IN SUPREMÆ DIGNITATIS 1343

Dipartimento di Linguistica "Tristano Bolelli"

Dottorato di Ricerca in Linguistica Generale, Storica, Applicata, Computazionale e delle Lingue Moderne

Tesi di dottorato:

A quantitative approach to the study of syntactic evolution

Candidato:	*Tutori*:
Gabriele Rigon	Prof. Giuseppe Longobardi
	Dott. Alessandro Lenci

Presidente: Prof. Giovanna Marotta

Triennio 2006-2008

Acknowledgments

I am grateful to many people that have supported me in the last years. First I wish to thank Prof. Giuseppe Longobardi and Dr. Alessandro Lenci, the supervisors of my work, for their constant presence and for having guided the improvement of my skills, often pointing me the way for the development of the research project.

I am especially grateful to Prof. Giovanna Marotta, Prof. Franco Fanciullo and Prof. Romano Lazzeroni as well, for having always welcomed me and for having read this work since its early stages, providing me with valuable feedback and explanations.

Many times at the Department of Mathematics and Informatics of Trieste I have also taken advantage of the helpful assistance of Prof. Andrea Sgarro and Dr. Luca Bortolussi, who have frequently offered me elucidations on mathematical questions; Moreover, Luca has written important components of the novel programs employed in the experimentation.

In addition, I would like to express my thankfulness to some persons I met at the Department of Linguistics of UCLA, in particular to Prof. Edward Stabler, who dedicated much time and attention to me, and to Prof. Edward Keenan, Prof. Marcus Kracht and Prof. Martin Prinzhorn, from whom I received important comments and suggestions on diverse issues. Among the people that have been visiting (or are visiting) Trieste, instead, I wish to gratefully acknowledge the contribution of Dr. Marco René Spruit and of Prof. Tania Kuteva, who kindly sent me their papers, and that of Prof. Robert Frank, to whom I have recently had the opportunity to introduce the work.

My deepest thanks are to my family and, above all, to Silvia, for the inestimable support and encouragement they have given me every day.

Contents

1. Introduction

This dissertation covers the experimentation of quantitative algorithmic procedures for the study of language evolution. In particular, the inquiry is based on the application of quantitative methods originally designed within molecular biology and population genetics to a syntactic comparative dataset (Longobardi 2003a; 2003b; Guardiano and Longobardi 2005; Gianollo *et alii* 2008; in press; Longobardi and Guardiano 2009): The goal is to infer hypotheses regarding genealogical relationships between a specific set of languages, accounting for the role of areal convergence in linguistic variation, and to evaluate them in light of the traditional accounts provided by historical linguistics.

This dissertation is organized as follows. Chapter 2 comprises the comparison between language evolution and biological evolution. Notice that here the term "language evolution" refers to the diachronic development of languages, not to the evolution of the language faculty, which is a completely different issue (see section 2.2.1). The idea is that some important features of language development may also be identified drawing a parallel with the biological domain, in particular referring to a general model of cultural evolution as an intermediate step. In section 2.1 the "modern synthesis" of biological evolution is presented: The subject is introduced by an historical overview of the development of the evolutionary theory starting from the pre-Darwinian explanations of species diversity. Section 2.2 provides an outline of some of the most relevant parallels between language evolution and biological evolution put forward so far and the discussion is preceded by an historical account of the use of evolutionary ideas in linguistics as well. Section 2.3 is dedicated to the analogical comparison: The question asked is whether the fundamental factors that drive biological evolution may also apply to cultural evolution in general and to linguistic evolution in particular. On the whole, this comparison seems to show that language evolution and biological evolution are considerably different in some respects. However, the dissimilarities do not prevent the application of quantitative reconstruction procedures.

The evolutionary account outlined in chapter 2 appears to be perfectly compatible with the most recent generative views on syntactic change, which are taken into consideration in chapter 3. Section 3.1 covers the basic methodological points and concepts of the parametric approach developed within generative grammar. In particular, in section 3.1.1 the methodological plan defined by Chomsky (1964) and Longobardi (2003a) is presented. Then section 3.1.2 and section 3.1.3 are intended to respectively illustrate the basic notions regarding the cognitive-biolinguistic and the formal aspects of generative grammar. Once the parametric perspective on synchronic language variation is clarified, section 3.1.4 is dedicated to the illustration of some current hypotheses on the forms of parameters, while section 3.1.5 covers the issue of their interaction. The views on syntactic change developed from the synchronically-based framework of the Principles and Parameters Theory are introduced in section 3.2. After having elucidated the link between the possible descriptive typologies of syntactic change and the general process of reanalysis, sections 3.2.2 and 3.2.3 are dedicated to the extension of the parametric approach to the explanation of historical facts and to the issue of gradualness. On this ground, some recent attempts to model population-level (or better E-level) parametric evolution using computational simulation are taken into account. "Syntactic borrowing" is discussed throughout section 3.3: In particular, in section 3.3.1 the meaning of "borrowing" that is relevant for the present work is defined in light of past and present proposals regarding the typology of contact phenomena and evidence in favour of the possibility of syntactic convergence is introduced. Finally, the focus of section 3.3.2 is the interpretation of syntactic borrowing within the parametric perspective.

Having outlined an appropriate evolutionary-oriented description of diachronic syntax, chapter 4 is then dedicated to the clarification of the aims of the present work and to the presentation of diverse methods of comparison adopted in historical linguistics and population genetics. Therefore in section 4.1 the meaning and the role of “genealogical classification” in linguistics are taken into account, together with the formal description of the tree-like and network-like representations of evolutionary paths. Then a critical survey of the relevant comparative procedures is illustrated, including the classical comparative method (section 4.2.1), lexicostatistic analysis and mass comparison (section 4.2.2), and the comparative approach of population genetics (section 4.2.3). The introduction of the “Parametric comparison method”, instead, is covered in section 4.2.4: The parallel between it and the procedures of investigation used in molecular biology paves the way to the discussion regarding the quantitative techniques of phylogenetic reconstruction that are relevant here, which is done in section 4.3. In particular, sections 4.3.1 and 4.3.2 concern the presentation of character-based and distance-based methods, while some quantitative approaches that can be used to detect contact and events of convergence between genealogically-separated lineages are illustrated in section 4.3.3.

Chapter 5 is dedicated to the presentation of the data. In section 5.1 an overview of the principal datasets used so far to perform quantitative investigations on the history of languages is provided, separating between lexical-based (5.1.1) and non-lexical-based (5.1.2) lines of inquiry. This serves to ground the introduction of the parametric dataset adopted in the present work (section 5.2). In particular, the data in question concern the parametric analysis of the nominal phrase of a specific set of languages (Indo-European and non-Indo-European, ancient and modern). In the course of the discussion the reason why the analysis is focused on a particular syntactic domain is explained and an account of the relevant hypothesis on nominal syntax is provided (section 5.2.1). The presentation of the parametric dataset, i.e. “Table A”, is covered in section 5.2.2. Section 5.3, instead, is dedicated to overview of “traditional” and quantitative-based proposals regarding in particular the genealogical classification of the languages included in Table A and provides a basis to evaluate the results of the experimentation.

The quantitative analyses are illustrated in chapter 6. Section 6.1.1 covers a preliminary character-based and distance-based review of the dataset, together with an account of the measure of distance between languages that is assumed. In section 6.1.2 the choice of the phylogenetic methods adopted to perform the investigation is justified, arguing that the procedures most apt to be used here are the distance-based ones, but that results obtained by means of parsimony analysis deserve to be carefully taken into account as well. Thus section 6.1.3 is intended to present the first outfit of phylogenies reconstructions on the full dataset, which is commented on in detail. Furthermore, section 6.2 is focused on possible strategies to account for homoplasy (i.e. chance and borrowing): While an empirically-based selection of parameters is proposed in section 6.2.1, section 6.2.2 comprises suggestions regarding the way in which parameters might be weighted according to their genealogical relevance. Finally, section 6.3 covers some tentative analyses concerning the possibility of detecting and accounting for borrowing in phylogenetic trees, the reconstruction of ancestral states and the mapping of syntactic distances onto the diachronic and the diatopic dimensions of variation.

2. Language evolution and biological evolution

This chapter is dedicated to the comparison between biological evolution and language evolution. Before introducing the reasons why it is relevant to go through this issue, it is essential to point out that here the term "language evolution" refers to the diachronic development of languages, not to the evolution of the language faculty, which is a completely different issue: The importance of this distinction is addressed in section 2.2.1. The aim of the comparison is in highlighting some important similarities and differences between the two processes, in order to draw an analogical parallel between them. This means that the comparison is intended neither to force an extensive mapping between biological evolution and language evolution nor to outline a full innovative theory of language change. Rather, the idea is to show how some important features of the evolutionary process undergone by languages can be identified also in light of this type of parallel. As shown in the next chapter, these features are perfectly compatible with the most recent theoretical views on syntactic change developed within the generative framework, which is the theoretical perspective assumed in the present dissertation (see section 3.1). Moreover, the comparison between biological evolution and language evolution also represents a fundamental starting point for the specific experimentation that is presented in this work. For the most part, the quantitative procedures that are taken into account here to investigate the history of languages (see section 4.3) are originally designed to analyze the history of biological entities, thus they often assume specific evolutionary models, which apply to the biological domain but not necessarily to other fields of inquiry. As a consequence, in order to fully appreciate these methods of reconstruction, it is necessary to have a thorough knowledge of the main concepts and terms adopted in the current theory of biological evolution and to understand to what extent they can apply to the study of the diachronic development of languages.

The term "evolution" has been used with different meanings in different times and fields (Bowler 1975; 2003: 8) and, although they will become clearer in the course of the discussion, it is useful to anticipate some important distinctions in addition to the already mentioned ones. The Latin word *evolutio* just means "to unroll", i.e. to develop a structure from its compact form, and in biology the first mention of "evolution" referred exactly to the occurrence of this kind of process in embryos. From then on "evolution" has often been intended as a process of change from simple to complex, from lower to higher states of existence. This description does not apply to the process, to which Darwin[1] and contemporary evolutionary biologists refer to. According to the "modern synthesis" (see section 2.1.3), changes that affect organisms and species can determine some kind of improvement of the evolving units, but not necessarily an advance through higher levels of complexity. What are the relevant evolving units in biology, why and with respect to what they show a relative improvement in the course of time are questions that are taken into account in section 2.1 and, in part, in section 2.2.2. However notice that the term "evolution" has to be used with an even more neutral meaning in the case of language development: In current literature the term "linguistic evolution" does not entail any kind of improvement, but just a process of diachronic change (see section 2.2.1). This distinction reflects a fundamental difference between the biological evolutionary process (as intended in the Darwinian paradigm) and the linguistic one, as clarified in section 2.3.2.

---

[1] Actually, in order to avoid misunderstandings with the embryological theory, Charles Darwin referred to a process of "descent with modification" in Darwin (1859) and he used the term "evolution" just in the closing paragraph of the book. Later "evolution" became popular thanks to Herbert Spencer and other biologists' works.

The chapter is organized as follows. In section 2.1 the modern perspective on biological evolution is presented: The subject is introduced by an historical account of the development of the evolutionary theory starting from the pre-Darwinian explanations of species diversity, with the aim of highlighting the controversial development of the basic concepts that are now included in the so-called "modern synthesis". Section 2.2 covers an outline of some of the most relevant suggestions that have been brought forward so far in order to draw a parallel between biological and linguistic evolution: Also in this case the discussion is preceded by an historical overview of various applications of evolutionary ideas in linguistics. Section 2.3 illustrates the attempt to draw an analogical parallel between biological evolution and language evolution: The question one is faced with is asking whether the fundamental factors that drive biological evolution can also apply to cultural evolution in general and to linguistic evolution in particular. Notice that the issue regarding the differences between the quantitative "evolutionary" analysis of biological and linguistic data is not covered here, but throughout chapter 4, 5 and 6.

2.1. Biological evolution

Before Darwin (1859) the traditional and most popular belief in Western thought was that species are fixed and do not change into others over time. This view was part of the established Christian vision, which followed from the literal interpretation of the Book of Genesis: According to the book, in six days God created the whole world as we now it, including all living things, and from then on the product of His will remained unchanged. This non-scientific explanation of species diversity, which also relates to the origin of universe, Earth and human beings, can be referred to as "biblical creationism". The creationist world representation was influenced by the recovery of some concepts of classical Western philosophy and their integration within the Christian thought during the Middle Ages. The Christian myth of the existence of a fixed typology, imagined as a collection of eternal and transcendent patterns of all material things in the mind of God, is also a medieval reinterpretation of Plato's "theory of Forms", according to which on Hyperuranium there should be a set of archetypical ideas, whose corrupted shadows (i.e. projections) are the physical things of our world (Bowler 2003: 5). Moreover, if on the one hand Aristotle's contribution to zoological classification was fundamental for the development of modern taxonomy, on the other hand his "ladder of life", the hierarchical order of organisms respect to their increasing complexity and vitality, is also at the core of the Christian "great chain of being", which shows God at the top, followed by angels and then by material and corruptible creations, starting from humans.

The debate between creationists and evolutionists is still going on in our time (its role in contemporary American culture is discussed for instance by Witham 2005), reflecting the everlasting contrast between scientific and religious belief, and it originated a number of intermediate positions (a survey of the development of creationist theories can be found in Numbers 1987; Witham 2005). As pointed out by Bowler (2003: 1-3), the "Darwinian revolution", which actually began before Charles Darwin's birth, faced the same resistance that was encountered by the "Copernican revolution" in the sixteenth century: As the Church opposed the new Copernican cosmology, it did so with the idea that living creatures on Earth are not part of an unchangeable divine project, but are continuously subject to transformations, and also that human beings, considered Lords of creation in the biblical perspective, should be downsized to "simple" evolved apes.

It is not relevant here to probe into deeper detail the creationist-evolutionist debate: This premise is just intended to sketch the framework in which the first evolutionary theories were conceived and to underline their innovative role in contemporary scientific thought.

2.1.1. Towards Darwin's synthesis

Even if innovative in many respects, especially because it unified a number of previously unrelated ideas in a unique framework, Darwin's theory was not the first attempt to explain the origin and the diversity of species in evolutionary terms. Actually some of its core ideas go back to ancient Western philosophy (see among others Guthrie 1957). For instance, if Anaximander is recognised as the first thinker, who around the sixth century BC hypothesized that life originated into water and then moved to the land, almost two centuries later Empedocles believed that organisms were originally (but are not in the present) subject to a process of selection, which was responsible for the survival of the most adapted for life. A similar idea was presented also by Titus Lucretius Carus in the first century BC: Most importantly, the Roman philosopher developed the naturalistic-atomistic view of the world, according to which living things came from the progressive interaction between their parts, and supported the idea of inheritance of traits from parents to offspring.

It is only after the medieval age, which was dominated by the Christian view, that one can find more articulated hypotheses about the origin and the classification of organisms, even if they are still mostly proposed in philosophical rather than proper scientific terms. A complete and accurate presentation of them may be found in Bowler (2003: 27-134). Here the discussion is focused only on the most relevant views and ideas, according to which it is possible to highlight the essential historical pattern that lead to Darwin's original synthesis. One can start this survey considering both the discoveries made in early geology and palaeontology and the methodological improvements of biological classification.

The history of Earth, as described by the Holy Scripture and accepted as orthodoxy until the Enlightenment, is quite short. In the middle of 1600 James Ussher, archbishop of Armagh, coped with the problem studying the generations of the Hebrew patriarchs: According to his calculations, the world should have been created in 4004 BC. By the end of the eighteenth century geologists had shown that the earth was much older than that and palaeontologists had started providing evidence in favour of the existence of a vast period of time in which humans did not yet exist and our world was inhabited by unimaginable creatures. After the suppression of the medieval cosmological system and the diffusion of the Copernican representation, from the seventeenth century on Galileo and Descartes' mechanical philosophy gave new impulses to the naturalistic explanations regarding the origin and the history of Earth. However, the first cosmological and geological (*ante litteram*) theories were still conditioned by the need of agreeing with the biblical version. For instance, early theories had to account for the occurrence of the Great Flood. One can think about William Whiston's theory: Though he proposed a model that was coherent with the mechanicist and Newtonian framework, the English scholar argued that the deluge actually occurred and that it was caused by a comet that hit Earth and melted on its surface. Among the effects of the flood he included also the stratification of rocks. After the term "geology" had been popularized in Diderot's *Encyclopédie* in the mid-eighteenth century, geological studies became always more familiar and diffused. In any case the first rigorous analyses were presented only in the beginning of 1800, when scholars like William Smith, Alexander Broignart and Charles Cuvier introduced fossil-based stratigraphy. At that point it was clear for a large part of the scientific community that the biblical estimate of the age of Earth was unreliable and that empirically-based hypotheses needed to include much larger amounts of time, but the debate about the processes of change undergone by the planet was still open. Under the influence of the Great Flood myth, many geologists were convinced that changes were caused by catastrophic events, which caused rapid mutations of climate, environment and geography. Others did not believe in this model: In particular, Charles Lyell supported the alternative "uniformitarian perspective", arguing that changes have occurred gradually throughout the history of Earth until our times. Moreover, he also argued that the same slow processes are still in action. Lyell's uniformitarism had a great influence on Darwin's thought, who applied a similar model of gradual change to the evolution of species. Moreover, from then on the debate on the issue of gradualism has become increasingly heated, so that nowadays it still represents a central question not only in evolutionary biology, but also in current theories of language change (see in particular sections 2.1.4, 2.3.2 and 3.2.2).

The contribution of geological studies was fundamental for the rise of the Darwinian theory and the advances in taxonomy had an important role as well. Starting from the traditional "chain of being" paradigm, in the middle of the eighteenth century the Swiss naturalist Charles Bonnet developed a first tentative classification of material things (from ethereal matter to men), which was still dominated by the idea of a fixed hierarchical order of complexity. In accordance with this view, he also supported the "theory of germs", which asserted that all creatures are present on earth from the beginning in form of miniatures stored in females' ovaries: Bonnet explained replication arguing that every single living thing was designed by God at the beginning. A similar line of reasoning was for example Jean-Baptiste Robinet's: He also tried to join the chain of being representation with the germ theory, but he assumed a principle of continuity between creatures, which he considered all linked one to the other, even if we perceive them as distinct species because the intermediates may have not developed yet. As Lyell's uniformitarian principle in geology, also Robinet's continuity principle might be seen as a partial anticipation of Darwin's gradualism. However the fixed hierarchical classifications based on the chain of being appeared ever more inadequate as many new organisms, whose hierarchical position was obscure, were continuously discovered.

An alternative perspective was that proposed by Carolus Linnaeus (Karl von Linné) in the same decades: Even if he was still convinced that living things had to be organized according to a divine plan, which is rational and unchangeable, he noticed that a linear classification was not sufficient to describe it. He argued that organisms are bound by precise ecological relationships, which are responsible for fluctuations of the ecological system and are included within the divine plan of creation. From this premise the Swedish botanist provided a relationship-oriented taxonomy based on the resemblances between organisms: Starting from plants and extending the same method to animals, he grouped the most similar organisms into species, then species into genera, genera into orders and orders into classes. In Linnaeus's view, similarities between organisms are not signs of common ancestry (i.e. evolution), but rather reflexes of fixed relationships in the mind of God. Although presented in a non-scientific framework, Linnaeus' non-linear classification of organisms appeared to be enlightening: Later biologists inserted other taxonomic levels[2], but essentially Linnaeus' original representation is still in use. Moreover, the scholar introduced the binomial nomenclature of organisms, which are sill identified by the Latin name of the genus followed by that of the species they belong to. In the course of time Linnaeus came to consider the possibility of formation of new species, but he explained it resorting exclusively to hybridization processes, since he could not conceive spontaneous transformations inside the same species. Anyway, his conservative approach did not prevent him to reach extraordinary taxonomic conclusions: The conviction that classification must rely exclusively on the comparison of physical traits led him to find a close relationship between humans and apes, which he included together in the order *Antropomorpha* (primates). This result was revolutionary for that time and represented another little step towards the definition of a naturalistic explanation of species diversity.

Though founded on evident observations, the hypotheses of cognacy between humans and apes immediately found transversal resistance. For instance, in the second half of the

---

[2] The hierarchical classification currently used by biologists counts a succession of more inclusive levels (also called "taxonomic ranks"). For instance, the main ranks of zoological classification are: Species, genus, family, order, class, subphylum, phylum and kingdom.

eighteenth century the French naturalist George-Louis Leclerc Buffon, who was independently working on biological classification in a materialistic framework, refused this idea, because he believed in the mental superiority of human beings respect to the rest of the animal kingdom. In the same period James Burnett, Lord Mondboddo, a Scottish judge that was also interested in the study of the origin of languages, claimed that apes must represent a primitive form of human beings, but he suggested keeping primates distinct from other kinds of animals as well. Burnett's is also considered a precursor of other concepts that later would be included in Darwin's theory, above all the idea of natural selection, since he noticed that transformations occurred in organisms are closely related to the environment in which they live. Rough versions of the same principle had already been proposed in materialistic frameworks, for instance by Diderot in his encyclopaedia, by the French scholar Pierre Louis Maupertuis some decades before and by Buffon too: However Burnett's hypothesis was founded on more accurate observations and on wider naturalistic evidence.

As has been shown so far, evolutionary thought underwent a rapid and complex development in the decades that preceded Charles Darwin's birth. This was made possible thanks to the interaction between a series of advances in different but related fields of knowledge. Nonetheless it is hard to find more than isolated and incidental mention of evolutionary concepts before the first half of the nineteenth century.

The first widespread theory of evolution (or, more properly, of transformation) of species is Jean-Baptiste Lamarck's, which is known precisely as the "theory of transmutation of species". The French naturalist illustrated his model in the *Philosphie Zoologique* (Lamarck 1809): He hypothesized that each species has his own independent lineage and that lineages can neither branch nor become extinct. From this point of view, its rigid perspective on species typology was not so different from that of the early taxonomists. However, it is notable that Lamarck also thought that species undergo transformations in the course of time, due essentially to two factors. The first one is the "internal force" of organisms: Such transcendent substance was assumed to make organisms produce always slightly different offspring, so that after many generations the amount of small modifications can make species become significantly changed. Of course the nature of this "internal force" remains obscure to scientific inquiry. The second and fundamental factor that drives transmutation in Lamarck's view is the inheritance of acquired characters: This means that all characters, that an individual acquires in the course of its life due to diseases, accidents and, most importantly, the use or disuse of organs, are inherited by offspring. The famous example given by the naturalist regards the giraffe's neck, which in his opinion got longer one generation after the other because giraffes tended to stretch it to reach leaves higher up in trees. The idea of an " internally-driven improvement" of organisms was not new, since it had been already illustrated, among others, by Erasmus Darwin, Charles Darwin's grandfather, in the work entitled *Zoönomia* (1794-1796). Though notice that, unlike Lamarck, Erasmus Darwin was in favour of the idea of a common ancestry of at least all warm-blooded animals. The inheritance of acquired characters has come to be known as "Lamarckian inheritance" and, in the course of time, it was shown to be completely untenable (as illustrated for instance in Maynard Smith 1989: 8-12). Anyway, even if Lamarck's theory got a cold reception after its publication (see for instance in Ridley 1993: 9), it remained influential until the second half of the nineteenth century as an alternative to the Darwinian evolutionary theory.

2.1.2. Charles Darwin's synthesis

Charles Darwin formed his first ideas on evolutionary theory during a voyage in the Galapágos islands (1837-1838), in which he was working over his collection of birds. He soon understood that the finches he found in different islands did not belong to the same species, but rather that each island had its own species of finch. At that point the hypothesis that all finches derived from a common ancestor was straightforward for him. The model he had in mind was rather unconventional, because it was substantially different from, say, Lamarck's: A tree structure in which all lineages were related and represented by progressively divergent branches. Nevertheless, even if from then on he was convinced that unique species can split into different ones in the course of time, he still had to explain under which forces changes occur.

The enlightenment came in 1838 while he was reading Thomas Malthus' *An Essay on the Principle of Population* (1798-1826). He realized that, in the course of the "struggle for existence" that goes on everywhere, favourable variations in the form are more likely to be preserved than unfavourable ones, which conversely tend to be eliminated. The core of this process, which later took the name of "natural selection", is the interaction between inheritance of traits, spontaneous variations and adaptation to the environmental: If an organism is born in a new form that is better adapted for survival in a specific environment, it has a greater probability to leave offspring and in further generations the frequency of its form will increase in the population. As the variations progressively accumulate, in Darwin's opinion the result might be the gradual formation of a new species. Apart from other previous incidents, when the naturalist published his theory in *On the Origin of Species* (Darwin 1859), a similar idea had already been independently developed by Alfred Russel Wallace.

Many biologists reacted to Darwin's theory agreeing with the general model of evolution, i.e. with the assumption that species are genealogically related and undergo changes over time, but the greater part of the scientific community rejected natural selection. Some fundamental objections were (and are still) raised against this hypothesis. One of them concerned the twin assumptions that changes were supposed to occur gradually and by chance and that they are mainly conserved due to natural selection. In the critics' opinion, there are gaps in the evolutionary history of organisms that cannot be explained assuming a gradual model in which only natural selection acts. What about coadaptations and rudimentary stages in the evolution of organs? Common examples are respectively the giraffe's neck, whose lengthening would imply the occurrence of many apparently unrelated changes regarding the length of bones, muscles, veins and so on, and the development of wings, whose early stages did not necessarily represent an advantage for ancestral "proto-birds". As for the first case, i.e. complex adaptations, modern evolutionary biologists reply that one can assume that the evolution of many parts of an organ is under common genetic control. The second case can also be explained in the Darwinian paradigm, because on the one hand there are many ways in which also rudimentary development of characters can represent a real advantage (e.g. "proto-winged birds might have glided from cliff tops or between trees - as many animals, such as flying foxes, do now" Ridley 1993: 328), on the other hand earlier stages could be "preadaptations" for later stages, which means that an organ that was developed by chance for one use can come to serve a different function with little adjustments (e.g. the evolution of tetrapods from lobe-finned fishes described in Ridley 1993: 329): This "recycling" process is called exaptation after the definition of Gould and Vrba (1982). As pointed out is section 2.2.3, in linguistics the term "exaptation" is also used with a purely analogical value, among others, by Lass (1990; 1997).

Another objection concerned the model of heredity assumed by Darwin, as explained for instance in Ridley (1993: 32-36). The naturalist did not have a clear idea about it and chose to adopt the model of blending inheritance, according to which the traits of offspring are a blended mixture of the traits of parents. For instance, suppose that the character "long neck" represents an advantage for the organisms and that it arises by chance in an individual, how can it be transmitted to future generations and increase in frequency in the population? When the long-neck individual mates together with a normal short-neck one, their offspring will have an average length of the neck under the blending inheritance model: This entails that one generation after the other the length of the neck, though favourable, is expected to progressively decrease in the population instead of increasing. The lack of an adequate model of inheritance represented a big problem: The theory could not work without it. The solution came only in the beginning of the twentieth century, when Mendelian genetics was integrated with the evolutionary theory originating the so-called "modern synthesis" or neo-Darwinism.

2.1.3. The modern synthesis

The most important innovation of the Mendelian theory of genetics is that it allows the description of organisms in terms of discrete traits, whose combination determines the apparent continuous variation of morphological traits observed in populations by so-called biometricians. As shown in particular in section 4.2, which is dedicated to a critical overview of comparative methods adopted in linguistics and biology, the shift of focus from continuous to discrete characters represents a major achievement for theoretical and empirical research aimed at reconstructing genealogical relationships between *taxa*. The basic concepts of molecular and Mendelian genetics are illustrated further in this section. Before proceeding, it is relevant to point out that only in 1918 the biologist Ronald Aylmer Fisher showed how biometric features can be derived within the Mendelian framework, while "neo-Darwinism" was elaborated even later. The theoretical synthesis between the Darwinian idea of natural selection and the Mendelian theory of heredity is owed to the work independently done by Ronald Aylmer Fisher, John Burdon Sanderson Haldane and Sewall Wright in theoretical population genetics during the '20s: It was called "modern synthesis" after the publication of Julian Huxley's book entitled *Evolution: The Modern Synthesis* in 1942. Essentially, the first works "demonstrated that natural selection could work with the kinds of variation observable in natural populations and the laws of Mendelian genetics" (Ridley 1993: 16). From then on the interest in empirical genetic research progressively increased.

Notice that the concept of species underwent a substantial change around the '40s. Until then it was still used with a typological meaning, i.e. Linnaeus': A set of similar individuals that share some common features, which define the standard type. When geneticists came to analyze populations, i.e. groups of interbreeding organisms, it was clear that there were no "typical genotypes" (see later discussion) and that the most meaningful criterion to define the relevant units of classification was interbreeding, regardless of the morphological similarity of individuals: Thus "species" was also defined as the (maximum) set of interbreeding organisms. Currently one can refer to the typological concept of species using the term "phenetic species", because it refers to "phenotypic" properties of organisms, while the unit of classification that is relevant for geneticists can be indicated as "biological species" (for a survey of the different meaning of "species" see Ridley 1993: 383-445). Even if in the beginning the modern synthesis concerned especially population genetics, i.e. the study of genetic variation in groups of interbreeding organisms, it was soon applied to biological evolution of species. The idea was that if changes accumulate in geographically separated populations of initially similar individuals, the populations can diverge so much that they become different species.

Decisive insight in molecular genetics were provided since the middle half of the twentieth century, in particular thanks to James Watson and Francis Crick' contributions in the '50s. The current version follows. Many terms (e.g. acid, phosphate, protein) are just briefly mentioned, since their exact description is beyond the aims of the present work: A detailed explanation of them can be found in any introductory book on the subject (e.g. Weaver and Hedrick 1997).

The heredity material of multicellular organisms is contained in their cells[3]: During reproduction a new organism originates from a single cell, the zygote, which derives from the fusion of its parents' egg and sperm (for animals, or ovules and pollen for plants), the gametes[4]. The heredity material has the form of a molecule of DNA, which is the acronym of deoxyribonucleic acid. A molecule of DNA is a sequence of units, called nucleotides: Each nucleotide consists of a phosphate, a sugar group and a base. While phosphate and sugar are always the same, nucleotides differ with respect to the base: There are four types of nucleotides corresponding to four different bases, i.e. adenine (A), cytosine (C), guanine (G) and thymine (T). According to the chemical group, adenine and guanine are purines, while cytosine and thymine are pyramidines. The full DNA molecule is the combination of two complementary strands that are bound together to form a double helix: They are complementary because an A in one strand always pairs with a T on the other, while a C pairs with a G. For example, the complementary strand of AACTTGA would be TTGAACT. Moreover, the total length of a DNA molecule can be divided into regions: Some of them are spacer regions, others are called genes. Roughly, one can imagine that each gene encodes a specific protein. It is not relevant here to illustrate the form of proteins and the way in which they are exactly encoded. It suffices to say that proteins are at the base of the formation of organisms, in that each part of a body is built from a different kind of protein. Therefore genes are the fundamental units of information for designing organisms.

The DNA is also organized into a set of structures, called chromosomes, whose number varies across species. For instance, humans have 46 chromosomes organised in two sets of 23 distinct pairs: 22 of them correspond to non-sex chromosomes (autosomes), while one pair is constituted by sex chromosomes that are similar in females (XX) and rather different in males (XY). While zygotes, from which adult individuals develop, have these two parallel sets of chromosomes (they are diploid), gametes have only one (they are haploid): The consequence of reproduction is that the zygote has one set of chromosomes that derives from the female gamete and the other that comes from the male one. Thus, since the new organism that grows from the zygote owns two parallel sets of chromosomes, it also owns two parallel sets of corresponding genes, which are not necessarily identical. The genetic locus is the place in which a gene lies in a chromosome: As a consequence, an organism has two genes at each genetic locus and the description of the pair of genes for each locus is called "genotype". These two genes may be identical or slightly different: The variants of genes that are found in a specific locus are called alleles[5]. A population in which more alleles of a single gene exist is called "polymorphic" for that gene and its condition is called "polymorphism". If the two genes of a locus are identical the genotype is called homozygote, conversely one has an heterozygote genotype. For example, consider a specific genetic locus with just two alleles, say *G* and *g*. In this case, there are three possible genotypes: Two of them are homozygote, i.e. *GG* and *gg*, and one is heterozygote, *Gg*.

How do organisms with different genotypes appear? Which are the "phenotypes", i.e. the observable forms, that derive from the genotype? This point is fundamental, because it is at the core of the Mendelian theory of inheritance. Consider genotypes *GG*, *gg* and *Gg* and imagine that the alleles *G* and *g* encode a specific property of an organism, say, again, the

---

[3] More specifically, the heredity material is contained inside the nucleus of the cells if the organism is eukaryotic, i.e. it has complex cellular structure (e.g. higher animals), while it has no specific location in the cells of prokaryote organisms (e.g. bacteria).

[4] This description applies only to sexual reproduction, which has to be distinguished from asexual reproduction. In asexual reproduction females' gametes develop into female offspring without the contribution of males (see Ridley 1993: 269-270).

[5] The concept of allele in molecular biology recalls that of allophones and allomorphs in linguistics: As alleles are concrete variants of abstract genetic loci, in a similar way allophones and allomorphs are respectively context-dependent instantiations of abstract phonemes and morphemes.

length of the neck: *G* correspond to "long neck" and *g* to "short neck". It is expected that a *GG* individual has a long neck and that a *gg* one has a short neck, but what about an individual that has the heterozygote genotype *Gg*? There are two possibilities: Either this individual has an average phenotype, i.e. an average length of the neck, or it has a phenotype that resembles one of the homozygote phenotypes. If the case is the former, unlike the phenotype (and unlike the predictions of the blending inheritance theory) the genes *G* and *g* are not blended, but they remain unchanged and they will be transmitted without change to the following generation. If the case is the latter and, for instance, the phenotype of *Gg* is the same of *GG*, the allele *G* is called dominant and *g* recessive: In this way only the information carried by *G* surfaces, but that of *g* is not lost, it just remains silent waiting to be transmitted. In other words, there can be a one-to-one correspondence between genotypes and phenotypes or not, if one gene is dominant and the other recessive. However, the key concept is that, even if phenotypes can appear to be blended, genes remain unchanged (unless they are affected by spontaneous mutation, as explained later in this section).

The so-called "Mendelian ratios" describe the probabilities of occurrence of different genotypes in offspring. If parents have both a homozygote genotype, say *GG*, their offspring is 100% *GG* too, as shown in case 1 of table 1[6]. In case 2, where one parent has genotype *GG* and the other *Gg*, the offspring is expected to be 50% *GG* and 50% *Gg*. Case 3 illustrates what happens if both parents' genotypes are heterozygote: 25% of the offspring has *gg* and another 25% *GG*, while 50% of offspring is *Gg*.

Thus the Mendelian theory of inheritance can complete Darwin's evolutionary theory: Accidental variations of genes are preserved through generations, instead of being blended and reabsorbed in the course of time. On this premise, natural selection can work favouring the more fitting phenotypes[7]. For instance, referring back to the previous example, assume that neither *G* nor *g* is dominant, but that having a longer neck represents an advantage for individuals in a certain ennvironment. At the beginning the population is 100% *gg*, i.e. all individuals have a short neck. Then a variation of the genotype spontaneously arises, so that a *Gg* individual is born, which has a slightly longer neck. Mating together with a *gg* type, the probability that offspring will be *Gg* is 50%. Then, since the phenotype that derives from *Gg* is favoured on that of *gg*, *Gg* individuals will have more probabilities to survive and to produce their own offspring. If one projects this mechanism over generations, the process is likely to produce a 100% *GG* population in the course of time. The resulting population may have accumulated so many differences with respect to the original one that it may have become a different species.

A final remark before summing up the neo-Darwinian framework: Notice that phenotypes do not only derive from genotypes, but are also influenced by external factors, like accidents and environmental conditions, which are responsible for the acquisition of new characteristics, the same that were erroneously believed to be inherited by Lamarck.

2.1.4. Evolutionary factors and individual transmission

According to the "modern synthesis", there are four main factors that intervene in evolutionary processes, as summarized for instance by Cavalli-Sforza (2001: 74-81): Mutation, natural selection, migration and genetic drift. Inheritance instead is a basic premise

---

[6] Tables and sketches are collected in the appendix.

[7] It is relevant to point out that the term "fitness" has a specific meaning in evolutionary biology. It is intended as "the average number of offspring produced by individuals with a certain genotype, relative to the number produced by individuals with other genotypes. When genotypes differ in fitness because of their effects on survival, fitness can be measured as the ratio of a genotype's frequency among the adults divided by its frequency among individuals at birth" (Ridley 1993: 634).

of the model, in that it determines a specific kind of transmission of genetic information, and is taken into account at the end of this section.

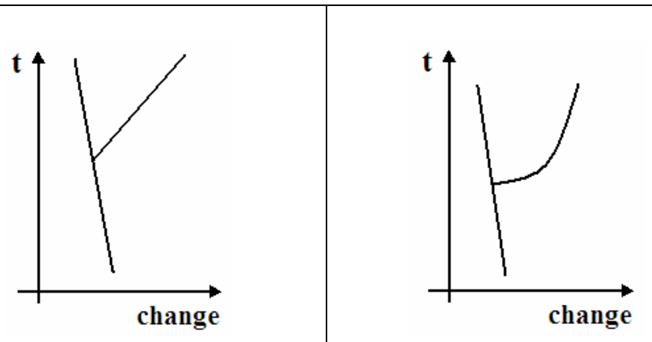
Mutations are accidental changes that affect the genetic material: They are relatively rare and mostly harmful, because organisms are complex machines in which slight variations can easily have lethal effects. Deleterious mutations are likely to be eliminated by natural selection and only few mutations are preserved: The favourable ones and the neutral ones, i.e. those which do not have negative effects on phenotype with respect to selection. Mutations arise by chance and cannot be driven by individuals, neither consciously nor unconsciously: In other words, there is nothing similar to Lamarck's "internal force". They can be caused either by external factors (e.g. radiations, viruses or mutagenic chemicals) or by internal processes (e.g. transpositions or errors that occur during meiosis and replication).

Most importantly, mutations are assumed to gradually accumulate until the resulting population becomes significantly different from the original one. Some models have been proposed against gradualism: Apart from theories that move from geological catastrophism, which was already contrasted by Lyell's uniformitarian principle in the first half of the eighteenth century (see section 2.1.1), one can remember the so-called "saltationism": According to this view, derived from the ideas proposed by the biologist Richard Goldschmidt in the '40s, the process of speciation (i.e. the derivation of new species from ancestral ones) is produced by rapid macromutations, rather than by progressive accumulation of little genotypic variations. This theory remains purely hypothetical for the moment. Another debate is taking place inside the gradualist perspective and regards the rate of accumulation. The fossil records show that on the one hand species show up rapidly, exist for a period and then become extinct, on the other hand there is often little evidence of transition between the ancestor species and their descendent ones. Is this entirely due to the incompleteness of fossil records? If the answer were "yes", one could assume an evolutionary model, in which the rate of change is constant both during and between speciation events. Eldredge and Gould (1972; 1988) call this view "phyletic gradualism" and they do not accept it. They propose the "punctuated equilibrium" model (Maynard Smith 1989: 282-284; Ridley 1993: 511-531), according to which mutations are assumed to gradually accumulate, but also to have much higher rates of occurrence at splitting events (i.e. speciation events) than during the periods, in which species persist: Thus they think that there are long periods of equilibrium punctuated by rapid episodes of variation. The difference between the two perspectives is clarified in figure 1: Representing the evolutionary process in a two-dimension graph with time on the vertical axis and change on the horizontal one, phyletic gradualism could be represented as in graph 1, while an extreme version of punctuated equilibrium would correspond to the graph 2 (Ridley 1993: 513): Indeed in the former graph the divergence between the two branches of the original lineage progressively occurs in a certain time *t*, while in the latter the same amount of divergence between the branches accumulates almost instantaneously. Notice that these models represent extreme possibilities of the same process and for the moment decisive evidence in favour of one or the other has not been provided[8]. As explained in section 2.3.3 and 3.1.5, the punctuated equilibrium hypothesis is taken into account also to describe the diachronic development of languages, but, at least in one case, with a substantially different meaning.

The idea of natural selection is essentially the same anticipated in section 2.1.2. Superficially one may assume that this process acts on phenotypes (but see the discussion regarding the "units of selection" in section 2.2.2): In a specific environment the individuals that have a more advantageous phenotype also have more chances to survive and to reproduce, i.e. they have an higher fitness (see note 7). Apart from features of the territory (e.g. climate), the

---

[8] A possible approach could be to admit that evolutionary processes can occur in different ways and that the punctuated model applies better to cases, in which an abrupt geographic split of the original population or rapid changes of the environmental conditions occur.



Graph 1: Phyletic gradualism.

Graph 2: Extreme punctuated equilibrium.

Figure 1: Different gradualist models of evolution.

environment includes also ecological relationships that link conspecifics and individuals that belong to different species: For instance, while the former may compete for the same resources, the latter may also compete, if they occupy the same ecological niche, or have a predator-pray relationship. Thus natural selection is the process that begins with adaptation to a specific environment: If there is no evidence for adaptation, one cannot assume that natural selection had a role in the evolution of a certain lineage. Moreover, similar phenotypic traits are likely to be selected by similar environments, thus they can provide misleading evidence if they are compared in reconstructing genealogical patterns: This point is taken into consideration in sections 4.2 and 4.3, within the discussion regarding the identification of suitable comparative characters in linguistics and population genetics.

Migration is also important in population genetics, since it determines the transfer of genetic material between different populations. That is why it is referred to also as “genetic flow”: When individuals of population *A* migrate in the area occupied by population *B*, they introduce their own alleles in that area. Of course, this process acts only if the populations can interbreed, i.e. they belong to the same biological species. While mutations increase the difference between separated groups and can cause significant genotypic divergence over generations, migrations have the contrary effect, because they make different populations converge. Thus mutation and migration are opposite factors in models aimed at studying the genetic similarity of populations as pertains to the geographic distance that separates them: The more two groups are distant, the more they are likely to diverge, while if they are closer, they can more easily be involved in migration processes and increase their genetic similarity. One of these models, conceived by Malècot (1959) and then developed by Kimura and Weiss (1964), is the “stepping-stone-model”: The name derives from the fact that human groups are represented as a series of steps that go across a pool. Cavalli-Sforza and Wang (1986) apply the same model to the linguistic analysis of some communities living in a chain of Micronesian islands: Their study is mentioned again in section 4.3.3, where some important quantitative perspectives aimed at studying the role of genealogical evolution and areal convergence in language development are presented, and also, incidentally, in section 5.1.1, where the quantitative encoding of the lexical data used by the scholars is considered.

Finally, the genetic drift[9] (see for instance Maynard Smith 1989: 24-27; Ridley 1996: 126-127; Cavalli-Sforza 2001: 74-78) is driven exclusively by chance, thus it does not favour

---

[9] Notice the term “drift” in population genetics has a different meaning respect to that used in linguistics by Sapir (1927), who distinguished between short-term and long-term drift: Both these processes are intended by the scholar as differentiating processes, i.e. processes of change that determine language divergence. Conversely, as explained in this section, the genetic drift has a

genotypes with respect to qualitative criteria like, for instance, natural selection. The drift can be explained as follows. Given parents' genotypes, by means of the Mendelian ratios one can exactly predict the probabilities of occurrence of offspring's genotypes. Nevertheless these probabilities are just theoretical, as in tossing a coin the probability of obtaining heads is 0.5. In ten tosses, one may observe five heads and five tails, but also seven heads and three tails or even ten successive heads. Only increasing the number of tosses the frequency of heads and tails approaches the theoretical prediction. In the same way, referring back to case 2 of table 1, it may happen that, out of six children born by parents that respectively have genotypes *GG* and *Gg*, three are *GG* and three are *Gg*, which would perfectly match the theoretical prediction of the Mendelian ratios. But, even if it is improbable, it is possible that, due to the effect of chance, all six children are *GG* and in this case the allele *g* would be lost. This is the chance effect of the genetic drift, which determines a non-predictable fluctuation of the genotype frequencies in the passage from one generation to the next. The genetic drift acts faster in smaller populations, because in these cases the genotype frequencies are more likely to deviate from the theoretical expectations, as the observed frequencies of heads and tails are more likely to be different from expectations with a small number of tosses. Therefore, if no alternative genotype for a specific locus is favoured by natural selection and no genetic flow from external populations occurs, due to the genetic drift one genotype will arrive to prevail on the others in the course of time. The final result of this process, if other factors do not intervene, is the convergence of the population to any genotype. Cavalli-Sforza (2001: 76-77) gives the following example. It has been observed that the frequency of blood group O is 100% in native Americans: This situation may be explained supposing that the first individuals that immigrated in the continent were few and that either they all had group O, or they also had frequencies of other groups, such as A or B, but these groups were eliminated by the effect of the genetic drift after some generations[10].

In higher plants and animals the inheritance of genetic material occurs from parents to offspring thanks to the mechanism presented in section 2.1.3. In epidemiological terms, this kind of transmission may be called "vertical": Thus a path of vertical transmission corresponds to a proper genealogical path. In nature there are also cases of horizontal transmission, in which an organism receives genetic material from another one even if there is not a parent-offspring relationship between them. This kind of transmission is known as "lateral gene transfer" and it has been studied only in the last decades (among others Syvanen 1985, 1994; Howe *et alii* 2001). It plays an interesting role in the evolution of unicellular organisms, but its relevance in the evolution of multicellular ones is debated and it may be considered rather marginal here. The terms "vertical transmission" and "horizontal transmission" may be also adopted to describe two important types of evolutionary paths of language evolution: Nevertheless using these terms with more than a simple descriptive value may give rise to conceptual problems (see the end of section 2.3.3). However, the same observation is valid for many other ideas that linguistics has borrowed from evolutionary biology: The following discussion is exactly aimed at showing the danger of strict parallels between biological evolution and language evolution.

---

homogenizing effect on a population. Here "drift" is always used with the meaning it has in population genetics.

[10] This is only a possible explanation. Another hypothesis is that the O group was selected because it provided better resistance to syphilis (Cavallis-Sforza 2001: 76-77).

2.2. Evolutionary ideas in historical linguistics

Many attempts have been made to apply evolutionary ideas to theories of language change so far. This section provides an overview of them with a twofold aim: It is intended on the one hand to anticipate some important points that are taken into account in section 2.3, on the other hand to show that too strict parallels between biological and linguistic evolution may pose some difficulties. It is important to recall the observation pointed out in the introduction of the chapter: This discussion concerns the evolutionary approaches adopted to analyze the diachronic development of languages, not the studies regarding the evolution of the language faculty: This distinction is clarified in section 2.2.1.

Croft (2000: 10-13) effectively points out that current evolutionary approaches to the study of language change may be divided into three categories: Literal, "generalizing" and analogical approaches. According to Croft's definition, the premise of literal views is that "language is a genetic capacity, and hence obeys certain principles of biology", thus "[a] literal approach to language diversity would amount to claiming that the differences among languages reflect genetic differences among their speakers", which is "patently false" (Croft 2000: 10)[11]. The last claim is indisputably supported by well-known empirical evidence and there can be no doubt that language change is culturally and not biologically driven (see section 2.3), therefore literal approaches (if there are any nowadays) do not require more attention here. The supporters of the "generalizing" approaches, instead, claim that there is a general theory of evolutionary processes, "which applies to the evolution of species and their traits in biology, to language change in linguistics, and to other domains as well" (Croft 2000: 11). Croft's (1996; 2000) theory is probably the most complete and extensive example of this category. Finally, analogical approaches are those for which "there are analogies between certain biological processes as described by evolutionary theory and certain processes of language change that call for description" (Croft 2000: 11). The following discussion starts from a brief outline of the role that evolutionary concepts have played in the history of comparative linguistics and then it focuses on some of the most recent and significant proposals, i.e. a "generalizing" view and some important analogical perspectives. Notice that the approach adopted in the present work may also be regarded as an analogical approach.

2.2.1. Historical overview

McMahon (1994: 314-340) provides an accurate outline of the application of evolutionary ideas in the history of linguistics. The scholar points out that a creationist view also existed for languages, as it did for species (McMahon 1994: 316): The idea that languages, like species in the biblical perspective, were created by God and are consequently fixed is suggested by the myth of the Tower of Babel described in the book of Genesis. Nonetheless, while the suppression of creationist positions represented a great challenge for evolutionary theories in biology, the idea that languages are subject to transformations was easily accepted. The main reason for this is that language change acts much faster than population diversification and speciation, so that sometimes its effects can be observed even by

---

[11] Croft misleadingly claims that the literal approach "is associated with Chomskian linguistics, because Chomsky argues for the biological basis of quite specific linguistic properties (e.g. certain syntactic structures and constraints)" (Croft 2000: 10). Actually there is nothing in the biolinguistic program (see section 3.1) that is even remotely intended to support a genetically-based interpretation of language change and development: The core idea of the Chomskian framework is rather that there is a set of biologically-determined grammatical features that are shared by all members of our species and that are consequently universal (i.e. the innate language faculty, or Universal Grammar, whose origin can then be studied from a biological-evolutionary perspective as that of any other organ), while language change has to be analyzed as a cultural and historical process.

uneducated speakers in the course of their lifetime, while the same evidence can be less effortlessly provided in support of biological evolution. From this point of view linguistics had an advantage over evolutionary biology in the beginning: A transformational approach was already implicitly assumed by Sir William Jones suggesting that Latin, Classical Greek, Sanskrit, Gothic, Celtic languages and, less safely in his opinion, Old Persian could have the same origin. This proposal dates back to 1786 and only after two decades Lamarck published his "theory of transmutation of species".

Thus the idea of language change was common since the origin of comparative linguistics and one could claim that it has always represented its core premise. Nevertheless the process of change described by nineteenth-century linguists was substantially different from the Darwinian one and more similar to the "evolutionary process" hypothesized by early embryologists and naturalists. In fact it was usually assumed that languages are subject either to advances or to decay, depending on each scholar's perspective, i.e. that the evolution of languages always entails some kind of improvement or decline: In particular, the approaches supporting the idea that languages undergo advancement in the course of time can be said to be in line with the "Growth Principle" using Harris and Campbell's (1995: 17) terminology. For instance, in the beginning of the nineteenth century Friedrich Schlegel adopted the typological division between affix and flexive languages, under the conviction that flexive languages are above all perfect: Since Sanskrit shows the most regular and rich inflectional system, he claimed that all Indo-European languages derived from it, i.e. are degraded forms of that perfect mother language. This line of reasoning is in accordance with the Romantic tendencies of the time in which Friedrich Schlegel was writing: On the one hand the interest in Oriental cultures was spreading into European thought, as the myth of the intellectual superiority of those cultures, on the other hand it was often assumed that the level of intelligence and civilization of a nation is reflected also by the language it adopts[12]. After Friedrich Schlegel, the advances of pre-scientific linguistic studies allowed to overcame the idea that Sanskrit was the mother of all Indo-European languages, assuming the existence of the unattested Proto-Indo-European, but the prejudice that some languages are superior to others was not abandoned. The morphological typology initiated by Friedrich Schlegel was refined by his brother August Wilhelm Schlegel, by Wilhelm von Humboldt and by August Friedrich Pott, and with slight variations this kind of classification (which distinguishes between isolating, agglutinative, flexive, incorporating and/or polysynthetic languages) is still in use. But with a different premise. While now the morphological typological classification has more a general descriptive value (no language is completely isolating, flexive and so on, which actually was already clear in Humboldt's mind), at the time it was conceived it was believed to correspond to higher or lower levels of perfection. But how was progress toward (or decay from) perfection measured? "[T]he usual nineteenth-century answer was that greater complexity, specifically in the morphology, signalled a more advanced and highly valued language" (McMahon 1994: 320). This approach closely resembles the "chain of life" paradigm adopted by early taxonomists.

The same idea of "evolution" was still rooted in August Schleicher's thought, even if the scholar is often mentioned as the first linguist most directly influenced by Darwin's theory. He also adopted the idea that languages are subject to change and he widely used biological terms in his works, so that he explicitly equated "language families with genera, languages with species, dialects with races, and idiolects with individual organisms" (McMahon 1994:

---

[12] Friedrich Schlegel concluded that German is the less corrupted derivation from Sanskrit, in that it preserves a rich inflectional system, therefore the German people should be considered superior from the intellectual point of view. In the same years (1807-1808) Fichte's *Addresses to the German nation*, in which the philosopher argues in favour of the cultural and spiritual superiority of Germans, was published.

319), but his first classification attempts were neither formalized as genealogical trees, they were rather taxonomic tables of Linnaeus's kind. However, the graph presented in Schleicher (1853) is conventionally assumed to be the first genealogical tree of Indo-European languages and Schleicher is rememberd as the initiator of the so-called *Stammbaumtheorie*. This representation of language diversity, which exclusively showed the development of languages over time, was soon criticized by Johannes Schmidt (Schmidt 1872), who proposed a model of language change focused on space variation, the *Wellentheorie* (see among others Fanciullo 2007: 144-148). The discussion regarding the opposition between these theories and their possible integration is fundamental in the present work and it is presented in section 2.3.3. For the moment, simply notice that the main objection that was rightly raised against the *Stammbaumtheorie* was that it cannot be considered an exhaustive representation of language change, since it accounts only for purely genealogical evolution, but it cannot represent contact-induced transformations, i.e. cases of areal convergence of genealogically separated lineages. As anticipated, Schleicher's model of evolution entails a direction from primitive to higher forms of languages as well. In particular, the scholar assumed that languages are subject to a three-stage history: Origin, progression (i.e. pre-Darwinian "evolution") from simple to complex and decay. Again, complexity corresponds to advance, and modern languages are supposed to be corrupted versions of the ancient ones.

In typological terms, according to Schleicher's hierarchy synthetic languages (roughly, those in which words tend to be constituted by more morphemes) represent an improvement respect to analytic languages (in which the tendency is that each single morpheme corresponds to a word). This view encountered the opposition of the Neogrammarians: In particular, Osthoff and Brugmann (1878) were convince that "the same types of language change apply to all phases of linguistic history" (Harris and Campbell 1995: 18) and laid the foundations of the current principle of "uniformitarism". Before introducing it, it is important to recall that the schleicherian view still persisted for a long time, so that after some decades Otto Jespersen proposed exactly the opposite idea. He hypothesized that the progress of languages is directed from complex to simple, i.e. from synthetic to analytic forms, which are more efficient: "[T]hat language ranks highest which goes furthest in the art of accomplishing much with little means, or, in other words, which is able to express the greatest amount of meaning with the simplest mechanism" (Jespersen 1922: 324; quoted in McMahon 1994: 323). Of course many exceptions to Schleicher and Jespersen's progressive scales of evolution are known.

It is not necessary to delve deeper into this question at this time, but two important points have to be to underlined: First, that all the approaches mentioned so far can only superficially be associated with the Darwinian theory, because they all make use of the pre-Darwinian idea of evolution; Second, that, agreeing with the Neogrammarian hypothesis, nowadays historical linguists do not interpret language development as processes entailing advancement or decay, but just "transformation". In other words, "[a]lmost all linguists take it for granted that, in some important sense, all languages are equal", i.e. that "there is no such thing as a 'primitive' language – the language of a nomadic tribe of hunter-gatherers is no less complex than the language spoken in an industrialized society" (Christiansen and Kirby: 2003: 4). This basic principle is called "uniformitarianism" (which does not have to be confounded with Lyell's uniformitarian view on geological history, see section 2.1.1).

At this point it is necessary to open a parenthesis. There is another issue implicated by the "evolutionary" approaches considered so far: The question regarding the origin of the human linguistic faculty. Even if this problem has nothing to do with the investigation covered in this dissertation, and, as pointed out since the beginning of the chapter, it rather has to be clearly kept separated from it, its historical development is relevant in order to complete the overview presented in this section. As the inquiry regarding the explanation of species diversity, the investigation of the origin of language began in ancient times. For instance, Swadesh (1971: 2) recalls that the idea that originally languages must have been "imitative" with respect to the objects of the world was already discussed in Plato's *Cratylus*. During the Enlightenment the German philosopher Johan Gottfried Herder made a similar proposal, claiming that humans developed language when they were "mentally mature enough" to do it and the originally imitative sounds they used became unrecognizable (i.e. arbitrary in later Saussurian terms) in the course of time. In some sense, the whole already mentioned debate from Schlegel to Jespersen followed the same line of reasoning, because all these eighteenth century scholars started from the implicit premise that there must have been a "primitive" stage in the development of languages, but notably they tried to project their hypothesized hierarchical processes of change onto an historical framework. This improper overlap of historical and "biological" inquiry caused many misunderstandings and lead to the proliferation of non-scientific and "creative" theories regarding the origin of language/s.

The debate, which was also stimulated by Darwin's theory of evolution, came to a temporary stop when in 1866 the Linguistic Society of Paris decided to ban from its session papers any work on the question. The investigation of the origin of the language faculty was revived only in the twentieth century by scholars like Hockett (see for instance Hockett 1960), who studied the issue taking into consideration animal forms of communication, and Swadesh (1971). In particular, the conference on "Origins and Evolution of Language and Speech", sponsored by the New York Academy of Science in 1975, contributed to the scientific rehabilitation of the theme. However, Christiansen and Kirby claim that "[t]he landmark paper 'Natural language and Natural Selection' published in 1990 by Steven Pinker and Paul Bloom in the respected journal *Behavioural and Brain Science*, is considered by many the catalyst that brought about the resurgence of interest in the evolution of language (faculty, editor's note)" (Christiansen and Kirby: 2003: 3). In the last decade the study of the origin has become increasingly popular with the awareness that the problem needs to be faced with regard to the contributions that come from disparate scientific fields, like biology, neuroscience, anthropology and so on. Some of the most recent works that cover the problem in such multi-field framework are, for instance, Kirby (1998), William and Bickerton (2000), the already mentioned Christiansen and Kirby (2003) and Piatelli-Palmarini and Uriagereka (2004). The issue has been addressed also within Chomskian biolinguistic program, see for instance Hauser *et alii* (2002), Fitch *et alii* (2005), and Chomsky (2008).

At this point the need to distinguish the linguistic perspectives that make use of evolutionary ideas to model historical language development from those regarding the evolution of language faculty should be clear: The two issues entail the study of different processes, which involve different periods of time (centuries the former and probably geological ages the latter), different levels of analysis (cultural and natural) and different fields of study[13].

Coming back to the history of the applications of evolutionary ideas in historical linguistics, McMahon (1994: 334-341) observes that the first significant attempts of parallelism in modern theories go back to the '50s and the '60s. Some biological metaphors begun appearing in creole studies also in the '70s, but, according to Briscoe (2002: 2), only from the '80s interest in the study of languages as evolving systems increased, in his instance taking advantage of the progresses both of the "modern synthesis" and of the quantitative algorithmic studies of dynamical systems (see for instance Lindblom 1986; Hurford 1987; Keller 1994). However, the most important insights have been provided in the last decade, "perhaps partly because work has only recently begun to address questions seen as central to (generative) linguistic theory" (Briscoe 2002: 2). In the next sections some of these recent

---

[13] Notice that the same question is faced by Longobardi (2003a: 2-4), when he outlines the extension of Chomsky's (1964) original levels of adequacy adding the level of actual historical adequacy and that of evolutionary adequacy. See section 3.1.1 for a detailed account on the issue.

proposals are reviewed, starting from Croft's "generalizing" approach and then illustrating some of the most interesting analogical views.

2.2.2. A "generalizing" approach

Croft's (1996; 2000) "Theory of Utterance Selection" probably represents the most broad-based "generalizing" approach aimed at explaining language change. The scholar's starting point is Hull's (1988) "generalized theory of selection", which has to do with a fundamental question of evolutionary biology: Which are the "units of selection"? The issue may be explained as follows. Life is organized in different levels, which are bound together, and each level includes different units: For instance, one may study the processes that life undergoes referring either to the genetic level of organization, in which case the relevant units are genes, or to the organismic level, whose units are organisms, or, climbing up the hierarchy of organization, to the level of species. Thus, in evolutionary biology the question is "whether natural selection really acts to produce adaptations to benefit one level, with benefits at other levels being incidental consequences, or whether it acts to benefit all levels" (Ridley 1993: 304).

Maynard Smith (1987) observes that the only units that can benefit from adaptation, and therefore may be direct units of selection, are those that show inheritability. Which are those units? A first answer could be "organisms", while groups, e.g. species, should be discarded, because for groups the inheritability from generation to generation is low. One should also notice that, in order to claim that natural selection increases the frequency of a unit in the course of time, not only that unit has to show inheritability, but it also has to be permanent enough, so that its frequency can be effectively altered by selection and be significantly measured. Williams (1966) and Dawkins (1976) argue that in this sense the relevant units are genes, not organisms, which are not "permanent": While each organism is unique and exists for a finite and limited period of time, a gene perfectly replicates itself and each replication is likely to be an identical copy of the same unit. Therefore, since genes show almost perfect inheritability and in principle may persist indefinitely, Dawkins (1976) claims that genes, that he call "replicators", are properly favoured or disfavoured by natural selection, while their phenotypic projections, i.e. organisms, are "vehicles", whose adaptation to the environment determines either their survival or their disappearance. An extensive account of the question of the unit of selection and of the gene-centred view of evolution may be found among others in Ridley (1993: 303-322). What is relevant here is that on this basis Hull (1988) develops his generalized theory of selection, in which the following components play a role:

1- "'Replicator – an entity that passes its structure largely intact in successive replications'
2- 'Interactor – an entity that interacts as a cohesive whole with its environment in such a way that this interaction causes replication to be differential' (Hull's interactor corresponds to Dawkin's "vehicle", editor's note)
3- 'Selection – a process in which the differential extinction and proliferation of interactors causes the differential perpetuation of the relevant replicators'
4- Lineage – an entity that persists indefinitely through time either in the same or an altered state as a result of replication'" (Hull 1988: 408-409; quoted in Croft 2000: 22).

Thus, according to this model, in biological evolution paradigm replicators are assumed to be genes, while paradigm interactors would be organisms: The former are assumed transmit their forms to descendents with no modifications, thus creating largely inalterable lineages; The latter carry instantiations of replicators and, if in the course of their interaction with the environment they die, together with them the specific instantiations that they carry disappear and the overall frequency of the relevant replicators in the population is modified.. Indeed "altered replication" and "differential replication", which are respectively mentioned in the definition of "lineage" and "interactor", are defined as follows: "[A]n altered replication produces variants of a structure" and a differential replication is "the shift in the frequencies of a variants of a structure" (Croft 2000: 3). In other words, in biological evolution the former process would correspond to a mutation, that creates an altered allele, and the latter to the propagation of an altered allele in a population. If the replication does not cause the occurrence of changes, it is called "normal replication". Moreover, another distinction is highlighted in the generalized theory of selection: While replications are processes that entail the creation of new entities, there is a process of change that acts on a single object, that changes over time: This process is called "inherent change". An example of inherent change may be the development of an organism in the course of its lifetime. As anticipated, Hull (1988) claims that this model of evolution is general, i.e. it can be applied in any case in which one needs to describe the evolution of a system. In particular, the scholar tries to apply the generalized theory of selection to conceptual change in the history of science (see for instance the summary provided by Croft 2000: 20-30).

Croft (2000) frames this model within a "usage-based theory of language". The scholar defines an utterance as "a particular, actual occurrence of the product of human behaviour in communicative interaction (i.e. a string of sound), as it is pronounced, grammatically structured, and semantically and pragmatically interpreted in its context", thus as "a spatiotemporally bounded individual" (Croft 2000: 26). In his opinion, only utterances can be considered "basic tokens", i.e. specific and concrete entities, whose historical development can be traced. In this sense, a language, say English, would not be a token[14], because it is an abstract entity removed from a specific historical context. According to Croft, a language would rather be "the population of utterances in a speech community" and a grammar is intended as "the cognitive structure in a speaker's mind that contains her knowledge about her language, and is the structure that is used in producing and comprehending utterances" (Croft 2000: 26). Moreover, the particular "linguistic structure" embodied in an utterance (e.g. a phoneme, a morpheme or a "syntactic construction") is called "lingueme" by Croft[15].

On the basis of these definitions, Croft assigns to each component of Hull's model a linguistic counterpart: The overview of the essential concepts is presented in table 2. According to the Theory of Utterance Selection, while a gene is the paradigm instantiation of the replicator in biological evolution, the same role is played by a lingueme in language evolution. Thus linguemes are assumed to be structured into utterances as genes are organized in the DNA. Normal replication would then correspond to regular utterance production and altered replication to innovative utterance production, which are respectively paralleled to normal reproduction and mutation of the biological model. Croft assigns the role of interactor to the speaker, who notably includes the grammar. Moreover, in the scholar's view the selection process is driven by convention, i.e. the set of the arbitrary (non-functionally and non-biologically determined) grammatical rules shared by a speech community, in any social-communicative context. Finally, for Croft differential replication corresponds to propagation of linguistic change: In his opinion, while the mechanisms that determine the innovation (the altered replication) is functionally based, i.e. it "involves the form-function mapping", the propagation "is a selection mechanism, in the evolutionary sense, and it is social" (Croft 2000: 8).

---

[14] This point of Croft's model becomes less clear when the scholar says that "[f]rom these basic tokens (i.e. utterances, editor's note), we can describe more complex tokens, such as a specific language or a speech community" (Croft 2000: 2).

[15] This term is intended to parallel Dawkins' "meme", which is intended as "a unit of cultural inheritance" (Croft 2000: 239).

The Theory of Utterance selection does not need to be covered in greater detail here. What is important to highlight is that Croft (2000) makes use of at least two useful and important distinctions. The first one is that between inherent change and replication, the second one is that between altered and differential replication. In particular, notice that in the scholar's view only replications are relevant in language evolution, because what actually "evolve" for him are utterances: If languages were assumed to develop as "organisms", as in some nineteenth century views, their diachronic transformation should be more properly called "inherent change" instead, because only this characterization of the process could account for the continuum of states observed in the genealogical development of specific languages (see section 2.3.1). Thus Croft's model is perfectly coherent in this sense. As for the second distinction, i.e. that between altered and differential replication, Croft again provides a useful observation saying that the former is functionally (but maybe one could more generally say "internally") driven and the latter depends on social selection. The fact that the propagation of an innovation has to be studied from the sociolinguistic point of view is undoubted, as the importance of studying the innovation as an internal linguistic process. Finally, one may also agree with Croft's idea that altered and differential replications are respectively synchronic and diachronic phenomena, because the former occurs "in a speaker action in a given point in time" and the latter "over a very long period of time, even centuries" (Croft 2000: 5). However, notice that the need to distinguish between synchronic innovation and diachronic diffusion is well-known in sociolinguistic studies at least since Labov (1966) and Weinreich *et alii* (1968).

Croft's view is debatable in other respects. His usage-based theory of language, which assigns a central role to utterances, is completely incompatible with the generative framework adopted here and one could object that it leads to bizarre conclusions. The following discussion is focused only on two fundamental points, since the exhaustive critic review of Croft's theory is beyond the aims of the present work. A first observation could regard the fact that, according to Croft's line of reasoning, since linguemes should be replicators and speakers, together with their grammar, interactors, utterances result to be associated with genetic material and grammar with phenotypic appearance. This parallelism appears to be misleading from the point of view of the perspective adopted here: It does not take into account the fact that there is also a causal relation between the genotype and the phenotype in biology, according to which the phenotype is (partially) determined by the genotype, while the common belief in generative linguistics is that grammar generates utterances (or better, sentences), not the other way around. In other words, looking at the instantiation of the evolutionary process in biology, one should expect not only that the replicator replicates itself and the interactor interacts with the environment, but also that the replicator projects itself onto the interactor: If one assumed that the interactor is the grammar and the replicators are the linguemes, the grammar would be a result of a projection of the linguistic usage. Croft is aware of this objection and claims that "the fact that the genotype is expressed in the phenotype, but a grammar generates an utterance, has no bearing on the mechanism or processes involved in replication, interaction and evolution" because "the generalized theory of selection does not apply only to the levels of the gene and the organism in biology" and "[i]t is independent of the levels of organization of biological entities" (Croft 2000: 40). In other words, the replicator and the interactor are assumed to be very specialized roles, which abstract away from the concrete relationship that exists between their instantiations. But then to what extent could one benefit of a model of language evolution, in which the application of the categories of replicator and interactor is so abstract?

Another possible objection to the Theory of Utterance selection regards the exact generalization of the process of selection, which is claimed to be the same process instantiated in different ways in biological evolution and language evolution. Actually natural selection and social selection are substantially different processes (see section 2.3.1): Putting it in

Croft's terms, while the former is driven by the environment and the interactors have a passive role, the latter is actively (and often consciously) determined by the interactors. Again, convention (which notably the scholar considers to play a more decisive role in language change than functional and formal features of language) as a mechanism of selection may probably be claimed to be too abstract to account for the actual social process of propagation, unless one considers "convention" as a transcendent internal force of the speech community. Thus one might wonder whether Croft's need to insert an extensive mapping between biological evolution and language evolution leads him to neglect some fundamental aspects of language change, in favour of debatable conceptual generalizations.

2.2.3. Analogical approaches

Analogical approaches are those, in which linguists make subsidiary use of biological terms and concepts to describe or to shed light on specific facts regarding the process of language change. As illustrated in particular in section 2.2.1, the use of biological metaphors in linguistics has led to controversial conclusions in the last two centuries, because this practice has often entailed an "overenthusiastic appropriation (of biological concepts, editor's note) with insufficient sense of subtlety or precise applicability of the originals" (Lass 1990: 79; quoted in McMahon 1994: 314). Nonetheless, many scholars have adopted cautious analogical approaches in recent times, being aware that such attempts require a careful reflection on the consequences of their comparison: Due to this prudence, in some cases the analogies do not go much beyond the terminological borrowing.

Croft (2000: 11) asserts that some recent perspectives in creole studies may be considered analogical. In particular, one may think about Whinnom (1971), who attempts to make use of the biological concept of hybridization to explain some dynamics of language contact, and Mufwene (2000), who declares that there are some informative similarities between the concepts of language and species. Actually, the scholar's perspective has become more "generalizing" in later works (see for instance Mufwene 2002; 2008), whose review is not covered here. A recent analogical proposal made in historical linguistics is for instance Lass' (1990; 1997) adoption of the biological concept of exaptation (see section 2.1.2) to refer to a diachronic process of reanalysis alternative to grammaticalization (see section 3.2.1). However, the scholar points out that "while claiming that the notion of exaptation seems useful in establishing a name and a descriptive framework for a class of historical events", he remains "fully aware (even insistent) that languages are not biological systems in any deep sense" (Lass 1990: 96; quoted in Croft 2000: 11).

One may state that another group of analogical approaches includes all evolutionary-oriented works that "utilize a methodology of computational implementation and simulation of (more or less explicit) formal models" (Briscoe 2002: 2). A collection of representative papers based on this kind of perspective is exactly that introduced by Briscoe (2002). Among them, the most interesting from the point of view of this dissertation are those dedicated to the simulation of the evolution of parametric systems, i.e. Niyogi (2002) and Turkel (2002). These models and in particular the development of Niyogi's framework are presented in the next chapter (section 3.2.4), since they require a full understanding of the Principles and Parameter Theory (section 3.1.3). For the moment, simply notice that Niyogi's (2002) starting point is Cavalli-Sforza and Feldman's (1981) model of cultural change, which is taken into account in section 2.3.1.

Analogical approaches are also more or less implicitly adopted in quantitative studies of language evolution, as the one presented in this dissertation, which may be referred to as "linguistic studies in quantitative phylogenetics". Recalling the definition provided in the introduction (chapter 1) and anticipating the detailed one illustrated in section 4.3, the term "quantitative phylogenetics" refers to the research field, whose aim is of automatically inferring genealogical (or, more generally, historical) relations between sets of *taxa* by means of algorithmic procedures applied to comparative datasets. The adjective "quantitative" refers to a basic property of the techniques used in this field, i.e. the fact that the reconstruction process is (almost) completely independent of the nature of the analyzed data and relies only (or mostly) on the numerical aspects of the comparison. Thus the same procedures can be (cautiously) exported through disparate disciplines, even if the basic concepts of the current quantitative phylogenetic paradigm were originally conceived in molecular biology in the '50s (see section 4.3.1). The implicit analogy adopted by the quantitative studies in linguistics regards the comparative data, which actually varies a lot depending on each scholar's view: As illustrated in section 5.1, many experiments are exclusively conducted on lexical and morpho-phonological datasets, while only a minority of them are focused on syntactic comparison. However, all of these works assume that the quantitative analysis of linguistic characters can be equated to that of genetic characters: In other words, linguistic traits, for example lexical encodings of meanings or syntactic features, are considered comparatively equivalent to genetic traits, i.e. genes, for the aims of the quantitative reconstruction. As already pointed out, this equivalence holds only if the evolutionary processes undergone by biological and linguistic characters are similar enough respect to some relevant aspects of the models that are implemented in the quantitative reconstruction procedures. This specific question is addressed later on the basis of the parallel provided in the next part of this chapter.

2.3. Analogies and differences

As claimed in the beginning of this chapter, providing a detailed evolutionary model of language evolution is beyond the aims of the present work. In this section an analogical approach is adopted, i.e. it is assumed while some ideas developed within the biological theory of evolution can be usefully adopted to highlight some important features of language change, diversely, the differences between biological and language evolution prevent drawing a strict parallel between them. The intent is to achieve these goals starting from the presentation of a quantitative model of cultural evolution. Section 2.3.1 illustrates in which sense language evolution has to be considered a particular instantiation of cultural evolution. In the course of the discussion presented in section 2.3.2 some points of convergence between the dynamics of biological evolution, cultural evolution and language evolution are underlined, as the essential features that differentiate them. Finally, section 2.3.3 covers a fundamental issue regarding language evolution: The distinction between pure genealogical evolution and areal convergence.

2.3.1. Cultural evolution

Before outlining the comparison between the factors that intervene in biological evolution and those that are relevant in language evolution, it is useful to look at the "cultural" aspect of language evolution and to illustrate how the evolution of "cultural" traits can be modelled. What is meant as "culture"? Abstracting away from detailed sociological and anthropological descriptions, one may assume the broadest possible definition of "culture", meaning "what is learned from others, in opposition to what is learned by oneself, in isolation" (Cavalli-Sforza 2001: 249; editor's translation). Consequently, a cultural system may be viewed as the set of cultural traits that are typically ascribed to a human group[16]: Cultural traits concern, for

---

[16] However notice that "culture is important for animals too. The parents' teaching, above all that provided by the mother for mammals and birds, is essential for most species" (Cavalli-Sforza 2001: 250; editor's translation).

instance, religious or political belief, family structure, diet, moral attitude and so on, i.e. they are all those features that affect individuals' habits and social interactions.

To understand which aspects of language evolution have to be regarded as cultural, one may start asking the following question: How is a specific language learned? According to the perspective developed since the '50s within generative grammar, there is a specific innate component of the human mind, which may be generally referred to as "language faculty", "that yields a particular language through the interaction with presented experience, a device that converts experience into a system of knowledge attained: knowledge of one or another language" (Chomsky 1986a: 3). The generative approach to the problem of language acquisition is broadly illustrated in section 3.1, for the moment just notice that the acquisition of a specific language is assumed on the one hand to be biologically determined, i.e. driven by a specific innate component of the human mind, on the other hand to require the child to be exposed to a specific linguistic environment. This approach is in accordance with the well-known empirical observation that a child does not natively acquire the language of its parents, but the language that is presented to its experience in the first years of life (in particular, during the so-called "critical period for acquisition", see section 3.1), which do not necessarily coincide: For instance, a child that is born of Japanese-speaker parents, is separated from them and is brought up in Italy, will natively acquire Italian, not Japanese. Thus, roughly speaking, while language faculty is genetically inherited, knowledge of a specific language is culturally transmitted, because a specific language is fundamentally learned by others.

It is possible to observe that, as (in general) biological and linguistic features, also cultural traits change over time. The first models aimed at shedding light on the dynamics of cultural change in an evolutionary perspective appeared in the '80s. Among them, probably the most famous and extensive is Cavalli-Sforza and Feldman's (1981). The scholars design a quantitative analysis of cultural change taking advantage of the main concepts developed in population genetics and with particular reference to the mechanisms of individual transmission of cultural traits. The premise of the model is that a cultural trait (or cultural unit) can be considered comparable to a genetic trait (a gene). According to this view, what do biological and cultural evolution have in common? Do the evolutionary factors identified in the theoretical framework of the modern synthesis apply in the same way to cultural evolution?

Mutation is the first evolutionary factor that has to be taken into account. The fact that, as genetic mutations, cultural innovations have a gradual diversification effect on cultures seems to be indisputable. However, while genetic mutations occur by chance due to chemical changes or errors in the copy of the original genetic material and more frequently they are damaging for organisms, Cavalli-Sforza and Feldman affirm that, "in the cultural process, the change is not necessarily a copying error, but can often be directed innovation, that is, innovation with a purpose, and might therefore appear to be non-random" (Cavalli-Sforza and Feldman 1981: 66). In other words, in principle any individual can consciously evaluate the possibility of adopting cultural change (e.g. her religious belief) with respect to personal preferences, values, advantages and so on. Thus the main difference between genetic mutation and cultural change seems to regard the fact that while the former is random, the latter is mostly directed and intentional.

This does not mean that cultural changes are always advantageous for individuals or groups. As natural selection influences the frequency of alleles in a population and leads to the disappearance of those, that are originated from harmful mutations, in the same way it can determine the deletion of cultural traits that are injurious to the existence of an individual or a group (e.g. smoking). However, any innovative cultural trait has to be accepted before undergoing natural selection: In particular, it has to be approved by single individuals in order to increase its frequency in the social group. Thus one may think that there are two levels of selection for cultural changes: The first one is social and the second one is natural.

Also migration plays an important role in cultural evolution: The passage of knowledge, techniques and habits between different human groups is widespread. Notice that, while in biology migration has effect only if the contact occurs between interbreeding populations, i.e. populations that belong to the same species, in principle there are no barriers against the flow of cultural units, at most there is (often temporary) resistance. Moreover, Cavalli-Sforza and Feldman point out that, unlike in the biological evolutionary model, migration can take two different forms: There can be migration of people, which carry their own cultural traits, and migration of ideas on their own. These migrations are respectively referred to as "demic" and "cultural" (Cavalli-Sforza and Feldman 1981: 68): Conversely, it s clear that gene flow can be only demic. Finally, the process of drift that intervenes in biological models (the genetic drift) perfectly applies to cultural evolution as well: In fact the drift just determines a random fluctuation of the frequency of any type of traits in a population and from this point of view it acts purely quantitatively, i.e. independently of the intrinsic meaning of the traits. Also in cultural evolution the drift is expected to have an homogenizing effect on the population and to act faster in smaller groups of individuals (see section 2.1.4).

The modalities of cultural transmission between individuals are rather different instead. While the individual transmission of genes that occurs during reproduction is vertical (apart from the marginal cases of lateral gene transfer mentioned in the end of section 2.1.4), Cavalli-Sforza and Feldman (1981) argue that the situation is more complex in cultural evolution. In particular, they identify four types of individual transmission. The first one may be assumed to be regularly (from the biological point of view) vertical, in that it occurs from one generation to the next between persons that have either a biological or a strong and lasting social relationship, e.g. from parents to children. Then the scholars recognize the possibility of individual horizontal transmission between two persons that do not have either a biological or a strong and lasting social relationship (e.g. friends): In particular, they call this transmission "oblique" if the transmitter belongs to an older generation respect to the recipient. The other two types of individual transmission introduced by the scholars are one-to-many, from one transmitter to many recipients (e.g. from a teacher to her students), and many-to-one, from many transmitters to one recipient (e.g. from a whole family to a new member)[17]. What about inheritance? The Mendelian model adopted in evolutionary biology only marginally applies to cultural evolution. On the one hand cultural units are often neither discrete nor "atomic", thus nothing prevents them from being blended (e.g. a child brought up by parents used to different diets may decide to mix them). On the other hand, cultural units are also subject to change and may be acquired at any moment in the course of an individual's life (apart from few cases in which a "critical period" for the acquisition is assumed, see Cavalli-Sforza 2001: 281-285), and what is effectively transmitted is the state of these units as they appear in the specific period in which the transmission occurs. In this sense, cultural transmission is more similar to a Lamarckian than a Mendelian process, because in practice it always entails inheritance of acquired characters.

2.3.2. An evolutionary perspective on language development

As for language evolution, three premises have to be pointed out before proceeding. The first one is an observation regarding the above mentioned four evolutionary factors: Only one of them refers to the individual and specific occurrence of an innovation, i.e. mutation, while the others, namely selection, genetic drift and migration, are relevant to describe the diffusion of

---

[17] The terms "vertical" and "horizontal" are borrowed from epidemiological studies, while the terms that indicate other types of individual transmission are introduced for the first time by Cavalli-Sforza and Feldman (1981) (see Cavalli-Sforza 2001: 320).

the innovation. In Croft's (2000) terms, mutation corresponds to altered replication, whereas selection, genetic drift and migration regard the mechanisms of differential replication, or propagation.

The second premise concerns the following question, that represents a central issue of linguistics: What is meant to be "language"? Compatibly with the generative perspective adopted in this work (again, see section 3.1 for a detailed account), here it is assumed Chomsky's (1986a: 15-50) distinction between Externalized-language and Internalized-language. The scholar observes that in modern linguistics scientific approaches have developed technical concepts of language that depart from commonsense ones, e.g. definitions of language that rely on socio-political or normative-teleological descriptions: Usually all these scientific approaches (e.g. structural and descriptive linguistics, behavioural psychology) tend "to view a language as a collection of actions, or utterances (e.g. Croft, see section 2.2.2; editor's note), or linguistic forms (words, sentences) paired with meanings, or a system of linguistic forms or events" (Chomsky 1986a: 19). The scholar proposes considering such concepts as instances of E-language, which is defined independently of the properties of individuals' mind/brain. Conversely, according to him, I-language can be intended as "some element of the mind of the person who knows the language, acquired by the learner, and used by the speaker-hearer" (Chomsky 1986a: 22). Chomsky argues that, while I-language can be a real object of scientific inquiry, E-language is epiphenomenal and represents an artificial construct from which generative grammar has shifted the focus of attention. How is I-language related to the above mentioned language faculty? As already claimed, the language faculty is assumed to represent a distinct system of the mind/brain, which is innate and common to the species: According to Chomsky, "given the appropriate experience (which is culturally transmitted; editor's note), this faculty passes from the state S_0 to some relatively steady state S_s, which then undergoes only peripheral modifications (say, acquiring new vocabulary items)" (Chomsky 1986a: 25). The point is that I-language should be incorporated within S_s. This (originally) synchronic approach to the study of language can be applied also to the diachronic study of language evolution, in particular of diachronic syntax. A detailed presentation of the issue is provided in the next chapter, with particular reference to Longobardi's (2003b: 172-174) model regarding the historical relations between I-languages (section 3.2.3). For the moment it is important to point out that the above mentioned framework represents a fundamental premise for all "child-based theories of language change", i.e. theories of language change which assume that the change originates in the stage of language acquisition. Moreover, notice that in the following discussion whenever the term "language" is used without further specifications, it is meant to be E-language.

Finally, the third premise is that the term "linguistic traits" is too general to be useful when one studies language change[18]. Each level of analysis of language (say, lexical semantics, phonology, morphology and syntax) should to be kept distinct from the others, because at different levels different kinds of changes occur. To provide a detailed account of all the levels of analysis of linguistic traits is beyond the aims of the present work: Thus the discussion that follows in this chapter is intended to be as general as possible, even if brief explanations regarding the different evolutionary dynamics of the specific levels have to be often provided. Again, notice that the specific modalities of "syntactic evolution" from the point of view of Principles and Parameters Theory are illustrated is detail in section 3.1, while section 5.1 covers the critic overview of the role of lexical and morpho-phonological comparative data in quantitative phylogenetic studies.

How do linguistic innovations arise? Are they more similar to genetic innovations or cultural innovations? In Chomskian terms, as it is more or less implicitly known since Schmidt (1872),

---

[18] Actually the same observation may apply also to the concept of "cultural trait".

innovations do not abstractly originate in E-language, but they rather show up in I-language and then spread in the speech community in the course of time. Notice that the distinction between innovation and diffusion would be lost, if one considered languages organisms, and language change could be described only as "inherent change", which is too general and unrealistic. Moreover, one may assume the idealization according to which some innovations mainly arise during an individual's attainment of the steady state (see the above mentioned quotation of Chomsky 1986a: 22): This hypothesis can be considered valid at least for what concerns the "narrow language faculty" as defined by Hauser *at alii* (2002), i.e. syntax, and nothing more is said here for the moment.

It has been shown in which sense genetic mutation is random and often harmful while cultural innovation is directed and intentional. As for language evolution, a characterization of changes in terms of "harmful" or "advantageous" does not apply, as explained later discussing the role of natural selection. Moreover, it seems necessary to admit that a unified characterization of language change in terms of intentionality cannot be provided: For instance, while on the one hand syntactic innovations could be thought of as unintentional changes, i.e. not consciously directed by the speaker, as they are described in the current generative framework (i.e. as processes of parameter resetting, see section 3.2.1), on the other hand there are changes that affect peripheral lexicon, which are more likely to be consciously driven by individuals (see for instance the introduction to the concept of "etymological motivation" in Benedetti 2003: 215-216). Some aspects of this issue are confronted in more detail later in this section, when the question of inheritance is taken into account.

For the moment, notice that something more general can be said about randomness instead: Even if language changes remain unpredictable[19] , they often appear to be not random, but rather "structurally driven", i.e. somehow influenced by "structural" properties of the linguistic system (in the sense intended, for instance, by Kiparsky in press). This is simply another presentation of a fundamental issue investigated by diachronic linguistics: The identification of universals and paths of change. Indeed "[u]no degli aspetti del mutamento linguistico più singolari […] è la frequente insorgenza – già nota ai neogrammatici – di un certo grado di simmetria (one of the most peculiar aspects of language change is the frequent occurrence – already known by Neogrammarians – of a certain degree of symmetry" (Lazzeroni 2006: 975), which, according to Lazzeroni (2006), represents the outcome of diachronic directed processes. To identify directions of change appears to be hard for lexical semantic-changes, since they are particularly subject to apparent arbitrariness (due to the arbitrariness of the relationship between *signifiant* and *signifié*) and they are sensible to factors external to language, e.g. specific social and historical contexts of use[20]. However, significant generalizations about paths of semantic change, e.g. the irreversibility of shifts from concrete to abstract, have been clarified (see among others the summary provided by Lazzeroni 1996: 24-30; Traugott and Dasher 2002; Giacalone and Mauri to appear). In phonology different types of "structure-dependent" changes have been investigated at least since the '20s (see for instance Kiparsky 1995: 3). Some paths of change and constraints

---

[19] As claimed by Croft, there are two positions regarding the possibility of constructing a predictive model of language change. According to the optimistic one, if linguists knew in enough detail all relevant linguistic facts, they would be able to exactly predict the occurrence of changes. Oppositely, for the pessimistic view "we would never be able to predict the change because there is at least some element of randomness in the process, as is the case with the randomness of mutation in biological reproduction and the random factors for survival of individual organisms" (Croft 2000: 3). However, the existence of a large set of structurally driven changes might suggest that, provided a full understanding of "the relevant linguistic facts", in principle it could be possible to predict them at least in probabilistic terms.

[20] Among the strangest cases, see for instance the etymological derivation of Italian "*cotillon*" ("gift that is offered at the end of a ball"), which is a borrowing from French "*cotillon*" ("skirt").

appear to be related to the inventory of phonemes of specific languages. For instance, having studied the inventories of phonemes on a wide sample of languages, Maddieson (1984) affirms that "[t]he total number of consonants in an inventory varies between 6 and 95 with a mean of 22.8" and "[t]he total number of vowels varies between 3 and 46 with a mean of 8.7" (Maddieson 1984: 9): Looking at these typological data, one may hypothesize that there are upper and lower limits to the number (and type) of phonemes[21] that can be included in a specific inventory, thus the phonetic changes that lead to create or merge phonemes (namely, splits and mergers) might tend to be somehow constrained. Moreover, mergers are well known to be irreversible processes, since the original phonemes are usually not restored by successive changes. Other paths of phonological change concern articulatory constraints on segmental contiguity, which are also well-studied within natural phonology (Dressler 1976). For instance, the consonantal cluster /tl/ appears to be marked in some languages, e.g. Italian, as it is shown by the following paradigmatic development: Latin "*vetulus*" ("old") > "*vetlus*" > "*veclus*" > Italian "*vecchio*". This fact is also related to the specific distribution of coronals[22] that is observed in Latin as well as in other languages (Marotta 1993). As for morphology, one of the most famous collection of works about universal paths of change is probably Greenberg *et alii* (1978). However, as pointed out by Giannini (2003: 94), morphological change cannot be easily characterized on its own, because the boarder between morphology and other levels of linguistic analysis, namely phonology and syntax, often appears "mobile" in diachrony. Finally, structurally-dependent change is particularly clear in syntax, especially from the point of view of the generative analysis: The structural dependency follows from the hypothesis that not all syntactic rules are possible and that the space of parametric variation is limited. The detailed explanation of these assumptions is covered in section 3.1.

The structural dependency of linguistic change may be interpreted as some kind of "structural pre-selection" in evolutionary terms, in the sense that, even before occurring, some changes are favoured over others, i.e. they have a higher probability of occurrence than others. Notice that similar mechanisms intervene also in biological evolution and, for instance, they regard the occurrence of specific types of nucleotides substitutions in mutation. Two distinct types of DNA substitutions are recognized: When either a purine is replaced by another purine nucleotide (A ↔ G) or a pyramidine by another pyramidine nucleotide (C ↔ T) the mutation is called transition, while when a purine changes to a pyramidine nucleotide or the other way around, a transversion occurs. Transitions are much more common than transversions. Thus nucleotides have a direct influence on the occurrence of the substitutions and, even if they do not prevent the genetic system to be altered in one way or another, their distribution makes some paths of mutation more or less probable than others[23].

---

[21] In the '40s Jakobson (1941) pointed out that "[a]n *a* emerges as the first vowel, and generally a labial as the first consonant, in child language. The first consonantal opposition is that of nasal or oral stop (e.g. *mama-papa*), which is followed by the opposition of labials and dentals (e.g. *papa-tata* and *mama-nana*). These two oppositions form the minimal consonantal system of the languages of the world" (Jakobson 1941: 47-48). Moreover, the scholar proposed a hierarchy of implicational rules ("laws of solidarity") based on evidence provided by the study of child-language acquisition and synchronic typology: For instance, according to the first of these rules "[t]he acquisition of fricatives presupposes the acquisition of stops in child language; and in the linguistic systems of the world the former cannot exist unless the latter exist as well" (Jakobson 1941: 51).

[22] Among others, see Kiparsky (1985) for the unmarkedness of coronal place of articulation ("coronal syndrome").

[23] As for cultural evolution, generalizations and predictions on patters and possibilities of cultural evolution can probably be provided only in a very tentative and approximate way. Nonetheless, if one looks at the interaction between two specific cultural subsystems, i.e. the religious and the moral one, some reasonable predictions can be made. For instance, it is likely that a community, in which the vast majority of people profess Catholicism, would be less prone to accept an innovation regarding the

What about gradualness? Language change in general appears to be gradual, since evidence shows that “different linguistic features shift at different times, and the different variants coexist, sometimes in the same text” (Croft 2000: 49). There are many examples that seem to support this claim: One of them may be, for instance, the fact that, studying early middle English texts, Allen (2000) does not find evidence “of an abrupt shift from OV to VO word order in that period; in fact, the coexistence of both orders must be accommodated” (Croft 2000: 51). The gradualness of change seems to be particularly problematic for child-based theories of language change: In particular, parametric theories of syntactic change predict an abrupt arise of innovations, or, in Lightfoot’s (1979; 1991; 1999) terms, changes are assumed to be catastrophic. The issue of gradualness respect to parameter-resetting models is presented in detail in section 3.2.3. For the moment, notice that in effect the discussion of gradualness of change might be biased by the neglect of a fundamental distinction, that has been implicitly assumed speaking about “gradualness” in biological and cultural evolution so far: Again, the distinction between facts that concern the synchronic occurrence of innovations and those that regard the diachronic diffusion of innovations. Take for instance biological evolution. Genetic changes are gradual only from the point of view of diffusion: Mutations are sudden changes at organismic level, because they affect discrete units (i.e. genes), but what is assumed to be gradual is their accumulation in populations and the consequent process of speciation. The discussion about gradualness of biological change presented in 2.1.4 precisely regarded the process of diversification of populations. In the same way, one should expect language change at the level of E-language diachronic development to appear gradual, i.e. to determine a gradual diversification of E-languages, due to its progressive spread in the speech community. Nonetheless, nothing prevents innovations that occur in I-languages to be abrupt: The gradualness of individual innovations depends only on the type of linguistic trait one is considering, which means that the fact that discrete linguistic traits, e.g. syntactic parameters, are assumed to undergo sudden change at individual level, does not contradict the evidence, according to which language change is gradual at E-language level. Nothing more specific is said about this aspect of linguistic innovations here, but notice that to account for gradualness of language change requires to distinguish on the one hand between innovation and diffusion, on the other hand between linguistic traits that are discrete (e.g. the syntactic ones within the generative framework) and those that show a continuous variation (e.g. sounds in a phonetic perspective).

Apart from one exception, proper natural selection does not intervene in language evolution, since there is no evidence for adaptation of linguistic traits: One may not claim, for instance, that a specific climate influenced the diffusion of a specific syntactic innovation, say the already mentioned shift from OV to VO word order in early Middle English. For the same reason, linguistic innovations cannot be regarded as “harmful” or “advantageous” for a language: There is no advantage in using a specific word order instead of another or in having 36 consonants instead of 37 in an inventory of phonemes in consequence of a merger. At most, similar hypotheses could have been taken in consideration until Jespersen (1922), but not in modern linguistics (see section 2.3.1). The exception for what concerns the effect of natural selection regards lexical-semantics, which is known to conserve traces of the territory inhabited by a speech community: For instance, this observation is particularly relevant for archaeological-linguistic studies aimed at reconstructing the geographic location of proto-languages (e.g. Gimbutas 1970). Social selection, instead, applies to language evolution (even if differently from culture it is not followed by natural selection): The effect of social factors on diffusion of linguistic innovation is studied in sociolinguistics at least since the already

---

moral acceptability of, say, euthanasia. In this sense religious belief could modify the probability of change of a specific moral trait and one might claim that the process of (pre-)selection is internally driven.

mentioned Weinreich *et alii*'s (1968) work and Labov's (1966; 1977) quantitative analysis. In particular, Labov introduces the concepts of "diastratic" (depending on social class) and "diaphasic" (depending on contexts of use) dimensions of linguistic variation. For the aims of the present work, it s not important to go deeper in detail into the issue of sociolinguistic variation, but it is important to consider that social selection is relevant in language evolution as it is in culture.

Moreover, the process of drift applies to language evolution a well, for the reasons outlined in section 2.3.1: Essentially, because the drift is a purely quantitative process. The role of migration in language evolution, instead, requires a discussion on its own, in that it is inextricably bounded to the issues of language contact. At first glance, it seems that the distinction provided for culture between demic and cultural migration could be a good starting point for the definition of a typology of language contact: Actually, much more than this has to be said. The next section is focused on the general distinction between genealogical evolution and areal convergence, while a detailed overview on the (past and) current proposals of contact typology is provided in section 3.3.1.

