Muong, Austro-Asiatic	Vietnam
White Hmong	Hmong-Mien	Laos
Yaqui	Uto-Aztecan	Mexico
Wichí	Mataco-Mataguayan	Argentina, Bolivia
Zinacantán Tzotzil	Mayan	Mexico

Table A.2: The semantic fields

	Semantic Field	Number of meaning
1	The physical world	75
2	Kinship	85
3	Animals	116
4	The body	159
5	Food and drink	81
6	Clothing and grooming	59
7	The house	47
8	Agriculture and vegetation	74
9	Basic actions and technology	78
10	Motion	82
11	Possession	46
12	Spatial relations	75
13	Quantity	38
14	Time	57
15	Sense perception	49
16	Emotions and values	48
17	Cognition	51
18	Speech and language	41
19	Social and political relations	36
20	Warfare and hunting	40
21	Law	26
22	Religion and belief	26
23	Modern world	57
24	Miscellaneous function words	14
	total	1,460

Table A.3: The lexical borrowing rates

Borrowing Type	Languages	Total words	Loanwords	Loanwords as % of total
Very high borrowers	Selice Romani	1,431	898	62,7%
	Tarifiyt Berber	1,526	789	51,7%
High borrowers	Gruindij	842	384	45,6%
	Romanian	2,137	894	41,8%
	English	1,504	617	41,0%
	Saramaccan	1,089	417	38,3%
	Ceq Wong	862	319	37,0%
	Japanese	1,975	689	34,9%
	Indonesian	1,942	660	34,0%
	Bezhta	1,344	427	31,8%
	Kildin Saami	1,336	408	30,5%
	Imbabura Quechua	1,158	350	30,2%
	Archi	1,112	328	29,5%
	Sakha	1,411	409	29,0%
	Vietnamese	1,477	415	28,1%
	Swahili	1,610	447	27,8%
	Yaqui	1,379	366	26,5%
	Average borrowers	Thai	2,063	539
Takai		1,123	291	25,9%
Lower Sorbian		1,671	374	22,4%
Hausa		1,452	323	22,2%
Mapudungun		1,236	274	22,2%
White Hmong		1,290	273	21,2%
Kanuri		1,427	283	19,8%
Dutch		1,513	289	19,1%
Malagasy		1,526	267	17,5%
Zinacantán Tzotzil		1,217	195	16,0%
Wichí		1,187	188	15,8%
Q'eqchi'		1,774	266	15,0%
Iraqw		1,117	162	14,5%
Kali'na		1,110	156	14,0%
Hawaiian		1,245	169	13,6%
Oroqen		1,138	137	12,0%
Hup		993	114	11,5%
Gawwada		982	111	11,3%
Seychelles Creole		1,879	201	10,7%
Otomi		2,158	231	10,7%
Low borrowers	Ket	1,030	100	9,7%
	Manange	1,009	84	8,3%
	Old High German	1,203	70	5,8%
	Mandarin Chinese	2,042	25	1,2%