However, before proceeding it is important to point out an observation. Until here, the term "transmission" has been used with the same meaning intended by evolutionary biology and Cavalli-Sforza and Feldman (1981), specifying it as "individual transmission" in the relevant contexts. In other words, it has been intended as the passage of traits between individuals, and, for what concerns language evolution, the passage of knowledge that occurs in the course of language acquisition. In this sense, individual transmission is a culturally-driven process that determines the formation of I-language. One may wonder how similar is this process with respect to the biological ones, but the comparison does not lead too far. A first statement about individual transmission of linguistic traits might be that it does not entail a Mendelian-like process of inheritance: The process appears to be memoryless, in the sense that, unlike genetic transmission, previous states of the linguistic traits are not preserved and they neither surface nor do they influence successive stages of diachronic development of I-language (as for syntax, see for instance Longobardi 2003b: 180-182). A clear example of this idea is provided by the already mentioned irreversibility of phonological mergers. Nonetheless, the mechanism of inheritance does not even appear to be Lamarckian in general, because no significant modifications should affect (at least) narrow components of linguistic knowledge in Chomsky's (1986a: 25) sense after the attainment of the "steady state", which entails that there is no generalized inheritance of traits that are acquired in the course of the entire life, but only of those that are acquired during the "critical period". In other words, there is a substantial difference between individual language transmission and individual biological transmission, since the former is assumed to be spread during the whole critical period while the latter is a punctuated event: However, this difference does not contrast with the fact that once the process is completed (after the critical period for knowledge of language and after reproduction for genetic material), in both cases the relevant traits (narrow components of I-language and genes) do not undergo further modifications in the course of an individual's life[24], so that they cannot be inherited in Lamarckian sense. The situation is rather different for non-narrow aspects of language: Lexical acquisition and change, especially in the case of non-basic vocabulary items as intended for instance by Swadesh (1952), seem to overcome the limits of the critical period, so that the transmission of these items would properly entail "inheritance of acquired (in Lamarckian sense) characters". Thus no clear and general

---

[24] Apart from cases of traumatic neurological and physiological alterations, which are not related to the language faculty, but to biological acquired modifications that, according to the Mendelian model, are not passed to successive generations.

mapping between biological and linguistic inheritance mechanisms can be provided[25]. As for the different modalities of individual cultural transmission outlined by Cavalli-Sforza and Feldman (1981) (vertical, horizontal, one-to-many and many-to-one), they could probably be of some use in the classification of the modalities of individual language transmission, in that they reflect possible instantiations of the process, but the discussion of this issue is not relevant here.

What is important now is that the term “transmission” has been used with a different meaning in recent literature. For instance, Wang and Minett’s (2005) work is dedicated to the quantitative analysis of the role of “vertical and horizontal transmission in language evolution”: Nevertheless with “vertical transmission” the scholars refer to the kind of language evolution that is due exclusively to the accumulation of innovations independently arisen in the history of a language, i.e. independently of contact situations, while with “horizontal transmission” they mean the evolutionary paths in which contact-induced changes occurred. In other words, Wand and Minett’s intention is to distinguish between pure genealogical evolution and evolution due to areal convergence. This use of “transmission” is not equivalent to Cavalli-Sforza and Feldman’s (1981) one: If it were, it would entail Wang and Minnet’s analysis to concern how individual transmission occurs in languages assuming that languages are like organisms, which undergo inherent change and processes somehow similar to biological lateral gene transfer. Actually this is not the case and Wang and Minnet (2005) do not even mention such problematic aspect of their terminological choice. Rather, the purely descriptive use of terms like “vertical transmission” and “horizontal transmission” appears to be particularly useful to classify two important types of evolutionary paths of language evolution, namely those described in the next section. Thus, provided that “vertical transmission” and “horizontal transmission” do not entail any conceptual parallelism between languages and organisms and that they have a high descriptive value, from now on these terms are used to indicate respectively pure genealogical evolution and areal convergence, while the cases in which it is indicated a passage of traits (being them genetic, cultural or linguistic) between individuals is referred to with “individual transmission”.

2.3.3. Genealogical evolution and areal convergence

As anticipated in section 2.2.1, the occurrence of contact-induced changes in the history of language is problematic for a model of language evolution, in which only pure genealogical development is represented. In particular, this question is well-known in linguistics at least since Schmidt (1872), who argued against the usefulness of Schleicher’s (1853) *Stammbaumtheorie*, because in his opinion it was insufficient to represent the actual dynamics of language change. Thus Schmidt proposed an alternative model, the *Wellentheorie*, which was focused on the diachronic spatial variation of changes, instead of the pure genealogical one. In current sociolinguistic terms, this means that he proposed to look also at the “diatopic” dimension of variation. The consequence of considering the spatial dimension of language variation is that one has to account also for the fact that linguistic innovations spread in linguistic areas in the course of time: Consequently, the contact between neighbouring languages may lead them to share some innovations, i.e. to converge with respect to some linguistic traits, even if there is not a close genealogical relationship between them. Thus the effect of language contact is perfectly comparable with that of migration in population genetics: While innovations tend to differentiate similar languages, contact tends to homogenize different ones.

---

[25] A general account for the “blending” aspects of language transmission would entail a confusing digression here, thus it is not provided. However, relevant observations about this point will be pointed out in introducing the details of the generative perspective on syntactic change (section 3.1).

Notice that the debate regarding the usefulness of tree-like models of language evolution is still open today and has become even more heated in recent times, as a consequence of the increasing popularity of quantitative approaches of reconstruction in historical linguistics. This is because quantitative phylogenetic techniques typically infer genealogical relationships in a "Schleicherian" tree-graph form (see section 4.1.2). Thus in linguistics, while some scholars seem to have much faith in the potential of the quantitative phylogenetic analysis, others believe that this approach is mainly inadequate or even hopeless (e.g. Moore 1994; Terrell 1988; Terrell *et alii* 2001; Heggarty 2004): Because it implies an oversimplified version of the actual dynamics of cultural evolution[26]. In other words, in critics' opinion "a family-treelike diagram does not adequately demonstrate the many kinds of historical and current relationships between languages" (Aikhenvald and Dixon 2001: 6). Haspelmath (2004b) and McMahon and McMahon (2005: 18) observe that family trees are certainly not sufficient to represent all the possible historical relations occurred between languages, nonetheless these structures are useful to highlight one specific kind of classification, i.e. the genealogical one, and if "a language has a single parent, then it follows that (apart from the truly exceptional cases) constructing a family tree of languages sharing a common ancestor is in principle possible" (Haspelmath 2004b: 214). Of course there are many problems related to this task: One of them is to distinguish between features that are present in a language due to pure genealogical evolution or to areal convergence. Such old-age requirements of the comparative method come from the awareness that the presence of contact-induced change between linguistic varieties is biased to the clean pattern of vertical transmission of the relevant genealogical traits (see among others Hübschmann 1875; Thomason and Kaufmann 1988: 206-211; Dench 2001; Haspelmath in press: 2; Wang and Minett 2005: 122): In fact not only areal convergence homogenizes different languages, but it also causes the differentiation of genealogically-close ones. The consequences of this fact for the adoption of quantitative methods of reconstruction are illustrated in section 4.3.3. For the moment, a relevant answer that would need to be addressed, as effectively pointed out by Haspelmath, is "to what extent are the relationships [between languages] shaped by areal diffusion rather than genealogical relatedness?" (Haspelmath 2004b: 209). In Wang and Minett's (2005) terms, how much do vertical and horizontal transmission respectively contribute to language evolution?

Dixon's (1997) diachronic model provides a radical answer: The scholar argues that genealogical development of languages is the exception, being the result of just momentary "punctuation" events (such as the introduction of agriculture or other relevant technologies), while the norm should be an "equilibrium" situation, in which languages coexist and repeatedly interfere with one another. In other words, in Dixon's "punctuated equilibrium model", which recalls only superficially the homonymous evolutionary model mentioned in section 2.1.4, the weight of the horizontal transmission is much greater than that of the vertical one in determining linguistic diversification: That is why the "family-tree" model should be substantially inadequate (Aikhenvald and Dixon 2001: 6). However the evidence presented in Dixon (2001) in support of this hypothesis is debated. He refers to the Australian linguistic area, providing a list of features that are mostly common to the languages of that area, but even if "this is a rich list, this procedure falls short of demonstrating that Australia is a linguistic area, because linguistic areas need not only be internally coherent, but also distinctive with respect to languages outside that area" (Haspelmath 2004b: 211). This kind of evidence is not provided by Dixon (2001), thus the model cannot be seriously tested. Dench's (2001) study on 17 languages from the wider Pilbara region (western Australia) is intended to support the "punctuated equilibrium" hypothesis as well, but it also appears weak, in the sense

---

[26] Notice that the problem of non-genealogical evolution of *taxa* is also known in philology with the name of "contamination" (e.g. Windram *et alii* 2006). See section 4.3.3 for further details.

that it does not definitely exclude genealogical transmission from the development of these varieties.

Ross's (2001) reply to Aikhenvald and Dixon's (2001) position is that areal and genealogical transmission do not provide mutually exclusive evidence: For instance, he claims that the influence of Waskia on Takia (see section 3.3.1) made the Oceanic variety significantly diverge from its genealogical group, and this fact could be perfectly represented into a tree structure. Watkins's (2001) criticism to the "punctuated equilibrium" hypothesis, instead, is based on the data regarding ancient Anatolian languages, which include non-IE (Hattic and Hurrian) and IE (Hittite, Palaic and Luvian, among others) varieties: The fact that in a relatively short period (approximately between 2200 and 1700 BC) all these languages converged phonologically and syntactically contradicts Dixon's model, according to which genealogical splits are supposed to be rapid "punctuation" events, while areal diffusion occurs during equilibrium periods. Similar cases reported by Watkins are those of the Balkan peninsula and of the South Asian area. Therefore, on the whole Dixon's scenario, though intriguing, appears to have little empirical support and, as claimed by Haspelmath, probably "concepts of equilibrium and punctuations are too general and too vague to be of much use when particular languages and historical situations are examined" (Haspelmath 2004b: 214).

Therefore horizontal transmission does not appear to have an overwhelming role in linguistic diversification with respect to genealogical inheritance, with the exception of languages which underwent abnormal formation and development: These include cases of massive and reiterate contact, abrupt creoles and pidgins (Thomason and Kaufmann 1988: 211). However, also extreme positions on the opposite side appear to be too optimistic. For instance, Nakhleh *et alii* (2005) are convinced that outright borrowing other than lexical between significantly different speechforms is a rather marginal phenomenon, thus they expect it to have very little influence on the genealogical reconstruction. But the typology of language contact, especially for syntax, can hardly be restricted into such absolute generalizations (see section 3.3.1), and for the moment the actual impact of areal diffusion of linguistic traits can hardly be estimated *a priori*. As a consequence, in general, Thomason and Kaufman's observation is always valid: It is necessary to "consider the areal situation as well as the genetic [or better, genealogical (Haspelmath 2004b: 222)] picture in carrying out any comparative reconstruction" (Thomason and Kaufmann 198: 208). So the answer to the question posed in the middle of this section is "it depends". The contribution given by horizontal and vertical transmission to linguistic evolution may vary a lot, according to different social and structural factors regarding the type of contact, the languages in question and the "linguistic traits" one is considering. This does not mean that either the *Stammbaumtheorie* or the *Wellentheorie* approach are absolutely inadequate, but that both of them can represent different models to extract different information from the history of languages: Depending on the language group one is examining, either of them can then provide more or less meaningful insights.

2.4. Conclusion

In this chapter some features of language evolution have been identified drawing a parallel with biological evolution. Moreover, an intermediate step toward the investigation of analogies and differences has been the illustration of a general model of cultural evolution. An overview of the factors and features that, according to this analogical approach, appears to be relevant in modelling biological, cultural and linguistic processes of evolution is provided in table 3. In particular, in light of this comparison the most relevant differences between biological and language evolution seem to concern the role of selection, which in the case of languages is not natural (i.e. not caused by the interaction with the environment, except for certain part of lexicon), but rather structural and social, and the mechanisms of individual transmission and inheritance. Nonetheless, some analogies between the two processes might concern other evolutionary factors, i.e. mutation/innovation and migration/contact: These factors, instantiated with different names in the two domains, entail precisely the same effects on variation, in that the former occurs punctually and has a diversification effect on a population/language, the latter is a projected over a certain time span and has a homogenizing effect instead.

Notice that genetic mutation and linguistic innovation might be assumed to intervene in a similar way in biological evolution and language evolution only if one considers linguistic traits that are included in the "narrow linguistic faculty", e.g. syntactic traits, while population migration and language contact are not perfectly corresponding processes: On the one hand the former is exclusively demic, while the latter has to be analyzed according to a much more complex typology (basically, it may be either demic or cultural), on the other hand migration applies only between interbreeding populations, while language contact may always cause areal convergence of different languages, even if the extent of its homogenizing effect may vary a lot according the type of contact, the languages in question and the "linguistic traits" one is considering.

On the whole, this comparison seems to show that only analogical parallels between biological evolution and language evolution can be outlined, since the two processes are substantially different in many respects. However, these dissimilarities do not prevent the application of some quantitative reconstruction procedures, as shown in sections 4.2 and 4.3 and chapter 6: Essentially, the most problematic aspect of language evolution from the point of view of a quantitative analysis is that one has to consider not only vertical, but also horizontal transmission of innovations in the history of languages. Moreover, the dynamics of language development identified in light of the parallel presented in this chapter appear to be perfectly compatible with the most recent generative views on syntactic change. The next chapter covers a detailed account of these theories.

3. Syntactic evolution

The use of the term “evolution” in historical linguistics has been considered in detail in the last chapter. It is important to recall that here “evolution” refers to the general process of diachronic development of languages, regardless if it occurred due to genealogical or to contact-induced change: Thus “evolution” may be considered a synonym of “diachronic variation”. As pointed out throughout section 2.3, one cannot provide theoretical generalizations about language evolution abstracting away from the development of specific types of “linguistic traits”: The dynamics of language change and language contact may appear to be very different if one considers, for instance, lexical-semantic “traits” and phonological “traits”. In this chapter the specific modalities of syntactic evolution from the point of view of the Principles and Parameters Theory are illustrated. Since the aim of the dissertation is to show how hypotheses regarding the history of languages can be reconstructed comparing them respect to their syntactic systems, it is important to understand how these systems developed in the course of time. Moreover, the focus on the parametric approach to the study of diachronic syntactic variation derives from the fact that the data, on which the experimentation presented in this work is based, are exactly parametric. The following discussion profits from the most recent applications of generative grammar to the study of diachronic syntax: The theory developed in the last years is based on the premise that the theoretical advances achieved in providing a parametric account for syntactic variation may also be exploited to shed light on the dynamics of diachronic phenomena. However, it is important to underline that many concepts and ideas regarding the study of syntactic change used within this (relatively) innovative approach derive from the tradition of work on the subject: Throughout section 3.2.1 the history of the study of diachronic syntax is briefly outlined in order to highlight the role of the traditional achievements in current theory.

The chapter is organized in three main sections. In the first one the basic methodological points and concepts of the parametric approach are taken into account. Section 3.1.1 is dedicated to the introduction of the methodological plan designed by Chomsky (1964) and expanded by Longobardi (2003a): According to it, the study of language may be organized in successive levels of adequacy of the grammatical descriptions. After this clarification of the directions of inquiry, section 3.1.2 and section 3.1.3 are thought to provide an overview respectively of the cognitive-biolinguistic and of the formal aspects of generative grammar: These points represent a fundamental ground to understand the discussion regarding the parametric approach to linguistic variation. Sections 3.1.4 and 3.1.5, instead, illustrate some current hypotheses on the forms of parametric variation and interaction. The views on syntactic change developed from the synchronically-based framework of the Principles and Parameters Theory are covered in section 3.2. Section 3.2.1 gives a basis to link the possible descriptive typologies of syntactic change to a coherent explanatory account. On this ground, the extension of the parametric approach to the explanation of historical facts is covered in section 3.2.2, while section 3.2.3 is dedicated to the issue of gradualness. These subjects are followed by the presentation of some recent attempts to model population-level (or better E-level) parametric evolution using computational simulation. The issue of contact is addressed in section 3.3: After the introduction of (past and) current general hypotheses on contact typology, the meaning of “borrowing” that is relevant for the present work is defined and evidence in favour of the possibility of syntactic borrowing is introduced. Finally, in section 3.3.2 it is explained how syntactic borrowing can be integrated in the parametric perspective.

3.1. The parametric framework

This section is intended to bring in the methodological and theoretical foundations of the parametric perspective developed within generative grammar. Some ideas have already been mentioned in the previous chapter: In particular, in section 2.3.1 a simplified account of the "cultural aspect" of language evolution and acquisition has been provided, while in section 2.3.2 both the E-language/I-language distinction and the question regarding the structural constraints that limit syntactic variation have been introduced. Now these points have to be reconsidered within the general picture.

Some terms have to be clarified first. For the aims of the following discussion, "grammar" may be defined as "the study of the principles (in a general sense, i.e. "rules", editor's note) which govern the formation and interpretation of words, phrases and sentences" (Radford 2002: 1). This definition of grammar is as broad as possible, in that it includes semantics, phonology, morphology and syntax. Instead, the grammar of a language has to be understood as the description of how words, phrases and sentences are formed and interpreted in that language. In the same way, while the term "generative grammar" is intended as the theory of language proposed by Chomsky[1] and it indicates the corresponding field of study, a specific generative grammar refers to the description of a specific language.

These distinctions are made clearer in the following section, which opens with a careful consideration of the aims of the linguistic inquiry, i.e. of the methodological plan designed since Chomsky (1964) in the form of a hierarchy of successive "levels of adequacy". Longobardi (2003a: 2) claims that "[a] good deal of misunderstanding in modern linguistics seems to me the consequence of misconstruals of the relations between the various questions and levels (of adequacy, editor's note)": Indeed, some harmful consequences of the misleading overlapping of methods and levels of study have already been highlighted in section 2.2.1 stressing the need to distinguish the linguistic perspectives that make use of evolutionary ideas to model historical language development from those regarding the evolution of language faculty. In the following section the current methodological paradigm is outlined in detail: To understand it will be essential in order to appreciate the cognitive-biolinguistic and formal aspects of generative grammar respectively presented in sections 3.1.2 and 3.1.3. Finally, sections 3.1.4 and 3.1.5 are dedicated to the overview of the hypotheses aimed at shedding light on the format of parameters and on their interaction, i.e. on the formal characterization of the parameter space: Notice that, in order to provide a clear explanation of the theoretical approach adopted here, in this part of the discussion the debate regarding some of the most controversial issues is largely left out to be introduced in the notes.

3.1.1. Levels of adequacy

In the '60s, Chomsky outlined a series of "levels of success" that "might be attained by a grammatical description associated with a particular linguistic theory" (Chomsky 1964: 28). These hierarchical levels are three in the original Chomskian account: the first one is that of the observational adequacy, which is accomplished when the grammatical description rightly represents the observed linguistic data; The second level corresponds to the descriptive adequacy and its satisfaction follows from the capability of the grammar to correctly describe native speakers' intuitions and to provide generalizations that show hidden regularities in the specific language under study; The third and highest (among these) levels, that of explanatory

---

[1] However notice that, according to Chomsky, generative grammar "is sometimes referred to as a theory advocated by this or that person", but actually it "is not a theory any more than chemistry is a theory" and it has rather to be considered "a topic, which one may or may not decide to study" (Chomsky 1986a: 4).

adequacy, is attained by a linguistic theory that "aims to provide a principled basis, independent of any particular language, for the selection of the descriptively adequate grammar of each language" (Chomsky 1964: 29). As pointed out by Longobardi (2003a), these three levels of adequacy correspond to an equal number of fundamental questions that may (and should) help to direct the linguistic investigation. These questions are respectively: "What are the recorded samples of human linguistic behaviour? What are the *actual* human languages? What are the *biologically possible* human languages?" (Longobardi 2003a: 2).

These levels of adequacy entail many relevant shifts of focus for linguistic inquiry. Longobardi (2003a: 3) says that the shift from the first to the second level parallels the shift from the study of E-languages to that of I-languages. This claim may be understood if one considers the implied features of these two different objects of study. In section 2.3.2 E-language was merely intended as "Externalized-language" and I-language as "Internalized-language", meaning that the former is a collection of utterances produced by native speakers and the latter the native speaker's knowledge of a specific language. In other words, E-language may be assumed to correspond to the observed language behaviour of a speech community and it has to do with the actual performance of native speakers, while I-language is related to native speakers' competence. Actually "E" and "I" may be associated also with other features. E-language is not only an external object, but is only "Extensional", because it consists of a physically limited subset of the native speakers' linguistic intuitions and its observation relies only on the extensional recording of expressions. Conversely, I-language, as well as internal, may be thought of as "Individual" and "Intentional", in that its description allows for the definition of all possible native speakers' linguistic intuitions, which are infinite, by means of the interaction of a limited set of rules. Referring back to the definitions provided at the beginning of section 3.1, an I-language corresponds to a specific generative grammar. In Chomsky's (1986a) terms, "the study of generative grammar represented a significant shift of focus […] from behaviour or the products of behaviour to states of the mind/brain that enter into behaviour" (Chomsky 1986a: 3). Thus, while a recording of linguistic facts defines an E-language and can attain at most the level of observational adequacy, the descriptive adequacy may be achieved only by the correct description of an I-language.

Advancement to a level of explanatory adequacy implies a fundamental progress of the aims of the generative program itself instead. In particular, Boeckx and Hornstein (2007: 5-6) consider the shift to explanatory adequacy the first methodological premise for the beginning of what they call "Cognitive period" of generative grammar. According to the scholars' review of the history of the field of study, the program has passed three (overlapping) periods, or stages of development, each one influenced by a specific work of Chomsky and connectable with a distinct "better-developed science". According to the authors, the first stage, associated with Chomsky (1957) and the relation with engineering, was the Combinatoric one and aimed at finding out an explicit formalism for representing language facts. The second stage, initiated by Chomsky (1965), was determined by the shift of focus to the problem of language acquisition and determined the program to be framed into a biological perspective: This era is the Cognitive or "biolinguistic" one. After the "principles and parameters" approach (see section 3.1.3) developed since the '80s was proven to be suitable from a methodological and empirical point of view to overcome the problem of language acquisition, the main theme of generative grammar has become the minimalist reduction of the theory in order to achieve elegance and simplicity, i.e. the unification of the modular theoretical subsystems similar to that pursued in physics since the first half of the last century. This third period is the Minimalist one and has been opened by Chomsky (1995). Therefore, while in some sense the shift of focus from E-language and observational adequacy to I-language and descriptive adequacy marks the birth of generative grammar, the passage from descriptive to explanatory adequacy is reflected by the development of the biolinguistic stage: Indeed in the '60s "the focus of research is set more abstractly still in that we want to our descriptively adequate grammars to follow from descriptively adequate theories of grammar" (Boeckx and Hornstein 2007: 6).

How can the generative program go "beyond explanatory adequacy" (Chomsky 2004)? In the first place, this objective may be intended to concern the development of the minimalist approach (Boeckx and Hornstein 2007: 19), i.e. the attempt to satisfy the already mentioned minimalist criteria in the definition of the generative theory of language. However this is not the only possible direction of inquiry. Longobardi (2003a) proposes to look at two other questions that follow the ones already determined by the three Chomskian levels of adequacy: "Why do we have precisely these actual languages? Why do we have precisely these biologically possible languages?" (Longobardi 2003a: 2). In the scholar's terms, to answer the former question would entail achieving "actual historical adequacy", while the latter question has to do with "evolutionary adequacy": Both the levels, which follow from the satisfaction of the first three, are projected onto the historical dimension of inquiry. The exploitation of the achievements of generative grammar, in particular of the principles and parameters approach, which has been conceived within a synchronic comparative framework, in historical-comparative studies represents the core idea of the definition of the levels of actual historical and evolutionary adequacy. Notice that the distinction defined by the separation of the fourth and the fifth level is exactly that outlined in section 3.1.1 between research aimed at studying on the one hand language evolution and on the other hand the evolution of language faculty. However, in order to appreciate the development of the generative program within the historical perspective, it is necessary to discuss the cognitive-biolinguistic and formal aspects of the field of inquiry: This is done in the following sections.

3.1.2. Universal grammar

As already claimed, the central issue of the biolinguistic period of generative grammar is the problem of language acquisition. Why is the explanation of this process problematic? The question may be introduced taking into account an important observation: The acquisition of first language does not require systematic teaching as that of second languages in adulthood and it depends on contact of the child with a specific linguistic environment, i.e. "language develops spontaneously, by exposure to linguistic input, that is, on the basis of what children hear" (Guasti 2002: 3). Is it plausible to hypothesize that the process of language acquisition is determined exclusively by this input, i.e. by the "primary linguistic data"? The answer developed in the Cognitive period of generative grammar is that the corpus the child is exposed to is not sufficient alone to direct acquisition and that rather humans are born with an innate predisposition for language learning. In this sense, the aim of linguistic inquiry indicated in the '60s is of outlining the "innate linguistic theory that provides the basis for language learning" (Chomsky 1965: 25; quoted in Boeckx and Hornstein 2007: 6).

The innatist hypothesis is grounded on the "poverty-of-stimulus argument". Fundamentally this argument has two premises: First, that all children achieve the same competence in a relatively short period of time, even if they are exposed to different corpora under different circumstances, i.e. that "all speakers of a language know a given fairly abstract property" (Guasti 2002: 5); Second, that the rules acquired by the child cannot be originated by the primary linguistic data alone, i.e. "that this property cannot be induced from the evidence available to children" (Guasti 2002: 5). A closer look at the characterization of the primary linguistic data may help to clarify the second premise. As for positive evidence, the relevant corpus children are exposed to is a finite set of sentences, which as a first approximation one can assume to be well-formed and relatively simple. Is this kind of positive evidence sufficient to make a child converge to the right grammatical rule? It is not, essentially because "though there is decisive linguistic evidence concerning what the correct rule is, there need not be such evidence in the primary linguistic data" (Boeckx and Hornstein 2007: 11; see, for instance, the example reported in 9-11). Moreover, even if one wanted to claim that the primary linguistic data is sufficient to determine the formation of the right rule in that it can contain the relevant pieces of information, given the first premise of the argument he would also have to demonstrate that all children "encounter sufficient examples of the decisive data" and that "they […] do so a sufficient number o times" (Boeckx and Hornstein 2007: 12). This is not the case, as it is demonstrated by empirical quantitative analyses of the primary linguistic data such as that carried out by Legate and Young (2002). What about negative evidence instead, i.e. information about which rules do not apply to the specific language? It has been shown that negative evidence of any kind does not play any role in acquisition. There are two types of negative evidence to consider: The direct one, which includes the explicit corrections made by adults in response to children's mistakes[2], and the indirect one, which is provided by the absence of particular expressions in the primary linguistic data. As for direct negative evidence, it has been observed that on the one hand children do not even make some kinds of mistakes, whose correction might be useful to direct acquisition, on the other hand, when they make mistakes, they ignore the corrections (see among others McNeill 1966). Indirect negative evidence, instead, can hardly be assumed to intervene in the process: If it were, one should assume that children can only reconstruct rules on the basis of the expressions they hear, and that they exclude those entailed by expressions that are not explicitly and robustly included in positive evidence. But this hypothesis contrasts with the observation that in fact positive evidence, as it has been characterized above, appears to be too poor to include all the possible relevant grammatical structures. In other words, if children excluded all the grammatical patterns that they do not have at their disposal in the corpus, language acquisition (as we know it) would become impossible.

Thus the linguistic input available to children consists of a finite set of simple sentences, the positive evidence: It does not include negative evidence and it can vary a lot from case to case. Futhermore, the knowledge of language, or competence, attained by a single speaker is a system of rules that allows him to produce an infinite number of complex linguistic expressions. Moreover, all speakers of a specific language are able to judge whether a sentence is acceptable or not and, despite the different corpora on which they have formed their knowledge, they almost completely converge onto the same grammar. This asymmetry, which is clearly outlined by the poverty-of-stimulus argument, has come to be known as "logical problem of language acquisition" (Baker and McCarthy 1981) or "Plato's problem", after the name of the Greek philosopher, who first took into consideration the question of the gap between knowledge and experience. The fact that the primary linguistic data on which the acquisitional process is based do not reflect the linguistic competence achieved poses fatal problems to all non-innatist models of language acquisition, such as those based on the mechanisms of imitation, reinforcement and association (see the critical review of these models in Guasti 2002: 11-15). The poverty-of-stimulus argument, instead, appears to support both conceptually and empirically[3] the hypothesis that there is some kind of biologically-determined device for language learning, that guides the process and interacts with the environment.

The innatist hypothesis is also in agreement with other important facts. First, it is reported that the development of linguistic competence follows the same path for all children and it is characterized by some typical stages independently of the specific language in question (see later in this section). Moreover, it is documented that not only first language acquisition

---

[2] Another kind of direct negative evidence could be provided by adults correcting themselves when they make the relevant mistakes. Nonetheless the decisive mistakes, whose correction in principle could be useful to direct acquisition, can be constructed as examples by linguists, but they are not attested to in real corpora.

[3] However see Pullum and Scholtz's (2002) criticism to the empirical foundations of the poverty-of-stimulus argument.

occurs rapidly and spontaneously, without open supervision, but also that it has to take place within the first years of life: If a subject is not exposed to the relevant linguistic environment in time, proper acquisition will not take place. The fact that a "critical period" for language acquisition exists is an important cue in favour of the innatist perspective. For instance, Smith (2004) compares language acquisition with the development of the ability of see stereoscopically and that of producing origami. Both these abilities are not owned by the child when it is born, but need to be developed. Nonetheless, while the former is effortlessly acquired by all individuals grown in normal circumstances (specifically, in an environment which allows them to have normal visual stimuli) and can occur only within a critical period, the latter can be consciously learned at any moment in the course of a lifetime. This is because vision is mainly a biologically-determined ability and paper-folding is more a culturally-transmitted capability, even if it requires strongly genetically-based skills (e.g. those concerning object manipulation and coordination). Smith (2004) argues that "[t]he structure of the visual system is genetically determined to a greater extent than origami", but "[i]t is not that paper folding is learned and visual development is innate, both are both, but to different degrees", thus "[l]ikewise in language, we expect to find a balance between genetic and environmental factors in language acquisition" (Smith 2004: 38).
One can say that, according to the generative biolinguistic program, the process of language acquisition more closely resembles the development of 3D vision than the acquisition of paper-folding, because it is largely and fundamentally developed on the basis of an innate predisposition: The "language faculty". In other words, it is recognized that there must be a system responsible for language acquisition, which in some way guides and controls the process providing a specific set of constraints. Thus, exactly these constraints "have historically been termed 'innate disposition', with those underlying language referred to as 'universal grammar'" (Hauser *et alii* 2002: 1577). Intending the language faculty in the narrow sense as "the abstract computational system alone, independent of other systems with which it interacts and interfaces[4]" (Hauser *et alii* 2002: 1571), i.e. as the "narrow syntax", recent research suggests that this faculty is not only innate and common to all humans ("uniformitarian hypothesis", see section 2.1.1), but also species-specific, in that it appears not to be shared by other animals, as shown by the cross-species comparative methodology outlined in Hauser *et alii* (2002).

Having grounded the existence of an innate language acquisition device, what is relevant now is to show how Universal Grammar can be described. Notice that Universal Grammar has to be interpreted as an actual cognitive object, that is present in the speakers' mind since they are born as "language faculty" and that consists of a set of formal linguistic universals. Recall that the following discussion is exclusively focussed on syntax, even if in principle Universal Grammar is also supposed to include relevant rules regarding phonology, morphology and semantics[5].

---

[4] In Hauser *at alii* (2002) the "faculty of language in the broad sense" (FLB) includes the "faculty of language in the narrow sense" (FLN) together with two other cognitive systems: The conceptual-intentional system and the sensory-motor one. According to the authors, the set of syntactic rules that constitute the narrow syntax, which is assumed to be the key component of FLN, map syntactic representations onto the conceptual-intentional interface by means of the semantic system and onto the sensory-motor interface by means of the phonological system. See also the discussion about LF and PF in section 3.1.3.

[5] "Whether these aspects of language (those related to phonology, morphology and semantics, editor's note) are subject to parametric variation in the same way of syntax is an open question; there is some reason that is true of phonology and morphology […], while semantics may not be subject to variation" (Roberts 2007: 12). As for the hypothesized universality of formal semantics see also note 11.

Typological investigations such as those already suggested in Chomsky (1965) have revealed that some syntactic traits seem to be invariant across languages. These characteristics, that appear to "constraint the application of every grammatical operation in every language" (Radford 2002: 11), are called "principles of the Universal Grammar" and they are assumed to be universally valid. In the next section an example of one of these principles, the so-called "Structure dependence principle", is provided: It will serve to introduce the discussion regarding the formal aspects of generative grammar. For the moment, it is important to account for another aspect of the biolinguistic perspective, which deals with the fact that, despite the fact that all humans are supposed to be born equipped with the same language faculty, different languages have different grammars. How can syntactic variation be explained in the biolinguistic perspective? Roberts (2007: 20) suggests that, conceptually, one could make two hypotheses. The first one is that speakers of different languages are genetically different, so that they have different language faculties. But this idea is false, because it disagrees with the observation that a child natively acquires any language is presented to its experience in the first years of life (recall the example of the Japanese child brought up in Italy, section 2.3.1). The alternative hypothesis entails the assumption that, apart from the universal principles, there must be a part of Universal Grammar that interacts with the linguistic environment to form a complete knowledge of language. In other words, this solution leads to hypothesize that "associated with the invariant principles of Universal Grammar there may be certain limited options which remain open, to be 'filled in', as it were, by experience" (Roberts 2007: 21): These are referred to as "parameters". Thus parameters are thought to represent "loci of linguistic variation" (as genes are loci of genetic variation, see section 2.1.3), in that they are expected to define "the exact ways that one speaker's knowledge of language can differ from another speaker's knowledge of language" (Baker 2008: 352). This approach to the description of the Universal Grammar is exactly known as "Principles and Parameters Theory" and derives from the approach originally outlined in Chomsky (1981).

What does this theory entail for the explanation of language acquisition? The guidelines to interpret acquisition of syntax in particular within the parametric framework can be brought back at least to Hyams (1986). Of course the learning task becomes highly simplified for children, who have access only to the superficial order of relevant elements and, in some sense, can recover structure from principles and parameters. Since children have access only to positive evidence, it has been suggested that the learning process is guided by a condition, called "Subset Principle" (Berwick 1985; Manzini and Wexler 1987), which leads them to choose the most restrictive grammar compatible with the primary linguistic data, i.e. not to take into consideration grammars that generate the input they are exposed to plus expressions that would not be generated by the actual grammar that generates the input. In other words, this kind of "conservative learning strategy" prevents acquirers from hypothesising grammatical rules that are not strictly compatible with positive evidence. Furthermore, it is assumed that the pieces of information that determine the choice of the parametric options play the role of "triggers" or "cues" for the parametric values (see the critical review of the different parametric methods provided in Fodor: 2001). Trigger-based approaches aimed at modelling the learning process are adopted for instance by Gibson and Wexler (1994) as well as by Clark (1990; 1992) and Clark and Roberts (1993). In general one may think of a trigger as "a sentence (a word string) of the target language, perceived by the learner, which 'automatically' flips a parameter switch to the correct value" (Sakas and Fodor 2001: 173). More formally, Clark and Roberts (1993) provide the definition of "trigger" in relation to that of "parameter expression" introduced in (1).

(1) "Trigger: A sentence *S* is a trigger for parameter p_j if *S* expresses p_j.
Parameter expression: A sentence *S* expresses a parameter p_j just in case a grammar must have p_j set to a definite value in order to assign a well formed representation to *S*" (Clark and Roberts 1993: 317).

The approaches to language acquisition developed, for instance, in Lightfoot (1989; 1999) and in Dresher and Kaye (1990) and Dresher (1999) are "cue-based" instead. According to these models, cues, i.e. fragments of utterances, are pieces of evidence encountered by children in the course of acquisition. Nothing more is said here about cue-based models, but notice that Dresher (1999) hypothesises that each parameter has a default value, which is maintained until the child identifies the relevant cue and "flips" the parameter switch to the marked value. This idea, as well as the formalism adopted by Dresher (1999) to describe parameters, are taken into account again in the following section, which is dedicated to the formal characterization of parameters. For the moment it is relevant to anticipate that the parametric set is thought to correspond to a finite list of discrete options, in particular to binary choices of the type either/or. This characterization allows to represent even apparently multiple patters of syntactic variations (i.e. types of variations that do not appear to be strictly binary) as the result of the interaction among many binary parameters. Such interaction determines the formalization of the parametric space as a finite "network", in Roberts' (2007; 2008) terms, of connected parameters. Again, a careful account for this idea, which is not immediately understandable without a formal background, is provided in the discussion presented in section 3.1.5.

Following the Principles and Parameters Theory one may shed light on many important issues and reinterpret them in a coherent framework. First of all, as already argued, the parametric approach permits clearly characterizing the process of language acquisition: Within this perspective, for what concerns syntax it is possible to define the acquisitional process as a procedure in which the values of parameters are set one after the other, so that linguists may be able to develop expectations regarding the successive stages attained by the child. In this view, now it is possible to fully appreciate the idea implied by Chomsky's (1986a) quotation provided in section 2.3.2 and reported here for convenience: "[g]iven the appropriate experience (which is culturally transmitted; editor's note), this faculty passes from the state S_0 to some relatively steady state S_s, which then undergoes only peripheral modifications (say, acquiring new vocabulary items)" (Chomsky 1986a: 25). S_0 corresponds to the Universal Grammar and it is entirely genetically-determined, while S_s, which is intended as the mature competence of the speaker, derives from the progressive parameter setting process activated in response to the linguistic environment (at least for what concerns "narrow syntax"). Therefore the intermediate stages of acquisition are expected to reflect the successive levels of advancement of the process, each stage of language acquisition being characterized by the setting of specific parametric values.

The parametric approach also represents a starting point for the explanation of non-random variation of syntax, i.e. of the patterns of variation observed in syntactic typology. As for this, one may think about the implicational universals studied since Greenberg (1963) and generally formalized as "if a language *L* has the property *p*, then it also has the property *q*". These universals show, for instance, that word orders regarding different types of grammatical elements are largely correlated (e.g. a VSO language is also likely to be prepositional and to show the orders noun-genitive "NG" and noun-adjective "NA"). Therefore, in principle, if exhaustively developed, the parametric approach might also allow the prediction of the dimensions of language typology, i.e. to define the boarders of syntactic variation in relation with the limitedness of the space of parametric variation (see the idea of "structural pre-selection" presented in section 2.3.2 and the characterization of the parametric space provided in section 3.1.5). Moreover, since parameters often appear to be related to patterns of variation, they are expected to represent deep generalizations regarding syntactic properties of languages, in the sense that various superficial differences registered by cross-linguistic typological comparison may be related to a unique parametric choice. From this idea the thought follows that not only finite, but also relatively limited[6] sets of parameters might be responsible for a large set of surface syntactic phenomena, which at first sight may often seem to be unrelated. However, the actual extent of the "explanatory scope" of parameters, i.e. of their capability to generate multiple descriptive properties of languages, is currently under discussion in connection with the issue of the distinction between micro and macroparameters (see the end of section 3.1.5).

Finally, "parameters can tell us which aspects of syntax are subject to change in the diachronic dimension" (Roberts 2007: 24). Section 3.2 is intended to introduce the application of the parametric perspective to diachronic syntactic variation on the basis of the formal notions of generative grammar presented in the following discussion.

3.1.3. Grammatical categories, syntactic operations and structure

This section illustrates some formal aspects of the generative approach and it is grounded on a minimalist (Chomsky 1995) account of the so-called "X-bar theory", whose original version has been developed since Chomsky (1970) and Jackendoff (1977). Notice that this part of the work is not meant to provide a complete introduction to syntactic theory, which is covered for instance in Radford (2002), but only to basically and briefly mention some concepts that are fundamental to understand the discussion set up in the rest of the chapter, without any intention to provide a comprehensive account for the different issues.

Before going over the main points, it is necessary to initiate the general formalism of tree-graphs, which are widely used to represent syntactic structures in generative grammar. Some important definitions (e.g. those of "ancestor", "descendent", leaf" and "internal node") are not of much use in this section but are needed to appreciate the discussion of section 4.1.2, which regards the representation of genealogical and non-genealogical relationships between taxonomic units. Therefore for the moment the explanation of these terms is left aside in order to allow space for the information that is more relevant to this section.

As a mathematical object, a "simple tree", like the one in figure 1, is defined as a collection of nodes (in figure 1 they are labelled 1, 2, 3, 4, 5, 6, 7, 8) and dominance relations between these nodes, in which there is a single node, called root (*r*), which dominates all the others (node 1 in figure 1). The lines that connect the nodes of the tree are called branches, or edges, and a node *x* dominates (D) a node *y* when it is possible to read down along branches from *x* to *y* (e.g. 1 D 5). The formal details regarding the interpretation of tree-graphs are provided in (2).

(2) "A simple tree *T* is a pair (*N*, *D*), where *N* is a set whose elements are called nodes and *D* is a binary relation on *N* called dominates, satisfying (a) – (c):

a. D is a reflexive partial order (i.e. reflexive, transitive and antisymmetric, editor's note) relation on *N*.

b. The root condition: There is a node *r* which dominates every node. […] This *r* is provably unique […] and called the *root* of *T*.

c. Chain condition. For all nodes *x*, *y*, *z*, if *x D z* and *y D z*, then either *x D y* or *y D x*." (Keenan 2007: 51).

---

[6] The issue of the quantification of the parametric set is taken into account again in section 5.2.1 within the discussion regarding the approach of the "Modularized Global Parameterization" (Longobardi 2003a).

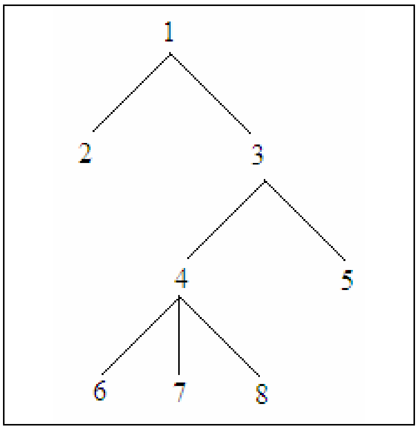


Figure 1: A simple (rooted) tree.

A node x is said to “strictly dominate” (*SD*) a node y if $x \neq y$ (e.g. 4 SD 7) and x is said to “immediately dominate” (*ID*) y if x *SD* y and there is no node z such that x *SD* z and z *SD* y (e.g. 3 *ID* 5). x is a “sister” of y (and y is a sister of x) if $x \neq y$ and $|\{ \exists z \in N \mid z\ ID\ x\ \&\ z\ ID\ y \}|$. Furthermore, x is said to “c-command” (*CC*) y if x and y are independent, i.e. neither dominates the other, and every branching node that strictly dominates x also dominates y. Finally, x is a “n-ary” branching node if and only if $|\{ y \in N \mid x\ ID\ y\}| = n$ (e.g. 1 is binary branching, 4 is ternary branching) and T is called n-ary branching if all nodes except for the leaves are n-ary branching (Keenan 2007: 55). It will soon become clear how these notions, in particular the relation of c-command, apply to the understanding of the representation of syntactic structures.

A central principle of Universal Grammar is the “Structure dependence principle”, according to which “[a]ll grammatical operations are structural dependent” (Radford 2002: 12). From this principle also follows that all grammatical operations are category-based, i.e. “apply to whole categories of words or phrases rather than to individual expressions” (Radford 2002: 29). The Structure dependence principle is reconsidered below in this section, after having introduced the complete framework. The first questions to take into account now are: What does “grammatical category” refer to? What is meant to be “structure”? A grammatical category is a group of elements that have a common set of grammatical properties, either morphological (e.g. the property of having a specific suffix, like the perfect suffix *–ed* of verbs in English) or syntactic (e.g. the property of occurring in certain positions of the sentence with respect to other elements, like the prenominal position of adjectives in English) or both. There are two types of categories: Lexical and functional. Fundamentally, while a lexical category includes words that have a lexical content, e.g. nouns (N), verbs (V) and adjectives (A), a functional category consists of words that lack it and have only a grammatical function, e.g. determiners (D), inflections (I) and complementizers (C). The structure derives from the combination of different elements into phrases or sentences by means of different “syntactic operations”. All syntactic operations assumed by generative grammar to explain the derivation of structure work on the basis of discrete elements (e.g. words and phrases) and are recursive, in the sense that under the right conditions they “can apply to their own output, thus in principle creating infinite structures from a finite set of symbols and types of operations” (Roberts 2007: 4).

The basic operation that allows to combine elements is “Merge”, which “takes a pair of syntactic objects (SO_i, SO_j) and replaces them by a new combined syntactic object SO_{ij}” (Chomsky 1995: 226). A syntactic object may be referred to as a “constituent”, defined in general as “an expression which is one of the components out of which a phrase or a sentence is built up” (Radford 2002: 256). Notice that the categorical label of the syntactic object produced by Merge is determined by that of one of the objects that are combined. For instance, if one has the noun (N) “book” and merges it with the determiner “the”, the result is the determiner phrase (DP) “the book”, which then can be merged with the verb (V) “read” to form the verbal phrase (VP) “read the book”. The structure outlined by this double application of Merge can be represented in the labelled bracketing form as in (3).

(3) $[_{\text{VP}}\ [_{\text{V}}\ \text{read}]\ [_{\text{DP}}\ [\ _{\text{D}}\ \text{the}]\ _{[\text{N}}\ \text{book}]]]$

Alternatively, a labelled tree-graph in figure 2a shows the same structure: As usual in the current theoretical framework, the syntactic tree is binary branching. The “head” of a phrase is the element that determines its categorical classification: In the example, the head of VP is the verb “read”, while the head of DP is the determiner “the”. These heads are said to “project” their phrase, which, in the above example, consists of the head itself (V and D) and a complement (respectively DP and N), the head and its complement being sister nodes in the tree. The minimal projection corresponds to the head itself, while the maximal projection (XP, using “X” as a generic categorical label) is a constituent that cannot be included within any larger constituent of the same head and constitutes a complete phrase. The term “intermediate projection” (X’)[7], instead, indicates a constituent that is neither minimal nor maximal, so that it is necessarily incomplete. For instance, “reads the book” is an intermediate projection of the head “reads” (it is a V’) and it is ungrammatical in English because it lacks a subject. Inserting the subject “he”, one obtains the grammatical sentence “he reads the book”, that, basically, can be interpreted as a complete VP. Thus the pronominal determiner “he” is called “specifier” of the head and it allows to complete the maximal projection: The resulting labelled bracketing form is given in (4) and it is represented as a tree graph in figure 2b.

(4) $[_{\text{VP}}\ [_{\text{D}}\ \text{He}]\ [_{\text{V'}}\ [_{\text{V}}\ \text{reads}]\ [_{\text{DP}}\ [\ _{\text{D}}\ \text{the}]\ _{[\text{N}}\ \text{book}]]]]$

Therefore a complement (Compl) of a head is a grammatical object that joins with a head to project it into a larger structure of the same category, while a specifier (Spec) corresponds to “the grammatical function fulfilled by certain types of constituent which (in English) precede the head of their containing phrase” (Radford 2002: 271). The specifier and the intermediate projection are sister nodes. The structural roles of complement and specifier also require to be taken into account within the semantic structure of sentences. In order to understand how, it is necessary to introduce more terminology originally borrowed from propositional logic. Expressions denoting activities or events, prototypically verbs, may be referred to as “predicates”: A specific set of “arguments” is associated with a predicate, each argument corresponding to a participant to the activity or the event. Abstracting away from any detailed account of the case theory and of the theory of thematic-roles (originally outlined in Chomsky 1981; 1986a), here it suffices to say that the complement of a predicate expression (e.g. a V-head), i.e. the “object”, may be referred to as “internal argument”, because its position is inside the intermediate projection of the head (e.g. V’), while the specifier, i.e. the “subject”, may be called “external argument” for the opposite reason. In this sense, both the positions of the object and of the subject are “argumental positions”.

In addition to Merge, another fundamental operation is “Agree” (or “Attract”), which is “the operation that manipulates combinations, by establishing a relation between lexical items within a syntactic space” (Roberts and Roussou 2003: 17). Referring back to the example “he reads the book”, one can see that Agree determines the agreement between the subject “he” and the verb “read”, which comes to end in the third-person suffix *–s*.

---

[7] The intermediate projection X’ may also be indicated by the label $\overline{X}$, from which the X-bar theory derives its name.

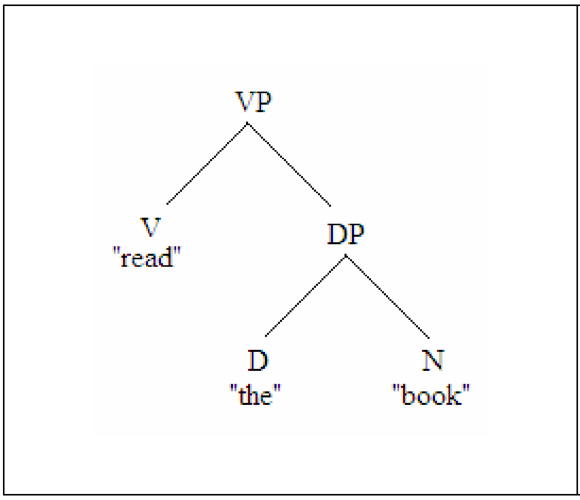


Figure 2a: Structure of “read the book”.

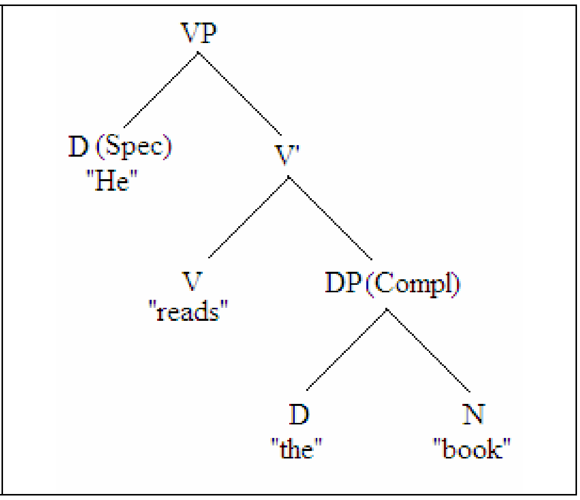


Figure 2b: Structure of “He reads the book”.

The derivation of structure may be determined also by “Move”. This operation is responsible of the movement of a word or a phrase from one structural position to another. For instance, consider the sentence “He is reading a book”: The position of the auxiliary “is” is the head of inflection (IP) and the subject appears to precede it. Nonetheless the subject “he” is assumed to have been originated as specifier of VP and then to have moved, or “raised”, to Spec IP, as illustrated in figure 3a. In the position left empty by the subject a trace remains, which is a “covert duplicate”, i.e. a phonologically silent copy, of the element that underwent movement: The trace is indicated with *t* plus an index, e.g. t_i, that indicates the relation with the moved element. What relation is this? Using terminology derived from the “Binding Theory” (designed since Chomsky 1981), a trace is said to be “bound” by its antecedent (the element that has been moved), which means that the antecedent determines the semantic and syntactic properties of its trace. In general, a constituent *x* is said to bind a constituent *y* if *x* and *y* refer to the same object/entity/state of affairs of the world and *x* c-commands *y*.

Operations that involve the movement of a phrase from subject or complement position to a higher subject position are called “Argument-movements”, or “A-movements”: This class comprises the already mentioned subject-raising, as well as, for instance, passivization. In contrast, “A-bar-movements” (or “Ā-movements”) refer to cases in which the movement targets a non-argument position, such as the case, for instance, of operator movements (e.g. wh-movement, which is considered in the next section) and scrambling (for further details see Radford 2002: 130-150). When the movement just affects a single head-word, it is termed “head-movement”. Consider for instance the interrogative sentence “Is he reading the book?”, which is derived from the affirmative “He is reading the book”. In this case the auxiliary “is” moves from inflection (head of IP) to complementizer (head of CP) to form the question. Figure 3b shows the I-to-C head movement in action. Some constraints apply to movements: “These include locality, structure preservation, and the requirement that movement leaves a well-formed trace” (Roberts 2001: 113). According to the locality condition, movement cannot occur across more than one intervening phrasal boundary. For instance, a head can move only to the directly higher head position: The operation can be iterated, in which case the movement is said to be “cyclic” and the head can progressively reach positions higher up the structure, but if a possible target head position is already filled by another element the operation is blocked (“Head Movement Constraint”, Travis 1984). From the structure preservation condition, instead, follows that, for instance, head movement can target only a head position. Finally, since, as already pointed out, the movement is said to leave a well-formed trace if and only if the trace results to be c-commanded by its antecedent, the

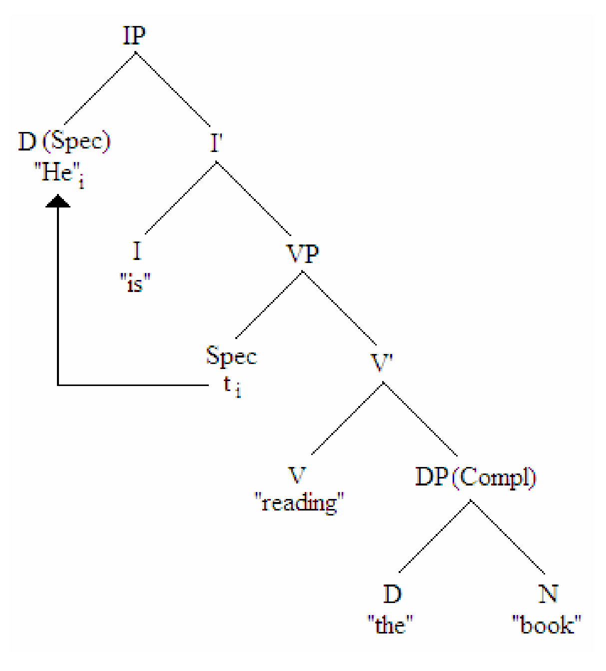


Figure 3a: Structure of “He is reading the book”.

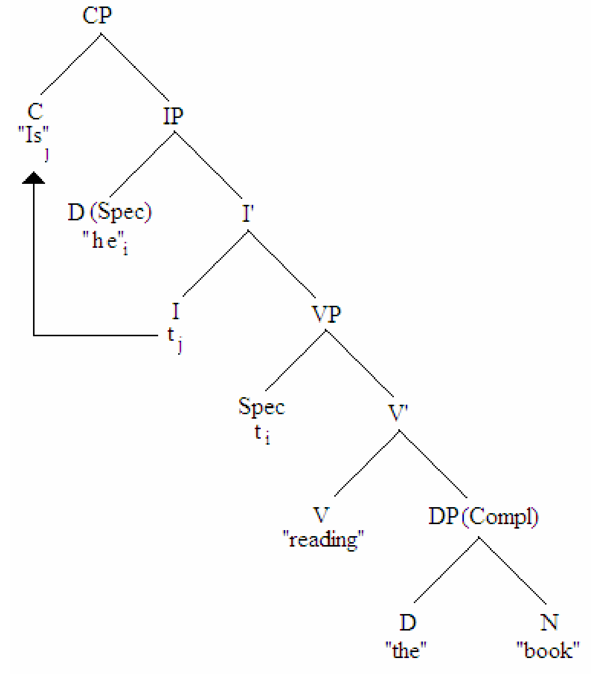


Figure 3b: Structure of “Is he reading the book?”.

movement has “to take place in an upward direction, where ‘upward’ is understood as defined by c-command” (Roberts 2001: 113).

Following Chomsky (1995; 2000), it is assumed that Merge and Agree are responsible for the connection of syntax with the “Phonological Form” (PF) and the “Logical Form” (LF): The former corresponds to the representation of sound, i.e. of how a certain linguistic expression is pronounced, and it is connected with the Articulatory-Perceptual system, or Sensory-Motor system in Hauser *et alii*’s (2002) terms, the latter includes information about its meaning and it is related with the Conceptual-Intentional system. Each syntactic object may or may not be mapped onto PF and LF, depending on the information provided in the lexicon[8]: If it is, it is said to be “interpretable” at the relevant form of representation (either PF or LF or both), otherwise it is considered “uninterpretable”. This means that in principle an object may be (+p/l), i.e. interpretable at PF/LF, or (-p/l), i.e. uninterpretable at PF/LF. For instance, “book” is interpretable at both levels (it is +p and +l) and corresponds to /buk/ at PF and to the object that it denotes at LF. Focussing on the interpretability at LF, notice that the process of mapping is assumed to work in accordance to Chomsky’s (1995) “checking theory”, which can be briefly summarised as follows. Each word comes with a set of features, which may have a semantic content (semantic features, e.g. person, number and gender for nouns) or lack it (e.g. case-features and complement/specifier-features). In the course of the derivation features are “checked” one with the other and non-semantic features are erased after having being checked: For instance, the specifier-feature of a verbal head (“reads”) is checked against the head-feature of a pronominal determiner (“he”) and then they are both erased. The point is that only semantic features are interpretable at LF, while non-semantic features are not: If after a derivation a structure possesses only semantic features, it is well-formed and it is said to “converge”, if uninterpretable features are still present it “crashes” (“principle of full interpretation”). Summing up, the principle of full interpretation applied to LF “entails that

---

[8] The key concepts of the mechanism of interpretation are explained for instance in Radford (2002: 69-74).

every element of the representation have a (language-independent) representation" (Chomsky 1995: 27).
The meaning and value of the Structure dependence principle should appear clearer now: The core idea implied by this principle is that syntactic operations can apply to syntactic objects with exclusive reference to the position and the role that these objects play within the syntactic structure they belong to. No other criterion, for instance based on the linear order of words in sentences, is relevant for syntactic operations. In other words, "syntactic rules cannot refer either to the number of words in a sentence or to the position of a word in a word-string" (Moro 2006: 197; editor's translation). Such a principle alone largely restricts the possibility of syntactic variation, given that the description of structure and of the mechanisms of its derivation, as they have been outlined in this section, are assumed to be universally valid in the generative perspective. It is not among the goals of this work to introduce other principles of the Universal Grammar, but rather it is important to focus on the parametric boundaries of syntactic variation. The next section is intended to introduce some hypotheses regarding the general form of parameters.

3.1.4. Parameter formats

Consider a concrete example, the so-called "wh-movement parameter" (or wh-parameter, Huang 1982). In certain languages the wh-movement occurs in direct wh-questions (i.e. those introduced by wh-words, e.g. "what" and "where") and it determines wh-expressions (i.e. expressions containing wh-words), which are originated as complements of V, to move to the Spec-CP position. For instance, the wh-movement is assumed to intervene in the derivation of the English direct interrogative sentences in (5a) and (5b)[9].

(5) a. [What]$_i$ do you want t$_i$?
b. [Where]$_i$ are you going t$_i$?

Notice that, for instance, wh-movement takes places in Italian direct interrogative sentences too, as shown in the examples in (6a) and (6b).

(6) a. [Che cosa]$_i$ vuoi t$_i$?
b. [Dove]$_i$ stai andando t$_i$?

The operation does not apply in the case of Mandarin Chinese instead: This is clear after the observation of sentences like (7a) and (7b).

(7) a. Nĭ yào [shénme]?
You want what
b. Nĭ qù [năr]?
You go where?
(Po-ching and Rimmington 2004: 51-52)

Cross-linguistic comparison shows that either languages are like English and Italian, which means that they have wh-movement (e.g. Spanish, German and Welsh), or they are like Mandarin Chinese, so that they do not allow wh-movement and the wh-expression remains *in-situ* (e.g. Japanese, Thai, Korean, Turkish) (Roberts 2007: 83). The variation can be brought back to the effect of a specific parameter, exactly the wh-parameter, which, as expected,

---

[9] In the examples included in (5)-(7) only the position of the wh-expression and that of its trace are shown, while all other details of the structural representation of the sentences are left out.

entails a binary option: Does the wh-expression move to Spec-CP in interrogative sentences or not? Alternatively, does the wh-movement occur overtly (before the spell-out) or covertly (after the spell-out)? The parametric choice applies systematically to all interrogative sentences, independently of the verbs and of the wh-phrases involved: "[T]here is no language in which the counterpart of 'who?' undergoes wh-movement but not the counterpart of 'what?'" and "there is no language in which wh-complements of some verbs can undergo fronting (i.e. wh-movement, editor's note), but not wh-complements of other verbs" (Radford 2002: 18). These claims, supported by the examples in (5)-(7), are in perfect accordance with the assumption involved by the Structure dependence principle, according to which syntactic operation are category based. Moreover, the same pattern of variation recorded in direct interrogative sentences is found in indirect ones: In (8) the relevant examples in English (8a), Italian (8b) and Chinese (8c) are given.

(8) a. I wonder [what]$_i$ Lisi bought t$_i$.
b. Mi chiedo [che cosa]$_i$ comprò Lisi t$_i$.
c. Wo xiang-zhidao Lisi mai-le [shenme].
I wonder Lisi buy-ASP[10] what.
(Huang 1982: 188)

Thus the value of the wh-parameter licenses or not the existence of a category-based operation (the wh-movement) in two distinct syntactic contexts: Direct and indirect interrogative sentences.

On the basis of the observations put forward so far, is it possible to account for the syntactic variation determined by the wh-parameter more formally? The starting point in accomplishing this is to define a generally valid form for parameters. To this end an important premise is that, at least since Borer (1984) and Manzini and Wexler (1987), it is recognised that parameters are associated with lexical items, so that they are reduced "to the one part of a language which clearly must be learned anyway: the lexicon" (Borer 1984: 29). More specifically, in current literature the lexical items, which are subject to parameterization, are assumed to be those belonging to functional categories (see for instance Chomsky 1995: 6, 55). Baker (2008) calls this idea "Borer-Chomsky conjecture" and defines it as in (9).

(9) "Borer-Chomsky conjecture
All parameters of variation are attributable to differences in the features of particular items (e.g., the functional heads) in the lexicon" (Borer 2008: 353)[11].

---

[10] In general terms, "le" indicates perfect verbal aspect in Mandarin Chinese.

[11] The "Borer-Chomsky conjecture" is one of the arguments put forward by Boeckx (2008) to support the so-called Strong Uniformity Thesis (SUT), according to which adopting a strict minimalist approach one could not assume that narrow syntax is subject to parametric variation and that it is "affected by lexical parameters". In other words, SUT would entail that narrow syntax should be regarded as fully uniform, i.e. fully identical in all possible human languages. The "Borer-Chomsky conjecture" would lead to sustain a SUT-like approach because "once parameters are confined to the lexicon, and much of the lexicon is relegated to post-syntactic components, there is little chance left for (narrow) syntax to be subject to parametric variation" (Boeckx 2008: 6). The scholar considers other two arguments in favour of the SUT hypothesis: The first one is related to the recent idea of connecting parameters to lexically invariant "edge features" (starting from Chomsky 2004), the second one deals with the asymmetry between LF and PF, where only the latter may be surely assumed to vary among languages, but not the former, and currently narrow syntax tends to be viewed as optimized more for the conceptual-intentional system than for the sensory-motor one (Chomsky 2005: 3). However, the scholar specifies that all these arguments do not represent decisive proof in support

The general format for describing parameters proposed by Roberts and Roussou (2003) and Roberts (2007) follows from specific hypotheses regarding the interpretation of functional categories and from the notion of "markedness"[12]. As for the former issue, the question posed by Roberts and Roussou (2003) is: Do functional categories have a proper syntactic representation, that allows them to be subject to Merge and Agree and to be interpreted as proper heads of projections? Given the premise that parameters are expressed in functional categories, this question is fundamental. Robert and Roussou (2003) demonstrate that, even if functional categories are subject to cross-linguistic variation in their realization, they can project their properties just like lexical categories, but differ from them "in being a closed class and in being inherently 'defective' in some way"[13] (Robert and Roussou 2003: 23). In the scholars' perspective, categories that have N and V features are always +l and +p, while functional categories are always +l, but they show cross-linguistic variation for what concerns the mapping onto PF. In other words, it is assumed that the same set of functional features is present in all languages, therefore it is universally mapped onto LF[14], but each language may or may not provide specific functional features with a phonetic representation. The resulting universal characterization of the lexicon is given in (10).

(10) "a. Lexical items, specified as ±V, ±N with PF and LF properties given
b. Substantive universals encoded as interpretable features of functional heads
c * assigned in a language-particular fashion to (b) (the asterisk indicates that the relevant functional head is interpretable at PF, editor's note)" (Roberts and Roussou 2003: 29).

According to Roberts and Roussou (2003), there are two operations by means of which functional feature may be mapped onto PF and become *F: Either Merge or Move. From this claim follows that a minimal system of parametric variation should be formulated in terms of two choices, introduced in (11).

(11) "a. F*? Yes/No
b. if F*, is it satisfied by Move or Merge?" (Roberts and Roussou 2003: 30)

Apart from the restrictions determined by the lexicon of each language, in general Merge should be "preferred", because it is more "economical" than Move according to the theory of markedness assumed by Roberts and Roussou (2003). The scholars suggest that markedness may be related to complexity, in the sense that the structures that are more complex should be considered more marked and, adopting Longobardi's (2001a: 94) idea, they assume that complex structures contain more feature syncretism, which means that "[g]iven two structural

---

of SUT, but rather tentative indications aimed at sketching a possible new line of inquiry, which is not clearly outlined. Therefore this approach is not taken into further account here.

[12] The proper introduction of the concept of "markedness" in linguistic theories would require a discussion on its own, which is beyond the aims of the present work. Here it suffices to say that the first use of this term refers back at least to the phonological paradigm of the Prague School and to Jakobson (1941). See Battistella (1996) for an introduction of the history of the concept.

[13] Something more about the "defectiveness" of functional elements is said in section 3.2.1 introducing grammaticalization.

[14] "There is, however, another possible alternative that the authors do not discuss: it is conceivable that some interpretable features may be lacking completely from LF in (certain constructions of) certain languages (e.g. definiteness in Latin and perhaps Russian DPs), not just from PF. This line of approach would be quite naturally compatible with the frequent observation that speakers of languages in which a feature is never formally represented rarely master the use of such a feature in other languages, certainly not simply the consequence of difficulties with the phonological representation" (Longobardi and Rigon 2008: 429).

representation R and R' for a substring of input text S, R is simpler than R' if R contains fewer formal features than R'" (Roberts 2007: 235). A syntactic object that undergoes movement must have both a feature that allows it to merge in its original position and another feature that determines movement, plus it must be associated with a phonological representation, which consists of features as well. An element that is not moved, instead, lacks the feature that triggers movement, thus it may be considered "simpler". Moreover, the simplest option is the case in which a feature does not have a phonological representation at all. From this line of reasoning Roberts and Roussou (2003) determine a markedness hierarchy whose simplified formulation is provided in (12).

(12) "$F^*_{Move} > F^*_{Merge} > F$"[15] (adapted from Roberts and Roussou 2003: 210).

This approach to markedness relates to parametric variation because it permits the evaluation of the values of each parameter and allows the determination of which one constitutes a marked choice[16]. Consequently, the unmarked value may be thought to represent the default choice of the learner, i.e., as a first approximation, the parametric setting resulting from the absence of a relevant triggering experience in the primary linguistic data. Fundamentally this view is in agreement with the basic assumption of Dresher's (1999) model (see section 3.1.2). As a consequence, the general format for describing parameters proposed in Roberts (2007: 269) and presented in (13) is specifically adapted from that proposed by Dresher (1999):

(13) "a. *Parameter*: A (functional) head H {has/does not have} feature F (in a given formal relation).
b. *Default*: F is absent.
c. *Cue/expression*: properties of inflectional morphology and linear order of elements" (Roberts 2007: 269).

Summing up, according to this format it is assumed that, given a binary parameter, one of its two values represents the default options and it is unmarked, in the sense that it involves a smaller number of features with respect to the marked one, which need to be triggered by a relevant piece of positive evidence in the course of acquisition.
Thus, coming back to the definition of the wh-parameters, Roberts (2007) reformulates it as in (14):

(14) "a. *Parameter*: [+wh] C {has/does not have} an EPP feature triggering movement of a wh-phrase to its specifier.
b. *Default*: EPP is absent.
c. *Cue/expression*: "displaced" wh-phrase, wh-marking on D" (Roberts 2007: 272).

Following Chomsky (2000; 2001), in Roberts (2007: 192) "EPP feature" means "Extended projection Principle feature" and it is intended as the specific feature owned by a syntactic object, the "Goal". This feature allows the Goal to "attract" another syntactic object, the "Probe", which, in order the movement to take place in accordance with the requirements of the checking theory, must also have the relevant uninterpretable morphosyntactic features. Nothing more is said here about this point and the older interpretation of the EPP (originally conceived in Chomsky 1982). What is relevant instead is to observe that in Robert's (2007) formalization the target position of the wh-movement ([+wh] C) is assumed to lack the EPP

---

[15] ">" means "is more marked than".

[16] Other approaches to markedness appear to be not of much use for determining a default assignment to parameters, as shown, for instance, from the empirical evaluation of a theory of markedness derived from the hypothesis of the Subset principle (Roberts 2007: 256-261).

feature in the default setting of the parameter and that in the course of the acquisition the child must be exposed to specific positive evidence (consisting in the wh-marking of the relevant elements and their fronting) to be able to set the wh-parameter like in English and Italian. From this follows that, conversely, the unmarked parametric value should be that found in Mandarin Chinese-like languages.

An alternative but closely similar set of abstract formats for parameters is that proposed by Longobardi (2005): The scholars calls them "parameter schemata". These formats are derived from the cross-linguistic empirical study of the DP-internal structure that allowed the collection of the parametric data, whose current version is used in the quantitative experimentation introduced in the present work (see chapters 5 and 6). In particular, Longobardi (2005) claims that the vast majority of the parameters included in this dataset, referred to as "Table A" (Gianollo *et alii* 2004), falls into one of the four schemata included in (15)[17].

(15) "a. Is F, a functional feature, grammaticalized?
b. Is F, F a grammaticalized feature, checked by X, X a lexical category?
c. Is F, F a grammaticalized feature, spread on Y, Y a lexical category?
d. Is F, F a grammaticalized feature checked by X, strong (i.e. overtly attracts X)?" (Longobardi 2005: 410).

In (15) "grammaticalized" indicates a situation in which "the (functional, editor's note) feature must obligatorily occur and be valued in a certain structure" (Longobardi 2005: 410). For instance, according to the relevant *a*-schema parameter, definiteness may or may not be obligatorily expressed on the DP of a language by means of a morphological marking: Only if the case is the former, as for instance in English, definiteness has to be accounted for in checking relations, thus it is said to be grammaticalized. In a language like Basque definiteness is instead not grammaticalized, therefore in this case the parameter is assumed to be set to the opposite value respect to English. Furthermore, a *b*-scheme parameter is relevant to determine whether a grammaticalized feature is checked by elements belonging to a specific lexical category X. Therefore, assuming that in a certain language, say again English, definiteness is grammaticalized, a relevant parameter of type *b* may ask whether this feature is marked on demonstratives (as it is in English) or not (as in Arabic). The schema in *c*, instead, refers to parameters that determine "if a feature which is interpreted in a certain structural position has uninterpretable occurrences, depending in value on it, on other categories" (Longobardi 2005: 411). Taking Y to be adjectives, this means that a specific parameter may exist to ask whether definiteness spreads on adjectives, which happens in Greek, but not in English. Finally, the *d*-schema indicates the possibility that a certain grammaticalized feature is able to overtly attract the lexical category X, i.e. to trigger its movement before the spell-out. Taking again definiteness as F and demonstratives as X, the resulting *d*-form parameter outlines a distinction between Bulgarian- and Norwegian-like languages, because in the former demonstratives are overtly attracted, while in the latter they are not. Further and more detailed explanations regarding the parameters of Tale A that fall into the schemata are provided in chapter 5 together with relevant examples.

The possibility of identifying such abstract general formats for parameters paves the way to achieve a deeply minimalist account for the process of language acquisition. For instance, Longobardi (2005) argues that in light of his model, it suffices to hypothesize that the initial

---

[17] Longobardi (2005: 411) mentions the possibility of adding a fifth parameter schema, which could be relevant in non-nominal syntactic domains and could be related to the phonetically null realization of certain features or categories, e.g. *pro*.

state of mind of the child (S_0) is provided with a narrow set of schemata, instead of a wide inventory of specific parameters: These schemata could then generate all observed actual parameters by means of the interaction with the functional features and the lexical items found by the child in the primary linguistic data. Therefore, one could reduce the principles and parameters model to a minimalist principles and schemata hypothesis designed as in (16).

(16) "Principles&Schemata model: UG = principles and parameter schemata. Parameter schemata at S_0, closed parameters at S_S" (Longobardi 2005: 412).

Even if not fully developed, a similar perspective may be applied to Roberts and Roussou's (2003) formalism as well. In any case, except for this advance, do the two models have some points in common? One may say that a grammaticalized feature, as it is intended in Longobardi's (2005) schemata, corresponds to a feature that has a phonological representation in Roberts and Roussou's (2003) basic formalism reported in (12), thus to F*. Moreover, while in Roberts and Roussou's (2003) model the parameterization only concerns the choice between Merge and Move in the actual realization of the feature, Longobardi (2005) distinguishes between three possibilities: The grammaticalized feature may be just checked by a lexical category X that is merged to the DP, it may spread to another category Y and, if checked by X, it may attract X and consequently entail movement. Thus the account provided by the parameter schemata appears to cover in detail a wider range of possible syntactic properties and phenomena than Roberts and Roussou's (2003) formalism. Furthermore, no theory of markedness is entailed by Longobardi's (2005) proposal: This choice may be thought to derive from the fact that, although the theoretically foundations of the markedness approach may be considered convincing and the theory may be shown to correctly account for specific cases, in general it is difficult and uncertain to apply it to an extensive empirically-based parameterization as that pursued in Table A. Finally, it is important to underline a key similarity between the models: In both of them the actual parameters derivable from the abstract formats appear to be related one to the other, i.e. they interact in some way. The next section is dedicated in covering this issue.

3.1.5. Interaction between parameters

The main form of interaction that relates parameters one to the other is their implicational ordering. This structural characterization of the parametric space is immediately understandable if one considers a common property of the models outlined in the previous section. As for Roberts and Roussou (2003), parameters governing the phonological realization of features condition the application of parameters that determine their instantiation (by means of Merge of Move), in Longobardi's (2005) approach the setting of an *a*-schema parameter allows or forbids *b*- and *d*-schema parameters to be relevant in a language, because these three types of parameters are significant only if the functional feature is grammaticalized. Moreover, parameters of type *d* are assumed to depend on *b*-schema ones, in that they require that F is checked by the lexical category X. Therefore, one may project the parameters generated by both abstract formalisms onto specific implicational hierarchies, as those represented as tree-graphs in figure 4 and figure 5. In these tree-graphs each parameter type is included within a type-node (e.g. that labelled with F*? in figure 4) and each value of the parameter type corresponds to a value-node (i.e. Yes or No). Therefore the existence of implications between parameters outlines a set of relationships, whose simplest instantiation may be characterized as in (17).

(17) A parameter p_2 is implied by the value v_i of a parameter p_1 in the case in which if p_1 has v_i, then p_2 may be set either to v_i or to v_j, otherwise p_2 cannot be set.

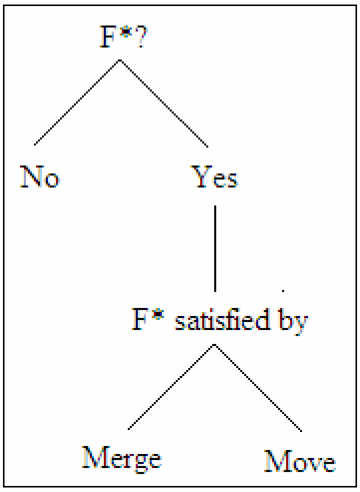


Figure 4: Implicational hierarchy entailed by Roberts and Roussou’s (2003) parametric model.

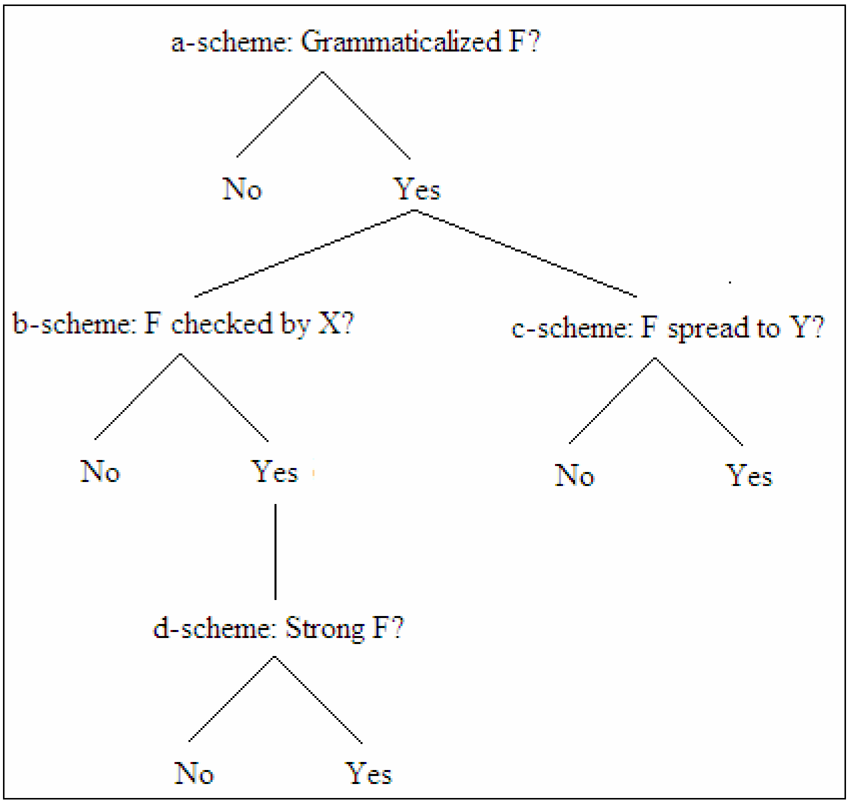


Figure 5: Implicational hierarchy entailed by Longobardi’s (2005) parameter schemata.

For instance, assuming Longobardi’s (2005) model, p_1 may correspond to the *a*-scheme parameter which asks whether definiteness is grammaticalized in a certain language: If the value of p_1 is “Yes” (or “+”, anticipating the coding adopted in Table A, see chapter 5), then the *b*-scheme p_2 “Is definiteness checked by demonstratives?” is relevant in the language and may be set either to “Yes” or to “No” (i.e. either to “+” or to “-”), if not p_2 is not applicable and cannot be set to any value (i.e. it has a 0-value, following the coding of Table A). According to Longobardi (2005), it is imaginable that implied parameters, which do not have their condition satisfied and appear to be not set in a certain I-language, “have actually never been present at any state in the mind attaining that I-language” (Longobardi 2005: 412). An alternative view on this matter is put forward by Roberts (2007), who hypothesizes that if a certain implicational condition is not satisfied, then “all the subordinate parameters (depending on that condition, editor’s note) automatically take on the default value” (Roberts 2007: 280). Of course this idea works only if one adopts a theory of markedness, such as that

previously discussed, that permits determining which parametric values are given by default and which ones are instead marked: Again, this is not the case for the interpretation of the parameters included in Table A.
The empirical investigation put forward at least since Gianollo *et alii* (2004) shows that actual implications between parameters may be more complex than those derived by (17). However, any further discussion about this point related to the data of Table A as well as its consequences on syntactic change in general and on the quantitative comparison between languages in particular, is left behind to be introduced starting from section 5.2.3. What is fundamental to highlight here is that the interactions between parameters play a fundamental role in language acquisition, because the existence of a hierarchical ordering of parameters entails that a parameter p_1 should be acquired before its subordinate parameter p_2. This ordering is outlined also by Baker (2001), who describes it introducing the concept of "ranking" and the formalization given in (18).

(18) "Parameter X ranks higher than parameter Y if Y produces a difference in one type of language defined by X, but not in the other" (Baker 2001: 163).

The scholar develops his approach providing a concrete example of ordered parameters, a real "Parameter Hierarchy" (PH, Baker 2001: 183; 2003: 352), whose punctual description is not covered here. Boeckx (2008: 4) appears to be doubtful about the validity of implicational orderings of parameters like PH, which he refers to as "clusters of parameters", and raises a question that remains open: "Where does the hierarchy come from? That is, What are the design principles that would make this specific hierarchy emerge?" (Boeckx 2008: 4). A tentative answer could be that the parameter space is designed in this way in function of a general economy principle that applies to the acquisitional process: The existence of implicational relationships between parameters highly constraints the possible grammars, whose number is dramatically decreased, thus the search for the right grammar pursued by the child comes to be simplified. However, the PH implementation of the ranking structure appears to be falsified by empirical data in some cases, in the sense that it does not correctly predict the existence of some language types, which are actually attested. Reviewing the empirical validity of PH, Newmeyer (2005: 85-87) discovers many problematic cases: For instance, according to PH "only a subset of SVO languages are permitted to have a positive value for the Null Subject Parameter (Chomsky 1981; Rizzi 1982), even though null subject languages can be SOV (Turkish) and VSO (Irish)" (Newmeyer 2005: 87). Nonetheless, the existence of these divergences between the theoretical prediction derived from the structure of PH and the typological data does not mean that the hierarchy approach to parametric variation is inapplicable in general and that one should necessarily assume that "[n]o hierarchy of the general form of the PH is capable of representing the parametric choices that the child is hypothesized to make" (Newmeyer 2005: 87). Roberts and Holmberg (2005) reply to Newmeyer's (2005) critique arguing that even if the details of the implementation of an ordering like PH need to be revised, the concept of parameter hierarchy deserves to be refined and developed. On the other hand, the empirical validity of this perspective appears to be strongly supported by the data of Table A, as it is argued in section 5.2.3.

A further development of this line of inquiry is related to the distinction between micro and macro-parameters, which may be brought back to the distinction between micro and macrocomparison. It is problematical to give a formal account for these notions, but in general "microcomparative syntax can be thought of as comparative syntax work done on a set of very closely related languages or dialects" Kayne (2005: 6), which thus largely show identical syntactic patterns and just small points of variation, while conversely macrocomparison should be aimed at identifying large-scale clusters of divergent properties between languages that belong to different groups or families. The separation of these two domains of study is ultimately problematic as that between "language" and "dialect" and for the moment can only rely on the pre-scientific idea of "similarity" and "diversity" (Chomsky 1986a: 15-19), but it can be of some use when, as a first approximation, it is projected onto the parametric framework. In particular, Kayne (2005) argues that "apparently microparametric differences might all turn out to dissolve into arrays of microparametric ones (i.e. into differences produced by the additive effects of some number of microparameters)" (Kayne 2005: 8). Following this view, Baker (2008: 354-355) sketches a parallel with biological evolution: He claims that the accumulation of microparametric differences between closely related languages resembles the process of gradual divergence between populations assumed in the standard paradigm of the modern synthesis, while abrupt events of macroparametric deviation may be regarded as analog to the process described by Eldredge and Gould's (1972; 1988) punctuated equilibrium model (see section 2.1.4)[18]. Of course, as argued throughout section 2.2 and 2.3, such analogy can only be considered useful on a purely descriptive level of analysis, but it has no deep explanatory pretension with respect to the actual dynamics of syntactic evolution. Roberts (2008) adopts the micro/macroparametric distinction in relation to the implicational ordering of parameters: He shows that the positions

---

[18] Referring to macroparametric variation, Baker (2008) says that, "[f]or example, there might be a single parameter in the statement of Merge that induces the core difference between head-initial and head-final languages (Stowell, 1981)" (Baker 2008: 355). Notice that the case of the head-parameter cannot be straightforwardly included in a macroparametric option determined by an implicational ordering of the type discussed so far, but that, as discussed by Roberts (2007: 92-102), the large scale typological variation, that it is often drawn back to the effect of a single head-parameter, should be studied more in detail to account for category-specific word-orders of languages like, for instance, German. In such cases the cross-categorial generalization entailed by the hypothesized head-parameter does not immediately apply. It does in other languages: For example "in English [a regular head-initial language, editor's note], a *right-branching language*, all heads precede their complement, while in Japanese [a regular head-final language, editor's note], a *left-branching language*, all heads follow their complements; the order is determined by one setting of the *head parameter*" (Chomsky 1995: 35). Such regularity determines that in English one observes the orders VO, AuxV and P-Obj (i.e. the existence of prepositions), while in Japanese one finds the orders OV, VAux and Obj-P (i.e. the existence of postpositions). In German instead the orders OV and AuxV are shown (in infinitival contexts, which are the unmarked ones) like in Japanese, but the language is clearly prepositional, like English. Roberts (2007: 97-98) proposes different options to explain why the parameter in its general formulation fails to satisfy the empirical verification in German-like languages. Among the others, the scholar takes into account the possibility of splitting the head-parameter into category-dependent subparameters, which, as claimed here, are not necessarily structurally dependent, but in regular head-initial and head-final languages could be thought to be set to coherent values. Perhaps this might happen in accordance with some kind of "(structurally-independent) cross-category harmony" of the type defined by Hawkins (1983: 134), which could be related to a general "regularizing" tendency of the language system. One could consider the effect of this phenomenon in comparison with that of analogy, intended as the cause of "irregular" phonological change recognized since the second half of the XIX century (see section 4.2.1). What is important to underline is that, since this kind of cross-category harmony would not derive from internal structural properties of the parametric space, it would not shed light on the deep design of the Universal Grammar, therefore it should be kept distinct from the interaction between parameters entailed by proper implicational ordering. This characterization of cross-category harmony seems to be compatible with Boeckx's (2008) idea that "clustering results from the child seeking to maximize the similarity across parameter settings, harmonize their values, and thereby economize what must be memorized (via the formation of a generalization across similar parameters)" (Boeckx 2008: 11), but with the difference that the scholar claims that any form of interaction between parameters should be reduced to this mechanism, which does not agree with the perspective adopted here.

occupied by parameters within the "networks"[19] resulting from the implicational relationships could reflect their macro/micro nature. This means that the more a parameter "ranks" higher in the structure, the more it should be considered a macroparameter: For instance, referring to Longobardi's (2005) schemata, a parameter of this type would be that determining whether definiteness is grammaticalized or not in a language. Conversley, moving down from the top of the network Roberts expects parameters to become more micro, which involves that they usually have a longer description and that they "are further along a learning path" (Roberts 2008: 9).

Having outlined the main current hypotheses regarding both the format and the interaction between parameters, from the next section on it is possible to show how the synchronically-based parametric perspective may be applied to highlight the dynamics of diachronic variation. This is done starting from the overview of the pre-generative and descriptive approaches to syntactic change.

3.2. Syntactic change

On first approximation, the term "syntactic change" may be intended to indicate any event of diachronic variation occurring in the history of a language. This general designation covers a wide range of possible types of change. An early fundamental criterion to classify these types consists of separating changes that derive from contact phenomena of any kind (see the discussion about the typology of language contact in section 3.3.1) from those occurred independently of areal convergence, e.g. purely genealogically. Of course this distinction fits together with that between horizontal and vertical transmission outlined in the end of the previous chapter. Notice that, from the point of view of the quantitative comparative analysis, to which this work is dedicated, this binary typology might be considered the only relevant one: In particular, the detection of contact-driven syntactic change is so important here that a specific section (section 3.3) is dedicated to the clarification of the definition and the role played by "syntactic borrowing" [20] in the history of languages.
Nonetheless, though satisfactory in a quantitative perspective, this classification is outlined with exclusive reference to a specific characterization of the causes of change, i.e. asking whether these causes are to be regarded as lineage-internal or external. To understand the actual dynamics of diachronic syntax necessitates providing a finer account for the different processes that may intervene. But is it possible to design a typology of syntactic change in light of the description of a limited set of fundamental mechanisms? This is an open and problematic question: Currently there is no agreement and just little explicit debate about the typology of syntactic change. In order to illustrate this situation and, furthermore, to clarify some widely-used specific terminology (e.g. terms like reanalysis, grammaticalization and so on), in the next section a brief outline of some proposals is provided, together with the historical contextualization of the most important notions. Section 3.2.1 is intended provide a basis to link the possible descriptive characterizations of syntactic change to the explanatory account, which can be given applying the parametric theory to the understanding of syntactic evolution: In particular, the extension of the parametric approach to the explanation of

---

[19] Actually the structures Roberts (2007: 277-282; 2008) refers to as "networks of parameters", e.g. Baker's (2001) PH, are tree-graphs. Anticipating the account provided in section 4.1.2, networks differ from trees in that they violate the chain condition (recall 2c).

[20] It is relevant to anticipate that here the terms "borrowing" is used in a general sense, i.e. it refers to any kind of possible non-genealogical influence between languages. Such a neutral and general connotation of "borrowing" differs from the specific ones that some scholars give to it (see section 3.3.1).

historical facts is covered in section 3.2.2 and section 3.2.3 illustrates some proposals to deal with the issue of gradualness. Finally, recent models designed to simulate the evolution of parametric systems are introduced in section 3.2.4.

3.2.1. Describing syntactic change

In the tradition of study of historical syntax there are rare proposals regarding the definition of a general and exhaustive descriptive typology of syntactic change. The debate has been mainly focused on the nature of single diachronic processes and on hypothesises related to their nature and theoretical status. For instance, an important controversy has affected grammaticalization, which is introduced below in this section. The question is: Should grammaticalization be assumed to instantiate a basic mechanism of diachronic syntax or might it be considered epiphenomenal respect to other processes? Going beyond descriptive typology, such matters may be included within the framework of attempts aimed at finding deeper explanatory generalizations of syntactic change (see for instance Faarlud's 1990: 31-46).

It is important to underline that saying that historical syntax has not yet provided a unified account for the typology of change does not entail that the entire field of study has been completely neglected in previous comparative linguistics. Longobardi (2003b) opens his programmatic discussion on historical syntax highlighting the fact that in traditional comparative linguistics this area of interest favoured the formulation of a large amount of hypotheses, but did not encourage the same methodological and empirical debate as phonology, morphology and lexicon, so that "in ben pochi casi […] generalizzazioni storico-comparative sono emerse in sintassi (there are few cases in which historical-comparative generalizations have risen in syntax)" (Longobardi 2003b: 165). In order to identify the causes of this "minority status", the scholar introduces some possible explanations, which are left out here to be taken into account in section 4.2.4 within the presentation of the Parametric Comparison Method. However, misunderstandings on the status of historical syntax studies can easily arise. For instance, disagreeing with Lightfoot (1988: 305-307), according to whom before the '70s syntactic change was an issue rather ignored by the mainstream discussion, Harris and Campbell (1995) argue that many present themes are deeply rooted in the history of the linguistic science. This means that, in the scholars' opinion, "[m]ost recent claims are not new, but were anticipated and are the continuation of lines of thought which have endured in the history of linguistics" (Harris and Campbell 1995: 14). Thus the summary of some relevant uses of notions regarding syntactic change in the history of linguistics sketched throughout this section has a twofold aim: It is intended on the one hand to show that, as claimed by Harris and Campbell (1995), it is indisputably true that important and still topical proposals related to diachronic syntax started appearing much before the '70s, on the other hand to support Longobardi's (2003b) claim, according to which these suggestions appear to be largely isolated, i.e. not outlined within autonomous and syntactically-specific methodological frameworks.

The descriptive typology of syntactic change adopted by Roberts (2007: 121-205) includes processes of grammaticalization and word-order change[21]. The author uses this classification with no explicit claim of being descriptively exhaustive, but just to show how apparently different cases of diachronic variation involve reanalysis and can ultimately be explained as

---

[21] Argument-structure change and complementation change are very specific descriptive sub-types of syntactic change that are also taken into account by the scholar and reduced to reanalysis-like processes. In any case, they are both left out here in order to preserve the clarity of the discussion and to focus it on the notions that are more relevant from the point of view of the history of the typology of syntactic change.

parametric changes (see sections 3.2.2 and 3.2.3). Indeed, nowadays reanalysis is typically studied within child-based theories of language change and it may be defined as the "process which changes the underlying structure of a syntactic pattern and which does not involve any modification of its surface manifestation" (Harris and Campbell 1995: 50). The earliest mention of a reanalysis-like change in a pre-acquisitional approach is probably Bopp's (1816) idea that infinitive morphology originated from "abstract" nominal forms reinterpreted as verbal endings. Moreover, reanalysis can be understood as a relevant underlying mechanism also, for instance, for the different possible patterns of syntactic change proposed by Brugmann's (1925: 7). Paul's (1898) explanation of cases of reanalysis in the history of German, instead, already included the hypothesis that this process derives from an analogy-driven bias of the learning process: This idea appears to anticipate the theoretical achievements attained in the second part of the XX century. Currently it is broadly accepted that this process "is clearly a major mechanism of syntactic evolution which we must understand in depth if we wish to understand how and why syntactic change occurs" (Langacker 1977: 57; quoted in Harris and Campbell 1995: 61). In particular, Harris and Campbell's (1995: 62-65) definition appears to cover a wide range of possible instantiations of reanalysis, which is assumed to affect constituency and hierarchical structure (e.g. the rise of the construction *for* + *to* in English from the reanalysis of a construction *for* + NP), category labels (e.g. the case of the Twi verb *wo*, which means ""be at", reanalysed as a preposition meaning "at"), grammatical relations (e.g. the loss of inversion constructions in the history of English) and cohesion (e.g. the reanalysis of the verb *nemi*, "to live/to walk" in Nahua, as clitic and affix). Therefore the authors hypothesize that reanalysis is one of the core mechanisms of syntactic change, together with extension/analogy, which "results in changes in the surface manifestation of a pattern and which does not involve immediate or intrinsic modification of underlying structure" (Harris and Campbell 1995: 51), and borrowing (see section 3.3.2). What is important to underline here is that the reduction of many sub-types of syntactic change (i.e. types defined on more descriptive grounds) to reanalysis appears to represent an important generalization on the way to find the extensive explanatory account provided within the parametric framework. In the following discussion three sub-types of syntactic change are considered in turn: Grammaticalization, exaptation and word-order change.

Among the earliest mentions of a grammaticalization-like process in the XIX century comparative linguistics there is Bopp's (1816) proposal that Indo-European endings may have been originated from independent elements that came to be adjoined to verbal roots: This idea is at the core of the German linguist's "Agglutination Theory". Nevertheless the identification of grammaticalization-like processes was not unknown since the "proto-linguistic" studies developed by Ancient Greek and Arabic grammarians (see for instance Stankiewicz 1974). Similarly, in the tradition of Sanskrit grammarians it is possible to find mention of at least one important idea related to the interpretation of syntactic change: That entailed by the "Root Theory", according to which all words (or, more precisely, all morphemes) derive from verbal roots[22]. The first proper designation of the term "grammaticalization" is provided by Meillet

---

[22] This theory is considered by Harris and Campbell (1995: 17) an example of the general "Growth Principle" approach, i.e. the idea that changes allow languages to attain successive levels of improvement, being them either from little to bigger, from concrete to abstract or from simple to complex. The discussion regarding this point has already been covered in section 2.1.1 dealing with the overview of the role that evolutionary concepts have played in the history of comparative linguistics: As stressed there, the improper interpretation of the evolutionary process undergone by languages exactly consists in assuming that diachronic development implies some form of improvement or decay. Nonetheless, different versions of the Growth Principle (e.g. Schleicher 1853;

(1912), who defines it as “the passage of an autonomous word to the role of grammatical element” (Meillet 1912: 131). Anticipating a hypothesis that is still important in the current generative framework, the French scholar associated with grammaticalization at least three important facts: The “loss of expressivity”, the weakening of phonological form and that of concrete meaning (see the examples reported in Hopper and Traugott 1993: 22-25). In the ‘60s there was a resurgence of interest in grammaticalization processes, so that Kuryłowicz (1965) said that it could be defined as consisting “in the increase of the range of a morpheme advancing from a lexical to a grammatical or from a less grammatical to a more grammatical status, e.g. from a derivative format to an inflectional one” (Kuryłowicz 1965: 52). From then on grammaticalization has been a focus of inquiry in many perspectives and it has been often connected to other phenomena and ideas, for instance “univerbation” (Givón 1984), semantic bleaching and phonological reduction. However, an important point is that “some (of these perspectives, editor’s note) refer explicitly to the output of grammaticalization as being ‘more abstract’ than its input (e.g. Traugott 1980: 46; 1990; Lehmann 1982: 128)” (Harris and Campbell 1995: 20). As already claimed, here grammaticalization is regarded as a sub-type of change with respect to reanalysis, following the ideas concerning its theoretical status outlined in the debate raised in the last decades. For instance, Hopper and Traugott (1993) are convinced that “reanalysis and analogy are the major mechanisms in language change” and “[t]hey do not define grammaticalization, nor are they coextensive with it, but grammaticalization does not occur without them” (Hopper and Traugott 1993: 61). This assumption is in line with Campbell’s (2001) view, according to which “‘[g]rammaticalization theory’ has no explanatory value because what it claims to explain is explained already by other well-understood mechanisms which lie behind it and, as is generally agreed, it cannot ‘explain’ without appeal to these other mechanisms and kinds of change” (Campbell 2001: 151). Therefore, grammaticalization may be regarded as a useful term, as it describes a specific instantiation of other basic mechanisms of syntactic change, but only once it has been established that it has no independent status with respect to them (Newmeyer 2001: 202-203): On the whole, it may be considered the process by means of which “new grammatical forms and constructions arise” or that “whereby items become more grammatical through time” (Hopper and Traugott 1993: 1-2; quoted in Roberts and Roussou 2003: 1). Some of the most important problems of grammaticalization concern the empirical support to its directionality. On the one hand, it seems like it is not true that the same lexical categories create the same grammatical ones in all languages, so that “[…] grammaticalization fails to evince the most important distinguishing feature of a distinct process – the unfolding of its component parts in a determinate sequence in which one step of the sequence inevitably engenders the following one” (Newmeyer 2001: 195). On the other hand, the validity of the unidirectional characterization of grammaticalization is questioned: If, for instance, Haspelmath (2004a: 21-23) regards counter-examples as misleading, in the last years “a growing body of literature is devoted to investigate various forms of ‘degrammaticalization’ (e.g. Giacalone Ramat 1998; van der Auwera 2002; Norde 2002)” (Narrog 2007: 2).

Exaptation (Lass 1990; 1997) has already been mentioned in the previous chapter (sections 2.1.2 and 2.2.3) in connection with the use that this term has in evolutionary biology, where it was originally conceived (Gould and Vrba 1982). It may be indicated as a process that is apparently similar to grammaticalization, but tends to contradict its general form. The two processes are assumed to have the same goal, i.e. the formation of new grammatical material, but they differ because while grammaticalization performs it starting from lexical elements, exaptation is assumed to recycle former grammatical elements that has become functionally unclear giving them a new grammatical role. In other words, exaptation may be thought of as

---

Jespersen 1922) persisted at least until the beginning of the XX century, even if contrasted by the Neogrammarian school (see in particular Osthoff and Brugmann 1878).

"a kind of conceptual renovation, as it were, of material that is already there, but either serving some other purpose, or serving no purpose" (Lass 1990: 316). One of Lass' examples that clearly illustrates exaptation is the construction of the complex case system that is observed in Modern Standard Finnish out of the Proto-Uralic affix system (Lass 1997: 320-324). In general, exaptation appears to be quite rare and specific, i.e. the process is not cross-linguistically replicated (however see the attempt to find a cross-linguistic account for grammaticalization in Narrog 2007), but due to its characterization this process requires to be distinguished by grammaticalization (see the general review provided by Traugott 2005).

While exaptation probably represents the newest typological proposal in the framework of diachronic syntax, the interest in issues related to word order can be brought back to the Ancient Greek and Roman grammatical traditions (Scaglione 1972: 155). One may say that the interest in word-order change started to draw the attention of European pre-scientific language studies at least since the XVII and the XVIII centuries, especially as a consequence of the overt comparison between the word order of Latin and that of its descendent languages. Interestingly, around the middle of '600 the French philosopher Geraud Cordemoy suggested that the observed word order reflects the order in which children acquire the language: Subjects and their qualifications, actions and objects. Herder (1772), thinking in a Growth Principle perspective, was convinced that the "logical" word order SVO represented an advance with respect to Latin's "inverted" one SOV, while Humboldt still considered acceptable to claim that languages showing inversion "favour clarity" (Harris and Campbell 1995: 22). In the end of the XIX century Delbrück (1878) proposed an idea that was developed in further Neogrammarians' work, i.e. that the original Indo-European word-order was SOV and then changed into SVO for reasons connected to stress requirements. In the same years Behagel (1878) was arguing that the original Germanic word-order was that currently found in subordinates, i.e. SOV, as it is reflected by old-aged compounds like for instance *übertreiben*, where *über*, interpretable as an adverbial modifier before the formation of the compound, precedes the verb *treiben*. Another fundamental suggestion was that provided by Wackernagel (1892), who observed that in Indo-European languages the "determining modifier" precedes the "determined head" and formulated the "law", according to which in Indo-European languages clitics tend to appear in sentence second position. As for the XX century, probably the most discussed paradigm of investigation regarding word-order typology is that initiated by Greenberg (1963), Lehman (1973) and Venneman (1973): It is exactly in these works that the first hypotheses on the already mentioned "implicational universals" are proposed. In particular, Vennemann (1973) developed the thought that changes tend to lead languages to show cross-categorial word-order consistency with respect to the implicational universals. Though rejected by many scholars (see for instance Mallinson and Blake 1981: 434-435), a similar idea (the "Cross-Category Harmony Principle") was put forward also by Hawkins (1983: 134) (about this point see also note 18). Within the generative perspective, there have been many significant attempts to account for word-order change in the last decades. Among them, Roberts (2007: 180-198) recalls van Kemenade's (1987) pre-minimalist approach to the explanation of the change from OV to VO in the history of English based on the idea that it was a consequence of the variation of the underlying head-complement order. This proposal was in some respects refined in Lightfoot (1991): His general view on syntactic change is the subject of the next section, in which it is shown how adopting the parametric model in diachronic syntax allows the reduction of the vast and sometimes baffling descriptive typology to a unique explanatory mechanism that is even deeper than reanalysis: Parameter resetting.

3.2.2. The parametric explanation

The idea that the explanation of language change in general has to be connected with language acquisition, i.e. that language change takes place in the course of language acquisition, is common since the early generative approaches to historical linguistics, i.e. at least since Klima (1964), Traugott (1965), King (1969) and Lakoff (1968). However, one may take Lightfoot (1979) as the first extensive account for diachronic syntax within the generative perspective. In this work the scholar outlines some proposals regarding the application of the Extended Standard Theory (Chomsky 1973) to syntactic change, putting forward two main questioned hypotheses. The fist one is that syntactic change is independent of any connection with semantic relations, pragmatic considerations and discourse function. This suggestion has been relaxed by the author in the following works. According to the second debated point, instead, Lightfoot (1979) assumes that grammatical changes largely occur as catastrophic restructuring events that follow from the gradual accumulation of small "environmental" changes ("Transparency Principle"). Even if relevantly modified, the core of this idea is maintained in Lightfoot's thought until the '90s and its discussion may be better framed within that concerning the issue of "gradualism". Much has been said in chapter 2 about it (see in particular sections 2.1.1, 2.1.4 and 2.3.2): Below the problem is reconsidered again to be regarded within the parametric perspective. In Lightfoot (1991) the scholar reinterprets his hypotheses regarding diachronic syntax within the Principles and parameters theory. Essentially, according to the scholar's proposal the main mechanism underlying syntactic change is the change of the value of parameters from on generation to the other, i.e. parameter resetting. Again, this idea confirms the thought that syntactic change occurs during language acquisition, i.e. the period in which parameter values are fixed[23]. As for the general design of Lightfoot's framework, two main questions deserve attention: The first is related to the logical problem of language change, the second one to the problem of gradualness, which is faced in the next section.

The logical problem of language change arises if one considers language acquisition as a deterministic process. If one accepts this view, they should consequently expect that the grammar acquired by a child precisely matches the grammar that generated the corpus, to which the child has been exposed to, in other words that the system of parametric values of the child converges with that of the speakers that produced the primary linguistic data. But the deterministic thesis is contrasted by the observation that changes occur from one generation to the following one, which means that one or more parameters have been set differently in the two generations. Thus the logical problem of language change can be stated as: "If the trigger experience of one generation, say g_1, permits members of g_1 to set p_k to value v_i, why is the trigger experience produced by g_1 insufficient to cause the next generation to set p_k to v_i?" (Clark and Roberts 1994: 12). The problem is schematized in figure 6 (adapted from Roberts 2007: 227): In this representation the speakers of generation 1 have a certain grammar *G1*, which generates the linguistic corpus 1, which, in an apparent contradictory way, is responsible for the acquisition of a grammar *G2* in generation 2, being *G2* different from *G1*. A possible solution to the apparent paradox is to relax the deterministic thesis regarding language acquisition and to suppose that in the course of the process "all parameter values must be fixed, but there is no requirement for convergence with the adult grammar (although this happens most of the times)" (Roberts and Roussou 2003: 13). In this sense, the aim of language acquisition should primarily be to approximate adult grammar, not of replicating it, even if the latter is largely the most frequent case (Roberts 2007: 231). What about the cases

---

[23] In particular, Lightfoot (1991: 22) assumes that children are degree-0 learners, which means that that the data that they use to acquire a language is exclusively constituted by clauses that are at a 0-degree of embedding, i.e. main clauses. Thus only modification of the degree-0 primary linguistic data can cause the child to converge to a grammar that is different respect to the adult grammar.

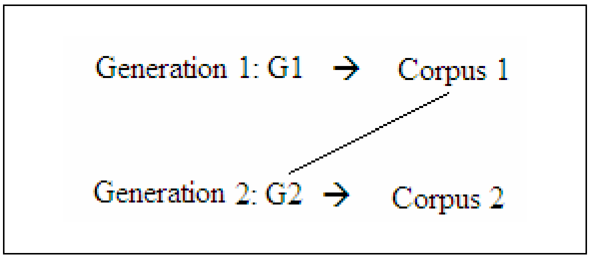


Figure 6: The logical problem of language change (adapted from Roberts 2007: 227).

of divergence? How can they be accounted for? One may think that, even if generally children's grammar perfectly overlaps with the parental one, "[s]ometimes […] minor changes in the relevant childhood experience cross a threshold and have consequences for the grammars that emerge." (Lightfoot 1999: 78-79). More specifically, these critical minor changes, which are ultimately responsible for change, may be thought to be those that affect the relevant triggering experience, i.e. the parameter triggers or cues (see the discussion is section 3.1.2): Following Roberts and Roussou (2003: 12-13), the critical changes are those that cause some parts of the trigger experience produced by the adult generation to become ambiguous for the next generation of acquirers, which then may reset the relevant parameters. This view significantly relies on Keenan (1994; 2002) and Longobardi's (2001a) Inertial Theory. According to it, syntax is assumed to be inert in diachrony, that is to be entirely unaffected by spontaneous internal change. In general terms, in syntax it is expected that "[t]hing stay as they are unless acted on by an outside force or decay" (Keenan 2002: 2). In the view of the Inertial Theory the "outside forces" correspond to changes that affect other linguistic levels, i.e. phonology, morphology and semantics, so that Longobardi (2001a) specifies the inertial behaviour of diachronic syntax as in (19).

(19) "[…] syntactic change should not arise, unless it can be shown to be caused – that is, to be a well-motivated consequence of other types of change (phonological changes and semantic changes, including the appearance/disappearance of whole lexical items), or recursively, of other syntactic changes […]" (Longobardi 2001a: 278).

On the one hand, to regard syntactic change as an indirect consequence of extra-syntactic changes permits solving the logical problem of language change that is raised within the acquisitional approach: Indeed, no paradox becomes evident if one considers that the adult grammar and the learner's are not directly connected, but are related by means of the primary linguistic data, and nothing prevents this data to be subject to, say, morphological erosion, semantic shift and so on. On the other hand, the idea that a syntactic change may recursively derive from another one is coherent with the hypotheses that (many) parameters interact one with the other, as shown in section 3.1.5. For the moment nothing more is said about the latter issue, which is reconsidered further after the introduction of Table A (section 5.2).

Roberts and Roussou (2003) apply the markedness hierarchy in (12) to parametric change, as it has been described so far, essentially arguing that whenever the trigger of a certain parameter has become ambiguous and obscure, "the learner will opt for the default option as part of the built-in preference of the learning device for simpler representations" (Roberts and Roussou 2003: 17). This proposal seems to be corroborated by a number of cross-linguistic examples of grammaticalization, but, if not better specified, it also carries a problematical theoretical consequence that cannot be straightforwardly accepted: That there is a universal tendency to default parametric values. Before facing this issue, one of the scholars' examples is illustrated in order to clarify their approach: The development of the Romance definite article.

It is well-known that Romance definite articles (e.g. Italian *il*, Spanish *el*, French *le* and so on) mostly developed out of the Latin demonstrative *ille* (and from *ipse* in varieties that are not considered here, like Sardinian and, partially, Catalan and Gascon) (see for instance Tagliavini 1999: 257-259). Latin had no definite articles, but it is documented that in Vulgar Latin (say around the V century) *ille* started occurring with nominals as definite article. One of the hypotheses that Roberts and Roussou (2003) take into account and aim at reinterpreting within their framework is Giusti's (2001) one, whose premise is that definite articles are directly merged in D, while demonstratives are generated lower in the structure and then move to merge to SpecD. Thus, considering the reanalysis undergone by *ille* from Latin to Vulgar Latin, one should account for the fact that there was a passage from the structure given in figure 7a to that presented in figure 7b. According to the approach adopted by Roberts and Roussou (2003), "[t]he trigger for the reanalysis was the phonological weakening of *ille* coupled with the loss of case morphology" (Roberts and Roussou 2003: 135). The background idea is that in nominal phrases that lack a definite article the nominal head (N) is obligatorily subject to N-to-D movement, which is assumed to be aimed at checking the Case features that are found in D. The rich case morphology of Latin allowed N-to-D movement to be not entirely expressed at PF, i.e. to optionally occur after the spell-out, because the N-to-D chain would have been visible anyway. Giusti (2001) claims that the examples like those in (20a) and (20b) (Giusti 2001: 168; reported in Roberts and Roussou 2003: 135) support the hypothesis that the overt N-to-D movement was largely optional in Latin, in that they show N in different intermediate positions of its raising to D.

(20) a. *vir* ille fortissimus
man-sg-Nom that-sg-Nom very strong-sg-Nom
That very strong man
(40, Seneca, *Ad Marciam de consolatione* I.3)

b. ultimam illam *faciem* rerum
last-sg-Acc that-sg-Acc aspect-sg-Acc things-pl-Gen
The last aspect of things
(40, Seneca, *Ad Marciam de consolatione* V.4)

Since the inflectional endings in general and case morphology in particular were deteriorating starting from Vulgar Latin, in Romance languages the insertion of a definite article in D could have represented a solution to make the N-to-D chain visible and, therefore, to allow case checking. Summing up, Giusti (2001) and Roberts and Roussou (2003: 131-136) are convinced that *ille* came to play this role because on the one hand, as former demonstrative, it was already independently +definite and on the other hand its phonological weakening facilitated its reanalysis as functional element.

In this case of grammaticalization Roberts and Roussou's (2003) markedness hierarchy in (12) seems to fit, in that their analysis entails a structural simplification of the nominal phrase. In particular, an instantiation of Move is lost (i.e. the movement of *ille* to the D area) and the demonstrative is reanalyzed as an element that is directly merged in D. Nonetheless, notice that the structural simplification does not necessarily correspond to the resetting of a parameter to an unmarked/default value. Nothing is said about this point by the authors, but rather in this example the parameter resetting appears to proceed in the opposite direction: Indeed, this case of grammaticalization seems to be reducible to the resetting of an *a*-scheme parameter, in Longobardi's (2005) terms, from the value "not grammaticalized" to "grammaticalized". Putting it differently, if it is true that Longobardi's (2005) *a*-scheme corresponds to a feature that has a phonological representation in Roberts and Roussou's (2003) basic formalism in (10), then the grammaticalization of definiteness from Latin to

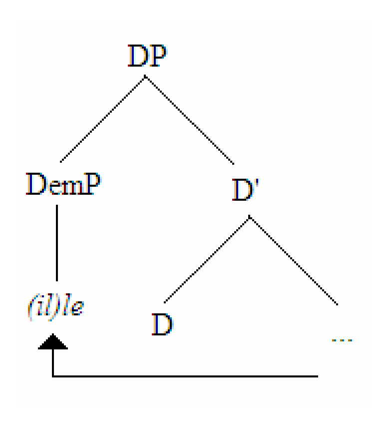


Figure 7a: Position of the demonstrative *ille* in Latin (adapted from Giusti 2001).

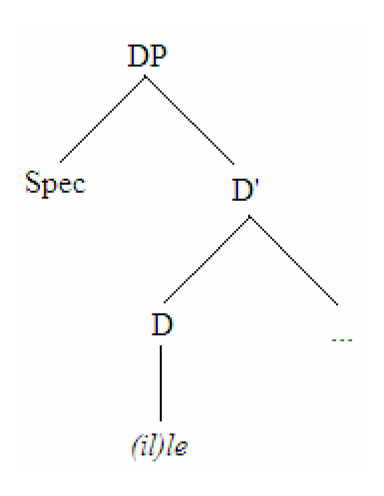


Figure 7b: The ranalysis of *ille* in Vulgar Latin (adapted from Giusti 2001).

Romance languages could be regarded as the resetting of the relevant parameter to the marked value. Therefore, one cannot make the general prediction that, since changes tend to trigger the less complex options, all languages should tend to have all the same unmarked parametric values: As claimed by Roberts and Roussou (2003), “the simplifications effected by changes are always local, and may increase the complexity elsewhere in the system” (Roberts and Roussou 2003: 17). However, Roberts (2007) observes that even if “a general move toward more and more unmarked systems […] is certainly possible in principle”, “whether it is actually happening it is an empirical question, albeit a rather difficult one to answer with any certainty” (Roberts 2007: 265).

3.2.3. Gradualness

The parametric interpretation of syntactic change leads one to ask to what extent syntactic evolution can be seen as a (more) gradual or a (more) punctuated diachronic process. To this end something has already been said in section 2.3.2, but now the matter can be reconsidered within the specific framework outlined so far. In the last section it has been pointed out that the problem is explicitly taken into account at least since Lightfoot (1979), who in the last decades has supported the hypothesis that syntactic change may be regarded as a “catastrophic” event, i.e. that “there must be points in history in when grammars differed in terms of some structural parameter setting” and “[i]n that case, changes are abrupt and catastrophic, with many surface effects” (Lightfoot 1999: 88). One can say that this view necessarily follows from any child-based approach to syntactic evolution, because if one assumes that changes take place in language acquisition, then they should be regarded as individual and punctual innovations, which primarily cause the knowledge of language of a single acquirer to diverge from that owned by the adult generation. Moreover, adopting a parameter-resetting model to explain language evolution, the abrupt characterization of syntactic change naturally arises, in that parameters represent discrete (binary) options, thus there should be nothing similar to a “partial” resetting of the values: Either a parameter is fixed to a certain value or to the other, no continuity or gradualness is expected between the two choices. It is important to recall what has been pointed out in section 2.3.2, i.e. that the discussion about gradualness might be biased by the neglect of a fundamental distinction, that should be explicated speaking about “gradualness” both in language and in biological evolution: The distinction between facts that concern the synchronic occurrence of

innovations and those that regard their diachronic diffusion. In biology mutations are sudden changes that affect the discrete units (i.e. genes) owned by individual organisms, but what is assumed to be gradual is their accumulation in populations and the consequent process of speciation. In the same way, parametric changes have to be thought of as innovations that arise punctually in single speakers' I-language of and then spread in the linguistic community in the course of time, thus determining the diachronic gradual variation between E-languages. The same separation is provided for example by Hale (1998), who distinguishes between change, that *per se* does not have temporal properties, and diffusion.
Compatibly with this perspective, the concept of "significant historical relationship" between I-languages has been formalized by Longobardi (2003b) as "σ relation", which is aimed at clarifying that "a certain language derives from another one" within the parametric view of diachronic syntax. The definition of this notion is introduced in (21).

(21) "a. L_2 diachronically derives from L_1 iff
  1. L_2 is acquired on the basis of a corpus generated by L_1 or
  2. L_2 derives from L_3 and L_3 derives from L_1.

b. L_2 immediately derives from L_1 iff L_2 is acquired on the basis of a corpus generated by L_1 (in which case L_1 and L_2 are diachronically contiguous).

c. Two linguistic objects *X* and *Y* (I-languages of fragments of I-languages) are related by a σ relation iff one derives from the other or *Z* exists, from which both *X* and *Y* derive" (adapted from Longobardi 2003b: 173).

The consequences of this formalization for the aims of the biolinguistic approach applied to historical comparative syntax are outlined in section 4.2.4, where the Parametric Comparison Method is taken into account. For the moment, it is important to observe that, as pointed out by Longobardi (2003b: 173-174), this definition of "derivation" fits both the cases in which the derivation occurs in a "regular" genealogical way, i.e. vertically, and that in which it includes horizontal transmission between different lineages, i.e. it entails the interference between different original I-languages, whose "fragments" (see 21c) are found in the resulting knowledge of language and are therefore historically related.

Thus the parametric approach to syntactic change leads to expect that syntactic evolution is locally, that is, at I-language level, abrupt. However this model does not directly support Lightfoot's (1991) idea that also at E-language level grammatical changes are expected to occur (quite) rapidly as a consequence of a large-scale parameter resetting, which had happened in a given community., i.e. that they follow a "S-curve" (among others see Osgood and Sebeok 1954: 155; Winreich *et alii* 1968: 113; Kroch 1989: 203; Denison 1999). Essentially, it seems clear that a parametric approach to syntactic evolution entails an "abruptness of grammar change in individuals of different generations", but "there is nothing in this cognitive model which requires the spread about which Lightfoot speaks of a new parameter setting throughout a speech community to be rapid or S-shaped" (Harris and Campbell 1995: 41). The debate regarding the validity of this application of the punctuated-equilibrium hypotheses to E-language evolution[24] closely resembles the discussion that has developed in evolutionary biology since Eldredge and Gould (1972) and it is mainly based on the empirical evaluation of historical data provided by different scholars. Moreover, it is fundamentally connected to the dynamics of diffusion of parametric change within speech

---

[24] Recall that this application of the punctuated equilibrium hypothesis is substantially different from Dixon's (1997), which instead concerns the identification of the different contribution of vertical and horizontal transmission in diachronic variation (see section 2.3.3).

communities. To provide an extensive account of this matter is beyond the aims of the present work, in that it would require a detailed digression into the sociolinguistic aspects of syntactic variation (see for instance the review provided in Roberts 2007: 315-340). Furthermore, once it is assumed that parameter resetting is a satisfying explanatory mechanism for syntactic change and that I-languages are diachronically linked by means of a significant historical relation (σ relation), in principle nothing more needs to be said here. Nevertheless, considering the involvement in quantitative and automatic analysis of parametric data of this dissertation, it is relevant to go through a brief exemplification of some recent experiments regarding the algorithmic simulation of the population-level evolution of parametric systems. This is done in the following section.

3.2.4. Simulation of parametric evolution

In the last years interest in the simulation of population-level (i.e. E-language level) diachronic development of languages has significantly increased and much experimentation has been proposed (see for instance Briscoe 2002; Turkel 2002; Gong *et alii* 2007). Here the focus is on the approach drawn in particular in Niyogi and Berwick (1995) and Niyogi (2002)[25], which covers specifically the investigation of the development of parametric systems. The simulation of parametric evolution entails considering the diachronic development of the set of parametric values owned by a certain language as the progression of a dynamical system, i.e. basically as the transformation undergone by a system of traits that changes in the course of time. The simplest model of this kind is one in which "the population initially consists of a fixed number of 'adult' language agents with pre-defined I-languages, and their output constitutes the data from which the next generation of 'child' learning agents acquires new I-languages" (adapted from Briscoe 2002: 5). Such a model perfectly fits the biolinguistic framework. Anyway, the strong idealization that is made is that the adult generation and the child generation are not overlapping, which means that they are considered clearly separated and ordered in the timeline in which the dynamical parametric system evolves.

Niyogi and Berwick's (1995) point is to show how theoretical hypotheses regarding specific parametric changes can be used to model a dynamical system. They argue that the trajectory of this system may be compared to the historical one that is observed in the lineage of the relevant language: Their idea is that if the trajectory of the simulated system and the actually documented one converge, this would corroborate the theoretical hypotheses under which the simulated parametric system has been designed. In other words, the authors want to "pose a formally criterion for the adequacy of grammatical theories" (Niyogi and Berwick 1995: 16) Within their experiment the ordered generations are linked by means of a trigger-based memoryless learning algorithm, which models the acquisitional process that occurs from one generation to the next: The memoryless characterization of the learning algorithm is very important, because it reflects the claim that previous states of the parametric settings are neither preserved nor influence successive stages of diachronic development of I-languages (Longobardi 2003b: 180-182). Niyogi and Berwick (1995) apply their procedure of analysis to examine the loss of Verb Second position from Old French (spoken between XXI and XIV century) to Modern French (since XVII century). In particular, they test the validity of Gibson and Wexler's (1994) and Clark and Robert's (1993) parametric accounts for this phenomenon. According to the former the change involved the resetting of three distinct parameters, heretofore mentioned: p_1 (related to the order head-specifier, either SpecX or XSpec), p_2 (related to the order head-complement, either XCompl or ComplX), p_3 (verb second, yes or

---

[25] Among others, see also Niyogi and Berwick (1997; 1998) and the overview of this line of research presented in Niyogi (2006).

no). Clark and Robert (1993) instead adopt a 5-parameter system to explain the loss of V2 in the history of French: p_1 (case assignment under agreement, yes or no), p_2 (case assignment under government, yes or no; relevant triggers may be the orders AdvVS and SVO), p_3 (nominative clitics, yes or no), p_4 (null subject, yes or no; relevant trigger may be the order WhVSO), p_5 (verb second, yes or no; relevant triggers may be the orders SVO and AdvVS). Niyogi and Berwick's (1995) simulation shows that the Gibson and Wexler's (1994) 3-parameter system allows the parameter related to verb second to be more regularly reset and to remain more stable than in Clark and Robert's (1993) 5-parameter system. Thus the scholars tentatively suggest that the loss of V2 "might have more to do with the choice of parameterizations than with learning algorithms" (Niyogi and Berwick 1995: 16).
In Niyogi (2002), instead, the scholar shows how Cavalli-Sforza and Feldman's (1981) model of cultural evolution (see section 2.3.1) may be linked to Niyogi and Berwick's (1995; 1997) model of parametric evolution: In particular, his focus is to adapt Cavalli-Sforza and Feldman's (1981) model of individual transmission to a case in which a child acquires a language starting from eventually contrasting adult grammars, i.e. two grammars in competition. On this basis, using a revised version of Niyogi and Berwick's (1995) trigger-based learning algorithm, the author outlines a simulation of the evolution of syntax from Old to Middle and to Modern English, with particular reference to the debated change from VO to OV order, showing how his approach can help to verify the plausibility of the various hypotheses regarding this specific change. Even if its insight is not decisive, it seems clear that the population-level simulation pursued in this line of research "demonstrates that the consequent predictions about the trajectory and direction of change are often surprising, very varied, and always sufficiently complex that mathematical modelling and/or computational simulation are essential tools in deriving them" (Briscoe 2002: 6).

Now that also the mechanisms of parametric change have been characterized in light of the abrupt process of parameter resetting at I-language level and some proposals have been introduced regarding the possibility of modelling the gradual diachronic spread of the innovation through the linguistic community, it is important to look closer at a specific type of change: That caused by language contact. The issue is covered in the next section.

3.3. Syntactic borrowing

Starting from chapter 2 the discussion on genealogical evolution and areal convergence of languages (i.e. vertical and horizontal transmission of linguistic innovations) has been based an obvious premise, which states that contact-induced phenomena are possible and, at some levels of analysis, above all in lexicon, they are also common enough to play an important role in diachronic variation. These assertions are not straightforwardly true in historical syntax, since the very possibility and the degree of "syntactic borrowing" is debated. The next sections are dedicated to the clarification of the issue: On the one hand, the concept of "borrowing" that is relevant for the present work is explained and evidence in favour of the possibility of syntactic borrowing is provided; On the other hand it is explained how borrowing can be accounted for in the parametric framework.

3.3.1. Possibility of syntactic borrowing

A general overview of the typology of contact-induced phenomena is useful to introduce the specific topic of syntactic borrowing. The first observation to point out is that social and demographic factors play a fundamental role in processes of language contact, thus it is impossible to consider them abstracting from the sociolinguistic perspective. A distinction provided by Thomason and Kaufmann (1988) is that between “borrowing” and “shift-induced interference”: In the authors’ model, the former corresponds to a situation of bilingualism, in which interference is incorporated from a second language into the speakers’ first language and above all lexicon is affected; The latter instead refers specifically to cases of imperfect learning, in which phonology and syntax are primarily involved, “while lexicon and morphology […] lag behind” (Thomason 2001: 4). In this taxonomy (reported similarly also by other authors, see for instance Kroch 2000: 176) typical borrowing arises when native speakers consciously import words from another language, while typical shift-induced interference is created by non-native speakers, who unconsciously impose features of their native language to the recipient language. Also Roberts (2007) distinguishes between two main kinds of possible contact situations, direct and indirect, which approximately resemble in their definition the notion of “borrowing” and “imperfect learning”[26]: “[…] in the indirect case the PLD [Primary Linguistic Data] simply contains a quantity of tokens from a distinct system, while in the indirect case the older group uses a second language in interaction with the younger group” (Roberts 2007: 391). The scholar’s representation of direct and indirect contact are respectively reported in figure 8a and 8b. Ross (1999, 2001) refines the specificity of Thomason and Kaufman’s (1988) classes of borrowing and shift-induced interference referring to them as “exo-borrowing” and “substrate”. Furthermore, he distinguishes two other complementary types of contact situations: “Metatypy”, which, according to his model, occurs when native speakers impose on their own language semantic-syntactic features from a dominant language, and “endo-borrowing”, referred to a situation in which a minority group imports words from its own language into the majority language, in order to express its cultural identity (e.g. Yiddish words used in English by Jewish Americans). Other authors focus the attention specifically on the typology of lexical borrowing (e.g. Haugen 1950; Myers-Scotton 2002), which is less relevant here. On the whole, the typology of contact situations appears to be potentially very complex and its exhaustive discussion is not among the aims of the present work. Therefore, taking inspiration from Thomason’s (2001: 3) notion of general contact-induced change, a clearer definition of what is meant here with the term “borrowing” can be that given in (22).

(22) Borrowing is intended to be any kind of linguistic change that is less likely to have occurred outside a specific contact situation.

This designation is broad enough to include any kind of at least suspected non-genealogical events of contact-induced change, regardless of the specific processes that they entail. Ultimately, referring to syntactic change, this is the only notion of “borrowing” that is substantially relevant with respect to the quantitative analysis introduced in chapter 6.

In light of such definition, the question that needs to be answered now is whether borrowing can actually affect syntax. The debate regarding the possibility of grammatical interference leads back to the first decades of the twentieth century, when contradictory views started appearing in the literature. Meillet (1921) and Sapir (1927) respectively argue that “[t]he grammatical systems of two languages are impenetrable to each other” (Meillet 1921: 82) and “[n]owhere do we find any but superficial morphological interinfluencings” (Sapir 1927: 217). Conversely, Schuchardt (1928) is convinced that “[e]ven closely knit structures, like inflectional endings, are not secure against invasion by foreign material” (Schuchardt 1928: 195). Weinreich (1979) observes that those clashing perspectives could essentially be brought back to “the lack of agreement between them on fundamental terms and concepts”, since to

---

[26] “[…] the direct case [of contact] does not really have to result in ‘borrowing’ on the part of either the younger or the older group: the Alien Corpus merely has an effect on the younger group’s PLD” (Roberts 2007: 391).

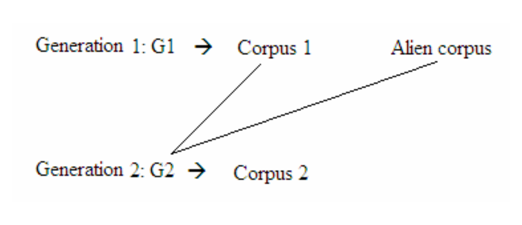


Figure 8a: Direct contact (taken from Roberts 2007: 390)

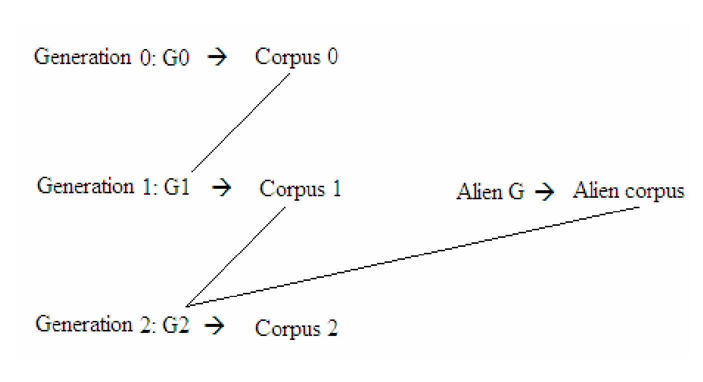


Figure 8b: Indirect contact (taken from Roberts 2007: 391)

the day the authors were writing there was “little uniformity in the drawing of lines between morphology and syntax, grammar and lexicon” (Weinreich 1979: 29). However, the author supports the idea that “morphemes and grammatical relations belonging to one language can occur in the speech of another language as ‘borrowings’” (Weinreich 1979: 30).

More recently, as pointed out by Sankoff (2001: 13), the same opposite views regarding syntactic borrowing have been discussed in the literature about language contact: Some scholars hypothesise that morphosyntactic elements and rules can be subject to genuine and direct borrowing as any other linguistic element (e.g. Thomason and Kaufmann 1988; Campbell 1993; Thomason 2001), others believe that this kind of transfer is almost impossible, and they account for internal syntactic change only as a consequence of lexical or pragmatic inter-influence (e.g. Lefebvre 1985; Prince 1988; King 2000; Nakhleh *et alii* 2005a). In particular, according to Thomason (2001) “[…] no absolute constraint against direct rule[27] transfer can be maintained” and “[…] the continuing popularity of the no-rule-borrowing position results in part from a pervasive tendency to underestimate speakers’ ability to manipulate their languages’ structures deliberately” (Thomason 2001: 2). Nakhleh *et alii*’s (2005a) position instead is in line with the most restrictive model and relies on the evidence provided by recent research on language contact, in which it is demonstrated that the inter-linguistic transfer of “closed-class” items occurs via processes different from those of typical lexical borrowing (e.g. Rayfield 1970:103-107; Prince and Pintzuk 2000): “[…] we accept the hypothesis that borrowing into native dialects from languages or dialects that are not closely related is tightly constrained, lexemes being virtually the only type of linguistic unit that is borrowed outright” (Nakhleh *et alii* 2005a: 4).

Within historical syntax “for a long time syntactic interference was considered either impossible or marginal” (Bowern, in press: 13), even if in some lines of research it has been believed that any kind of change is a type or a result of language contact (e.g. Poussa 1982; Mufwene 2001). Harris and Campbell (1995) include borrowing among the basic mechanisms that would allow to explain syntactic change and they define it as “a change in which a foreign syntactic pattern […] is incorporated into the borrowing language through the influence of a donor pattern found in a contact language” (Harris and Campbell 1995: 122). The authors distinguish it from contact, which according to them should be referred to cases in which the speakers of one of the languages in question is “familiar” with the other. Furthermore, they are convinced that “given enough time and intensity of contact, virtually anything can (ultimately) be borrowed)” (Harris and Campbell 1995: 149). Some important ideas have been pointed out also within the biolinguistic paradigm, but they are left out to be

---

[27] “Rule” is intended as any kind of grammatical generalization, including morphosyntactic and word-order patterns.

presented in the next section. The question that has to be posed now is: Which kind of evidence can be provided in favour of the existence of syntactic borrowing?

For instance, Roberts (2007) analyzes Trips's (2002) hypothesis, regarding the word-order change from OV to VO in Old English attributed to the contact with Old Norse around the tenth century, and a possible case of substratum influence undergone by Hiberno-English (the English of Ireland) due to the presence of Welsh English (Thomas 1994: 137): He concludes that in both studies the contact situation represents a not necessary and barely sufficient account to explain the changes. Though such specific instances of contact provide weak evidence in favour of contact-induced syntactic change, the author points out that other significant examples have been reported in this sense and can be regarded within the parametric approach. One of these, i.e. the borrowing of Preposition Stranding from English to Prince Edward Island French, is precisely taken into consideration in the next section. Thomason (2001) reports many cases of hypothesized rule borrowing, mainly regarding morphological patterns, but including also examples of syntactic-features transfer. Among the best documented ones she considers, there is the situation of Kupwar (systematically studied by Gumperz and Wilson 1971), an Indian village that lies on the border between the Indic and the Dravidian speaking areas: Four linguistic communities are settled there, two of them are Indic and two Dravidian. A massive syntactic borrowing has occurred among these varieties without any mediation of lexical transfer, since no shared morphemes are concerned in the relevant constructions. In particular, the rules involved in the transfer are: Subject/verb agreement in patterns in four different constructions, the syntax of the yes/no question marker and rules for the use of demonstratives and possessive forms in both attributive and predicative constructions.

The evidence provided by Gumperz and Wilson (1971) seems to support Ross' (1999, 2001) idea of metatypy. Perhaps Ross' most famous account for it is based on the study of the massive adoption of semantic-syntactic patterns from the Papuan language Waskia into the Oceanic language Takia, which appears to have occurred without any lexical or phonological diffusion of any kind: In the scenario drawn by the scholar, the semantic shift and the syntactic restructuring are part of the same underlying process. In particular, the transfer from Waskia to Takia is assumed to consists of: "SOV order, postposed demonstrative, postpositions, preposed full-NP possessors with a binary alienability distinction, and cosubordinate clause linkage. […] semantic patterns and idioms" (Haspelamth 2004b: 217). Ross' explanation of metatypy relies on the psychology of the bilingual speakers, who try to fill up the prestige gap between their own language and the dominant language assimilating to the first one relevant traits of the second one, but without modifying the lexicon, in order to keep intact their cultural identity. Nonetheless there are cases in which metatypy does not apply so strictly and straightforwardly: For instance, the syntactic influence from Southern Min on Taiwanese Mandarin (Chappell 2001) works in the other way around respect to Ross' expectation, since the dominant variety is precisely Taiwanese Mandarin. Even if the explanation proposed by Ross does not account for similar cases, the evidence he has collected seems to indicates that outright syntactic interference is possible. Other attested cases of significant syntactic borrowing regard, for instance, inter-influence among the Tamangic languages (Noonan 2006), that regarding the Indian (Emeneau 1964) and the Balkan *Sprachbund* (Joseph 1983), as well as the transfers described from Finnish into Russian (Thomason and Kaufman 1988), from north-western Afro-asiatic into Insular Celtic (Gensler 1993) and from Cushitic into Ethiopian Semitic (Weninger 2001). Moreover, an hypothesis that will be particularly relevant in the discussion introduced in the following chapters, in particular in chapter 6, is that regarding the supposed contact-induced diffusion of definite articles throughout the European area (Heine and Kuteva 2006: 97-139).

In conclusion, the existence of syntactic borrowing seems to be corroborated by a number of studies. There are cases in which this event occurs alone and others in which it is accompanied by lexical, phonological or morphological calquing, cases in which it is direct and others in which it stands as a derivative phenomenon. All these options have been attested, with infinite degrees of strength of the intervening variables. Concerning the aims of the present research, the only important fact is that syntax is not immune to contact-induced change, being the latter of any kind. In the next section some ideas regarding the possible interpretation of syntactic borrowing within the parametric perspective are presented.

3.3.2. A parametric account for syntactic borrowing

Even if in the biolinguistic framework the problem of syntactic borrowing seems to have spawned little extensive methodological research so far, some important observations and accounts have been proposed. In particular, within the already mentioned Longobardi's (2001a) Inertial Theory paradigm syntactic borrowing is not excluded from the process of diachronic development, but it is claimed that its theoretical *status* has to be distinguished from that of proper change: "[…] interference would not be strictly speaking a case of change from, say, grammar 1 to grammar 2, the latter built on a primary corpus generated by the former; rather, it would be the case of the deterministic construction of grammar 2 on a mixed corpus generated by grammar 1 and by other significantly distinct grammars" (Longobardi 2001a: 278). Moreover, it has been pointed out in section 3.2.3 that Longobardi's (2003b) formal definition of "historical derivation" and that of σ relation explicitly includes the case in which a certain knowledge of language is build up from fragments of different I-languages. Roberts (2007: 383-406) largely investigates the role of language contact in the diachronic parameter-resetting. He claims that indisputably the intervention of an alien grammatical system can significantly contaminate the primary corpus on which a new generation acquires a target language, thus causing a shift in the fixation of the parametric values. In synthesis, his point is that "[…] the evidence of language contact can be integrated into a parametric, acquisition-driven model of syntactic change quite unproblematically […]" (Roberts 2007: 406).

As already pointed out, among the clearest cases of syntactic borrowing, one may consider the direct contact between English and Prince Edward Island French (see also the overview given in Roberts 2007: 238-242), which resulted in borrowing of Preposition Stranding (PS), along with some English prepositions, into the insular Romance variety. What is meant with PS? PS indicates the rarely attested rule, which allows the complement of a prepositional phrase to raise to a higher position leaving the prepositional head behind, i.e. "stranded" in its original position. The examples in (23) show how this rule applies to English. In both cases the stranded preposition is *to*: (23a) illustrates the situation in which the element that raises is "who", while in (23b) "John" undergoes movement.

(23) a. [Who]$_i$ did you speak to t$_i$?
b. [John]$_i$ was spoken to t$_i$.
(adapted from Roberts 2007: 238)

In standard modern French PS is not a valid rule, in that it generates the ungrammatical sentences in (24a) to be compared with that in (24b); As known at least since King and Roberge (1990) and illustrated in (25), PS is applicable in Prince Edward Island French instead (as well as in other North American varieties of French, even if less robustly).

(24) a. *[Qui]$_i$ as-tu parlé à t$_i$?
Who have-you speak to
b. [A qui]$_i$ as-tu parlé t$_i$?
To who have-you speak
(adapted from Roberts 2007: 238)

(25) [Où]$_i$ ce-qu'elle vient de t$_i$?
Where that-she come from?
(adapted from King 2000: 136)

Notice that Prince Edward Island French borrowed many prepositions from English, for instance *off* and *about*, which is demonstrate in the example in 26.

(26) [Quoi]$_i$ ce-qu'il a parlé about t$_i$?
What that-he have speak about ?
(adapted from King 2000: 143)

The first hypothesis that can be made is that PS has been regularly imported in the Romance language together with prepositions. This would be confirmed also by other cases, in which there has been large borrowing of prepositions from English to North American French varieties, which now allow PS: For instance, one may think about Monctoc French (Roy 1979) and Nova Scotia Acadian varieties (Flikeid 1989). Where the lexical borrowing has been more limited, e.g. in Newfoundland varieties of Acadian (King 2000: 144), the rule has not been transmitted. Nonetheless notice that the new syntactic patterns of Prince Edward Island French are not identical with English patterns: Rather, it seem like the rule of PS has been generalized by the Romance speakers, so that it can apply also to cases in which in English it would generate (almost) ungrammatical sentences. So, the examples in (27) show the contrast between English (27a) and Prince Edward Island French (27b).

(27) a. *[Who]$_i$ did Pugsley give a book yesterday to t$_i$?
b. [Quoi]$_i$ ce-que tu as parlé hier à Jean de t$_i$?
What that-that you have speak yesterday to Jean about
(adapted from King 2000: 146)

King (2000) suggests that "there is a causal relationship between the borrowing of prepositions and the emergence of Preposition Stranding" (King 2000: 145) and that, since PS is found in the Romance variety without the constraints that the rule has in English, "an account of Preposition Stranding in Prince Edward Island French in terms of direct syntactic borrowing is untenable" (King 2000: 147). Anyway, whatever the exact explanation of this syntactic borrowing, being it indirect or direct, i.e. mediated or not by lexical borrowing, being it a syntactic calque or not, it is clear that the change occurred in Prince Edward Island French can be significantly brought back to a contact situation and, according to the definition in (22), it may be considered an actual instance of "syntactic borrowing".

The contact between English and the Romance variety can be accounted for in general terms within the parametric perspective claiming that at a certain point English expressions that incorporated PS (and therefore prepositions) must have been included in the primary linguistic data of the insular French acquirers. As a consequence, in that way the value of the relevant parameter that governs PS was changed respective of standard French. Roberts (2007) does not indicate which parameter should be responsible of PS. For that matter, a proposal that deserves to be mentioned is that of Snyder (2001) and Sugisaki and Snyder (2002), who consider PS (and, following Stowell 1981; 1982, other relevant constructions, e.g. V +

particle constructions) as a rule determined by the setting of a specific noun-compounding parameter. This idea is based on the observations that languages that have endocentric compounds allow PS and that during the acquisition of English children start applying PS only after they have learned how to produce noun compounds. If true, this hypothesis would entail the identification of the exact parameter that underwent resetting in the history of Prince Edward Island French, but, as shown by Gebhardt (2005), it is not: Indeed English speakers productively use noun-compounding, but speakers of both French and Prince Edward Island French use endocentric compounds "at similar minimal rates". Even if this seems not to be the right parametric model and the case in question remains open, what is relevant to underline is that syntactic borrowing, as an instance of syntactic change, can be regularly interpreted within the biolinguistic approach.

3.4. Conclusion

In this chapter an explanatory account for syntactic evolution within the parametric perspective has been provided. After having outlined the basic methodological and formal characterization of the biolinguistic program, the discussion has been dedicated to the introduction of some current hypotheses regarding the parameter formats and the forms of interaction between parameters. Then it has been shown how the biolinguistic approach, originally designed to explain cross-linguistic synchronic variation, can be applied to the investigation of diachronic syntax. In particular, it has been argued that the complex and debated descriptive typology of syntactic change can be largely reduced to reanalysis and that reanalysis is interpretable in the acquisition-based paradigm. This means that this process has been assumed to reflect a deeper explanatory mechanism of syntactic change, i.e. the diachronic process of parameter resetting. This account raises the problems regarding the "logical problem of language change" and the gradualness of syntactic evolution: While the former may be coherently solved in an inertial perspective, i.e. assuming that syntactic change indirectly follows from morpho-phonological and semantic changes that affect the primary linguistic data, the latter needs to be faced separating the I-language level from the E-language one. From the point of view of the parametric analysis, the syntactic change necessarily appears to be abrupt at I-language level, while in principle nothing prevents the consideration of its diachronic diffusion at E-language level more gradual or more punctuated. Thus the concept of "significant historical relation" between I-languages has been outlined and some recent proposals regarding the computational simulation of the population-level diffusion of parametric changes have been briefly illustrated. Finally, it has been stressed again the importance of distinguishing between "regularly genealogical" syntactic change and contact-induced syntactic change, which has been generally referred to as "syntactic borrowing". After having shown how syntactic borrowing actually plays a relevant role in the history of languages, it has been claimed that it can be ultimately analyzed as a case of diachronic parameter resetting too.

On the basis of this discussion, it is now possible to reconsider the specific characterization of syntactic evolution in comparison with the model of biological evolution introduced in chapter 2. Table 1 illustrates an overview of the relevant factors and features that have been attributed to syntactic evolution so far. Notice that this model shares many important similarities with that describing biological evolution. In particular, the relevant units that are taken into consideration in both the processes are discrete (parameters/genes), thus they are assumed to abruptly change at individual level (I-language/organism) but to gradually accumulate at E-language or population level. Furthermore, both the processes are assumed to be memoryless. As already claimed in the previous chapter, the most evident difference between language evolution in general and biological evolution regards the modalities of selection of changes: In particular, in syntactic evolution there is nothing similar to natural selection, but rather structural pre-selection may be clearly defined in light of the limits of parametric variation and, of course, sociolinguistic variables may intervene to socially select changes. As the gene flow that occurs due to the process of demic migration, syntactic borrowing represents an actual possibility of areal interference between syntactic systems and it has the effect of making these systems more similar. Furthermore, the discussion about genetic drift provided in chapter 2 perfectly fits the specific case of syntactic evolution. Finally, the dynamics of individual transmission appear to be significantly similar. As argued in section 2.3.2, there is a substantial difference between individual transmission of syntax and individual biological transmission: The former occurs during the whole critical period of language acquisition, while the latter is a punctuated event. Nonetheless, it is important to notice that "once the process is completed (after the critical period for knowledge of language and after reproduction for genetic material), in both cases the relevant traits (narrow components of I-language and genes) do not undergo further modifications in the course of an individual's life" (section 2.3.2).

On the whole, the analogical parallel between the model of syntactic evolution assumed here and that of biological evolution seems to reinforce the conviction that, from the quantitative point of view, the two processes may be studied by means of similar methods of investigation. The next chapter covers the presentation of these methods and is introduced by an extensive discussion regarding the different comparative methods that can be used in historical linguistics.

4. Comparative methods

The suggestion provided in the end of the last chapter is that syntactic evolution and biological evolution may be studied by means of similar methods of investigation from the quantitative point of view. This statement needs to be elucidated in many regards providing a more accurate account for some notions and subjects already considered in the previous discussion. First, what is exactly intended here saying "to study language evolution (and biological evolution)"? Moreover, which are the relevant methods of investigation that can be adopted to pursue this objective? Finally, which is the exact meaning of the term "quantitative"? These three questions are treated in turn in this chapter, whose main focus is the critical illustration of the possible methods that can be used to shed light on the historical development of languages, i.e. of the "comparative methods".

The contents are organized as follows. Section 4.1 is dedicated to the presentation of the aim of the investigation, that is to the explanation of the meaning and the role of "genealogical classification" in linguistics. This section of the chapter also includes a digression on the formal description of the tree-like and network-like representations of evolutionary paths. In section 4.2, instead, a critical review of diverse comparative methods used in linguistics and molecular biology is provided: Given that essentially the relevant differences between these methods concern what is compared and how it is compared, the idea is to contrast the guidelines of three linguistics-specific comparative methods (i.e. the classical comparative method, lexicostatistic analysis and mass comparison) with those of the method of comparison used in population genetics. The goal is to lay the foundations to appreciate the introduction of the syntactic-specific method adopted here: The "Parametric comparison method". The characterization of this approach and its explicit parallelism with the procedures of investigations used in molecular biology paves the way to the presentation of the quantitative techniques of "phylogenetic reconstruction", which is covered in section 4.3. This last section is then intended to provide some basic technical notions of the field of study that may be referred to as "quantitative phylogenetics": In particular, it is dedicated to the outline of some character-based and distance-based phylogenetic methods that are reconsidered within the linguistic framework both in chapter 5 and, above all, in the experimentation presented in chapter 6.

4.1. Genealogical classification and beyond

For the aims of the present work, in the first place to investigate language evolution means to study the genealogical classification of languages, i.e. "identify instances of genetic relatedness amongst languages" (Harrison 2003: 241), which, according to Harrison (2003), is the first goal of comparative historical linguistics. The genealogical classification "is […] comparison with a particular goal in view – the goal being the discernment of meaningful resemblances in the data, which in turn may allow us to identify those languages which descend from a single common ancestor" (McMahona and McMahon 2005: 1). Notice that, as in Harrison (2003: 241), in many works (e.g. Greenberg 1987; Thomason and Kaufmann 1988; the articles collected in the volumes introduced by Baldi 1990, Durie and Ross 1996 and Aikhenvald and Dixon 2001; Campbell 2004: 184-210; McMahon and McMahon 2005) the term "genetic" is used instead of "genealogical": However, here "genealogical" is preferred "in order to avoid confusion with biological genetics" and in light of the fact that "the term 'genealogical' is more transparent than 'genetic', because the linguistic relationships in question are indeed family-like, and not merely ‘related to genesis /origin’ in a general sense” (Haspelmath 2004b: 222)[1]. Another term that in the following discussion is used as synonymous of “genealogical” is “phylogenetic”: Its exact origin and use are taken into account starting from section 4.1.1 and, more extensively, in section 4.3.1, which is dedicated to the historical overview of “quantitative phylogenetics”.

The surface meaning of “genealogical relatedness” appears to be unproblematically understandable even for non-linguists. Nonetheless at least two points require attention: The first one is related to the different kinds and the value of the groupings that may be defined on the basis of genealogical relationships, the second one concerns how genealogical relationships can be formally defined and represented. Reflection on the former issue are then provided in section 4.1.1. Something important regarding the latter point, instead, has been already stated in section 3.2.3, when Longobardi’s (2003b) characterization of “σ relation” between I-languages has been introduced. With this conceptual premise in mind, section 4.1.2 is thought to bring in some basic notions and terminology regarding the tree-graph (and network-graph) formalism applied to the representation of genealogical relationships.

4.1.1. Grouping languages

Languages can be grouped on the basis of different criteria and what is primarily relevant here is to order them with respect to the degree of their genealogical relatedness. Another possible kind of classification is the already mentioned typological one (see in particular sections 2.2.1 and 3.2.1), which is based on the descriptive categorization of languages according to the similarities they show with respect to their phonological, morphological or syntactic systems. Both these kinds of classification fundamentally rely on the comparison between languages, but they adopt different methods of investigation and they have different aims. For the moment, it is important to anticipate that “the goal of typological comparison is to determine the universal parameters along which languages vary” and “attempts to use typological comparison to establish genetic relationships simply don’t work” (Ross and Durie 1996: 5). To point out this observations is relevant here because the distinction between typological and genealogical classification in linguistics might be thought to reflect that between strictly taxonomic and “phylogenetic” classification in biology. In section 2.1.1 the hierarchical classification used by biologists has been introduced, claiming to derive from Linnaeus’ original relationship-oriented taxonomy based on the resemblance between organisms: According to it, organisms are categorized into species, genus, family, order and so on. These taxonomic ranks are not chiefly intended to include information regarding the evolutionary history of organisms, but simply to order them with respect to their likeness. Thus the aim of biological taxonomy appears to be perfectly in line with that of language typology. As language typology, this field of study primarily has the formal objective of descriptively grouping its objects of study and “does not necessarily cover evolutionary aspects” (Cavalli-Sforza *et alii* 1994: 31). Conversely, the classification referred to as genealogical in linguistics corresponds to the phylogenetic one in biology: In this case, the focus is on the enlightenment of the evolutionary history of languages, species or populations, that is on the recovery and the analysis of the genealogical paths that connect them.

In historical linguistics language families are the central units of genealogical classification: They may be defined as the main groups “of genetically related languages, that is, languages which share a linguistic kinship by virtue of having developed from a common ancestor” (Campbell 2004: 187). As it known since the first half of the XIX century (see section 2.2.1), the proto-language, or mother language, out of which the grouped languages have developed

---

[1] Haspelmath (2004b) also points out that “[a] different terminological choice is to differentiate between ‘linguogenetic’ and ‘biogenetic’ (e.g. McConvell 2001)” but he doubts that “geneticists can be persuaded to rename their science to ‘biogenetics’” (Haspelmath 2004b: 222).

(e.g. Proto-Indo-European) does not necessarily have to be known or attested: The way in which the genealogical relatedness of languages that belong to the same family can be hypothesized even in absence of an attested proto-language (and how proto-languages can be reconstructed) is briefly taken into account in section 4.2.1, whose topic is the introduction of the classical comparative method. The terminology adopted to identify family-internal genealogical units may vary greatly depending on the scholars' choices: For instance, the term "subgroup" (or "subfamily" or "branch") may be "used to refer to a group of languages within a language family which are more closely related to each other than to other languages of that family [...]" (Campbell 2004: 187), while Graffi and Scalise (2002: 57-58) assume that families are divided into groups (or classes), which, in turn, include subgroups (or branches). However the labelling of the sub-family groupings is largely conventional and actually it is not of much interest here. For the aims of the present work and, in particular, of the illustration of the relevant hypotheses provided in section 5.3 and in chapter 6, where the genealogical classification of a specific set of languages is taken into account, the terms "group" and "subgroups" are considered sufficiently clear to respectively indicate larger and smaller sub-family units, whose exact range will be clarified from time to time. Therefore, for instance, the Germanic languages (e.g. English, German and Norwegian) will be assumed to constitute a group, whose hypothesized mother-language is Proto-Germanic, while the West-Germanic varieties (e.g. English and German) will be considered a subgroup of Germanic. Observe that many terms have been adopted to refer to debated higher-order genealogical units as well, for example "stock, phylum and macrofamily" (Campbell 2004: 187) (see also the problematic definition of Nostratic in section 5.3.1) and there are also cases in which there is no agreement about the grouping in which a language should be included, in which cases the language in question (e.g. Basque) is said to be "(genealogically) isolated".

It is worth underlining that the use of genealogical classification terminology is essentially conventional in historical linguistics, in the sense that the distinction between the hierarchical positions of the various genealogical units is not based on formal criteria of any kind. This means that projecting the evolutionary history of any set of languages onto a genealogical (or, more appropriately here, phylogenetic) tree there is no formal difference between, for instance, a family and a subgroup with respect to the depth at which the internal nodes representing the respective proto-languages (e.g. Proto-Indo-European and Proto-Germanic) are placed in the structure. In other words, abstracting away from tree-graph representations, "one should not believe that the various 'mother languages' may be placed at the same 'genealogical level'" (Fanciullo 2007: 24; editor's translation). Moreover, observe that to formalize the distinction between more and less inclusive/deep genealogical units ultimately poses the same difficulties that one encounters trying to find a valid criterion to differentiate micro and macrocomparison (see section 3.1.5 and the quotations provided there), because the issue is closely bound to the impossibility of reliably quantify the degree of similarity amongst languages. Of course this does not prevent the studying of diachronic development of languages and to reconstruct the genealogical paths that bind them, but it is just a reminder of the fact that genealogical units labelled with the same name (e.g. family) may have diverse extents and refer back to non-coeval mother languages. As will be clarified in the next section, the recognition of genealogical units on phylogenetic trees is essentially a matter of interpretation and labelling of the internal nodes of the graphs.

A concluding remark that has to be pointed out before proceeding is that, besides the proper genealogical and typological criteria of classification, languages can also be grouped with respect to the similarity they show due to their "areal relatedness". The issue of areal convergence has already been covered in section 2.3.3: In particular, it has been claimed that the contact between neighbouring languages may lead them to converge with respect to some linguistic traits, even if there is not a close genealogical relationship between them. When this happens and a relevant set of features comes to be shared among a number of geographically close languages, these languages are said to constitute a "*Sprachbund*". Among the clearest examples of this unit of areal classification one may think about the Balkan *Sprachbund*, which includes many languages that will be considered later, e.g. Rumanian, Greek, Serbo-Croatian and Bulgarian. As already explained, areal convergence poses an insurmountable obstacle if one wants to model language evolution exclusively relying on tree-graphs. This is why, in order to account for both genealogical and areal relatedness modelling the evolutionary history of languages, it is necessary to overcome the constraints of tree-graphs and to adopt network-like representations. The illustration of the formal properties of both trees and networks is the theme of the next section.

4.1.2. Trees and networks

Tree-graphs have been taken into consideration many times thus far. In chapter 2 (see in particular section 2.1.2) it has been pointed out that since Darwin's original synthesis biologists represent evolutionary paths as tree structures in which all lineages are related and represented by progressively divergent branches. The same type of representation was entailed by Schleicher's (1853) model of language development and, as underlined in section 2.3.3, the debate regarding its usefulness started in the second half of the eighteenth century and it is still viable today. Following for instance Haspelmath (2004: 214) and McMahon and McMahon (2004: 18), it has been said that nothing prevents the use of tree-structures to display genealogical classification, even if it is necessary to bear in mind that these representations cannot exhaustively describe the historical development of languages. In particular, on the basis of a tree-graph one cannot account for events of contact between lineages. Not only can areal convergence not be included in trees, but its presence in the history of languages may also easily bias the accuracy of the genealogical classification. This last issue is reconsidered in section 4.3.3, where quantitative methods designed to detect non-genealogical transmission of features are illustrated. For the moment, what is relevant to consider is that, from the point of view of the representation of the evolutionary histories, in principle it is possible to model the diachronic development of languages within genealogical structures, in which the possibility of influence between different branches is not excluded *a priori*: This would entail the use of evolutionary histories that are no more strictly tree-like, but rather network-like.

In section 3.1.3 the general formalism of trees has been introduced with the objective of showing how these graphs can be employed to represent syntactic structures. In particular, the basic formal details regarding the interpretation of tree-graphs are recalled in (1) for convenience:

(1) "A simple tree *T* is a pair (*N*, *D*), where *N* is a set whose elements are called nodes and *D* is a binary relation on *N* called dominates, satisfying (a) – (c):

   d. D is a reflexive partial order (i.e. reflexive, transitive and antisymmetric, editor's note) relation on *N*.

   e. The root condition: There is a node *r* which dominates every node. […] This *r* is provably unique […] and called the *root* of *T*.

   f. Chain condition. For all nodes *x*, *y*, *z*, if *x D z* and *y D z*, then either *x D y* or *y D x*." (Keenan 2007: 51).

All the definitions provided in section 3.1.3, in particular the explanation of the terms "strictly dominate" (*SD*), "immediately dominate" (*ID*), "sister node" and "n-ary tree", are applicable here too. From them follows that a "branch", or "edge", is exactly a pair (*x*, *y*) such that either *x ID y* or *y ID x*. Moreover, *x* is an "ancestor" of *y* (and, conversely, *y* is a "descendant" of *x*) if and only if *x SD y*: For instance, referring to the graph provided in figure 1a, it is clear that

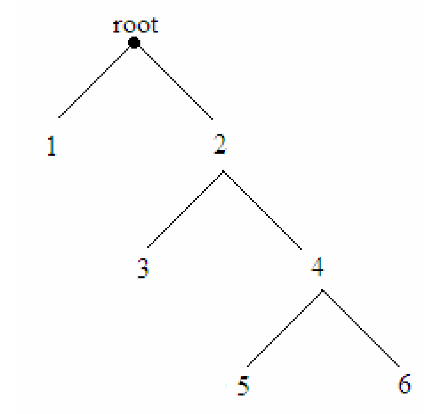


Figure 1a: Rooted tree.

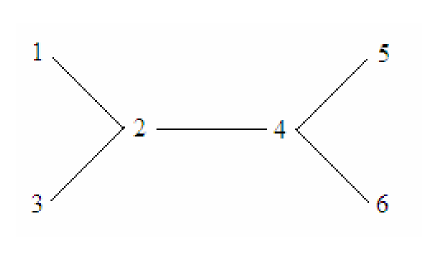


Figure 1b: Unrooted tree.

node 2 is an ancestor of nodes 4, 5 and 6, which, conversely, are all descendents of 2. x is a “leaf”, or “terminal node”, if and only if the set of nodes strictly dominated by x is empty (i.e. $\{z \in N \mid x\ SD\ z\} = 0$) (e.g. node 6 of figure 1a). Finally, nodes different from the root and the leaves are called “internal nodes” (e.g. node 4 of figure 1a). Simple trees are always rooted, i.e. provided with a root, and it is often the case that the classification they convey is provided in the so-called “Newick format” (Felsenstein 2004a: 590-591), i.e. the “well-known representation of trees in terms of parenthesis patterns which is due to the famous mathematician Arthur Cayley, and which has been around for over a century” (Felsenstein 2004b). For instance, the Newick-format representation of the tree in figure 1a is given in (2). Notice that only the *taxa* positioned in leaf nodes are included in this representation, while the labels of the internal nodes cannot be recovered from the Newick format.

(2) (1, (3, (5, 6)));

Tree-graphs can be used to represent both taxonomies and phylogenies: If the case is the former, trees are sometimes referred to as “dendrograms” (or “phenetic trees”), while in the latter case they may be called “cladograms” (or “phylogenetic trees”). Anyway, though conceptually fundamental (see the discussion in section 4.1.1), this distinction becomes less relevant within the “tree-reconstruction practice”, so that “the use of these terms (dendrograms and cladograms, editor’s note) may be deceptive for it reflects more often the aims and hopes of the research worker than the result achieved” (Cavalli-Sforza *et alii* 1994: 31). Thus this issue is left aside here and from now on the interpretation of tree-graphs is implicitly assumed to be the phylogenetic one, which is straightforward. Adopting the term “*taxon*” or “taxonomic unit” to refer to any kind of object or unit that is classified (i.e. a language, a species or a population), the root represents the common ancestor of all *taxa* that lie in the leaves (e.g. languages), while branches stand as paths of genealogical development and internal nodes as intermediate proto-*taxa* (e.g. proto-languages), or alternatively as points of genealogical divergence, e.g., referring to biological evolution, as past speciation events (see for instance Semple and Steel 2003: 19-20).

Thus here the tree-graphs that are relevant to represent genealogical classification are rooted binary trees, which, as shown throughout section 4.3, are the most commonly used phylogenetic structures in quantitative phylogenetics. Nonetheless these are not the only relevant ones: Unrooted binary trees (i.e. binary trees not provided with a root), like that in figure 1b, are important as well, above all because they usually represent the original output worked out by many quantitative methods of phylogenetic reconstruction (see section 4.3.2). Nonetheless the information that can be extracted from unrooted trees is not properly genealogical, in the sense that in these structures the direction in which development occurred is not specified throughout the structure, i.e. dominance relations cannot be determined at all: In other words, since there is no root, there is no "starting point" that determines the origin of the dominance relations, which means that in unrooted trees "there is no arrow of time" (Dawkins and Wong 2005: 4). Take for instance the unrooted tree in figure 1b: One cannot decide from this graph if 2 *ID* 4 or the other way around. Unrooted trees should then be intended to map relations of similarity between *taxa*, rather than their actual genealogical relatedness. However, it is always possible to provide a root to an unrooted tree. Among the diverse methods that is possible to employ in this aim (see section 4.3.1 and Felsenstein 2004a: 6-8), it is worth briefly anticipating the mention of the so-called "outgroup criterion". The outgroup consists in a *taxon*, which is known (or supposed) to be monophyletic, i.e. to have undergone an independent evolution with respect to the rest of the *taxa* included in the tree. In figure 2 a graphical explanation of the procedure is provided: In the first case, the node 2 (or, alternatively, node 4) is identified as monophyletic, while in the second 5 is chosen. It is clear that the results of rooting a tree with the outgroup criterion can be genealogically very different, thus one has to rely on solid external evidence to do it.

As for network-like graphs, it has already been pointed out that they can help to overcome the constraints posed by trees representing the actual history undergone by languages. In particular, these graphs can be used to model the diachronic development of languages representing not only their "regular" genealogical evolution, but also events of areal convergence, i.e. contact. Using the terminology introduced in the end of section 2.3.2 in Wang and Minett's (2005) sense, networks can be used to display both vertical and horizontal patterns of language variation. Recall that the problem of non-genealogical transmission of traits is (still) considered rather marginal in evolutionary biology but it is typically ascribed to cultural evolution in general, thus, as it is claimed in section 4.3.1, within humanities linguistics is not the only field of study in which scholars are interested in phylogenetic reconstruction and interference events between lineages have to be accounted for. Also philologists have recently adopted phylogenetic methods to classify manuscript traditions and in philology the issue of horizontal transmission is known as the problem of "contamination" (e.g. Windram *et alii* 2006). Therefore "[…] in the study of linguistic and textual history the fact that networks are favoured over trees seems to have become more and more accepted […]" (Stolz 2006: 210).

Networks differ from trees in that they violate the chain condition. For instance, in the structure presented in figure 3, both nodes 3 and 4 dominate node 5, but neither 3 nor 4 dominates the other. As for the phylogenetic interpretation of the graph in figure 3, the information contained in the network is the following: The ancestors of the node 5 may be thought to be both 3 and 4, so that, supposing that the branch (4, 5) is chosen to represent the vertical transmission and the branch (3, 5) the horizontal transmission of features, the proper genealogical ancestor of 5 is 4. Nonetheless 5 also retains a significant amount of features which do not derive from genealogical inheritance, but rather from a non-genealogical transmission from 3. In other words, in figure 3 may be encoded a situation in which 5 is a genealogical descendent of 4, but at the same time its history cannot be fully accounted for without supposing that a contact occurred between its lineage (4, 5) and that of *taxon* 3: The event is represented by the branch (3, 5), i.e. the branch (3, 5) is a "contact" edge drawn on the underlying phylogenetic tree.

Until here the focus of the discussion has been the explanation of the aim of the present investigation, i.e. how genealogical classification can be represented using trees and how non-genealogical paths of transmission can be accounted for using networks. The next point to illustrate is the methods by means of which the above mentioned objective can be pursued:

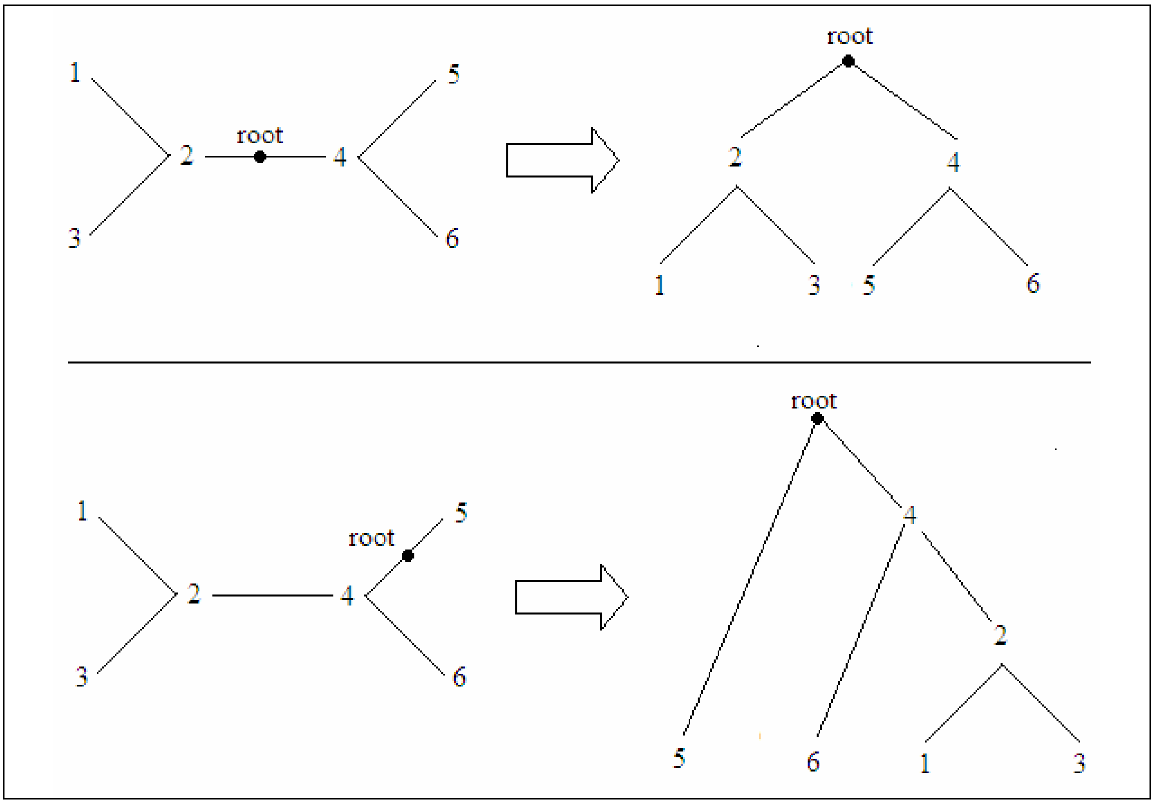


Figure 2: Two possible ways of rooting an unrooted tree by means of the outgroup criterion.

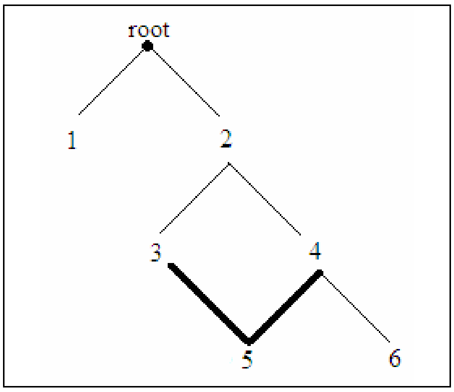


Figure 3: A network, which violates the chain condition.

This is done starting from section 4.2, which is dedicated to the presentation of some relevant comparative methods.

4.2. Comparative methods in linguistics and biology

In this section diverse comparative methods adopted in linguistics and biology are illustrated and critically reviewed. In historical linguistics the term “comparative method” without further specifications commonly refers to the “classical comparative method”, which “is typically seen as the gold standard by comparative historical linguists” and “has been under development for more than a hundred years [...]” (McMahon and McMahon 2005: 5).

Observe that the reliability and the usefulness of the classical comparative method in pursuing its aims within the limits of its effectiveness are not under discussion here. What will be suggested in the course of the discussion is that a conviction like Harrison’s (2003: 213) one, according to which this method is “the only tool available to us for determining genetic relatedness amongst languages", should probably be relaxed. Indeed, even if the scientific rigor and the consistency of other methods proposed in the second half of the XX century (i.e. lexicostatistic analysis and mass comparison), have been demonstrated to be far from those of the classical one (see section 4.2.2), an alternative view that has been developed at least since Longobardi (2003a) appears to represent an innovative way to shed light on language evolution: The "parametric comparison method". This recent approach is essentially grounded on the extension of the biolinguistic paradigm of generative grammar to the diachronic dimension, which has been covered in chapter 3; More precisely, it moves from a fundamental shift in the choice of the *comparanda*, which closely recalls that which occurred in biology thanks to population genetics in the '50s. In section 4.2.3 it will be argued that the parametric comparison properly allows to adopt quantitative phylogenetic methods conceived to be used in molecular biology with the aim of studying language evolution. The introduction of these last techniques is the focus of section 4.3, while the survey presented in this part of the dissertation begins to take into consideration the classical comparative method.

4.2.1. The classical comparative method

According to Ross and Durie (1996: 3), the classical comparative method (henceforth in this section simply "comparative method") is usually intended as the method itself and its associated theory, which is the "Neogrammarian Hypothesis" outlined at least since Osthoff and Brugmann (1878)[2]. Thus Ross and Durie (1996) discern between what they refer to as the comparative method in the "narrow sense" (the comparative procedure on its own) and in the "wide sense" (the method plus the underlying theory). If no further specifications are provided, following the mainstream literature here the discussion is largely focused on the latter definition.

One may assume that the comparative method "has two indivisible parts: the demonstration of linguistic relatedness, and the reconstruction of a hypothetical common ancestral system" (McMahon and McMahon 2005: 5). Among others, Thomason and Kaufmann (1988: 201-202) and Harrison (2003: 215) argue along the same lines. For instance, Harrison (2003) claims that "[t]he primary role of the comparative method is in developing and testing hypotheses regarding genetic relatedness" and "[i]ts secondary, and subsequent, role […] is in recovering antecedent language states through reconstruction" (Harrison 2003: 215). As a first approximation these statements might be considered suitably correct to introduce the issue. However Nichols (1996), reviewing the theory and the practice adopted by comparativists until the XIX century (in particular that of Meillet, see for instance Meillet 1958: 88-97), shows how it is possible to more accurately analyze and understand the comparative procedure. Before introducing her approach and the relevant theoretical background, it is worth underlining a general observation. At all stages of its application, the comparative method relies on a basic attribute of language: The arbitrariness of linguistic signs, i.e. the arbitrary form-meaning combination[3]. This means that "two different languages are unlikely

---

[2] This claim should not lead one to think that only after the Neogrammarian school a method of systematic comparison aimed at shedding light on the history of languages began to be carried out. As commonly known, systematic comparisons started to be performed since the first half of the XIX century by scholars like Rasmus Rask, Franz Bopp, Jakob Grimm and August Schleicher (e.g. Belardi 2001: 281-298). Nonetheless the "Neogrammarian hypothesis" represents the fundamental conceptual shift thanks to which the comparative method could be theoretically grounded and scientifically formalized.

[3] Of course the arbitrariness of linguistic signs does not concern specific classes of words and expressions. Thus certain kinds of paradigms cannot be used as valid individual-identifying evidence, using Nichols' (1996) terminology: For instance, personal pronouns are "a good example of a systematically structured and phonologically filled lexical field", but it can be shown that "forms of

to have by chance more than a minuscule percentage of (nonmarginal) morphemes which match each other in form and meaning” (Ross and Durie 1996: 6) and that when such “matching” is found, it may be plausibly drawn to the retention of a trait that revels common ancestry of the languages in question. Though the comparisons underlying the identification of “matching” “do not have scientific value if one cannot demonstrate that they are not accidental, but are ordered into schemata that are outlined on the basis of the comparisons themselves” (Tagliavini 1999: 5; editor’s translation). This preliminary explanation has to be considered as a rough indication of the overall *ratio* of the procedure, whose methodological and theoretical details are taken into account in the following discussion.

The core of the comparative method relies on the Neogrammarian hypothesis of the “regularity of sound change", i.e. on the assumption that “if sounds of a language change over time, they do so in a largely regular manner […], such that sound *x* becomes sound *y* under statable conditions not just in some morphemes but in all morphemes that meet these conditions” (Ross and Durie 1996: 6). The “conditions” the scholars refer to are the phonetic contexts in which the sounds occur. The expectation that follows from this hypothesis is to find traces of the regular development of sounds in “genealogically related” languages, i.e. to identify a systematic correspondence throughout lexicon and morphology between the certain sounds of one language and the sounds of another language in the same contexts. Therefore, in essence the idea underlying the comparative method is that when comparing a series of languages these systematic phonetic correspondences are found, they provide strong evidence in favour of the genealogical relatedness between the compared languages. Take for instance the series of correspondence in (3a-d)[4].

(3) a. ClG *‘έξ* – Lat *sex* – OHG *sehs* – Scr *ṣáṣ* (“six”)
b. ClG *‘επτά* – Lat *septem* – Got *sibun* – Scr *saptá* (“seven”)
c. ClG *‘αλς* – Lat *sāl* – Got *sal-t* – OS *solĭ* (“salt”)
d. ClG *‘ημι* – Lat *sēmi* – Scr *sāmi* (“half”)
(adapted from Fanciullo 2007: 73)

What should be noticed is that while ClG has [h] followed by a vowel at the beginning of all the words in (3a-d) (i.e. in the phonetic context noted as #_V), in the same context the other languages included in the series have [s]. As well as shedding light on genealogical relatedness, this observation allows making hypotheses about the original Proto-Indo-European form, i.e. to reconstruct it. Something more about this latter point is said below. For the moment notice that examples of systematic correspondences of this kind may be found in any introductory (and also advanced) book of historical linguistics, in particular for what concerns the systematic comparison of Indo-European languages, even if the same method of analysis has been fruitfully applied to non-Indo-European languages as well (see in particular the collection of articles introduced by Baldi 1990). Summing up, the point is that systematic phonetic correspondences are extremely unlikely (say, almost impossibly, see below for some probability estimates) due to chance, so that they require an historical explanation. However, according to Nichols (1996) “demonstration of relatedness through systematic correspondences in vocabulary is not the operating procedure for the classic application of the

---

first and second person, and of singular and plural numbers, are not independent” because “the relation of paradigmaticity to coding phonological form is not arbitrary” (Nichols 1996: 54). A similar observation also applies to the choice of the items to include in the successive stages of the comparative method, in particular in that of the identification of systematic sound correspondences: In this case, at least onomatopoetic and nursery words must be excluded (see among others Ringe 1990: 3; McMahon and McMahon 2005: 15).

[4] List of the abbreviations used in (3) and (4): classical Greek (ClG), Latin (Lat), Old High German (OHG), Sanskrit (Scr), Gothic (Got), Old Slavic (OS), German (D).

comparative method to the Indo-European languages going back to the late eighteenth century" (Nichols 1996: 41). In other words, in the scholar's opinion working out systematic correspondences in the lexicon represents an advanced stage of application of the comparative method, that is usually performed only after one has already assumed genealogical relatedness.

Nichols (1996) argues that actually the first step of the comparative method consists of an heuristic component, which would allow linguists to establish that a significant genealogical relationship exists. This part of the procedure entails collecting "diagnostic linguistic evidence", that "is primarily grammatical and includes morphological material with complex paradigmatic and syntagmatic organization" (Nichols 1996: 41). More precisely, the scholar refers to diagnostic linguistic evidence as "individual-identifying evidence", because it permits determining the existence of a unique individual proto-langue[5]. How valid are the forms or the paradigms in question as pieces of diagnostic evidence? In other words, is it possible to calculate a statistical threshold for individual-identifying, according to which one can decide whether the evidence can be largely (or exclusively) explained as proof of genealogical relatedness and not in terms of chance? For the reasons explained in Nichols (1996: 49), the scholar assumes that "a probability of occurrence (of, for instance, a certain morphological paradigm, editor's note) of one in a hundred thousand or less is individual identifying at a statistically significant level, and a probability of one in ten thousand is at least interesting and borderline useful" (Nichols 1996: 49). Among the other cases, the author takes into account the occurrence of the miniparadigm "good – better – best", which is found in English and its sister Germanic languages, e.g. German (*gut – besser – best*), Dutch (*goed – beter – best*), Swedish (*god – bättre – bäst*), Norwegian (*god – bedre – bedst*) and so on. The case has already been taken into consideration by Newman (1980) as diagnostic evidence of genealogical relatedness. Nichols (1996) assumes that, considering that the greater part of Indo-European languages have an inventory of around 20 consonants and 5 vowels, the probability that in a phoneme sequence a certain consonant occurs in a particular position is 0.05, while the probability to find a certain vowel in a specific position is 0.2. Thus the scholar takes into account the two roots that appear in the paradigm, i.e. that of the positive ("good") and that of the comparative/superlative ("bett-"). She claims that, comparing, say, English and Dutch, the probability that in German by chance the initial sound of the word for "good", i.e. *goed*, is exactly [g] is 0.05. Nichols (1996) follows this line of reasoning for all the other occurrences, which, she points out, are independent one from the other, in the sense that the occurrence of [g] in the initial position should not influence the occurrence of a specific vowel in the second position. Thus she concludes that "each of the roots 'good' and 'bett-' has chance of 0.05 × 0.2 × 0.05", i.e. of 0.0005, that is still not sufficient to overcome the threshold of 0.000 001. Nonetheless, the probability that the entire paradigm occurs corresponds to the chance of the positive form multiplied for that of the comparative/superlative one, that is to 0.0005 × 0.0005 = 0.000 000 125, and this probability largely satisfies the threshold requirement. Of course such a result has to be intended as an indicative estimate, especially because the probabilities of the single occurrences are given on the basis of an approximate calculation of the number of consonants and vowels that are included in the different phonemic inventories. If one considers each language in detail, the total number of segments may significantly deviate: For instance, Dutch has an inventory of 33 phonemes, that is 21 consonants and 12 vowels Maddieson (1984). Thus a more realistic assessment could be made on the basis of the analysis of a wider sample of languages, exactly as that pursued by Maddieson (1984), who concludes that the total number of consonants in an average phoneme inventory is 22.8, while for vowels this value should be around 8.7

---

[5] Individual-identifying evidence is contrasted with type-identifying evidence, which "include features found in enough unrelated language families that comparative structural linguistics views them as typological" (Nichols 1996: 48).

(Maddieson 1984: 9; see also section 2.3.2): In this view the probability of single occurrences of a certain consonant would change into 0.043 and that of a certain vowel into 0.115[6].
Apart from this observation, Nichols' (1996) idea is that before applying the comparative method throughout the lexicon, one needs to find relevant pieces of individual-identifying evidence, i.e. to demonstrate that the languages under consideration are genealogically related by means of the comparison of organized paradigmatic sets of elements. In this sense, above all individual identifying is, for instance, the common retention of irregular morphological paradigms, which are relatively rare and distinctive in a language system, because "they are not generally introduced in the process of language change" (Baldi 1990: 4). The latter claim will become clearer after the introduction of the role of "analogy" in language change, which is covered below. For the moment, observe that Nichol's (1996) idea is caught by other scholars too, for instance by McMahon and McMahon (2005), who claim that in the application of the comparative method "[…] there is an initial stage of working out whether certain languages are likely to be related, and in the development of the method on Indo-European this initially involved the observation of morphological similarities" McMahon and McMahon (2005: 5). Thus the individual-identifying principle appears to formalize a practice that "continued to be asserted in general and methodological pronouncements of later Indo-Europeanists" (Nichols 1996: 47; see in particular the quotation she provides from Meillet 1958: 91) and that can be applied to the study of new genealogical domains as well (e.g. see the analysis of sub-grouping in Tibeto-Burman in LaPolla 2000).

In light of the arguments outlined so far, the identification of systematic correspondences in "cognate lexical items (or roots)" should then be considered the second step of the comparative method. The term "cognate" is intended as "a word (or morpheme) which is related to a word (morpheme) in sister languages by reason of these forms having been inherited […] from a common word (morpheme) of the proto-language from which the sister languages descend" (Campbell 2004: 126). Then after having collected a set of supposed cognates in the languages under consideration, the comparativist has to identify sets of systematic correspondences of the type presented above in (3). At this point a question that has been left aside so far becomes evident, i.e. the fact that not only the form of the words/morphemes, but also their meanings seem to play an important role in the comparison. Actually, the semantic "similarity" of cognates poses a problem, because the evolution of lexical semantics appears to be hardly predictable, especially because it is particularly sensitive to factors external to language, e.g. specific social and historical contexts of use (see section 2.3.2). That is why Nichols (1996: 58) claims that semantics can be largely ignored by the comparative method and it is accepted that cognate sets comprise "considerable semantic latitude in root morphemes". Therefore, the correspondence of (3) is assumed to be satisfactorily instantiated in (4) too, where in addition it is also shown how it is found in the context #_r.

(4) ClG *ʻρέω* ("to flow") – Scr *sráv-ati* ("he flows") – OS *o-strov-ŭ* ("in the middle of currents, i.e. island") – D *Strom* ("large river, current")
(adapted from Fanciullo 2007: 73-74)

Nonetheless, relevant research has been done to better understand the dynamics of semantic shifts: Apart from the works already mentioned in section 2.3.2, Ross (1996) and Wilkins (1996) cover the issue exactly with the aim of a better accounting for it within the

---

[6] Anyway, more accurate and extensive statistical calculations regarding both the classical comparative method and mass comparison are provided, for instance, in Ringe (1992; 1996), while McMahon and McMahon (2005: 54-68) offer a guide to probability theory applied to language comparison.

comparative framework. Nichols' (2006) comments of about the degrees of freedom that can be accepted comparing meanings in different languages are recalled also in section 5.1.1 dealing with lexicostatistic comparison.

Besides corroborating the hypotheses of genealogical relatedness amongst the languages in question, the work that is done on cognate sets and systematic correspondences in the lexicon might be intended to have essentially a twofold aim: On the one hand it would help to draw a more detailed picture of the internal sub-groupings of the already identified genealogical units (Nichols 1996: 41); On the other hand it would represent a fundamental way to carry out the reconstruction of the assumed proto-language, i.e. to recovery the ancestral forms. As already pointed out, the extensive illustration of how comparative reconstruction of proto-languages is performed in historical linguistics is beyond the aims of the present work[7] (on this issue see among others the introductory notions provided in Baldi 1990; Fox 1995; Campanile 1996: 115-133; Campbell 2004: 122-167). Though it is worth mentioning how the method would proceed in a case like that of the systematic correspondence given in (3) and (4), where three scenarios might be hypothesised. One may assume that in the context #_V/r proto-Indo-European had either [h] like ClG, or [s] like the majority of the other languages, or another sound that has not been preserved in daughter languages. The first hypothesis entails admitting a passage from proto-Indo-European [h] to [s] in all the languages under discussion except from ClG, while the second scenario would lead to reconstruct [s] in proto-Indo-European, which changed into [h] in ClG and was preserved in the other languages. Since it is known that the change [s] > [h] is very common and that [h] > [s] is almost unattested, the second solution appears to be preferable. Furthermore, since this reconstruction of [s] is sufficiently plausible, it is not necessary to hypothesis a more complex situation like that involved by the third hypothesis (Fanciullo 2007: 74-75). Notice that this reconstruction is also in agreement with another important criterion, i.e. the "majority rule", according to which should be favoured the proto-form which is identical to the form instantiated in the majority of the daughter languages in question: This rule is related to the parsimony of the reconstruction (which will be reconsidered within the framework of quantitative phylogenetics as well, see section 4.3.2), i.e. to the reasonable idea that the fewer are the changes one has to postulate, the better it is.

However, even if the Neogrammarian assumption applies to a number of relevant cases, the regularity requirement of sound change seems to be not always satisfied and exceptions to the "exceptionless" phonetic rules, i.e. irregularity, often arises. In the Neogrammarian perspective all apparent contraventions to rules, i.e. all changes that cannot be explained resorting neither to a single rule nor to the interference between different rules, should be drawn back to secondary causes that impede the perfect regularity of sound change to instantiate: Analogy and borrowing. This idea entails assuming that *per se* sound change is regular, but its linear occurrence may be biased by the intervention either of a different kind of language-internal regularizing tendency, i.e. analogy (see also chapter 3, note 18) or of a language-external phenomenon, i.e. borrowing: Examples of these processes are provided for instance by Fanciullo (2007: 153-156). What should be noticed here is that even in the recent literature a similar, even if more extensive, account for irregularity is provided. Ross and Durie (1996: 13-14) observe that the dawn of the sociolinguistic paradigm in the '60s (Labov 1966; Weinreich *et alii* 1968) paved the way for the acceptance of speaker-oriented explanations for irregularity, such as those connected, for instance, to speech articulation (Blust 1996), politeness (Newman 1996) or, again, contact (Campbell 1996; Ross 1996). In this way "[c]omparativists find themselves moving from the quasi-Darwinian Neogrammarian

---

[7] All the more so is the illustration of internal reconstruction, the procedure initiated by De Saussure (1879) and meant to investigate the history of a language exclusively relying on the evidence provided by that specific language alone (see for instance Fox 1995: 145-216; Campbell 2004: 225-251).

paradigm which underlies the regularity hypothesis to a speaker-oriented paradigm which has a place for both regularity and irregularity" (Ross and Durie 1996: 13).

Until here it has been shown that the comparative method is primarily a formal procedure of investigation focused on the identification of systematic sound correspondences, but that, especially when it comes to deal with cognate sets, it cannot abstract away from considering the meaning, as well as the form, of its units of comparison. Thus, due to the problems related to diachronic semantics, the lexical-oriented part of the comparative procedure appears to be not entirely safe against the risk of providing imprecise measures of similarity between languages. Notice that, summarizing the factors that constrain the application of the technique, McMahon and McMahon (2005) claim that exactly due to this problem, as well as to the interference of contact phenomena and to cases such as those mentioned in note 3, "not all linguistic material is suitable as input to the comparative method" (McMahon and McMahon 2005: 14). This statement is mainly related to the selection of cognate sets, which will be reconsidered briefly in the next section and, more specifically, in section 5.1. However, given an accurate choice of the *comparanda*, the consistency of the comparative method is not under discussion, in that, it "has the major epistemological advantage of providing a sharp and much-needed demarcation criterion between science and pseudo-science in etymology and historical linguistics" (Longobardi and Guardiano 2009: 3).

As has been shown, the method is grounded on the detection of extremely improbable facts, i.e. individual-identifying evidence and systematic sound correspondences, which are due almost impossibly to chance and have a strong claim for historical explanations. However, such fundamental evidence appears to be quite rare when one comes to compare languages that are not closely genealogically related. This fact represents an inconvenience for the universal applicability of the procedure and it is connected to the fact that the comparative method is "intrinsically limited in the time depth over which it can operate" (McMahon and McMahon 2005: 17). This non-immediately understandable comment may become clearer if one considers that "[w]e cannot recover things about the proto-langue via the comparative method if the daughters simply do not preserve evidence of them" (Campbell 2004: 166). More precisely, Ross and Durie (1996) affirm that "[t]he comparative method relies on the retention of archaic features in daughter languages" and "[w]hen something is independently lost in all the attested languages of a family, there can be no way of reconstructing it" (Ross and Durie 1996: 28). This may happen not only as a consequence of borrowing and analogy, but also, for instance, when successive changes obscure the context in which previous changes occurred, so that it is impossible to recover regularity. Harrison (2003: 230) clearly states that when "too much" time has passed, there may be no more available evidence to apply the comparative method, so that the limit of its usability is not intrinsically temporal, but rather practical and statistical. Thus the method can hardly provide consistent insights within the so-called long-distance (or long-range) comparison, i.e. comparison between languages and families that may be hypothesized to be "distantly" genealogically related and to have undergone 6 or more millennia of separation: In these cases, extensive evidence of regular sound correspondences can no longer be found (Longobardi 2003a: 120).

To overcome the limitations of the classical comparative method, especially that concerning its time-depth effectiveness, other comparative procedures have been proposed. The illustration of the most debated ones, i.e. lexicostatistics analysis and mass comparison, is covered in the next section.

4.2.2. Lexicostatistics and mass comparison

Among others, Ross and Durie (1996: 4-5) underline that the classical comparative method should be kept distinct from lexicostatistics and mass (or multilateral) comparison. Both these methods have been subject, to different extent, to hard criticism and, since they move from apparently similar premises, they are covered together in this section.

The classical lexicostatistical paradigm was explicitly designed since the '50s by Morris Swadesh[8] in a series of works focused on American Indian languages and soon started to be confused with one of its specific applications, that is glottochronology. The distinction between the main method (lexicostatistics) and its extension (glottochronology) has been neglected by many scholars especially in the '60s (Fox 1995: 279-280), but this overlapping is actually misleading and "leads to confusion – and often debate at cross-purposes – over the goals or achievements of the particular method used in a particular piece of research" (Embleton 2000: 160). However, if in Swadesh (1952) one may find one of the earliest mention of the methodological premises of the comparative procedure, it is in Lees (1953) that its statistical basis started being worked out with particular reference to glottochronology.

In general terms, lexicostatistics may be defined as "the use of standard meaning lists to assess degrees of relatedness among languages" (McMahon and McMahon 2005: 33). So the first step of the method consists of defining an "appropriate" word list, or, more precisely, a list of meanings, which are translated in all the languages in question: The lists of each pair of languages are then evaluated and "scored for cognates and non-cognates" (Swadesh 1955: 122). In this way it would be possible to determine the percentage of cognates that are shared by each pair and provide a measure for the genealogical relatedness. It is clear that so far two points require attention: The choice of the meanings to include in the list and the criterion by means of which cognacy is established. The former issue is specifically taken into account in Swadesh (1952), where the scholar argues that "[s]uitable items for a test list must be universal and non-cultural" and they must be easily identifiable broad concepts, which can be matched with simple terms in most languages" (Swadesh 1952: 457). As is well-known, this entails that the meanings included in the list should be those belonging to the "core vocabulary" of languages, for instance those related to family relations (e.g. "father", "mother"), basic numerals, body parts (e.g. "hand", "nose") and so on (see also the examples provided in section 5.1.1). This choice is motivated by the fact that the words expressing these meanings are expected not only to be relatively safe from "non-universality, duplication, ambiguity, sound imitation" (Swadesh 1955: 125), but also to have the same likelihood to persist (the importance of this feature is clarified later) and to be more resistant to borrowing, which would represent the most dangerous bias in the calculation of the cognate percentages. The original formulation of the so-called Swadesh list comprises 215 meanings, 15 of which were already recommended for discard in Swadesh (1952) but were used by Lees (1953) as well. The list proposed since Swadesh (1955), instead, was reduced in order to more precisely follow the methodological premises and consists of 100 items[9]. Therefore, basically the calculation of the cognate percentages is primarily intended to measure the genealogical proximity between languages, in the sense that the more two languages share a common "derivation" of their vocabulary, the closer they are thought to be related. Essentially, also for lexicostatistics cognacy between words has to be determined on the basis of systematic sound correspondences of CVC roots (see for instance the summary provided in Fodor 1982: 65-66), so that the method can be applied only after the word lists have been analysed by means of a classical comparative procedure and the relevant languages have been shown to be genealogically related (Trask 1996: 362). As a consequence, according to McMahon and McMahon (2005: 35), it is incorrect to think that lexicostatistics relies on the identification of cognate words based on superficial form-meaning matching. This misunderstanding may lead to associate this method with mass comparison, which is substantially mistaken.