B The Swadesh 100-word list

Swadesh's 100-word list

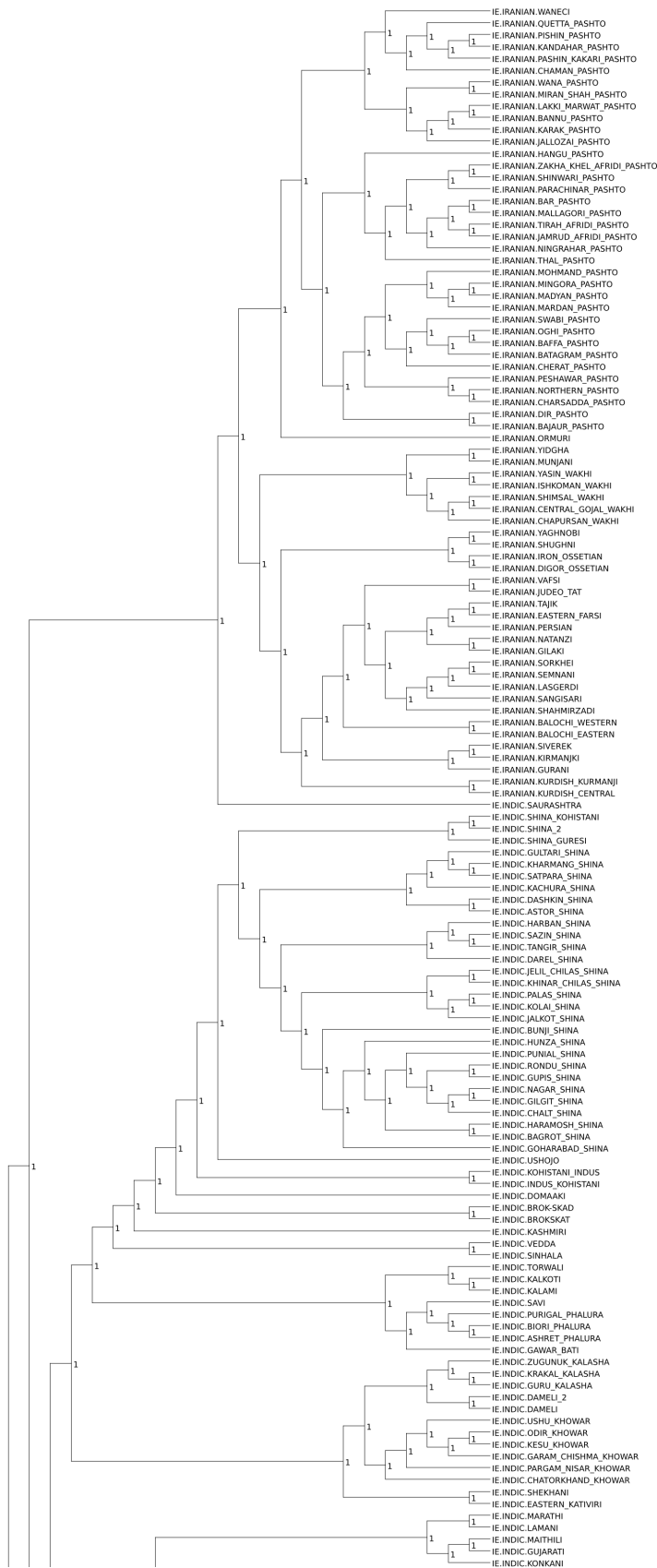
1. I	31. bone	61. die	91. black
2. thou	32. grease	62. kill	92. night
3. we	33. egg	63. swim	93. hot
4. this	34. horn	64. fly	94. cold
5. that	35. tail	65. walk	95. full
6. who?	36. feather	66. come	96. new
7. what?	37. hair	67. lie	97. good
8. not	38. head	68. sit	98. round
9. all	39. ear	69. stand	99. dry
10. many	40. eye	70. give	100. name
11. one	41. nose	71. say	
12. two	42. mouth	72. sun	
13. big	43. tooth	73. moon	
14. long	44. tongue	74. star	
15. small	45. fingernail	75. water	
16. woman	46. foot	76. rain	
17. man	47. knee	77. stone	
18. person	48. hand	78. sand	
19. fish	49. belly	79. earth	
20. bird	50. neck	80. cloud	
21. dog	51. breasts	81. smoke	
22. louse	52. heart	82. fire	
23. tree	53. liver	83. ash	
24. seed	54. drink	84. burn	
25. leaf	55. eat	85. path	
26. root	56. bite	86. mountain	
27. bark	57. see	87. red	
28. skin	58. hear	88. green	
29. flesh	59. know	89. yellow	
30. blood	60. sleep	90. white	