---

[8] See Embleton (2000) for references to works that may have predated the overt introduction of lexicostatistics.

[9] More precisely, to draw the 100-item list the scholar cancelled 108 meanings and added 8 more, i.e. *say*, *moon*, *round*, *full*, *knee*, *claw*, *horn* and *breast*.

Glottochronology represents the attempt to extend the use of lexicostatistics to the estimation of divergence times, i.e. to the use of cognate percentages to determine the time scale of the hypothesized genealogical relationships. In order to reconstruct dates Swadesh (1950) proposes the formula in (5).

(5) $$t = \frac{\log C}{2 \log r}$$

In (5) t is the elapsed time in millennia, which is derived from C, the percentage of shared cognates, and r, the "glottochronological constant" that expresses the rate of word retention in 1000 years. Of course the element that deserves particular attention here is the latter, in that it represents the strongest assumption of the method, i.e. that the rate of retention (or, conversely, change) is constant in all time scales for all languages: In this sense the choice of a word-list, in which all items may be thought to have the same likelihood to persist, is fundamental. Lees (1953) reconsiders the equation in (5) within a broader statistical account and provides a step-by-step introduction to the passages that lead to its formulation (Lees 1953: 113-117). In particular, he shows how r can be calculated considering 13 pairs of control languages, each one including an attested mother language and one of its daughters (e.g. Old English and Modern English, Classical Latin and Modern Catalan and so on). Given the percentage of shared cognates between each of these pairs and the separation time, the pair-specific r results as the proportion of vocabulary items that are retained in the relevant lineage in 1000 years. For instance, according to the scholar in the case of Old English (AD 900-1000) and English C is 76.6 and t is 1, so the pair-specific r is exactly 76.6; As for Classical Latin (200 BC) and Modern Catalan, instead, C should be 60.6 and t to 2.15, therefore the pair-specific r would correspond to 72.9. The overall r is then the average of all the 13 pair-specific rs: Using the 215-item list Lees (1953) derives a value of 81%, while on the basis of the 100-item core vocabulary Swadesh (1955) recalculates r to 86%. Actually "this in itself might raise initial doubts over the method, since we are meant to be dealing with a constant, which turns out to be a variable" (McMahon and McMahon 2005: 182).

The main criticism of lexicostatistics has focused both on the possibility of defining a general word-list and on the attempt to estimate dates. The objections related to the former point are self-evident if one only considers the reformulation of the core vocabulary that was carried out by Swadesh himself in the early '50s and the difficulties concerning the translation of all the relevant meanings in all languages. Indeed there are many cases in which the same meaning have different possible translations in a language or some words cannot be supplied (Campbell 2004: 204-207): When the latter is the case, the cognate percentages for the pair of languages in question are not calculated on, say, 100 vocabulary items, but on the basis of the subset of items for which the translation is provided in both languages. Thus while since the beginning of the lexicostatistic enterprise some scholars rejected in general the idea of exclusively grounding the comparison on the lexicon, claiming that lexicostatisticians have "mistaken the dictionary for the language" (Teeter 1963: 644), others were doubtful whether a universal word list was achievable (e.g. Hoijer 1956; Gudschinsky 1956). As for glottochronology, the critics' attacks mainly regarded the assumtion that the retention rate of lexicon is the same at all times for all languages. In general, one may observe that hypothesizing the existence of a glottochronological constant is very hazardous, because "every language is subject to different influences and pressures, to which its vocabulary will respond in different ways" (Fox 1995: 289). Indeed, solving (5) for other pairs of coeval languages the divergence times appear to be unreliable in many cases: For example the separation date between French and Italian is too late and would result to be AD 1586, instead of being around the V century AD. Fox (1995: 289) offers a wide overview of the works that confirm the inadequateness of this dating technique. Anyway, among the clearest demonstrations of the inapplicability of a universal retention rate one may recall the work of Bergsland and Vogt (1962), who "lexicostatistically" proved that there are languages (e.g. Modern Icelandic and Modern Armenian) in which the conservativeness of lexicon is extremely high compared with the supposed average.

Therefore, provided the maintenance of a classical comparative methodology to determine cognacy, the main problem of lexicostatistics *per se* appears to be the proper choice of the *comparanda*. To what extent the limited number of lexical items included in a basic vocabulary can be assumed to be safe from, say, borrowing and semantic misinterpretations? This problem appears to be hardly solvable. Furthermore, looking at the way in which lexical comparison has been performed after the '60s, it seems that since Swadesh' lists little progress has been made on this point. For instance, the "comparative Indo-European data corpus" given in Dyen *et alii* (1992) essentially comprises the same 200-item word-list proposed by Swadesh (1952): In this version, the meanings are translated into 95 languages from the Indo-European family and all cognate pairs are identified. Observe that in the last years Dyen *et alii*'s (1992) corpus has become the standard dataset for lexically-based quantitative phylogenetic studies, as will be explained in particular in section 5.1.1: However, it is worth anticipating that it is often the case that these works only share this (fundamental) starting point with lexicostatistics, since in general they use much more sophisticated methods of analysis.

The first mention of "mass comparison", later referred to also as "multilateral comparison", can be found in Greenberg (1954: 406-408). A concise account for the method and for the objections that have been raised against it is provided below. In the meanwhile, notice that mass comparison allowed Greenberg (1963) to order African languages into four main families (Afroasiatic, Khoisan, Niger-Kordofanian, and Nilo-Saharan): In the course of time the majority of the scientific community accepted this specific classification and only sporadic occasional rejections persisted (e.g. Fodor 1982). The "Indo-Pacific hypothesis" that Greenberg put forward in the '70s (Greenberg 1971) got cold reception instead: On the basis of his mass comparison methodology, he suggested classifying the relevant languages into 14 groups with different internal sub-groups, but in a few years his hypothesis came to be almost entirely disregarded. Finally, as is well-known, the most debated application of Greenberg's technique concerns the classification of the languages of the Americas (Greenberg 1987). The scholar's hypothesis consisted of ordering all the languages in question into three families: The Amerind (comprising 11 groups), the Eskimo-Aleut and the Na-Dene. Even if flanked by some evidence collected within anthropometrics and population genetics (Greenberg *et alii* 1986; Cavalli-Sforza *et alii* 1988), Greenberg's proposal gave rise to a great controversy, not only about the classification itself, but also about the method of comparison in general, and the debate was still in progress when he tried to apply it to Eurasia (Greenberg 2000).

Since the first formulation, the guideline of mass comparison has been to take into account a wide set of languages belonging to the same linguistic area and then to compare them simultaneously with respect to a large number of "grammatical" and lexical forms. "Grammatical forms" is meant to indicate, for instance, morphological endings and pronominal paradigms, but, as demonstrated by McMahon and McMahon (1995: 210), these features appear to have a very marginal role in the classification of Greenberg (1987)[10], thus one can leave them aside and assume that actually the comparison is carried out exclusively on lexical items. As anticipated, mass comparison fundamentally differs from lexicostatistics

---

[10] For instance, among the grammatical features that according to Greenberg (1987) would provide a common match among all the languages included in the 11 Amerind subgroups there are the first person ending *–n* and the second person ending *–m*. However, as pointed out by Nichols (1996: 54) and recalled here in note 3, these forms are not independent and, moreover, they are very frequent worldwide (McMahon and McMahon 2005: 23).

with respect to the criterion of comparison, that is the way in which a significant relation is determined between the word-lists of different languages. According to Greenberg (1987: 1-37), the norm to recognize relatedness between words should be their "resemblance" in form and meaning: If word *A* of language *L1* is "similar" to word *B* of language *L2*, then the genealogical-relatedness score between *L1* and *L2* gains a point, otherwise not. How can "resemblance", i.e. a significant match between two words, be established? Greenberg (1987) does not provide cues in this sense (Fodor 1982: 71), in that he believes that similarities should be self-evident. "Is a form *A* more like *B* or like *C*? Given, for example, *pan*/*fan*/*ezuk*, who would hesitate?" (Greenberg 1987: 5). Of course phonological and semantic resemblance can only represent a vague indicator of shared ancestry, i.e. of the fact that two words may descend from the same proto-form. Among others, Wright (1991) notices that adopting a resemblance-based method to identify matches between lexical items the evaluation cannot be scientifically replicated, in that there is no objective principle one can refer to, particularly because not all linguists own "intuitive sense for linguistic affinity" (Wright 1991: 58). However, Greenberg's idea is that "the number of languages considered should probabilistically compensate for the lower precision of the compared entities and protect against the risk of chance similarity" (Longobardi 2003a: 120). Ruhlen (1994), a supporter of mass comparison, argues that, having two languages with a phoneme inventory of, say, 7 consonants and 3 vowels, the probability that in the two languages the same meaning is encoded by to the same CVC root by chance is one of 7 × 3 × 7 = 147, i.e. 0.0068. Thus, for instance, with three phonologically similar CVC roots with a similar meaning this probability should be 0.000,000,314, which in principle is already beyond Nichol's (1996) individual-identifying threshold. Nonetheless, Longobardi (2003a) shows that, since in mass-comparison matches vowels are allowed to vary a great deal, the basic probability should be increased to one of 7 × 7, i.e. 0.0204; What is more, due to the fact that in these comparisons also the semantic identity may be rather unconstrained (see for instance Goddard 1987: 657), so that on average a range of 10 different specific meanings can be assumed to satisfy the match each time (e.g. man, boy, male, brother, older brother and so on), the basic probability of chance resemblance between two lexical roots would become ten of 7 × 7, i.e. 0.204, which would provide scarce statistical evidence in favour of a historical explanation even comparing 4 lexical items[11]. Furthermore, Campbell (1988: 600) observes that it in mass comparison is often the case that only arbitrary segments of words are compared and, among others, McMahon and McMahon (1995: 183-185) list a number of errors in Greenberg's data themselves, including the use of non-existing forms, false cognates and references to wrong languages. On the whole, it seems that there is no way to obtain objective and reliable estimates of genealogical relatedness by means of Greenberg's procedure.

The neglect of any formal criterion of comparison between lexical items, and especially of those adopted by the classical comparative method, is the main cause of criticism against mass comparison. Greenberg appears to propose his procedure as an actual alternative to the classical comparative method (McMahon and McMahon 2005: 20), not just as a technique of heuristic investigation for long-range comparison, as it might be thought to have been originally conceived. In Greenberg's opinion "no one claims that we can devise a classification by regularity of sound correspondences, only that we can test hypotheses that have already been proposed" (Greenberg 1987: 6). Actually this is also (among others) Nichols' (1996) point, but Nichols' (1996) analysis shows that the recognition of systematic sound correspondences is preceded by the identification of diagnostic evidence, which is guided by well-defined scientific criteria as well. In Greenberg's opinion, instead, the evaluation of resemblance between words is exactly the procedure adopted by early

---

[11] See the references provided in note 6 and Fodor (1982: 80-96) for more extensive statistical calculations regarding mass comparison.

comparativists until Schleicher: He argues that, as these scholars were able to formulate important hypotheses regarding the classification of Indo-European languages even in absence of the Neogrammarian theoretical background, in the same way one should consider acceptable now investigating new linguistic families primarily relying on a general criterion of similarity. In other words, Greenberg's (1987) idea is that aiming at determining genealogical relatedness the usefulness of the principles of the classical comparative method, as well as that of the reconstruction of proto-forms, is illusory (Greenberg 1987: 2).

To provide a more detailed account for both the mass-comparison procedure and its (numerous) detractors' observations is not relevant here. What is important to underline is that, in line with Greenberg's intentions, in principle the method appears to be universally applicable, because there are no barriers to the list of words that can be included in the lists, but rather the more they are the better it is. At first glance this characteristic might be thought to represent an advantage of the multilateral procedure over the classical comparative method and lexicostatistics, neither of which can be said to be universally applicable in practice: The former because it relies on the detection of extremely rare evidence, the latter because it is *a priori* focused on the analysis on a limited list of meanings, whose "universality" is disputed. However, any possible gain of mass comparison is cancelled out by the fact that Greenberg is not able to establish a principle to carry out non-subjective and scientifically rigorous comparisons, thus the method does not yield a sharp demarcation criterion (e.g. Longobardi and Guardiano 2009: 4).

Summing up, in the last two sections it has been argued that, among the comparative methods considered, the classical is the only one that allows the investigation of the history of languages according to scientifically precise and statistically reliable criteria. Notice though, that its scope of application is relatively limited, for the reasons worked out in the end of section 4.2.1, with respect to the set of languages and, consequently, the time depth at which it can be successfully employed. The attempts to overcome these constraints by means of lexicostatistics and mass comparison have been demonstrated to be (at least) dubious as for their efficacy and consistency. Therefore is it possible to design a method to perform a reliable long range comparison? Following the framework outlined at least since Longoardi (2003a), in the next section arguments will be provided in favour of the idea that a significant procedure of deep historical investigation may be designed adopting a syntax-based (or better, parameter-based) comparative approach. In order to highlight the relevance of the shift of focus from (mainly) lexical to syntactic comparison, the introduction of the latter will be preceded by an overview of the comparative method used in population genetics, where long-distance genealogical relationships between human groups are precisely the main focus of interest.

4.2.3. From population genetics to the "new synthesis"

In biology the importance of adopting a rigorous and reasoned comparative methodology is clear at least since Linnaeus' taxonomic studies and, as noticed by Hauser *et alii* (2002), the comparative method was also "the primary tool used by Darwin (1859; 1871) to analyze evolutionary phenomena and continues to play a central role throughout modern evolutionary biology" (adapted from Hauser *et alii* 2002: 1572). Thus Harvey and Pagel (1998: 2) notice that "[f]orm Darwin's time to the present, the comparative method has remained the most general technique for asking questions about common patterns of evolutionary change" (Harvey and Pagel 1998: 2). The first part of this section briefly covers the basic notions that are relevant to provide hypotheses of phylogenetic classification in population genetics. Of course the following discussion is grounded on the framework outlined in chapter 2 (see in particular sections 2.1.3 and 2.1.4).

The units of comparison that are relevant in population genetics are ultimately genes. Recall that in the study of biological evolution the shift of focus from morphological (or phenetic) traits to genes, i.e. the integration of the Mendelian model with the Darwinian paradigm, gave rise to the modern synthesis and, consequently, to population genetics in the first half of the XX century. Apart from the already mentioned progress that this achievement allowed in the explanation of the mechanisms underlying the evolutionary process, what advantage did it entail precisely from the point of view of the comparison between taxonomic units? Cavalli-Sforza and Feldman (2003: 266) claim that the first study about molecular genetic variation was provided by Hirszfeld and Hirszfeld (1919), who described the human gene ABO, from which A, B and O blood groups derive (see below in this section). Notice that from the '20s to the early '90s, when the techniques for automated DNA sequences were developed, for the most part the analyses of genetic variation did not refer to the direct identification of genes, but mainly to the indirect evidence provided by the observation of the proteins encoded by the genes. Thus "[t]he first book of allele frequencies in populations, published in 1954 (Mourant 1954, editor's note), was limited almost completely to serological variation" (Cavalli-Sforza and Feldman 2003: 266) and nowadays there is still a larger amount of data from protein markers (also referred to as 'classic' markers') than of data from DNA. However, until now the results obtained by means of the classical (i.e. protein-based) analysis of polymorphism seem to have been in agreement with those derived from the study of pure DNA markers. It is beyond the aims of this dissertation to delve into the history of the various approaches and achievements of population genetics in the last fifty years, but it is worth pointing out that the first phylogenetic reconstruction of the evolution of our species drawn on the basis of genetic data is that suggested by Cavalli-Sforza (1963); Among the latest one, instead, one may remember the "phylogeographical" hypothesis reported in Underhill *et alii* (2001). All the proposals advanced in this line of inquiry are based on the premise that the evolutionary process undergone by our species should be projected onto both the historical and the geographic dimensions, so that they aim to provide both dates of separations between populations and information regarding the paths of migration of human groups throughout the continents. According to Underhill *et alii* (2001), the most plausible scenario would entail an initial expansion in Africa of a population of about 1000 individuals 100,000 years ago. Then between 60,000 and 40,000 years ago a second spread should have occurred to Asia in two routes: A southern one, which led humans to reach Oceania between 60,000 and 40,000 years ago, and a central one, which involved an expansion "through Middle East, Arabia or Persia to Central Asia, from where migration occurred in all directions reaching Europe" (Cavalli-Sforza and Feldman 2003: 270) and, finally, the Americas between 35,000 and 15,000 years ago.

Which are the relevant features of (some) genetic markers, that make them so suitable for comparing human populations and achieving consistent hypotheses of reconstruction? First of all, genes "are drawn from a universal list of discrete biological options" (Longobardi and Guardiano 2009: 6). The universal applicability of such a set of *comparanda* is safe, in that the same genetic markers are expected to be found in all individuals that belong to the human species. Furthermore, being that these markers are well-identified and discrete, it is guaranteed that one is always comparing like with like and questions regarding the interpretation of the comparative units (such as those that concern the semantics of lexical items) cannot arise[12]. Finally, the finiteness of the list of biological options can always be obtained and the size of this list is determined starting from the empirical evaluation of each scholar's specific needs. The last point may be understood better if one considers how the genetic data is used by population geneticists. In general the first step of the comparative

---

[12] However, to align multiple DNA sequences in order to compare them is not a straightforward operation, but is a common task in automatic sequence analysis. For an introduction to the relevant theory and techniques see for instance Higgins (2003) and Higgins and Salemi (2003).

procedure is the estimate of the frequencies that the alleles of a certain gene have in each population. For instance the already mentioned gene ABO has exactly the alleles A, B and O and an individual may have one of the six possible genotypes: AA, AO, BB, BO, AB, OO. Notice that the genotypes AA and AO on the one side and BB and BO on the other surface respectively as blood group A and as blood group B in the phenotype, so that when, say, an A individual is found, it is necessary to carry out specific mathematical calculations to determine whether she is AA or AO. Anyway, given a sample of subject it is possible to count the frequency of the alleles in populations: For example in the Basques the allele A has frequency 23%, the allele B 2% and the allele O 75% , while in the Italians A has 20%, B 7% and O 73% (Cavalli-Sforza 2001: 36). Therefore, once these quantities are measured, it is possible to derive an objective measure of similarity between the populations in question, or more precisely, a measure of distance (see section 4.3.3 and the references provided there). The more allele frequencies, i.e., ultimately, *comparanda*, the more reliable is the measure.
The choice of the genetic markers that are used to compare populations is not random and it depends on their "selective neutrality". This means that the genetic traits that are above all significant in population genetics are the "biologically inactive" ones, i.e. those that do not directly undergo natural selection (Cavalli-Sforza 2001: 108-109). Why is this important? Because similar environments favour the frequency of similar genotypes, so that the degree of relatedness between two populations calculated on the basis of "naturally selected" genes might be irrelevant from the strict genealogical point of view. The outcome would be comparable to the result that one would obtain comparing populations on the basis of anthropometric (or morphological) characters (e.g. the size of the skull, the colour of the skin and so on), which, however, are subject not only to natural selection, but also to environmental conditions that exclusively affect the phenotype in the course of an individual's life (e.g. the diet). Cavalli-Sforza (1963) attempts to reconstruct a phylogenetic trees of populations on the basis of morphological characters and, as expected, he finds that in this classification the Africans appear to be very closely related to the Aborigines, which is false. Therefore, choosing selective neuter genetic markers allows one to avoid the reconstruction of incorrect classifications biased by the presence of traits that have parallely evolved in different lineages. The question of parallel evolution will be reconsidered in section 4.3.3.

Notice that in the last decades an increasing number of works have tried to integrate the investigation of human prehistory that is pursued in human genetics and linguistics, but also in archaeology and anthropology. The objective of these attempts is to achieve a "new synthesis"[13] between the disciplines of historical linguistics, prehistoric archaeology and molecular genetics, an aim that has been more explicitly formulated, among others, by Renfrew (1987: 7; 1992; 2008). The progress of this innovative framework is particularly challenging for a number of reasons. First because, as stated by Nichols (1996: 39), many non-linguists are often unable to distinguish between linguistic classifications obtained by means of the safe classical comparative method and those derived from other non-secure procedures of historical investigation. For instance, Cavalli-Sforza *et alii* (1988) consider by the same yardstick the classification of the Indo-European and the Uralic family and that of the Indo-Pacific and New-World macrofamilies identified by Greeberg (1971; 1987). Renfrew (1991: 6) believes that the Nostratic hypothesis (see section 5.3.1) is grounded on the equivalent comparative evidence on which Proto-Indo-European is hypothesised. Moreover, Cavalli-Sforza *et alii* (1994: 23) declare to have made profit from Ruhlen's (1987) classification of the languages of the world (determined by means of mass comparison) to

---

[13] Of course, this "new synthesis" should not be confused with the modern synthesis occurred in evolutionary biology in the first half of the XX century.

draw a parallel with their genetically-based classification of human populations, and the same certainty of the reliability of Greenberg and Ruhlen's classifications can be found in Cavalli-Sforza (2001: 201-215) and Cavalli-Sforza and Feldman (2003: 270-273). In no way is this survey intended to blame the scholars that have been mentioned, but rather it serves to highlight how the attention in parallels between non-linguistic and linguistic classifications is increasing. As for the synthesis between historical linguistics and population genetics in particular, this interest is motivated by the fact that "[t]here is generally a great correlation between linguistic families and the genetic tree of major populations" (Cavalli-Sforza and Feldman 2003: 273). One might cautiously state that, whenever safely obtained, this correlation would reflect "[t]he frequent expectation is that if two (or more) languages are phylogenetically related, then genes of the populations speaking these languages will likely be similar", but actually "[w]e cannot assume, *a priori*, that linguistic history and human biological history will correlate" (Campbell 2006: 1). Thus caution is motivated exactly for the reasons worked out in Campbell (2006), which, summing up, start from the premise that to trace back relevant paths of historical development of human groups is more difficult when one is dealing with languages, that are subject to cultural transmission, that with genes, that undergo demic transmission (see also the observations put forward in section 2.3.1)[14], and divergences between insights obtained in, say, historical linguistics and population genetics may easily arise (McMahon 2004).

The intersection between historical linguistics and archaeology, instead, has been characterized by various explicit controversies (Heggarty 2007: 311). Probably the most famous one concerns the debate that is still dividing supporters of the "Kurgan" hypothesis on the origin of Indo-Europeans (e.g. Gimbutas 1970, see section 2.3.3) and followers of Renfrew's (1987) "Anatolian hypothesis": Summing up, while the former is motivated by archeo-linguistic evidence, the latter is primarily based on archaeological proofs. The debate regarding this specific topic, that has been revived by recent quantitative investigations (e.g. Gray and Atkinson 2003), is reconsidered in section 5.3, where other relevant proposals regarding the genealogical classification of some Indo-European and extra-Indo-European languages are illustrated. For the moment it is preferable to abstract away from them and to comprehend whether, as an early tentative step towards the achievement of the "new synthesis", it is possible at least to design a method of linguistic comparison that acts closer in accordance with the principles of the procedures adopted in population genetics. This issue is exactly the topic of the following discussion, where a recent method of syntactic comparison is introduced.

4.2.4. The Parametric comparison method

In general, one may say that the fate of syntactic comparison has been largely connected to that of the study of syntactic change (see section 3.2.1), therefore only in the last years has the issue been revived thanks to the projection of the theoretical advancements of generative syntax onto the historical dimension. Anyway, besides recalling the discussion presented in the last chapter together with the references provided there, it is worth briefly mentioning here some specific details regarding the *status* of comparative studies in syntax.

---

[14] For instance, the following arguments are pointed out by Campbell (2006): "(1) […] while a person has only one set of genes (for life), a person can be multilingual, representing multiple languages; (2) […] individuals (and communities) can abandon one language and adopt another, but people do not abandon their genes nor adopt new ones – *language shift* (language replacement) is a common fact of linguistic life; there is no deterministic connection between languages and gene pools. Languages become extinct in populations which survive genetically (language replacement and extinction are frequent)" (Campbell 2006: 1).

Fox (1995: 104) observes that the possibility of applying the classical comparative method to syntax has often appeared far-off to many linguists (e.g. Jeffers 1976: 4-5; Winter 1984: 618), essentially because, in their opinion, "syntactic change is essentially pattern replacement rather than evolution of the same entity, and hence the concept of a correspondence set is meaningless" (Fox 1995: 105). In other words, the likelihood of finding comparable syntactic units in different languages has often raised disbelief (Longobardi 2003b: 167) and the historical overview provided in Thomason (2004) confirms that comparative syntax has been largely disregarded until recent times. As well as the difficulty concerning the isolation of appropriate syntactic *comparanda*, another problem, mainly related to the possibility of reconstructing proto-forms, concerns the directionality of change: Some comments about this issue, that are less pertinent now, are stated in the end of this section. Anyway, in the XX century probably the most extensive attempts to provide an account for this issue have been made within the typological morphosyntactic-oriented framework initiated by Greenberg (1963) and put forward, among others, in Lehman (1973) and successive works (see also section 3.2.1)[15]. The general validity of typological comparison was soon criticized (e.g. Watkins 1976). In particular, its method of shedding light on the historical development of languages still appears to be improper, in that it relies on the detection of superficial syntactic patterns (e.g. word orders), which do not necessarily reveal anything about the similarity of the underlying structures (Thomason 2004: 7; Lightfoot 2002). Thus it is reasonable to claim that "attempts to use typological comparison to establish genetic relationships simply do not work" (Ross and Durie 1996: 5). In the '90s Harris and Campbell (1995: 344-376) proposed outlining the concept of "syntactic correspondence", that was meant to follow that of "systematic sound correspondence", arguing that two sentences in two related languages correspond if the have the same meaning, the grammatical morphemes they include are cognates and they are "responses to identical or essentially identical stimuli in two stages of a language" (Harris and Campbell 1995: 350). Apart from the fact that the last point seems not to be supported by psycholinguistic evidence (Thomason 2004: 6), the scholars' suggestion appears to be heavily dependent on the fact that the evaluation of cognacy between the relevant morphemes should already have been established by means, say, of the application of the classical comparative method. Moreover, as already claimed, any reference to the "meaning" comparing forms of different languages is risky and, again, nothing prevents the "same" sentence in different languages to be generated by different grammars. Stating it differently, as pointed out by Lightfoot (2002), the superficial correspondence between two syntactic patterns does not necessarily reveal that these patterns are produced by the same underlying structure, so that Harris and Campbell's (1995) notion of "syntactic correspondence" cannot provide reliable insights regarding the actual relatedness between the expressions in question, and neither about the reconstruction of proto-forms (Roberts 2007: 361).

As argued throughout section 3.2 and 3.3, the synchronically-based parametric perspective may be effectively applied to explain the dynamics of syntactic change. In the same way, it is argued at least since Longobardi (2003a) that the shift of focus to parameters as *loci* of syntactic variation may allow solving the problem of the choice of the *comparanda* in historical syntax as well. The idea, supported also by Roberts (1998), is that parameters appear to be perfectly suitable for comparing languages, because they are a universal finite list of discrete biological options (Longobardi 2003a: 106). Thus this set of *comparanda* is applicable to any group of possible human languages, in that parameters (or, better, parametric values) are assumed to be incorporated into any individual's knowledge of language. Moreover, since these cognitive objects correspond to precise discrete grammatical

---

[15] However, important typological insights may already be found in Sapir (1921: 120-146).

choices, on the one hand there is no doubt that one is always comparing "like with like", on the other hand ambiguous parameter settings are thought to be impossible, in the sense that each language has to be provided with a clear and exhaustive list of parametric values, each of them corresponding precisely to one of the two possible options entailed by each parameter (of course this observation applies to the cases of parameters that are relevant in each language, see the discussion about the possible consequences of unsatisfied implications in section 3.1.5 and section 5.3.1). Finally, for the reasons introduced especially in section 3.1.2, parameters are assumed constitute a set that is finite, as, consequently, the number of grammars that they can generate is finite, and also limited, at least relatively to the number of surface syntactic features that they control. To sum up, parameters as comparative characters are intended to share fundamental properties with genetic markers, i.e. universality, discretness, non-ambiguity and finitness.

On this premise it is possible to outline a proper procedure of syntactic comparison, i.e. the Parametric comparison method, whose definition and significance have been analyzed within a specific line of research developed in the last years (see in particular Longobardi 2003a; 2003b; Guardiano and Longobardi 2005; Gianollo *et alii* 2008; in press; Longobardi and Guardiano 2009). Notice that this method overcomes both the difficulties entailed by the lexical comparison and those connected to the typological one. As for the former, parameters are designed to correspond to deep structural properties of syntax, thus misleading superficial similarities as, say, those between epiphenomenal word orders, are left out from the comparison. Furthermore, unlike lexical-based evaluations of relatedness, the parametric estimates are not only safer from the potentially subjective interpretation of similarity between the comparative units, which may arise from precise sources of vagueness concerning the match between forms and meanings (see Longobardi and Guardiano 2009: 7-8), but also precisely measurable, as is explained later in this section and, more accurately, in section 6.1.1. Therefore it appears clear that "[l]ike the classical method, though by different means PMC (the Parametric comparison method, editor's note) overcomes the intrinsic uncertainty about the appropriate identification of *comparanda* which undermines (say, editor's note) mass comparison" (Longobardi and Guardiano 2009: 7). What is more, in principle the parametric comparison seems to have at least two advantages over the classical procedure: First, as already claimed, it is universally applicable, which means that it does not rely on the identification of (relatively) rare evidence (i.e. individual identifying evidence and systematic sound correspondences); Second, according to Keenan (1994; 2002) and Longobardi's (2001a) inertial reading of syntactic evolution (section 3.2.2), syntax is expected to be (among) the most conservative parts of languages, i.e. to be more likely to retain traces of chronologically deep common ancestry (e.g. Longobardi 2003a: 122)[16] than phonology, morphology and, ultimately, lexicon. In this sense, the Parametric comparison method may be seen as particularly apt to shed light on long-range genealogical relationships and thus to provide a contribution to overcome the main limitation of the classical approach.

In any case the verification of the validity of the last statements has to be preceded by other considerations. Indeed, having established that the parametric comparison is in principle possible, one has to understand whether it is also historically significant, i.e. whether it actually permits shedding light on the genealogical relatedness of languages and their historical evolution. Moreover, it is also fundamental to comprehend how its results relate with those obtained by means of non-syntactic methods of genealogical classification. The answers to these questions are not entirely predictable *a priori* and rather the "tacit assumption" that is sometimes taken for granted in the literature is that "the classification of

---

[16] This is true "at least with respect to those parameters which are not primarily set (hence diachronically resettable) on the basis of phonological (segmental or prosodic) evidence" (Longobardi 2003a: 122). On the connection between parameters and functional elements recall the discussion in section 3.1.4.

languages based on lexical arbitrariness (which is assumed to be genealogically relevant) and that based on syntactic properties (alleged to be only typologically significant) are essentially orthogonal" (Longobardi and Guardiano 2009: 4). This would mean that syntax is often supposed not to contain relevant genealogical information. Such a critical position is shared, among others, by Newmeyer (2005) and, with particular reference to the possibility of reconstructing proto-syntax, by Lightfoot (2002) (see later in this section). In synthesis, the issue may be summarized in the question provided in (6), which is the most recent statement of the so-called "Humboldt's problem"[17] provided by Longobardi and Guardiano (2009).

(6) "Are syntactic and lexical classifications of language significantly isomorphic?" (Longobardi and Guardiano 2009: 5)

The scholars argue that three different scenarios might occur: Either syntax provides weaker genealogical insights (i.e. with a shorter chronological range of efficacy) than lexicon, or they both provide the same results, or syntax provides stronger insights (i.e. with a longer chronological range of efficacy) than lexicon. Of course one can determine which from among the possible answers to (6) is closer to reality only on empirical grounds: The pertinent experiments regarding this specific issue will be presented in chapter 6, together with other relevant basic and more advanced tests based on the actual dataset presented in chapter 5.
However, as suggested at least since Longobardi (2003b: 202-205), apart from the empirical evaluation, the consistency of the method may be corroborated also from the theoretical point of view. In this sense, as it has been done since (Longobardi 2003a), it is important to ask what is the probabilistic reliability of the parametric comparison and whether it is below Nichol's (1996) threshold for individual-identifying. Thus, what is the probability that two languages, say *L1* and *L2*, have a certain number of identical parameter values by chance? Being parameters binary choices, this probability is 1 out of 2 (0.5) considering a single parameter, 1 out of 2^2 (0.25) with 2 parameters and so on; In general, having a number of parameters *n*, the probability that two languages converge on the same string of values by chance is 2^n. What happens with a "partial" convergence, i.e. the case in which one has a list of *n* parameters and *L1* and *L2* show the same parametric setting only in *k* cases, being $k < n$? For instance, if *L1* and *L2* are compared with respect to 45 parameters and they have 40 identical values, what is the probability that this occurred by chance? The estimate of this probability can be calculated using a basic formula of combinatorial mathematics, the "binomial coefficient", which is given in (7a)[18]. Notice that the binomial coefficient may equally be expressed as *C(n,k)* or in the bracketing form with *n* over *k*.

(7) a. $C(n,k) = \binom{n}{k} = \frac{n!}{k!(n-k)!}$

b. $\sum_{k=0}^{n} \binom{n}{k} = 2^n$

Given a set *S* of *n* elements, the binomial coefficient allows to obtain the number of all the possible subsets of size *k* derivable from *S*, that is the number of all the possible combinations without repetitions of the elements included in *S*. This number corresponds precisely to *C(n,k)*.

---

[17] "[T]he very inspiration for raising this question is rooted in Wilhelm von Humboldt's original distinction between several possible levels of language classification" (Longobardi and Guardiano 2009: 5; other relevant references are mentioned by the authors in the same page of the paper).

[18] In (7ab) and (8) the symbol "!" is used to indicate the factorial: For instance, 3! = 3 ×2 × 1, 12! = 12 × 11 × 10 × 9 × 8 × 7 × 6 × 5 × 4 × 3 × 2 × 1, and so on.

Notice that the number of all the possible subsets of any size that can be drawn from *S* is thus the sum of all the possible *C(n,k)*, where only *k* varies, and, as shown in (7b), it corresponds to 2^n. Coming back to *L1* and *L2*, one can say that in principle the two languages might converge for any number of parameters: This means that there could 5 identities, 18 identities, 28 identities and so on, without regard to the actual parametric values for which each identity is observed. Since each convergence defines a subset *k* drawn from the total number *n* of the parameters in question, which are 45, the number of the possible convergences is calculated as in (7b) and corresponds to 35,184,372,088,832. However, the identities between *L1* and *L2* are neither 5, nor 18, nor 28, but they are 40. Therefore the following objective is to understand how many convergences of 40 parametric values are possible[19] (i.e. how many subsets of 40 elements can be drawn out of a set of 45 elements), which, in this case and according to (7a), are 1,221,759. Finally, the probability that *L1* and *L2* share 40 identities out of 45 parametric settings by chance is obtained calculating the ratio between the last value and the number of all the possible convergences: The ratio means that only *C(n,k)* convergences out of the possible 2^n ones are exactly of the type observed, i.e. include 40 identities. The formula to calculate the probability *P* in question is given in (8).

(8) $$P = \frac{C(n,k)}{2^n} = \frac{\binom{n}{h}}{2^n} = \frac{\frac{n!}{h!(n-h)!}}{2^n}$$

As for *L1* and *L2*, the probability of their convergence is 0.000000034724479, which largely satisfies Nichol's (1996) threshold requirement for individual identifying. At this point it is worth pointing out that an identity score like that between *L1* and *L2* is not theoretical, but it is actually recorded from the data that will be presented in the next chapter, i.e. Table A: In particular, the situation of *L1* and *L2* corresponds to that of English and German. In sum, the parametric comparison method appears to pass the test of probabilistic consistency without complication, which means that the parametric correspondences that are found among languages appear to be due not to chance, but rather to represent significant evidence of common ancestry.

What about reconstruction instead? Is it possible to provide insights on the syntax of proto-languages within the parametric perspective? Among others, Lightfoot's (2002) answer to the last question is "no". Fundamentally he argues that reconstruction requires one to know which is the likely direction of changes (e.g. the fact that, referring back to section 4.2.1, the sound change [s] > [h] is more likely than [h] > [s]) and according to his approach catastrophic reanalyses, which are the primary mechanisms underlying syntactic change, would "constitute cutoff points to reconstruction" (Lightfoot 2002: 257) because they have no intrinsic directionality, thus in his opinion the entire syntactic system would evolve in a chaotic unreconstructable fashion like, for instance, weather. More optimistic (but not parameter-oriented) views on the possibility of reconstructing syntax are, for example, that of Miranda (1978), who draws a brief parallel between phonological and syntactic procedures, and that of Harris and Campbell's (1995: 353), who are convinced that once adequate correspondences were found, reconstruction would become achievable. Longobardi (2003b) points out that nothing prevents outlining a method of syntactic reconstruction, because it "appears to pose the same problems – and to require the same methodological choices – as phonetic reconstruction does" Longobardi (2003b: 199; editor's translation). Thus, assuming parameters as satisfactory *comparanda*, Roberts (2007: 360-368) agrees with this claim and,

---

[19] This value would remain the same if one calculated it referring to the number of differences that distinguish *L2* from *L2* (i.e. 45 – 40 = 5). For instance, given $n = 45$, to solve (7a) with $k = 40$ or $k = 5$ (i.e. $n - k$) one obtains the same *C(n,k)* (1,221,759).

referring to Vincent and Roberts's (2007) proposal, suggests dealing with the issue of directionality raised by Lightfoot (2002) resorting to theory of markedness developed since Roberts and Roussou (2003) (see sections 3.1.4, 3.1.5 and 3.2.2). However, this approach, as already claimed, is hardly applicable for the moment to an actual extensive parametric dataset (see also Longobardi and Guardiano 2009: 12). Of course to abstract away from markedness and, consequently, directionality, significantly restricts the number of principles that may help one to accurately perform and validate reconstructions: This is why reconstruction is not among the central topics of this work. Nonetheless, with this caution in mind, some early and tentative results regarding this issue will be suggested in section 6.3.2.

On the whole, parameters appear on the one hand to be perfectly adequate *comparanda*, on the other hand, to share some key features with the comparative units that are relevant in population genetics, i.e. genetic markers. This observation provides further support to the final claim of the last chapter, according to which parametric evolution and biological evolution might be studied by means of similar methods of investigation. In particular, these methods are those conceived in the field of quantitative phylogenetics, which is introduced in the following discussion.

4.3. Quantitative phylogenetic methods

"Quantitative phylogenetics" (also called "cladistics" at least since Ashlock 1974) is intended here as the research field, whose objective is to mechanically infer genealogical relationships between *taxa* (i.e. basically phylogenetic trees) using algorithmic procedures, called "phylogenetic methods", applied to comparative datasets. As anticipated in section 2.2.3, "quantitative" precisely indicates the fact that the techniques, by means of which the reconstruction of the relevant genealogical paths are derived, fundamentally rely on the numerical aspects of the comparison, that is they work independently of the nature of the taxonomic units in question, which might be, for example, species, populations, languages and so on. Of course this statement oversimplifies the matter, because actually most methods are designed according to precise evolutionary hypotheses, which may be applicable to some evolutionary domains but not to others, so that actually one has to carefully choose the procedures that are more suitable with respect to his field of inquiry.

Of course the use of the same methods of analysis in different domains with similar objectives may give an important contribution to the development of the "new synthesis". The first phylogenetic procedures started been conceived in the '50s within systematic biology to draw genealogical trees of species and, later, human populations. Felsenstein (2004a: 123) claims that Michener and Sokal's (1957) paper may be considered the first work on numerical inference of phylogenies: The aim of the clustering algorithm (see section 4.3.2) designed by the scholar was precisely of inferring phylogenies, not just of drawing taxonomies. This at least in Michener's view, while Sokal was less convinced about the evolutionary interpretation of the structure and preferred to attribute to it a taxonomic interpretation. Besides Michener and Sokal (1957), Edwards and Cavalli-Sforza (1964) "has at least an equal claim to be the founding paper for the numerical inference of phylogenies" (Felsenstein 2004a: 128): Indeed in this work both the parsimony method, the likelihood method and the statistical inference approach to inferring phylogenies were introduced for the first time (see section 4.3.1 for clarifications about "parsimony" and "likelihood"). In the last decades quantitative phylogenetic methods have been exported through disparate disciplines. In particular, one may think about the increasing number of attempts to automatically generate phylogenies using quantitative procedures in linguistics (e.g. Lohr 1999; Gray and Jordan 2000; Ringe *et alii* 2002; Forster and Toth 2003; McMahon and McMahon 2003; 2005; Gray and Atkinson 2003; Bryant 2004; Ben Hamed 2005; Nakhleh *et alii* 2005a; 2005b; Atkinson *et alii* 2005; Forster *et alii* 2006; Atkinson and Gray 2006a; 2006b; Gray *et alii* 2009). While most of these experiments have been exclusively conducted on lexical and morpho-phonological datasets, a minority of them focused on syntactic comparison (e.g. Longobardi and Guardiano 2005; 2009; Spruit 2005; 2008; Ryder 2006). An accurate account of all these works is given in the next chapter. For the moment it is worth anticipating that the quantitative techniques should be thought as "an addition to the historical linguist's toolkit, not a replacement for the historical linguist" (McMahon and McMahon 2005: VII). Notably, within humanities linguistics is not the only field of study in which the quantitative perspective is seriously taken into consideration: Also philologists have recently adopted phylogenetic methods to classify manuscript traditions (e.g. Barbrook *et alii* 1998; ; Spencer *et alii* 2004; Macé and Baret 2006). As pointed out introducing trees and networks in section 4.1.1, historical linguistics and philology share not only the objective of shedding light on the evolutionary history of their respective objects of study (i.e. languages and texts), but also the problem of the interference between different lineages, that philologists call "contamination" (Windram *et alii* 2006): The quantitative approach to the issue of horizontal transmission in philology is mentioned in section 4.3.3.

The next sections are meant to present diverse types of phylogenetic methods, with particular reference to the algorithmic procedures that concern the experimentation presented in chapter 6: Section 4.3.1 covers the illustration of two character based methods, i.e. those based on the criteria of maximum parsimony and maximum likelihood, while the introduction of some distance-based procedures, based on clustering and on least-squares, is provided in section 4.3.2. Notice that in the course of the discussion the specific applications that the methods in questions have had in historical linguistics so far are only incidentally mentioned in order to preserve the clarity of the discussion: This aspect is left aside to be covered in section 5.1. The same observation does not apply to section 4.3.3, which is dedicated to the survey of the principal techniques that may be utilized to automatically detect horizontal transmission in quantitative analyses: In this case some specific linguistic studies, in which many of the techniques in question have been designed, will be taken into consideration.

4.3.1. Character-based methods

Abstracting away from detailed conceptual insights (e.g. Colless 1985), "character" here is meant to be synonymous of "*comparandum*", i.e. a feature or trait that may be instantiated in different ways, that is assuming different "values" or "states", in different individual units, so that it serves to group these individual units into different *taxa*. Characters are *loci* of variation which, according to òeach field and perspective of study, are significant to differentiate relevant objects: For instance, in biology they may be biometrician dimensions (e.g. the length of the neck) or genetic markers (e.g. the ABO gene), in linguistics meanings (e.g. the meaning "father" in a lexicostatistic list) or syntactic parameters (see section 5.2.3). In the following discussion only discrete characters are considered, i.e. those that vary in a non-continuous fashion and whose values indicate precise separated states. For instance, discrete characters are adopted to define the taxonomic units of the dataset provided in figure 4a, which may be referred to as "character matrix": In this case four taxonomic units (*L1, L2, L3* and *L4*), that correspond to the rows, are defined and differentiated by means of five binary characters (*A*, *B*, *C*, *D*, *E*), whose values are + and -, so that, for instance, $A(L1) = A(L2) = +$ and $A(L3) = A(L4) = -$ while $B(L1) = B(L3) = +$ and $B(L2) = B(L4) = -$.
Phylogenetic character-based methods are procedures aimed at inferring phylogenetic trees of taxonomic units analysing the evolutionary history undergone by the set of characters by means of which the units are described. This definition will soon become clearer. For the

	A	B	C	D	E
L1	+	+	+	+	+
L2	+	-	+	+	+
L3	-	+	+	-	+
L4	-	-	-	-	+

Figure 4a: Character matrix of *L1, L2, L3, L4*.

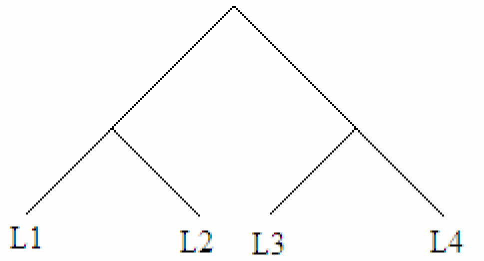
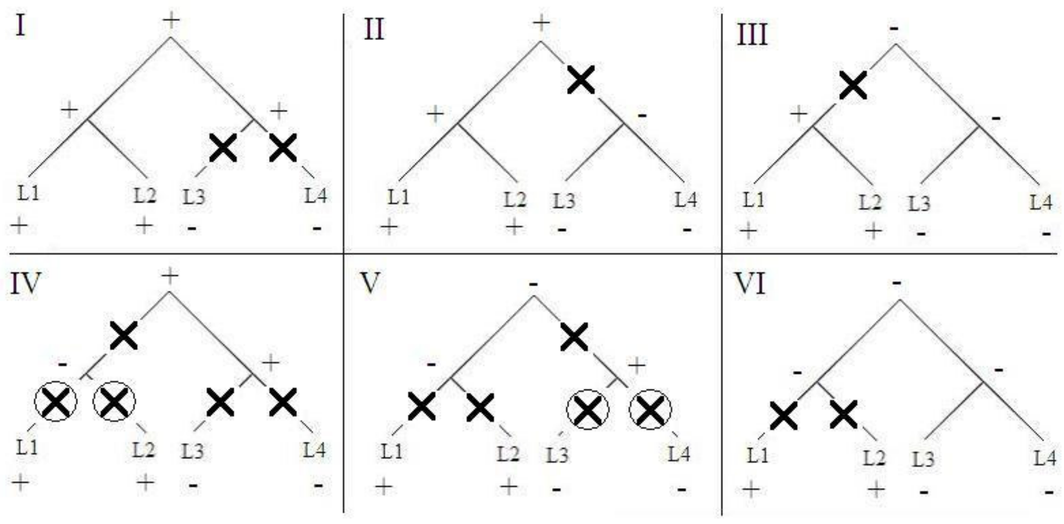
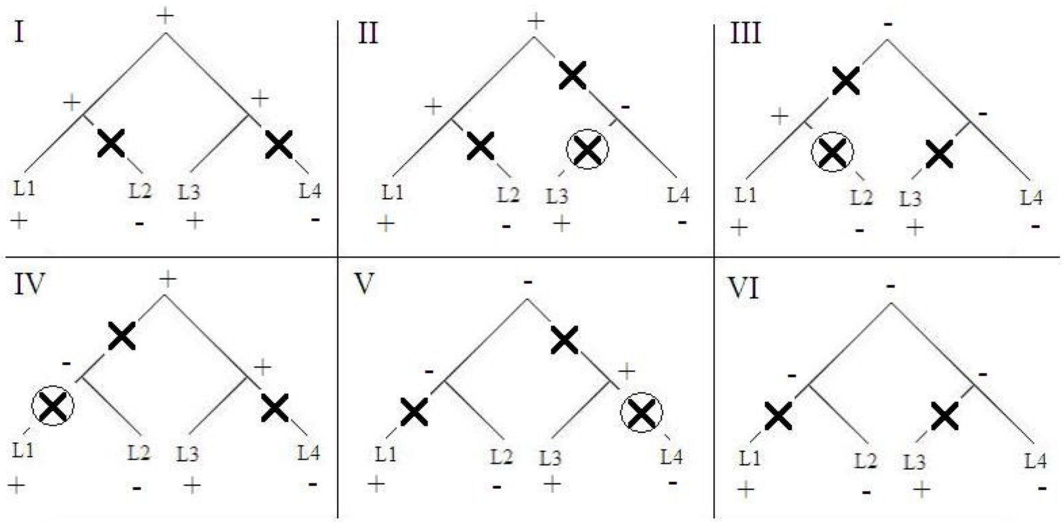


Figure 4b: Phylogenetic tree to be tested.

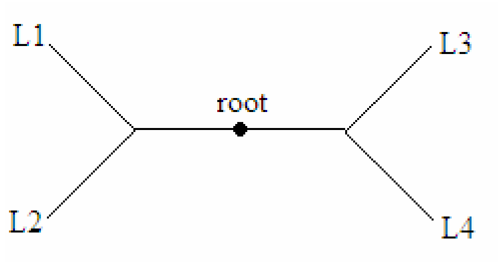


Reconstructing the history of character A onto the phylogenetic tree.



Reconstructing the history of character B onto the phylogenetic tree.

Figure 5: Character-state histories.



Parsimony score of best tree found = 5.

Character *status* summary:

- 1 character is constant (E);
- 1 character is parsimony uninformative (C);
- 3 characters are parsimony informative (A, B, D).

Figure 6: The most parsimonious tree obtained by the dataset of figure 4a.

moment notice that a condition that is taken for granted by the procedures in question is that characters are inheritable: Basically, they are supposed to be transmitted from ancestors to descendents preserving their states, except for the cases in which changes occur. The series of events of transmission undergone by a certain character on a phylogenetic tree may be called “character-state history”. Moreover, another important but less evident premise of these methods is that characters are assumed to evolve independently, both one with respect to the other and in different lineages (Felsenstein 2004b; for the maximum likelihood method in particular see Felsenstein 2004a: 251): This means that the phylogenetic methods in question are not guaranteed to infer the best genealogical tree whenever a character interferes with another, e.g. the state + of a character *X* determines the occurrence of state + in character *Y*, or there is a contact event between distinct lineages. The latter requirement is connected to the recurrent observation that has been made so far, i.e. that horizontal transmission biases the estimate of genealogical relatedness between *taxa*; The former, instead, may be more easily understood on the basis of a concrete example, thus it is reconsidered in section 5.1.1 introducing the lexical dataset used by Atkinson and Gray (2003; 2006). In any case, notice that the degree of bias in the results depends on how much the data departs from the optimal conditions and, to a certain extent, deviations are tolerated (again, see in particular the explanation provided in section 5.1.1).

The first mention of “parsimony” can be traced back to Edwards and Cavalli-Sforza (1963), who said that the ideal phylogenetic tree should be that in which “the minimum net amount of evolution” occurred. However, as already claimed, it was only in Edwards and Cavalli-Sforza (1963) that this notion was integrated for the first time within a specific procedure of reconstruction, i.e. a “maximum parsimony method”. The most parsimonious phylogenetic tree is defined as the tree “on which, when we reconstruct the evolutionary events leading to our data, there are as few events (of change, editor’s note) as possible” (Felsenstein 2004a: 1). This requirement is like an “Occam’s razor” for the phylogenetic reconstruction, in that it says that no more changes of state should be postulated in the evolutionary history of the *taxa* than those that are strictly necessary to explain the observed distribution of the character-states. It is worth noting that the principle of parsimonious evolution is debated greatly, since some scholars argue that it does not rely on a scientifically valid hypothesis (e.g. Cavalli-Sforza *et alii* 1994: 34-35). However, it can be cautiously taken into account in the domains in which it is reasonable to suppose a relatively high level of conservativity of the character states, or at least a less pronounced attitude to change.

The maximum parsimony methods require that the history of each character is reconstructed onto the phylogenetic tree. An example of this procedure based on the dataset of figure 4a and the tree of figure 4b is provided in figure 5. The parsimony of the tree *((L1, L2), (L3, L4))* can be tested mapping the observed states on the leaves and then reconstructing all the possible character-state histories throughout the graph: The events of change are represented with a *X* (in some cases a “circled” *X*, which is not relevant for the moment). As for the character *A*, the most parsimonious scenarios are (II) and (III), because they require just one change: Depending on the state that is hypothesized for the root *r*, this change can be projected either on the branch (*r, L1+L2*) or on the branch (*r, L3+L4*). The other reconstructions entail a higher number of changes, thus within the parsimonious perspective they are discarded. Character *B*, instead, presents a less parsimonious reconstruction: The best scenarios are (I) and (VI), for which two changes are required, while the maximum number of mutation events, postulated by all the other reconstructions, is three. Therefore, until now the phylogeny in 4b requires a total of three changes of state, one for character *A* and two for character *B*.

Observe that, even if the history of characters *A* and *B* has been projected onto a rooted tree, the rootedness of the graph is not required to perform the reconstruction of the ancestral states. The parsimony scores of *A* and *B* would be the same wherever one decided to put the root, as, consequently, the global parsimony of the unrooted trees that can be generated from the taxonomic units. The step-by-step demonstration of this assertion is not provided here, since it may be checked simply on the unrooted tree in figure 6, which is actually the most parsimonious tree that can be drawn for *L1*, *L2*, *L3* and *L4*: In particular, the parsimony score of this phylogeny amounts to 5 and it is calculated summing the changes included in the history of *A* (1 change), *B* (2 changes), *C* (1 change), *D* (1 change) and *E* (0 changes). In order to determine that this tree is the most parsimonious, the same analysis performed for *A* and *B* has been repeated with respect to all the characters included in the dataset on all the tree-graphs that can be generated by the taxonomic units (see among others Felsenstein 2004a: 19-36). This structure needs to be rooted then by means of the "outgroup criterion", already mentioned in section 4.1.2: The tree given in figure 4b may be thought to have been drawn assuming that the *taxon* (*L1*, *L2*) is monophyletic, i.e. evolved separately from *L3* and *L4*, in other words that (*L1*, *L2*) plays the role of the outgroup in this graph[20].

In this case the output of the parsimony analysis is rather simple, because it consists of a unique tree that maximizes the parsimony criterion, i.e. that has the lowest parsimony score. However, it often happens that there is more than a single tree that attains the highest parsimony score, so that the outcome of the analysis is an entire set of phylogenies. Concrete examples of this situation are among the experiments presented in chapter 6. For the moment, notice that when this happens it is necessary to understand what is the "trend" of the output, i.e. to identify a so-called "consensus tree", that summarizes the different branching options that are found in the set of the inferred trees. Nothing more about this point is said here in order to maintain the focus of attention on the method under discussion: The criteria by means of which consensus trees are derived are introduced in section 6.1.3, where their relevance with respect to the "bootstrap procedure" is shown.

Apart from the phylogenetic tree and its parsimony score, figure 5 includes other information about the parsimony analysis. In particular, it is indicated that one character of the dataset is "constant" (*E*) and one is "parsimony uninformative" (*C*), so that only 3 characters out of 5 (*A*, *B* and *D*) are "parsimony informative". These specifications may be understood referring to the concept of "split", that is defined in (9).

(9) A split *s* is a division of the set of taxonomic units into *n* non-empty parts, where *n* is the number of values observed in the set for a single character.

For instance, the split entailed by character *A* is represented as $s(A) = \{(L1, L2), (L3, L4)\}$: In other words, *A* determines the division of the original set of languages into two non-empty subsets, i.e. $A_1=\{L1, L2\}$ and $A_2=\{L3, L4\}$, and $n(s(A)) = 2$. Therefore, a character is constant if the distribution of its states does not entail a distinction between the taxonomic units, so

---

[20] Needless to say that, assuming that the taxon (*L3*, *L4*) evolved separately from *L1* and *L2* and therefore (*L3*, *L4*) is the outgroup, one would obtain the same result.

that it does not provide information regarding their classification and it is "transparent" to the phylogenetic analysis: This is the case of *E*, which entails a unique subset that corresponds exactly to the entire set of taxonomic units, i.e. *s(E) = {(L1, L2, L3, L4)}*. A character is parsimony uninformative if it divides the set of taxonomic units into two subsets and one of them contains only one unit. For instance, *s(C) = {(L1, L2, L3),(L4)}*. Why is the split entailed by *C* irrelevant for the parsimony analysis? Essentially, because in any possible tree generated by *L1-L4* this split would entail the postulation of one and only one change of state, that is the change that differentiates *L4* from the other units in question. In other words, under the parsimony criterion the distribution of the states of *C* does not reveal anything relevant about the phylogeny (apart from the evident fact that *L4* is separated from the rest), thus it is ignored by the analysis. Of course parsimony informative characters are those that do not fall under the previous definitions.

Are all changes equal? The answer is "no". When multiple changes are identified for a single character, either back-mutation or parallel evolution or both might have occurred. In the example of figure 5, in any case in which more than one *X* is drawn in a scenario, there has been a parallel development, i.e. the same change of state has occurred in two different lineages. Moreover, the "circled" *X*s indicates cases of back-mutation, i.e. for instance situations in which in a single path from the root to a leaf there has been a change from + to – and then a return to +. The events of back-mutation and parallel evolution are designated by a single term in quantitative phylogenetics: Homoplasy (e.g. Felsenstein 2004a: 3). At this point it is useful to highlight the meaning of homoplasy contrasting it with that of homology, as it is done in (10ab).

(10) a. Homology: Similarity (between two states) due to common ancestry.
b. Homoplasy: Similarity (between two states) due to convergent evolution (e.g. parallel evolution or back-mutation).

Therefore, the corresponding states of a certain character instantiated in two different taxonomic units are considered "homologous" if they are identical as a consequence of deriving from the same proto-character state. If, instead, they are identical because they have undergone parallel but independent changes, they are said to be "homoplastic". An example of homoplasy in the biological domain may be that regarding the development of wings: Both birds and bats have wings, but this feature developed independently and parallely in the two lineages. The application of the same notion in historical linguistics will be clarified in section 4.3.3. The adjective "homoplastic" may also be used to refer to characters, whose evolutionary history onto a certain tree includes events of non-genealogical convergence.

From this picture follows the concept of "character compatibility". By definition a character is said to be compatible with a given phylogenetic tree when it is not homoplastic with respect to it (see for instance Felsenstein 2004a: 87-96). Returning to the example of figure 5, it is clear that only the history of character *A* can be projected onto the tree without homoplasy (scenarios (II) and (III)), thus it is compatible. Character *B*, instead, is not, since there is no way to reconstruct a non-homoplastic evolution for it. The concepts of parsimony and compatibility are strictly connected: It is easy to show that, if the number of states of a character is k, that character will be compatible with a tree if and only if it is possible to reconstruct its most parsimonious history with a minimum number of changes c such that $c \leq (k-1)$ (Felsenstein 2004a: 87). In fact in the example of figure 5 only character *A* satisfies this requirement.

Estabrook *et alii* (1976) provide two theorems regarding character compatibility: According to the first one, two binary characters are jointly compatible if and only if they do not possess all four possible character-state distributions (which is exactly the situation shown in the dataset of figure 5), or, if all four relations exist, they form a circuit and the characters are not compatible. From the point of view of the perspective offered by the split-based interpretation of a character matrix, this means that given two characters, say *A* and *B*, and the splits that they determine, $s(A)=(A_1, A_2)$ and $s(B) = (B_1, B_2)$, in order *A* and *B* to be jointly compatible one of the four intersections $A_1 \cap B_1$, $A_1 \cap B_2$, $A_2 \cap B_1$ and $A_2 \cap B_2$ must be empty. This means that there is no phylogenetic tree with respect to which both A and B are simultaneously compatible. The second theorem proposed by Estabrook *et alii* (1976), instead, says that two sets of binary characters are compatible if and only if they are pair-wise compatible. Essentially, this entails that, given the set of all the splits generated by the characters included in the dataset, say $\sum = \{s(A), s(B), \ldots, s(N)\}$, $\sum$ is said to be compatible with a tree if all the splits included in $\sum$ are compatible: In other words, if two characters are pair-wise incompatible and they are both comprised in $\sum$, $\sum$ is also incompatible with respect to a certain tree. From this follows the definition of "perfect phylogeny", which is intended as a tree that is based on a fully compatible split set $\sum$ (Huson 1998: 68). Finally, compatibility is also fundamental to define the notions of "clique" (or "maximal clique"), which is intended as largest set of compatible characters, i.e. the largest set of characters which are all pair-wise compatible. (Le Quesne 1969; Estabrook *et alii* 1976a; 1976b; see also Felsenstein 2004a: 91-96).

Before proceeding, notice that an extensive account about the possible algorithmic implementations of parsimony and its variants is provided for instance by Felsenstein (2004a: 11-18 and 73-86) and will be mentioned in section 6.1.2, dealing with the problems related to the applicability of these techniques to the parametric dataset. Something more about compatibility, instead, is stated in the discussion regarding the quantitative approaches designed to detect contact between different lineages in section 4.3.3.

The account for the maximum likelihood method that is given here is very brief and simplified. Indeed, the illustration of this techniques has just two objectives: On the one hand it serves as a basic and concise introduction to the discussion of some quantitative studies that in the last years have been proposed in historical linguistics and that are covered in chapter 5; On the other hand, as this method won't be used to carry out the experimentation for the reasons worked out in section 6.1.2, it is important to shed light on its essential mechanism in order to motivate the choice to exclude it. However, further details may be found for instance in Edwards (1972), Pagel (2000), Felsenstein (2004a: 248-274) and in the relevant works that will be mentioned henceforth.

The maximum likelihood approach was first outlined in statistics by Fisher (1922); Both Edwards and Cavalli-Sforza were been Fisher's students and they thought to apply the same method to the problem of inferring phylogenies (Edwards and Cavalli-Sforza 1964). Fundamentally, this technique requires three components: A character matrix, a model of character evolution and a set of trees on which to perform the likelihood evaluation. First of all, what is meant to by "model of character evolution"? Essentially, this is the set of probabilities of all possible character-state changes, i.e. of the rate with which each character state is supposed to change into another one. These probabilities are collected within a rate matrix. An example of rate matrix is provided in Atkinson and Gray (2006: 95). Here it is relevant to recall that models of DNA evolution are widely used in molecular biology: The simplest one is Jukes and Cantor's (1969), which assumes equal rates of change between all pairs of nucleotides, i.e. equal rates both for transitions and for transversions (e.g. Felsenstein 2004a: 156-158; an overview of the different models of DNA evolution is provided, among others, in Strimmer and von Haeseler 2003). Therefore, given the character matrix and the rate matrix, which, notably, essentially incorporates information about the directions of changes in the form of probabilities, the successive steps of the maximum likelihood method consist of determining which phylogenetic tree maximises the likelihood. This is done considering (in principle) all the possible trees that can be generated from the dataset in turn. For each phylogeny the following operations are performed:

1. All the possible character-states histories of each character are mapped onto the tree. Notice that each character state-history of each character defines a different scenario;
2. To each scenario a certain likelihood is assigned. The likelihood depends on the probabilities given in the model;
3. The total likelihood of the tree in question precisely corresponds to the product of the likelihood scores of all the scenarios that have been taken into account.

Once this process is complete, the choice of the tree that entails the maximum likelihood score is straightforward. The main problem is to carry out these calculations on all the possible phylogenies, thus in general Bayesian inference is used, which allows "to draw inferences from a large amount of data using powerful probabilistic models without searching for the 'optimal tree'" combined with the Markov Chain Monte Carlo algorithm (Metropolis *et alii* 1953) "to generate a sample of trees in which the frequency distribution of the sample is an approximation of the posterior probability distribution of the trees" (Atkinson and Gray 2006: 96; on these topics see in particular Huelsenbeck *et alii* 2001). It is not relevant here to delve into these techniques, since they would require an out-of-line digression. Rather, a more extensive account will be provided in the next section for a completely different type of procedures: The distance-based phylogenetic methods.

4.3.2. Distance-based methods

Sokal and Sneath's (1963) clustering algorithms represented the first attempt to design procedures of classification based on distances between taxonomic units, but proper distance-based phylogenetic methods were introduced only since Cavalli-Sforza and Edwards (1967) and they were then popularized by Fitch and Margoliash (1967). These techniques are blind to the distribution of character values and focuses only on the set of pair-wise distances that are given in input in the form of a matrix. Consequently the effectiveness of these procedures depends on the appropriateness and the accuracy of the distance measures that are calculated between each pair of taxonomic units. In particular, an insight into a specific measure of dissimilarity that is used in population genetics is provided in Cavalli-Sforza *et alii*' (1994: 29-30): The distance adopted by the scholars is calculated from the allele frequencies observed in different populations (see section 4.2.3). Detailed surveys of other possible distance formulas are offered for instance in Nei (1987) and Xia (2001: 179-196), with particular reference to sequence-based measures (i.e. based on nucleotide, codon and amino-acid sequences) designed starting from specific models of evolution, e.g., among others, the distance derived from the above mentioned Jukes and Cantor's (1969) model. It is worth anticipating that none of these formulas can be used to calculate the distances between the languages included in the parametric dataset: The pertinent measure will be introduced in section 6.1.1.

As a general principle, in any distance-matrix procedure the reconstructed branch-length that separates each pair of leaves in the phylogenetic tree, i.e. the expected distance, is supposed to reflect the distance between the taxonomic units given in input, i.e. the observed distance. In other words, "the best way of thinking about distance-matrix methods is to consider distances as estimates of the branch lengths separating that pair of species (or taxonomic unit, abstracting away from the biological context, editor's note)" (Felsenstein 2004a: 147). The problem is that in the full tree, in which more than just two units are considered, the set of the path lengths between each pair may (and often does) not exactly correspond to the set of the observed distances. Take for instance the unrooted tree in figure (7a), in which there are four

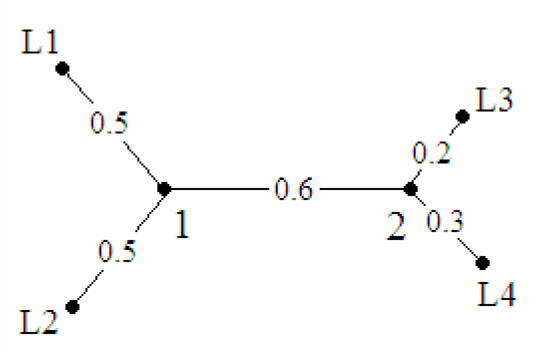


Figure 7a: An unrooted tree, whose branch lengths are specified.

	L1	L2	L3	L4
L1	0	1	1.3	1.4
L2	1	0	1.3	1.4
L3	1.3	1.3	0	0.5
L4	1.4	1.4	0.5	0

Figure 7b: Matrix of additive distances derived from the unrooted tree.

leaves (*L1*, *L2*, *L3* and *L4*) and two internal nodes (*1* and *2*)[21] and the branch lengths are specified as $l_{(L1,\,1)} = l_{(L2,\,1)} = 0.5$, $l_{(1,\,2)} = 0.6$, $l_{(2,\,L3)} = 0.2$ and $l_{(2,\,L4)} = 0.3$. Summing up the length of the branches, one obtains the set of distances between *L1*, *L2*, *L3* and *L4* that is given in figure (7b), i.e. $D_{L1,\,L2} = 1$, $D_{L1,\,L3} = 1.3$, $D_{L1,\,L4} = 1.4$, $D_{L2,\,L3} = 1.3$, $D_{L2,\,L4} = 1.4$, and $D_{L3,\,L4} = 0.5$. This set of distances is said to be perfectly additive[22], in that it can be projected onto a real graph represented in a two-dimension space. Thus if one is provided with a similar distance matrix, no problem arises and there is no approximation in the reconstruction of the tree. What happens if the set of distances is just slightly different? Suppose to start the reconstruction with a distance matrix identical to that of figure (7b), except for the value assigned to $D_{L1,L3}$, which is changed into 1.5 (figure 8a). In this case there is no way to perfectly match the set of observed distances with the set of branch lengths assigned to a real graph in a two-dimension space, as shown by the two tentative trees in figure (8b): If the set of branch lengths $l_{(L1,\,1)} = l_{(L2,\,1)} = 0.5$, $l_{(1,\,2)} = 0.6$, $l_{(2,\,L3)} = 0.2$ and $l_{(2,\,L4)} = 0.3$ is preserved, one loses the actual projection of $D_{L1,\,L3}$ onto the tree, which appears to be 1.2 instead of 1.5; If $l_{(L1,\,1)}$, $l_{(L2,\,1)}$, $l_{(1,\,2)}$, $l_{(2,\,L3)}$ and $l_{(2,\,L4)}$ are designed to perfectly match $D_{L1,\,L3}$, $D_{L2,\,L3}$, $D_{L2,\,L4}$ and $D_{L3,\,L4}$, for instance extending $l_{(L1,\,1)}$ from 0.5 to 0.7, then $D_{L1,\,L2}$ and $D_{L1,\,L4}$ cannot be projected onto the tree and appear to be respectively 1.2 and 1.6 instead of 1 and 1.4. One can try any possible assignment of branch lengths, but the set of the observed distances in figure 8a can never be exactly represented in the two-dimension space, therefore it is said to be a set of non-additive distances: In this situation a phylogenetic tree can be inferred, but it is necessarily an approximation of the data given in input.

As a matter of fact, normally any set of observed distances is hardly additive or even close enough to satisfy this condition. Distance-based phylogenetic methods are intended precisely to overcome the difficulties related to this issue by means of different criteria of tree inference. Here two types of approach are considered in turn: That of the clustering algorithms (in

---

[21] The number assigned to the internal nodes are not meaningful, i.e. they do not indicate any specific order. This observation will recalled also for the trees reconstructed on the parametric dataset.

[22] Notice that additive distances do not coincide with triangular distances. While additive distances are those, which can be derived from a tree-graph, triangular distances are those, which satisfy the triangle inequality condition. According to the theorem of triangle inequality, for any triangle the measure of a given side must be less than or equal to the sum of the other two sides but greater than or equal to the difference between the two sides. Thus in any set of trinagular distances D_n the conditions $D_1 + D_2 \geq D_3$ and $D_1 - D_2 \leq D_3$ must be satisfied. There are sets of triangual distances which are not additive, while additive distances are always triangular. In other words, the set of additive distances is a subset of the set of triangular distances.

	L1	L2	L3	L4
L1	0	1	1.5	1.4
L2	1	0	1.3	1.4
L3	1.5	1.3	0	0.5
L4	1.4	1.4	0.5	0

Figure 8a: Matrix of non-additive distances.

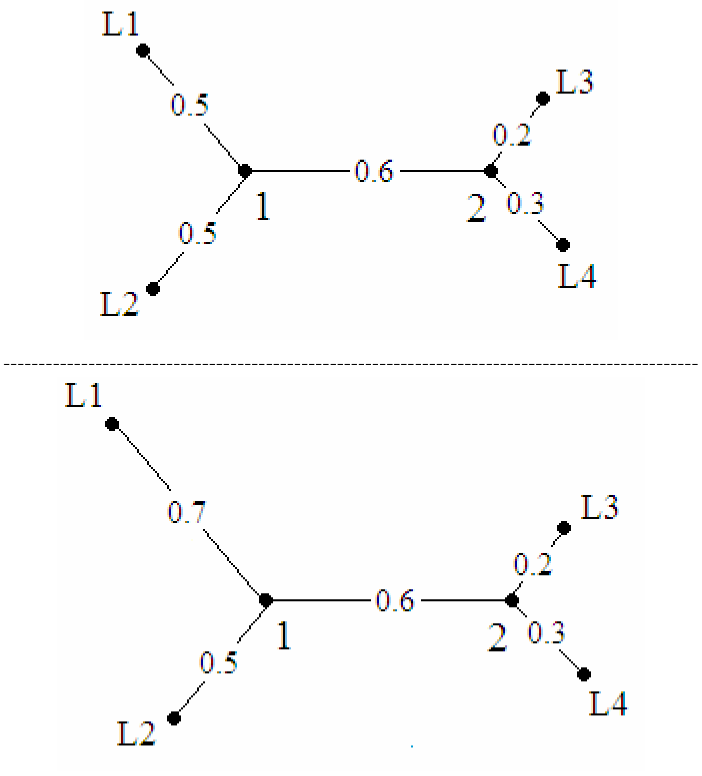


Figure 8b: Tentative unrooted trees inferred from the matrix of non-additive distances.

particular, the algorithms of Neighbor-Joining and UPGMA) and that of the least-squares techniques.

Clustering algorithms do not have an explicit criterion to deal with the problem of additive distances, but they are directly applied to distance matrices to mechanically produce successive groupings that, in the end, result in tree-graph structures. The first clustering procedure considered here is Neighbor-Joining (NJ, Saitou and Nei 1987). The main advantage of NJ is that it is fast compared to other procedures, e.g. the least squares methods, because it can be implemented in an algorithm, which has low computational costs. Given a set of *n* taxonomic units whose pair-wise distances *D* are specified in a distance matrix, this algorithm (as modified by Studier and Keppler 1988) works essentially joining at each cycle the two taxonomic units *i* and *j* for which $D_{i,j}$ is the smallest, then deleting *i* and *j* from the distance matrix and substituting them with a unique unit (*i*,*j*), whose distance with any other unit *k* is calculated with a specific formula. The operations performed by the algorithm are the following:

1. “For each tip, it computes $u_i = \sum_{j:j\neq i}^{n} D_{ij}/(n-2)$. Note that the denominator is (deliberately) not the number of items summed.
2. It chooses the *j* and the *j* for which $(D_{ij} - u_i - u_j)$ is the smallest.
3. It joins items *i* and *j*, then it compute the branch length from *i* to the new node (*vi*) and from *j* to the new node (*vj*) as

$$v_i = \frac{1}{2}D_{ij} + \frac{1}{2}(u_i - u_j) \text{ --- } v_j = \frac{1}{2}D_{ij} + \frac{1}{2}(u_j - u_i)$$

4. It computes the distance between the node (*ij*) and each of the remaining tips as

$$D_{(i,j),k} = (D_{i,k} + D_{j,k} - D_{i,j}) / 2$$

5. It deletes tips *i* and *j* from the tables and replaces them by the new node, (*ij*), which is now treated as a tip.
6. If more than two nodes remain, the algorithm goes back to step 1. Otherwise, it connects the two remaining nodes (say, *l* and *m*) by a branch of length D_{lm}" (adapted from Felsenstein 2004a: 167)

The specific passages of the procedure are exemplified in sketch 1 (see the appendix). Notice that NJ is assured to infer the true tree for the set of taxonomic units included in the distance matrix if the distances are perfectly additive, i.e. precisely reflect the set of branch lengths reconstructed on the tree (Felsenstein 2004a: 166).

The other clustering algorithm that is relevant here is UPGMA (Sokal and Michener 1958), which is the acronym of "Unweighted Pair Group Method Using Arithmetic Averages". This procedure works in a similar way to which NJ does, but it performs the following operations:

1. "It finds the *i* and *j* that have the smallest distance D_{ij}.
2. It creates a new group, *(ij)*, which has $n_{(ij)} = n_i + n_j$ members;
3. It connects *i* and *j* on the tree to a new node,which corresponds to the new group *(ij)*. It gives the two branches connecting *i* to *(ij)* and j to *(ij)* each length $D_{ij}/2$.
4. It computes the distance between the new group and all the other groups (except for *i* and *j*) by using:

$$D_{(ij),k} = \left(\frac{n_i}{n_i + n_j}\right)D_{ik} + \left(\frac{n_i}{n_i + n_j}\right)D_{jk}$$

5. It deletes the columns and rows of the data matrix that correspond to groups *i* and *j*, and add a column and row for group *(ij)*.
6. If there is only one item in the data matrix, the algorithm stops. Otherwise it returns to step 1" (adapted from Felsenstein 2004a: 162).

Observe that in this case *n* indicates the number of members that are included in each of the groups that are created at each repetition of the passages 1-5. As was done for the NJ algorithm, the first operations performed by UPGMA on an actual distance matrix are illustrated in sketch 2.

It is important to point out that UPGMA produced "ultrametric trees". A tree is said to be ultrametric, or clocklike or reconstructed under the "molecular clock hypothesis", if it is rooted and has "the total branch length from the root up to any tip equal" (Felsenstein 2004a: 161): This means that in an ultrametric tree the total branch length from a tip down to the root (and also to any internal node, which are pseudo-roots from which sub-trees originate) corresponds to the average of the distances between all the pairs of taxonomic units, which have the root as their most recent common ancestor. Take for instance the ultrametric tree in figure (9a), in which the branch lengths are specified: The graph is designed so that from the root (node 1) to any leaf the length of the path is the same, i.e. $l_{(1, L1)} = l_{(1, L2)} = l_{(1, L3)} = l_{(1, L4)} = 0.6$. The resulting set of distances provided in figure 9b is $D_{L1, L2} = D_{L3, L4} = 1$ and $D_{L1, L3} = D_{L1, L4} = D_{L2, L3} = D_{L2, L4} = 1.2$. If the matrix in figure 9b is given in input to UPGMA, the

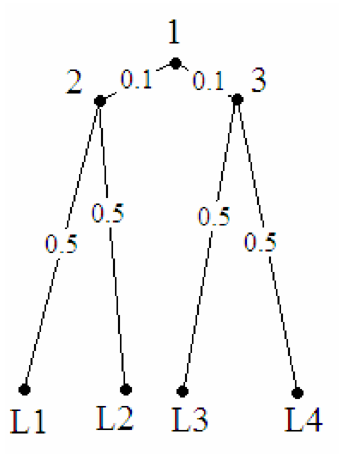


Figure 9a: An ultrametric tree, whose branch lengths are specified

	L1	L2	L3	L4
L1	0	1	1.2	1.2
L2	1	0	1.2	1.2
L3	1.2	1.2	0	1
L4	1.2	1.2	1	0

Figure 9b: Matrix of distances derived from the ultrametric tree in figure 9a.

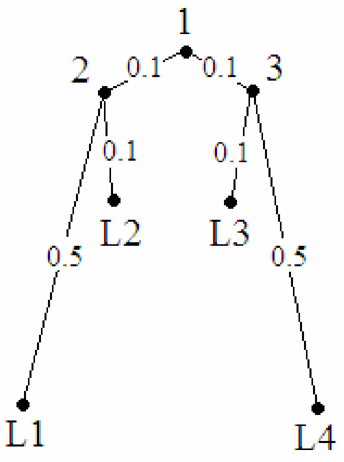


Figure 10a: A non-ultrametric tree.

	L1	L2	L3	L4
L1	0	0.6	0.8	1.2
L2	0.6	0	0.4	0.8
L3	0.8	0.4	0	0.6
L4	1.2	0.8	0.6	0

Figure 10b: The distance matrix derived from the non-ultrametric tree.

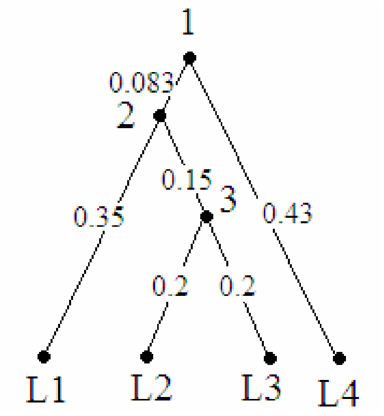


Figure 10c: The (wrong) ultrametric tree reconstructed by UPGMA.

method reconstructs exactly the original tree of figure 9a. Now take the non-ultrametric tree in figure 10a: In this case the length of the paths from the root to the leaves is not constant, in fact $l_{(1, L1)} = l_{(1, L4)} = 0.6$ and $l_{(1, L2)} = l_{(1, L3)} = 0.2$. This means that *L2* and *L3* are genealogically closer to the root than *L1* and *L4*. From this non-ultrametric tree the distance matrix in figure 10b can be derived. What happens if this matrix is given in input to UPGMA? The algorithm reconstructs the tree in figure 10c, which is incorrectly forced to be ultrametric and consequently wrong. Therefore, if the set of distances given in input to UPGMA is significantly non-ultrametric, or, at least, does not approximate this requirement, the algorithm, as any clocklike reconstruction procedure, may infer an incorrect phylogeny.

What is the evolutionary interpretation of an ultrametric tree? First of all, observe that the molecular clock hypothesis, under which this type of graphs is reconstructed, is the only alternative to the outgroup criterion if one wants to identify a root: Indeed, the root node is placed precisely at the same distance from all the leaves. Furthermore, adopting a clocklike approach entails assuming that the rate of change is approximately equal in all lineages. How is this assumption derived? Of course the amount of evolution a that accumulates in a certain lineage i increases through time: Knowing how much evolution a_i occurred and how much time t_i elapsed in the lineage, it is possible to obtain an estimate of the rate of evolution, i.e. r_i,, as shown in (11).

(11) $r_i = a_i / t_i$

Consider that a certain branch of a tree, say (*1*, *2*), reflects the expected (or reconstructed) amount of evolution a^{Exp} occurred from node *1* to node *3*, i.e. $a^{Exp}{}_{12}$. Thus $a^{Exp}{}_{12} = t_{12} \times r_{12}$. From this follows that, given a node *3* such that *3* and *2* are coeval, i.e. such that the same time elapsed from *1* to *2* and from *1* to *3* ($t_{12} = t_{13}$), if the same branch length separates *1* and *2* on the one side and *1* and *3* from the other, i.e. $a^{Exp}{}_{12} = a^{Exp}{}_{13}$, one has to conclude that the rate at which evolution occurred in the two lineages has to be the same as well, as illustrated in (12).

(12) $\frac{a_{12}}{t_{12}} = \frac{a_{13}}{t_{13}} \rightarrow r_{12} = r_{13}$

However it is important to underline that, even if theoretically the restriction imposed by the molecular clock hypothesis appears to be very strong, in practice it may be satisfied also if the rates of evolution in the different branches of a phylogeny are all nearly equal, i.e. show at least a tendency in this sense. Thus, for instance, UPGMA can give “misleading results if the distances actually reflect a substantial nonclocklike tree” and “[f]or this problem to arise, evolutionary rates on different branch lengths must differ by at least by a factor of two” (Felsenstein 2004a: 165-166, see the example provided by the author).

Unlike clustering algorithms, the least-squares methods, conceived by Cavalli-Sforza and Edwards (1967) and popularized after Fitch and Margoliash (1967), cope directly with the problem of non-additive distances and in their basic implementation do not assume the molecular clock hypothesis[23]. The aim of these procedures is of minimizing over the whole tree the difference between the set of distances that are given in input and those that are reconstructed in the graph, i.e. the lengths of the branches that separate the leaves. The tree chosen by a least squares method is that, in which the sum, say *Q*, of all variations between each observed distance ($D^{Obs}{}_{i,j}$) and its correlated expected distance ($D^{Exp}{}_{i,j}$) in the tree has the lowest value. The formula in (13) formalizes this calculation:

(13) $Q = \sum_{i,j} w_{i,j} (D^{Obs}_{i,j} - D^{Exp}_{i,j})^2$

For the moment consider only the core of the formula, which is ($D^{Obs}{}_{i,j}$ - $D^{Exp}{}_{i,j}$), without taking into account the squaring and the term w_{ij}. The central part of the formula means that, given an unrooted tree whose leaves correspond to the taxonomic units of the dataset, for any pair of units (*i* and *j*) the algorithm considers their distance given in the matrix and a possible distance projected onto the tree, then it calculates the difference between these measures, i.e. the error occurred in the reconstruction, say $e_{i,j}$. The calculation is repeated by the algorithm until any possible pair of taxonomic units of the dataset is examined and a complete set of $e_{i,j}$ is obtained. Finally, all these values are summed up to derive *Q*, which provides a score for the phylogenetic tree under study.

Now consider the term w_{ij} , which stands for “weight”. According to (13), $e_{i,j}$ has to be normalized with a specific weight w_{ij}, which characterizes each least squares method: “The

---

[23] Notice however that in PHYLIP (Felsenstein 2004b) the Fitch-Margoliash method is implemented also in a clocklike fashion (i.e. including a molecular clock) in the program Kitsch. See section 6.1.1 for further observations.

difference between the methods is how they weight departures of observed from expected" (Felsenstein 2004b). The Fitch-Margoliash method (Fitch and Margoliash 1967) assumes $w_{ij} = 1/D_{i,j}^2$. Why? This assumption can be explained starting from some basic concepts of probability theory. Suppose one has n observation (O_n), each one taken from a different normal probability distribution. If he wants to confront (e.g. to sum) the individual deviations of these observations from the average of each distribution (M_n), i.e. $(O_n - M_n)$, he needs to normalize them with respect to the standard deviation of each distribution (s_n): Thus a normalized individual deviation is $(O_n - M_n) / s_n$. In order to avoid negative values of the individual deviations, he also has to square the fraction: $(O_n - M_n)^2 / s_n^2$. Notice that the squared standard deviation (s_n^2) corresponds by definition to the variance V_n of the distribution. After having normalized the individual deviations, one may for instance sum them to obtain a value Q' as in (14).

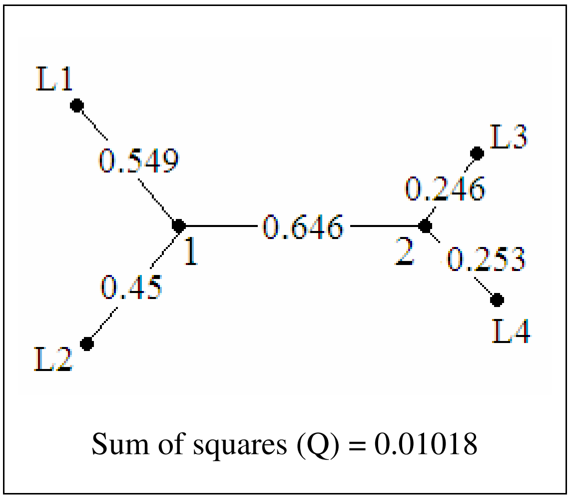
$$(14) \quad Q' = (O_1 - M_1)^2 / V_1 + (O_n - M_n)^2 / V_n + \ldots + (O_n - M_n)^2 / V_n = \sum_n \frac{(O_n - M_n)^2}{V_n}$$

As for the least squares methods, the observed distance ($D^{Obs}_{i,j}$) plays the role of the observation (O_n) and it is assumed that each observed distance derives from a different normal probability distribution. Therefore if one wants to sum the individual deviations of each observed distance from its corresponding expected distance ($D^{Exp}_{i,j}$), i.e. what was called error ($e_{i,j}$), he has to normalize the individual deviations with respect to the standard deviation of each distribution. The Fitch-Margoliash method assumes that the standard deviation is proportional to the observed distance, i.e. it has the same trend, because it is plausible to assume that the standard deviation is not the same for small distances as it is for large, but it is higher in the case of larger distances, which relate taxonomic units that are genealogically more distant and deviated earlier in the past from a common ancestor. In other words, according to the Fitch-Margoliash method the noise accumulated in the long period is expected to be greater that the noise accumulated in the short period. As a consequence, the standard deviation (s_n) of (14) is replaced with $D^{Obs}_{i,j}$, obtaining the normalized individual deviation $(D^{Obs}_{i,j} - D^{Exp}_{i,j}) / D^{Obs}_{i,j}$. Finally, since the individual deviations are squared to avoid negative values, the $D^{Obs}_{i,j}$ that lies in the denominator is squared too, and one obtains the squared normalize individual deviation, i.e. the squared normalized individual error: $(D^{Obs}_{i,j} - D^{Exp}_{i,j})^2 / D^{Obs}_{i,j}{}^2$. Observe that in the context of the least-squares methods the variance corresponds to the squared observed distance: $V_n = D^{Obs}_{i,j}{}^2$. Referring back to the formula in (13), now the reason why the Fitch-Margoliash method assumes $w_{ij} = 1/D_{i,j}^2$ should be clearer. Indeed, the formula in (13) adapted for the Fitch-Margoliash method corresponds to the formula in (15).

$$(15) \quad Q = \sum_{i,j} \frac{(D^{Obs}_{i,j} - D^{Exp}_{i,j})^2}{(D^{Obs}_{i,j})^2}$$

Among possible alternatives would be that proposed by Cavalli-Sforza and Edwards (1967), who assume $w_{ij} = 1$: The scholars hypothesize that the standard deviation of each distribution is equal, i.e. it is the same for short and long distances. This means that Cavalli-Sforza and Edwards (1967) consider each error ($e_{i,j}$) independent of the distance, but such hypothesis may be thought to be less realistic than Fitch and Margoliash's (1967) assumption.

The tree inferred using the Fitch-Margoliash method on the basis of the non-additive distance matrix of figure 8a is given in figure 11, together with the corresponding sum of squares $Q = 0.01018$. Notice that, a NJ, also this procedure basically (see note 22) reconstructs unrooted trees, which then may be provided with a root using the outgroup criterion.



Sum of squares (Q) = 0.01018

Figure 11: The unrooted tree inferred by means of the Fitch-Margoliash method from the non-additive distance matrix of figure 8a; the sum of squares *Q* is specified.

All the methods presented thus far permit inferring phylogenies in form of trees, that, more specifically, are binary branching graphs: In other words, these methods are exclusively aimed at identifying genealogical paths, but on their own they cannot assist one to identify contact events between different lineages. A summary of the specific techniques that can be used to detect events of horizontal transmission is offered in the next section.

4.3.3. Detecting horizontal transmission

An overview of the history of the quantitative procedures used to detect horizontal transmission of features can be found in Wang and Minett (2005: 122-126). One partially qualitative method derives from Hübschmann's (1875) "stratification technique", which was applied by the XIX-century scholar to Armenian: It consists of "stratifying" the correspondences between two lineages and recognizing only the oldest stratum as bearer of the vertical transmission. This method is used by Wen (1940) and Sagart and Xu (2001) on a Sino-Tibetan corpus, as well as by Wang (2004) in the so-called "distillation procedure".

Sankoff (1972) and Embleton (1981; 1986), working in a lexicostatistic paradigm, suggest to model borrowing in terms of geographic neighbourhood between languages: According to this approach, the more two languages are spatially close, the more probable productive contact occurred between them, thus a factor can be inserted in the analysis to correct their similarity index. However this procedure does not properly detect borrowing, but rather is supposed to improve the normal genealogical inference. At this point it is worth noting that Cavalli-Sforza and Wang (1986), on the basis of the analysis they carry out on a body of lexical data (see section 5.1.1) from a chain of Micronesian islands, show that actually there is an important correlation between the geographic distance of human groups and the lexical diversity of the languages spoken by these groups, although this correlation may vary with respect to the different rates in which diverse word classes tend to change.

Another lexicostatistic technique designed to detect borrowing is Wang's (1989), who takes into consideration the values encountered in the "error matrix" as possible indicators of borrowing: This table of values includes the deviations between the observed distances and the optimal distances that are expected from the ideal phylogenetic tree. Though intriguing, the reliability of this idea has not yet been proved. Within the same field of study, Heine (1974) and Hinnebusch (1999) propose looking at the "skewing" in the lexical similarities between languages in order to detect possible horizontal transmission. The scholars' premise is the observation (corroborated by much Africanist literature) that the lexical similarity between closely genealogically related languages is inclined to be equal when in the history of the *taxa* there has been only vertical transmission, but tends to significantly diverge in the case of contact-induced interference. Thus Hinnebusch (1996) suggests measuring the difference that is observed between the similarity indexes, i.e. the “skewing”, and to use it in order to estimate the possibility that borrowing occurred. This idea is statistically tested in Wang and Minett (2005), where it is shown to be a fairly robust method. In Minett and Wang (2003), instead, it is statistically falsified the hypothesis, according to which branches of negative length inferred by distance-based phylogenetic methods can indicate horizontal transmission.

Within the field of population genetics, Cavalli-Sforza *et alii* (1994: 35) propose to use the bootstrap (Bradley 1979; Felsenstein 1985), a technique that tests the robustness of the genealogical information contained in a dataset, to find out areas of the phylogeny in which “mixtures” between human groups occurred. In each cycle of this procedure, from the original character-based dataset a new one is created, which has the same size of the original, but in which some characters have been deleted and some others multiplied to replace the deleted ones. It is possible to demonstrate that the random variation of “bootstrapped” data corresponds to the variation that it is expected from collecting new data in the same domain (Felsenstein 2004a: 338). The bootstrapping technique is widely used in the experimentation presented in chapter 6, even if with a different aim, and an extensive account for it is provided in section 6.1.3. For the moment, what is important to say is that Cavalli-Sforza *et alii* (1994: 35) argue that the more a taxonomic unit manifests instability, i.e. it shifts from one sub-group to another at each cycle, the more it is likely to have undergone horizontal transmission from other taxonomic units in its history, because it shares significant amounts of features with multiple lineages. Ogura and Wang (1996) use this method on a IE lexical dataset and correctly detect the borrowing occurred from French and the Scandinavian languages into English.

Other relevant quantitative procedures are those based on the concepts of “parsimony” and “character compatibility”, which play an important role in the quantitative methods aimed at detecting borrowing. For instance, Minett and Wang (2003) suggest finding the maximally parsimonious trees and assume that each innovation should arise independently only once, considering all innovations after the first as contact-induced. This method has been applied to seven main dialectal sub-groups of Chinese obtaining uncertain results. Nakhleh *et alii* (2005a), instead, propose looking at horizontal transmission in the history of incompatible characters: The reason why homoplastic character histories may be suspected of borrowing is essentially that their observed state distribution can be explained also without back-mutation and parallel evolution, if one assumes an edge that connects the relevant lineages, i.e. a convergence due to contact between different lineages. Recall for instance the reconstructed history of the homoplastic character *B* represented in figure 5: In the first scenario a connecting edge between the branch that leads to *L2* and the branch that leads to *L4* can be drawn, so that the apparent parallel mutation is substituted with the horizontal transmission of the – state between *L2* and *L4*. Nakhleh *et alii*’s (2005a) aim at finding the largest set of characters that are all pair-wise compatible with respect to the same phylogeny of the Indo-European languages (i.e. the “clique”). The scholars’ research is reconsidered in section 5.1.2, within the discussion aimed at presenting quantitative experiments in historical linguistics. Observe that the maximum-compatibility procedure is currently being investigated also within the answer-set-programming framework (Erdem *et alii* 2003; Brooks *et alii* 2005), showing promising results. The question that has to be posed concerns its fitness respect to linguistic data: Is it plausible to assume that the vertical transmission of linguistic characters occurs mainly in a compatible fashion? In other words, is it empirically grounded to argue that apparent back-mutation and parallel evolution are possible indicators of borrowing? Actually

Nakhleh *et alii*'s (2005a) method is particularly relevant respect to the experiment presented in this dissertation, the issue discuss this issue in more detail in section 6.3.1.
Finally, Bandelt and Dress (1992) present a character-based method called "split decomposition", which is based on the notion of split presented in section 4.3.1. Recall that a perfect phylogenetic tree should be based on a fully compatible split set ∑ (Huson 1998: 68), that, as pointed out, means the same as saying that a perfect phylogenetic tree should be based on a clique of compatible characters. Bandelt and Dress' (1992) split decomposition method does not force a tree-like topology, but rather it transforms the set of splits ∑ into a "splits graph". When ∑ is compatible the result is a normal tree, otherwise it is a network: The more is the "split incompatibility" detected in the dataset, the more marked is the departure from a tree-like topology. Computational implementations of this method are provided by Bandelt *et alii* (1995), who designed the software "Network"[24], Huson (1998), who created the interactive program "SplitsTree", and by Bryant e Moulton (2004), who combined aspects of the split decomposition analysis and of the standard NJ method to developed the algorithm "NeighborNet". Bryant *et alii* (2005) apply NeighborNet to the lexical dataset of Dyen *et alii* (1992), showing that according to this method the evolution of IE languages appears to have occurred in a largely tree-like fashion, while Ben Hamed (2005) uses it on Chinese dialects, finding a situation of pervasive dialectal continuum. In philology, both SplitsTree and NeighborNet are used to reconstruct the evolution of a simulated textual tradition in Baret *et alii* (2006): The results obtained by the two approaches are very close to each other, but hardly comparable to the real stemma. The reasons why the split decomposition technique may be inadequate to deal with specific datasets are worked out in section 6.3.1.

4.4. Conclusion

In this chapter a survey of various methods by means of which it is possible to study evolutionary paths, with particular reference to those concerning languages, has been offered. The first step has been to define the direction of inquiry, i.e. what is meant to say "to study language evolution", explaining the role of genealogical classification in historical linguistics. Moreover, it has been illustrated the way in which genealogical hypotheses and events of areal convergence may be represented providing a formal account for tree-graphs and network-graphs. On this basis, it has been possible to illustrate diverse methods of comparison that have been developed and applied in historical linguistics: The classical comparative method, the lexicostatistic analysis and mass comparison. It has been argued that, among these, only the classical method may be considered a scientifically reliable and consistent procedure, but that it can operate only under specific conditions. In particular, it appears to be not universally applicable due to the fact that it is grounded on the detection of relatively rare evidence, and, as a consequence, it is limited in the time-depth at which it can operate. In order to offer a solution to overcome these limitations, a different comparative method has been presented, whose design is in line the recent advances achieves in diachronic syntax. The *comparanda* of this method are precisely syntactic parameters, which entail various advantages for the comparative procedure: As the characters used in population genetics, parameters constitute a universal biologically-determined finite list of discrete options, thus, also in virtue of the inertial characterization of syntactic evolution, appear to be particularly apt to shed light on long-distant genealogical relationships. Moreover, parameters provide a precise and confident method of measuring the degree of similarity (or, conversely, dissimilarity) between languages and their discrete nature makes them good candidates as input for quantitative procedures of phylogenetic reconstruction.

---

[24] See McMahon and McMahon (2005: 144-145) for a sketch of the program and further references.

Thus the last section of the chapter has covered the introduction of some relevant quantitative techniques aimed at reconstructing evolutionary histories. It has been pointed out that these phylogenetic methods were developed since the ‘50s in systematic biology and population genetics and have been recently applied to different fields of study, such as philology and, above all, historical linguistics. In particular, the role that this transfer of technology may have for the development of the “new synthesis” between population genetics, archaeology and linguistics has been underlined. Both character-based and distance based methods have been considered and the focus has been on those that have been used to carry out the experiments presented in this dissertation. Finally, an overview of some quantitative procedures specifically designed to detect horizontal transmission has been provided. Throughout section 4.3 only incidental mention has been made regarding the application of the techniques in question to actual linguistic datasets: To carefully illustrate them is the first objective of the next chapter.

5. The dataset

The topic of this chapter is the illustration of the data on which the experimentation is based, which are comprised in the so-called “Table A”. However, the further discussion entails important digressions from the main theme and, in particular, it begins with a survey of other relevant linguistic datasets collected and quantitatively analyzed in the last decades. This summary is intended not only to help one to better appreciate the introduction of Table A, but also to offer a more accurate insight in the line of research in which this dissertation may be contextualized. Moreover, after having described the parametric data, it will be fundamental to carefully take into account which hypotheses regarding the genealogical classification of the languages in question have been put forward so far. This is done especially with the aim of grounding the qualitative evaluation of the phylogenetic results presented in chapter 6.

The discussion is organized into three main sections. The first one covers an overview of the principal datasets used so far to perform quantitative investigations on the history of languages. These studies, which have been carried out mainly but not exclusively by linguists, are ordered according to the type of data the scholars have employed: In particular, section 5.1.1 is specifically dedicated to the illustration of lexical-based analyses, while section 5.1.2 shows how relevant experiments have been conducted on the basis of other types of linguistic data. In section 5.2, instead, the parametric dataset is illustrated: Since, as will be explained, it consists of the comparative analysis of the DP across various languages, its introduction is preceded by a review of the status of nominal syntax in generative grammar, as well as by the presentation of the specific theoretical background on which the parameter set incorporated in Table A has been designed. Finally, section 5.3 illustrates some relevant past and current proposals regarding the genealogical classification of the languages that are considered in the present investigation: This summary concerns both hypotheses conceived within the traditional paradigm and those derived from the recent quantitative perspectives adopted in some of the works already considered in section 5.1. Moreover, a brief account is provided for more controversial suggestions regarding long-range genealogical relationships.

5.1. Linguistic datasets in quantitative studies

The importance that the quantitative perspective is assuming in historical linguistics has already been underlined in the beginning of section 4.3, with particular reference to the consequences that this methodological shift may have for the achievement of the new synthesis (section 4.2.3) and referring to a number of experiments carried out by different scholars. At this point some of the key works on the subject require reconsideration in light of the type of comparative data on which they are based. In particular, the objective pursued in the following sections is to distinguish between the experiments performed on the basis of lexical datasets and those carried out relying on non-lexical *comparanda*. Actually the former are much more common than the latter, and this fact might be drawn back to the reasons worked out (especially) in section 4.2.4. In sum, the tradition of studies in historical linguistics has been prominently focused on lexicon, which permits identifying evident units of comparison. Thus, on the one hand a great amount of comparative lexical data has been collected within lexicostatistics since the ‘50s and it is ready to be employed within the quantitative perspective, on the other hand one might think that lexical datasets are more easily understandable and manageable also for non-linguists, who play an important role in the field of study in question. However, as shown starting from section 5.1.2, relevant quantitative insights on language variation and evolution can be, and actually have already been, provided also by means of morphological, phonological and, above all, syntactic comparison.

5.1.1. Lexical datasets

Recall from section 4.2.2 that the first lexicostatistic meaning list dates back to Swadesh (1952) and comprises 200 items plus 15 already considered for disposal by the scholar; In Swadesh (1955) some more glosses are eliminated, in order to more consistently follow the methodological criteria of the lexicostatistic paradigm and, in particular, to avoid the situations described in Swadesh (1955: 124-126): The result is what is currently known as the 100-item list. This inventory is referred to as "high rank" list by Wang and Wang (2004), in the sense that the items it contains appear to be particularly stable and loan-resistant[1]. However, notice that the data source for many recent works is actually the original 200-item list, in particular the version outlined by Isidore Dyen for 95 Indo-European languages in the '60s and then worked out to obtain pair-wise cognate percentages. The main reference to this data is the dataset drawn by Dyen *et alii* (1992), who provide extensive lexicostatistic information regarding Indo-European languages but do not infer genealogical trees by themselves.

As already argued, not only the choice of meanings, but also the estimate of cognacy between the words in which they are encoded are fundamental. Essentially, this evaluation can be made only by applying the classical comparative method to the word lists. To this end some recent proposals designed in a quantitative-oriented perspective deserve special attention. For instance, one may think about Kondrak's (2002) automatic procedure to determine recurrent sound correspondences in bilingual wordlists: His method is developed within the line of inquiry aimed at designing cognate-identification algorithmic techniques (see the references provided in Kondrak 2002). Another relevant insight is that offered by Nichols (2006), who outlines diverse criteria to evaluate the degrees of freedom, both in form and meaning, that can and cannot be accepted when one wants to establish that two words are cognates: More specifically, she defines the concept of "quasi-cognates", intended as pairs of words that may not be immediately identified as etymologically related, but for which specific criteria of resemblance are satisfied.

How are lexical data effectively encoded in quantitative phylogenetic studies? In order to clarify the matter, it is opportune to start taking into account how they are translated into character matrices, and only afterwards to consider the derivation of distance measures. As for the former point, one of the clearest examples is provided by Atkinson and Gray (2006b), who carefully introduce the issue. In this work, that synthesizes the methodological and experimental line of inquiry pursued in Gray and Atkinson (2003), Atkinson *et alii* (2005) and Atkinson and Gray (2006a), the scholars focus on two alternative views on the prehistory of the Indo-European languages, already mentioned in section 4.2.3: The "Kurgan hypothesis" (Gimbutas 1970) and Renfrew's (1987) idea of the Anatolian origin. The specific conclusion at which Atkinson and Gray (2006b) arrive in their study will be considered in section 5.3.2. Here it is important to observe that the authors' experiment relies on Dyen *et alii* (1992) dataset and, in particular, it contains the 200 meanings instantiated in 87 Indo-European languages. In the coding adopted by Atkinson and Gray (2006b) each meaning is associated with different lexical roots and each root constitutes a binary character by means of which

---

[1] Wang and Wang (2004) contrast the 100 "high rank" words with the "low rank" ones, that are those remaining from the original 200-item list after the exclusion of the high rank items. The scholars empirically support this distinction showing that high rank words appear to be more apt to genealogically classify Chinese dialects.

languages are described: If a certain root is found in a language, the relevant character-state in that language is 1, otherwise it is 0. For instance, they consider that the meanings for "here", "sea", "water" and "when" are instantiated in English, German, French, Italian and Hittite as shown in (1).

(1) English: *here*, *sea*, *water*, *when*;
German: *hier*, *See/Meer*, *Wasser*, *wann*;
French : *ici*, *mer*, *eau*, *quand*;
Italian: *qui/qua*, *mare*, *acqua*, *quando*;
Modern Greek: *edo*, *thalassa*, *nero*, *pote*;
Hittite: *ka*, *aruna*, *watar*, *kuwapi*.
(adapted from Atkinson and Gray 2006b: 94)

The scholars attribute four different roots to the meaning "here", four to "sea", three to "water" and one to "when": In this way they are able to draw the character matrix reported in table 1. On the whole, the number of characters that they derive from the original 200-item list is 2449. Atkinson and Gray (2006b) prefer not to directly deal with meanings, i.e. not to associate to each gloss a multiple-state character, because "coding the data as semantic categories makes it difficult to deal with polymorphisms" (Atkinson and Gray 2006b: 94). Polymorphisms would arise due to the fact that languages may have more lexical roots for the same meaning (e.g. *See/Meer* in German). Even if the authors' solution permits overcoming this difficulty, it is problematic in other respects. First, because it creates dependence between characters. Indeed, as observed by Evans *et alii* (2006), once in a language a meaning m_i is instantiated in a lexical root l_1 (i.e. the state of the character l_1 is 1), the probability that in the same language another root, say l_2, is used to express the same meaning (i.e. the probability that the character l_2 has the state 1) decreases. This aspect represents a difficulty because Atkinson and Gray (2006b) analyze the data by means of a maximum likelihood procedure, which, as pointed out in section 4.3.1, requires characters to evolve independently with respect to the other. The scholars are aware of the objection and reply that actually "the assumption of independence does not hold for nucleotide or amino acid sequence data either" but "biologists still get reasonably accurate estimates of phylogeny despite violations of this assumption" (Atkinson and Gray 2006b: 103). Moreover, the binary coding of individual multi-state semantic features, i.e. meanings, creates redundancy in the dataset, and one may ask how much this influences the phylogenetic reconstruction. On this point the scholars claim that, according to Pagel and Mead's (2006) demonstration, redundancy is not suppose to influence the structure of the phylogeny in itself, i.e. to modify the genealogical classification, but just to cause the inference of the same tree with different branch lengths: "In other words, the binary and multi-state trees are just scaled versions of one another" (Atkinson and Gray 2006b: 103).

At this point should be noted that the binary encoding assumed by Atkinson and Gray (2006b) is not novel, since it is already entailed by Dyen *et alii*'s (1992) formalization of the data into cognate sets, and it represents the standard in quantitative character-based experiments grounded on lexical comparison. One may recall for instance Gray and Jordan (2000), who attempt to shed light on the controversy between two competing hypothesis regarding the settlement of Austronesian-speaking groups throughout the Pacific: The "express-train" (Diamond 1988) and the "entangled-bank" (Terrell 1988; Terrell *et alii* 1997) hypotheses. They compare 77 Austronesian languages with respect to 5185 lexical items, which are collected in Robert A. Blust's Austronesian Comparative Dictionary (the work is in progress): Each one is provided with a set of cognate terms and lists of the languages in which these cognates appear. Gray and Jordan (2000) perform a maximum parsimony analysis of the dataset and the phylogeny they obtain appears to be more compatible with Diamond's (1988) reconstruction than with Terrell's (1988). The recent insight provided by Gray *et alii* (2009) essentially confirms this result: In particular, Gray *et alii*'s (2009) experimental reconstruction is done using a refined version of Gray and Jordan's (2000) dataset and Bayesian analysis.

It might be argued that the problems discussed above do not derive from the binary coding, but rather from the nature of the data. This assertion may be better understood in light of a problem (in particular) that arises when one tries to quantify precisely the similarity between languages, which, since the dawn of lexicostatistics, is intended as the percentage of cognates shared by each pair of languages. This assumption is made, for instance, by Cavalli-Sforza and Wang's (1986), who aim to determine the influence of geographic distance on lexical similarity between different communities applying a specific mathematical model (see section 2.1.4) to an actual linguistic area. In order to derive a measure of "linguistic similarity" between each pair of the 17 Micronesian languages they consider, Cavalli-Sforza and Wang (1986) adapt the lexical data collected by Quackenbush (1966) and obtain a character matrix comprising 571 glosses (i.e. meanings). As done by Atkinson and Gray (2006b), the different lexical roots that are instantiated for each gloss in the 17 languages are considered separately in order to create the binary-character matrix. In particular, drawing a parallel with population genetics, Cavalli-Sforza and Wang (1986) explicitly assume the convention according to which "a gloss is taken as the equivalent of a specific gene, and sets of cognatic morphemes of that gloss are taken as the alleles" (Cavalli-Sforza and Wang 1986: 43). Then the scholars calculate the similarity between each pair of languages as a percentage of shared cognates, and find that actually lexical variation is strongly (but peculiarly) connected to an areal neighbourhood. Among the others, McMahon and McMahon's (2003) experiment follows a similar line: The authors use Dyen *et alii*'s (1992) cognate percentages to build a distance matrix including the relevant 95 Indo-European languages. Distance measures are calculated as complements of the percentages: For instance, if the percentage of shared cognates between Italian and French is 80,3%, the corresponding distance between the Romance varieties is assumed to be 100 - 80.3 = 19.7. They can then employ these data as input for distance-based phylogenetic methods, i.e. the Fitch-Margoliash method and Neighbor-Joining[2]. Further comments on McMahon and McMahon's (2003) results are not provided here, but in the next chapter a brief comparison between the same dataset adopted by the scholars (or better, a part of it) and the parametric one will be introduced.

However, what problem occurs with a distance based on cognate-percentages? Essentially, that all cognates have the same weight while comparing two languages, regardless if in a certain language they are one of the possible instantiations of the same meaning (i.e. they have synonyms) or not. For instance, recall from the previous discussion that German has two words to express "sea": The first, *See*, is a cognate of English *sea*, the second, *Meer*, is etymologically related, among others, with Italian *mare*. In this way the same meaning makes the similarity of German both with Italian and with English to increase by one point (of course not in percentage, but in the absolute value). The same increase derives, for example, from the match between German *hier* and English *here*, which in both languages are the only words available to express the relevant meaning. Nonetheless it is not straightforward that one should quantify the increase of the similarity score between German and English entailed by *See/sea* on the one side and that derived from *hier/here* on the other in the same way. Actually the matter is already considered by Swadesh (1955: 125), who simply suggests not to insert certain words in the meaning lists (e.g. precisely *sea-river-lake*, all replaced by *water* in the 1000-item list) in order to avoid duplications due to the occurrence of synonyms.

---

[2] The algorithmic implementations of the methods used by McMahon and McMahon (2003; 2005) are those provided in Felsenstein's (2004b) package of program PHYLIP. See section 6.1.2 for further details.

5.1.2. Non-lexical datasets

The non-lexical datasets presented in this section include phonological, morphological and syntactic *comparanda*. The first relevant example is Ringe *et alii*'s (2002), which covers 24 well-attested ancient and modern Indo-European languages (the reason why they are allowed to mix non-coeval languages in the same experiment is clarified in section 6.1.2). Following the traditional paradigm, these varieties are already assumed to be classified according to 10 genealogical groups (Anatolian, Indo-Iranian, Greek, Italic, Germanic, Armenian, Celtic, Tocharian, Balto-Slavic, Albanian) and the authors' aim is precisely to investigate the relationships between them. Actually the dataset they use is not entirely non-lexical, but hybrid, in that it comprises 22 phonological characters encoding regular sound changes, 15 morphological characters (mainly) related to word inflection and 333 lexical characters. In (1a), (1b) and (1c) instances, respectively, of phonological, morphological and lexical characters are introduced; The same traits are also reported in Table 2 together with the states that they show in the languages included in the experiment.

(1) a. "Character P16 = Sequence of changes (a) Grimm's law; (b) Verner's law; (c) initial-syllable stress; (d) merger of unstressed _*e_ with _*i_ except before _*r_.
Character states = 1: absent (ancestral); 2: present.

b. Character M8 = Most archaic superlative suffix.
Character states = 1: _*-isto-_; 2: _*-ismo-_; 3-13: other or none.

c. Lexical character 348: "grind"
Character states = 1: _*molh_$_2$- or _*melh_$_2$-; 2: _h_$_2$_elh_$_1$- or _alh_$_1$-; 3-13: other or none" (adapted from Ringe *et alii* 2002: 113-125).

Table 3 summarizes the authors' evaluation of each class of characters. In synthesis, they expect phonological characters to provide clear evidence regarding the direction of innovations, because as already argued, phonemes often follow known patterns of change and are typically not subject to back-mutation (e.g. recall the irreversibility of mergers); Nonetheless, "the individual (phonological, editor's note) changes are usually so 'natural' that they can be repeated in different lines of descent" (Ringe *et alii* 2002: 66), so that parallel development of these traits is rather common. On the contrary, morphological characters can offer strong proof of proper genealogical relatedness, essentially for the same reasons outlined by Nichols (1996) discussing the individual identifying value of inflectional paradigms (section 4.2.1), but, even if morphological homologies are clear, they do not reveal anything about the direction of shared innovations. Finally, lexical characters are regarded by Ringe *et alii* (2002) as the weakest ones, in the sense that semantic innovations may easily develop parallely in different lineages, directions of change are rarely recoverable and the history of these traits is often biased by the presence of borrowing. Moreover polymorphism affects many lexical characters comprised in this dataset as well, as in the case of those considered in the previous section. Ringe *et alii* (2002) treat this issue explicitly (e.g. in connection to the forms *small* and *little* in English, both meaning "not large"), observing that actually it is hard to deal with, thus often they resolved polymorphic characters into pairs of monomorphic ones, i.e. splitting them (more details are provided in Ringe *et alii* 2002: 83-85). On the whole this dataset appears to be markedly grounded on many traditional insights of the classical comparative paradigm: It has the same reliability but also the same range of application, thus it is hardly replicable for other linguistic domains. On the other hand, it fundamentally differs from the lexicostatistic-based ones because, including qualitatively different characters, it incorporates knowledge of the direction of evolutionary paths and, in the phonological and morphological components, it is relatively protected against homoplasy, i.e. back-mutation and parallel evolution. As for the experiment carried out by Ringe *et alii* (2002), the scholars use the whole character matrix of 370 characters as input for a specifically designed maximum-compatibility procedure of phylogenetic reconstruction, in that they are interested in finding what they call "perfect phylogeny", i.e. essentially the ideal tree of the 24 languages in question onto which the history of all the characters can be projected without homoplasy. The outcome they obtain shows that this condition is largely satisfied, except for about 18 characters (all of which are lexical) that appear to have undergone homoplastic evolution. However, an accurate account of the results is offered in section 5.3.2, where some relevant hypotheses derived from other studies performed in the same line of research and on the basis of a largely identical dataset are also introduced (i.e. Nakhleh *et alii* 2005a).

Examples of syntactic datasets employed in recent quantitative-oriented works may be found, among others, in Dunn *et alii* (2005), Gray 2005, Ryder 2006 and in particular in the special issue of *Lingua* introduced by Nerbonne (2009b). Here the focus is only on a specific investigation, i.e. that performed by Spruit (2005; 2008) on the variation of Dutch dialects. It is important to point out that the scholar's investigation is properly grounded on a quantitative comparison methodology, nonetheless its aim is not of obtaining a genealogical classification of the varieties in question, but rather of studying their variation in connection with their geographic distribution, i.e. of performing a dialectological analysis of the Dutch area[3]. Observe that Spruit's (2005; 2008) work represents a contribution to a wider line of research, which includes geographic-oriented inquiries on the diachronic patterns of diffusion accomplished by means of data concerning diverse linguistic levels (among others, see for example Heeringa 2004; Nerbonne and Kleiweg 2007; A survey on quantitative dialectology is provided in Nerbonne 2009a). The data considered by the scholar are taken from the Syntactic Atlas of the Dutch Dialects (SAND1, Barbiers *et alii* 2005), which includes 145 maps that show the geographic distribution of 507 syntactic variables and describe 267 Dutch dialects: More specifically, SAND1 "covers syntactic domains related to the left periphery of the clause and pronominal reference" (Spruit 2008: 48). Thus the data concern complementizers, fronting phenomena, subject pronouns, reflexive and reciprocal pronouns and so on. A representative example of the way in which they are provided in the atlas is given in (2).

(2) "Context: Weak reflexive pronoun as object of inherent reflexive verb.
Variables: {*zich*, *hem*, *zijn eigen*, *zichzelf*, *hemzelf*}
Example: Jan herinnert zich dat verhaal wel.
Jan remembers himself that story AFFIRM
John certainly remembers that story"
(adapted from Spruit 2008: 48)

Table 4 shows how the variables in (2) are coded by Spruit (2008: 49). Each variable is a possible instantiation of a certain syntactic element (in this case, a specific type of reflexive pronoun) and it corresponds to a distinct character in the matrix: Different dialects (e.g. Lunteren and Veldhofen) may or may not realize the syntactic element in question by means of one of the available forms, thus they are assigned respectively a + or a -. The distances are

---

[3] "It is worth keeping in mind that dialectology and historical "classification" may not be exactly the same endeavour […]. One important difference between these two is their relation to geography, which influences the distribution of dialectal varieties massively, but not necessarily discretely. Thus Heeringa and Nerbonne (2001) show how the dialectal analysis provides an analytical foundation for the notion 'dialect continuum', in which classification into discrete groups, the very heart of phylogenetic analysis, plays no role" (Nerbonne 2007: 1).

calculated comparing the strings of binary values of each pair of dialects: When in a locus different states of a certain variables are found, the score of the distance between the linguistic varieties is incremented by 1. This measure of dissimilarity, that will be taken into account again in section 6.1.1, is among the simplest possible, in that it requires counting the number of differences between two strings of equal length, and is called "Hamming distance" (Hamming 1950). Spruit (2008) then draws the distance matrix that collects the entire set of pair-wise distances and applies to it the "multidimensional scaling" procedure (Torgerson 1952; For an introduction of the technique see among others Manly 2005: 163-176). In this way "the set of 267 dialectal dimensions for each dialect is scaled down to a coordinate in a three-dimension space" (Spruit 2008: 50) and, in synthesis, it is possible to obtain a map of the geographic area in which the location of each dialect is associated with a different colour and colours gradually fade into others, i.e. the dialectal continuum is represented.

Leaving aside a detailed account of the advanced steps of the analysis, other observations about the data are relevant here. Consider that the variables described and encoded as in (2) and in Table 4 appear to be highly morphologically-oriented and, in virtue of their specificity, they are referred to by Spruit (2008) as "atomic variables", which are opposed to the so-called "feature variables". The latter are designed starting from the same basic data, but coding it at a higher level of abstraction from morphological variation. For example, in Table 5 atomic variables and feature variables respectively head rows and columns: A + at the intersection between a row and a column means that the relevant feature variable, say "reflexive", is instantiated by a certain atomic variable, i.e. either *zich* or *zichzelf*. Thus if in a dialect one finds either *zich* or *zichzelf*, it means that for the linguistic variety in question the feature variable "reflexive" has the value + (otherwise – would be assigned). Table 6 illustrates how the feature variables of Table 5 are set in Lunteren, which has both *zich* and *zijn eigen*, and in Veldhofen, whose only available form is *zich*: Notice that the Hamming formula is adopted in this case as well to calculate distances. Even if it leads to obtaining essentially identical results (Spruit 2008: 58-59), in this study the passage from atomic to feature variables appears to be very important, essentially because it permits acquiring a dataset that is less dependent on morphology and more focused on proper syntactic variation. To understand whether this variation should be considered relevant at typological or at genealogical level is beyond the aim of this introduction and, actually, it is not a primary issue within Spruit's (2005; 2008) dialectological investigation as well. On the contrary, as explained in chapter 4, the syntactic dataset adopted in this work is explicitly designed under the hypothesis that the comparison on which it is grounded can effectively help to recover genealogical information and to shed light on the history of languages. The following sections are dedicated to its illustration.

5.2. The parametric dataset

The parametric dataset has been developed since Longobardi (2003a) and it is grounded on the theoretical criteria presented in section 4.3.4. Thus the choice of the type of *comparanda* has been motivated, but nothing has been said so far regarding the choice of the specific parameters. As shown illustrating the examples in the previous sections, this issue requires attention because it is fundamental if one aims at a sound application of a comparative procedure. Longobardi (2003a: 111-112) points out that ideally a comprehensive syntactic comparison between languages should be grounded on the whole list of parameters. Nonetheless the global parameterization of syntax, even if achievable in principle, is still beyond the possibilities of current investigations. On the other hand, Longobardi and Guardiano (2009: 9) argue that to rely on a random list of parameters, or in general on a set designed in light of non-structural criteria, does not guarantee obtaining "balanced" results. Therefore their suggestion is to attempt to base the application of the comparative method on

the exhaustive account of a significant domain of grammar, which is required to be relatively independent from others, to show enough cross-linguistic distinctions and, for practical reasons, to be particularly well-studied especially within the generative paradigm (Longobardi 2003a: 112). Observe that choosing to investigate syntactic variation within a specific closed set of *comparanda* also allows one to pursue the objective of finding proper individual-identifying evidence, in Nichols' (1996) sense (Longobardi and Guardiano 2009: 9).
This method of inquiry is referred to as "modularized global parameterization" (Longobardi 2003a: 112) and has lead to focusing the attention on the domain of nominal syntax. This grammatical module appears to appropriately satisfy the requirements outlined above, thus it has been chosen to carry out an exhaustive parametric investigation that has permitted to draw the dataset presented in section 5.2.2. However, before coming to its illustration, it is opportune to introduce the theoretical background on which it is based.

5.2.1. The structure of DP

The parametric dataset is primarily designed following recent theoretical descriptions of the formal representation of nominal phrases, that is, in particular, starting from specific hypotheses regarding the structure of DPs. On the basis of the account of the formal aspects of the generative approach provided in section 3.1.3, the importance of framing nominal domain within a DP-oriented, instead of an NP-oriented perspective might not be fully clear, therefore a brief historical contextualization of the issue is worth being provided. Since Chomsky (1970) the idea that "the structure related to a Verb is the same as that related to the corresponding noun" (Giorgi and Longobardi 1991: 1) appears to be rather natural, but until the '80s extensive formal explanations of the internal structure of nominal phrases are not put forward. Chomsky's (1970) assumption starts to be worked out, for example, in Anderson (1979) and Cinque (1980), but nominal phrases continue to be represented as NPs, i.e. essentially as lexical projections, which entails that determiner elements (e.g. definite articles) are assumed to be generated in SpecN. The revision of the X-bar theory proposed in Chomsky (1986b) and its extension to the domain of NPs allows shedding new light on nominal structures. Indeed Chomsky (1986b) suggests that not only lexical elements (e.g. V and N), but also functional ones can project their own phrases. As recalled by Bernstein (2001: 537), this hypothesis is soon recognized to be incompatible with the idea that determiners should be hosted in SpecN: On the one hand nothing prevents these elements to be assumed to be proper phrasal heads, as for instance complementizers (C) and verbal inflections (I), on the other hand to associate determiners with SpecN contrasts with the idea that specifier positions are typically occupied by phrasal categories (e.g. wh-phrases for what concerns SpecC). Therefore, developing some relevant insights already put forward by Brame (1982) and Szabolsci (1983), Abney (1987) "provided theoretical and empirical arguments for the idea that a functional category, Determiner Phrase (DP), is the maximal projection of the lexical head, the noun" (Bernstein 2001: 537), so that he proposes a structure like that given in (3), where NP plays the role of complement with respect to the D head.

(3) $[_{DP}$ SpecD $[_{D'}$ D [NP]]]

Abney's (1987) "DP Hypothesis" paves the way to many new insights on nominal syntax, because it allows drawing an explicit parallel between the organization of clauses and that of known phrases, in particular for what concerns argument structure and word-order. A comprehensive overview of the studies dedicated to these points is given for instance by Bernstein (2001), while, in order to focus the attention on the line of inquiry that is more relevant to the aims of this section, the following discussion mainly covers the concise review of the DP characterization provided by Longobardi (2001b).

The scholar introduces his analysis observing that the argumental structure of nouns is not only analogous to that found in clauses (among early mention of this idea see, among others, Ritter 1988; Picallo 1991; Valois 1991), but also appears to be largely ordered in the same way, so that agents (i.e. thematic subjects, S) are hosted in a higher position than themes (i.e. direct objects, O) and other complements of N. Furthermore, in nominal phrases another specific type of argument, which is not relevant at clause level, is assumed to intervene, that is "possessor" (P). Longobardi (2001b: 563-566) claims that, according to evidence related to possessivization (see below) and binding, one can determine that the hierarchical order of nominal arguments is P > S > O. Thus, considering both P and S external arguments and O internal argument of N, the nucleus of the nominal phrase is hypothesized to be structured as in (4).

(4) [P [S [N O]]]

The author observes that "many languages tend to use a special Case, Genitive, normally the same employed to express P, for the arguments of nouns whose verbal thematic correspondents bear Nominative and Accusative" (Longobardi 2001b: 566). In particular, he points out that cross-linguistic analyses show that as a first approximation there are two distinct ways in which Genitive is assigned to nominal arguments: Either the case is expressed in a prepositional (e.g. as commonly happens in Italian) or in a possessivized form. The typology of the latter realization appears to be rather rich, in that it includes various morphological options, i.e. final phrase suffix (e.g. English *'s*), final word suffix (e.g. German *s*), a dedicated inflectional ending (e.g. Latin), phi-feature agreement with the noun (e.g. German possessives) and zero-realization (e.g. Hebrew construct state). On the whole, what is underlined by Longobardi (2001b) is that all possessivized genitival forms occur high in the nominal structure, because they often appear before both nouns and adjectives or, if they appear in a postnominal or postadjectival position, in any case they precede prepositional genitives. Thus, considering the variability of the (high) positions in which possessivized forms are hosted and the fact that in German two of them may co-occur (one prenominal and preadjectival, the other postnominal and postadjectival, e.g. *Hannos schönes Porträt Evas* "John's beautiful portrait of Eva") the scholar hypothesizes that the structure in (4) may be extended as shown in (5).

(5) [1 GenS 2 AP 3 GenO [P [S [N O]]]]

In the parenthesized formulation above GenS and GenO correspond to the positions that are thought to be occupied by possessivized, or more properly, structural genitives, while 1-2-3 are intended as the sites in which nouns surface in different languages and AP indicates the adjectival phrase. At this point a relevant parametric generalization may already be put forward, i.e. that "[l]anguages make the parametric choice of activating just one or the other or both" (Longobardi 2001b: 568). So far it is clear that German is a language that activates both, while, say, English sets off only GenS; A language in which exclusively GenO is relevant, instead, is for instance Welsh, as shown in the example in (6).

(6) car y meddyg
car the doctor
"the doctor's car"
(King 2003: 38)

Within this framework, the relative order of adjectives is assumed to be particularly relevant to shed light on movement phenomena inside the phrase. For what concerns the clause-level, consider that since Edmonds (1978), and then, in particular, Pollock (1989), adverbs are assumed to provide information regarding the movements undergone by verbal heads: E.g. comparing French, where adverbs follow verbs, with English, that shows the opposite pattern, in synthesis the above mentioned authors hypothesize that while in the Romance variety movement occurs, so that V raises over adverbs, in English this does not happen. During the ‘90s the analogy between the structure of clauses and nominal phrases naturally leads drawing a parallel between the roles played by adverbs and adjectives also in this respect: The hypothesis developed by many analyses performed in this line of inquiry (e.g. Crisma 1991; 1996; Bernstein 1993; Cinque 1994)[4] is that within the nominal domain the unmarked order is AN, while the order NA found, for example, in Romance, would be the outcome of N raising to higher positions. This proposal appears as a generalization of Longobardi’s (1994) idea, according to which in Romance N-raising may be responsible for the order NA when N are determinerless proper names. The last assertion is reconsidered below in this section, for the moment it is opportune to notice that a finer description of the NP-internal movements of nominal heads is developed adopting a universal hierarchy for adjectives, i.e. assuming that they are ordered with respect to their semantic properties (e.g. Cinque 1994). For instance, Bernstein (2001), in line with the investigation carried out since Bernstein (1991), provides evidence regarding adjective order comparing Walloon with other Romance languages: She concludes that there is a “relatively robust noun movement in the south-eastern Romance varieties and relatively weak noun movement in the north-western varieties” (Bernstein 2001: 549), as shown in (7), where examples from Walloon (7a) and Sardinian (7b) are reported.

(7) a. one bèle bleûve cote (Walloon)
a nice blue dress
“a nice blue dress”
b. tempus malu est faendi (Sardinian)
weather bad is making
“the weather is bad”
(Bernstein 2001: 548)

Longobardi’s (2001b) account, instead, relies on Crisma’s (1991; 1993) proposal according to which, paralleling the order of adverbs suggested by Jackendoff (1972), adjectives are assumed to be organized into the hierarchy illustrated in (8).

(8) Subject-(or Speaker-)oriented adjectives > Manner adjectives > Argument adjectives (Crisma 1991: 1993)

More specifically, then, among manner adjectives Longobardi (2001b) takes into account the distinction between appositive (M1) and restrictive (M2) adjectives (already considered with reference to N-raising in Romance for instance by Bernstein 1993 and Crisma 1996), pointing out that in general Romance prenominal adjectives are M1 (e.g. Italian *il bel vestito*, the nice dress), while Romance postnominal adjectives are typically M2 (e.g. *il vestito bello*, “the nice dress”); Of course, as appears clear from the English translations, in Germanic the distinction between the positions of M1 and M2 relative to N is obscured due to the fact that all adjectives must occur prenominally. Therefore, in light of this characterization, until this point the structure of the nominal phrase may be schematized as in (9), where numbers indicate possible landing sites for N-raising.

---

[4] Further references about this topic may be found in Bernstein (2001: 542) and Longobardi (2001b: 579).

(9) [1 GenS 2 [Subject-oriented [M1 [3 [M2 [4 [Argument [5 [GenO [P [S [N O]]]]

And concerning the D area? As a first approximation, one may say that commonly the position of D and SpecD are assumed to host elements such as articles, demonstratives, possessives, quantifiers and cardinal numbers, as exemplified in section 3.2.2 reporting Giusti's (2001) hypothesis on the development of Romance definite articles out of Latin *ille*. Determiners are thus responsible on the one hand for the definite/indefinite interpretation of nominals, on the other hand for the mass/count reading of singular Ns. Starting from Szabolsci (1987) many scholars have adopted the idea that only DP can be (verbal) arguments, while NPs can not (see also Stowell 1989; 1991). In particular, Longobardi (1994) develops this hypothesis providing evidence from Romance, like the Italian examples in (10ab), in which it is clear that the determinerless noun phrase cannot function as head of the relative clause, but just as predicate of a the copula *è*.

(10) a. Gianni è medico (*che…)
  John is doctor (that)
 b. Gianni è un medico (che)
  John is a doctor (that)
 (Longobardi 1994: 618-619)

This assumption needs to be reconciled with the fact that across languages there are many cases of determinerless nominals that may appear as arguments. First of all, Longobardi (2001b) focuses on a specific class of determinerless nominals, which are headed by mass or plural nouns and are referred to as Bare Nouns (BNs). Notice that the first insights on English BNs date back to Carlson (1977) and the issue has been taken into account by many scholars in the last decades (e.g. Chierchia 1985; 1998). BNs show an interesting cross-linguistic variation, which concerns both their semantic interpretation and their syntactic distribution. The sentences in (11a-d) serve to illustrate that English BNs may be interpreted either as indefinites or as kind-referring names, while in the Romance variety only the former interpretation is available and the latter requires the insertion of a definite article. Moreover, these examples show that the syntactic distribution of Italian BNs is more constrained than that of English, in that in the Romance language these objects are allowed to appear only as verbal complements, while they cannot be hosted in a pre-verbal subject position.

(11) a. (I) *Pomodori furono introdotti in Europa dopo il 1492.
  (*The) Tomatoes were introduced in Europe after 1492.
 b. Tomatoes were introduced in Europe after 1492.
 (adapted from Longobardi 2001b: 583)
 c. Ieri ho comprato pomodori.
  Today (I) have bought tomatoes
 d. Yesterday I bought tomatoes.

The available positions for Italian BNs seems to be only the lexically-governed ones and "[e]specially since Contreras (1986) it has been speculated that such distribution in Romance could precisely be due to the presence of an actual empty category as the invisible determiner" (Longobardi 2001b: 582). This means that Italian BNs may be hypothesized to be actual DPs, but headed by an empty D. Moreover, it is observed by Longobardi (2001b: 583) that in many languages (e.g. Icelandic, Welsh, Hungarian, Hebrew) also singular count nouns (i.e. Bare Singulars, BSs) can appear in argument positions, thus, summing up, cross-linguistic variation permits distinguishing at least five different types: Languages with no BNs (e.g. French), languages with stricter BNs (e.g. Italian), languages with freer BNs (e.g.

English), languages with indefinite BSs (e.g. Hebrew) and, finally, languages with ambiguous BSs, i.e. entirely lacking articles (e.g. Latin). Some parametric-oriented explanations for this cross-linguistic variation are provided in Longobardi (2001b: 584-585). What is relevant here is to understand how English BNs can be accounted for: Observe that due to their freer distribution these nominals cannot be described assuming, as in Italian, that they are introduced by an empty determiner, because, as seen above, this would constraint them to be exclusively verbal complements.

The condition of English BNs may be clarified considering that regarding another relevant group of determinerless nouns, i.e. proper names. It is argued since Longobardi (1994; 1996) that in Romance determinerless argument proper names are subject to N-to-D raising and that "several traditional semantic properties associated with object reference (e.g. transparency in intensional contexts, rigidity of designation) are indeed a necessary correlate of precisely this syntactic raising" (Longobardi 2001b: 589). This generalization is supported by Italian examples like those given in (12), where it is shown that, being N proper names in this case, movement is indicated by the order N-A.

(12) a. Maria sola ha letto questo libro.
   Mary only has read this book.
 b. *Sola Maria ha letto questo libro.
   Only Mary has read this book.

The formulation in (13) summarizes the point.

(13) If N overtly moves to a phonetically empty D then it will be object-referring (Longobardi 2001b: 589).

However, with proper names in Germanic the order N-A is not permitted, and only A-N surfaces. This variation between Romance and German is accounted for by Longobardi (2001b; 2008) claiming that in the former group of languages D is "strong", in Chomsky's (1995) sense, and triggers overt N-raising, while in the latter subfamily D is assumed to be "weak", thus the movement occurs "covertly", i.e. after the spell-out at LF. Recall that the "strength" is precisely involved also in Longobardi's (2005) parameter schemata, in particular in the d-schema (see section 3.1.4). Thus the strong/weak D distinction is assumed to constitute a relevant parametric difference among languages. Moreover, it serves to clarify the opposition between, say, Italian and English not only with respect to the distribution of proper names, but also to that of BNs. Indeed, while Italian proper names are assumed to raise to D in virtue of the strength of the functional head and BNs are hypothesized to be introduced by an empty determiner, Longobardi (2008) suggests that in English, as proper names are subject to covert movement, so BNs may be thought to be. This is related to the generalization in (15).

(15) A language has kind-referring (i.e. referential generic) bare nouns iff D is not Strong (Longobardi 2001b: 595).

A synthesis of the variation entailed by determinerless nominals in Romance and Germanic is provided in Table 7. However, observe that in Romance N-to-D raising is not the only possible strategy available to satisfy the requirement of full D: An alternative is to use an "expletive" article, in Zubizarreta, and Vergnaud's (1992) sense, that determines the order Art-A-N to surface, as exemplified in (16).

(16) a. La sola Maria ha letto questo libro.
The only Mary has read this book.
b. *La Maria sola ha letto questo libro.
The Mary only has read this book.

There are languages (e.g. Greek) in which D is strong, as demonstrated by the fact that, in accordance with (15), kind referring BNs are not allowed, but expletive articles are obligatory with all proper names, as, on the contrary, happens optionally in Italian. Longobardi (2001b: 595-596) suggests that this situation may derive from the interaction with an independent parametric condition which deals with the limit that is imposed to N-raising, which in Greek is blocked to an intermediate landing site H before reaching D due to the "barrierhood of Hmax".
Which is the trait that determines D to be strong? A recent insight on this question is provided by Longobardi (2008) starting from Crisma's (1997) development of the already mentioned Szabolsci's (1987) principle, according to which only DP can be (verbal) arguments, while NPs can not. In particular, Crisma (1997) claims that DPs can denote individuals, while NPs can denote only properties, and a certain nominal expression can function as an argument only if it denotes an individual. Longobardi (2008) hypothesizes that individuals are denoted by the trait "Person" and this trait is associated with D. Indeed this is the position in which pronouns occur both in languages like Romance and in those like Germanic, as shown in (17).

(17) a. Noi ricchi / *Ricchi noi
b. We rich / *Rich we

This conclusion, independently achieved along similar lines also by Bernstein (2008), entails to claim that the strong/weak characterization of the D position corresponds to the strength/weakness of the Person feature. Nothing more is said here about this theoretical perspective, which would require a discussion on its own, but notice that it is at the core of diverse important parametric insights. The same observation applies to the illustration of the debate regarding the postulation of other functional categories inside the DP (see among other Cinque 1994), intended as counterparts of the numerous functional projections suggested at clause level: A brief overview of some relevant ideas on this issue may be found for instance in Bernstein (2001: 554-559). What is important at this point is to provide an overview of the DP structure, on which the parameterization of the nominal domain assumed in this work is grounded: Following Longobardi's (2001b: 597) outline, this is done in (18).

(18) [D [GenS [Num [H1 [S-oriented [M1 H2 [M2 H3 [Arg H4 [GenO [P [S [N O]]]]]]]]]]]]

D: Determiner (e.g. definite article)
GenS: High position of structural genitives.
Num: Numerals.
H1: Highest landing site for N-raising.
S-oriented: Subject-oriented adjective.
M1: Manner 1 adjectives.
H2: Medium-high landing site for N-raising.
M2: Manner 2 adjectives.
H3: Medium landing site for N-raising.
Arg: Argumental adjectives.
H4: Medium-low landing site for N-raising.
GenO: Low position of structural genitives.

H5: Low landing-site for N-raising.
P: Possessor (external argument of N).
S: Subject (external argument of N).
N: Head noun.
O: Object (internal argument of N).

5.2.2. Table A

The parametric dataset focused on the analysis of the DP conceived and developed since Longobardi (2003a) is called "Table A". The parameters included in Table A have been selected largely on the basis of the formal account for nominal syntax sketched in the previous section, but some of them are also grounded on lines of inquiry that have not been explicitly mentioned there and, above all, many of them have been designed in light of novel empirical research. However, what syntactic variation is precisely encoded in the dataset? In general terms, three phenomena have been considered relevant to encode a certain morpho-syntactic difference in the form of a relevant parameter: "[T]he position of a category, the variable form of a category depending on the syntactic context, or the presence of obligatory formal expression for a semantic distinction (i.e. the obligatory valuing of a semantic distinction)" (Longobardi and Guardiano 2009: 10).

The current set of 63 binary parameters is presented in Table 7: Each parameter is associated with a number from 1 to 63 and is described with a name (e.g. *p1* = ± *gramm. person*), which may be followed by a formula that indicates whether and how it is subject to an implicational condition. This point is reconsidered after having introduced the coding adopted in the dataset, also referred to as "parametric grid". The parameters may be (approximately) grouped according to different sub-modules of the nominal domain as shown in (19):

(19) - (*p1 – p6*): Status of features Person, Number and Gender;
- (*p7 – p16*): Definiteness;
- (*p17 – p24*): Countability and related concepts;
- (*p25 – p31*): The grammar of Genitive case;
- (*p32 – p41*): Properties of adjectival and relative modifications;
- (*p42 – p50*): The relative position of N inside the DP and its raising;
- (*p51 – p55* and *p60 – p63*): The status of demonstratives and other determiners;
- (*p56 – p59*): The grammar of possessive pronouns.
(adapted from Guardiano and Longobardi 2009: 10)

The grid introduced in table 8 collects the parametric configurations of the nominal syntax of 28 languages, whose genealogical relationships are particularly well-studied within the traditional historical linguistics paradigm: There are 22 Indo-European ("IE") varieties (Italian "It", Salentin "Sal", Spanish "Sp", French "Fr", Portuguese "Ptg", Rumanian "Rum", Latin "Lat", Classic Greek "ClG", New Testament Greek "NTG", Grico "Gri", Modern Greek "Grk", Gothic "Got", Old English "OE", English "E", German "D", Norwegian "Nor", Bulgarian "Blg", Serbo-Croatian "SC", Russian "Rus", Irish "Ir", Welsh "Wel", Hindi "Hi"), 2 Semitic ("Sem") varieties (Hebrew "Heb", Standard Arabic "Ar"), a language belonging to the Niger-Congo family (Wolof "Wo"), 2 Uralic ("Ur") varieties (Hungarian "Hu", Finnish "Fin") and Basque ("Bas").

Observe that among these 28 languages 23 are modern and 5 are ancient, namely Lat (I century BC – I century AD), ClG (IV century BC), NTG (I century AD), Got (IV century AD) and OE (IX – XII century AD)[5]. Two "dialectal" varieties, instead, may be less known to some readers, i.e. Salentino and Grico: Both are spoken in Italy in Southern Apulia, but the former is an Italo-Romance variety located in the provinces of Brindisi and Lecce (represented in Table A by the dialect of Cellino San Marco), the latter is a Greek found South of Lecce (represented here by the dialect of Calimera). A remark about the modalities in which data have been collected: The parameter states regarding the modern languages included in Table A have been assigned relying on existing literature and pursuing extensive investigations among linguistically trained native speakers. In particular, these informants have been interviewed on the basis of a precise questionnaire, which includes a set of potential triggers (see section 3.1.2) in the form of specific sentences and, more generally, linguistic expressions written using English as a meta-language: The so-called "Trigger list"[6]. From the analysis of the relevant literal translations provided by the native speakers parameter settings have been assigned to the different languages. The parametric study of ancient languages, instead, has been performed by different scholars relying of closed written corpora: The references to the works in are provided in Longobardi and Guardiano (2009: 11, note 29). The coding of the parameter states adopted in Table A is given in (19).

(19) a. Basic states (+/-): They indicate opposite significant parametric settings, neither of them stands as an ancestral or marked state;
b. 0-states: These values are used in the situations in which a basic state cannot be assigned to a parameter, due to the dependence relation that exists between this parameter and the value/s of one/more preceding one/s (or an external condition);
c. ?-states: Question marks refer to cases in which the setting of a parameter is uncertain. This happens in particular with ancient varieties, whose corpora are closed and hard to analyze.

It is opportune to underline that the symbols + and – that code basic parametric states are intended to have no ontological, but just oppositional value: This is a consequence of the fact that designing Table A no theory of markedness, such as, for instance, that proposed by Roberts and Roussou (2003) and Roberts (2007) (see section 3.1.4), has been assumed. One of the most interesting features of the dataset is the existence of implications between parameters: Their theoretical foundation has already been discussed in section 3.1.5, where it has been also anticipated that the necessity of their postulation appears to be empirically supported by these data. The general (and simplest) form of implication illustrated in (17) in chapter 3 is recalled here in (20) for convenience.

(20) A parameter p_2 is implied by the value v_i of a parameter p_1 in the case in which if p_1 has v_i, then p_2 may be set either to v_i or to v_j, otherwise p_2 cannot be set.

As for the instantiation of this abstract formula in Table A, consider for example that, for the reasons worked out below, parameter 3 (*p3 = ± gramm. gender +2*) determines whether or

---

[5] For what concerns Lat, the parameter values are assigned on the basis of analyses conducted by Gianollo (2005) on Classical Latin texts (mainly Cicero's *Ad Familiares* letters and Petronius' *Satyricon*) and of traditional descriptions provided by scientific grammars of Latin. The settings regarding ClG and NTG are grounded on the research work of Guardiano (2003) on Plato's *Apology*, *Symposium* and *Cratylus*, and on Mark and John's Gospels. As for Got, of course the corpus taken into account is Wulfila's translation of the Bible, while the parameter values of OE are determined according to original searches performed by Paola Crisma on the YCOE corpus (Taylor *et alii* 2003).

[6] The list, together with the translations of the triggers in each of the languages included in Table A and references to the informants involved, is going to be published in the website of the research project in question (http://www.units.it/linglab).

not Gender is grammaticalized within the nominal syntax of a certain language, but can be "meaningfully" set if and only if the value of parameter 1, which concerns the grammaticalization of Number, is set to +, otherwise it gets 0. Thus *p2(+)* → *p3(±)* & *p2(-)* → *p3(0)*, which means *p3(±) iff p2(+)*. Conceptually there are two types of implications: Those that are motivated on empirical grounds and rely on already known typological generalizations, and those that originate from logical restrictions on the settings caused by the hierarchical ordering of parameters. The former type can be exemplified by the implication *p3(±) iff p2(+)*, that derives from Greenberg's (1966) universal number 36, according to which "if a language has the category of gender, it always has the category of number" (Greenberg 1966: 94). This generalization entails that only when in a language Number is grammaticalized (*p2*), so can be Gender (*p3*), from which follows that, if in a certain language the value of *p2* is different from +, the – setting of *p3* is entirely predictable, i.e. to assign – to *p3* would be redundant. The implication *p46(±) iff p47(+)*, instead, is logically-based, in that it is clear that one has to determine whether or not N raises over M2 adjectives (*p46*), before asking whether or not it raises over the position of M1 (*p47*), and also in this case the state of *p47* is predicable from *p46*(-). Finally, note that the implication between *p32* and *p33* is logical as well, but involves a different situation. Indeed if in a certain language adjectives are not hierarchically structured, the value of *p33*, which determines whether features spread on structured adjectives, is not properly predictable, but rather it cannot be decided at all: From the acquisitional point of view, this means that *p33* represents a choice for which there is no evidence in the primary linguistic corpus. Longobardi (2003a) observes that the distinction between redundant and pointless parameter settings "may end up being irrelevant in an abstract typological or historical theory of language variation, but it is not in a theory of the actual setting algorithm and in a theory of LAD as a discovery procedure" (Longobardi 2003a: 110). Thus, concerning the present work, the conceptual typology of implications is considered useful in the description of the data, but has no consequences on the experimentation.

From the formal point of view, instead, two types of implications can be distinguished: Simple, in the cases in which a single parametric value is the antecedent of the implication (e.g. *p3(±) iff p2(+)*), and complex, when more parametric values, related by logical operators (AND, OR and NOT), are required to satisfy the condition (e.g. *p35(±) iff (p6(+) AND (p33(-) OR p34(+)))*). In Table A the order of precedence of the operators is OR > AND (the latter is represented with a comma), so that, if parentheses are not inserted to indicate a different sequence of operations, disjunctions are solved before conjunctions.

Further comments on the structure of the data are left aside to be discussed in section 6.1.1. Now it is important to provide a more precise account for the set of languages included in Table A with particular reference to some important hypotheses that have been suggested about their genealogical classification.

5.3. The languages of Table A

A brief sketch of the families, according to which the 28 linguistic varieties comprised in the parametric dataset are traditionally grouped, has been offered in the last section. However, in order to accurately evaluate the results of the experimentation presented in chapter 6 such a summary is not sufficient and a more detailed overview of relevant proposals regarding in particular the genealogical classification of the languages involved has to be provided. This is done in the following discussion, which begins with an outline of the hypotheses conceived within the "traditional" comparative framework and then covers the suggestions that have been put forward in some of the works already mentioned in section 5.1.

5.3.1. “Traditional” hypotheses

The following discussion is meant to sketch the general genealogical framework in which the languages comprised in the parametric dataset are traditionally included and to focus more specifically on the identification of some specific issues that are relevant dealing with the sample of varieties in question. It has already been stated that the four families involved in the present inquiry are the Indo-European, the Semitic, the Uralic and Niger-Congo; Moreover, an isolated language is included in the dataset, i.e. Basque. Observe that all the non-Indo-European groupings are represented at most by two varieties in Table A, thus in the experimentation there is no concern about their internal articulation, but only about the fact that in an inferred phylogeny they are kept clearly distinguished one from the other and, above all, from Indo-European. Consequently, it is not necessary to provide any further account for their genealogical status and the following discussion concerns two main topics: The Indo-European internal classification and controversial attempts that have been made to relate families according to higher-ranking phylogenetic units.

Introductory descriptions of the internal articulation of the IE family, together with historical overviews regarding the development of the diverse groups, may be found, among other works, in Meillet (1937), Lockwood (1972), Baldi (1983), Szemerényi (1985: 30-33), Martinet (1994: 55-98), Villar (1997), Graffi and Scalise (2002: 61-64), Campanile *et alii* (2005) and Fanciullo (2007: 199-232). As a first approximation, it is useful to frame the IE languages included in Table A within Campbell’s (2004: 190-191) graph, that is accurate enough for the aims of the present discussion: An adapted version of the scholar’s representation, with arrows pointing at the varieties that are relevant here, is given in figure 1. Some comments deserve to be pointed out about this drawing. First, among the ancient varieties of the parametric dataset, only Lat and Got are overly indicated, and ClG, NTG and OE are left out. Moreover, due to their “dialectal” status, Sal and Gri are not comprised in the graph as well, but their place would unquestionably be close to, respectively, Italian and Greek. Second, all the groups that are relevant in the present work are clearly illustrated by Campbell (2004). More spcificaly, the descriptions regarding the internal structure of Celtic, Slavic and Germanic may be considered sufficient, in light of the fact that in the parametric dataset only 2 Celtic, 3 modern German and 3 Slavic languages are included, so that the distribution outlined in (21) may be considered the optimum to achieve.

(21) Celtic → (Ir, Wel)
(Modern) Germanic → West Germanic (E, D), North Germanic (Nor)
Slavic → South Slavic (SC, Blg), East Slavic (Rus)

The articulation of Romance, instead, deserves more attention, because in the dataset this group is represented by 5 varieties, whose classification might be particularly delicate. Campbell (2004) distinguishes between a West Romance branch, including Sp, Ptg and Fr, and an East Romance lineage, to which It and Rum are assumed to belong. This taxonomy appears to be perfectly in line with Diez’s (1836), who, however, proposed distinguishing between a proper West Romance section (with Sp and Ptg) and a North West Romance subgroup (with Fr and Provencal). In any case, following Tagliavini (1999: 349-356), who provides a detailed overview of the tradition of study in this domain, a well-defined classification seems to be particularly strict for Romance, which may be better regarded as a *continuum* from the western to the eastern part of the *Romania*, i.e. from Balcano-Romance to Italo-Romance and from Gallo-Romance to Ibero-Romance.

As for the genealogical relationships between different groups, observe that the relative order in which they appear in Campbell’s (2004) tree might not be strictly significant, in the sense

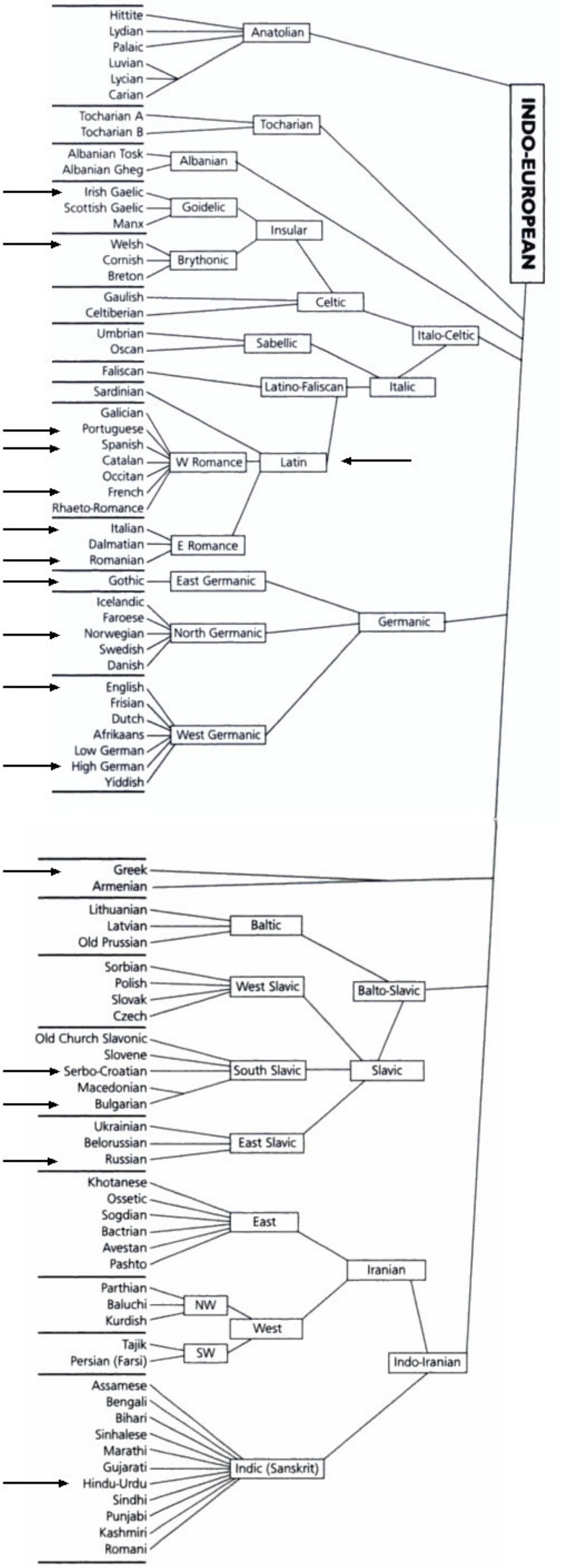


Figure 1: Traditional classification of IE languages (adapted from Campbell 2004: 190-191).

that in the structure the common origin of each of them is encoded, but not necessarily the fact that some groups are more closely related to each other than others. Actually the main problem dealing with Indo-European and, in general, language families, is not to identify groups, but to determine how these groups are related and how these groups are internally articulated. To properly elucidate this issue a concise digression in the history of the attempts of Indo-European classification should be outlined based on Villar's (1997: 597-616) overview. As already stated, Schleicher's (1853) *Stammbaum* is considered the first explicit representation of the relationships among the different groups in the form of a genealogical tree. The proposal of the scholar is presented in figure 2a. In this phylogeny at least two important macro-groups require particular attention, since they are largely considered significant in current literature as well: The Indo-Iranic and the Balto-Slavic. For the rest nothing substantial needs to be pointed out about the scholar's suggestion: Indeed, the other hypotheses entailed by his representation already underwent relevant changes in the following decade (Schleicher 1963), especially due to the influence of Lottner's (1858) idea about the supposed close relatedness of Celtic and Latin. As already claimed, in the second half of the XIX century, above all after the diffusion of Schmidt's (1872) *Wellentheorie*, it became clear that providing an exhaustive account of all the possible relations between the Indo-European varieties exclusively by means of tree-graphs was unrealistic, so that proper phylogenetic structures ceased to be produced. Therefore illustrations like Schmidt's (1872), reported in figure 2b, started becoming popular: These pictures are meant to show the diatopic variation of languages precisely in the original area in which they separated and from which they spread following different directions, but they lack appropriate information regarding relative chronology of the divergence events (Villar 1997: 603). In the early XX century many "dialectal diagrams" intended to reconcile the tree- and the wave-oriented perspectives were designed: Among others, Antoine Meillet's is presented in figure 3. Villar (1997: 604) argues that the problem with these representations is that they appear to be more a synthesis of a complex net of multilateral isoglosses than actual historical accounts of the development of the languages in question: Indeed, they do not provide indications about geographical and chronological locations of the different varieties.

Thus even comprehensive diagrams came to a stop, but in the meantime much work was done to shed light on specific problems of Indo-European classification. While groupings like Indo-Iranic and Balto-Slavic continued to be regarded as well-established, others continued to be subject to discussion: For instance, in the '30s Lane (1933) was convinced that "[t]he exact relation between the Germanic and Celtic dialects is still unsettled, though doubtless no one will ever dare maintain anything like a Germane-Celtic unity" (Lane 1933: 244), while Dillon (1944) provided an overview of the comparative evidence regarding the Italo-Celtic unit, which also summarizes diverse points debated until the first half of the XX century. Moreover, at least two main issues started to be taken into account in the same period: The first was that of the secondary convergence, the second had to do with the awareness that the separation of the Indo-European languages did not occur abruptly, but lineages progressively diverged in diverse times. Among the clearest insights that include the latter idea there are, for instance, Meid's (1975) and Adrado's (1979). Both the scholars hypothesize three main periods in the evolution of IE varieties: More precisely, Meid (1975) argues that all languages that diverged in the "primitive period" (before 3500 BC) became extinct, while only Anatolian survived from the "middle period" (between 3500 and 2500 BC) and all the other IE languages gradually separated in the last stage in accordance with a precise chronology, so that he suggested the successive splits involving the following groups: Armenian, Indo-Iranic and Greek before 2000 BC, Tocharian and "South-Western European" before 1800-1500 BC and then, between 1500 and 500 BC, Germanic and Balto-Slavic.

The debate regarding the majority of the relationships between IE groups (e.g. Italo-Celtic, Germano-Celtic, Gemano-Balto-Slavic and so on) is still open in the traditional paradigm.

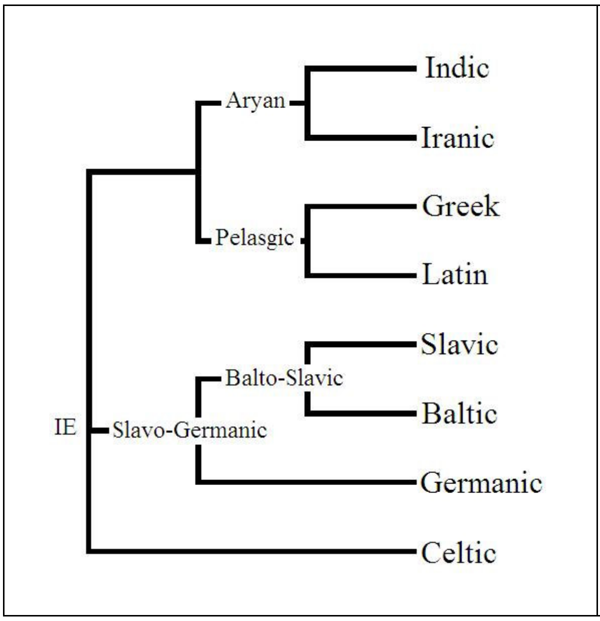


Figure 2a: Schleicher’s (1853) original classification of IE groups (adapted from Villar 1997: 599).

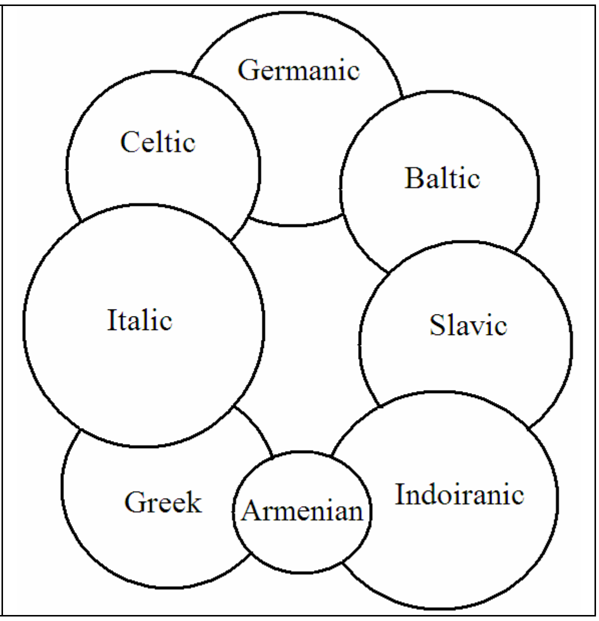


Figure 2b: Schmidt’s (1872) classification (adapted from Villar 1997: 603).

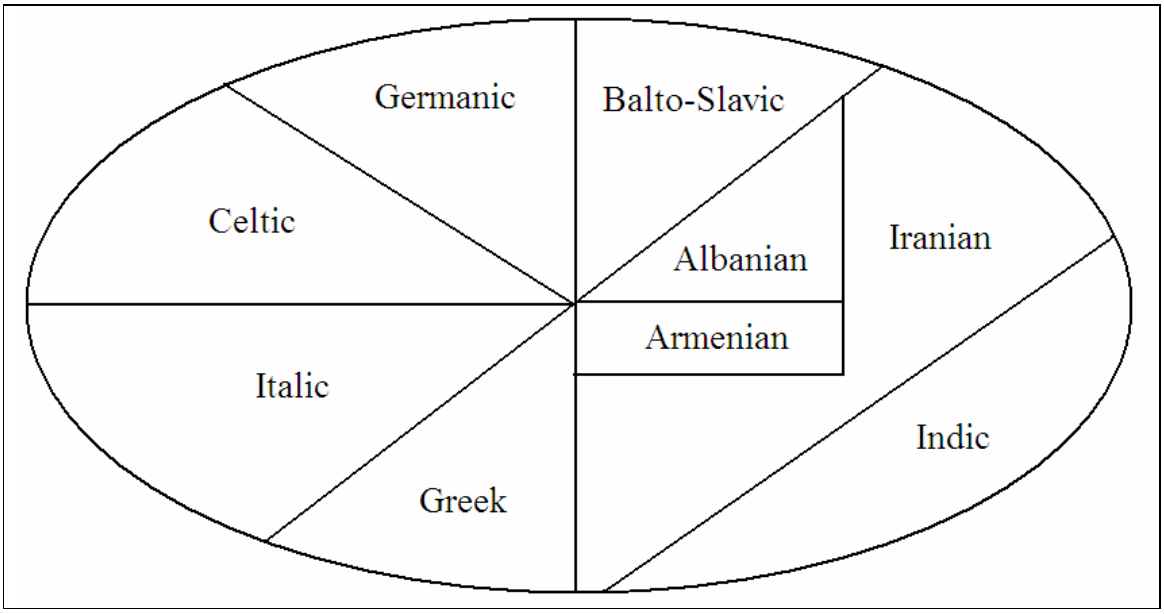


Figure 3: Meillet’s classification of IE languages (adapted from Villar 1997: 605).

All attempts made until the XX century to reconstruct long-range relationships between families, i.e. macro-families, have been disputed and have raised scepticism among the large majority of the scholars. As already claimed in the last chapter and, in particular, in section 4.2.2 presenting mass-comparison, to investigate deep time depths by means of the classical comparative method appears to be almost unfeasible, due to lack of relevant evidence. Furthermore the reliability of alternative methods proposed to pursue this objective is at least dubious and, for a large part of the scientific community, almost null. What is more important in this discussion is to recapitulate the status of the issue, with particular reference to the hypotheses that have been put forward regarding the genealogical relatedness of Indo-European with other families.

Historical surveys on this subject may be found, for instance, in Villar (1997: 643-655) and Greenberg (2000: 1-23). One may observe that disparate groups or languages have been proposed as close relatives of Indo-European, including among others Korean (e.g. Eckardt 1966; Koppelmann 1928; 1933), Sumerian (Autran 1925) and Ainu (Koppelmann 1928). However, in all such cases no proofs of relatedness other than relatively limited sets of lexical items, whose low probative value has been already underlined, have been used. Other hypotheses are grounded on more interesting evidence and deserve to be taken cautiously into account. For example, since Anderson (1879) a number of studies have been dedicated to the identification of a genealogical relationship between Indo-European and Uralic also relying on correspondences between closed morphological classes, especially that of pronouns. In any case, on the one hand the evidence that supports an Indo-European-Uralic macro-grouping seems to be insufficient (e.g. recall the fact that certain kinds of paradigms cannot be used as valid individual-identifying evidence, note 3 in chapter 4), on the other hand the hypothesis would disagree with notable differences observed in other key domains (i.e. numerals and parts of core vocabulary).

Many attempts have been made also to join Indo-European and Semitic. This line of research is rooted in Møller (e.g. 1907) and Cuny's (e.g. 1924) well-known proposal of relating the list of phonemes postulated since De Saussure (1879) as "sonant coefficients" and, later, as *ə elements, to laryngeal articulations, from which the theory that has come to be known as "laryngeal theory". The scholars' aim was precisely of identifying a relevant common trait between Proto-Indo-European and Proto-Semitic, in light of the observation that in Semitic languages laryngeal sounds are outstanding. Even if the deciphering of Hittite in the first decades of the XX century corroborated the idea that the phonetic content of the *ə elements could be regarded as laryngeal, doubts remained regarding the possibility of demonstrating that the two language families in question derived from the same common ancestor (e.g. Fanciullo 2007: 13-134). The evidence on which the hypothesis was grounded continued to be primarily "pregrammatical", i.e. strongly lexical-based. This did not prevent the topic being overtly reconsidered by Pedersen, who argued in favour of the possibility of defining a vast genealogical unit comprising not only Indo-European and Semitic, but also Uralic and other relevant families. To refer to this macro-family the scholar adopted the name "Nostratic" (Pedersen 1903; 1933).

The question related to the identification of a large cross-familiar *taxa* including Indo-European was revived in the '60s mainly as a consequence of the proposals made by two Soviet linguists, Illič-Svytič and Dolgopolsky (1986). The former hypothesized a common historical origin of Indo-European, Afro-Asiatic, Uralic, Altaic, Kartvelian and Dravidic, which should date back to 12,000 – 15,000 years ago; The latter arrived independently along similar lines, trying to isolate systematic sound correspondences and, as lexicostatisticians, to focus on a subset of the lexicon that he thought was particularly resistant against borrowing. Among others, Kaiser and Shevoroshkin (1988) provide a summary of the evidence collected in favour of these extended Nostratic hypotheses. In the following decades the inquiry continued to develop: Among recent contributions to the debate one may recall for example Bomhard and Kerns (1994) and Salmons and Joseph (1998). A discussion on its own, instead, should be dedicated to the attempts made within the mass-comparison paradigm, e.g. Ruhlen (1994) and Greenberg (2000), and those closely connected to archaeological investigations, e.g. some of the papers belonging to the collection edited by Renfrew and Nettle (1999). In any case, to sum up, the general current thought of the mainstream scientific community is probably reflected by Campbell's (1998) assessment, according to which it is worth remaining open-minded about the Nostratic theme, but for the moment no convincing evidence has been provided in favour of a Nostratic genealogical unit. Observe that what is questioned is not the existence of proto-stages, with respect to which the languages comprised in the observed families would represent a set of related descendents, in that it is natural to think that there must have been language evolution also beyond the time depth covered by classical comparative method (e.g. Villar 1997: 643), but rather the possibility of scientifically studying this proto-stages and of providing a consistent historical account for them.

5.3.2. Quantitative hypotheses

For the reasons explained in section 5.1, only the results achieved by two lines of quantitative study are considered in the following discussion: That of Ringe *et alii* (2002) and Nahleh *et alii*'s (2005a) research group, and that pursued by Gray and Atkinson (2003) and Atkinson and Gray (2006b). Both the experimentations are concerned with the phylogenetic classification of sets of Indo-European languages and, moreover, fundamentally rely on classical comparative accounts, the former explicitly, the latter indirectly due to the fact that it is grounded on lexicostatistic data, in which cognate estimates are derived following the traditional comparative procedure. As a consequence in any case one would not expect that these inquiries overcome the time-depth of the family-level, because they are subject to the same chronological limitation that affects the classical comparative method.

The tree reconstructed by Ringe *et alii* (2002) on the basis of their maximum-compatibility procedure is illustrated in figure 4. Apart from the fact that all the relevant groups are clearly outlined, observe that some specific macro-groups originate: Italo-Celtic, Albano-Germanic and Greco-Armenian together with Balto-Slavic and Indo-Iranic. According to the authors this phylogeny is not supported only by 18 lexical characters out of the 322 on which the experiment has been carried out (data already known to be homoplastic is left out of the experiment by the scholars) and only one lineages appear to be particularly problematic: That of Germanic, which shares many characters states with other non-directly related groups (e.g. one possible coding of the meaning "young" with Celtic, against the common coding of the same gloss in Italic and Greek). On the whole, the evolutionary tree is considered stable by the authors (Ringe *et alii* 2002: 98-106). In Nakhleh *et alii* (2005a) the compatibility of the character set is tested on five candidate genealogical structures: One is precisely Ringe *et alii*'s (2002), two are derived modifying it in "linguistically plausible ways" and two correspond to independently suggested classifications. Putting it simply, a score is assigned to each tree according to the number of characters with which it is incompatible, i.e. the number of characters that are homoplastic: Of course the higher this score, the less the tree is supported. Moreover, each homoplastic character-state history is evaluated in order to understand whether the homoplasy may be considered as an effect of parallel development (i.e. chance) or contact between different lineages (i.e. borrowing or horizontal transmission). Nakhleh *et alii*'s (2005a) goal is to understand what may be a "perfect phylogenetic network", i.e. a graph that wholly explains the variation of character-states observed in the sample of languages in question by means of both a series of lineages in which proper vertical development occurred and (a small set of) plausible contact edges that connect some lineages and represent paths of secondary convergence. In light of this intent, the best scenario that the scholars find to account for historical diversification of the Indo-European languages is that based on Ringe *et alii*'s (2002) phylogeny. This perfect phylogenetic network is presented in figure 5: Nakhleh *et alii* (2005a) point out that it entails "at most three historically real episodes of contact between the relevant language groups" (Nakhleh *et alii* 2005a: 407), two involving Germanic and another connecting Proto-Italic and Proto-Greco-Armenian.

Atkinson and Gray (2006b) are less interested in the classification of the Indo-European groups, even if this aspect of the outcome of their experiments is relevant as well, than they are in the question regarding the chronology of the splits occurring in the family. As already claimed, the scholars' objective is to provide support in favour of one of the two main competing theories on the origin of Indo-Europeans: The "Kurgan hypothesis" (Gimbutas 1970) and the "Anatolian hypothesis" (Renfrew 1987). In essence, according to Gimbutas (1970) the birthplace of the languages (and the people) in question is located in the steppe of

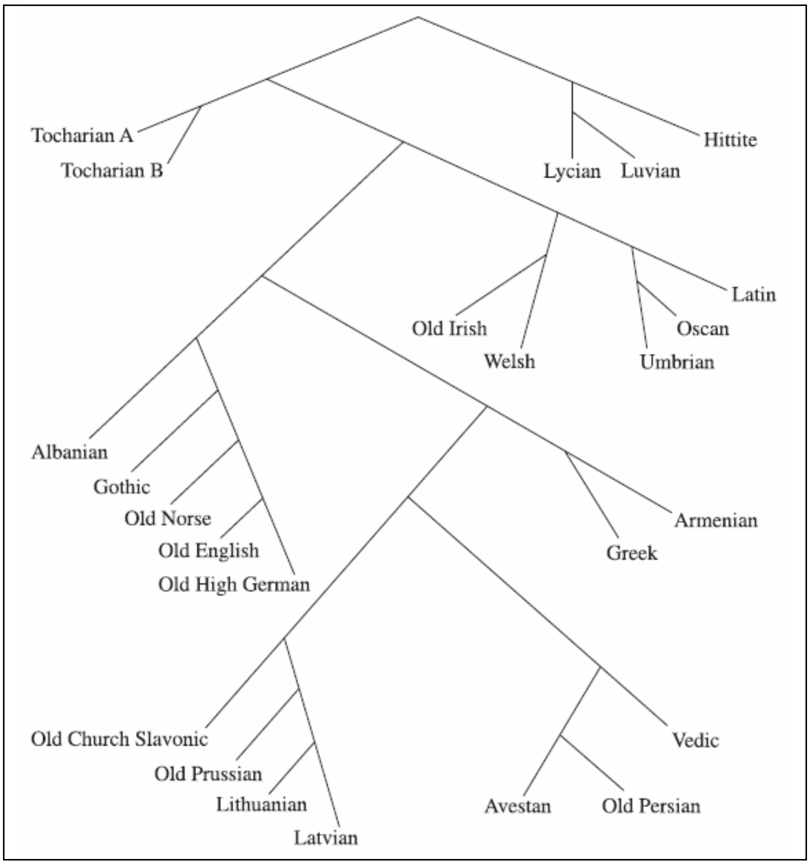


Figure 4: Ringe *et alii*'s (2002) phylogeny (from Ringe *et alii* 2002: 87).

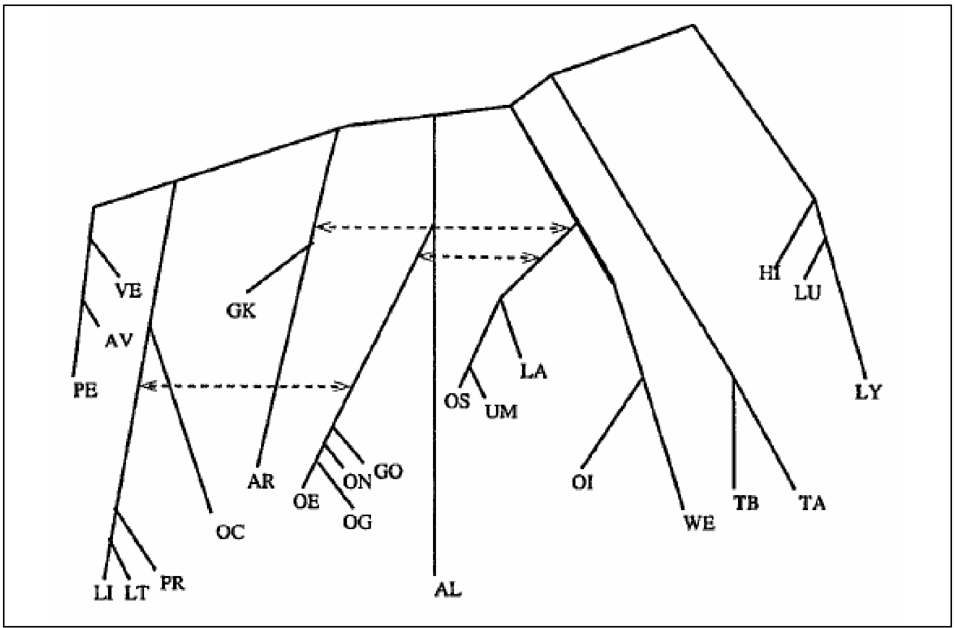


Figure 5 : Nahleh *at alii*'s (2005a) perfect phylogenetic network[7] (from Nahleh *et alii* 2005a: 403).

---

[7] This graph includes Old Persian (PE), Avestan (AV), Vedic (VE), Lithuanian (LI), Latvian (LT), Old Prussian (PR), Old Church Slavonic (OC), Ancient Greek (GK), Classical Armenian (AR), Old English (OE), Old High German (OG), Old Norse (ON), Gothic (GO), Albanian (AL), Oscan (OS), Umbrian (UM), Latin (LA), Old Irish (OI), Welsh (WE), Tocharian B (TB), Tocharian A (TA), Hittite (HI), Luvian (LU) and Lycian (LY) (Nakhleh *et alii* 2005a: 392).

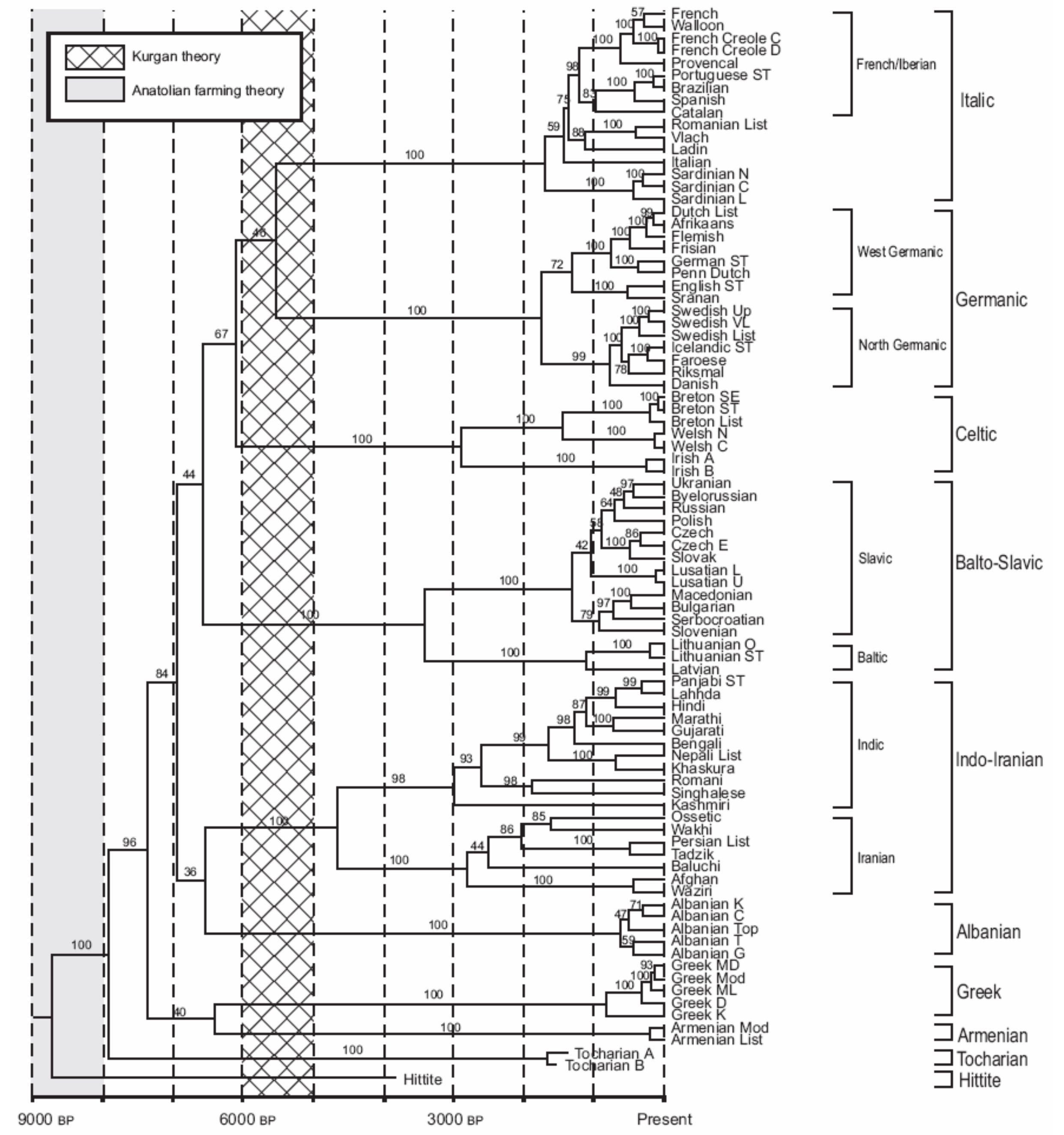


Figure 6: Gray and Atkinson's (2006) phylogeny.

South Russia and the expansion from that area should have begun around the V/VI millennium BC; Renfrew (1987), instead, is convinced that the Asia Minor should be regarded as the Indo-European ancestral homeland, and that from there the different groups spread following the expansion of agriculture at least since the VII/VIII millennium BC. It is worth pointing out that the mainstream historical linguistics literature is largely concordant with the former proposal, strongly supported by archeo-linguistic evidence, since the latter almost exclusively relies on archaeological data and does not include a satisfactory linguistic account (see for instance Villar 1997: 48-75). However, the dating problem is fundamental in Atkinson and Gray's (2006b) account precisely because the two hypotheses entail diverse chronologies. In figure 6 the scholars' reconstruction is shown: it is notable that in this case no Italo-Celtic unity is inferred, while Blato-Slavic and Indo-Iranian are confirmed and, among the languages that are relevant here, Greek lies in one of the most external branches of the graph, meaning that its lineage diverged rather early (around 5500 BC according to Atkinson and Gray 2006b). The dates projected onto the phylogeny appear to support the Anatolian hypothesis: Actually this point is not of much interest here, since, as explained in section 6.3.2, glottochronological estimates cannot be performed on the basis of the parametric data. Anyway, the debate connected to this line of research is sketched in Atkinson and Gray (2006b: 102-104).

5.4. Conclusion

In this chapter the parametric dataset and, in general, the present investigation has been framed within the quantitative phylogenetic research carried out in linguistics. The survey has been initially focused on the presentation of other relevant datasets, comprising lexical, morphological, phonological and non-parametric syntactic *comparanda*, which have been critically illustrated. Then Table A has been introduced, starting from the outline of the theoretical background on which it is based: In particular, it has been shown how the parameters listed in the dataset are drawn from specific accounts provided in the last decades on the form of DPs. Then, after having provided a detailed description of the coding and the underlying structure of the data, a specific section has been dedicated to the introduction of "traditional" and quantitative-oriented hypotheses on the genealogical classification of the languages included in Table A, with the aim of preparing the ground for the evaluation of the results illustrated in the next chapter.

6. Quantitative analyses

This chapter covers the illustration of the quantitative experimentation carried out on Table A. The investigation has been designed in order to pursue two main objectives. The first intent is to develop the program of empirical testing of the parametric dataset initiated in the line of research of Longobardi and Guardiano (2009) (see further references below and in section 4.2.4) and to provide evidence in favour of the idea that syntactic and, more generally, language evolution may be successfully studied adopting a method of parametric comparison. In other words, the initial aim is to show that, if parametrically encoded, comparative syntax may be effectively used to reconstruct language phylogenies. To this end it is necessary to provide an accurate quantitative account for the data, especially in light of the structure that lies behind them, from the point of view of both a character-based and a distance-based perspective. On this ground, it is then possible to justify the choice of using a specific set of phylogenetic methods, whose adoption depends on the characterization of the data in question, and finally to properly apply these techniques to Table A and to obtain a first outfit of results, i.e. of genealogical trees.

The second objective of the experimentation is to test specific procedures of analysis that, inferring hypotheses of historical relatedness between linguistic varieties, may help to account for the occurrence of homoplastic events (in particular parallel evolution and horizontal transmission, i.e. convergence due to chance and contact) in the diachronic development of languages. Above all the development of this direction of inquiry is related to the achievement of two (overlapping) goals: On the one hand the isolation of a limited selection of parameters thought to comprise a higher amount of genealogical information on the basis of empirical considerations; On the other hand the identification and the evaluation of diverse quantitative criteria used to weight parameters according to their “genealogical relevance”. Both these tasks represent attempts to reduce the bias caused by homoplastic character-state histories in the reconstruction of phylogenies. Nonetheless, neither of them explicitly deal with the analysis of specific non-homologous similarities between languages: Thus a further advancement of the investigation consists of identifying a method to detect single cases of non-genealogical convergence of linguistic features and then of designing an evaluation procedure that can assist in distinguishing between the diverse types of homoplastic similarities, i.e. separating those that are chance effected from those that derive from language contact.

Before proceeding, a fundamental premise of the work, already mentioned in the introduction, needs to be recalled and clarified. The quantitative experimentation presented here is primarily intended to have an empirical, rather than an inferential, statistical value, i.e. the outcome of the present work have to be regarded largely as descriptively, and only tentatively explanatory, adequate. This means that the hypotheses introduced regarding specific paths of syntactic evolution should be regarded as entirely substantial and reliable with respect to the sample of data under study; Nonetheless there is no guarantee that the procedures of analysis adopted here to reconstruct the paths of historical development of the languages in question may be universally applicable, because they are not validated on the basis of a proper statistical corroboration of the results. Roughly, this would require building them relying on a general model describing the evolutionary process in question, which could then be put to the test determining how much the choice of the set of languages and of the set of parameters influence the inference of phylogenetic hypotheses. This is only marginally completed and, as a matter of fact, it is beyond the aims of the work: Actually some tentative explanatory suggestions are briefly taken into account and the reliability of the phylogenies that are proposed is supported by means of a specific procedure of perturbation of the data on different subsets of languages. In any case here no explicit general evolutionary model is assumed. As a consequence, the inquiry is not specifically aimed at grounding any inference regarding the mechanisms that led to formation of the genealogical relationships that are described, even if this does not prevent the results presented here from providing indications for the development of generalizations about the model that underlies the relevant process of diachronic development.

The chapter is organized as follows. Section 6.1.1 includes the preliminary quantitative analysis of the dataset and comprises both a character-based evaluation of the parameters included in Table A and the account of the measure of distance between languages that is used throughout the experimentation. Moreover, the specific set of syntactic distances used in the experiments is introduced and compared with a corresponding set of lexical-based distances. In section 6.1.2 the choice of the phylogenetic methods adopted to carry out the empirical investigation is motivated, arguing that, in light of the characterization of the parametric data provided above, the procedures most aptly used here are the distance-based ones, but that results obtained by means of parsimony analysis are worth being taken into account (albeit cautiously) as well. Thus in section 6.1.3 the first outfit of phylogenies reconstructions on the full dataset is presented and commented on in detail. The discussion included in section 6.2, instead, is specifically dedicated to accounting for homoplasy in the dataset: In particular, an empirically-based selection of parameters is proposed in section 6.2.1, while in section 6.2.2 some suggestions regarding the manner in which parameters might be weighed according to their genealogical relevance are put forward. And finally, the last part of the chapter covers certain advanced and tentative analyses performed on the parametric dataset, concerning the possibility of detecting borrowing in phylogenetic trees, the reconstruction of ancestral states and the correlation of the syntactic distances with separation times and geographic distances.

6.1. Inferring the trees

The first step of the present inquiry consists of testing whether or not the dataset contains relevant genealogical information, i.e. whether or not there is a relatively robust vertical transmission of the syntactic traits. Ultimately, this point is put to the test using the data as input for the quantitative-based reconstruction of phylogenetic structures. Similar attempts have already been presented in Guardiano and Longobardi (2005), Longobardi and Guardiano (2009) and Gianollo *et alii* (to appear): In these works it has been demonstrated that notable genealogical insights can be provided on the basis of the parametric comparison method and, as anticipated in section 4.2.4, a plan of preliminary conceptual and empirical tests of the dataset has been outlined. The initial objective of the following discussion is to follow the scholars' program, but, before introducing it, a brief character-based description of the parameter set is provided.

6.1.1. Quantitative analysis of the dataset

Recall from section 5.1.2 that the fact that parameters are related one to the other by means of implicational conditions represents one of the most important features of the dataset: In particular, from the point of view of the quantitative analysis, the fact that the characters are not isolated, but rather framed within precise structures, is above all important. The consequences of this situation are taken into account later, for the moment a useful way to visualize it may be the graphical representation of the parameter space presented in sketch 1[8]. The labelled boxes correspond to parameters or external conditions, while basic values (+, -) are included into small circles and logical operators (AND, OR) into larger ones. If an implication exists between two parameters (say, *p1* and *p2*), an edge connects the relevant value of the antecedent (*p1(+)*) with the box that represents the consequent (*p2*). At least two interesting remarks arise from this representation. First, except for five parameters that are isolated (namely *p20*, *p31*, *p37*, *p41*, *p59*), all the others are grouped into a unique implicational structure, which may be formally regarded as a network: Indeed, in this graph the chain condition (section 4.1.2) is violated due to the convergence of multiple edges to the nodes that correspond to logical operators. Second, the parameters can be assigned to different "ranking levels", in Baker's (2001) terms, according to the depth at which they are placed in the structure: Given that independent parameters (e.g. *p1*) may be considered level-1, the level of an implied parameter may be intended to correspond to the level of its deepest antecedent increased by one. Nothing more needs to be said about these notes here; However, the notion of ranking level will be reconsidered for practical reasons in section 6.3.2 dealing with the reconstruction of the ancestral parameter states.

Table 1 includes the list of the 63 parameters of Table A, as they are indicated in the first column: The second column provides the collection of the implicational conditions and the third the ranking-level values. The rest of the table is dedicated to the character-based analysis of the dataset and is divided into two parts: One covers the whole group of 28 languages comprised in the data, the other is focused on the subgroup containing exclusively the 23 modern languages (i.e. Lat, ClG, NTG, Got and OE are excluded). The relevance of this distinction is clarified in the next section. For the moment, notice that in each part the number of +, -, 0 and ?-states recorded in the relevant sample of languages is specified, together with the parsimony-status (PS) of each parameter. As explained in section 4.3.1, according to PS a character may be constant (C), parsimony uninformative (U) or parsimony informative (I). The summary of the latter analysis is provided in (1ab).

(1) a. Parsimony-based evaluation of the parameter set, all languages:

- Constant parameters: 5 (*p1*, *p2*, *p24*, *p35*, *p46*);
- Parsimony uninformative parameters: 12 (*p6*, *p9*, *p10*, *p11*, *p13*, *p16*, *p30*, *p32*, *p37*, *p48*, *p54*, *p58*);
- Parsimony informative parameters: 46 (the others);

b. Parsimony-based evaluation of the parameter set, modern languages:

- Constant parameters: 7 (*p1*, *p2*, *p8*, *p24*, *p26*, *p35*, *p46*)
- Parsimony uninformative parameters: 12 (*p6*, *p9*, *p10*, *p11*, *p13*, *p16*, *p30*, *p32*, *p37*, *p48*, *p54*, *p58*);
- Parsimony informative parameters: 44 (the others);

The PS evaluation has been drawn on the basis of the set of the splits of the language set determined by the parameters of Table A, which is presented in table 2. The splits that are considered relevant are only those defined on the basis of an alternation between the two basic states (+ and -). For example, the split derived from *p3* (± *gramm. gender*) is rather straightforward, in that in this case all the languages are set to a meaningful state (i.e. 0s are not recorded for *p3*) and there are only three languages (Hu, Fin and Bas) out of 28 in which gender is not grammaticalized, i.e. whose state for *p3* is -. The split entailed by *p6 (± numb. on N*) instead, is intended to include only 26 languages, because in Wo and Bas this parameter

---

[8] The implicational structure has been drawn with the program *Graphviz*.

is set to 0 due to the fact that for both the linguistic varieties the implication *p6(±) iff p5(+)* is not satisfied.

It is opportune to underline that a parameter whose PS is constant or uninformative, i.e. it is irrelevant from the point of view of the character-based phylogenetic analysis, may still be informative when considered within the distance-oriented approach. Is it possible to calculate a set of distances from Table A? This point is particularly important not only from the perspective of adopting distance-based methods of phylogenetic reconstruction, but also to carry out the program of preliminary testing of the data, as it is proposed by Longobardi and Guardiano (2009: 13). Their points are reconsidered below, after having introduced the procedure by means of which syntactic distances may be derived. First of all, consider that each language is identified by a specific string of parametric values: These include both basic and 0/?-states. Since Longobardi (2003a) it is assumed that the first step to outline a measure of syntactic dissimilarity between languages may be to take into account a precise pair of coefficients, <*i*,*d*>, where *i* is the number of identities between two strings and *d* the number of differences: If two languages show the same basic value (either + or -) in a parametric setting, *i* is increased by one, otherwise it is *d*. At the same time, the 0-states and the ?-states are assumed to be completely uninformative, i.e. blanks in the strings: In other words, whenever at least one of the two compared languages shows either a 0-state or a ?-state, neither *i* nor *d* is increased.