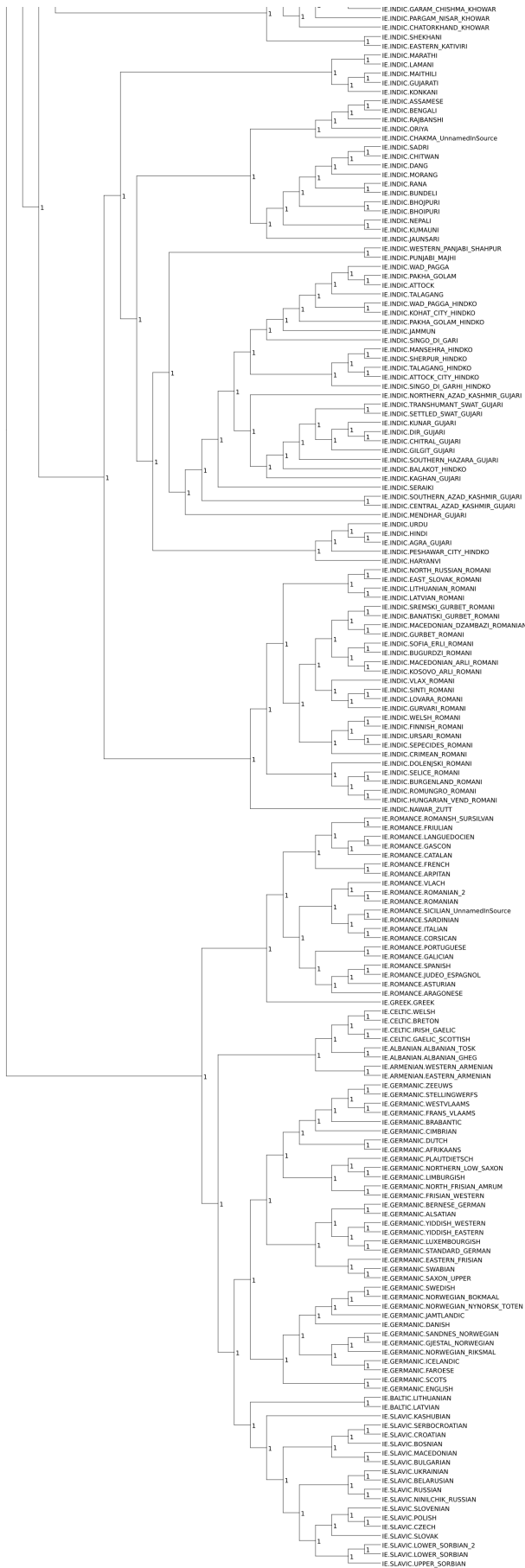
C The Leipzig-Jakarta 100-word list

Leipzig-Jakarta 100-word list

1. ant	34. to go	68. rope
2. arm/hand	35. good	69. to run
3. ash	36. hair	70. salt
4. back	37. hard	71. sand
5. big	38. he/she/it/him/her	72. to say
6. bird	39. to hear	73. to see
7. to bite	40. heavy	74. shade/shadow
8. bitter	41. to hide	75. skin/hide
9. black	42. to hit/to beat	76. small
10. blood	43. horn	77. smoke
11. to blow	44. house	78. soil
12. bone	45. I/me	79. to stand
13. breast	46. in	80. star
14. to burn (intransitive)	47. knee	81. stone/rock
15. to carry	48. to know	82. to suck
16. child (reciprocal of parent)	49. to laugh	83. sweet
17. to come	50. leaf	84. tail
18. to crush/to grind	51. leg/foot	85. to take
19. to cry/to weep	52. liver	86. thick
20. to do/to make	53. long	87. thigh
21. dog	54. louse	88. this
22. drink	55. mouth	89. to tie
23. ear	56. name	90. tongue
24. to eat	57. navel	91. tooth
25. egg	58. neck	92. water
26. eye	59. new	93. what?
27. to fall	60. night	94. who?
28. far	61. nose	95. wide
29. fire	62. not	96. wind
30. fish	63. old	97. wing
31. flesh/meat	64. one	98. wood
32. fly	65. rain	99. yesterday
33. to give	66. red	100. you (singular)
	67. root	

D Expert Tree of the Indo-European languages





E Concept Tree “Mountain” of the Indo-European languages