How can the coefficients be used in order to obtain an appropriate and accurate distance *D*? It is important to note that different pairs of languages may be compared on the basis of a different number of characters, which corresponds to ($d + i$): For instance, considering the comparison between It and Sp on the one hand and SC and Hi on the other, the outcome in terms of coefficients in each case is respectively <48,5> and <29,5>, thus the pair (It, Sp) is confronted according to 53 parameters, while (SC, Hi) according to 34. It is clear that *d* alone is not a sufficiently informative measure of distance: In the case of (Sal, Sp) and (Sal, SC) the number of differences is the same (5), but there is no doubt that It is genealogically closer to Sp than SC to Hi, thus the fact that in the first pair there is a higher number of identities is fundamental and it must be reflected by an accurate *D*. Incidentally, observe that *d* precisely corresponds to the Hamming distance (Hamming 1950) used, among the works examined in this dissertation, by Spruit (2005; 2008) to compare Dutch dialects (section 5.1.2): In the scholar's case the difficulty encountered here does not arise, because in his dataset there is no meaningless character state, as 0s and ?s are. However, a solution that has been adopted in the case of parametric comparison consists in assuming that certainly *D* should be directly proportional to *d*, but it should also be related to the number of effectively compared characters, i.e. normalized. Thus the simplest manner in which to account for this fact can be provided by a *D* that equals the number of differences *d* out of the number of compared characters. The formula is presented in (2).

(2) $$D = d / (d + i)$$

According to (2), the distances corresponding to the pairs (It, Sp) and (SC, Hi) are $D_{It,\ Sp}$ = 0,0943396 and $D_{SC,Hi}$ = 0,147059: Thus It and Sp result to be less distant that SC and Hi, and, besides the fact that this satisfies the expectation derived from the knowledge of the history of these languages, it is also concordant with the observation that the first two languages share a higher number of identical parametric settings than the second ones. The entire set of syntactic distances that separate each pair of languages of Table A is collected in the matrix in table 3. It is worth pointing out that the formula given in (2) defines a measure of distance that may be considered the complement the so-called "Jaccard similarity index" (Jaccard 1901), introduced by the Swiss botanist Paul Jaccard to quantify the resemblance between different sample sets. In linguistics, for instance, the Jaccard coefficient is taken into account within phonology as a possible way to calculate distances between phonemes (e.g. Pucher *et alii* 2007)[9]. Moreover, observe that the reason why PS-constant and PS-uninformative parameters may carry significant information within a distance-based analysis should be clear now: The former may be relevant in the case in which they present at least one ?/0-state among their settings, which would lead to increase *i* by 1 for all the languages provided with the constant basic state, but not for the languages whose setting is uncertain/irrelevant (e.g. *p24*); PS-uninformative parameters, instead, are important to increase the distance between isolated languages from all the others (e.g. *p32*, that separates Wo from all the other languages of the sample).

Now that the measure assumed here to quantify syntactic distance has been introduced, it is possible to take into account Longobardi and Guardiano's (2009) testing program. In particular, in order to understand whether Table A encodes a significant amount of genealogical information, the scholars propose to verify whether the set of distances drawn from Table A can satisfy some specific conceptual expectations: Two of them are summarized in (3ab).

(3) a. The probability of chance resemblance for the most similar languages must attain an individual-identifying level in Nichol's (1996) sense;
b. The syntactic distances derived from the parametric comparison should be scattered across different degrees of dissimilarity.
(adapted from Longobardi and Guardiano 2009: 13)

The point in (3a) has already been considered in section 4.2.4: In the same section in (8) the formula to calculate the probability that two languages share a certain number of identical parameter states by chance has been provided, together with a relevant example. Taking into account the coefficients *i* and *d*, the same calculation carried out for all the pairs of languages of the sample shows that in 82 cases out of 378 (about 21.7%) the highest threshold value given by Nichols' (1996: 49) for individual-identifying evidence (1/100,000) is largely overcome: These correspond to pairs whose similarity strongly claims an historical explanation. Other 50 cases out the remaining 296 (about 13.22% of the total), instead, satisfy Nichols' (1996: 49) lower threshold requirement (1/10,000) and may be thus regarded as particularly interesting as proofs of genealogical relationship between the languages in question.

Note that the pairs of languages, whose closeness is demonstrated to be far from being chance effected, are precisely those for which the distance is lower. In particular, Longobardi and Guardiano (2009: 21-22) collect the entire list of language pairs in increasing order of distance in the so-called "Table D": For each pair, they indicate whether, according to independent knowledge derived from traditional historical linguistic accounts, the relation between the two languages should be expected to be strong (common ancestry, at most 4,000 years of separation time), looser (safe derivation from a common ancestor, e.g. Proto-Indo-European, more than 4,000 years of separation time) or weak (no independently assessed genealogical relation). The scholars' observation is that "most strong pair cluster in the topmost part of the table (Table D, editor's note), while weak relations tend to occur from the middle to the bottom" (Longobardi and Guardiano 2009: 14). Further comments on this matter are not strictly required here, since they are already offered in the above mentioned work. However, it is worth providing a sample of Table D (in particular, the highest and the lowest sections) in table 4, which includes the list of language pairs in the first column: For each pair the outcome of the comparison is given, i.e. the number of identities (*i*), differences

---

[9] Other possible distance measures adopted in phonology, in particular the so-called Levenshtein or edit distance (Nerbonne *et alii* 1999, Heeringa 2004) are discussed, among others, by Nerbonne and Kleiweg (2007) with particular reference to dialectal phonological variation.

(*d*), single 0s (*0*) and double 0s (*00*) (i.e. the number of cases in which comparing the two strings either one or two 0-states are found in a certain locus), the probability that the number of identities/differences recorded is due to chance and, finally, the syntactic distance (i.e. assumed to separate the two languages) are specified.

As required by the test given in (3a) the distances included in Table D actually appear to be scattered, in the sense that they do not converge on the same average value, say 0.5, but rather "they range from <40,1> (distance 0.024) to <13,8> (distance 0.64), assuming many intermediate values" (Longobardi and Guardiano 2009: 13). The distribution of the distances can be also appreciated in the scattered plot in figure 1: The values of distance are projected onto the vertical axis, while the pairs of languages are mapped onto the horizontal axis following the order of Table D. At this point it is important to observe that the uppermost value achieved by the distances is about 0.5: Only the distances of 7 pairs of languages (1.85% of the total), namely (Fr, Bas), (NTG, Bas), (ClG, Wo), (Ar, Bas), (Ar, Wo) (Ir, Bas) and (Wel, Bas), slightly overcome this limit. This is in agreement with a basic assumption that is stated since Guardiano and Longobardi (2005), i.e. the "Anti-Babelic Principle", which is illustrated in (4).

(4) "Anti-Babelic Principle: similarities among languages can be due either to historical causes (common origin or, at least, secondary convergence) or to chance; differences can only be due to chance (no one ever made languages diverge on purpose" (Longobardi and Guardiano 2009: 13).

According to this principle, it is expected that the distance that separates two genealogically unrelated languages should be precisely around 0.5, i.e. that *i* should be almost equal to *d*: This assumption appears to be largely compatible with the parametric data.

A final test that can be made before proceeding to the proper phylogenetic analysis of the data is to compare the set of syntactic distances with a corresponding list of lexical distances. This can be done taking into account Dyen *et alii*'s (1992) estimates based on the complement of the percentage of cognates shared by each pair of languages, i.e. the same measures of dissimilarity used, for example, by McMahon and McMahon (2003) (see section 5.1.1). Only 15 languages are common to both the parametric database and the relevant lexical one: The list is comprised in table 5, which illustrates the matrix of the latter. The two sets of distances, whose correlation is significant and amounts to 0.74418, are plotted on the same graph in figure 2, projecting the lexical measures on the horizontal axis and the syntactic ones on the vertical axis. As observed by Longobardi and Guardiano (2009: 17), two main clusters of points originate from this representation: In the leftmost the increase in both types of distances seems to be proportional; In the other, instead, there is still syntactic variability, while the lexical distances appear to be less scattered. This impression is confirmed by the statistical calculations performed by the above mentioned authors: For example, the standard deviation of the sample of syntactic distances is 0.051 and that of their corresponding lexical distances does not rise above 0.035. Furthermore, the scholars point out that "syntactic distances are considerably smaller than lexical ones", which could be an indication of the fact that syntax evolves slower than lexicon. This fact is also clearly visible from the comparison of the phylogenies they derive from the two sets of distances on the relevant 15 languages using the Fitch-Margoliash method with molecular clock (see the next section): The two trees appear to provide widely similar genealogical classifications, but the total length of the branches of the lexical-based structure is significantly greater than that of the syntactic-based one. On the whole, these observations suggest that syntax appears to be "a potential good indicator of relative taxonomy among sets of distant languages whose vocabularies display too few cognates to make solid clusters identifiable" (Longobardi and Guardiano 2009: 17).

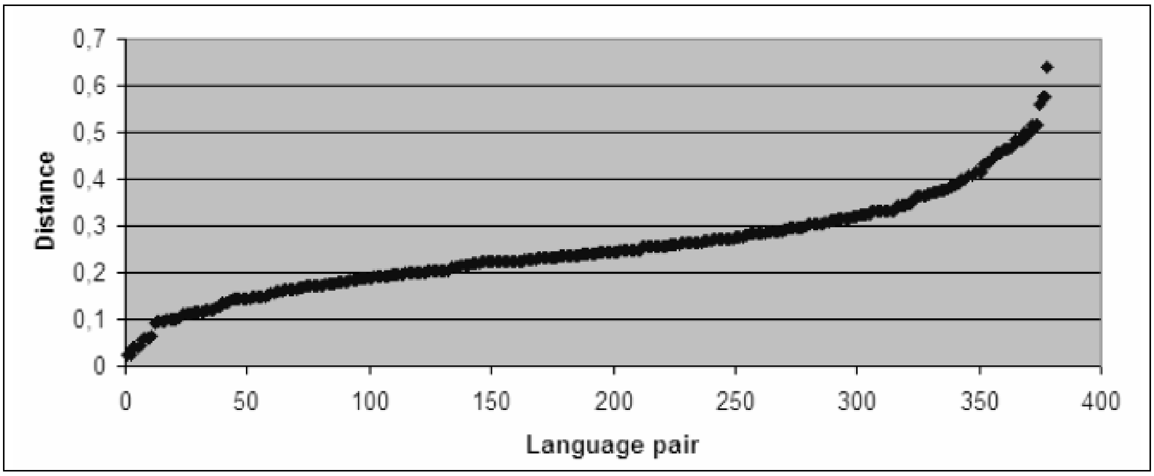


Figure 1: Scattered plot of the syntactic distances derived from Table A.

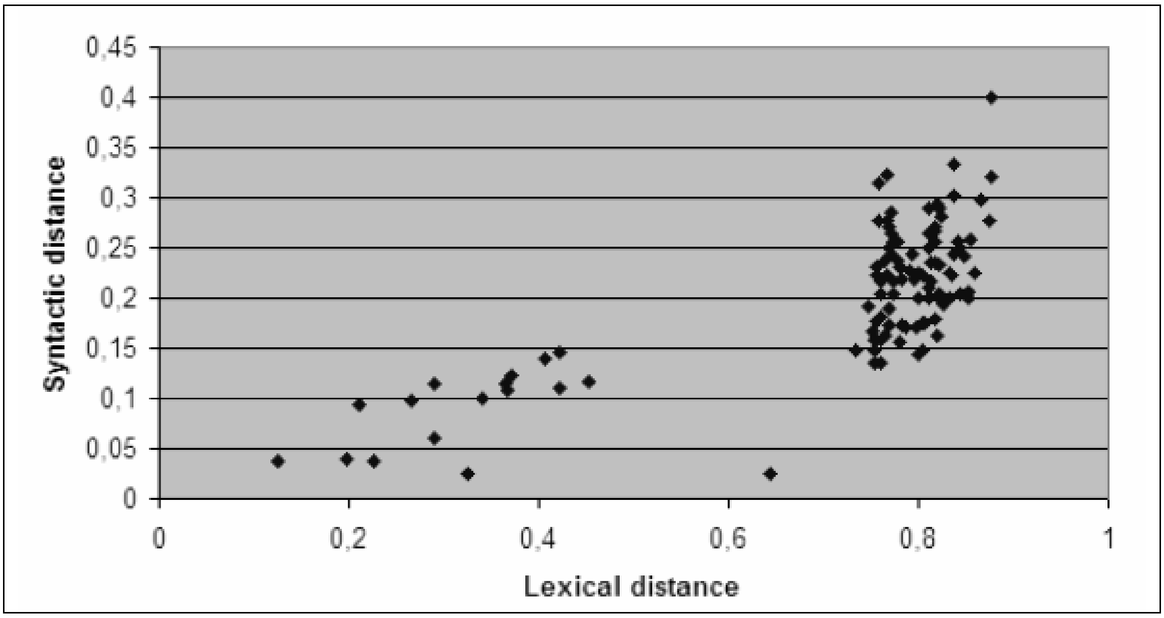


Figure 2: Syntactic distances plotted against lexical distances.

The further empirical test that is carried out on Table A to support this assumption consists of attempting to infer hypotheses on the genealogical relationships between the languages of the sample by means of a specific group of phylogenetic methods. The reasons behind their choice are explained in the next section.

6.1.2. The choice of the phylogenetic methods

What procedures of phylogenetic reconstruction appear to be most apt to properly work out the data of Table A? As pointed out in 4.3, in general each method underlies specific evolutionary hypotheses, thus the choice depends on the assumptions that one can make regarding his own dataset. Of course the more detailed these hypotheses are, the more justified the choice of the method will be and the higher the expectation about the results. Conversely, if one has little knowledge about the evolutionary process that lies behind the observed data, the choice of the proper phylogenetic technique is restricted to those for which the preliminary hypotheses are less specific.
In section 4.3.1 it was pointed out that an important premise of character-based methods is that characters evolve independently, both one with respect to the other and in different lineages (Felsenstein 2004b): Thus if they interact or there are contacts between distinct

branches of the real phylogeny, the procedures in question are not guaranteed to infer the best genealogical tree. In particular it has been demonstrated by many simulation studies in the last decades that "[i]f characters evolve in a correlated manner, the characters in the correlated suite are effectively overweighted (de Queiroz 1993; Chippindale and Wiens 1994)" (O'Keefe and Wagner 2001: 657). The requirement regarding the non convergence of genealogical branches does not poses problems dealing with the parametric dataset, rather the testing of the validity of this condition may be thought to be one of the aims of the analysis. As for the independence of characters, instead, the situation appears to be challenging: Indeed parameters are designed to interact, so that in the sample under study they are almost all related to form a unique large network.

Even leaving aside this difficulty, to what extent are the character-based techniques described in chapter 4 applicable? As for the maximum-parsimony method, it has been anticipated that, apart from the basic version of the procedure outlined in section 4.3.1, different algorithmic implementations and variants of parsimony exist: In particular, Felsenstein (2004a: 73-86) lists several different proposals, e.g. the Camin-Sokal parsimony (Camin and Sokal 1965) and the Dollo parsimony (Le Quesne 1974), whose detailed explanation is not necessary here. What should be detailed, instead, is that each of these variants assume further hypotheses on the evolutionary model underlying the data, which specifically have to do with knowledge of the direction of change: For instance, using Camin-Sokal parsimony it is assumed that, given a set of binary characters for which two states are possible, say + and -, one knows which is the ancestral state , say -, and that change can only happen from state – to state +. In other words, in the evolutionary model entailed by this variant of parsimony the state shift + > - is considered not only more marked, i.e. less parsimonious, by the algorithm, but even impossible. Such an assumption permits the procedure to be more effective when searching for the best phylogeny: Indeed, weighing some changes more than others (in the case of Camin-Sokal - > + has parsimony cost 1, while the cost of - > + is infinite), the parsimony score attained by the different character-state histories, and consequently the global scores attributed to the possible trees tested, may be more accurately estimated, than they would be in a non-weighted approach. This is precisely Ringe *et alii*'s (2002) point, when they argue that the contribute of phonological characters to the phylogenetic reconstruction is very important, in that (often) these characters encode precise information regarding the direction of changes (see sections 2.3.2 and 5.1.2). In any case, as explained throughout the preceding discussion, no directionality hypothesis applies to the parametric dataset due to the fact that no theory of markedness, like Roberts and Roussou (2003) and Robert's (2007), has been assumed in encoding the parameter states in Table A. Therefore, in principle only the basic variant of the parsimony method can be used dealing with the data in question, i.e. precisely the procedure described in section 4.3.1: This entails assuming a unique hypothesis regarding the underlying evolutionary process, i.e. that it is "parsimonious". Nonetheless this restriction alone appears to be rather deficient in allowing the algorithm to perform an accurate and reliable reconstruction. This point, in addition to the fact that parameters actually evolve in a correlated manner, leads one to think that the maximum parsimony method is certainly not the best technique of phylogenetic analysis that could be applied to Table A. However, even without high expectations about the results, some relevant parsimony-based experiments have been carried out.

The maximum-likelihood method, instead, is not applicable at all to Table A at the present stage of development of the research plan. Recall that in order to use this procedure, one has to design a specific evolutionary model on the basis of which the likelihood of phylogenies with respect to a certain dataset is estimated. Putting it simply, this model collects the transition probabilities between the possible character states, so it inevitably includes hypotheses regarding the relative markedness of certain changes or, in other words, knowledge about the probabilities that the characters in question evolve following specific directions. Again, such a model could be outlined relying only on a theory of parametric markedness.

As to what concerns distance-based methods, in section 4.3.2 it has been pointed out that the effectiveness of these procedures mainly depends on the appropriateness and the accuracy of the set of distances that are estimated. In the case of Table A this means that the problem of the implications, which generate the 0s in the strings of parametric values, can be independently treated before the activation of the reconstruction procedure, that is, in the phase in which the distances are calculated. Thus, given that the measure provided in (2) is regarded as sufficiently suitable for the aims of the present work, the issue is to evaluate the alternative distance-based techniques and to determine which one(s) are expected to provide a better phylogenetic account from the parametric data. In chapter 4 an important distinction has been made between procedures that project the original set of distances on an ultrametric tree, i.e. that assume the molecular clock hypothesis (MCH), and those that infer phylogenies without this constraint. In particular, only UPGMA has been shown to reconstruct "clocklike" trees, even if incidentally (see note 23 in chapter 4) it has been pointed out that in actuality least-squares methods can be implemented into algorithms designed to produce ultrametric structures as well.

Does MCH represent a valid assumption dealing with the parametric data? More precisely, does the MCH apply to the evolutionary process from which the data originated? First, there are reasons to claim that nothing prevents assuming MCH here. Recall that, following Felsenstein (2004a: 165-166), a procedure of phylogenetic reconstruction in which MCH is integrated is thought to provide consistent results if the set of distances do reflect at least a quasi-clocklike tree, that is, if the distances derive from an evolutionary process in which the underlying rates of change on distinct lineages do not differ by more than a factor of two (Felsenstein 2004a: 165-166). As a result the requirement is not strict and some degrees of deviation from the optimal condition are tolerated. In any case, as a matter of fact nothing sound can be said for the moment about the rates of parametric change from the sample of data under study: Certainly to hypothesize that the rate is universal (and constant), as it is done in lexicostatistics, appears to be unproven, as well as it would be to claim that in some branches of the phylogeny the rate is more than twice as much as it is in others. Actually to attempt measuring the rate of parametric change in the short period is problematic, again, because of the interactions between the characters. Consider the example given in sketch 2, in which three coeval languages (*L4*, *L5* and *L6*) and five parameters are taken into account, so that the second and the third parameters are implied by the first, i.e. *p2(±) iff p1(+)* and *p3(±) iff p1(+)*. Suppose one knows that the common ancestor of *L4*, *L5* and *L6* is *L0*, but he ignores the specific evolutionary history undergone by the observed languages, i.e. that he is aware of the fact that *L0* is comparable with *L4*, *L5* and *L6* with respect to three loci of variation (*p1*, *p4* and *p5*), but he lacks information regarding the specific series of changes that led from the mother language to the three descendents. In this situation, the only possible estimate that one can make is that in a time *t* (the time that separates *L0* from *L4*, *L5* and *L6*) one difference produced in each of the three lineages, thus the rate of evolution is identical in the three branches and corresponds to *1/t*. Nonetheless, in light of the real evolution occurred, this is false, since in (*L0*, *L4*) and in (*L0*, *L6*) $r = 2/t$, while in (*L0*, *L5*) $r = 1/t$. This example is intended to show that due to the appearance (or disappearance) of relevant parameter states in the course of language development it is not possible to infer a meaningful value for the rate of change. Note that in the case presented in sketch 2 the problem in calculating rates derives from the alternation between +/- and 0-states, not from homoplastic events of any kind, which alone may independently pose difficulties. Anyway, the issue is taken into account again on empirical grounds in section 6.3.2, which is dedicated to the analysis of the correlation of syntactic distances with separation times and geographic distances.

Besides the fact that there is no evidence against MCH, there are also reasons to believe that, to a certain extent, this hypothesis may help one to better work out the parametric data. Indeed, an ultrametric tree reconstructed in a clocklike perspective entails that the quantity of evolution occurred from the root to all the leaves is equal, and this characteristic appears to be particularly compatible with a distinctive property of the parametric distances: The fact that, according to the Anti-Babelic principle, beyond a certain degree of common ancestry all distances tend to an equal average value, i.e. 0.5. This means that given a sample of coeval languages that may be ideally thought to be very "distantly" genealogically related (or completely unrelated, see the discussion in the next section), e.g. E, Wo and Bas, the syntactic distances between each of these varieties and their (ideal) mother are expected to be all approximately equal, i.e. about 0.5, and this situation would be suitable to represent in an ultrametric phylogeny, in which by definition the total length of the branches that connect each leaf to the root is the same.

Moreover, concerning the distinction between clustering and least-squares procedures, nothing precise can be said *a priori* about which of the two groups of methods could permit obtaining more reliable results: For instance, NJ "has sometimes been claimed to perform better than the Fitch-Margoliash method" but "this seems not to be the case, although the difference is not great (Kuhner and Felsenstein 1994)" (Felsenstein 2004a: 168). Nahleh *et alii* (2005b) compare the effectiveness of UPGMA and NJ[10] on Ringe *et alii*'s (2002) dataset (see sections 5.1.2 and 5.3.2), showing how the former technique is clearly inferior with respect to the others they test. However, as clearly stated by the scholars, this is not due to any deficiency of UPGMA in itself, but rather to the fact that it is applied to a set of non-coeval languages, which is quite uncertain dealing with a MCH-based technique: Indeed one cannot expect the quantity of evolution being the same in all lineages if some lineages are already known to be shorter than others.

Actually to deal with a set of non-coeval languages may be problematic not only within a MCH approach, but also in general performing a distance-based analysis due to two inconveniences. The first one is that, as implicitly assumed until here, usually all phylogenetic methods do not allow the automatic insertion of *taxa* in the internal nodes of the trees, but only in the leaves: This constraint is largely not an issue in biology, since in general experiments are conducted only on contemporary species and populations, but it represents a serious limitation for research conducted in linguistics and philology (Macé and Baret 2006: 103-104; Baret *et alii* 2006: 276), where ancient languages and copies of texts commonly provide relevant empirical evidence. Suggestions have recently been proposed regarding quantitative procedures that should help in determining whether or not and at which depth of the structure a certain leaf should be internalized: For instance, Spencer *et alii* (2004: 507) propose to consider the internalization of the terminal nodes whose branch lengths are "very small". In any case, it seems reasonable to claim that the internalization problem can hardly be solved on exclusively quantitative grounds, in that independent evidence, that is, previous knowledge of the actual chronological placement of the *taxa*, appears to be fundamental in dealing with it. The topic is revisited in section 6.3.2 in connection with that regarding the reconstruction of ancestral parametric states.

The second difficulty posed by sets of non-coeval taxonomic units regards the calculation of distances. Suppose one faces a scenario in which there are a modern language *L1* (e.g. It), its direct ancestor *L2* (e.g. Lat) and a sister of *L2*, say *L3* (e.g. ClG): This means that *L2* and *L3* are parallel descendents from a protolanguage *L4* (e.g. proto-IE). If *L2* and *L3* are close

---

[10] Actually besides UPGMA and NJ Nakhleh *et alii* (2005b) take into account other methods as well, which however are all character-based and consequently less relevant here, i.e. those based on maximum compatibility, maximum compatibility, maximum weighted compatibility and Atkinson and Gray's (2006b) likelihood-oriented approach.

enough to *L4*, it may be the case that they show more similarity between each other than that between *L1* and *L2*, because it is possible that the amount of change occurred from *L1* to *L2* is higher than the amount of change that differentiated *L2* from *L3*. In other words, if one relies only on a quantitative comparison, two internal sister nodes may appear to be genealogically closer to each other than they are to their respective descendents (see the discussion in 6.3.1). This is another serious complication that has to be faced and, again, it cannot be solved without the support of some prior knowledge about the units included in the dataset.

Since both the problems may significantly bias the phylogenetic inference, it has been chosen to take into account only the majority set of coeval languages of Table A, i.e. the modern ones, in all the experiments regarding proper phylogenetic reconstruction. The ancient varieties are left aside to be considered only in particularly relevant cases of character-based analyses (see section 6.3.2).

To sum up the contents of this section, on the whole it is expected that distance-based methods are more apt than character-based ones to reconstruct evolutionary histories on the basis of parametric data. In particular, among distance based methods those that integrate MCH may be thought to provide even slightly better results with respect to plain algorithmic procedures, those that do not entail any specific hypothesis on the underlying process of diachronic development. The overview of the methods that can be, and actually have been, applied in the present experimentation is offered in table 6, where four distance-based methods and the only relevant character-based technique, the maximum parsimony method, are listed: Fitch-Margoliash with MCH (henceforth FMC), Fitch-Margoliash without MCH (FM), UPGMA, Neighbor-Joining (NJ) and maximum parsimony (MP). In the same table also the names of the programs in which the above mentioned methods are implemented and that have been adopted to carry out the experiments are specified: In particular, the algorithmic implementations of the Fitch-Margoliash method (in its two variants), NJ and UPGMA, are provided in the package PHYLIP (Felsenstein 2004b) in the programs called Kitsch (FMC), Fitch (FM) and Neighbor (which includes both NJ and UPGMA); The program by means of with the MP analysis have been carried out, instead, is PAUP (Swofford 1998). Starting from the next section the application of these procedures is illustrated.

6.1.3. Inferring trees from full Table A

The phylogenies presented in this section as tree-graphs are intended to be read following the formalism outlined in section 4.1.2. Two points about the interpretation of the genealogical structures require particular attention: On the one hand the branches that connect the different nodes should be seen as proper “σ relations”, in Longobardi’s (2003b) sense; On the other hand, the roots that are reconstructed in the diverse phylogenies should be regarded as ideal abstract proto-languages from which all the varieties comprised in the sample descend, but not necessarily as actual historical objects. This means that the analyses introduced here are not intended to provide evidence in favour of a monogenetic view of language evolution, i.e. of the idea that all the languages in question (and, in general, all the languages of the world) are genealogically related. Even if this hypothesis is not unconceivable within a biolinguistic perspective on language variation and in light of a likely monogenesis of the language faculty in the human species (e.g. Longobardi 1999: 6), to corroborate its plausibility appears to be difficult in general and certainly much beyond the scope of this investigation.

The majority of the trees illustrated in the following discussion are derived starting from a preliminary procedure of perturbation of the parametric data: The bootstrap[11] (Bradley 1979;

---

[11] Other techniques used to empirically test the variability of estimates, including for instance the so-called “jackknife” (e.g. Felsenstein 1985; Wu 1986), which largely has the same properties of the

Felsenstein 1985). This technique has already been mentioned in section 4.3.3, claiming that it is proposed, among others, by Cavalli-Sforza *et alii* (1994: 35) to test the robustness of the genealogical information contained in a dataset and to find cues of horizontal transmission between different lineages. Essentially, given the original character matrix containing 63 parameters, the bootstrap runs for *n* cycles to create a set of *n* resampled datasets, so that for each one the procedure draws 63 times, and each time a parameter is chosen randomly from the original list: In this way the same parameter may be included in a resampled dataset many times or excluded at all. What should be done after having obtained the set containing *n* resampled character matrices? If one needs to use the outcome of the bootstrap as input for a distance-based phylogenetic method (as is the case here), from each new dataset a distance matrix has to be derived, then these *n* matrices can be worked out by the appropriate algorithms to infer *n* trees. At this point, the last passage of the process entails summarizing the results into a "consensus tree" (Margush and McMorris 1981), in which each fork is provided with a number, that indicates how many times the group of languages to the right of the fork has been inferred in the *n* trees: Higher numbers correspond to better supported groups. A basic criterion to draw a consensus tree is to build it including all the groups that occur at least 50% of the time ("majority rule"), but in the present experimentation the "extended majority rule" has been adopted, so that the consensus trees illustrated in the following discussion comprise all the groups of languages reconstructed in the relative majority of the times in the set of *n* trees. The entire procedure, from the bootstrap to the consensus tree, is illustrated in sketch 3 by means of a simplified dataset including 6 languages and five characters. The programs used to carry out the bootstrap and to calculate the distance matrices from the resampled datasets have been specifically designed for this work (taking advantage of the valuable support of Luca Bortolussi of the Department of Methematics of the Università di Trieste in the early stages of development of the algorithms), as all the programs used to perform the operations in question in the following discussion if no different indications are provided. The consensus trees, instead, are derived by means of the program Consense of the package PHYLIP (Felsenstein 2004b).

Observe that in the case of distance-based methods, that reconstruct a unique tree, the consensus procedure is precisely applied to synthesize the multiple outcomes of the bootstrapped dataset, which have always been resampled 1000 times. The maximum parsimony method, instead, may produce on its own many diverse phylogenies that optimize the parsimony score (see section 4.3.1): Therefore in the following discussion the consensus trees derived from the parsimony-based analysis are not intended to have been obtained bootstrapping Table A, but just as the summary of the outcomes produced by the algorithm directly from the original dataset. It is opportune to specify that bootstrap has not been applied performing experiments with maximum parsimony because, as explained, this method already poses various difficulties when used on the parametric data and combining it with the perturbation of the character matrix easily leads one to attain inconsistent results.

The FMC tree of the 23 modern language of Table A is provided in figure 3. Observe that for the moment only the relative classification of languages is taken into account, while consensus scores are reconsidered in section 6.1.4, since they require a discussion on their own. All the groupings that were expected are correctly identified. In particular, Bas stands alone as leaf of the most external branch of the phylogeny, meaning that in this sample it is essentially genealogically unrelated to all the other languages. Note that the method automatically poses the root in the path that divides it from the cluster including the rest of the varieties: This means that this lineage may be thought to have been the first that diverged in

---

bootstrap, and various more or less controversial permutation tests, are accounted in detail in Felsenstein (2004a: 335-363).

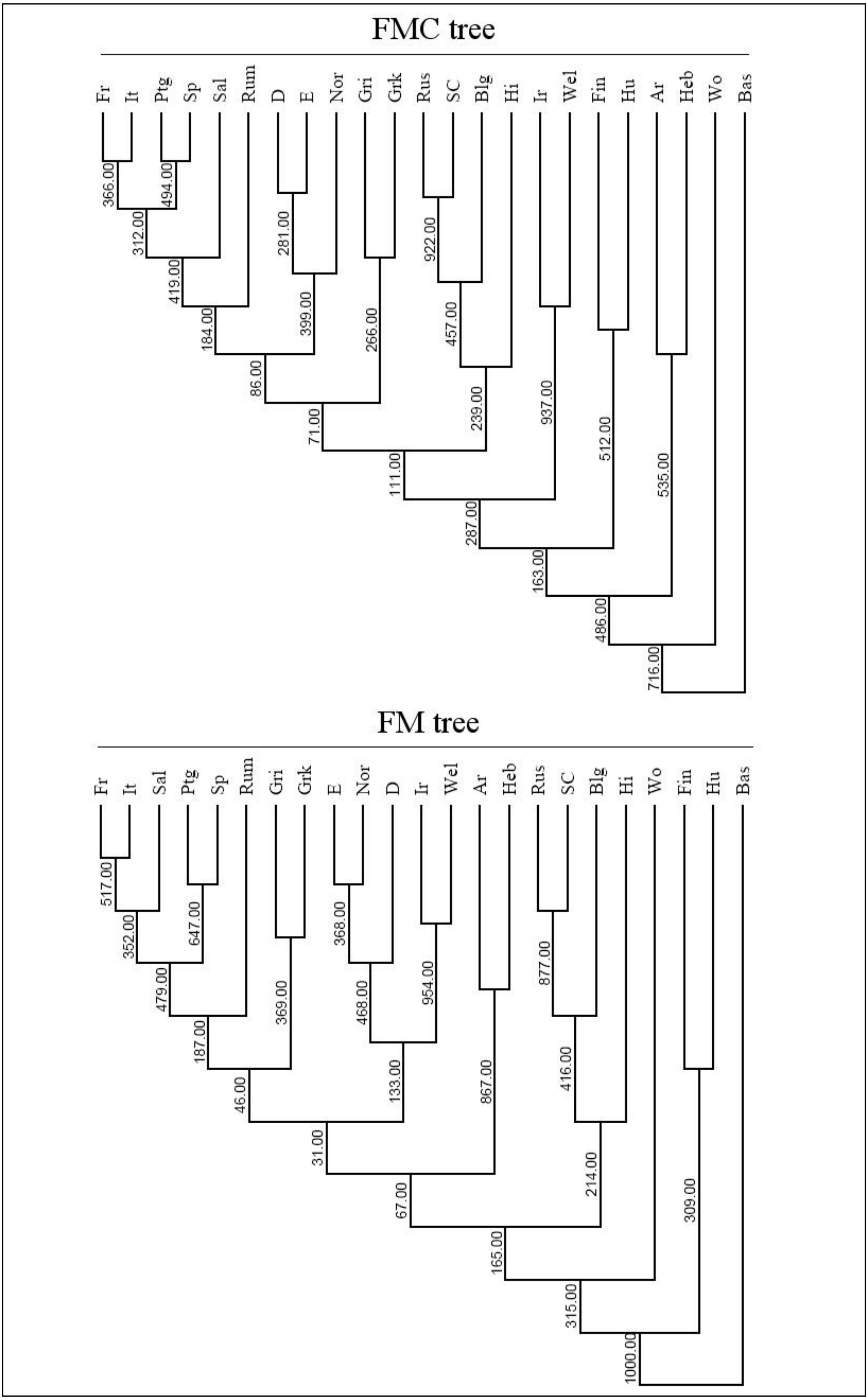


Figure 3: FMC and FM trees derived from the full dataset, all modern languages.

the evolution of an ideal ancestor common to all the varieties of Table A. Similar observations apply to Wo, whose branch separates immediately after that of Bas, so that the Niger-Congo variety appears to be isolated in this sample. These successive splits lead to the identification of an internal node from which Semitic, Uralic and Indo-European originated: This might be intended to reflect some version of the Nostratic hypothesis, but actually it is not very probative, in the sense that the relevant node is defined only as a result of the comparison of two languages (Bas and Wo), than one could already expect to be outliers. However, in the FMC tree a specific branch indicates the divergence of the Semitic family, represented by Heb and Ar, that is followed by the separation of the Uralic, comprising Hu and Fin. The Indo-European family is then correctly recognized and all its members are positioned in a compact cluster: The groups that are recovered by means of the quantitative analysis are, in order of progressive separation, Celtic, Indo-Slavic, Greek, Germanic and Romance. Note that no Italo-Celtic unity is recognized, which is more in accordance with Atkinson and Gray's (2006b) account than with the traditional one and Ringe *et alii*'s (2002). Nonetheless, the FMC phylogeny strongly suggests the same relative similarity of Slavic and Indic, the only "*satəm* languages" in this sample, that is indicated in Ringe *et alii*'s (2002) structure: This genealogical nearness is not captured by Atkinson and Gray's (2006b) classification, which, instead, in agreement with that entailed by the FMC analysis, indicates a more pronounced degree of relatedness between Germanic and Romance.

Three Indo-European groups show interesting internal articulations in the FMC tree (of course in the case of the other groups, that comprise no more than two languages, the question of the internal articulation is not relevant). That of Slavic is wrong, because the established distinction between Eastern (Rus) and Southern (Blg, SC) varieties is not reconstructed. This mistake coherently reflects the information contained in the dataset and encoded in the distance matrix: Indeed the parametric settings for the nominal syntax of SC and Rus are largely identical (on this point see also Longobardi and Guardiano 2009: 15). The classification of Germanic languages, instead, adequately reproduces the division between Northern (Nor) and Western (E, D) varieties, while that of Romance, even if slightly different from, say, the classical accounts derived from Diez (1836), may be regarded as widely satisfactory: Rumanian results to be the more genealogically distant language and it is opposed to a West-Romance cluster that in some way reproduces an East-West continuum, from Sal to (Fr, It) and (Sp, Ptg). The position of Sal, which is separated from It, is not surprising, since the South-Italy dialectal variety has undergone a long history of close contact and, consequently, interference with Gri, which has caused many of its important linguistic traits to significantly diverge from those of standard It.

The FM phylogeny is included in figure 3 as well. Since the outcome of this basic version of the FM method is an unrooted tree, the root has been assigned to the structure by means of the outgroup criterion following the indication provided in the precedent experiment, i.e. considering Bas as outlier. It is evident that this tree includes a fundamental mistake: The unity of the Indo-European family is not recognized, because Semitic is placed between the main Indo-European cluster (comprising Germano-Celtic and Greco-Romance) and the most external relevant group (Indo-Slavic). Even if the structure presents other important differences with respect to the FMC one, e.g. the different genealogical depth at which Uralic and Wo separate, further comments are not necessary to evaluate it, in that the fact that the method fails to recover the correct classification at family-level entirely disqualifies the result.

The trees inferred by means of the clustering techniques are shown in figure 4 and appear to parallel the FMC and the FM classifications. The UPGMA tree is identical to the FMC one, thus it is substantially satisfactory. The NJ phylogeny, instead, poses diverse problems: Again, the most relevant one, as a consequence of which the structure may be regarded as completely inadequate, is that concerning the insertion of Semitic into the Indo-European area.

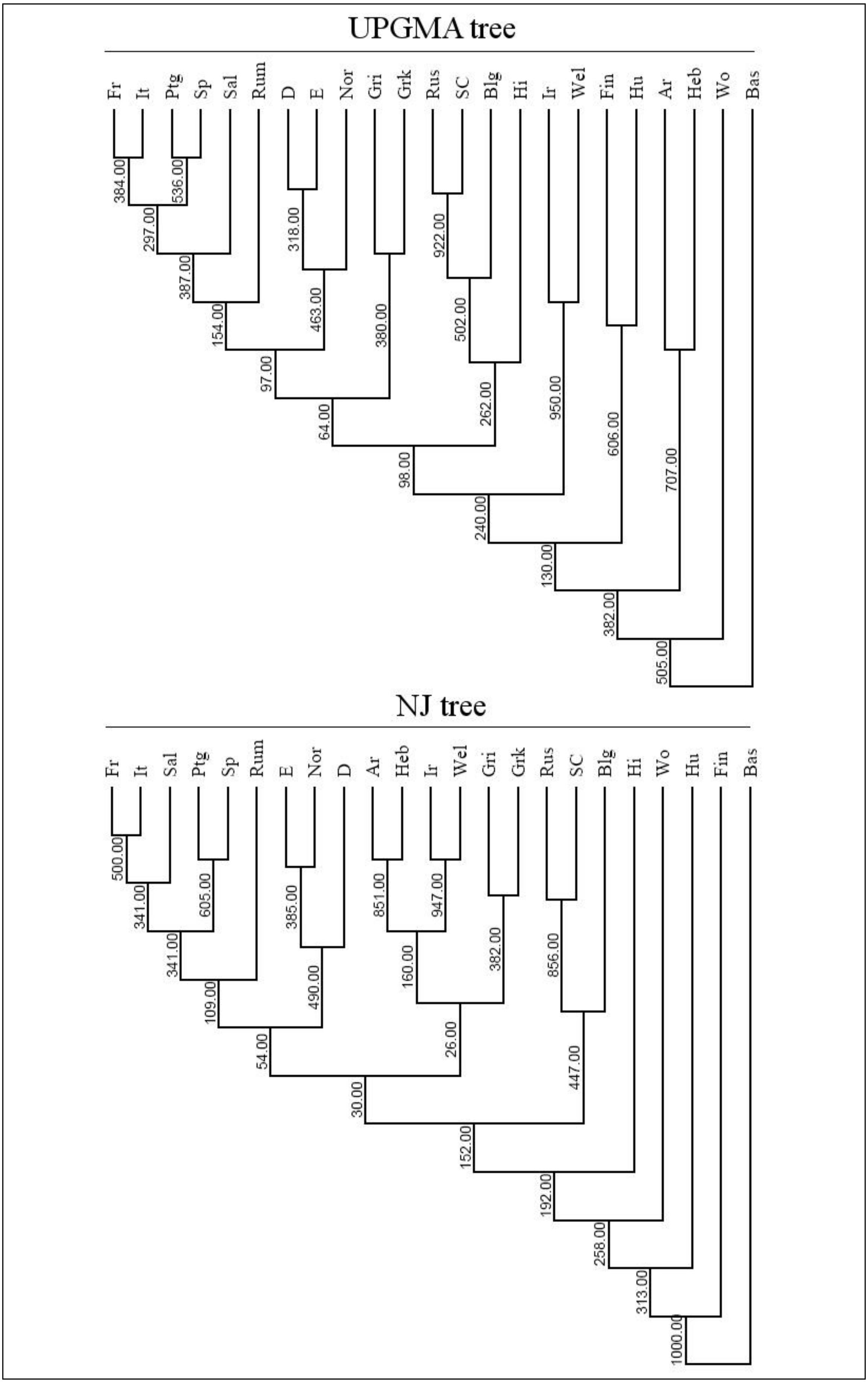


Figure 4: UPGMA and NJ trees derived from the full dataset, all modern languages.

The first set of results drawn from the entire set of modern languages seems to corroborate the suggestion put forward in section 6.1.2, i.e. that distance-based methods integrating MCH (i.e. FMC and UPGMA) are more apt to work out the parametric data than those that do not entail a clocklike evolutionary model (FM and NJ). In particular, note that the primary problem of the FM and the NJ phylogenies precisely derives from the fact that a long-range relationship, like that, which might exist between Indo-European and Semitic, is incorrectly accounted for in the classification.

The same series of methods has been used to investigate in greater detail the classification of Indo-European languages. Indeed it is not granted that, by performing the analyses exclusively focusing on the languages that belong the specific family in question, one may obtain the same distribution of groups inferred in the whole cross-family phylogeny. Recall from section 4.3.2 that all the distance-based phylogenetic techniques basically aim at projecting the complete set of observed distances given in the matrix onto the trees in the form of branch lengths, i.e. at deriving a set of additive distances: This means that the overall structure that is provided as output of the procedures is intended to represent the best possible approximation of the whole set of observed distances, which may not coincide with the best possible approximation of a subset of the set. Putting it simply, inferring the tree comprising all the modern languages of Table A it is possible that the relative position of some Indo-European groups is modified by the presence of non-Indo-European languages. For example, it has been observed in the previous NJ tree that Celtic and Semitic occur in the same cluster and there are reasons to believe that their "reciprocal attraction" may also be in part responsible for the erroneous classification obtained in the FM tree: Thus what would be the reconstructed genealogical status of Celtic also in the FMC and UPGMA trees, if its position were not influenced by the presence of Semitic?

The phylogenies derived by means of FMC, FM, UPGMA and NJ for the 17 modern Indo-European languages of Table A are introduced in figure 5. In each analysis a "null" taxonomic unit has been inserted to provide an independent outgroup, i.e. Out, whose distance is assumed to be the same with respect to all the other relevant languages and much beyond the Anti-Babelic limit, i.e. 1: This assures that Out always has the role of (ideal) outlier without influencing the distribution of the other languages. Also in this case the FMC and the UPGMA trees are equal and they largely correspond to the subtrees included in the cross-family FMC and UPGMA phylogenies of figure 3 and 4, except for the fact that in this case the internal genealogical hierarchy of Germanic is incorrect. Since the same group shows an alternative, and equally wrong, articulation in the FM and NJ classifications, it is reasonable to claim that this is due to fundamentally ambiguous parametric evidence, that leads the quantitative analysis to provide unstable outcomes. This is not surprising, because the history of E in particular includes many events of secondary convergences that may have led this variety to show an uncertain genealogical status within Germanic: One may think, for instance, about the contact, which occurred between Old English and Old Norse from the IX to the XI century in the Danelaw[12], or about the well-known introduction of Anglo-Norman, a northern dialect of Old French, in England starting from the XI century, whose consequent changes, at least for what concerns lexicon, are still visible in Modern English. Therefore at this point it seems reasonable to argue that the present sample of parameters cannot help to define a precise genealogical account for the classification of the Germanic languages in question within their group, even if on the basis of the FMC and UPGMA trees derived from the 28-language set no problems appeared to arise. Looking closer at the FM and NJ trees of figure 5, instead, it may be noticed that in the former, due to the absence of Semitic, the Germano-Celtic macro-group appears to be closer to Romance than Greek, while in the latter

---

[12] About the syntactic influence of Old Norse on Old English see for instance the references provided in Roberts (2007: 391-399).

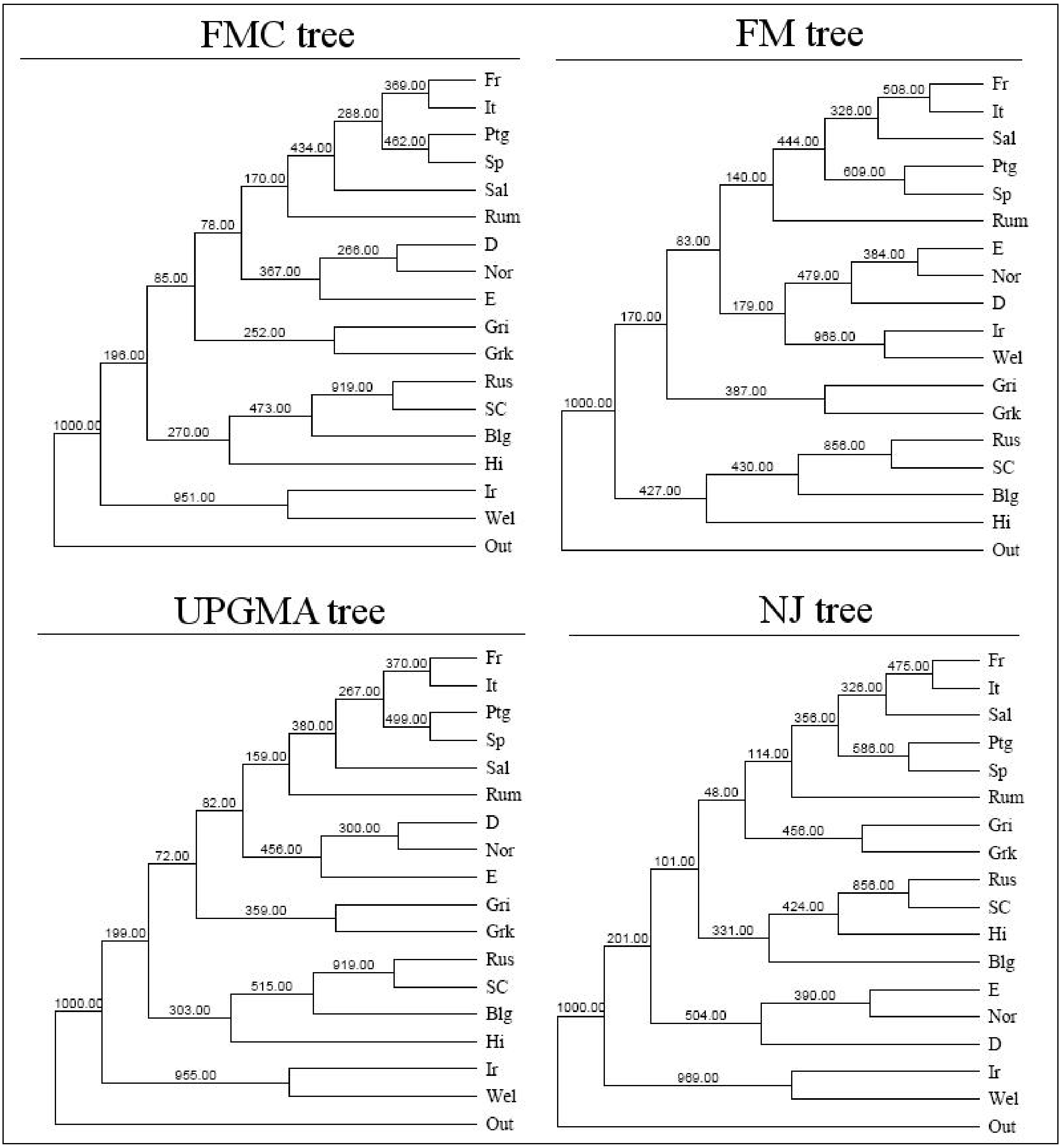
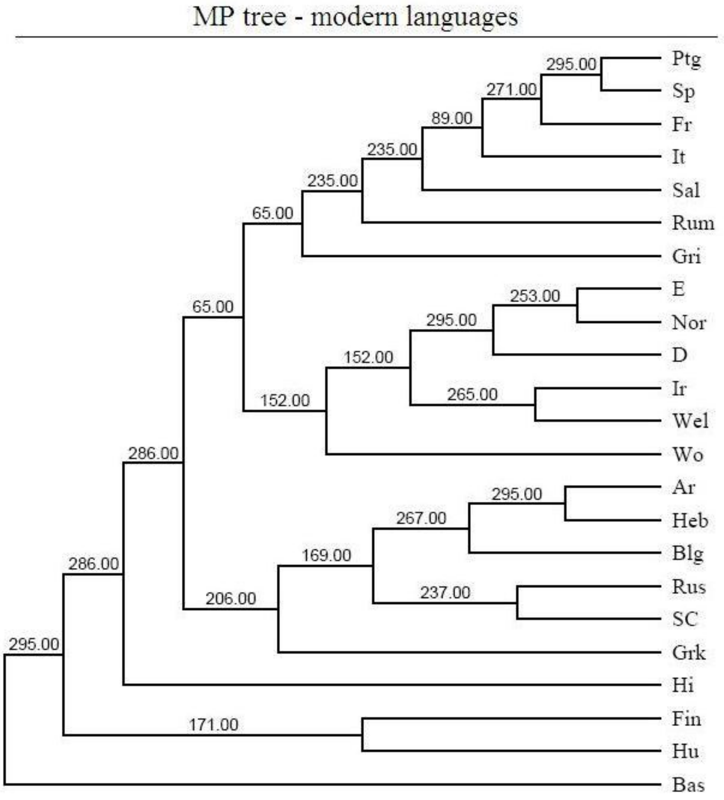


Figure 5: FMC, FM, UPGMA and NJ trees derived from the full dataset, only IE languages.

the relative order of the lineages is largely modified with respect to that found in figure 4. What is relevant to point out is that these outcomes seem to support the idea that MCH may provide an assumption that is applicable to the parametric dataset: Indeed the genealogical hypotheses drawn by means of methods that integrate it appear to show a consistent convergence, which, at this stage, is not recorded with FM and NJ.

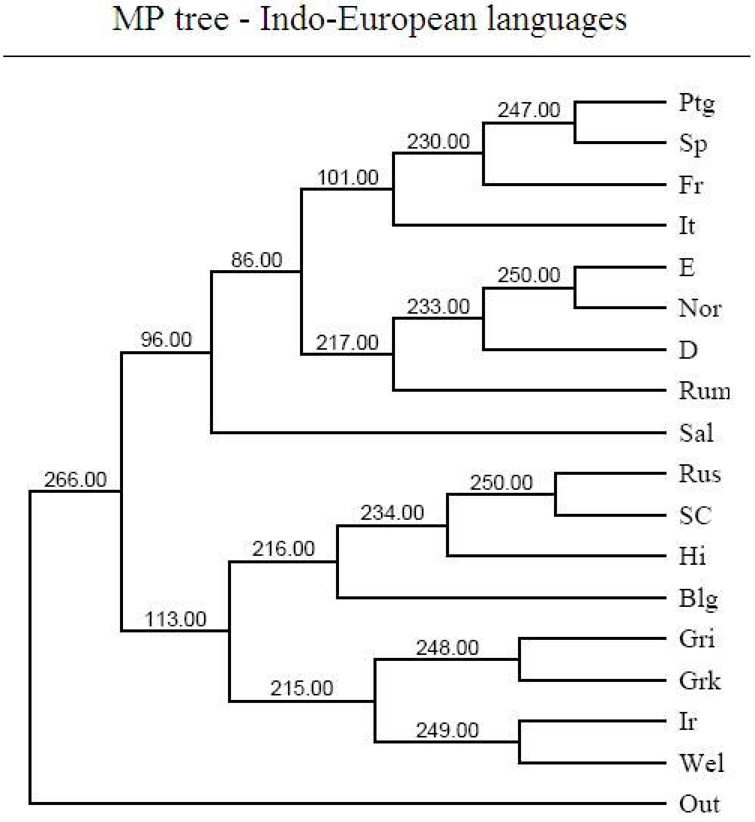
The maximum-parsimony analysis performs clearly worse that distance-based methods on the parametric data. The MP trees reconstructed on the whole set of modern languages and on the Indo-European are provided in figure 6. The former phylogeny is the consensus of 295 different trees, all entailing 114 changes of state of the 63 parameters, which is the most parsimonious global character-state history found by the procedure: Some limited parts of the


MP tree - modern languages

Number of parsimony-informative characters: 44

Score of best tree(s): 114

Trees retained = 295


MP tree - Indo-European languages

Number of parsimony-informative characters: 27

Score of best tree(s): 64

Trees retained = 266

Figure 6: MP trees derived from the full dataset, modern and IE languages.

structure do not represent wrong inferences (e.g. Romance, Germanic and Celtic are recognized, as well as Semitic and Uralic), but the entire picture is evidently inadequate. The MP consensus tree reconstructed on the Indo-European modern languages (synthesizing 266

MP trees whose parsimony score is 64) is less problematic, but still many wrong genealogical relationships are hypothesized: Rum appears to be closer to Germanic than to Romance, Sal is represented as the outlier of the Germano-Romance group and the organization of the "*satəm*" varieties is confused.

Such results were expected and may be thought to be due to various causes, that make the MP analysis hardly applicable to this version of the dataset. First, as reported in section 6.1.1, there is a limited number of informative parameters for the character-based analysis: In the case of the entire set of modern languages they are 44 and they decrease to 27 focusing on the Indo-European varieties. In particular, many PS-uninformative parameters are fundamental to ground long-range relationships, i.e. to isolate specific languages, as Bas and Wo (e.g. *p37* and *p32*), and in this cases the procedure loses fundamental pieces of evidence regarding specifically the family-level grouping. This inconvenience has the same negative impact on the phylogenetic inference as the other two above mentioned problems, those concerning the interaction between parameters and the lack of information on the direction of parametric changes. Finally, another possible difficulty may be related to the occurrence of cases of homoplastic development in the subset of PS-informative parameters: However, homoplasy is not only an issue for character-based analysis, but it also influences the effectiveness of the distance-based methods. The next section precisely covers some proposals to deal with homoplasy in the phylogenetic reconstruction.

6.1.4. Consensus scores

The consensus scores reported at each fork of the distance-based trees considered so far[13] deserve particular attention, in that they suggest interesting observations regarding the solidity of the inferred groupings. A specific section is dedicated to comment them at this point because some general preliminary remarks, that will be useful also for the evaluation of the results presented in the following sections, need to be made.

First of all, it is opportune to point out that no absolut consensus threshold, according to which one could establish that a certaing group is indubitably well-supported by the data, may be assumed. In light of a preliminary overview of the scores obtained in the diverse experiement, here it appears reasonable to consider 400 (which, for example, is about the score attained by West Romance and Germanic in the FMC tree) a satisfactory threshold, over which the identification of a certain group may be thought to be acceptably sound. Consensus scores below (or near) the value of 100, instead, may be considered too weak, i.e. to be completely unreliable genealogical hypotheses. Having established these limits, one might consider all scores ranking between 100 and 400 not entirely safe, but still sufficiently interesting at least as tentative proposals of classification.

The consensus scores of the different forks of the FMC tree significantly vary: The highest values are assigned to Celtic (whose unity is reconstructed in 937 trees out of the 1000 derived from the bootstrap), (SC, Rus) (922) and non-Basque (716) (recall that (SC, Rus) represents the unique wrong inference included in the structure). Other scores that overcome (or are in the neighbourhood of) the threshols of 400 in this phylogeny are those assigned to the group (Sp, Ptg) and to West Romance (see in any case the observations pointed out in section 6.1.3), Germanic, Slavic, Uralic and Semitic. On the whole this picture may be considered significant in light of the relatively small set of characters to which the bootstrap has been applied. The identification of a "Nostratic" grouping appears to be well-supported as well. Thus all forks corresponding to conventional classification hypotheses are identified without uncertainty by the analysis, with two exceptions. One is that of the Indo-European

---

[13] The consensus scores concerning the MP trees, which should be regarded as largely deficient on ther own, do not require to be discussed deeper in detail.

family, wich attains the value 287: As revealed by the NJ tree in figure 4 and already specified in the previous discussion, this is plausibly due to the “attraction” between Celtic and Semitic. The other exception regards the cluster (Gri, Grk), identified in 266 trees out of 1000: In this case the peculiar status of Gri, whose history includes massive contact with It and Sal, may help to understand the relatively weakened genealogical relation with Grk. The lowest support, instead, is provided to forks that define relationships between Indo-European groups, i.e. those corresponding to the separation of Indo-Slavic (111) and Greek (71) and that indicating the closeness between Germanic and Romance (86): This situation appears to reflect the problems that are found in the traditional perspective dealing with family-internal grouping (see section 5.3.1).

If in the other phylogeny obtained under MCH, i.e. the UPGMA tree, the distribution of the consensus scores largely parallels that given in the FMC structure, to some extent the situation is different in the case of the FM and the NJ ones. Even if in both of them the overall structure is not acceptable due to the mistake regarding the insertion of Semitic in the Indo-European family, it is worth observing that in these cases the support in favour of short-range relationships as those between It and Fr and between Gri and Grk is slightly increased: In any case, these trees are perfectly in line with the preceding ones with respect to the fact that groups are identified with higher confidence, while the relations between them are inclined to remain more uncertain. The same observation applies to the trees derived on the 17 modern Indo-European languages of Table A.

The average consensus scores of the eight trees illustrated in figure 3, 4 and 5 is given (5ab).

(5) a. All modern languages:
    FMC tree: 392.52
    FM tree: 389.05
    UPGMA tree: 399.24
    NJ tree: 371.95
  b. Indo-European modern languages:
    FMC tree: 411.25
    FM tree: 426
    UPGMA tree: 427.19
    NJ tree: 409.13

Observe that there are not significant variations of the average scores depending on the specific methods, but only a tendency towards an increase of support for the different inferred groups when dealing with the subset of Indo-European languages. In the following discussion the average scores collected in (5) will be compared with those derived from further experimental results.

6.2. Selecting and weighting characters

In the last section it has been demonstrated that Table A contains substantial genealogical information, which means that the vertical direction of language evolution may be successfully recovered from parametric comparison. The phylogenies derived under MCH have been shown to be satisfactory especially with respect to the classical established classifications and consistent both when obtained by means of different procedures (FMC and UPGMA) and on different samples of languages. Nonetheless working without MCH various problematic classifications are inferred. At this point it is important to consider how much homoplasy biases the phylogenetic reconstruction and if it is possible (at least) to limit its influence on the phylogenetic reconstruction. In the following discussion these points are considered in turn, starting from the outline of an empirically-grounded selection of genealogically-relevant parameters and then taking into account diverse criteria to weigh parameters according to their phylogenetic importance. These lines of investigation are primarily intended to draw a more detailed account of the parameter set and to show how it is possible to improve the performance of the quantitative analyses on the basis of more accurate characterizations of the data.

6.2.1. Empirically-based selection of parameters

To select characters and to exclude those, that are known or believed to have undergone an homoplastic evolution, is a common practice both in traditional and in quantitative phylogenetic-oriented studies. Indeed it is the first reasonable approach that one can adopt to reduce the effect of parallel development and contact-driven convergences in the data. It has already been pointed out that within the classical comparative method independent knowledge may often help to distinguish between reliable and unsound evidence, for instance in the case in which specific morphological paradigms are demonstrated to be not arbitrarily encoded (see note 3 in chapter 4). On the other hand, at least since Swadesh (1952) in the entire lexicostatistics perspective great efforts have been made to filter meaning lists with the goal of finding the "core vocabulary" of languages, which is thought to be particularly resistant to borrowing, conservative and so on. Such a caution is rather far from the mass-comparison practice instead. Furthermore, among the examples that can be found within the quantitative perspective, one may recall the hybrid dataset used by Nakhleh *et alii* (2005a), which is precisely a "screened version" of that adopted in Ringe *et alii* (2002) and Nakhleh *et alii* (2005b) and has been built excluding polymorphic lexical characters and, in general, all characters that clearly exhibited parallel development in the preceding investigations (Nakhleh *et alii* 2005a: 394; Further references are provided by the authors). On the whole, choosing characters appears to be necessary not only to reconstruct better phylogenetic trees, but also to derive important information regarding the characterization of the data and deeper accounts of the domain from which data have been collected. Both the latter objectives have been primarily pursued outlining the selection of parameters introduced in this section and one main question has been posed: Is it possible to identify a parameter subset that comprises more genealogical evidence and less homoplasy than the entire set (or, ideally, exclusively genealogical evidence and no homoplasy)?
To answer this question requires recalling the sub-modules of nominal syntax according to which the parameters of Table A may be grouped, which is done in (6) for convenience.

(6) - (*p1 – p6*): Status of features Person, Number and Gender;
- (*p7 – p16*): Definiteness;
- (*p17 – p24*): Countability and related concepts;
- (*p25 – p31*): The grammar of Genitive case;
- (*p32 – p41*): Properties of adjectival and relative modifications;
- (*p42 – p50*): The relative position of N inside the DP and its raising;
- (*p51 – p55* and *p60 – p63*): The status of demonstratives and other determiners;
- (*p56 – p59*): The grammar of possessive pronouns.

(adapted from Guardiano and Longobardi 2009: 10)

One of the sub-modules in (5) may be suspected of being particularly homoplastic *a priori*: That concerning definiteness. This idea mainly derives from Heine and Kuteva's (2006: 97-

139) areal-typological description of the geographic expansion of definite articles in Europe[14]. The scholars are convinced that language contact played a decisive role in triggering or accelerating the grammaticalization of articles in the majority of the cases they discuss, which are focused in particular on Slavic varieties. In synthesis Heine and Kuteva (2006) claim that the majority of the languages of the world (about two thirds) lack articles Dryer (1989: 85), as the largest part of the "Old European" languages essentially did (see the references in Heine and Kuteva 2006: 98-99), and that this fact plus, among other reasons, the geographic distribution that articles currently exhibit in the European area, suggests that the diffusion of these elements has been largely driven by secondary convergence. Assuming the spread of definite articles in Europe to match with that of definiteness, in light of Heine and Kuteva's (2006) account one should expect to find large homoplastic patterns of variation in the relevant nominal sub-domain, therefore parameters belonging to it should be more cautiously inserted in a genealogical-focused selection.

As for the other sub-modules, for the moment no data-external reasons can be found to believe that they are likely to have developed in an extensive non-genealogical fashion. In any case, a survey of the set of splits collected in table 2 allows identifying areas of the data, in which homoplasy manifestly appears to pose more problems. Thus the first criterion adopted to perform the selection has been to exclude the parameters determining splits evidently recognized as "genealogically inadequate", i.e. those entaling divisions of the set of languages that group together varieties known to be unrelated or to have undergone secondary convergences. The clearest and simplest examples of these cases are provided by *p40* (± *overt Mod°*) and *p12* (± *def checking N*), whose respective splits are given in (7)[15]:

(7) s(*p40*) = {(It, Sp, Fr, Ptg, Lat, Gri, E, Blg, Sc, Rus, Heb, Ar), (Sal, Wo)
s(*p12*) = {(It, Sal, Sp, Fr, Ptg, Gri, Grk, Got, E, D, Ir, Wel, Heb, Ar, Wo, Hu), (Rum, Nor, Blg)}

The + setting of *p40* shared by Sal and Wo may be safely drawn back to parallel development, since the two languages have never come in contact (see also the discussion regarding the procedure to test the plausibility of borrowing in section 6.3.1), while, for what concerns *p12*, it appears sound to hypothesize that the presence of postnominal clitic elements to express definiteness in both Rum and Blg represents the instantiaion of a typical secondary convergence occurred in the Balkan area (e.g. Tagliavini 199: 259), while the same trait seems to have independently emerged in Nor (Longobardi 2001b: 587-588). Apart from these well-grounded cases, the other parameters, whose splits have been regarded as inadequate to provide genealogically-useful information, are indicated in (8).

(8) s(*p4*) = {(Sp, Grk, Ar, Bas), (It, Sal, Fr, Ptg, Rum, Gri, E, Fr, Nor, Blg, SC, Rus, Heb, Wo, Hu, Fin, Hi)}
s(*p7*) = {(SC, Rus, Hu, Fin, Bas), (others)}
s(*p17*) = {(It, Sal, Sp, Fr, Ptg, Rum, Gri, Grk, E, D, Nor, Hu, Hi, Bas), (SC, Rus, Wel, Heb, Ar, Fin)}
s(*p20*) = {(It, Sal, Sp, Fr, Ptg, Rum, Gri, Grk, E, D, Nor, Blg, SC, Rus, Heb, Wo, Hu, Fin, Hi, Bas), (Wel, Ar)}
s(*p23*) = {(It, Sal, Sp, Fr, Ptg, Rum, Gri, Grk, E, D, Nor), (Blg, Bas)}
s(*p25*) = {(It, Sal, Sp, Fr, Ptg, E, D, Nor, Blg, Ir, Wel, Heb, Ar), (Rum, Gri, Grk, SC, Rus, Wo, Hu, Fin, Hi, Bas)}

---

[14] Heine and Kuteva (2006) consider indefinite articles as well, but for what concern them they claim: "[W]e are not aware of any comprehensive areal-typological study dealing with the languages of Europe. Our account of indefinite articles in Europe is therefore based on individual publications on individual languages and – in isolated cases – on elicited data from informants" (Heine and Kuteva 2006: 97).

[15] Differently from table 2, the splits reported in this section include only modern languages.

s(*p27*) = {(It, Sal, Sp, Fr, Ptg, Rum, E, Blg, Heb, Ar, Hu), (Gri, Grk, D, SC, Rus, Ir, Wel, Fin, Hi, Bas)}
s(*p28*) = {(It, Sal, Sp, Fr, Ptg, Gri, Grk, Blg, SC, Rus, Ir, Wel, Wo), (Rum, E, D, Nor, Heb, Ar, Hu, Fin, Hi, Bas)}
s(*p29*) = {(Rum, Gri, Grk, D, SC, Rus, Ir, We, Heb, Ar, Wo, Hu, Fin), (E, Nor, Hi, Bas)}
s(*p53*) = {(Sp, Rum, Grk, Blg, Ar), (It, Sal, Fr, Ptg, Gri, E, D, Nor, SC, Rus, Wo, Hu, Fin, Hi, Bas)}
s(*p55*) = {(It, Sal, Sp, Fr, Ptg, Rum, Gri, E, D, Blg, Heb, Wo, Bas), (Grk, Nor, Ar, Hu)}
s(*p61*) = {(Sp, Ptg, Bas), (It, Sal, Fr, Gri, Grk, Heb, Ar)}
s(*p62*) = {(Ir, Wo), (Wel, Heb, Ar)}

Having excluded the 15 parameters mentioned so far, an extensive empirical testing has been carried out on the remaining subset of the data to refine the selection, with the aim of obtaining the largest possible convergence of results by means of the different distance-based phylogenetic methods. The best parameter selection found, i.e. the one that has allowed attaining the desired convergence, comprises 40 parameters and is presented in table 7. Considering it in light of the devision of Table A into syntactic sub-domains, it may be noticed that many parameters regarding definiteness are left out (*p7*, *p8*, *p11*, *p12*, *p13*), as well as most regarding Countability (*p17*, *p20*, *p23*, *p24*), Genitive (*p25*, *p26*, *p27*, *p28*, *p29* and, to a certain extent, *p58*) and demonstratives or other determiners (*p53*, *p55*, *p60*, *p61*, *p62*). The sub-modules that have been almost entirely preserved are instead those concerning the status of features Person, Number and Gender (except for *p4*), adjectives (except for *p35*, that has a strong morpho-syntactic content), N-raising (except for *p40*, whose formulation is currently under discussion) and possessives (except for *p58*). As shown in the implicational structure in sketch 4, the selected parameters are largely independent of the interaction with settings of parameters that have been excluded, whose nodes are represented as circles.

The same outfit of results provided on the basis of the full dataset is proposed here starting from figure 7, which covers the FMC and FM trees of the modern languages, and figure 8, which includes the UPGMA and NJ trees. Of course, since they have been used precisely as cues of the consistency of the selection, these phylogenies widely converge on the same genealogical classification and do not present mistakes, apart from the persistent inconvenience regarding the internal articulation of Germanic. On the other hand, the Slavic languages come to be classified in agreement with the established account. Only two details differentiate the four trees: The first corresponds to the identification of a Germano-Celtic macro-group inside the Indo-European family, that is given in the FM and NJ trees, but not in those reconstructed under MCH; The second has to do exclusively with the FM phylogeny, in which no “Nostratic” unity is recognized. As pointed out in section 6.1.4, the highest consensus scores, i.e. those approaching or overcoming the threshold of 400, are those attributed to established groups, while the scores assigned to the relations between the groups tend to be low, even if superior with respect to those obtained on the basis of the full dataset on average. Moreover, it is important to note that the relative distribution of the Indo-European groups appear to be stable in these structures: The Italo-Celtic cluster is always identified, in accordance with some of the already mentioned traditional hypotheses, and it is followed by Slavic and Hi, Germanic and Celtic. The same distribution of groups is largely confirmed in the results obtained with the distance-based techniques on the Indo-European family alone, as illustrated in figure 9. Essentially the situation parallels that already described for the trees inferred on the whole sample of modern languages: The trees reconstructed under MCH overlap and the only important difference between these and the other two regards the alternative distribution of Celtic and Germanic. However, notice that in this series of outcomes the internal articulation of the Indo-European family is preserved both in presence and in absence of languages belonging to other families: This represents a further piece of

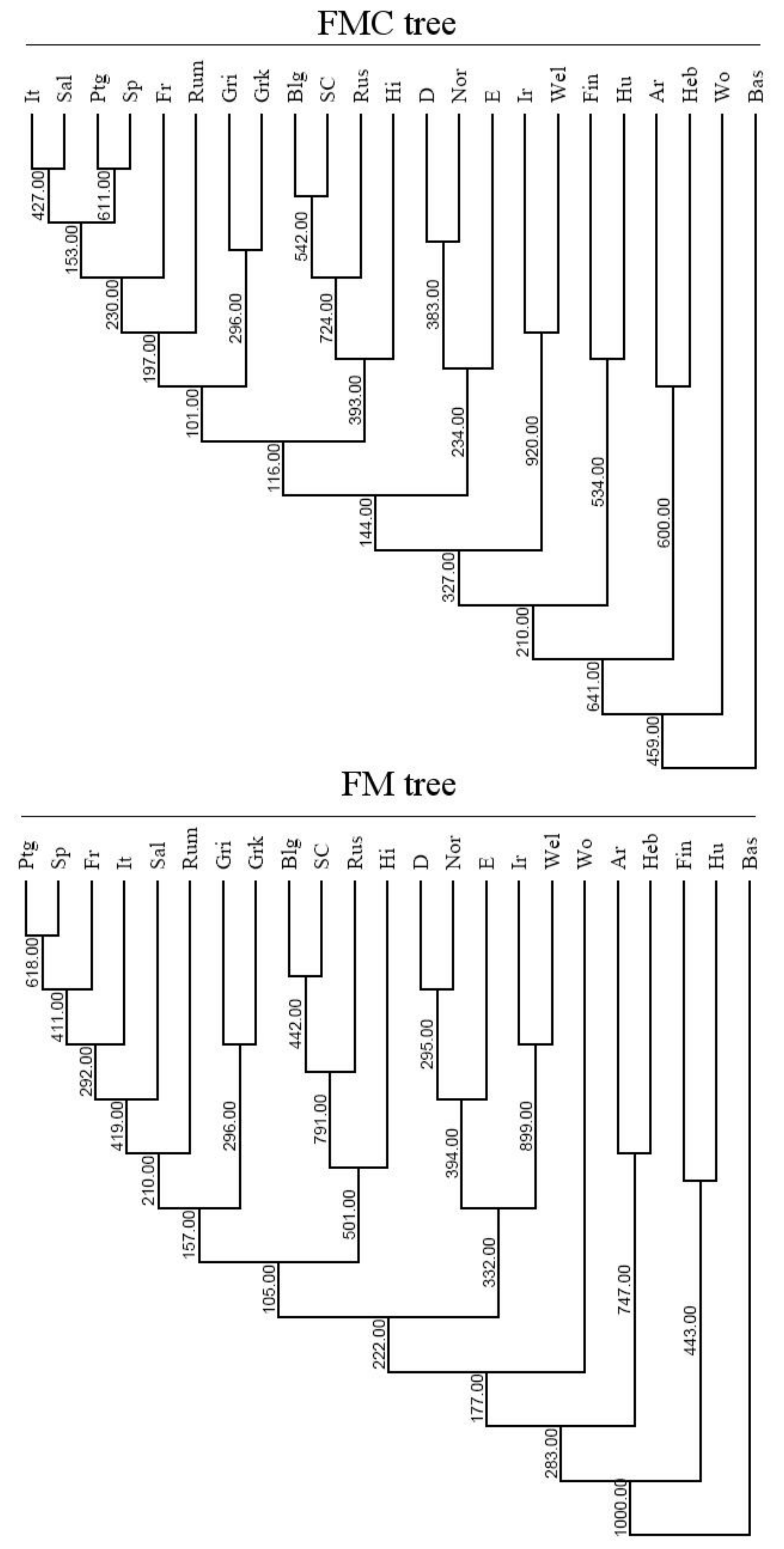
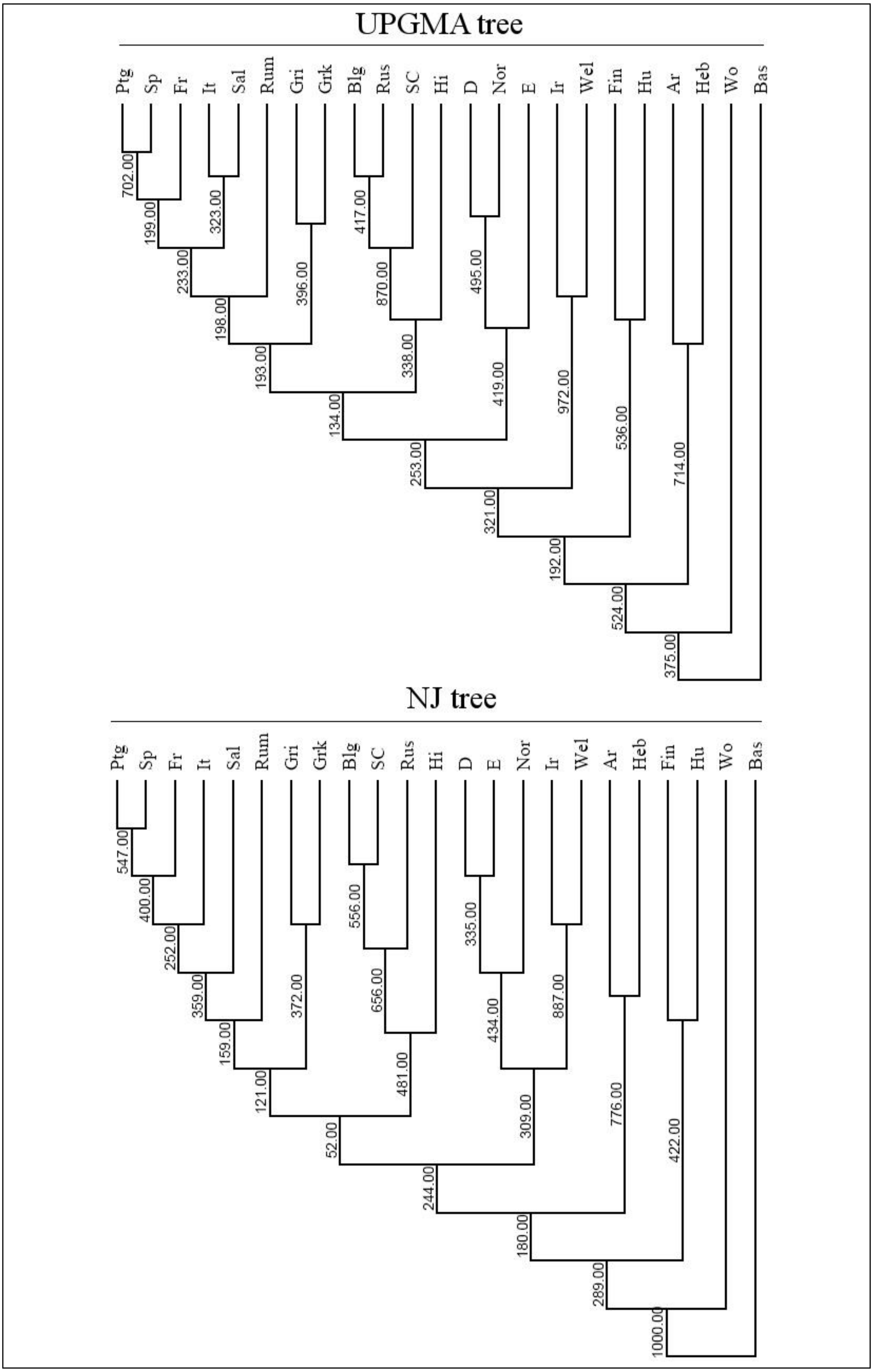


Figure 7: FMC and FM trees derived from the parameter selection, all modern languages.



UPGMA tree

NJ tree

Figure 8: UPGMA and NJ trees derived from the parameter selection, all modern languages.

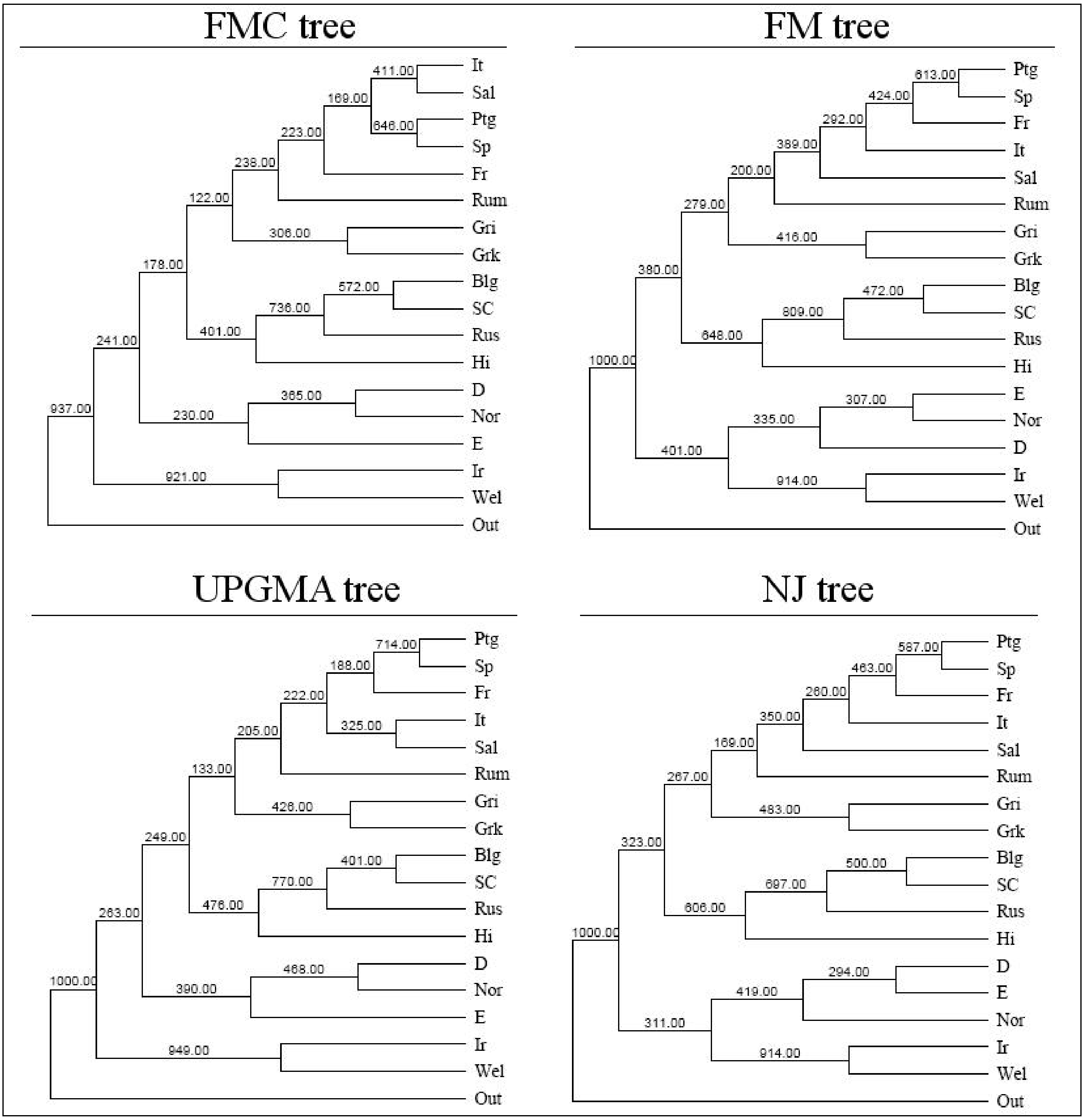


Figure 9: FMC, FM, UPGMA and NJ trees derived from parameter selection, only IE languages.

evidence in favour of the suggested solidity of the selection, with particular reference to the possibility of applying it to different samples of languages maintaining a high level of performance of different phylogenetic methods. The summary of the consensus scores of the whole series of trees is given in (9ab).

(9) a. All modern languages:
FMC tree: 392.4762
FM tree: 401.7
UPGMA tree: 419.2381
NJ tree: 391.55

b. Indo-European modern languages:
FMC tree: 418.5
FM tree: 458.6
UPGMA tree: 448.6875
NJ tree: 442.8667

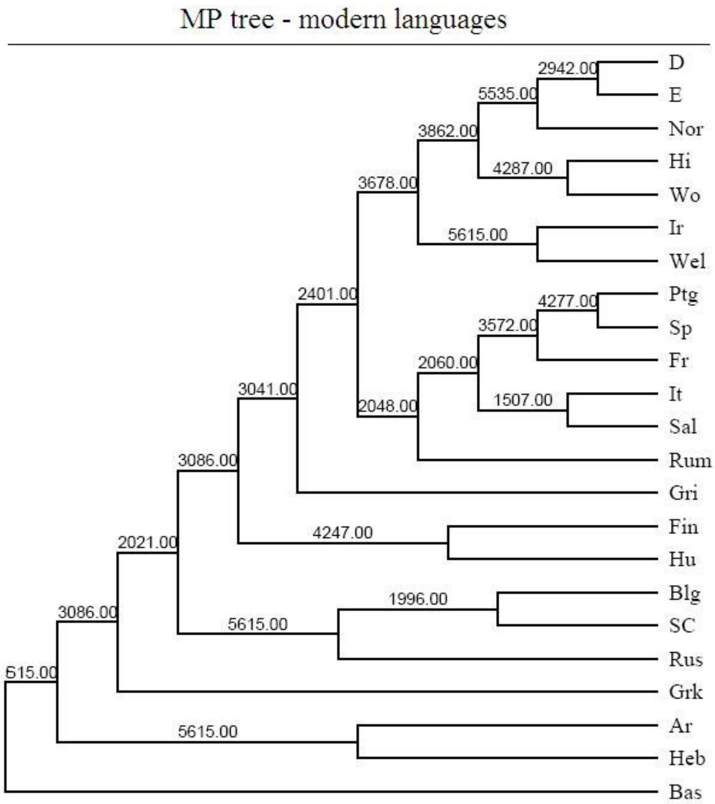
With respect to the values collected in (5) only a slight increase is recorded in the average consensus score of (almost) all the trees, particularly for the cases in which the FM and the NJ methods have been used.

The MP trees, instead, may be found in figure 10. While, for the same reasons worked out in section 6.1.3, the phylogeny comprising all the modern languages is still problematic, especially due to the fact that the MP procedure does not correctly infer long-range relationships, the tree of the Indo-European family is satisfactory in this case, even if it entails a different distribution of the groups with respect to that reconstructed with the distance-based methods: The Greco-Celtic unit has not been recognized in other experiments so far, as well as the closeness of Germanic and Indo-Slavic, whose cluster appears as outlier of the family. This hypothesis differs considerably from that provided by the distance-based techniques, which tend to indicate Celtic. In any case this result appears to be very significant, also because it is obtained on the basis of only 15 parsimony-informative parameters (namely *p22*, *p34*, *p36*, *p38*, *p39*, *p41*, *p43*, *p45*, *p47*, *p50*, *p52*, *p56*, *p57*, *p59*, *p63*), which served to classify 17 languages.

In sum, the selection-oriented experiments appear to corroborate the idea that a first adequate subset of genealogically relevant parameters should primarily include the data regarding the status of features Person, Number and Gender, adjectives, N-raising and possessives, while parameters connected in particular to definiteness, Countability and genitives should be largely left out. Nonetheless, to what extent is the identification of such a subset dependent on the specific sample of languages under study? This is a difficult question, that can be properly answered only within a statistical investigation of the data, which, as claimed in the beginning of this chapter, is beyond the objectives of the present work. However, here the selection in question has been shown to work effectively on two diverse samples of languages, applying five different phylogenetic methods and obtaining 8 convergent phylogenies plus one that is acceptable in any case. It is reasonable to claim that most homoplasy contained in the full dataset has been filtered, and that the outcome is a group of "more genealogically representative" parameters, i.e. of parameters whose evolutionary histories include, on the whole, a relatively small amount of areal convergence and of parallel evolution and therefore tend to be "purely genealogical" and "genealogically typical" of distinct lineages. This distinction is reconsidered in section 6.3.1. For the moment, in the next section a quantitative criterion to "select" characters is illustrated, together with other procedures aimed at "weighting" comparative characters, that is, ultimately, at measuring homoplasy.

6.2.2. Weights

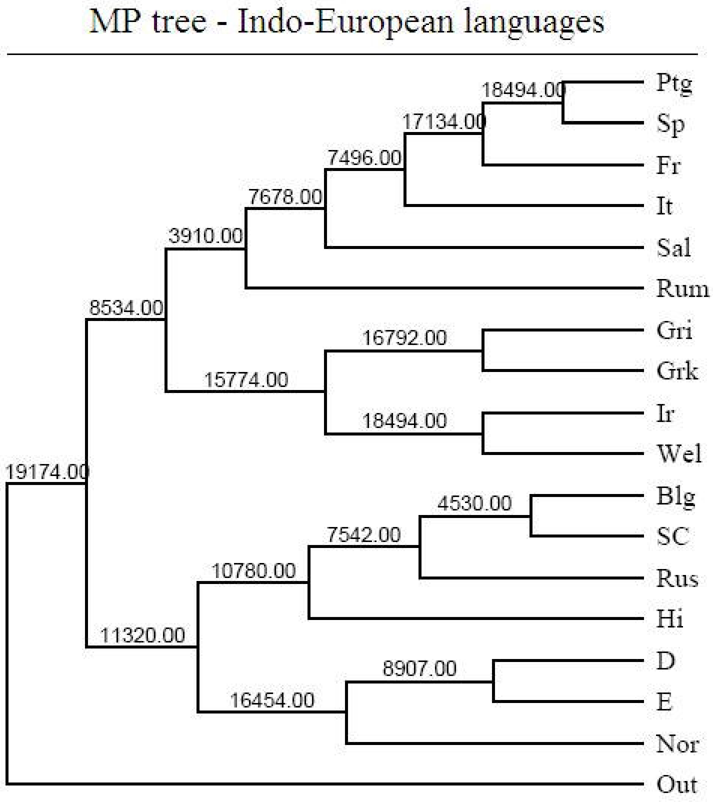
To determine which parameters are included in a certain selection and which ones are excluded corresponds to assigning to each of them a discrete binary weight, which is 1 if the parameter is in and 0 if it is out. This is simply another way to encode the choice that has been made, in that 0-weight parameters are assumed to be irrelevant in the analysis carried out to infer the phylogenetic tree(s): This encode is used for instance in table 8 to summarize the results of the empirical-based selection, which are provided in the column "selection". The following columns ("clique", "CI" and "RI"), instead, collect the values of other weights



Number of parsimony-informative characters: 28

Score of best tree(s): 61

Trees retained = 5615



Number of parsimony-informative characters: 15

Score of best tree(s): 30

Trees retained = 19174

Figure 10: MP trees derived from the parameter selection, modern and IE languages.

assigned to parameters according to the criteria explained throughout this section. It is worth pointing out that in general the aim of a weight, as it is intended here, should be to increase the comparative relevance of parameters that are assumed to have undergone a typical and regular genealogical development and to decrease the importance of those, whose history is suspected to include homoplasy: If binary 01 weights are assigned, the former will be regularly considered and the latter will be simply ignored.

A common procedure used in quantitative phylogenetics that allows determining the 01-weighting of a set of characters is grounded on the concept of clique, that has been introduced in section 4.3.1. Consider, for instance, that a discrete binary weighting of this type is indirectly derived by Ringe *et alii* (2002) and Nakhleh *et alii* (2005a), who intend to recover the tree, with respect to which, the majority of the characters are simultaneously compatible: In synthesis, the scholars aim at finding the phylogeny that is supported by the largest clique of characters. When they state that 18 characters are homoplastic with respect to the tree they have inferred, they implicitly claim that these characters have no genealogical value in their sample, i.e. may be assigned a genealogical weight 0, while the other characters, that are included in the clique, may be assigned a regular genealogical weight 1. Observe that, as the characters included in any given selection, those covered by a clique are assumed to be purely genealogical and genealogically typical. Thus what hypotheses does the clique-based analysis entail about the underlying evolutionary process that produced the dataset? Basically, it involves assuming that the greatest amount of variation is due precisely to purely genealogical and genealogically typical evolution. Moreover, since the clique-based analysis derives from a compatibility evaluation of the character set, which in turn is grounded on the parsimony-based examination of the character-state histories, as a matter of fact the clique-based analysis also incorporates the hypothesis that evolution occurs in a parsimonious fashion. If these conditions do not apply to the actual dynamics of the process from which the data originated, the phylogeny inferred on the basis of the clique does not reflect the actual tree.

What is relevant here is primarily deriving the largest clique of compatible parameters from Table A (in the version comprising only the modern languages) and then to compare it with the selection designed in the previous section. The procedure used here to obtain the clique is grounded on the notions provided in section 4.3.1 and in particular Estabrook *et alii*'s (1976a;1976b) theorems. The first step is to draw a compatibility matrix in which the pair-wise compatibility of each parameter to the others is specified: As shown in the sample provided in table 9, the value "1" indicates that two characters are compatible and "0" the opposite state. The character matrix is then elaborated to obtain the clique in (10), that includes 36 characters.

(10) Clique (36): {*p1*, *p2*, *p5*, *p6*, *p8*, *p9*, *p10*, *p11*, *p13*, *p16*, *p19*, *p20*, *p22*, *p24*, *p26*, *p30*, *p32*, *p35*, *p36*, *p37*, *p40*, *p41*, *p43*, *p44*, *p46*, *p47*, *p48*, *p49*, *p51*, *p52*, *p54*, *p58*, *p59*, *p60*, *p61*, *p63*}

The consensus tree derived from this subset of parameters by means of the FMC method, which until now has provided the best results together with UPGMA, is presented in figure 11. This phylogeny, whose average consensus score is particularly low and amounts to 225.619, is evidently wrong in many parts that do not require discussion now. Indeed this experiment is only intended to show how the selection / 01-weighting outlined relying on the clique-based analysis is substantially inadequate. Such a result may be in part a consequence of the fact that this procedure is grounded on a parsimonious character-based analysis, but perhaps even more it derives from the inapplicability of the strict maximum compatibility approach to Table A. This point will become clearer after the introduction of two further weights in the course of the evaluation of the whole picture drawn in section 6.2.

For the moment, one may note that the clique-based discrete weight is problematic. Not all clique-external parameters are equally homoplastic respect to the phylogenetic tree identified by the clique, because the projection of their history onto the tree may require the postulation of just one homoplastic event or of many different events. In other words, the history of a parameter may be perfectly genealogical for all languages except for a single contact event

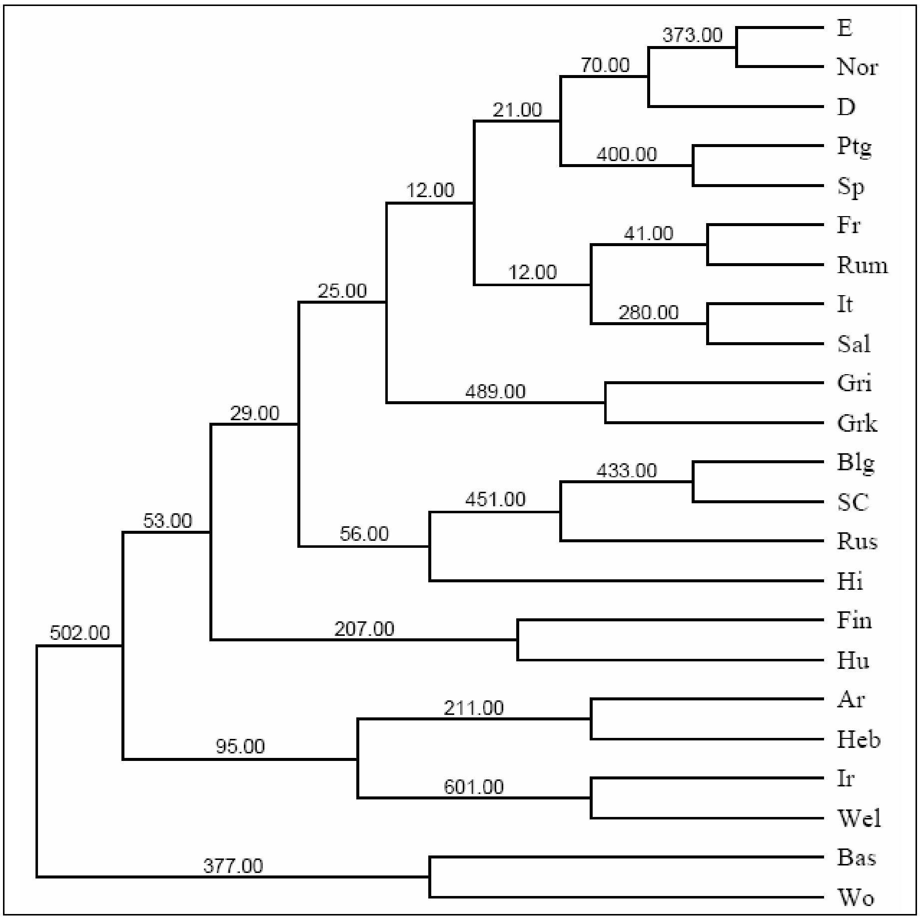


Figure 11: Consensus tree derived from the clique-based analysis, inferred with FMC.

occurring between two lineages, which causes the homoplasy and the exclusion from the clique, while the history of another parameter may require the postulation of many contacts between different lineages: A discrete binary "compatibility weight" would not be able to distinguish between these two situations, i.e. the fact that the former parameter appears to be subject less to contact-induced change than the latter and consequently it appears to be more genealogically informative. Thus in order to account for these differences one should rely on a continuous measure of weight instead of a discrete one. Actually such measures are available and here two of them are put to the test: The consistency index (CI, Kluge and Farris 1969) and the retention index (RI, Farris 1989). CI and RI are calculated starting from a specific phylogenetic tree under the hypothesis that the evolutionary process occurred in the most parsimonious way. The relevant variables for the calculation of CI and RI are listed in (11a-c).

(11) a. Minimum number of changes (*m*): Is given by the minimum possible number of changes of a parameter (i.e. 1, being parameters binary characters);
b. Maximum number of changes (*M*): Corresponds to the number of changes counted in the least parsimonious scenario, i.e. to the number of *taxa* included in the smallest set defined by the split entailed by the relevant character;
c. Most parsimonious number of observed changes (*Obs*).

The consistency index and the retention index of a character *i* are calculated according to the formulas in (12ab):

(12) a. $$\begin{cases} Obs_i = 0 & CI_i = 1 \\ Obs_i \neq 0 & CI_i = \dfrac{m_i}{Obs_i} \end{cases}$$

b. $$\begin{cases} Obs_i = 0 & RI_i = 1 \\ Obs_i \neq 0 & RI_i = \dfrac{M_i - Obs_i}{M_i - m_i} \end{cases}$$

The formula in (12a) means that the value of CI is 1 any time the maximum parsimony score of the character is 0 (i.e. the character is constant), otherwise it is calculated as in the second line. (12a) has to be read along the same lines. Both CI and RI represent possible measures of the genealogical relevance of a certain character on a given phylogeny, because they encode homoplasy as the proportion of changes that exceed the minimum. These weights have been calculated on the most consistent tree found in the outfit of results derived from the full dataset, i.e. the MCP one (figure 3), and they have been applied to Table A, focusing on whole sample of the modern languages, in order to obtain weighted measures of distance. These have been collected in the relevant matrices and given in input to the FMC and the FM methods, i.e. the procedures that in the basic trees (figures 3 and 4) have produced the best and the worse phylogenies (in any case respectively paralleling UPGMA and, partially, NJ), in order to test whether CI and RI increase or decrease the precision of the opposite inferences. The results are provided in figures 12 and 13: Note that both with CI and RI the accuracy of the FMC phylogenies is slightly compromised, because in the former case Sal is classified as outlier of Romance instead of Rum, while in the latter the correct articulation of Germanic is lost; The FM trees, instead, are significantly improved, and no more problems arise with the interference of Semitic with Indo-European-internal groups. Furthermore, in both the trees obtained using RI a novel cluster appears, which groups together Semitic and Uralic: This inference cannot be regarded as wrong *a priori*, because there are no clear established hypotheses by means of which to evaluate it, but certainly its uniqueness in the whole set of results makes it at least dubious. It is also important to observe that the average consensus scores of these last four phylogenies, illustrated in (13ab), have to some extent augmented, especially for what concerns the FM phylogenies.

(13) a. CI:
FMC tree: 444.8095
FM tree: 494.9
b. RI:
FMC tree: 467.8095
FM tree: 501.65

The distribution of the consensus scores on the forks of the trees is in line with that recorded in the previous reconstructions, entailing higher support for independently known groups and lower reliability for the relationships between them. Attempts have been made also to use CI and RI to infer better MP trees, but the results remain unsatisfactory, therefore they do not require to be reported here.

The experiments above are intended to demonstrate that CI and RI may represent valid indicators of homoplasy. Their evident limitation is that they can be obtained only on the basis of an already-known phylogeny, as a consequence they are not useful to properly infer genealogical classifications, but they can provide valuable suggestions to map homoplasy in the dataset. In this sense, it is interesting to compare the global average value of CI and RI

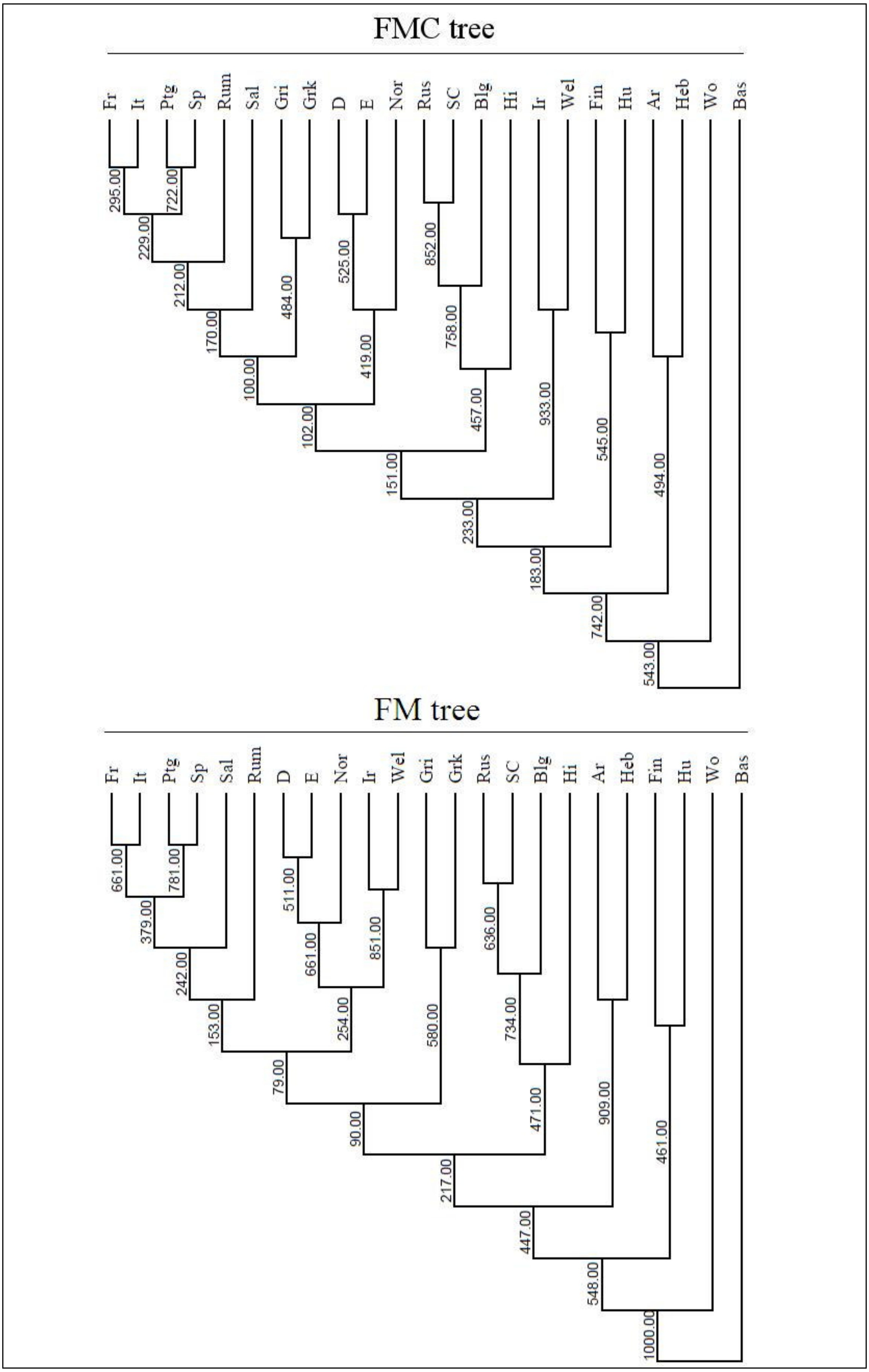


Figure 12: FMC and FM trees derived applying CI on full Table A, all modern languages.

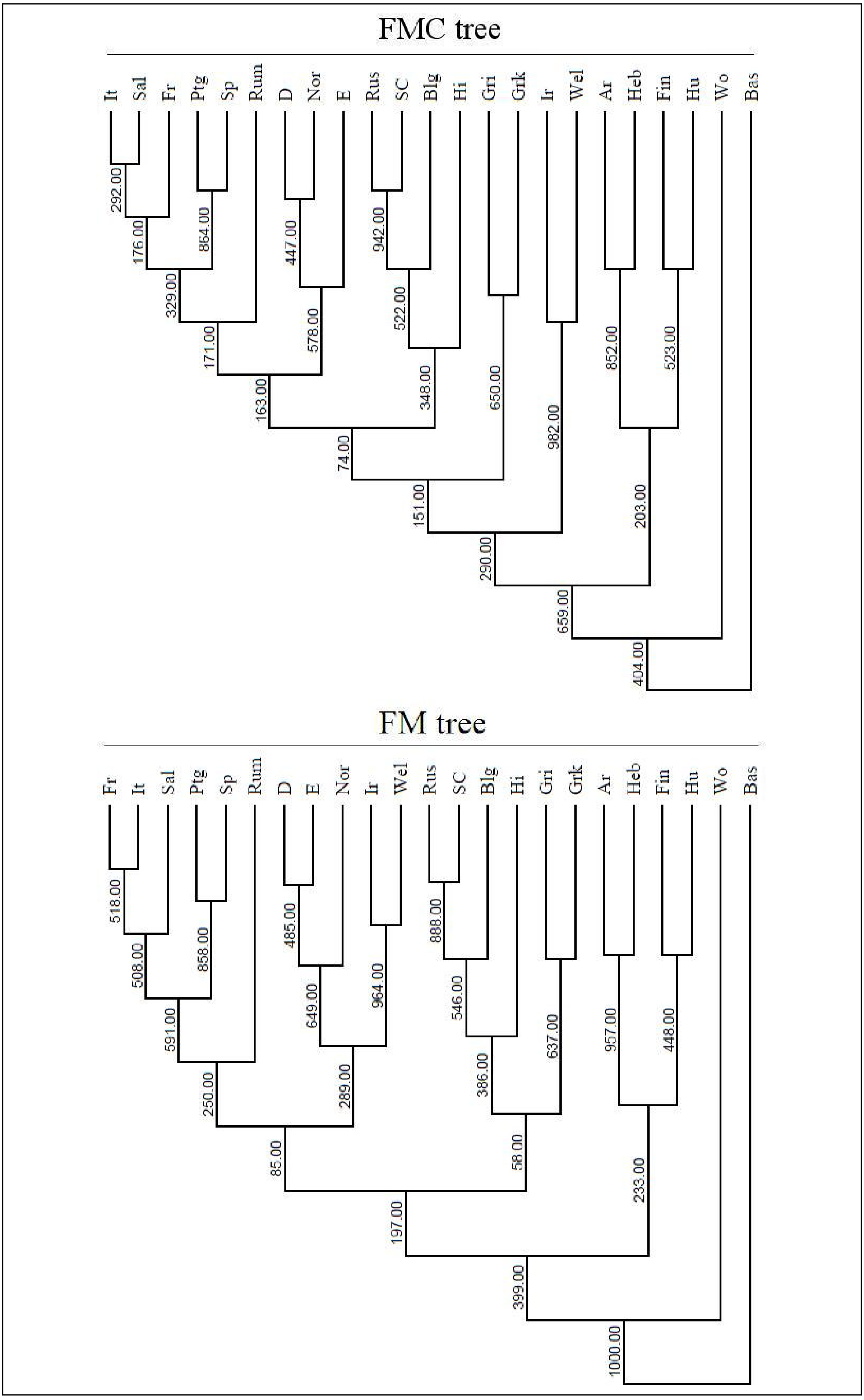


Figure 13: FMC and FM trees derived applying RI on full Table A, all modern languages.

with those recorded in the different sub-modules of Table A and in the empirically-based selection, as illustrated in table 10: Observe that actually the quantity of homoplasy recorded by CI and RI is lower in the set of selected parameters than in the whole sample and that, with the only exception of the sub-module concerning definiteness, the CI and RI scores are higher in the sections of Table A that have been regarded as genealogically more relevant designing the empirically-based selection (i.e. those regarding the status of features Person, Number and Gender, adjectives, N-raising and possessives). This is not surprising, since part of the evidence on the basis of which the selection has been outlined consists precisely of the direct evaluation of the splits determined by each parameter to find the most genealogically relevant ones. What is important to notice is that the CI/RI estimates allow quantifying the amount of homoplasy that has been eliminated and that entailed by the different subsets of parameters. Nonetheless, on the one hand CI and RI do not provide accurate insights as those attained selecting parameters in light of independent knowledge, on the other hand they do not permit distinguishing between the different causes of homoplasy. The discussion introduced in the next section is intended to propose an evaluation procedure that might help to pursue this goal.

6.3. Detecting borrowing and beyond

The following discussion includes more tentative insights and cues for future research than consistent results: Indeed it should be regarded as primarily programmatic, because it deals with issues that appear to be hardly solvable in general, and, in particular, to go beyond the scope of the present inquiry. The first theme that is taken into account concerns the definition of a procedure that may allow distinguishing between the diverse sources of homoplasy within the quantitative perspective. The second argument, instead, is connected to the problems of reconstructing ancestral states and of providing a chronological account of the genealogical classifications; Moreover, in section 6.3.2 the question of the correlation between syntactic and geographic distances is considered.

6.3.1. Detecting borrowing

The procedures to detect contact reviewed in section 4.3.3 need to be reconsidered here in light of the quantitative analyses of the parametric data that have been carried out so far: The objective is to evaluate which technique(s) appear(s) most apt to work out the parametric data. The summary of the available procedures is provided in table 11.

Sankoff (1972) and Embleton's (1981, 1986) idea of modelling borrowing following the paths of geographic neighbourhood between languages might be adequate, to a certain extent, dealing with lexical data, as shown for instance by Cavalli-Sforza and Wang (1986), but it appears to be too superficial to account for the structural characterization of syntactic evolution, as has been outlined in chapter 3; Moreover, the analysis of the correlation between syntactic and geographic distances presented in section 6.3 disfavours the adoption of this procedure. Another method is generally bnot valid, because its hypothesis has been proved to be false: The "negative branch length" test (Minett and Wang 2003). Two others are yet to be verified, thus they cannot be applied to the present work as well: These are the procedures based on the error matrix (Wang 1989) and on the parsimony analysis (Minett and Wang 2003). As for the "stratification technique" (Hübschmann 1875; Wen 1940; Sagart and Xu 2001) and the "distillation procedure" (Wang 2004), they require dating the different *strata* of features, which is a type of evidence not available for the parametric dataset. A similar problem arises with the "skewing test" proposed by Heine (1974), Hinnebusch (1999) and Wang and Minett (2005): Again, this method is based on statistical data, which are not accessible for parametric syntax. The "split decomposition" technique (Bandelt and Dress

1992), together with its various implementations (Bandelt *et alii* 1995; Huson 1998; Bryant and Moulton 2004), is a purely quantitative procedure that does not need to rely on external information, but it cannot help distinguishing between borrowing, back-mutation and parallel evolution: The graphs resulting from this method are usually rich in connecting branches, which are postulated on the basis of any of the three possible sources of character/split incompatibility. Of course, if the characters of a dataset are prone to show, for instance, parallel evolution, this will result in a number of wrong connecting edges throughout the graph. So the quantitative blindness of the split decomposition represents a decisive restriction to their applicability.

Cavalli-Sforza *et alii*'s (1997) bootstrap test, instead, has already been widely applied to Table A within this experimentation. Perhaps it is true that low consensus scores might suggest the occurrence of secondary convergences in the relevant lineages, but it may also be a signal of extensive parallel development of some traits. In sum, to use exclusively bootstrap to shed light on specific events in the history of languages appears to be insufficient as to adopt Bandelt and Dress' (1992) method alone or, probably, any other purely quantitative procedure of investigation. Another problem with Cavalli-Sforza *et alii*'s (1997) bootstrap test is related to the difficulty in establishing an objective threshold of the consensus scores below which one should start postulating the occurrence of non-genealogical convergence between lineages.

On the whole, Nakhleh *et alii*'s (2005a) compatibility test appears to be the most reasonable solution that one can adopt, because it allows clearly isolating all homoplastic events and, what is more relevant, evaluating them relying on independent knowledge, such as that introduced below. Nevertheless, to what extent does the maximum compatibility test conceptually apply to a linguistic inquiry?

Nakhleh *et alii's* (2005a) point is that character incompatibility is only marginally an effect of backmutation and parallel evolution, but rather it is mostly an indication of borrowing, at least for what concerns their dataset. The scholars argue that backmutation has a very limited role in linguistic diachronic processes and this "is partly the result of known properties of linguistic systems and language change and partly the result of probabilistic factors" (Nakhleh *et alii* 2005a: 2). It has already been observed in section 5.1.2 that the scholars carefully consider that phonological and morphological characters have complementary properties, the former entailing strong cues of directionality but also (relatively) significant probabilities of parallel development, the latter being individual-identifying but lacking directionality, while lexical characters should be regarded as the poorest source of genealogical information. With these expectations in mind, Nakhleh *et alii's* (2005a) perform an accurate preliminary selection of characters, so that if the quantitative analysis still detects problematic events, it is likely that they are contact effected.

What about syntactic characters? In general, assuming syntax to be inertial in Keenan (1994; 2002) an Longobardi's (2001a) sense, one should expect backmutation to be extremely marginal or not to exist at all in this domain, because the factors that cause a syntactic change, such as phonological erosion and semantic shift, are indirect and complex, i.e. they presumably act slowly and idiosyncratically. Parallel evolution, instead, should be regarded more cautiously, not only due to the awareness that typical patterns of syntactic change exist (see in particular section 3.2.1): Also in syntax each case of suspected parallel development needs to be independently evaluated, because this event cannot be excluded *a priori*. Thus a blind maximum compatibility approach is unsatisfactory, precisely as shown by Nakhleh *et alii*'s (2005a) analysis, and qualitative evaluations have to be adopted in each suspicious case in order to distinguish between parallel evolution on the one side and horizontal transmission on the other. As for the parametric dataset, this caution appears to be particularly relevant, because the possible states of the characters are just two by definition: Whenever a change occurs in one lineage, it can be exclusively either from + to – or from – to + (excluding the cases in which a 0-state is triggered or replaced by + or -), which means that in principle parallel development of the same parametric settings could be expected to be relatively widespread and to cover the vast majority of the cases of homoplasy. It is worth anticipating that the analysis of the correlation between syntactic and geographic distances provided in the next section appears to be in line with this belief.

Following Nakhleh *et alii* (2005a), it has been chosen to evaluate the compatibility of a set of parameters, in which a relevant amount of homoplasy had been already eliminated, i.e. that corresponding to the selection outlined in section 6.2.1. This sample has been tested on the most reliable and representative tree among those inferred in the previous discussion: The FMC phylogeny derived from the selection of parameters on the modern languages (figure 7). Table 12 presents the list of parameters that are compatible with this tree, i.e. those whose character-state history may be reconstructed on the tree without homoplasy, together with the split they determine in the set of languages. Take for instance *p5*: Its most parsimonious development can be projected onto the lineages postulating a – state in the root, that is preserved in Bas, in Wo and in the internal node from which Wo diverges; According to an MP reconstruction the state – is then believed to have changed into + just before the point of separation of Semitic, so that all the languages of the sample, except for Wo and Bas, currently exhibit the latter state. This possible scenario is illustrated in figure 14, together with one of those concerning the history of p47. The parameter-state distribution of p47 is incompatible with the tree, because at least two changes in different lineages (or a contact) are required to account for the observed distribution of the character-states: According to one of the scenarios that could be drawn in a MP perspective, the state + may be thought to represent an innovation independently occurred in Celtic and in Sal, while the rest of Romance might be assumed to have preserved an independently-acquired state -. On the whole 21 parameters out of 40 are compatible with the FMC phylogenetic tree, i.e. 52.5%, but notice that not all compatible parameters are also PS-informative. This is the case only for 9 of them, which corresponds to 42.85%, while 12 parameters are invisible for the compatibility analysis, since either they are constant (e.g. *p1*) or they are PS-uninformative (e.g. *p32*). Table 13 collects the 19 parameters (47.5%) that are not compatible with the FMC tree and their respective splits.

The procedure by which one could determine whether or not a certain incompatible parameter was involved in processes of contact is tentatively designed here as an algorithmic representation of the series of regular tests that historical linguists commonly perform. In essence, the procedure, whose first stage is synthesized in (14), covers three successive steps, in which, according to independent evidence, the geographic, chronological and structural plausibility of the contact-hypothesis is progressively estimated.

(14) Stage 1

Consider the whole list of incompatible parameters and their corresponding splits.

Step 1: Evaluate geographic plausibility of contact within the minority group;

→ If convergence is geographically plausible, then proceed to step 2, else lay aside the parameter for Stage 2.

Step 2: Evaluate chronological plausibility within the minority group;

→ If convergence is chronologically plausible, then proceed to step 3, else lay aside the parameter for Stage 2.

Step 2: Evaluate structural plausibility within the minority group;

→ If convergence is structurally plausible, then postulate horizontal transmission, else lay aside the parameter for Stage 2.

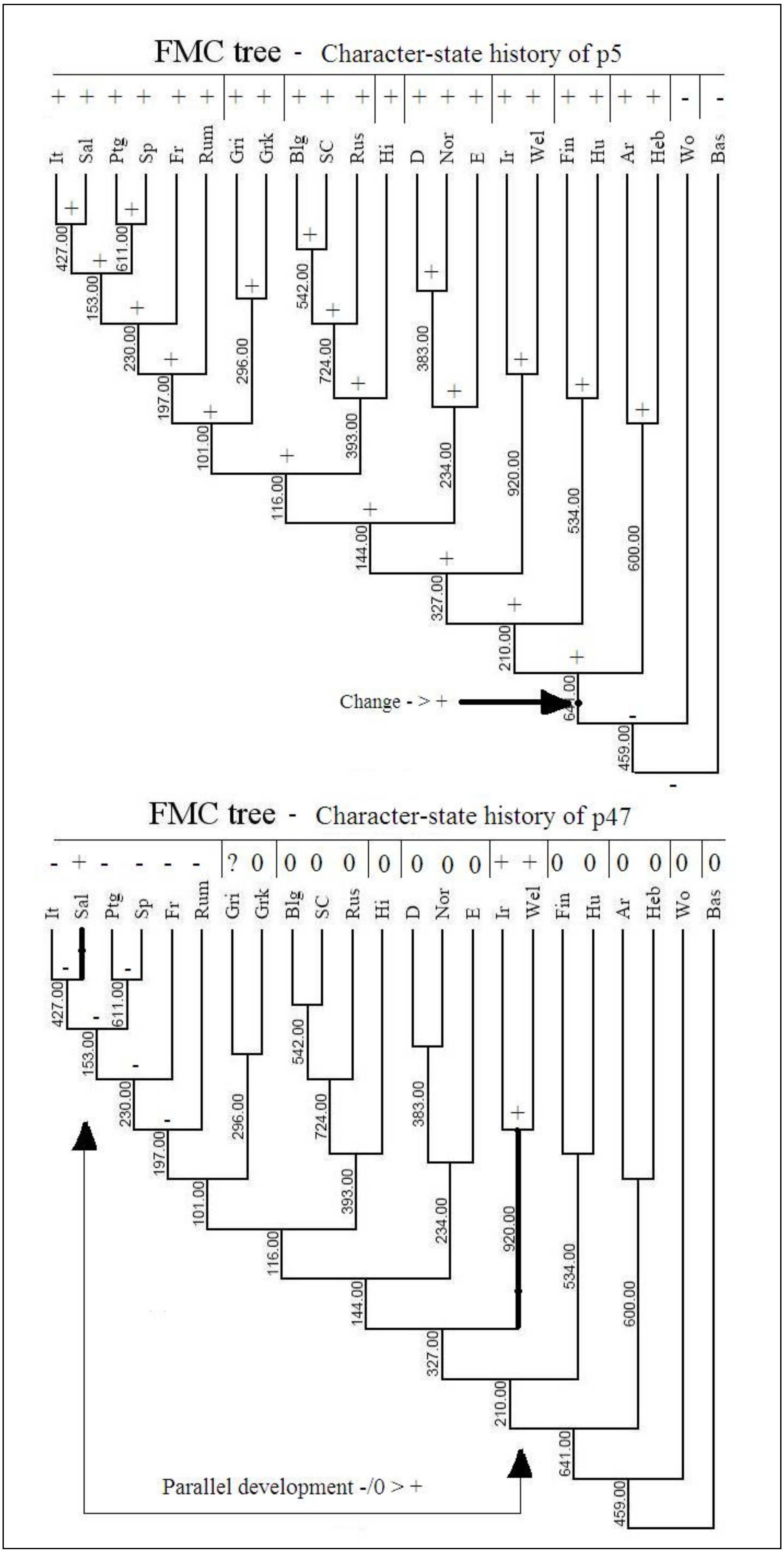


Figure 14: Parsimonious character-state history of p3 and p47, possible scenarios.

Thus for each incompatible parameter the first step of the evaluation procedure consists of verifying whether or not the languages included in the minority set of the split (which, according to the traditional majority rule, may be assume to be the innovative languages, that is the languages in whose lineage the change occurred) could be involved in relevant contact situations in their past: If the answer is "yes", it is geographically plausible that contact occurred, otherwise parallel development has to be postulated immediately. For instance, a parametric setting that groups together Ir and Wo, as in the case of *p62* (± *obl. def. inh.*), cannot be brought back to contact, since those languages have never coexisted in the same area. The second step regards the chronological plausibility of the convergence and entails knowing when certain traits started being attested in a phylogenetic branch, in order to compare the time of the contact with that of the possible borrowing: If, for instance, two languages already shared a certain trait before starting interaction, there are no reasons to postulate a contact edge between their lineages. The third stage of the procedure concerns the structural aspects that restrict the possibility of syntactic borrowing, in light of the fact that "what is replicated in language contact is constrained by a factor that may be called 'linguistic compatibility', that is, the relative degree of structural similarity among the languages involved" (Heine and Kuteva 2006: 140). To this end it is also opportune to recall Nakhleh *et alii*'s (2005) assumption, according to which "[…] borrowing into native dialects from languages or dialects that are not closely related is tightly constrained […]" (Nakhleh *et alii* 2005a: 4). It is reasonable to claim that languages that are genealogically distant and, as a consequence, structurally very different can hardly pass syntactic features to each other (on these argument recall the discussion about the possibility of syntactic borrowing in section 3.3.1). Thus, relaxing Nakhleh *et alii*'s (2005) criterion of "closeness", it might be assumed that at least among languages that belong to different families syntactic interference is highly improbable, even if not impossible and worth considering. Stage 2 (as the successive ones) is designed according to this idea and begins eliminating from the analysis the language that represents the outliers of the set, e.g. Bas. In this way a parallel development is automatically postulated in the splits containing Bas, which are all reduced or, in some cases, solved: For example, without Bas the split determined by *p3* (± *gramm. gender*), i.e. *s(p3) = {(Bas, Hu, Fin), (others)}*, becomes compatible with the FMC phylogeny and no homoplasy occurs.

What is relevant here is not to apply the procedure throughout the entire sample of parameters in question, which would be unfeasible at the present stage of advancement of the research project due to the lack of fundamental evidence regarding, in particular, chronological and structural plausibility estimates[16]. It is important to point out, instead, that the goal of shedding light on the diverse dynamics of language evolution appears to be hardly achievable relying exclusively on quantitative analyses, but requires one to be provided with extensive knowledge of the historical and qualitative characterization of the data.

6.3.2. Reconstruction and correlation with time and space

The observation pointed out in the end of the last section applies to the issues regarding the reconstruction of ancestral parameter states as well. In any case, in the following discussion it is shown how the quantitative analysis focused on ancient stages of language development permits obtaining interesting cues.

The parsimony-based analysis has been used to map the ancestral-parameter state distribution on the MP tree derived from the empirically-based selection and including only the Indo-European languages (figure 10). This phylogeny comprises 16 internal nodes, each one corresponding to an ideal proto-language and it is reported in figure 15 provided with labelled

---

[16] However, an early and tentative analysis of this type carried out on an old version of the data is provided in Rigon (2007).

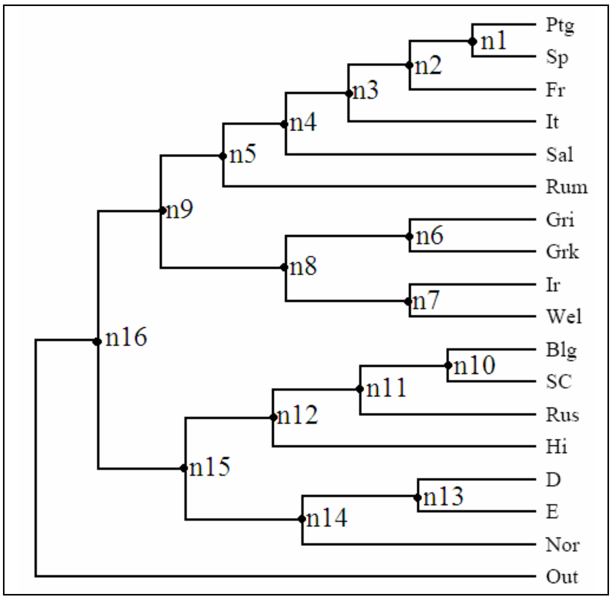


Figure 15: MP tree for Indo-European derived from the selection, labelled internal nodes.

internal nodes. For each of these "proto-languages" the values of the 40 parameters included in the selection have been tentatively set in accordance with the most parsimonious scenarios describing the evolutionary history of the languages in question. The reconstruction has been automatically carried out by the same program used for MP inference so far (i.e. PAUP) and has entailed taking into account the successive implicational ranking levels of parameters one after the other (see section 6.1.1). This means that first the MP reconstruction of level-1 parameter-states has been performed, i.e. the hypothetical MP states of all level-1 parameters have been mapped onto the internal nodes of the tree, and 0s have been assigned to level-2 parameters, in the cases in which the relevant implication were not satisfied[17]. Then the MP reconstruction of the level-2 parameter states has been carried out for the proto-languages to which a basic parameter state could still be assigned, and so on.

The resulting character-matrix containing the reconstructed parameter states for the 16 internal nodes is introduced in table 14, which includes also the states exhibited by the ancient languages of Table A (Lat, ClG, NTG, Got, OE). In order to evaluate the reliability of the reconstruction, using the formula in (2) the pair-wise distances between nodes and ancient languages have been calculated. Table 15 collects all the 90 pair-wise distances in which at least one member of the pair is a language in increasing order (node – node distances are not reported here) and the rows corresponding to the lowest five values of each ancient variety are highlighted in different degrees of grey-black. It is remarkable that the five nodes that are closer to Lat are precisely those of Romance, i.e. *n1*, *n2*, *n3*, *n4*, *n5*, which suggests that the reconstruction of the ancestral states in the Romance area may be considered appropriate. OE is also quite clearly attracted by Germanic: Indeed both *n13*, including (D, E) and *n14* (D, E, Nor) are listed among lowest values of distance of OE, as well as *n15*, from which Germanic and the "*satəm* group" originate. The latter node is also the point of the graph that is closest to Got, which is an acceptable inference in light of the fact that the East branch of Germanic might precisely be projected in the lineage (14, 15). Nonetheless, the other relevant distances of Got and those of ClG and NTG do not indicate a precise position for these varieties in the phylogeny and they are rather far from satisfying the expectations.

---

[17] As for the parameters included in the selection that are implied by states of parameters that are not comprised in the sample, the reconstruction of the ancestral states of the external parameters in question has been performed as well in a separate stage of the experiment in order to properly assign 0s.

A possible explanation for this fact may be provided by observing how the internal nodes and the ancient varieties are mapped onto the same graph. The experiment has been carried out using NJ and adopting the bootstrap perturbation test. The resulting tree in figure 16, even if conventionally rooted, should not be regarded as if it contained genealogical information, but only as a structure in which the units are mapped according to their similarity. It is important to observe that the position of Lat and OE is confirmed, in that they are grouped with, respectively, Romance- and Germanic-internal nodes. Got, ClG and NTG, instead, form a compact cluster that lies in the neighbourhood of the Romance area. There are at least two possible rationalizations for this picture. On the one hand it is plausible that the reconstruction of the ancestral states is likely to become less reliable going deeper in the past, which would be concordant with the fact that among the more ancient varieties only the position of Lat is adequately indicated. On the other hand, this situation seems to support the idea that proto-languages may tend to appear genealogically closer between each other than they are with their respective descendents, as claimed in section 6.1.2. The latter hypothesis is supported also by another quantitative analysis that has been performed. Recall from section 6.1.2 that nothing prevents using character-based methods to classify non-coeval varieties, because the limitations that constrain distance-based techniques do not apply to these procedures. Since the MP tree of the modern Indo-European languages derived on the basis of the empirically-based selection of parameters is the only suitable character-based phylogeny obtained so far, the same experiment has been repeated including the five ancient languages obtaining the tree offered in figure 17. Apart from the problematic position of Rum, whose classification appears to be biased, this outcome evidently shows that ancient languages tend to group against the modern ones. Moreover, their articulation is correctly recovered by the MP procedure.

Notice that the analyses presented so far may be thought to have a twofold objective, in that not only are they concerned with the issue of reconstruction, but may be also viewed as a general method to attain indications about the position of ancestral taxonomic units in phylogenetic trees, i.e. to deal with the “internalization problem” mentioned in section 6.1.2.

Having introduced ancient languages in the experimentation, it is interesting to consider them also in connection with the problem of the correlation of syntactic distances and separation times. For the reasons worked out in section 6.1.2, it is reasonable to expect that parametric distances, as they are derived at the present stage of the investigation, cannot be meaningfully related to a precise chronology, at least for what concerns the short period, and this is why estimates of rates of evolution have not been adopted here. In any case, for the moment a caution that can be used to increase the significance of the analysis is to plot syntactic distances only in relation with reliable chronological data, i.e. those concerning the separation time of descendents from attested ancestors, and to avoid using supposed divergence times of coeval languages, which would require accurate dates regarding the split of group-internal subgroups. This means that, for instance, the separation times regarding the pairs (Lat, It), (Lat, Sal), (Lat, Sp), (Lat, Fr,), (Lat, Ptg) and (Lat, Rum) are taken into account here, since the period in which the variety of Latin in question is attested is known, while it is less clear when the ideal divergence between, for instance, It and Sp, occurred, thus the divergence time of (It, Sp) is left out. Given this principle, only the 11 pairs listed in table 16 can be considered. In this sample the correlation between syntactic distances and separation times is 0.295537, ie. not significant, as revealed also by a brief look at figure 18, in which syntactic distances are plotted against separation times. Nonetheless, observing the rate of evolution of each lineage, the amount of evolution occurred from OE to E appears to be much greater than that recorded for the other pairs, which could follow from the peculiar contact-history of the Germanic variety. Thus excluding OE – E form the calculation, the correlation between syntactic distances and separation times increases to 0.575207, which in any case is still rather

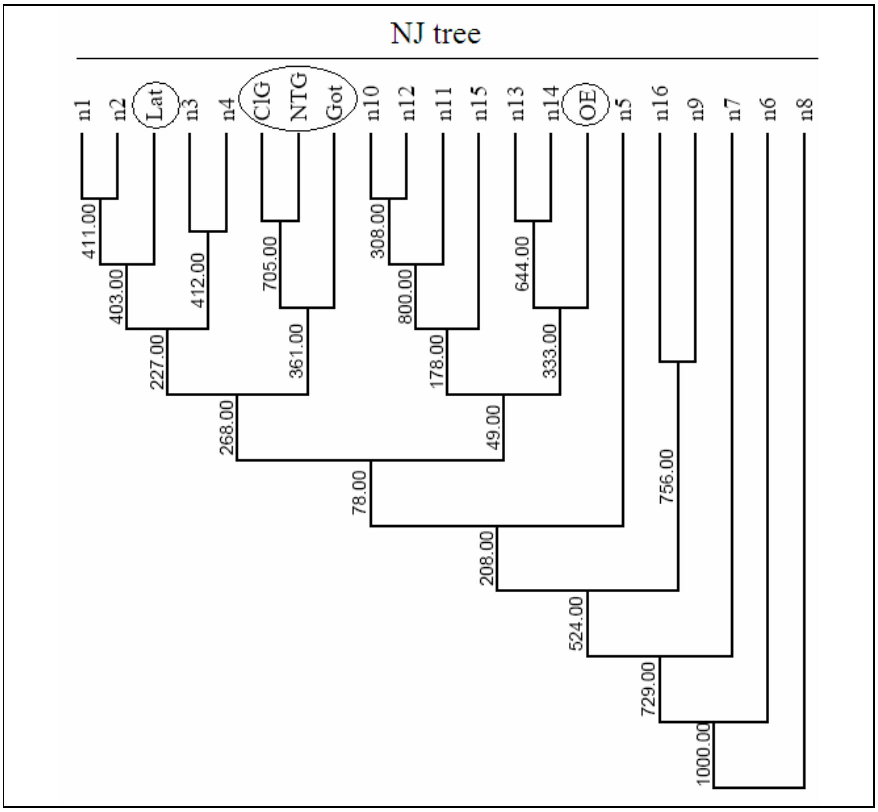


Figure 16: The internal nodes (MP tree in figure 15) and the ancient languages mapped onto a NJ unrooted tree.

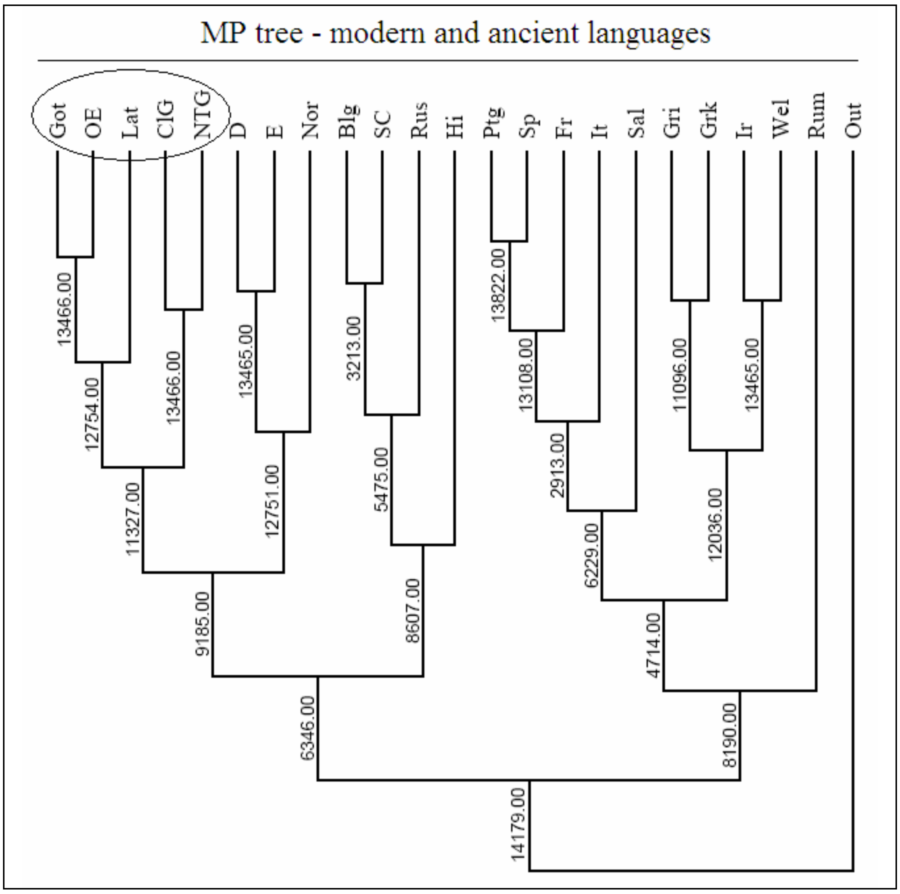


Figure 17: MP tree derived from the selection of parameters, modern and ancient languages.

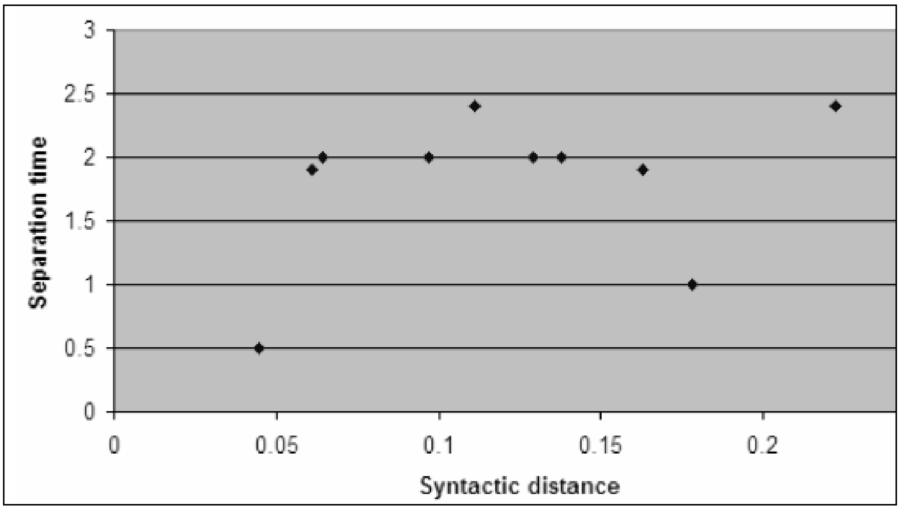


Figure 18: Syntactic distances plotted against separation times.

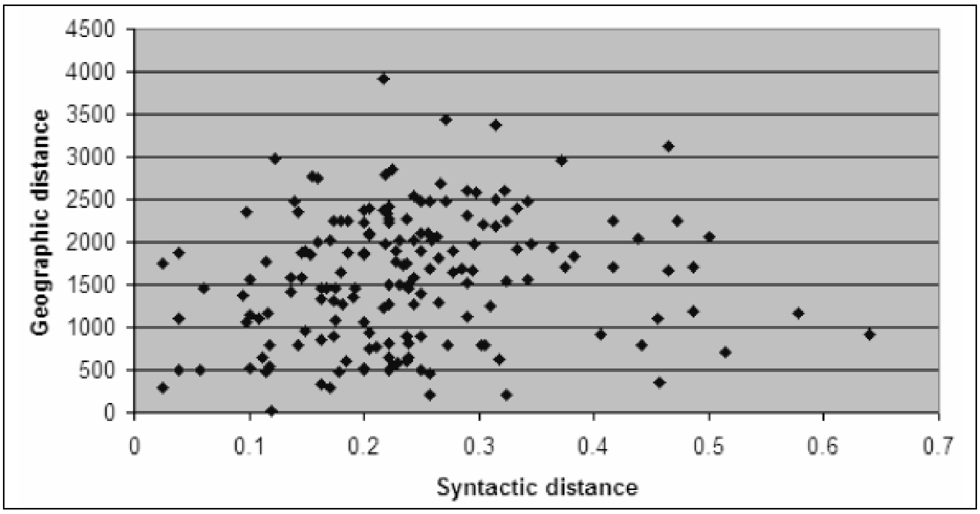


Figure 19: Syntactic distances plotted against geographic distances.

low. Nonetheless, even from these rough estimates it can be observed that there are no remarkable differences between the rates of change in the different lineages, which tends to be in line with Felsenstein's (2004a: 165-166) requirement about the assumption of MCH.
How do syntactic distances vary, instead, with respect to the geographic distribution of languages? Another cautious answer to this question can be provided comparing syntactic distances and geographic distances. This attempt has been made considering only the modern European languages of Table A, that are 19 including the non-Indo-European varieties (i.e. Bas, Hu and Fin). The first step of the analysis has been to define a list of representative locations (i.e. cities) for each language and to collect their geographic coordinates, as shown in table 17. These data have been elaborated to draw the matrix of geographic distances presented in table 18[18]. In figure 19 syntactic distances are plotted against geographic distances. The distribution of dots in the graph already suggests a lack of relatedness between the two measures, that is confirmed by the correlation value between the series, which corresponds to 0.13094. Bas and the Uralic varieties might be thought to bias the result due to the evident disproportion between their geographic location and their genealogical isolation in the European linguistic area, but even excluding them the correlation increases only to 0.316792, which remains unsignificant. On the whole, this analysis suggests

---

[18] Geographic distances have been derived with the Geographic distance matrix generator (Ersts 2008).

that the syntactic distances derived from Table A are relatively safe from the influence of areal convergence in general, which would corroborate the idea that the homoplasy observed in the parametric data is largely due to parallel development, i.e. chance, rather than by contact. It is opportune to underline that this observation applies to the parameter-based syntactic data in question, but not necessarily to syntactic data in general. For example, recall from section 5.1.2 that the syntactic distances obtained by Spruit (2005; 2008) are precisely shown to correlate with the geographic distances, i.e. the scholar demonstrates that the grammatical variation recorded in the relevant linguistic area parallels the continuous geographic distribution of the speech communities. It is reasonable to believe that such a result derives from the fact that Spruit's (2005; 2008) data are grounded on a dialectological perspective, which is focused on grammatical traits that tend to be closer to morphology, rather than to "narrow syntax". If correct, this idea would entail that grammatical "micro-variation", perhaps also that encoded in the form of micro-parameters, in Kayne's (2005) terms (see section 3.1.5), is more inclined to reflect patterns of areal relationships between languages, rather than carrying information regarding the vertical dimension of their evolution. As a consequence, it would be corroborated the hypothesis according to which traces of proper genealogical relatedness are more likely to be preserved, and should be more extensively searched, at the level of "macro-comparison", like that instantiated in the data of Table A.

6.4 Conclusion

The discussion covered in this chapter has focused on a set of quantitative analyses performed on Table A with the primary aim of testing its genealogical relevance. After having provided a character-based account for the data, a measure of syntactic dissimilarity has been proposed and used to derive the set of distances employed throughout the experimentation. On these grounds, some preliminary expectations regarding the distribution of variation in the sample have been verified and then the phylogenetic methods introduced in section 4.3.3 have been taken into account and their choice has been motivated in light of the characterisation of the data. The application of the quantitative techniques has followed and a first outfit of phylogenies has been illustrated: The trees in question have been derived on different samples of languages by means of both distance-based and character-based procedures and it has been shown that some methods, in particular those that entail an MCH-compatible evolution, perform better than others, that assume different hypotheses about the underlying process of development. Moreover, recalling that homoplasy may constitute a relevant source of bias, a selection of parameters, whose evolutionary histories may be thought to include less events of parallel evolution and areal convergence, has been designed on the basis of independent evaluation of the data and empirical considerations. This selection has been demonstrated to represent a genealogically consistent sample of parameters, as illustrated by the convergence of results of different phylogenetic techniques.

Alternative quantitative-based methods to reduce the effect of homoplasy on the phylogenetic inference have been put to the test: In particular, two types of character-weights have been taken into account, both discrete (clique-based) and continuous (RI and CI). None of them permit achieving the same accuracy of the empirically-based selection in the identification of homoplasy, but RI and CI can provide a useful account of its distribution on the data. The evaluation of homoplasy and the possibility of distinguishing between its diverse forms (i.e. chance and borrowing) have been the topic of the successive discussion. In essence, it has been argued that for the moment no purely quantitative technique appears adequate to aid in distinguishing between parallel evolution and contact, so that the plausibility of the latter on a given phylogeny can be estimated relying only on exhaustive external evidence. Thus the investigation has focused on the tentative reconstruction of ancestral parameter states and on the internalization of ancient languages in the phylogenies, observing that, as expected, proto-languages tend to appear genealogically closer to each other than they are with their respective descendents, but that the reconstruction procedure adopted allows obtaining some significant insights on the position of ancient taxonomic units in the trees as well. Finally, it has been shown that the mapping of syntactic distances on a precise chronological scale is probably early, but already worth considering; On the other hand projecting the same distances on the diatopic dimension offers no significant correlation between them and geographic distances, which is in agreement with the belief that the parametric data primarily encodes genealogical, and non areal, variation.

Further comments on the contents of this chapter are covered in the general conclusions of the dissertation.

7. Conclusion

In this dissertation a quantitative approach to the study of syntactic evolution has been adopted with the aim of illustrating its applicability, its effectiveness and its limits. A first step towards the achievement of this objective has been the comparison between language evolution, intended as diachronic development of languages, and biological evolution. On the one hand this parallel has been intended to interpret in a novel perpsective some important features of the evolutionary process undergone by languages. On the other hand, it has been outlined to introduce the question of the applicability of the quantitative procedures of analysis employed in the experimentation, in light of the fact that most of them are originally designed to analyze the history of biological entities, thus they often assume specific evolutionary models, which are appropriate to be used in the biological domain but not necessarily in other fields of inquiry. It has been shown that language evolution and biological evolution appear to be substantially different in some respects, but not in those that would compromise the application of the techniques of investigation in question. In particular, the most relevant differences between the two processes seem to concern the role of selection, which in the case of languages is not natural, but rather structural and social, and the mechanisms of individual transmission and inheritance. On the other hand some significant factors are common to both, i.e. those concerning the occurrence of innovations and the dynamics of diffusion, which are also related to certain important questions debated in historical linguistics and evolutionary biology, namely those regarding gradualism and contact between different lineages.

The following step of the discussion has been to focus on syntactic evolution, with the goal of illustrating that the specific analogical parallel between the parametric model of diachronic syntax and that of biological evolution seems to reinforce the conviction that the two processes may be studied by means of similar quantitative investigation methods, essentially because the parametric approach allows attaining both a satisfactory comparative description of the relevant linguistic traits and a suitable encoding of the data. In order to elucidate these points, it has been shown how the extension of the biolinguistic perspective of generative grammar from the cross-linguistic synchronic variation to the study of diachronic syntax may help shedding new light on the mechanisms of grammatical change and providing for them an appropriate explanatory account. In particular, the shift of focus to parametric syntax represents a fundamental premise of the present work, because the way in which the universal parameter set is characterized, i.e. essentially as a finite list of discrete biological options, permits overcoming the limits attributed to traditional syntactic comparison.

The latter claim has been explained framing the inquiry within the historical linguistics perspective and defining more precisely the objectives of the work. To this end the role of genealogical classification in historical linguistics has been taken into account, as well as a survey of diverse methods of comparison that have been adopted in historical linguistics studies. The classical comparative method, the lexicostatistic analysis and mass comparison have been considered in turn and it has been argued that, among these techniques, only the classical one may be considered a scientifically reliable and consistent procedure, but that it appears to be not universally applicable due to the fact that it is not effective in long-range comparisons. Thus the parametric comparison method has been introduced, claiming that, acting closer in accordance with the principles of the procedures adopted in population genetics, it provides a precise and confident method for measuring the degree of relatedness between languages. Moreover, it has been observed that the nature of parameters makes them good candidates not only for recovering evidence of distant genealogical relationships between languages, but also as input for quantitative procedures of phylogenetic reconstruction. It has been underlined that in the last decades these methods have been exported through disparate disciplines and this transfer of technology may play an important role in the development of the “new synthesis” between population genetics, archaeology and linguistics. Then the character-based and distance-based procedures adopted to carry out the investigation have been presented, as well as quantitative procedures specifically designed to detect horizontal transmission in the evolutionary history of languages.

Having grounded the applicability of the quantitative approach to the study of diachronic syntax, the attention has been focused on the illustration of the parametric dataset used in the experiments, introducing it with an account of the line of research in which this dissertation may be contextualized. Since the data consists of the comparative analysis of the DPs across various languages, the status of nominal syntax in generative grammar has been taken into account, with special reference to the specific theoretical background on which the parameter set incorporated in Table A is designed. Thus the encoding adopted in Table A and the structure of the dataset have been elucidated, and a specific section has been dedicated to the survey of past and current proposals regarding the genealogical classification of the languages considered in the investigation.

The last chapter has covered the illustration of the quantitative analyses performed on the parametric dataset. It is precisely at this point that the effectiveness and the limits of the relevant procedures have been tested. Summing up, after having offered an extensive character-based and distance-based description of the data, the choice of the phylogenetic methods adopted has been justified. Then the inquiry has been directed towards two main topics: The possibility of recovering genealogical information from the parametric comparison and the role of homoplasy (i.e. chance and borrowing) in the phylogenetic inference. Moreover, a tentative diachronic/diatopic mapping of syntactic distances derived form Table A has been offered. Some comments and suggestions have been pointed out throughout the discussion. For what concerns the evaluation of the data, it has been argued that:

→ The parametric dataset comprises sound genealogical information, as demonstrated both by preliminary empirical testing based on the examination of the syntactic distances and by the first outfit of phylogenies derived from the whole sample of modern languages included in Table A on the full list of unweighted parameters. Consequently the results support the idea that it is possible to successfully recover the vertical direction of syntactic evolution on the basis of the parametric comparison method. Moreover, the quantitative phylogenetic analyses carried out on the sample of linguistic varieties in question appear to provide good indications of the fact that the parametric comparison may effectively help shedding light on long-range genealogical relationships. In any case, it is safe that the data included in Table A allow identifying traditionally established genealogical groupings with a sufficiently high degree of confidence, while the solidity of the phylogenetic inferences regarding the relations between these groups is inferior.

→ The amount of genealogical information contained in the data may be increased reducing homoplasy (i.e. the effect of chance and contact) and assuming a theory of markedness in encoding the parameter states in Table A. As for the former point, it has been suggested that the main source of homoplasy that affects the parametric data is parallel development. This idea relies on the expectation that, being the possible basic parameter states just two by definition, in principle changes occurring in different lineages may easily lead languages to independently converge on the same setting, but it is also grounded on the empirical observation that there is no correlation between parametric and geographic variation in the European area. In the course of the experimentation it has been shown that it is possible to identify at least a subset of parameters, in which the amount of homoplasy is significantly diminished and which permits reconstructing well-supported genealogical classifications by means of diverse distance-based phylogenetic methods. The idea of assigning markedness estimates to the parametric values, instead, remains purely theoretical for the moment. However, it would be fundamental to introduce information regarding the directionality of changes, and, ultimately, to define a model describing the evolutionary process undergone by the parameter system, which would pave the way to the use of sophisticated character-based phylogenetic techniques and the identification of more accurate phylogenetic hypotheses.

→ Parameter-based syntactic distances appear to lack a significant correlation with separation times and their projection onto a precise chronology appears difficult due to the structured nature of the data (i.e. the presence of implicational conditions that relate parameters), which obscures the actual stages of development of the system. It has already been recalled that the patterns of parameter-based syntactic and geographic variation seem to be largely unrelated as well: As a first approximation, the contrast with the situation of syntactic (dialectal) "micro-variation", which instead appears to extensively follow the areal distribution of the speech communities, may lead one to think that traces of proper genealogical relatedness are more likely to be found looking at "macro-variation" of languages (see in any case the problems related to the definition of the micro/macro distinction reported in section 3.1.5).

The main observations regarding more specifically the effectiveness and the limits of quantitative techniques to the investigation of syntactic evolution are that:

→ The specific encoding of the data limits the application of some quantitative phylogenetic methods. While distance-based procedures do not pose problems given an accurate measure of dissimilarity between languages, character-based techniques encounter difficulties, again, due to the underlying structure of the data and the lack of a theory of markedness that helps drawing a precise model of parameter evolution. On the other hand, the hypothesis of a "clocklike evolution" (MCH) that is integrated in some distance-based techniques may be shown not only to be in principle applicable to the data in question, but also to allow obtaining largely correct phylogenetic inferences, especially regarding long-range genealogical relationships, plausibly due to the Anti-Babelic nature of syntactic distances.

→ Given that the bias of the inferred genealogical classification may be reduced limiting the influence of homoplasy, this goal may be achieved selecting/weighting characters on the basis of different criteria: For the moment no independent quantitative procedure appears to be apt to provide a sound set of weights, but, as already recalled, an adequate selection of parameters permits designing an appropriate subset of genealogically relevant parameters. Furthermore, the diverse dynamics of homoplasy are difficult to distinguish relying exclusively on quantitative analyses and the task requires one to be provided with extensive knowledge of the historical and qualitative characterization of the data.

→ The application of the character-based quantitative analysis to the issue of the reconstruction of the ancestral states may supply tentative indications for mapping ancient varieties onto phylogenies, i.e. to solve the problem of the "internalization" of the ancestors in the phylogenetic structures.

Many points, some of which have already been taken into consideration in the last years, deserve and require to be reconsidered in future research. Among others, providing a better quantitative account for the data in light of their underlying structure appears to be particularly relevant. This topic is precisely connected to the definition of a model of parametric evolution, which would entail incorporating knowledge about the directions and the probabilities of change in the reconstruction of the evolutionary history of parameters. The identification of a set of purely quantitative weights aimed at reducing homoplasy, instead, is already under study: The objective is to define not only a global measure of genealogical relevance, i.e. valid with respect to any possible lineage defined by the languages of the dataset, but also a local one, that applies to specific areas of the genealogical structure and may vary according to the specific languages that are compared. Many efforts have been made to pursue this goal, even if the procedures designed so far, as well as the results obtained, are not mature enough to be presented here. Finally, the analysis of homoplasy on geographical, chronological and structural grounds represents another further key line of inquiry, whose development requires attaining an extensive and accurate knowledge of the relevant historical and linguistic evidence.

References

Aikhenvald, A. Y. and R. M. W. Dixon. 2001. “Introduction” in Aikhenvald, A. Y. and R. M.W. Dixon (eds.) 2001. *Areal diffusion and genetic inheritance*: 1-26. Oxford: Oxford University Press.

Abney, S. P. 1987. *The English Noun Phrase in its Sentential Aspect*. Ph.D. dissertation, MIT.

Adrados, F. R. 1979. “Arqueología y diferenciación del Indoeuropeo” in *Emerita*, 47: 261-282.

Aikhenvald, A. Y. and R. M. W. Dixon. 2001. “Introduction” in Aikhenvald, A. Y. and R. M.W. Dixon (eds.) 2001. *Areal diffusion and genetic inheritance*: 1-26. Oxford: Oxford University Press.

Allen, C. 2000. “Obsolescence and sudden death in syntax: the decline of verb-final order in early Middle English” in Bermúdez-Otero, R. D. Denison, R. M. Hogg and C. B. McCully (eds.) *Generative theory and corpus studies: a dialogue from 10ICEHL*: 3-25. Berlin: Mouton de Gruyter.

Anderson, M. 1979. *Noun Phrase Structure*. Ph.D. dissertation, University of Connecticut.

Anderson, N. 1879. *Studien zur Vergleichung der indogermanischen und finnisch-ugrischen Sprachen*. Dorpat: H. Laakmann.

Ashlock, P. 1974. “The uses of cladistics” in *Annual Review of Ecology and Systematics*, 5: 81-99.

Atkinson, Q. D. and R. D. Gray. 2006a. “Are accurate dates an intractable problem for historical linguistics?” in Lipo, C., M. O’Brien, S. Shennan and M. Collard (eds.) *Mapping our Ancestry: Phylogenetic Methods in Anthropology and Prehistory*: 269–96. Chicago (IL): Aldine.

Atkinson, Q. D. and Gray, R. D. 2006b. “How old is the Indo-European language family? Progress or more moths to the flame?” in Forster, P. and C. Renfrew (eds.) *Phylogenetic Methods and the Prehistory of Languages*: 91-109. Cambridge: The McDonald Institute for Archaeological Research.

Atkinson, Q. D., G. Nicholls, D. Welch and R. D. Gray. 2005. “From words to dates: water into wine, mathemagic or phylogenetic inference?” in *Transactions of the Philological Society*, 103(2): 193–219.

Autran, C. 1925. *Sumerien et Indo-europen: l'aspect morphologique de la question*. Paris: Paul Geuthner.

Baker, C. L. and J. J. McCarthy. 1981. *The logical problem of language acquisition*. Cambridge, MA: MIT Press.

Baker, M. 2001. *The Atoms of Language: The Mind’s Hidden Rules of Grammar*. Oxford: Oxford University Press.

Baker, M. 2003. “Linguistic differences and language design” in *Trends in Cognitive Sciences*, 7: 349-353.

Baker, M. 2008. “The macroparameter in a microparametric world.” in Biberauer, T. (ed.) *The Limits of Syntactic Variation*: 351-373. Amsterdam: John Benjamins.

Baldi, P. 1983. *An Introduction to the Indo-European Languages*. Carbondale, IL: Southern Illinois University Press.

Baldi, P. 1990. “Introduction” in Baldi, P. (ed.) *Linguistic change and reconstruction methodology*: 1-16. New York: Mouton de Gruyter.

Bandelt, H.-J. and A. W. M Dress. 1992. “A canonical decomposition theory for metrics on a finite set” in *Advanced Mathematics*, 92: 47–105.

Bandelt H.-J., Forster P., Sykes B.C., Richards M.B. 1995. “Mitochondrial portraits of human populations” in *Genetics*, 141: 743-753.

Barbiers, S., H. Bennis, M. Devos, G. de Vogelaer and M. van der Ham (eds.). 2005. *Syntactic Atlas of the Netherlandic Dialects*. Volume 1. Amsterdam: Amsterdam University Press.

Barbrook, A. C., C. J. Howe, N. Blake and P. Robinson. 1998. “The phylogeny of The Canterbury Tales” in *Nature*, 349: 839.

Baret, P. V., C. Macé and P. Robinson. 2006. “Testing methods on an artificially created textual tradition” in Macé, C., P. V. Baret, A. Bozzi, L. Cignoni (eds.) *The Evolution of Texts. Confronting Stemmatological and Genetical Methods. Proceedings of the International*

Workshop held in Louvain-la-Neuve on September 1-2, 2004 (Linguistica 24-25): 255-283. Pisa-Roma: Istituti Editoriali e Poligrafici Internazionali.

Battistella, E. 1996. *The logic of markedness*. Oxford: Oxford University Press.

Behagel, O. 1878. "Die neuhochdeutschen Zwillingswörter" in *Germania: Vierteljahrsschrift für deutsche Alterumskunde*, 23(11): 257-292.

Belardi, W. 2001. *L'etimologia nella storia della cultura occidentale*. Roma: Il Calamo.

Ben Hamed, M. 2005. "Neighbour-nets portray the Chinese dialect continuum and the linguistic legacy of China's demic history" in *Proceedings of the Royal Society*, B, 272: 10015-1022.

Benedetti, M. 2003. "L'etimologia fra tipologia e storia" in Mancini, M. (ed.) *Il cambiamento linguistico*: 209-262. Roma: Carocci editore.

Bergsland, K. and H. Vogt. 1962. "On the validity of glottochronology" in *Current Anthropology*, 3: 115-153.

Bernstein, J. B. 1991. "DPs in French and Walloon: Evidence for parametric variation in nominal head movement" in *Probus*, 3: 1-26.

Bernstein, J. B. 1992. "On the syntactic status of adjectives in Romance", presented at the *1992 Annual LSA Meeting*, Philadelphia.

Bernstein, J. B. 1993. *Topics in the Syntax of Nominal Structure across Romance*. Ph.D. dissertation, City University of New York.

Bernstein, J. B. 2001. "The DP Hypothesis: Identifying Clausal Properties in the Nominal Domain" in Baltin, M. and C. Collins (eds.) *The Handbook of Contemporary Syntactic Theory*: 536-561., Cambridge, MA - Oxford, UK: Blackwell.

Bernstein, J. B. 2008. "English th- Forms" in Klinge, A. and H. Müller (eds.) *Essays on Nominal Determination*: 213-232. Philadelphia-Amsterdam: John Benjamins.

Berwick, R. 1985. *The acquisition of Syntactic Knowledge*. Cambridge: Cambridge University Press.

Blust, R. 1996. "The Neogrammarian Hypothesis and Pandemic Irregularity" in Durie, M. and M. Ross (eds.) *The comparative method reviewed: Regularity and irregularity in language change*: 135-156. Oxford: Oxford University Press.

Boeckx, C. 2008. *Approaching Parameters from Below*. Manuscript. Harvard University.

Boeckx, C. and N. Hornstein. 2007. "Les differents objectifs de la linguistique theorique [The varying aims of linguistic theory]" in Franck, J. and J. Bricmond (eds.) *Cahier Chomsky*: 61-77. Paris: L'Herne.

Bomhard, A. R. and J. C. Kerns. 1994. *The Nostratic macrofamily: A study in distant linguistic relationship*. Berlin-New York: Mouton de Gruyter.

Bopp, F. 1816. *Über das Conjugationssystem der Sanskritsprache in Vergleichung mit jenem der griechischen, lateinischen, persischen und germanischen Sprache, nebst Episoden des Ramajan und Mahabaharat in genauen, metrischen Übersetzungen aus dem Originaltaxte und einigen Abschnitten aus den Vedas*. Frankfurt: Andreäische Buchhandlung.

Borer, H. 1984. *Parametric Syntax*. Dordrecht: Foris.

Bowern, C. In press. "Syntactic Change and Syntactic Borrowing in Generative Grammar" in Ferarresi, G. and M. Goldbach (eds.), *Handbook of Historical Reconstruction*.

Bowler, P. J. 1975. "The Changing Meaning of 'Evolution'" in *Journal of the History of Ideas*, 36(1): 95-114.

Bowler, P. J. 2003. *Evolution. The History of an Idea*. Berkley – Los Angeles – London: University of California Press.

Bradley, E. 1979. "Bootstrap Methods: Another Look at the Jackknife" in *The Annals of Statistics*, 7(1): 1–26.

Brame, M. 1982. "The head-selector theory of lexical specifications and the nonexistence of coarse categories" in *Linguistic Analysis*, 10: 321-325.

Briscoe, T. 2002. "Introduction" in Briscoe, T. (ed.) *Linguistic evolution though language acquisition*: 1-21. Cambridge, UK: Cambridge University Press.

Brooks, D. R., E. Erdem, J. Minett, W. James and D. Ringe. 2005. "Character-based cladistics and answer set programming" in *Proceedings of the Seventh International Symposium on Practical Aspects of Declarative Languages (PADL 05)*: 37–51.

Brugmann, K. 1925. *Die Syntax des einfachen Satzes im Indogermanischen*. Berlin and Leipzig: Walter de Gruyter.

Bryant, D. 2004. “Radiation and Network Breaking in Polynesian Linguistics” in *Phylogenetic Methods and the Prehistory of Languages*. Cambridge: McDonald Institute for Archaeological Research.

Bryant, D. and V. Moulton. 2004. “NeighborNet: An Agglomerative Method for the construction of Planar Phylogenetic Networks” in *Molecular Biology and Evolution*, 21(2): 255-265.

Camin, J. H. and R. R. Sokal. 1965. “A method for deducing branching sequences in phylogenies” in *Evolution*, 19: 311-326.

Campanile, E. “La ricostruzione linguistica e culturale” in Lazzeroni, R. (ed.) *Linguistica storica*: 13-54. Roma: La Nuova Italia Scientifica.

Campanile, E., B. Compie and C. Watkins. 2005. *Introduzione alla lingua e alla cultura degli Indoeuropei*. Bologna: Il Mulino.

Campbell, L. 1988. “Review of ‘Language in the Americas’ (Joseph H. Greenberg)” in *Language*, 64: 591-615.

Campbell, L. 1993. “On proposed universals of grammatical borrowing” in Aertsen, Henk, and Jeffers, Robert J. (eds.) *Historical Linguistics1989: Papers from the 9th International Conference on Historical Linguistics 1989*. Rutgers University, 14-18 August 1989. Amsterdam: John Benjamins Publishing Co., pp 91-109.

Campbell, L. 1996. “On Sound Change and Challenges to Regularity” in Durie, M. and M. Ross (eds.) *The comparative method reviewed: Regularity and irregularity in language change*: 72-89. Oxford: Oxford University Press.

Campbell, L. 1998. “Nostratic: A personal Assessment” in Salmons, J. C. and B. D. Joseph (eds.) *Nostratic. Sifting the evidence*: 107-152. Amsterdam: John Benjamins.

Campbell, L. 2001. “What’s wrong with grammaticalization ?” in *Language Sciences*, 23: 113-162.

Campbell, L. 2006. “Languages and Genes in Collaboration: some Practical Matters”, paper presented at the *Language and Genes: an interdisciplinary conference, University of California, Santa Barbara, 8th-10th September 2006.*

Campbell, Lyle. 2004. *Historical Linguistics: An Introduction*. Cambridge, MA: MIT Press.

Carlson, G. 1977. *Reference to Kinds in English*. Ph.D. dissertation, UMass at Amherst.

Cavalli-Sforza, L. L. 1962. “The distribution of migration distances, models and applications to genetics” in Sutter, J. (ed.) *Human displacements: measurement, methodological aspects*: 139-158. Munich: Éditions Sciences Humaines.

Cavalli-Sforza, L.L. 2001. *Geni, popoli e lingue*. Milano: Adelphi.

Cavalli-Sforza, L. L and A. W. F. Edwards. 1967. “Phylogenetic analysis: models and estimation procedures” in *American Journal of Human Genetics*, 19: 233-257.

Cavalli-Sforza, L. L. and M. Feldman. 1981. *Cultural Transmission and Evolution. A Quantitative Approach*. Princeton, NJ: Princeton University Press.

Cavalli-Sforza, L. L. and S-Y. Wang. 1986. “Spatial distance and lexical replacement” in *Language*, 62: 38-55.

Cavalli-Sforza, L. L., A. Piazza, P. Menozzi and J. Mountain. 1988. “Reconstruction of human evolution: bringing together genetic, archeological and linguistic data” in *Proceedings of the National Academy of Sciences*, 85: 6002-6006.

Cavalli-Sforza, L. L., P. Menozzi and A. Piazza. 1994. *The History and Geography of Human Genes*. Princeton: Princeton University Press.

Cavalli-Sforza, L. L. and M. Feldman. 2003. “The application of molecular genetics approaches to the study of human evolution” in *Nature genetics*, 33: 266-275.

Chapell, H. 2001. “Language Contact and Areal Diffusion in Sinitic Languages” in Aikhenvald, A. Y. and R. M.W. Dixon (eds.) 2001. *Areal diffusion and genetic inheritance*: 291-327. Oxford: Oxford University Press.

Chierchia, G. 1985. “Formal Semantics and the Grammar of Predication” in *Linguistic Inquiry*, 16(3): 417-443.

Chierchia, G. 1998. “Reference to kinds across languages” in *Natural Language and Semantics*, 6: 339–405.

Chippindale, P. T. and J. J. Wiens. 1994. “Weighting, partitioning, and combining characters in phylogenetic analysis” in *Systematic Biology*, 43: 278–287.

Chomsky, N. 1957. *Syntactic Structures*. The Hague-Paris: Mouton and Co.

Chomsky, N. 1964. *Current issues in linguistic theory*. The Hague: Mouton and Co.

Chomsky, N. 1965. *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Chomsky, N. 1970. “Remarks on nominalization” in Roderick, J. and P. Rosenbaum (eds.) *Readings in English transformational grammar:* 184-221. Waltham, MA: Grinn.

Chomsky, N. 1973. “Conditions on transformations” in Anderson S. R. and P. Kiparsky (eds.) *A festschrift for Morris Halle*: 232-286. New York: Holt, Rinehart and Winston.

Chomsky, N. 1981. *Lectures on Government and Binding*. Dordrecht: Foris.

Chomsky, N. 1982. *Some Concepts and Consequences of the Theory of Government and Binding*. Cambridge, MA: MIT Press.

Chomsky, N. 1986a. *Knowledge of language*. New York: Praeger.

Chomsky, N. 1986b. *Barriers*. Cambridge, MA: MIT Press.

Chomsky, N. 1995. *The Minimalist Program*. Cambridge, MA: MIT Press.

Chomsky, N. 2000. “Minimalist inquiries: The framework” in Martin, R., D. Michael and J. Uriagereka (eds.) *Step by Step: Essays in Honor of Howard Laskin*: 89-155. Cambridge, MA: MIT Press.

Chomsky, N. 2001. “Derivation by Phase” in Kenstowicz, M. (ed.) *Ken Hale: A Life in Language*: 1-52. Cambridge, MA: MIT Press.

Chomsky, N. 2004. “Beyond explanatory adequacy” in Belletti, A. (ed.) *Structures and beyond*: 104-131. Oxford: Oxford University Press.

Chomsky, N. 2005. *On phases*. Manuscript. MIT.

Chomsky, N. 2008. *The biolinguistic program: Where does it stand today?* Manuscript, MIT.

Christiansen, M. H. and S. Kirby. 2003. “Language Evolution: The Hardest Problem in science?” in Christiansen, M. H. and S. Kirby (eds.) *Language Evolution: The States of the Art*: 1-15. Oxford: Oxford University Press.

Cinque, G. 1980. “On extraction from NP in Italian” in *Journal of Italian Linguistics*, 5: 47-99.

Cinque, G. 1994. “On the evidence for partial N-movement in the Romance DP” in Cinque, G., J. Koster, J.-Y. Pollock, L Rizzi and R. Zanuttini (eds.) *Paths towards Universal Grammar: Studies in Honor of Richard S. Kayne*: 85-110. Washington, DC: Georgetown University Press.

Clark, R. 1990. “Papers on Learnability and natural selection” in *Technical Reports on Formal and Computational Linguistics 1*, University of Geneva.

Clark, R. 1992. “The selection of syntactic knowledge” in *Language Acquisition*, 2(2): 83-149.

Clark, R. and I. Roberts. 1993. “A computational Approach to Language Learnability and Language Change” in *Linguistic Inquiry*, 24: 299-345.

Clark, R. and I. Roberts. 1994. *On complexity as the engine of variation*. Manuscript, University of Pennsylvania and Bangor.

Colless, D. H. 1985. “On ‘Character’ and related terms” in *Systematic Zoology*, 34: 229-233.

Contreras, H. 1986. “Spanish bare NPs and the ECP” in Bordelois, I, H. Contreras and K. Zagona (eds.) *Generative studies in Spanish syntax*: 25–50. Dordrecht: Foris.

Crisma, P. 1991. *Functional Categories inside the Noun Phrase: A study of the Distribution of Nominal Modifiers*. MA dissertation, Università di Venezia.

Crisma, P. 1993. “Adjective Placement in Romance and Germanic Event Nominals” in *Rivista di Grammatica Generativa*, 18: 61-100.

Crisma, P. 1996. “On the configurational nature of adjectival modification” in Zagona, K. (ed.) *Grammatical Theory and Romance Languages – Current Issues in Linguistic Theory*: 59-71. Amsterdam: John Benjamins.

Crisma, P. 1997. *L'articolo nella prosa inglese antica e la teoria degli articoli nulli*. Ph.D. dissertation, Università di Padova.

Croft, W. 1996. “”Linguistic selection: an utterance-based evolutionary theory of language change” in *Nordic Journal of Linguistics*, 19: 99-139.

Croft, W. 2000. *Explaining Language Change. An evolutionary Approach*. Harlow, Essex: Longman.

Cuny, A. 1924. *Études prégrammaticales sur le domain des langues indoeuropeénnes et chamito-sémitiques*. Paris.

Darwin, C. 1859. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. London: John Murray.

Darwin, C. 1871. *The Descent of Man and Selection in Relation to Sex*. London: John Murray.

Dawkins, R. 1976. *The Selfish Gene*. Oxford, UK: Oxford University Press.

Dawkins, R. and Y. Wong. 2005. *The Ancestor's Tale. A Pilgrimage to the Dawn of Evolution*. New York: Houghton Mifflin Harcourt.

De Queiroz, A. 1993. "For consensus (sometimes)" in *Systematic Biology*, 42: 368–372.

De Saussure, F. 1879. *Mémoire sur le système primitif des voyelles dans les langues indo-européennes*. Leipzig: Teubner.

Delbrück, B. 1878. *Die altindische Wortfolge aus dem Çatapathabrâ mana*. Halle and der Saale: Buchhandlung des Weisenhauses.

Dench, A. 2001. "Descent and diffusion: the complexity of the Pilbara situation" in Aikhenvald, A. Y. and R. M. W. Dixon (eds.) *Areal diffusion and genetic inheritance: problems in comparative linguistics*. Oxford: Oxford University Press.

Denison, D. 1999. "Slow, Slow, Quick, Quick, Slow: the Dance of Language Change?" in Bringas López, A. at alii (eds.) *'Woodenrous Ænglissce': SELIM Studies in Medieval English Language*: 51-64. Vigo: Universidada de Vigo (Servicio de Publicacións).

Diamond, J. M. 1988. "Express train to Polynesia" in *Nature*, 336: 307-308.

Diez, F. 1836. *Grammatik der romanischen Sprachen*. Bonn: Weber.

Dillon, M. 1944. "Italic and Celtic" in *The American Journal of Philology*, 65(2): 124-134.

Dixon, R. M. W. 1997. *The rise and fall of languages*. Cambridge: Cambridge University Press.

Dixon, R. M. W. 2001. "The Australian Linguistic Area" in Aikhenvald, Alexandra Y. and R. M.W. Dixon (eds.) *Areal diffusion and genetic inheritance*: 64-104. Oxford: Oxford University Press.

Dolgopolsky, A. B. 1986. "A probabilistic hypothesis concerning the oldest relationships among the language families of North Eurasia" in Shevoroshkin, VV and T. L. Markley (eds.) *Typology relationship and time*. Ann Arbour.

Dresher, B. E. 1999. "Charting the Learning Path: Cues to parameter setting" in *Linguistic Inquiry*, 30(1): 27-67.

Dresher, B. E. and J. D. Kaye. 1990. "A computational learning model for metrical phonology" in *Cognition*, 34: 137–195.

Dressler, W. 1976. "Fonologia naturale e cambio linguistico" in *Studi di Fonetica e Fonologia*. Roma: Bulzoni.

Dryer, M. S. 1989 "Article-Noun Order" in *Papers from the 25th Regional Meeting of the Chicago Linguistic Society*: 83-97.

Dunn, M. A. Terrill, G. Reesink, R. A. Foley, S. C. Levinson. 2005. "Structural Phylogenetics and the Reconstruction of Ancient Language History" in *Science*, 309(5743): 2072-2075.

Dyen, I., J. B. Kruskal and P. Black. 1992. "An Indoeuropean classification: a lexicostatistical experiment" in *Transactions of the American Philosophical Society*, 82(5): III-132.

Eckardt, A. 1966. *Koreanisch und Indogermanisch Untersuchungen über die Zugehörigkeit des Koreanischen zur indogermanischen Sprachfamilie*. Heidelberg: Julius Groos.

Edmonds, J. 1978. "The verbal complex V'-V in French" in *Linguistic Inquiry*, 9: 151-175.

Edwards, A. W. F and Cavalli-Sforza L. L. 1963. "The reconstruction of evolution" in *Annals of Human Genetics*, 27: 106-106.

Edwards, A. W. F. 1972. *Likelihood: An account of the statistical concept of likelihood and its application to scientific inference*. Cambridge: Cambridge University Press.

Edwards, A. W. F. and L. L. Cavalli-Sforza. 1964. "Reconstruction of evolutionary trees" in Heywood, V. H. and J. McNeill (eds.) *Phenetic and Phylogenetic Classification*: 67-76. London: Systematics Association Publ. No. 6.

Eldredge, N. and S. J. Gould. 1972. "Punctuated equilibria: an alternative to phyletic gradualism" in Schopf, T. J. M. (ed.) *Models in Paleobiology*: 82-115. San Francisco: Freeman and Cooper.

Eldredge, N. and S. J. Gould. 1988. "Punctuated equilibrium prevails" in *Nature*, 332: 211-212.

Embleton, S. 1981. *Incorporating Borrowing Rates in Lexicostatistical Tree Reconstruction*. Ph.D. dissertation, Department of Linguistics, University of Toronto.

Embleton, S. 1986. *Statistics in Historical Linguistics*. Bochum: Brockmeyer.

Embleton, S. 2000. "Lexicostatistics/glottochronology: from Swadesh to Sankoff to Starostin to future horizons" in Renfrew, C., A. McMahon and L. Trask (eds.) *Time Depth in Historical Linguistics*: 143-166. Cambridge: McDonald Institute for Archaeological Research.

Emeneau, M. B. 1964. "India as a linguistic area" in Hymes, D. (ed.) *Language in culture and society*. New York: Harper and Row.

Erdem, E., V. Lifschitz, L. Nakhleh and D. Ringe. 2003. “Reconstructing the Evolutionary History of Indo-European Languages using Answer Set Programming” in *Proceedings of the 5th International Symposium on Practical Aspects of Declarative Languages (PADL 03).*

Ersts, P. J. 2008. *Geographic Distance Matrix Generator (version 1.2.3).* American Museum of Natural History, Center for Biodiversity and Conservation. Available from http://biodiversityinformatics.amnh.org/open_source/gdmg. Accessed November 2008.

Estabrook, G. F., C. S. Johnson and F. R. McMorris. 1976a. “A mathematical foundation for the analysis of cladistic character compatibility” in *Mathematical Bioscience*, 29: 181-187.

Estabrook, G. F., C. S. Johnson and F. R. McMorris. 1976b. “An algebraic analysis of cladistic characters” in *Discrete Mathematics*, 16: 141-147.

Evans, S. N., D. Ringe and T. Warnow. 2006. “Inference of divergence times as a statistical inverse problem” in Forster, P. and C. Renfrew (eds.) *Phylogenetic Methods and the Prehistory of Languages*: 119-138. Cambridge: The McDonald Institute for Archaeological Research.

Faarlund, J. T. 1990. *Syntactic change. Toward a Theory of Historical Syntax.* Berlin: Mouton de Gruyter.

Fanciulllo, F. 2007. *Introduzione alla linguistica storica*. Bologna: Il Mulino.

Farris, J. S. 1989. “The retention index and the rescaled consistency index” in *Cladistics*, 5: 417–419.

Felsenstein, J. 1985. “Confidence limits on phylogenies: an approach using the bootstrap” in *Evolution*, 39: 783-791.

Felsenstein, J. 2004a. *Inferring phylogenies*. Sunderland, MA: Sinauer.

Felsenstein, J. 2004b. *PHYLIP (Phylogeny Inference Package) version 3.6b*. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Fisher, R. A. 1922. “On the mathematical foundations of theoretical statistics” in *Philosophical Transactions of the Royal Society of London, A*, 222: 309-368.

Fitch, W. M. and E. Margoliash. 1967. “Construction of phylogenetic trees” in *Science*, 155: 279-284.

Fitch, W. T., M. D. Hauser and N. Chomsky. 2005. “The Evolution of the Language Faculty: Clarifications and Implications” in *Cognition*, 97(2): 179-210.

Flikeid, K. 1989. “‘Moitié Français, moitié anglais?’ Emprunts and alternances de langues dans les communautés acadiennes de la Nouvelle-Ecosse” in *Reveu Québécoise de Linguistique Théorique et Appliquée*, 8 : 277-228.

Fodor, I. 1982. *A Fallacy of Contemporary Linguistics. J. H. Greenberg’s Classification of the African Languages and His ‘Comparative Method’*. Hamburg: Buske.

Fodor, J. D. 2001. “Setting Syntactic Parameters” in Baltin, M. R. and C. Collins (eds.) *Handbook of Contemporary Syntactic Theory*: 730-767. Oxford: Blackwell Publishers.

Forster, P. and A. Toth. 2003. “Toward a phylogenetic chronology of ancient Gaulish, Celtic, and Indo-European” in *Proceedings of the National Academy of Science*, 100: 9079-9084.

Forster, P., Polzin, T., and Rohl, A. 2006. “Evolution of English Basic Vocabulary within the Network of Germanic Languages” in Forster, P. and C. Renfrew (eds.) *Phylogenetic Methods and the Prehistory of Languages*: 131-137. Cambridge: McDonald Institute for Archaeological Research.

Fox, A. 1995. *Linguistic Reconstruction: An Introduction to Theory and Method.* Oxford: Oxford University Press.

Gebhardt, L. 2005. “Preposition stranding and the noun compounding parameter” presented at the *NWAV34, October 20th-23rd 2008, New York University.*

Gensler, O. D. 1993. *A typological evaluation of Celtic/Hamito-Semitic syntactic parallels*. Ph.D. dissertation, University of California at Berkley.

Giacalone Ramat, A. 1998. “Testing the boundaries of grammaticalization.” in Giacalone Ramat, A. and Hopper P. J. (eds.) *The limits of grammaticalization*: 107-127. Amsterdam: Benjamins.

Giacalone, A. and C. Mauri. To appear. “From cause to contrast. a study in semantic change” to appear in a Festschrift volume.

Giannini, S. 2003. “Il mutamento morfologico” in Mancini, M. (ed.) *Il cambiamento linguistico*: 89-163. Roma: Carocci editore.

Gianollo, C. 2005. *Constituent structure and parametric resetting in the Latin DP: a diachronic study.* Ph.D. dissertation, Università di Pisa.

Gianollo, C, C. Guardiano and G. Longobardi. 2004. “Historical Implications of a formal Theory of Syntactic Variation” n Anderson, S. and D. Jonas (eds.) *Proceedings of DIGS VIII.* Oxford: Oxford University Press.

Gianollo, C, C. Guardiano and G. Longobardi. 2008. “Three fundamental issues in parametric linguistics” in T. Biberauer, T. (ed.) *The Limits of Syntactic Variation*: 109-142. Philadelphia/Amsterdam: John Benjamins.

Gianollo, C, C. Guardiano and G. Longobardi. In press. “Syntactic measuring of language relatedness” in A. Kidwai et alii (eds.) *Proceedings of V Asian GLOW, New Delhi, JNU, 2005*.

Gibson, E. and K. Wexler. 1994. “Triggers” in *Linguistic Inquiry*, 25(3): 407-454.

Gimbutas, M. 1970. “Proto-Indo-European culture: The Kurgan culture during the fifth to the third millennia B.C.” in Cadorna, G. C., H. M. Hönigswald and A. Senn (eds.) *Indo-European and Indo-Europeans*: 155-198. Philadelphia: University of Philadelphia Press.

Giorgi, A. and G. Longobardi. 1991. *The syntax of Noun Phrases: Configuration, parameters and empty categories*. Cambridge, UK: Cambridge University Press.

Giusti, G. 2001. “The birth of a functional category: from Latin ILLE to the Romance article and personal pronoun” in Cinque, G. and G. Salvi (eds.) *Current Studies in Italian Syntax: Essays Offered to Lorenzo Renzi*: 157-171. Amsterdam: North Holland.

Givón, T. 1984. *Syntax: a functional-typological introduction, vol. 1*. Amsterdam: John Benjamins.

Goddard, I. 1987. “Review of ‘Language in the Americas’ (Joseph H. Greenberg)” in *Current Anthropology*, 28: 656-657.

Gong, T., J. W. Minett and W. S.-Y. Wang. 2007. “A Simulative Study of the Roles of Cultural Transmission in Language Evolution” in *Proceedings of 2007 IEEE Congress on Evolutionary Computation*: 843-850. Singapore.

Gould, S. J. and E. S. Vbra. 1982. “Exaptation – a missing term in the science of form” in *Paleobiology*, 8: 4-15.

Graffi, G. and S. Scalise. 2002. *Le lingue e il linguaggio*. Bologna: Il Mulino.

Gray, R. D. 2005. “Pushing the Time Barrier in the Quest for Language Roots” in *Science*, 309(5743): 2007-2008.

Gray, R. D. and F. M. Jordan. 2000. “Language trees support the express-train sequence of Austronesian expansion” in *Nature*, 405: 1052-1055.

Gray, R. D. and Q. D. Atkinson. 2003. “Language-tree divergence times support the Anatolian theory of Indo-European origin” in *Nature*, 426: 435-439.

Gray, R. D., Drummond, A. J., and Greenhill, S. J. 2009. “Language Phylogenies Reveal Expansion Pulses and Pauses in Pacific Settlement” in *Science*, 323(5913): 479-483.

Greenberg, J H., C. G. Turner II and Stephen Zegura. 1986. “The settlement of the Americas: a comparison of the linguistic, dental and genetic evidence” in *Current Anthropology*, 25: 477-97.

Greenberg, J. 1963. “Some universals of grammar with particular reference to the order of meaningful elements” in Greenberg, J. (ed.) *Universals of Grammar*: 73-113. Cambridge, MA: MIT Press.

Greenberg, J. 1966. “Some Universal of Grammar with particular reference to the order of meaningful elements” in Greenberg, J. (ed.) *Universals of Language*: 73-113. Cambridge, MA: MIT Press.

Greenberg, J. H. 1954. “Studies in African linguistic classification VIII. Further remarks on method; revisions and corrections” in *Southwestern Journal of Anthropology*, 10: 405-15.

Greenberg, J. H. 1963. *The languages of Africa*. Bloomington: Indiana University.

Greenberg, J. H. 1971. “The Indo-Pacific hypothesis” in *Current trends in linguistics*, 8: 808-871.

Greenberg, J. H. 1987. *Language in the Americas*. Stanford, CA: Stanford University Press.

Greenberg, J. H. 2000. *Indo-European and Its Closest Relatives*. Stanford, CA: Stanford University Press.

Greenberg, J., C. A. Ferguson, and E. A. Moravcsik (eds.). 1978. *Universals of human language, Vol. 3: word structure*. Stanford: Stanford University Press.

Guardiano, C. 2003. *Struttura e storia del Sintagma Nominale nel Greco Antico: ipotesi parametriche*. Ph.D. dissertation, Università di Pisa.

Guardiano, C. and G. Longobardi. 2005. “Parametric Comparison and Language Taxonomy” in M. Batllori, M., M. L. Hernanz, C. Picallo, and F. Roca (eds.) *Grammaticalization and Parametric Variation*: 149-174. Oxford: Oxford University Press.

Guasti, M.-T. 2002. *Language Acquisition: The Growth of Grammar*. Cambridge, MA: MIT Press.

Gumperz, J. J. and R. Wilson. 1971. “Convergence and Creolization: a case from the Indo-Aryan /Dravidian border” in Hymes, Dell (ed.) *Pigdization and creolization of languages*: 151-167. Cambridge: Cambridge University Press.

Gusschinsky, S. 1956. “Three distirbing questions concernine lexicostaistics” in *International Journal of American Linguistics*, 22: 212-213.

Guthrie, W. K. C. 1957. *In the Beginning: Some Greek Views on the Origins of Life and the Early State of Man*. Ithaca - New York: Cornell University Press.

Hale, M. 1998. “Diachronic Syntax” in *Syntax*, 1: 1-18.

Hamming, R. W. 1950. “Error Detecting and Error Correcting Codes” in *Bell System Technical Journal*, 26(2): 147-160.

Harris, A. C. and L. Campbell. 1995. *Historical syntax in cross-linguistic perspective*. Cambridge: Cambridge University Press.

Harrison, P. 2003. “On the limits of the classical comparative method” in Joseph, B. and Janda R. (eds.) *The Handbook of Historical Linguistics*: 213-243. Oxford: Blackwell Publishing.

Harvey, P. H. and M. D. Pagel. 1998. *The comparative method in evolutionary biology*. New York: Oxford University Press.

Haspelmath, M. 2004b. “How hopeless is genealogical linguistics, and how advanced is areal linguistics?” in *Studies in Language*, 28(1): 209-223.

Haspelmath, M. In press. “Loanword typology: Steps toward a systematic cross-linguistic study of lexical borrowability” to appear in Stolz, T. D. Bakker and R. Palomo (eds.) *Aspects of language contact: New theoretical, methodological and empirical findings with special focus on Romanisation processes*. Berlin: Mouton de Gruyter.

Haspelmath, Martin. 2004a. “On directionality in language change with particular reference to grammaticalization.” in Fischer, O., M. Norde and H. Peridon (eds.) *Up and down the cline: The nature of grammaticalization*: 17-44. Amsterdam: Benjamins.

Haugen, E. 1950. “The analysis of linguistic borrowing” in *Language*, 26: 210-31.

Hauser, M. D., N. Chomsky and W. T. Fitch. 2002. “The Faculty of Language: What Is It, Who Has It, and How Did It Evolve?” in *Science*, 298: 1569-1579.

Hawkins, J. 1983. *Word Order Universals*. New York: Academic Press.

Heeringa, W. 2004. *Measuring Dialect Pronunciation Differences using Levenshtein Distance*. PhD dissertation, Rijksuniversiteit, Groningen.

Heeringa, W. and J. Nerbonne. 2006. “De analyse van taalvariatie in het Nederlandse dialectgebied: methoden en resultaten op basis van lexicon en uitspraak” in *Nederlandse Taalkunde*, 11(3): 218-257.

Heggarty, Paul. 2004. “Interdisciplinary indiscipline? Can phylogenetic methods meaningfully be applied to language data – and to dating language?” in *Phylogenetic Methods and the Prehistory of Languages*. Cambridge: McDonald Institute for Archaeological Research.

Heine, B. 1974. “Historical linguistics and lexicostatistics in Africa” in *Journal of African Linguistics*, 11(3): 7–20.

Heine, B. and T. Kuteva. 2006. *The changing languages of Europe*. Oxford: Oxford University Press.

Herder, J. G. 1772. *Abhandlung über der Ursprung der Sprache*. Berlin: C. F. Voss.

Higgins, D. 2003. “Multiple alignment. Theory” in Salemi, M and Vandamme A.-M. (eds.) *The Phylogenetic Handbook*: 45-60. Cambridge: Cambridge University Press.

Higgins, D. and M. Salemi. 2003. “Multiple alignment. Practice” in Salemi, M and Vandamme A.-M. (eds.) *The Phylogenetic Handbook*: 61-71. Cambridge: Cambridge University Press.

Hinnebusch, T. J. 1996. “Skewing in lexicostatistic tables as an indicator of contact” paper presented at the *Round Table on Bantu Historical Linguistics. Universitè Lumière 2, Lyon, France, May 30–June 1, 1996*.

Hinnebusch, T. J. 1999. “Contact and lexicostatistics in comparative Bantu studies” in Hombert, J.-M. and L. M. Hyman (eds.) *Bantu Historical Linguistics: Theoretical and Empirical Perspectives*: 173–205. Stanford, CA: CSLI Publications.

Hirszfeld, L. and H. Hirszfeld. 1919. “Essai d’application des methods au problème des races” in *Anthropologie*, 29 : 505-537.

Hockett, C. F. “The origin of Speech” in *Scientific American*, 203: 88-96.

Hoijer, H. 1956. “Lexicostatistics: a critique” in *Language*, 32: 49-60.

Hopper, P. J. and E. C. Traugott. 1993. *Grammaticalization*. Cambridge: Cambridge University Press.

Howe, C. J., A. C. Barbrook, M. Spencer, P. Robinson, B. Bordalejo and L. R. Mooney. 2001. “Manuscript evolution” in *Trends in Genetics*, 17: 147-52.

Huang, J. 1982. *Logical Relations in Chinese and the Theory of Grammar*. Ph.D. dissertation, MIT.

Hübschmann, J. H. 1875. "Über die Stellung des Armenischen im Kreise der indogermanischen Sprachen" in *Zeitschrift für vergleichende Sprachforschung auf dem Gebiete der Indogermanischen Sprachen*, 23(1): 5-49.

Huelsenbeck, J.P., F. Ronquist, R. Nielsen and J.P. Bollback, 2001. "Bayesian inference of phylogeny and its impact on evolutionary biology" in *Science*, 294: 2310 – 2314.

Hull, D. L. 1988. *Science as a progress: an evolutionary account of the social and conceptual development of science*. Chicago: University of Chicago Press.

Hurford, J. 1987. *Language and Number*. Oxford: Blackwell.

Huson, D. H. 1998. "SplitsTree: Analyzing and Visualizing Evolutionary Data" in *Bioinformatics*, 14: 68–73.

Hyams, N. 1986. *Language acquisition and the theory of parameters*. Dordrecht: Kluwer.

Jaccard, P. 1901. "Étude comparative de la distribution florale dans une portion des Alpes et des Jura" in *Bulletin del la Société Vaudoise des Sciences Naturelles*, 37: 547-579.

Jackendoff, R. 1972. *Semantics and Cognition*. Cambridge, MA: MIT Press.

Jackendoff, R. 1977. *X-bar Syntax: A Study of Phrase Structure*. Cambridge, MA: MIT Press.

Jakobson, R. 1941. *Child Language, Aphasia, and Phonological Universals*. The Hague: Mouton de Gruyter.

Jeffers, R. J. 1976. "Syntactic change and syntactic reconstruction" in Christie, W. (ed.) *Current Progress in Historical Linguistics (ICHL 2)*: 1-16. Amsterdam: North Holland.

Jespersen, O. 1922. *Language: its Nature, Development and Origin*. London: Allen & Unwin.

Joseph, B. D. 1983. *The Synchrony and Diachrony of Balkan Infinitive: A Study in Areal, General and Historical Linguistics*. Cambridge: Cambridge University Press.

Kaiser, M. and V. Shevoroshkin. 1988. "Nostratic" in *Annual Review of Anthopology*, 17: 309-329.

Kayne, R. 2005. "Some notes on comparative syntax, with special reference to English and French" in Cinque, G. and R. Kayne (eds.) *The Oxford Handbook of Comparative Syntax*: 3-69. New York: Oxford University Press.

Keenan, E. 1994. *Creating anaphors. An historical study of the English reflexive pronouns*. Manuscript, UCLA.

Keenan, E. 2002. "Explaining the Creation of Reflexive Pronouns in English" in Minkova, D. and R. Stockwell (eds.) *Studies in the history of English: A millennia Perspective*: 325-355. Berlin: Mouton de Gruyter.

Keenan, E. 2007. *Mathematical Structures in Language*. Los Angeles: UCLA Academic Publishing.

Keller, R. 1994. *On Language Change: The Invisible Hand in Language*. London: Routledge.

Kimura, M. and G. H. Weiss. 1964. "The stepping-stone model of population structure and the decrease of genetic correlation with distance" in *Genetics*, 49: 561-576.

King, G. 2003. *Modern Welsh: a comprehensive grammar*. New York: Routledge.

King, R. 2000. *The Lexical Basis of Grammatical Borrowing: A Prince Edward Island French Case Study*. Amsterdam-Philadelphia: John Benjmains Publishing.

King, R. and Y. Roberge. 1990. "Preposition stranding in Prince Edward Island French" in *Probus*, 2(3): 351-169.

King, R. D. 1969. *Historical linguistics and generative grammar*. Englewood Cliffs NJ: Prentice Hall.

Kiparsky, P. 1985. "Some consequences of Lexical Phonology" in *Phonology Yearbook*, 2: 82-138.

Kiparsky, P. 1995. "The Phonological Basis of Sound Change" in Goldsmith, J. (ed.) *Handbook of Phonological Theory*: 640-670. Oxford: Blackwell.

Kiparsky, P. In press. "Universals constrain change; change results in typological generalizations" in Good, J. (ed.) *Language Universals and Language Change*. Oxford: Oxford University Press.

Kirby, S. 1998. "Fitness and the selective adaptation of language" in Hurford, J. R., M. Studdert-Kennedy and C. Knight (eds.) *Approaches to the Evolution of Language: Social and Cognitive bases*: 359-383. Cambridge: Cambridge University Press.

Klima, E. S. 1964. "Relatedness between grammatical systems" in *Language*, 40: 1-20.

Kluge, A. G. and J. S. Farris. 1969. "Quantitative phyletics and the evolution of anurans" in *Systematic Zoology*, 18: 1–32.

Kondrak, G. 2002. "Determining Recurrent Sound Correspondences by Inducing Translation Models" in *Proceedings of the Nineteenth International Conference on Computational Linguistics (COLING 2002), Taipei, August 2002*: 488-494.

Koppelmann, H. 1928. *Die Verwandtschaft des Koreanischen und der Ainu-Sprache mit den Indogermanischen Sprachen*. Wien.

Koppelmann, H. 1933. *Die Eurasische Sprachfamilie. Indogermanisch, Koreanisch und Verwandtes*. Heidelberg: Karl Winter.

Kroch, A. 1989. "Function and grammar in the history of English: Periphrastic 'do'" in Fasold, R. (ed.) *Language change and variation*: 133-172. Amsterdam: Benjamins.

Kroch, A. 2000. "Verb-complement order in Middle English" in Pintzuk, S., G. Tsoulas and A. Warner (eds.) *Diachronic Syntax: Models and Mechanisms*: 132-163. Oxford: Oxford University Press.

Kuhner, M. K. J. Y. and J. Felsenstein. 1994. "Maximum likelihood estimation of recombination rates from population data" in *Genetics*, 156: 439-447.

Kurylowicz, J. 1965. "Zur Vorgeschichte des germanischen Verbalsystems" in *Beiträge zur Sprachwissenschaft, Volkskunde und Literaturforschung: Wolfgang Steinitz zum 60. Geburtstag*: 242-247. Berlin: Akademie-Verlag.

Labov, W. 1966. *The Social Stratification of English in New York City*. Washington DC: Center for Applied Linguistics.

Labov, W. 1977. *Il continuo e il discreto nel linguaggio*. Bologna: Il Mulino.

Lakoff, R. 1968. *Abstract syntax and Latin complementation*. Cambridge MA: MIT Press.

Lamarck, J.-B. 1808. *Philosophie zoologique ou exposition des considérations relatives à l'histoire naturelle des animaux*. Paris : Dentu.

Langacker, R. W. 1977. WSyntactic Reanalysis" in Li, C. N. (ed.) *Mechanisms of Syntactic Change*: 59-139. Austin: University of Texas Press.

LaPolla, R. J. 2000. "Subgrouping in Tibeto-Burman. Can an individual-identifying standard be developed? How do we factor in the history of migrations and language contact?" Paper presented at the *33rd International Conference on Sino-Tibetan Languages and Linguistics, Bangkok, October 2-6, 2000.*

Lass, R. 1990. "How to Do Things with Junk: Exaptation in Language Evolution" in *Journal of Linguistics*, 26: 79-102.

Lass, R. 1997. *Historical linguistics and language change*. Cambridge: Cambridge University Press.

Lazzeroni, R. 1996. "Il mutamento linguistico" in Lazzeroni, R. (ed.) *Linguistica storica*: 13-54. Roma: La Nuova Italia Scientifica.

Lazzeroni, R. 2006. "I percorsi del mutamento linguistico. Un'ipotesi" in Bombi, R., G. Cifoletti, F. Fusco, L. Innocente, V. Orioles (eds.) *Studi linguistici in onore di Roberto Gusmani*: 975-983. Alessandria: Edizioni dell'Orso.

Le Quesne, W. J. 1969. "A method of selection of characters in numerical taxonomy" in *Systematic Zoology*, 18: 201-205.

Le Quesne, W. J. 1974. "The uniquely evolved character concept and its cladistic application" in *Systematic Zoology*, 23 : 513-517.

Lees, R. 1953. "The basis of glottochronology" in *Language*, 29: 113-127.

Lefebvre, C. 1985. "Grammaires en contact: définition et perspectives de recherche" in *Revue québécoise de Linguistique*, 14: 11-47.

Legate, J. and C. Yang. 2002. "Empirical re-assessment of stimulus poverty arguments" in *The Linguistic Review*, 19: 151-162.

Lehmann, C. 1982. *Thoughts on grammaticalization: a programmatic sketch*. Cologne: Institut für Sprachwissenschaft der Universität.

Lehmann, W. P. 1973. "A structural principle of language and its implications" in *Language*, 49: 47-66.

Lightfoot, D. 1979. *Principles of Diachronic Syntax*. Cambridge: Cambridge University Press.

Lightfoot, D. 1988. "Syntactic Change" in Newmeyer, F. (ed.) *Linguistics: the Cambridge survey, vol. 1: Linguistic theory: foundations*: 303-323. Cambridge: Cambridge University Press.

Lightfoot, D. 1989. "The Child's Trigger Experience: Degree-0 Learnability" in *Behavioural and Brain Science*, 12: 321-334.

Lightfoot, D. 1991. *How to Set Parameters: Arguments for Language Change*. Cambridge, MA: MIT Press.

Lightfoot, D. 1999. *The Development of Language: Acquisition, Change, and Evolution*. Oxford: Blackwell Publishing.

Lightfoot, David W. 2002. "Myths and the prehistory of grammar" in *Journal of Linguistics*, 38: 113-136.

Lindblom, B. 1986. "Phonetic universals in vowel systems" in Ohala, J. and J. J. Jäger (eds.) *Experimental Phonology*: 13-44. Orlando, FL: Academic Press.

Lockwood, W. B. 1972. *A panorama of Indo-European languages*. London: Hutchinson.

Lohr, M. 1999. *Methods for the Genetic Classification of Languages*. Ph.D. dissertation, University of Cambridge.

Longobardi, G. 1994. "Reference and Proper Names: A Theory of N-movement in Syntax and Logical Form" in *Linguistic Inquiry*, 25(4): 609-655.

Longobardi, G. 1996. "The Syntax of N-raising: a minimalist theory" in *UiL OTS Working Papers in Theoretical Linguistics*, 96(5). University of Utrecht: Utrecht Institute of Linguistics.

Longobardi, G. 1999. "Presentazione" in *Le scienze*, 18: 2-6.

Longobardi, G. 2001a. "Formal Syntax, Diachronic Minimalism, and Etymology: The History of French Chez" in *Linguistic Inquiry*, 32(2): 275-302.

Longobardi, G. 2001b. "The Structure of DPs: Some Principles, Parameters and Problems" in Baltin, M. and C. Collins (eds.) *The Handbook of Contemporary Syntactic Theory*: 562-603., Cambridge, MA - Oxford, UK: Blackwell.

Longobardi, G. 2003a. "Methods in Parametric Linguistics and Cognitive History" in *Linguistic Variation Yearbook*, 3: 101-138.

Longobardi, G. 2003b. "La sintassi storica" in Mancini, M. (ed.) *Il cambiamento linguistico*: 165-207. Roma: Carocci editore.

Longobardi, G. 2005. "A Minimalist Program for Parametric Linguistics?" in Broekhuis, H., N. Corver, M. Everaert and J. Koster (eds.) *Organising Grammar: A Festschrift for Henk van Riemsdijk*: 407-414. Berlin: Mouton de Gruyter.

Longobardi, G. 2008. "Reference to Individuals, Person, and the Variety of Mapping Parameters" in Klinge, A. and H. Müller (eds.) *Essays on Nominal Determination*: 189–211. Philadelphia-Amsterdam: John Benjamins.

Longobardi, G. and C. Guardiano. 2005. "Parametric Comparison and Language Taxonomy" in Batllori, M., M. L. Hernanz, C. Picallo, and F. Roca (eds.) *Grammaticalization and Parametric Variation*: 149-174. Oxford: Oxford University Press.

Longobardi, G. and C. Guardiano. 2009. "Evidence for Syntax as a Signal of Historical Relatedness" in Lingua, doi:10.1016/j.lingua.2008.09.012.

Longobardi, G. and G. Rigon. 2008. "Review of 'Syntactic Change - A minimalist approach to grammaticalization (Roberts and Roussou 2003)'" in *Language*, 84(2): 428-433.

Macé, C. and P. V. Baret. 2006. "Why phylogenetic methods work: The theory of evolution and textual criticism" in Macé, C., P. V. Baret, A. Bozzi, L. Cignoni (eds.) *The Evolution of Texts. Confronting Stemmatological and Genetical Methods. Proceedings of the International Workshop held in Louvain-la-Neuve on September 1-2, 2004 (Linguistica 24-25)*: 89-108. Pisa-Roma: Istituti Editoriali e Poligrafici Internazionali.

Maddieson, I. 1984. *Patterns of sound*. Cambridge: Cambridge University Press.

Malécot, G. 1959. "Quelques schemas probabilistes sur la variabilité des populations naturelles" in *Annales de l'Université de Lyon Sciences, Section A*, 13: 37-60.

Mallinson, G. and B. J. Blake. 1981. *Language typology: cross-linguistic studies in syntax*. Amsterdam: North Holland.

Manly, B. F. J. 2005. *Multivariate Statistical Methods. A primer*. Boca Raton – London – New York – Washington, D.C.: Chapman & Hall/CRC.

Manzini, M.-R. and K. Wexler. 1987. "Parameters, Binding theory and Learnability" in *Linguistic Inquiry*, 18: 413-444.

Margush, T. and F. R. McMorris. 1981. "Consensus n-trees" in *Bulletin of Mathematical Biology*, 43: 239-244.

Marotta, G. 1993. "Dental stops in Latin: A special class" in *Rivista di linguistica*, 2(5): 55-101.

Martinet, A. 1994. *L'indoeuropeao. Lingue, popoli e culture*. Roma-Bari: Laterza.

Maynard Smith, J. 1987. "How to model evolution" in Dupré, J. (ed.) *The Latest on the Best*: 119-131. Cambridge, MA: MIT Press.

Maynard Smith, J. 1989. *Evolutionary genetics*. New York: Oxford University Press.

McConvell, P. 2001. “Language shift and language spread among hunter-gatherers” in Panter-Brick, C., R. H. Layton and P. Rowley-Conwy (eds.) *Hunter-Gatherers: An interdisciplinary perspective*: 143–69. Cambridge: Cambridge University Press.

McMahon, A. 1994. *Understanding Language Change*. Cambridge - New York: Cambridge University Press.

McMahon, A. 2004. “Language, time and human histories” in Brisard, F, S. d’Hondt, and T. Mortelmans (eds.) *Language and revolution/Language and time*: 155-70. Antwerp: Univesity of Antwerp.

McMahon, A. and R. McMahon. 1995. “Linguistics, genetics and archaeology: internal and external evidence in the Amerind controversy” in *Transactions of the Philological Society*, 93: 125-225.

MacMahon, A. and R. MacMahon. 2003. “Finding families: Quantitative methods in language classification” in *Transaction of the Philological Society*, 101(1): 7-55.

McMahon, A. and R. McMahon. 2005. *Language Classification by Numbers*. Oxford: Oxford University Press.

McNeill, D. 1966. “Developmental Psycholinguistics” in Smith, F. and G. Miller (eds.) *The Genesis of Language*: 15-84. Cambridge, MA: MIT Press.

Meid, W. 1975. “Probleme der räumlichen und zeitlichen Gliederung des Indogermanischen” in Rix, H. (ed.) *Flexion und Wortbildung. Akten der V. Fachtagung der Indogermanischen Gesellschaft Regensburg*: 204-218. Wiesbaden.

Meillet, A. 1912. “L’évolution des formes grammaticales” in *Scientia*, 12 : 384-400.

Meillet, A. 1921. *Linguistique historique et linguistique générale* (= Collection linguistique publiée par la Société de Linguistique de Paris 8, 11). Paris.

Meillet, A. 1937. *Introduction à l'étude comparative des langues indo-européennes*. Paris: Hachette.

Metropolis, N., A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller and E. Teller, 1953. “Equations of state calculations by fast computing machines” in *Journal of Chemical Physics*, 21: 1087–1091.

Michener, C. D. and R. R. Sokal. 1957. “A quantitative approach to a problem in classification” in *Evolution*, 11: 130-162.

Minett, J. W. and W. S.-Y. Wang. 2003. “On detecting borrowing: distance-based and character-based approaches” in *Diachronica*, 20(2): 289-330.

Miranda, R. 1978. “On the Reconstructability of Syntactic Change” in Dressler, W., W. Meid, O. E. Pfeiffer, T. Herok (eds.) *Proceedings of the twelfth international congress of linguists, August 28 – September 2 1977, Wien: 504-507.* Innsbruck: Universität Innsbruck.

Møller, H. 1907. *Semitisch und Indogermanisch: Konsonanten*. Kopenhagen.

Moore, J. H. 1994. “Putting anthropology back together again: the ethnographic critique of cladistic theory” in *American Anthropologist*, 96(4): 925-948.

Moro, A. 2006. *I confine di Babele*. Milano: Longanesi.

Mourant, A. E. 1954. *The Distribution of Human Blood Groups*. Oxford: Blackwell Scientific.

Mufwene, S. 2001. *The ecology of language evolution*. Cambridge: Cambridge University Press.

Mufwene, S. S. 2000. “Language contact, evolution, and death: how ecology rolls the dice” in Kindell, G. and M. P. Lewis (eds.) Assessing ethnolinguistic vitality: theory and practice; selected papers from the Third International Language assessment Conference: 39-64. Dallas: SIL International.

Mufwene, S. S. 2002. “Competition and selection in language evolution” in Selection, 3(1): 45-56.

Mufwene, S. S. 2008. Language evolution: Contact, Competition and Change. New York: Continuum Press.

Myers-Scotton, C. 2002. Language contact: Bilingual encounters and grammatical outcomes. Oxford: Oxford University Press.

Nakhleh, L., D. Ringe and T. Warnow. 2005a. “Perfect Phylogenetic Networks: A New Methodology for Reconstructing the Evolutionary History of Natural Languages” in *Language*, 81(2):382-420.

Nakhleh, L., T. Warnow, D. Ringe, and S. N. Evans. 2005b. “A Comparison of Phylogenetic Reconstruction Methods on an IE Dataset” in *Transactions of the Philological Society*, 3(2): 171-192.

Narrog, H. 2007. “Exaptation, Grammaticalization and Reanalysis” in *California Linguistics Notes*, 32 (1).

Nei, M. 1987. *Molecular Evolutionary Genetics*. New York: Columbia University Press.

Nerbonne, J. 2007. Part II of “Review of ‘Language Classification by the Numbers’ (April McMahon and Robert McMahon)” in *Linguistic Typology*, 11: 425-436.

Nerbonne, J. 2009a. “Data-Driven dialectology” in *Language and Linguistics Compass*, 3(1): 175-198.

Nerbonne, J. 2009b. “Introduction” in Lingua , doi:10.1016/j.lingua.2008.09.010.

Nerbonne, J. and P. Kleiweg. 2007. “Toward a Dialectological Yardstick” in *Journal of Quantitative Linguistics*, 14(2): 148-167.

Nerbonne, J., W. Heeringa and P. Kleiweg. 1999. “Edit Distance and Dialect Proximity” in Sankoff, D. and J. Kruskal (eds.) *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*: V-XV. Stanford, CA: CSLI.

Nerbonne, P. and P. Kleiweg. 2007. “Toward a Dialectological Yardstick” in *Journal of Quantitative Linguistics*, 14(2): 148-167.

Newman, P. 1980. *The classification of Chadic wihin Afro-Asiatic*. Leiden: Universitair Pers.

Newmeyer, F. 2001. “Deconstructing grammaticalization” in *Language Sciences*, 23: 187-230.

Newmeyer, F. 2005. *Possible and Probable Languages: A Generative Perspective on Linguistic typology*. Oxford: Oxford University Press.

Newmeyer, F. 2005. *Possible and Probable Languages: A Generative Perspective on Linguistic typology*. Oxford: Oxford University Press.

Nichols, J. 1996. “The Comparative Method as an Heuristic” in Durie, M. and M. Ross (eds.) *The Comparative Method Reviewed: Regularity and Irregularity in Language Change*: 39-71. New York: Oxford University Press.

Nichols, J. 2006. “Quasi-cognates and Lexical Type Shifts: Rigorous Distance Measures for Long-range Comparison” in Forster, P. and C. Renfrew (eds.) *Phylogenetic Methods and the Prehistory of Languages*: 57-66. Cambridge: McDonald Institute for Archaeological Research.

Niyogi, P. 2002. “Theories of cultural evolution and their application to language evolution” in Briscoe, T. (ed.) *Linguistic Evolution through Language Acquisition: Formal and Computational Models*: 205-234. Cambridge: Cambridge University Press.

Niyogi, P. 2006. *The Computational Nature of Language Learning and Evolution*. Cambridge, MA: MIT Press.

Niyogi, P. and R. C. Berwick. 1995. *The Logical Problem of Language Change*. Technical report, MIT.

Niyogi, P. and R. C. Berwick. 1997. “Evolutionary Consequences of Language Learning” in *Linguistics and Philosophy*, 20(6): 697-719.

Niyogi, P. and R. C. Berwick. 1998. “The Logical Problem of Language Change: A Case Study of European Portuguese” in *Syntax: A Journal of Theoretical, Experimental, and Interdisciplinary Research*, 1(2): 192-205.

Noonan, M. 2006. “Contact-induced change in the Himalayas: the case of the Tamangic languages”. Paper presented at the *International Colloquium on Language Contact and Contact Languages, University of Hamburg, July 6-8, 2006.*

Norde, M. 2002. “The final stages of grammaticalization: Affixhood and beyond.” in Wischer, I. and G. Diewald (eds.) *New reflections on grammaticalization*: 45-65. Amsterdam: Benjamins.

Numbers, R. L. 1987. “Creationists” in *Zygon*, 22(2): 133-164.

O’Keefe, F. R. and P. J. Wagner. 2001. “Inferring and Testing Hypotheses of Cladistic Character Dependence by Using Character Compatibility” in *Systematic Biology*, 50(5): 657-675.

Ogura, M. and W. S.-Y. Wang. 1996. “Lexical diffusion and evolution theory” in Hickey, R. and S. Puppel (eds.) *Trends in Linguistics: Studies and Monographs 101. Language History and Linguistic Modelling: A Festschrift for Jacek Fisiak on his 60th Birthday, Vol. 1*: 1083–1098. Berlin: Mouton de Gruyter.

Osgood, C. and T. Sebeok. 1954. Psycholinguistic: A survey of theory and research problems” in *Journal of Abnormal and Social Psychology*, 49(4): 1-203.

Osthoff, H. and K. Brugmann. 1878. *Morphologische Untersuchungen auf dem Gebiete der indogermanischen Sprachen*. Leipzig: S. Hirzel.

Pagel, M. and A. Maede. 2006. “Estimating Rates of Lexical Replacement on Phylogenetic Trees of Languages” in Forster, P. and C. Renfrew (eds.) *Phylogenetic Methods and the Prehistory of Languages*: 173-182. Cambridge: McDonald Institute for Archaeological Research.

Pagel, M., 2000. “Maximum-likelihood models for glottochronology and for reconstructing linguistic phylogenies” in Renfrew, C., A. McMahon and L. Trask. (eds.) *Time Depth in Historical Linguistics*: 413-439. Cambridge: The McDonald Institute for Archaeological Research.

Paul, H. 1920 [reprint of the 1898 edition]. *Prinzipien der Sprachgeschichte*. Tübingen: Max Niemeyer.

Pedersen, H. 1903. "Türkische Lautgesetze" in *Zeitschrift der Deutschen Morgenländischen Gesellschaft*, 57: 535-561.

Pedersen, H. 1931. *Linguistic Science in the Nineteenth Century: Methods and Results*. Cambridge, MA: Harvard University Press.

Piatelli-Palmarini, M. and Uriagereka, J. 2004. "The immune syntax: the evolution of the language virus" in Lyle Jenkins, L. (ed.) *Variation and Universals in Linguistics*: 341-377. Oxford: Elsevier.

Picallo, C. 1991. "Nominals and nominalization in Catalan" in *Probus*, 3: 279-316.

Po-Ching, Y. and D. Rimmington. 2004. *Chinese: A Comprehensive Grammar*. London – New York: Routledge.

Pollock, J.-Y. 1989. "Verb movement, universal grammar, and the structure of IP" in *Linguistic Inquiry*, 20: 365-424.

Poussa, P. 1982. "The Evolution of Early Standard English: The Creolization Hypothesis" in *Studia Anglica Posnaniesia*, 14: 69-85.

Prince, E. F. 1988. "On pragmatic change: the borrowing of discourse functions" in *Journal of Pragmatics*, 12: 505-518.

Prince, E. F., and S. Pintzuk. 2000. "Bilingual code-switching and the open/closed class distinction" in *University of Pennsylvania Working Papers in Linguistics*, 6(3): 237-57.

Pucher, M., Andreas Türk, J. Ajmera and N. Fecher. 2007. "Phonetic Distance Measures for Speech Recognition Vocabulary and Grammar Optimization" in *Proceedings of the 3rd congress of the Alps Adria Acoustics Association, Graz, Austria.*

Pullum, G. and B. Scholtz. 2002. "Empirical assessment of Stimulus Poverty Argument" in *The Linguistic Review*, 19: 9-50.

Radford, A. 2002. *Syntax. A minimalist introduction*. Cambridge: Cambridge University Press.

Rayfield, J. R. 1970. *The languages of a bilingual community*. The Hague: Mouton.

Renfrew, C. 1987. *Archaeology and language: The puzzle of Indo-European origins*. Cambridge – New York: Cambridge University Press.

Renfrew, C. 1991. "Before Babel: Speculations on the origins of linguistic diversity" in *Cambridge Archaeological Journal*, 1(1): 3-23.

Renfrew, C. 1992. "Archaeology, genetics and linguistic diversity" in *Man*, 27: 445-478.

Renfrew, C. 2008. "Language, Genes and Archaeology: problems with the New Synthesis", paper presented at the *Max Planck Institute for Evolutionary Anthropology, Leipzig, 8th October 2008*.

Renfrew, C., D. Nettle (eds.). 1999. *Nostratic: examining a linguistic macrofamily*. Cambridge: McDonald Institute for Archaeological Research.

Ridley, M. 1993. *Evolution*. Boston: Blackwell Scientific Publications.

Rigon, G. 2007. *Towards the automatic detection of syntactic borrowing on a parametric dataset*. Technical report, Università di Pisa.

Ringe, D. 1996. "The mathematics of Amerind" in *Diachronica*, 13: 135-154.

Ringe, D. A. 1992. "On calculating the factor of chance in language comparison" in *Transactions of the American Philosophical Society*, 82(1): 1-110.

Ringe, D., T. Warnow and A. Taylor. 2002. "Indo-European and computational cladistics" in *Transactions of the Philological Society*, 100(1): 59-129.

Ritter, E. 1991. "Two functional categories in noun phrases: evidence from Modern Hebrew" in Rothstein, S. (ed.) *Syntax and Semantics*: 37-62. San Diego: Academic Press.

Rizzi, L. 1982. *Issues in Italian syntax*. Dordrecht: Foris.

Roberts, I. 1998. "Review of 'Historical syntax in cross-linguistic perspective' (Harris and Campbell 1995)" in *Romance Philology*, 51: 363-370.

Roberts, I. 2001. "Head movement" in Baltin, M. R. and C. Collins (eds.) *Handbook of Contemporary Syntactic Theory*: 113-147. Oxford: Blackwell Publishers.

Roberts, I. 2007. *Diachronic Syntax*. Oxford: Oxford University Press.

Roberts, I. 2008. "Macroparameters, Markedness and Typological Drift" presented at *Levels of Analysis in the History of Indo-European Languages, May 22nd-24th 2008, University of Trieste*.

Roberts, I. and A. Holmberg. 2005. "On the Role of Parameters in Universal Grammar: A reply to Newmeyer" in Broekhuis, H., N. Corver, M. Everaert and J. Koster (eds.) *Organising Grammar: A Festschrift for Henk van Riemsdijk*: 538-553. Berlin: Mouton de Gruyter.

Roberts, I. and A. Roussou. 2003. *Syntactic change. A minimalist approach to grammaticalization*. Cambridge: Cambridge University Press.

Ross, M. 1996. "Contact-induced change and the comparative method: cases from Papua New Guinea" in Durie, M. and M. Ross (eds.) *The comparative method reviewed: Regularity and irregularity in language change*: 180-217. Oxford: Oxford University Press.

Ross, M. 1999. *Exploring metatypy: how does contact-induced typological change come about?* Manuscript, Australian National University.

Ross, M. 2001. "Contact-induced change in Oceanic Languages in north-west Melanesia" in Dixon, R. M. W. and A. Y. Aikhenvald (eds.) *Areal diffusion and genetic inheritance*: 134-166. Oxford: Oxford University Press.

Ross, M. and M. Durie. 1996. "Introduction" in Durie, M. and M. Ross (eds.) *The Comparative Method Reviewed: Regularity and Irregularity in Language Change*: 3-38. New York: Oxford University Press.

Roy, M.-M. 1979. *Les conjonctions anglaises "but" et "so" dans le français de Moncton*. M. A. dissertation, Università du Québec à Montréal.

Ruhlen, M. 1987. *A Guide to the World's Languages*. Stanford, CA: Stanford University Press.

Ruhlen, M. 1994. *A Guide to the World's Languages. Volume 1: Classification*. Stanford, CA: Stanford University Press.

Ruhlen, M. 1994. *On the origin of languages*. Stanford, CA: Stanford University Press.

Ryder, R. J. 2006. *Grammar and Phylogenies*. Manuscript. Department of Statistics, University of Oxford.

Sagart, L. and S. Xu. 2001. "History through loanwords: the loan correspondences between Hani and Chinese" in *Cahiers de Linguistique — Asie Orientale*, 30(1): 3–54.

Saitou N. and M. Nei. 1987. "The neighbor-joining method: A new method for reconstructing evolutionary trees" in *Molecular Biology and Evolution*, 4: 406-425.

Sakas, W. G. and J. D. Fodor. 2001. "The Structural Triggers Learner" in Bertolo, S. (ed.) *Language Acquisition and Learnability*: 172-233. Cambridge: Cambridge University Press.

Salmons, J. C. and B. D. Joseph (eds.). 1998. *Nostratic. Sifting the evidence*. Amsterdam: John Benjamins.

Sankoff, D. 1972. "Reconstructing the history and geography of an evolutionary tree" in *American Mathematical Monthly*, 79: 596–603.

Sankoff, G. 2001. "Linguistic Outcomes of Language Contact" in Trudgill, Peter, J. Chambers and N. Schilling-Estes (eds.) *Handbook of Sociolinguistics*: 638-668. Oxford: Basil Blackwell.

Sapir, E. 1921. *Language*. New York: Harcourt Brace & World.

Sapir, E. 1927. *Language. An introduction to the study of speech*. New York: Harcourt, Brace and company.

Scaglione, A. D. 1972. *The Classical Theory of Composition: From Its Origins to the Present: a Historical Survey*. Chapel Hill: University of North Carolina Press.

Schleicher, A. 1853. "Die ersten Spaltungen des indogermanischen Urvolkes" in *Allgemeine Zeitung für Wissenschaft und Literatur*, August 1853.

Schmidt, J. 1872. *Die Verwandtschaftsverhältnisse der Indogermanischen Sprachen*. Weimar: H. Böhlau.

Schuchardt, H. 1928. *Hugo Schuchardt-Brevier: Ein Vademecum der allgemeinen Sprachwissenschaft*. Spitzer, L. (ed.). Halle/Saale: Niemeyer.

Semple, C and M. A. Steel. 2003. *Phylogenetics*. Oxford: Oxford University Press.

Smith, N. 2004. *Chomsky: Ideas and Ideals*. Cambridge: Cambridge University Press.

Snyder, W. 2001. "On the nature of syntactic variation: evidence from complex predicates and complex word-formation" in *Language*, 77: 324-342.

Sokal, R. R. and P. H. A. Sneath. 1963. *Numerical taxonomy*. San Francisco: W. H. Freeman.

Spencer, M., E. A. Davidson, A. C. Barbrook, C. J. Howe. 2004. "Phylogenetics of artificial Manuscripts" in *Journal of Theoretical Biology*, 227(4): 503-511.

Spruit, M. R. 2005. "Classifying Dutch dialects using a syntactic measure" in *Linguistics in the Netherlands 2005*: 166-179.

Spruit, M. R. 2008. *Quantitative perspectives on syntactic variation in Dutch dialects*. Utrecht: LOT.

Stankiewicz, E. 1974. “The dithyramb to the verb in eighteenth and nineteenth century linguistics” in Hymes, D. (ed.) *Studies in the history of linguistics: traditions and paradigms*: 157-190. Bloomington: Indiana University Press.

Stolz, M. 2006. “Linking the variance. Unrooted trees and networks” in Macé, Caroline, Philippe V. Baret, Andrea Bozzi, Laura Cignoni (eds.) *The Evolution of Texts. Confronting Stemmatological and Genetical Methods. Proceedings of the International Workshop held in Louvain-la-Neuve on September 1-2, 2004 (Linguistica 24-25)*: 193-216. Pisa-Roma: Istituti Editoriali e Poligrafici Internazionali.

Stowell, T. 1981. *Origins of phrase structure*. Ph.D. dissertation, MIT.

Stowell, T. 1982. “Conditions on reanalysis” in *MIT Working Papers in Linguistics*, 4: 245-269.

Stowell, T. 1989. “Subjects, specifiers, and X-bar Theory” in Baltin, M. and T. Kroch (eds.) *Alternative Conceptions of Phrase Structure*: 232-262. Chicago: Chicago University Press.

Stowell, T. 1991. “Determiners in NP and DP” in Leffel, K. and D. Bouchard (eds.) *Views on Phrase Structure*: 37-56. Dordrecht: Kluwer Academic Publishers.

Strimmer, K and A. von Haeseler. 2003. “Nucleotide substitution models“ in Salemi, M and Vandamme A.-M. (eds.) *The Phylogenetic Handbook*: 72-87. Cambridge: Cambridge University Press.

Studier J. A. and K. J. Keppler. 1988. “A note on the neighbor-joining algorithm of Saitou and Nei” in *Molecular Biology and Evolution*, 5: 729-731.

Sugisaki, Koji and Snyder, William 2002. “Preposition stranding and the compounding parameter: a developmental perspective” in Skarabela, Barbora et al. (eds.). *Proceedings of the 26th Annual Boston University Conference on Language Development*. Somerville: Cascadilla Press.

Swadesh, M. 1952. “Lexico-Statistic Dating of Prehistoric Ethnic Contact” in *Proceedings of the American Philosophical Society*, 96:453–463.

Swadesh, M. 1955. “Towards greater accuracy in lexicostatistic dating” in *International Journal of American Linguistics*, 21: 121-137.

Swadesh, M. 1971. *The Origin and Diversification of Language*. Chicago - New York: Aldine - Atherton.

Swofford, D. L. 1998. *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods).* Version 4. Sunderland, MA: Sinauer Associates.

Syvanen, M. 1985. “Cross-species gene transfer; Implications for a New Theory of Evolution” in *Journal of theoretical Biology*, 112: 333-343.

Syvanen, M. 1994. “Horizontal Gene Transfer: Evidence and Possible Consequences” in *Annuual Review of Genetics*, 28: 237-261.

Szabolcsi, A. 1983. “The possessor that ran away from home” in *The Linguistic Review*, 3: 89-102.

Szemerényi, O. 1885. *Introduzione alla linguistica indoeuropea*. Milano: Edizioni Unicopli.

Tagliavini, C. 1999. *Le origini delle lingue neolatine*. Bologna: Pàtron Editore.

Taylor, A., A. Warner, S. Pintzuk and F. Beths. 2003. *The York-Toronto-Helsinki Parsed Corpus of Old English Prose (YCOE).* Department of Language and Linguistic Science, University of York.

Teeter, K. 1963. “Lexicostatistics and genetic relationship” in *Language*, 39: 638-648.

Terrell, J. 1988. “History as a family tree, history as an entangled bank: constructing images and interpretations of prehistory in the South Pacific” in *Antiquity*, 62: 642-657.

Terrell, J., Hunt, T. L. and Gosden, C. 1997. “The dimensions of social life in the Pacific” in *Current Anthropology*, 38: 155-195.

Terrell, J., Kelly, K.M. and Rainbird, P. 2001. “Foregone conclusions? In search of 'Austronesians' and 'Papuans'” in *Current Anthropology*, 42: 97-124.

Thomas, A. 1994. “English in Wales” in Burchfield, R. (ed.) *The Cambridge History of the English Language - Volume V: English in Britain and Overseas, Origins and Development*: 94-147. Cambridge: Cambridge University Press.

Thomason, S. 2004. *Syntactic reconstruction*. Manuscript, University of Michigan.

Thomason, S. and T. Kaufman. 1988. *Language Contact, Creolization, and Genetic Linguistics*. Berkeley: University of California Press.

Thomason, S. G. 2001. *Can rules be borrowed?* Manuscript, University of Michigan.

Thomason, S. G. 2007. “Language contact and deliberate change” in *Journal of Language Contact, THEMA 1*: 41-62.

Torgerson, W. 1952. “Multidimensional scaling: I. Theory and method” in *Psychometrika*, 17: 401-419.

Traugott, E. and R. Dasher. 2002. *Regularity in semantic change*. Oxford: Oxford University Press.

Traugott, E. C. 1965. “Diachronic syntax and generative grammar” in *Language*, 41: 402-415.

Traugott, E. C. 1980. “Meaning-change in the development of grammatical markers” in *Language Science*, 2: 44-61.

Traugott, E. C. 2005. “Exaptation and grammaticalization” in Akimoto, M. (ed.) *Linguistic Studies Based on Corpora*: 133-156. Tokyo: Hituzi Syobo Publishing Co.

Travis, L. 1984. *Parameters and effects of word-order variation*. Ph.D. dissertation, MIT.

Trips, C. 2002. *From OV to VO in Early Middle English*. Amsterdam: Benjamins.

Turkel, W. J. “The learning guided evolution of natural language” in Briscoe, T. (ed.) *Linguistic Evolution through Language Acquisition: Formal and Computational Models*: 235-254. Cambridge: Cambridge University Press.

Turkel, W. J. 2002. “The learning guided evolution of natural language” in Briscoe, T. (ed.) *Linguistic Evolution through Language Acquisition: Formal and Computational Models*: 235-254. Cambridge: Cambridge University Press.

Underhill, P., G. Passarino, A. A. Lin, P. Shen, M. Mirazón Lahr, R. A. Foley, P. J. Oefner and L. L. Cavalli-Sforza. 2001. “The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations” in *Annals of human genetics*, 65: 43-62.

Valois, D. 1991. *The Internal Syntax of DP*. Ph.D. dissertation. University of California Los Angeles.

van der Auwera, Johan. 2002. “More thoughts on degrammaticalization” in Wischer, I. and G. Diewald (eds.) *New reflections on grammaticalization*: 19-29. Amsterdam: Benjamins.

van Kemenade, A. 1987. *Syntactic Change and Morphological Change in the History of English*. Dordrecht: Foris.

Vennemann, T. 1973. “Explanation in syntax” in Kimball, J. (ed.) *Syntax and semantics, vol. 1*: 1-50. New York: Academic Press.

Villar, F. 1997. *Gli indoeuropei e l’origine dell’Europa*. Bologna: Il Mulino.

Wackernagel, J. 1892. “Über ein Gesetz der indogermanischen Wortstellung“ in *Indogermanische Forschungen*, 1: 133-436.

Wang, F. 2004. *Language Contact and Language Comparison — the case of Bai*. Ph.D. dissertation, City University of Hong Kong.

Wang, F. and W. S-Y Wang. 2004. “Basic Words and Language Evolution” in *Language and linguistics*, 5(3): 643-662.

Wang, W. S.-Y. and J. W. Minett. 2005. “Vertical and horizontal transmission in language evolution” in *Transactions of the Philological Society*, 103(2): 121–146.

Wang, W. S.-Y., 1989. “The migration of the Chinese people and the settlement of Taiwan”, in Chang, K., K. Li, A. P. Wolf and A. C. Yin (eds.) *Anthropological Studies of the Taiwan Area: Accomplishments and Prospects*. Taipei: Department of Anthropology, National Taiwan University.

Watkins, C. 2001. “An Indo-European linguistic area and its characteristics: Ancient Anatolia. Areal diffusion as a challenge to the comparative method?” in Aikhenvald, A. Y. and R. M.W. Dixon (eds.) *Areal diffusion and genetic inheritance*: 44-63.. Oxford: Oxford University Press.

Weaver, R. and P. Hendrick. 1997. *Genetics*. Dubuque, IA: William C. Brown Publishing.

Weinreich, U. 1979. *Languages in Contact. Findings and Problems*. The Hauge: Mouton.

Weinreich, U., W. Labov and M. I. Herzog. 1968. “Empirical foundations for a theory of languge change” in Lehmann, W. and Y. Malkiel (eds.) *Directions for historical linguistics*: 97-195. Austin: University of Texas Press.

Wen, Y. 1940. “A study of synonyms in the Min-chia language” in *Anthology of the Research Institute of Chinese Culture, Huaxi Xiehe University*, 1(1): 1–27.

Weninger, S. 2001. “Vom Altätiopischen zu den neuäthiopischen Sprachen” in Haspelmath, M., H. König, W. Oesterreicher and W. Raible (eds.) *Language typology and language universals – Volume II*: 1762-1774. Berlin: de Gruyter.

Whinnom, K. 1971. “Linguistic hybridization and the ‘special case’ of pidgins and creoles” in Humes, D. (ed.) *Pigdization and creolization of languages*: 91-115. Cambridge: Cambridge University Press.

Wilkins, D. 1996. “Natural Tendencies of Semantic Change and the Search for Cognates” in Durie, M. and M. Ross (eds.) *The Comparative Method Reviewed: Regularity and Irregularity in Language Change*: 264-306. New York: Oxford University Press.

William, H. C. and D. Bickerton. 2000. *Lingua ex Machina: Reconciling Darwin and Chomsky with the human brain*. Cambridge, MA: MIT Press.

Williams, G. C. 1966. *Adaptation and Natural Selection*. Princeton: Princeton University Press.

Windram, F. H., M. Spencer, C. J. Howe. 2006. “Phylogenetic analysis of manuscript traditions, and the problem of contamination” in Macé, C., P. V. Baret, A. Bozzi, L. Cignoni (eds.) *The Evolution of Texts. Confronting Stemmatological and Genetical Methods. Proceedings of the International Workshop held in Louvain-la-Neuve on September 1-2, 2004 (Linguistica 24-25)*: 141-156. Pisa-Roma: Istituti Editoriali e Poligrafici Internazionali.

Winter, W. 1984. “Reconstuctional comparative linguistics and the reconstruction of syntax of undocumented stages in the development of languages and language families” in Fisiak, J. (ed.) *Historical Syntax*: 613-626. Berlin: Walter de Gruyter & Co.

Witham, L. A. 2005. *Where Darwin Meets the Bible. Creationists and Evolutionists in America*. Oxford, MS: Oxford University Press.

Wright, R. 1991. “Quest for the mother language” in *Atlantic Monthly*, April: 39-68.

Wu, C. F. J. 1986. “Jackknife, bootstrap and other resampling plans in regression analysis” in *Annals of Statistics*, 14: 1261-1295.

Xia, X. 2001. *Data Analysis in Molecular Biology and Evolution*. Norwell, MA: Kluwer Acamemic Publishers.

Zubizarreta, M. L. and J.-R. Vergnaud. 1992 “The Definite Determiner and the inalienable Construction in French and English” in *Linguistic Inquiry*, 23(4): 595-652.

Appendix

Chapter 2: Language evolution and biological evolution

<table>
<tr><th>Case 1</th><th>Case 2</th><th>Case 3</th></tr>
<tr>
<td>
<table>
<tr><td></td><td>*G*</td><td>*G*</td></tr>
<tr><td>*G*</td><td>*GG*</td><td>*GG*</td></tr>
<tr><td>*G*</td><td>*GG*</td><td>*GG*</td></tr>
</table>
GG = 4/4 (100%)
Gg = 0/4 (0%)
gg = 0/4 (0%)
</td>
<td>
<table>
<tr><td></td><td>*G*</td><td>*G*</td></tr>
<tr><td>*G*</td><td>*GG*</td><td>*GG*</td></tr>
<tr><td>*g*</td><td>*Gg*</td><td>*Gg*</td></tr>
</table>
GG = 2/4 (50%)
Gg = 2/4 (50%)
gg = 0/4 (0%)
</td>
<td>
<table>
<tr><td></td><td>*G*</td><td>*g*</td></tr>
<tr><td>*G*</td><td>*GG*</td><td>*Gg*</td></tr>
<tr><td>*g*</td><td>*Gg*</td><td>*gg*</td></tr>
</table>
GG = 1/4 (25%)
Gg = 2/4 (50%)
gg = 1/4 (25%)
</td>
</tr>
</table>

Table 1: Mendelian ratios.

Generalized theory of selection	**Paradigm instantiation of selection in biology**	**Paradigm instantiation of selection in language**
replicator	gene	lingueme
structured set of replicators	string of DNA	utterance
normal replication	reproduction by e.g. interbreeding	utterance production in communication
altered replication	recombination, mutation of genes	mechanisms for innovation
alternative replicators	alleles	variants
interactor	organism	speaker (including grammar)
environment	ecological environment	social-communicative context
selection	survival and reproduction of organisms	entrenchment of convention by speakers and its propagation in communication

Table 2: Overview of Croft's Theory of Utterance Selection (adapted from Croft 2000: 38).

Biological evolution	
Relevant units	• Genes.
Factors	Mutation: • Often harmful for the organism; • Random; • Gradual accumulation; • Diversification effect between populations.
	Selection: • Mainly natural selection, related to adaptation.
	Migration: • Relevant only between populations, i.e. interbreeding groups; • Only demic; • Homogenizing effect between populations.
	Genetic drift: • Random; • Homogenizing effect within a population.
Individual transmission	• Vertical (in higher animals and plants). • Mendelian inheritance model.
Cultural evolution	
Relevant units	• "Cultural traits/units".
Factors	Cultural change: • Often directed and intentional; • Gradual accumulation ; • Diversification effect between cultures.
	Selection: • Social selection, related to individual/social acceptance; • Social selection is often followed by natural selection.
	Migration: • Always relevant, i.e. there are no "interbreeding barriers"; • Demic or cultural; • Homogenizing effect between different cultures.
	Genetic drift: • Random; • Homogenizing effect within a population.
Individual transmission	• Vertical, horizontal (oblique), one-to-many or many to one; • No general inheritance model.
Language evolution	
Relevant units	• "Linguistic traits".
Factors	Innovation: • Neither "harmful" nor "advantageous"; • Random (at least for traits included in the "narrow linguistic faculty"); • Gradual accumulation; • Diversification effect between languages.
	Selection: • Structural pre-selection; • Marginal natural selection (only for some aspects of lexical-semantics); • Social selection.
	Language contact: • Always relevant (different effects); • Complex contact typology; • Homogenizing effect between different languages.
	Genetic drift: • Random; • Homogenizing effect within a language.
Individual transmission	• Vertical, horizontal (oblique), one-to-many or many to one; • No general inheritance model.

Table 3: Overview of the relevant factors and features of biological evolution, cultural evolution and language evolution.

Chapter 3: Syntactic evolution

Syntactic evolution	
Relevant units	• Parameters.
Factors	Syntactic change: • Determined by diachronic parameter resetting; • Parameter resetting is an indirect consequence of syntax-external changes (either morpho-phonological or semantic) or of other syntactic changes (Inertial Theory) • Neither “harmful” nor “advantageous”; • Punctual/abrupt at I-language level; • Gradual accumulation at E-language level; • Memoryless; • Diversification effect between I- and, ultimately, E-languages.
	Selection: • Structural pre-selection due to the limits of parametric variation; • Social selection.
	Syntactic borrowing: • Always possible; • At least binary typology (direct/indirect contact); • Homogenizing effect between different syntactic systems.
	Genetic drift: • Random; • Homogenizing effect within a language.
Individual transmission	• Vertical, horizontal (oblique), one-to-many or many to one; • Acquisition-driven.

Table 1: Overview of the relevant factors and features of syntactic evolution.

→ Consider the following distance matrix (the same given in 7b)

	L1	L2	L3	L4
L1	0	1	1.3	1.4
L2	1	0	1.3	1.4
L3	1.3	1.3	0	0.5
L4	1.4	1.4	0.5	0

1. From the original distance matrix, NJ calculates the values of u_i for each taxonomic unit i:

$$u_{L1} = u_{L2} = (1 + 1.3 + 1.4)/2 = 1.85$$
$$u_{L3} = (1.3 + 1.3 + 0.5)/2 = 1.55$$
$$u_{L4} = (1.4 + 1.4 + 0.5)/2 = 1.65$$

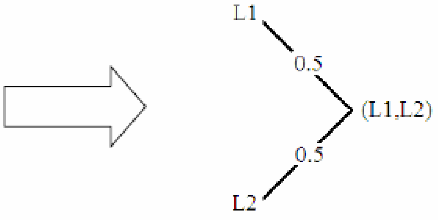
2. A new matrix containing the values ($D_{ij} - u_i - u_j$) is drawn:

	L1	L2	L3	L4
L1	0	-2.7	-2.1	-2.1
L2	-2.7	0	-2.1	-2.1
L3	-2.1	-2.1	0	-2.7
L4	-2.1	-2.1	-2.7	0

In this case, both for (*L1*, *L2*) and (*L3*, *L4*) the value ($D_{ij} - u_i - u_j$) is the smallest, thus either of them can be chosen. Suppose to choose (*L1*, *L2*).

3. The branch length from *L1* to the internal node (*L1*, *L2*) and from *L2* to the internal node (*L1*, *L2*) are calculated as:

$$v_{L1} = (0.5 \times 1) + 0.5 \times (1.85 - 1.85) = 0.5$$
$$v_{L2} = (0.5 \times 1) + 0.5 \times (1.85 - 1.85) = 0.5$$

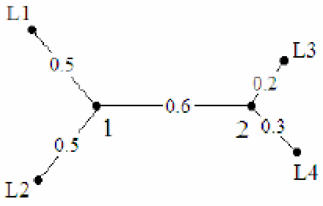


4. The distance between the new node (*L1*, *L2*) and the remaining taxonomic units is calculated and a new distance matrix is drawn:

	(L1,L2)	L3	L4
(L1,L2)	0	0.8	0.9
L3	0.8	0	0.5
L4	0.9	0.5	0

5. The algorithm goes back to step 1 until only two nodes remain to be joined.

→ When the procedure is complete, the tree inferred by the NJ algorithm is exactly the following:



Sketch 1: First operations carried out by the NJ algorithm on the distance matrix of figure 7b.

→ Consider the following distance matrix (the same given in 7b)

	L1	L2	L3	L4
L1	0	1	1.3	1.4
L2	1	0	1.3	1.4
L3	1.3	1.3	0	0.5
L4	1.4	1.4	0.5	0

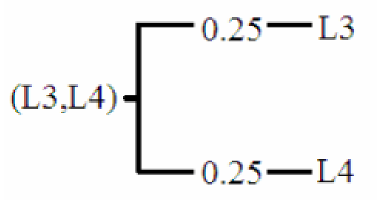
1. From the original distance matrix, UPGMA takes into account *L3* and *L4*, which are the taxonomic units whose distance is the smallest (0.5)

2. A new group is created, (*L3, L4*), which contains 2 members.

3. The branch length from *L3* to the internal node (*L3,L4*) and from *L4* to the internal node (*L3,L4*) are calculated as:

$l_{L3,(L3,L4)} = 0.5 / 2 = 0.25$
$l_{L4,(L3,L4)} = 0.5 / 2 = 0.25$

⇨



4/5. The distance between the new node (*L3, L4*) and the remaining taxonomic units is calculated and a new distance matrix is drawn:

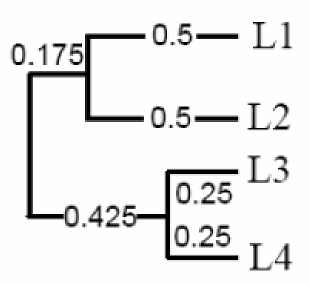
$D_{(L3,L4),L1} = 1/2 \times 1.3 + 1/2 \times 1.4 = 1.35$
$D_{(L3,L4),L2} = 1/2 \times 1.3 + 1/2 \times 1.4 = 1.35$

⇨

	L1	L2	(L3,L4)
L1	0	1	1.35
L2	1	0	1.35
(L3,L4)	1.35	1.35	0

5. The algorithm goes back to step 1 until only one node remains.

→ When the procedure is complete, the tree inferred by the UPGMA algorithm is the following:



Sketch 2: First operations carried out by the UPGMA algorithm on the distance matrix of figure 7b.

Meaning	**"here"**				**"sea"**				**"water"**			**"when"**
Cognate set	1	2	3	4	5	6	7	8	9	10	11	12
English	1	0	0	0	1	0	0	0	1	0	0	1
German	1	0	0	0	1	1	0	0	1	0	0	1
French	0	1	0	0	0	1	0	0	0	1	0	1
Italian	0	1	0	0	0	1	0	0	0	1	0	1
Greek	0	0	1	0	0	0	1	0	0	0	1	1
Hittite	0	0	0	1	0	0	0	1	1	0	0	1

Table 1: An example of the binary matrix used by Gray and Atkinson (2003; 2006) to encode Dyen *et alii* (1992) lexical data (adapted from Gray and Atkinson 2006: 94).

Language	**Group**	**P16**	**(…)**	**M8**	**(…)**	**"grind"**
Hittite	Anatolian	1	(…)	3	(…)	1
Luvian	Anatolian	1	(…)	9	(…)	8
Lycian	Anatolian	1	(…)	10	(…)	9
Old Indic (Early Vedic)	Indo-Iranian	1	(…)	1	(…)	5
Avestan	Indo-Iranian	1	(…)	1	(…)	6
Old Persian	Indo-Iranian	1	(…)	1	(…)	11
Classic Attic Greek	Greek	1	(…)	1	(…)	2
Umbrian	Italic	1	(…)	2	(…)	1
Oscan	Italic	1	(…)	2	(…)	13
Latin	Italic	1	(…)	2	(…)	1
Gothic	Germanic	2	(…)	1	(…)	1
Old High German	Germanic	2	(…)	1	(…)	1
Old English	Germanic	2	(…)	1	(…)	7
Old Norse	Germanic	2	(…)	1	(…)	1
Classical Armenian	Armenian	1	(…)	4	(…)	2
Old Irish	Celtic	1	(…)	2	(…)	1
(Modern) Welsh	Celtic	1	(…)	2	(…)	1
Tocharian A	Tocharian	1	(…)	6	(…)	10
Tocharian B	Tocharian	1	(…)	11	(…)	4
Old Church Slavonic	Balto-Slavic	1	(…)	7	(…)	1
Old Prussian	Balto-Slavic	1	(…)	12	(…)	12
(Modern) Lithuanian	Balto-Slavic	1	(…)	8	(…)	1
(Modern) Latvian	Balto-Slavic	1	(…)	13	(…)	1
(Modern) Albanian	Albanian	1	(…)	5	(…)	3

Table 2: A sample of Ringe *et alii*'s (2002) data, including phonological (P16), morphological (M8) and lexical characters ("grind") (adapted from Ringe *et alii* 2002: 113-125).

	Phon. characters	**Morph. characters**	**Lex. characters**
Direction of innovation	Known	Unknown	Unknown
Parallel development	Probable	Improbable	Probable
Borrowing	-	-	Yes

Table 3: Summary of Ringe *et alii*'s (2002) evaluation of characters.

	Lunteren	**Velhoven**	**distance**
zich	+	+	0
hem	-	-	0
zijn eigen	+	-	1
zichzelf	-	-	0
hemzelf	-	-	0
			Total: 1

Table 4: Sample of the character matrix, containing atomic variables, used by Spruit (2008), including the calculation of the distance value between two dialects (adapted from Spruit 2008: 49).

	personal	**reflexive**	**possessive**	**owness**	**focus**
	hem	*zich*	*zijin*	*eigen*	*zelf*
hem	+				
hemzelf	+				+
zich		+			
zichzelf		+			+
zijn			+		
zijn zelf			+		+
zijn eigen			+	+	
zijn eigen zelf			+	+	+

Table 5: Mapping of atomic variables to feature variables proposed in Spruit (2008: 56).

	Lunteren	**Veldhofen**	**distance**
	{zich, zijin eigen}	{zich}	
personal	-	-	0
reflexive	+	+	0
possessive	+	-	1
ownness	+	-	1
focus	-	-	0
			Total: 2

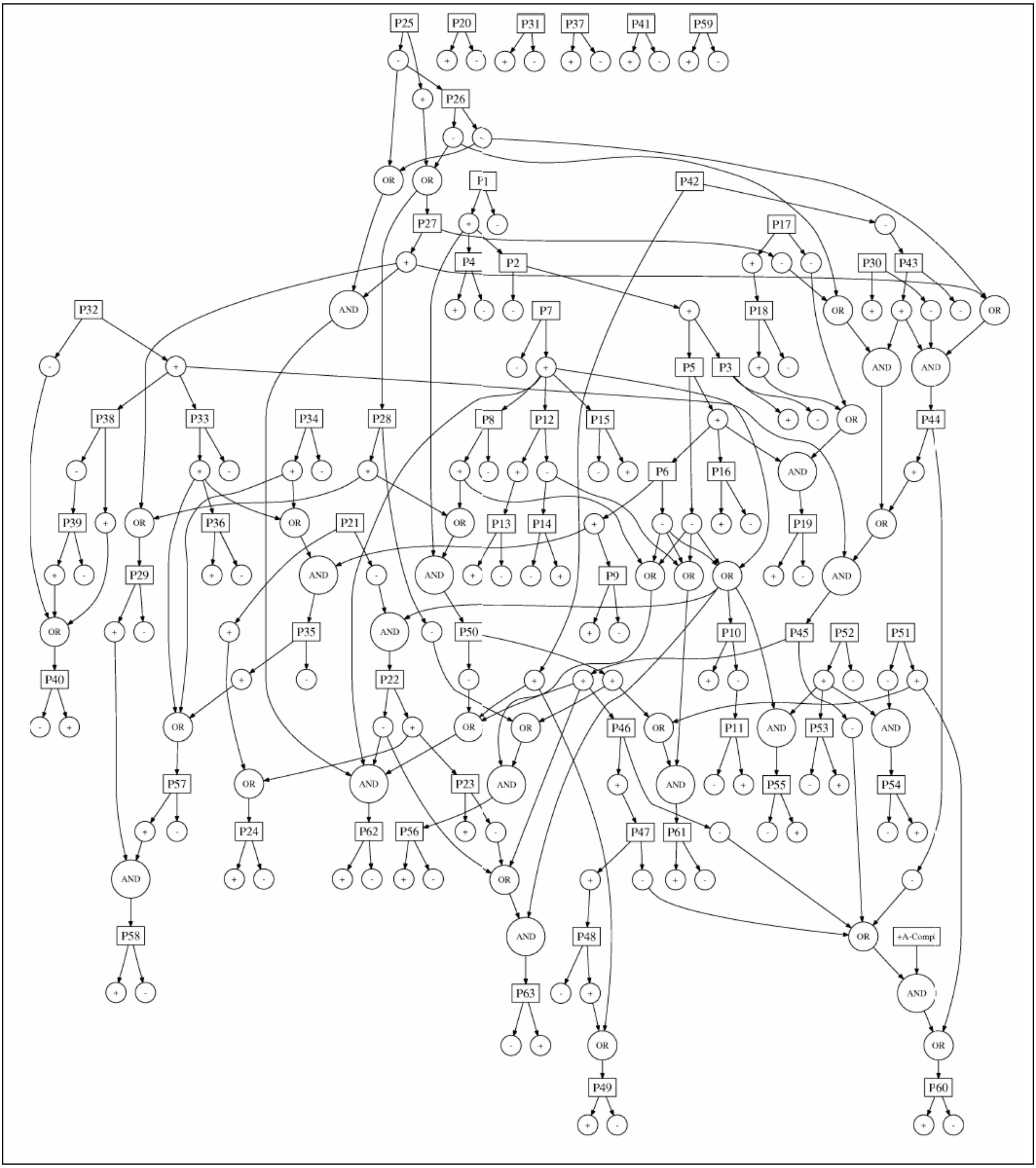
Table 6: Sample of the character matrix, containing feature variables, used by Spruit (2008), including the calculation of the distance value between two dialects (adapted from Spruit 2008: 57).

Number	**Parameter**	**Number**	**Parameter**
1	± gramm. person	33	± feature spread to structured APs +32
2	± gramm. number +1	34	± feature spread to pred. APs
3	± gramm. gender +2	35	± numb. on A +6, +33 or +34
4	± variable person on D +1	36	± D-controlled infl. on A +33
5	± feature spread to N +2	37	± DemP over relatives
6	± numb. on N (BNs) +5	38	± free APs in Mod +32
7	± gramm. partial def	39	± APs in Mod -38
8	± gramm. def +7	40	± overt Mod° -32 or +38 or +39
9	± free null partitive Q +6	41	± adjectival Gen
10	± gramm. dist. art. -5 or -6 or +7	42	± N-raising with pied-piping
11	± gramm. top. art. -10	43	± N over ext.arg. -42
12	± def checking N +7	44	± N over GenO +26 or +27, -30 ,+43
13	± def spread to N +12	45	± N over As +32, (-26 or -27, +43) or +44
14	± def on attrib. -12	46	± N over M2 As +45
15	± def on relatives +7	47	± N over M1 As +46
16	± D-controlled infl. on N +5	48	± N over high As +47
17	± gramm. cardinal nouns	49	± N over cardinals +42 or +48
18	± gramm. cardinal adjectives +17	50	± strong D (person) +1, +8 or +28
19	± plural spread from cardinals +5, -17 or +18	51	± NP over D
20	± gramm. mass-to-count	52	± strong deixis
21	± N-to-predicate incorporation	53	± strong anaphoricity +52
22	± gramm. partial count -5 or -6 or +7, -21	54	± DP over Dem -51, +52
23	± gramm. count +22	55	± D checking Dem -5 or -6 or +7, +52
24	± count checking N +21 or +22	56	± D-checking poss -5 or -6 or +8, +50 or -28
25	± prepositional Gen	57	± feature spread on poss +35 or +34 or +33
26	± free inflected Gen -25	58	± feature spread on postp. Gen +29, +57
27	± GenO +25 or -26	59	± enclitic possessives
28	± GenS +25 or -26	60	± Consisten. Princ. +51 or (-44 or … or -47, +A-Compl)
29	± postpositional Gen +27 or +28	61	± null-N-licensing art -5 or-6 or -12, +50 or +51
30	± Gen over DemP	62	± obl. def. inh. +7,-22, (-25, +26) or +27,+42 or +45 or -50
31	± poss checking N	63	± gramm. geogr. art. -5 or-6 or +7, -22 or -23 or +45
32	± structured APs		

Table 7: List of the parameters included in Table A.

	It	Sal	Sp	Fr	Ptg	Rum	Lat	ClG	NTG	Gri	Grk	Got	OE	E	D	Nor	Blg	SC	Rus	Ir	Wel	Heb	Ar	Wo	Hu	Fin	Hi	Bas
p1	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
p2	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
p3	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	+	-
p4	-	-	+	-	-	-	?	+	?	-	+	?	?	-	-	-	-	-	-	?	?	-	+	-	-	-	-	+
p5	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	-
p6	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	0	+	+	+	0
p7	+	+	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+	-	-	+	+	+	+	+	+	-	-	-
p8	+	+	+	+	+	+	0	+	+	+	+	-	+	+	+	+	+	0	0	+	+	+	+	+	+	0	0	0
p9	-	-	-	0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	-	+	-	0
p10	-	-	-	-	-	-	0	-	-	-	-	-	-	-	-	-	-	0	0	-	-	-	-	+	-	0	0	+
p11	-	+	-	-	-	-	0	-	-	-	-	-	-	-	-	-	-	0	0	-	-	-	-	0	-	0	0	0
p12	-	-	-	-	-	+	0	-	-	-	-	-	-	-	-	+	+	0	0	-	-	-	-	-	-	0	0	0
p13	0	0	0	0	0	-	0	0	0	0	0	0	0	0	0	+	-	0	0	0	0	0	0	0	0	0	0	0
p14	-	-	-	-	-	0	0	+	+	-	+	+	-	-	-	0	0	0	0	-	-	+	+	-	-	0	0	0
p15	-	-	-	-	-	-	0	-	-	-	-	-	-	-	-	-	-	0	0	-	-	-	+	+	-	0	0	0
p16	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	0	-	-	-	0
p17	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	?	+	+	?	+	+	+	?	-	+	-	-
p18	0	0	0	0	0	0	0	0	0	0	0	0	+	0	0	0	?	-	+	?	-	-	+	?	0	+	0	0
p19	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	?	0	+	?	0	0	+	0	-	-	+	0
p20	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	?	+	-	+	-	-	-	-	-
p21	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	+	-
p22	+	+	+	+	+	+	0	-	-	+	+	0	+	+	+	+	+	0	0	-	-	-	-	0	0	0	0	+
p23	+	+	+	+	+	+	0	0	0	+	+	0	-	+	+	+	-	0	0	0	0	0	0	0	0	0	0	-
p24	-	-	-	-	-	-	0	0	0	-	-	0	-	-	-	-	-	0	0	0	0	0	0	-	-	0	-	-
p25	+	+	+	+	+	-	-	-	-	-	-	-	-	+	+	+	+	-	-	+	+	+	+	-	-	-	-	-
p26	0	0	0	0	0	-	+	+	-	-	-	-	-	0	0	0	0	-	-	0	0	0	0	-	-	-	-	-
p27	-	-	-	-	-	-	0	0	+	+	+	+	+	-	+	?	-	+	+	+	+	-	-	+	-	+	+	+
p28	-	-	-	-	-	+	0	0	-	-	-	+	+	+	+	+	-	-	-	-	-	+	+	-	+	+	+	+
p29	0	0	0	0	0	-	0	0	-	-	-	-	-	+	-	+	0	-	-	-	-	-	-	-	-	-	+	+
p30	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	?
p31	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	-	-
p32	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+
p33	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+	+	?	+	+	0	-	+	+	-
p34	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	+	+	+	+	+	?	+	+	+	+	+	+	+
p35	+	+	+	0	+	+	+	+	+	+	+	+	+	0	+	+	+	+	+	+	?	+	+	0	+	+	+	0
p36	-	-	-	-	-	-	-	-	-	-	-	+	+	0	+	+	-	-	-	-	?	-	-	0	0	-	-	0
p37	+	+	+	+	+	+	?	?	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	?	-
p38	+	+	+	+	+	-	+	+	+	-	+	+	-	-	-	-	-	-	-	-	-	-	+	0	-	-	-	-
p39	0	0	0	0	0	-	0	0	0	-	0	0	+	+	-	-	+	+	+	-	-	+	0	0	-	-	?	?
p40	-	+	-	-	-	0	-	-	-	0	-	-	-	-	0	0	-	-	-	0	0	-	-	+	0	0	?	?
p41	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	-	-	-	-	-	-	-	?	-
p42	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-	+
p43	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	-	+	+	+	+	+	0	0	+	+	+	+	0
p44	0	0	0	0	0	0	-	-	+	+	+	+	+	0	+	0	0	+	+	+	+	0	0	+	0	-	0	0
p45	+	+	+	+	+	+	0	0	-	+	-	-	-	0	-	0	-	-	-	+	+	0	0	0	-	0	-	0
p46	+	+	+	+	+	+	0	0	0	+	0	0	0	0	0	0	0	0	0	+	+	0	0	0	0	0	0	0
p47	-	+	-	-	-	-	0	0	0	?	0	0	0	0	0	0	0	0	0	+	+	0	0	0	0	0	0	0
p48	0	+	0	0	0	0	0	0	0	?	0	0	0	0	0	0	0	0	0	-	-	0	0	0	0	0	0	0
p49	0	-	0	0	0	0	0	0	0	?	0	0	0	0	0	0	0	0	0	0	0	+	+	?	0	0	0	-
p50	+	+	+	+	+	+	0	+	+	+	+	0	-	-	-	-	+	0	0	-	-	+	+	-	?	0	0	+
p51	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	+
p52	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	+	+	+	+	+	+	+
p53	+	+	-	+	+	-	-	-	-	+	-	-	+	+	+	+	-	+	+	0	0	?	-	+	+	+	+	+
p54	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	0	+	?	0	-	-	-	0
p55	+	+	+	+	+	+	0	-	-	+	-	?	+	+	+	-	+	0	0	0	0	+	-	+	-	0	0	+
p56	-	-	+	+	?	-	0	-	-	-	-	0	0	0	0	0	-	0	0	+	+	-	-	+	?	0	0	-
p57	+	+	+	+	+	+	+	+	+	-	-	+	+	0	+	+	+	+	+	-	?	-	-	+	-	-	+	-
p58	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-	0	0	0	0	0	0	0	0	0	0	+	0
p59	-	-	-	-	-	-	-	+	+	+	+	-	-	-	-	-	?	-	-	-	-	-	-	?	-	-	-	-
p60	+	0	+	+	+	+	?	?	?	?	-	?	+	+	+	+	-	-	-	0	0	0	0	?	+	+	0	+
p61	-	-	+	-	+	0	0	+	+	-	-	0	0	0	0	0	0	0	0	0	0	-	-	?	?	0	0	+
p62	0	0	0	0	0	0	0	0	0	0	0	0	-	0	0	0	0	0	0	-	+	+	+	-	0	0	0	0
p63	+	+	-	+	-	+	0	+	+	+	0	0	0	0	0	0	-	0	0	?	?	-	-	0	-	0	0	-
	It	Sal	Sp	Fr	Ptg	Rum	Lat	ClG	NTG	Gri	Grk	Got	OE	E	D	Nor	Blg	SC	Rus	Ir	Wel	Heb	Ar	Wo	Hu	Fin	Hi	Bas

Table 8: Current version of Table A.



Sketch 1: Implicational structure(s) of Table A.

P	Implication	Ranking level	All languages						Modern languages				
			+	-	0	?	PS		+	-	0	?	PS
p1	-	1	28	0	0	0	C		23	0	0	0	C
p2	+1	2	28	0	0	0	C		23	0	0	0	C
p3	+2	3	25	3	0	0	I		20	3	0	0	I
p4	+1	2	5	17	0	6	I		4	17	0	2	I
p5	+2	3	26	2	0	0	I		21	2	0	0	I
p6	+5	4	25	1	2	0	U		20	1	2	0	U
p7	-	1	22	6	0	0	I		18	5	0	0	I
p8	+7	2	21	1	6	0	U		18	0	5	0	C
p9	+6	5	1	24	3	0	U		1	19	3	0	U
p10	-5 or -6 or +7	5	2	21	5	0	I		2	17	4	0	I
p11	-10	6	1	20	7	0	U		1	16	6	0	U
p12	+7	2	3	19	6	0	I		3	15	5	0	I
p13	+12	3	1	2	25	0	U		1	2	20	0	U
p14	-12	3	6	13	9	0	I		3	12	8	0	I
p15	+7	2	2	20	6	0	I		2	16	5	0	I
p16	+5	4	1	25	2	0	U		1	20	2	0	U
p17	-	1	7	18	0	3	I		6	14	0	3	I
p18	+17	2	4	3	18	3	I		3	3	14	3	I
p19	+5, -17 or +18	4	19	2	5	2	I		14	2	5	2	I
p20	-	1	2	25	0	1	I		2	20	0	1	I
p21	-	1	3	25	0	0	I		3	20	0	0	I
p22	-5 or -6 or +7, -21	5	14	6	8	0	I		13	4	6	0	I
p23	+22	6	11	3	14	0	I		11	2	10	0	I
p24	+21 or +22	6	0	17	11	0	C		0	16	7	0	C
p25	-	1	13	15	0	0	I		13	10	0	0	I
p26	-25	2	2	13	13	0	I		0	10	13	0	C
p27	+25 or -26	3	14	11	2	1	I		11	11	0	1	I
p28	+25 or -26	3	12	14	2	0	I		10	13	0	0	I
p29	+27 or +28	4	4	16	8	0	I		4	13	6	0	I
p30	-	1	1	26	0	1	U		1	21	0	1	U
p31	-	1	5	23	0	0	I		5	18	0	0	I
p32	-	1	27	1	0	0	U		22	1	0	0	U
p33	+32	2	23	3	1	1	I		18	3	1	1	I
p34	-	1	25	2	0	1	I		20	2	0	1	I
p35	+6, +33 or +34	5	23	0	4	1	C		18	0	4	1	C
p36	+33	3	4	19	4	1	I		2	16	4	1	I
p37	-	1	24	1	0	3	U		21	1	0	1	U
p38	+32	2	11	16	1	0	I		7	15	1	0	I
p39	-38	3	6	8	12	2	I		5	8	8	2	I
p40	-32 or +38 or +39	4	2	16	8	2	I		2	11	8	2	I
p41	-	1	3	24	0	1	I		3	19	0	1	I
p42	-	1	3	25	0	0	I		3	20	0	0	I
p43	-42	2	23	2	3	0	I		18	2	3	0	I
p44	+26 or +27, -30 ,+43	4	11	3	14	0	I		8	1	14	0	U
p45	+32, (-26 or -27, +43) or +44	5	9	10	9	0	I		9	7	7	0	I
p46	+45	6	9	0	19	0	C		9	0	14	0	C
p47	+46	7	3	5	19	1	I		3	5	14	1	I
p48	+47	8	1	2	24	1	U		1	2	19	1	U
p49	+42 or +48	9	2	2	22	2	I		2	2	17	2	I
p50	+1, +8 or +28	4	14	7	6	1	I		12	6	4	1	I
p51	-	1	2	26	0	0	I		2	21	0	0	I
p52	-	1	26	2	0	0	I		21	2	0	0	I
p53	+52	2	16	9	2	1	I		15	5	2	1	I
p54	-51, +52	2	1	22	4	1	U		1	17	4	1	U
p55	-5 or -6 or +7, +52	5	14	6	7	1	I		13	4	6	0	I
p56	-5 or -6 or +8, +50 or -28	5	5	11	10	2	I		5	9	7	2	I
p57	+35 or +34 or +33	6	18	8	1	1	I		13	8	1	1	I
p58	+29, +57	7	1	1	26	0	U		1	1	21	0	U
p59	-	1	4	22	0	2	I		2	19	0	2	I
p60	+51 or (-44 or … or -47, +A-Compl)	8	12	4	6	6	I		11	4	6	2	I
p61	-5 or-6 or -12, +50 or +51	5	5	7	14	2	I		3	7	11	2	I
p62	+7,-22, (-25, +26) or +27,+42 or +45 or -50	6	3	3	22	0	I		3	2	18	0	I
p63	-5 or-6 or +7, -22 or -23 or +45	7	7	7	12	2	I		5	7	9	2	I

Table 1: Summary of the character-state evaluation.

Parameter	**Split**
p1	-
p2	-
p3	(Hu, Fin, Bas) / others
p4	(Sp, ClG, Grk, Ar, Bas) / (It, Sal, Fr, Ptg, Rum, Gri, E, Fr, Nor, Blg, SC, Rus, Heb, Wo, Hu, Fin, Hi)
p5	(Wo, Bas) / others
p6	Fr / others
p7	(SC, Rus, Hu, Fin, Bas) / others
p8	(It, Sal, Sp, Fr, Ptg, Rum, ClG, NTG, Gri, Grk, OE, E, D, Nor, Blg, Ir, Wel, Heb, Ar, Wo, Hu) / Got
p9	(It, Sal, Sp, Ptg, Rum, Lat, ClG, NTG, Gri, Grk, Got, OE, E, D, Nor, Blg, SC, Rus, Ir, Wel, Heb, Ar, Hu, Hi) / Fin
p10	(Wo, Bas) / (It, Sal, Sp, Fr, Ptg, Rum, ClG, NTG, Gr, E, D, Nor, Blg, Ir, Wel, Heb, Ar, Hu)
p11	(It, Sp, Fr, Ptg, Rum, ClG, NTG, Gri, Grk, Got, OE, E, D, Nor, Blg, Ir, Wel, Heb, Ar, Hu) / Sal
p12	(It, Sal, Sp, Fr, Ptg, ClG, NTG, Gri, Grk, Got, OE, E, D, Ir, Wel, Heb, Ar, Wo, Hu) / (Rum, Nor, Blg)
p13	(Rum, Blg) / Nor
p14	(It, Sal, Sp, Fr, Ptg, Gri, OE, E, D, Ir, Wel, Wo, Hu) / (ClG, NTG, Grk, Got, Heb, Ar)
p15	(It, Sal, Sp, Fr, Ptg, Rom, ClG. NTG, Gri, Grk, Got, OE, Germ, Blg, Ir, Wel, Heb, Hu) / (Ar, Wo)
p16	(It, Sal, Sp, Fr, Ptg, Rum, Lat, ClG, NTG, Gri, Grk, Got, OE, E, D, Nor, Blg, SC, Rus, Ir, Wel, Heb, Ar, Hu, Fin, Hi) / Ar
p17	(It, Sal, Sp, Fr, Ptg, Rum, Lat, ClG, NTG, Gri, Grk, Got, E, D, Nor, Hu, Hi, Bas) / (OE, SC, Rus, Wel, Heb, Ar, Fin)
p18	(SC, Wel, Heb) / (Rus, Ar, Fin)
p19	(It, Sal, Sp, Fr, Ptg, Rum, Lat, ClG, NTG, Gri, Grk, Got, OE, E, D, Nor, Rus, Ar, Hu, Fin, Hi) / (Hi, Fin)
p20	(It, Sal, Sp, Fr, Ptg, Rum, Lat, ClG, NTG, Gri, Grk, Got, OE, E, D, Nor, Blg, SC, Rus, Heb, Wo, Hu, Fin, Hi, Bas) / (Wel, Ar)
p21	(Wo, Hu, Hi) / others
p22	(It, Sal, Sp, Fr, Ptg, Rum, ClG, NTG, Gri, Grk, OE, E, D, Nor, Blg, Ir, Wel, Heb, Ar, Bas) / (Ir, Wel, Heb, Ar)
p23	(It, Sal, Sp, Fr, Ptg, Rum, Gri, Grk, OE, E, D, Nor) / (Blg, Bas)
p24	(It, Sal, Sp, Fr, Ptg, Rum, Gri, Grk, OE, E, D, Nor, Blg, Wo, Hu, Hi, Bas)
p25	(It, Sal, Sp, Fr, Ptg, E, D, Nor, Blg, Ir, Wel, Heb, Ar) / (Rum, Lat, ClG, NTG, Gri, Grk, SC, Rus, Wo, Hu, Fin, Hi, Bas)
p26	(Rum, NTG, Gri, Grk, Got, OE, SC, Rus, Wo, Hu, Fin, Hi, Bas) / (Lat, ClG)
p27	(It, Sal, Sp, Fr, Ptg, Rum, NTG, E, Blg, Heb, Ar, Hu) / (NTG, Gri, Grk, D, SC, Rus, Ir, Wel, Fin, Hi, Bas)
p28	(It, Sal, Sp, Fr, Ptg, NTG, Gri, Grk, Blg, SC, Rus, Ir, Wel, Wo) / (Rum, Got, OE, E, D, Nor, Heb, Ar, Hu, Fin, Hi, Bas)
p29	(Rum, NTG, Gri, Grk, Got, OE, D, SC, Rus, Ir, We, Heb, Ar, Wo, Hu, Fin) / (E, Nor, Hi, Bas)
p30	(It, Sal, Sp, Fr, Ptg, Rum, Lat, ClG, NTG, Gri, Grk, Got, OE, E, D, Nor, Blg, SC, Rus, Ir, Wel, Heb, Ar, Wo, Hu, Fin) / Hi
p31	(Heb, Ar, Wo, Ur) / others
p32	Wo / others

Table 2 (to be continued): The set of splits determined by the parameters of Table A.

Parameter	Split
p33	(It, Sal, Sp, Fr, Ptg, Rum, Lat, ClG, NTG, Gri, Grk, Got, OE, D, Nor, Blg, SC, Rus, Ir, Heb, Ar, Fin, Hi) / (E, Hu, Bas)
p34	(It, Sal, Sp, Fr, Ptg, Rum, Lat, ClG, NTG, Gri, Grk, Got, OE, Nor, Blg, SC, Rus, Ir, Heb, Ar, Wo, Hu, Fin, Hi, Bas) / (E, D)
p35	(It, Sal, Sp, Ptg, Rum, Lat, ClG, NTG, Gri, Grk, Got, OE, D, Nor, Blg, SC, Rus, Ir, Heb, Ar, Hu, Fin, Hi)
p36	(It, Sal, Sp, Fr, Ptg, Rum, Lat, ClG, NTG, Gri, Grk, Blg, SC, Rus, Ir, Heb, Ar, Fin, Hi) / (Got, OE, D, Nor)
p37	(It, Sal, Sp, Fr, Ptg, Rum, NTG, Gri, Grk, Got, OE, E, D, Nor, Blg, SC, Rus, Ir, Wel, Heb, Ar, Wo, Hu, Fin) / Bas
p38	(It, Sal, Sp, Fr, Ptg, Lat, ClG, NTG, Grk, Got, Ar) / (Rum, Gri, OE, E, D, Nor, Blg, SC, Rus, Ir, Wel, Heb, Hu, Fin, Hi, Bas)
p39	(Rum, Gri, D, Nor, Ir, We, Hu, Fin) / (OE, E, Blg, SC, Rus, Heb)
p40	(It, Sp, Fr, Ptg, Lat, ClG, NTG, Gri, Got, OE, E, Blg, Sc, Rus, Heb, Ar) / (Sal, Wo)
p41	(It, Sal, Sp, Fr, Ptg, Rum, Lat, ClG, NTG, Gri, Grk, Got, OE, E, D, Nor, Ir, Wel, Heb, Ar, Wo, Hu, Fin, Bas) / (Blg, SC, Rus,)
p42	(It, Sal, Sp, Fr, Ptg, Rum, Lat, ClG, NTG, Gri, Grk, Got, OE, E, D, Nor, Blg, SC, Rus, Ir, Wel, Wo, Hu, Fin, Hi) / (Heb, Ar, Bas)
p43	(It, Sal, Sp, Fr, Ptg, Rum, Lat, ClG, NTG, Gri, Grk, Got, OE, D, Blg, SC, Rus, Ir, Wel, Wo, Hu, Fin, Hi)/ (E, Nor)
p44	(NTG, Gri, Grk, Got, OE, D, SC, Rus, Ir, Wel, Wo) / (Lat, ClG, Fin)
p45	(It, Sal, Sp, Fr, Ptg, Rum, Gri, Ir, Wel) / (NTG, Grk, Got, OE, D, Blg, SC, Rus, Hu, Hi)
p46	(It, Sal, Sp, Fr, Ptg, Rum, Gri, Ir, Wel)
p47	(It, Sp, Fr, Ptg, Rum) / (Sal, Ir, Wel)
p48	Sal / (Ir, Wel)
p49	(Sal, Bas) / (Heb, Ar)
p50	(It, Sal, Sp, Fr, Ptg, Rum, ClG, NTG, Gri, Grk, Blg, Heb, Ar, Bas) / (OE, E, D, Nor, Ir, Wel, Wo)
p51	(Wo, Bas) / others
p52	(Ir, Wel) / others
p53	(Sp, Rum, Lat, ClG, NTG, Grk, Got, Blg, Ar) / (It, Sal, Fr, Ptg, Gri, OE, E, D, Nor, SC, Rus, Wo, Hu, Fin, Hi, Bas)
p54	(It, Sal, Sp, Fr, Ptg, Rum, Lat, ClG, NTG, Gri, Grk, Got, OE, E, D, Nor, Blg, SC, Rus, Hu, Fin, Hi) / Heb
p55	(It, Sal, Sp, Fr, Ptg, Rum, Gri, OE, E, D, Blg, Heb, Wo, Bas) / (ClG, NTG, Grk, Nor, Ar, Hu)
p56	(It, Sal, Rum, ClG, NTG, Gri, Grk, Blg, Heb, Ar, Bas) / (Sp, Fr, Ir, Wel, Wo)
p57	(It, Sal, Sp, Fr, Ptg, Rum, Lat, ClG, NTG, Got, OE, D, Nor, Blg, SC, Rus, Wo, Hi) / (Gri, Grk, Ir, Heb, Ar, Hu, Fin, Bas)
p58	Nor / Hi
p59	(It, Sal, Sp, Fr, Ptg, Rum, Lat, Got, OE, E, D, Nor, SC, Rus, Ir, Wel, Heb, Ar, Hu, Fin, Hi, Bas) / (ClG, NTG, Gri, Grk)
p60	(It, Sp, Fr, Ptg, Rum, OE, E, D, Nor, Hu Fin, Bas) / (Grk, Blg, SC, Rus)
p61	(Sp, Ptg, ClG, NTG, Bas) / (It, Sal, Fr, Gri, Grk, Heb, Ar)
p62	(OE, Ir, Wo) / (Wel, Heb, Ar)
p63	(It, Sal, Fr, Rum, ClG, NTG, Gri) / (Sp, Ptg, Blg, Heb, Ar, Hu, Bas)

Table 2 (continued): The set of splits determined by the parameters of Table A.

	It	Sal	Sp	Fr	Ptg	Rum	Lat	ClG	NTG	Gri	Grk	Got	OE	E	D	Nor	Blg	SC	Rus	Ir	Wel	Heb	Ar	Wo	Hu	Fin	Hi	Bas
It	0	0.06	0.09	0.04	0.04	0.1	0.1	0.18	0.19	0.1	0.2	0.2	0.19	0.14	0.15	0.16	0.17	0.22	0.22	0.2	0.24	0.23	0.31	0.36	0.27	0.31	0.24	0.49
Sal	0.06	0	0.15	0.1	0.1	0.14	0.13	0.23	0.23	0.12	0.22	0.24	0.24	0.19	0.17	0.19	0.2	0.23	0.22	0.22	0.26	0.29	0.37	0.33	0.3	0.32	0.24	0.49
Sp	0.09	0.15	0	0.1	0.04	0.14	0.06	0.16	0.19	0.2	0.2	0.17	0.21	0.18	0.19	0.21	0.17	0.28	0.27	0.18	0.22	0.27	0.29	0.39	0.3	0.37	0.29	0.46
Fr	0.04	0.1	0.1	0	0.06	0.15	0.14	0.24	0.24	0.15	0.25	0.23	0.22	0.16	0.18	0.19	0.23	0.27	0.26	0.21	0.25	0.29	0.37	0.33	0.31	0.33	0.28	0.51
Ptg	0.04	0.1	0.04	0.06	0	0.12	0.1	0.19	0.2	0.14	0.22	0.2	0.19	0.14	0.15	0.16	0.16	0.22	0.22	0.18	0.22	0.23	0.32	0.34	0.25	0.31	0.24	0.44
Rum	0.1	0.14	0.14	0.15	0.12	0	0.1	0.17	0.17	0.12	0.2	0.15	0.19	0.21	0.17	0.17	0.17	0.24	0.23	0.24	0.29	0.23	0.33	0.39	0.24	0.24	0.19	0.42
Lat	0.1	0.13	0.06	0.14	0.1	0.1	0	0.06	0.12	0.22	0.15	0.12	0.21	0.25	0.23	0.2	0.14	0.19	0.18	0.23	0.3	0.29	0.28	0.46	0.33	0.28	0.17	0.41
ClG	0.18	0.23	0.16	0.24	0.19	0.17	0.06	0	0.04	0.22	0.11	0.13	0.26	0.32	0.29	0.26	0.21	0.27	0.27	0.26	0.31	0.29	0.26	0.52	0.33	0.36	0.26	0.5
NTG	0.19	0.23	0.19	0.24	0.2	0.17	0.12	0.04	0	0.16	0.06	0.09	0.21	0.34	0.24	0.27	0.19	0.16	0.16	0.23	0.27	0.3	0.29	0.38	0.3	0.33	0.24	0.52
Gri	0.1	0.12	0.2	0.15	0.14	0.12	0.22	0.22	0.16	0	0.12	0.21	0.18	0.23	0.16	0.22	0.22	0.18	0.18	0.14	0.21	0.27	0.38	0.29	0.24	0.24	0.22	0.42
Grk	0.2	0.22	0.2	0.25	0.22	0.2	0.15	0.11	0.06	0.12	0	0.11	0.24	0.33	0.27	0.29	0.2	0.21	0.2	0.23	0.3	0.27	0.24	0.42	0.29	0.34	0.28	0.47
Got	0.2	0.24	0.17	0.23	0.2	0.15	0.12	0.13	0.09	0.21	0.11	0	0.11	0.26	0.14	0.18	0.22	0.19	0.18	0.24	0.27	0.26	0.28	0.39	0.27	0.31	0.21	0.44
OE	0.19	0.24	0.21	0.22	0.19	0.19	0.21	0.26	0.21	0.18	0.24	0.11	0	0.18	0.1	0.18	0.21	0.15	0.12	0.2	0.23	0.27	0.36	0.26	0.22	0.23	0.17	0.36
E	0.14	0.19	0.18	0.16	0.14	0.21	0.25	0.32	0.34	0.23	0.33	0.26	0.18	0	0.11	0.12	0.24	0.32	0.31	0.26	0.26	0.25	0.4	0.44	0.24	0.38	0.26	0.41
D	0.15	0.17	0.19	0.18	0.15	0.17	0.23	0.29	0.24	0.16	0.27	0.14	0.1	0.11	0	0.11	0.25	0.24	0.23	0.18	0.16	0.29	0.41	0.3	0.22	0.29	0.2	0.46
Nor	0.16	0.19	0.21	0.19	0.16	0.17	0.2	0.26	0.27	0.22	0.29	0.18	0.18	0.12	0.11	0	0.26	0.29	0.28	0.22	0.24	0.3	0.38	0.43	0.24	0.31	0.21	0.44
Blg	0.17	0.2	0.17	0.23	0.16	0.17	0.14	0.21	0.19	0.22	0.2	0.22	0.21	0.24	0.25	0.26	0	0.11	0.11	0.27	0.3	0.2	0.33	0.47	0.32	0.36	0.23	0.5
SC	0.22	0.23	0.28	0.27	0.22	0.24	0.19	0.27	0.16	0.18	0.21	0.19	0.15	0.32	0.24	0.29	0.11	0	0.02	0.22	0.23	0.26	0.41	0.3	0.32	0.26	0.15	0.46
Rus	0.22	0.22	0.27	0.26	0.22	0.23	0.18	0.27	0.16	0.18	0.2	0.18	0.12	0.31	0.23	0.28	0.11	0.02	0	0.22	0.27	0.29	0.37	0.3	0.34	0.25	0.14	0.46
Ir	0.2	0.22	0.18	0.21	0.18	0.24	0.23	0.26	0.23	0.14	0.23	0.24	0.2	0.26	0.18	0.22	0.27	0.22	0.22	0	0.03	0.26	0.32	0.33	0.25	0.26	0.32	0.58
Wel	0.24	0.26	0.22	0.25	0.22	0.29	0.3	0.31	0.27	0.21	0.3	0.27	0.23	0.26	0.16	0.24	0.3	0.23	0.27	0.03	0	0.27	0.33	0.38	0.29	0.35	0.4	0.64
Heb	0.23	0.29	0.27	0.29	0.23	0.23	0.29	0.29	0.3	0.27	0.27	0.26	0.27	0.25	0.29	0.3	0.2	0.26	0.29	0.26	0.27	0	0.15	0.5	0.25	0.27	0.37	0.49
Ar	0.31	0.37	0.29	0.37	0.32	0.33	0.28	0.26	0.29	0.38	0.24	0.28	0.36	0.4	0.41	0.38	0.33	0.41	0.37	0.32	0.33	0.15	0	0.58	0.35	0.35	0.48	0.56
Wo	0.36	0.33	0.39	0.33	0.34	0.39	0.46	0.52	0.38	0.29	0.42	0.39	0.26	0.44	0.3	0.43	0.47	0.3	0.3	0.33	0.38	0.5	0.58	0	0.31	0.35	0.33	0.46
Hu	0.27	0.3	0.3	0.31	0.25	0.24	0.33	0.33	0.3	0.24	0.29	0.27	0.22	0.24	0.22	0.24	0.32	0.32	0.34	0.25	0.29	0.25	0.35	0.31	0	0.16	0.26	0.38
Fin	0.31	0.32	0.37	0.33	0.31	0.24	0.28	0.36	0.33	0.24	0.34	0.31	0.23	0.38	0.29	0.31	0.36	0.26	0.25	0.26	0.35	0.27	0.35	0.35	0.16	0	0.27	0.32
Hi	0.24	0.24	0.29	0.28	0.24	0.19	0.17	0.26	0.24	0.22	0.28	0.21	0.17	0.26	0.2	0.21	0.23	0.15	0.14	0.32	0.4	0.37	0.48	0.33	0.26	0.27	0	0.31
Bas	0.49	0.49	0.46	0.51	0.44	0.42	0.41	0.5	0.52	0.42	0.47	0.44	0.36	0.41	0.46	0.44	0.5	0.46	0.46	0.58	0.64	0.49	0.56	0.46	0.38	0.32	0.31	0

Table 3: Distance matrix derived from Table A.

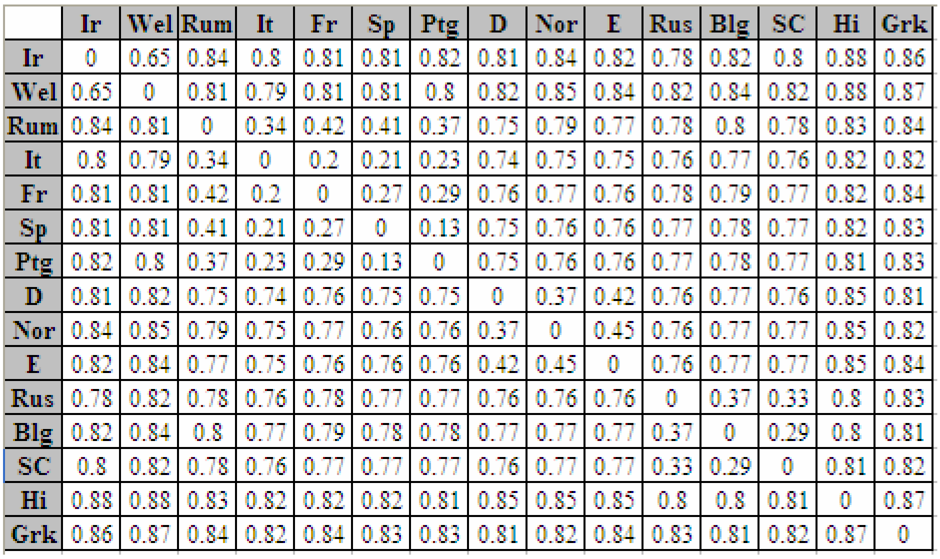
	i	d	0	00	Chance probability	Distance
SC - Rus	40	1	1	21	1,86446E-11	0,0243902
Ir - Wel	39	1	8	15	3,63798E-11	0,025
It - Ptg	50	2	1	10	2,94431E-13	0,0384615
Sp - Ptg	50	2	1	10	2,94431E-13	0,0384615
It - Fr	49	2	2	10	5,66214E-13	0,0392157
ClG - NTG	43	2	6	12	2,81375E-11	0,0444444
It - Sal	49	3	3	8	4,90719E-12	0,0576923
Fr - Ptg	47	3	3	10	1,74083E-11	0,06
Lat - ClG	31	2	13	17	6,14673E-08	0,0606061
NTG - Grk	46	3	5	9	3,27276E-11	0,0612245
Sp - Lat	29	2	24	8	2,16532E-07	0,0645161
NTG - Got	40	4	6	13	7,71655E-09	0,0909091
It - Sp	48	5	0	10	3,18599E-10	0,0943396
It - Lat	28	3	24	8	2,09315E-06	0,0967742
Ptg - Lat	28	3	23	9	2,09315E-06	0,0967742
Rum - Lat	28	3	25	7	2,09315E-06	0,0967742
Sal - Ptg	46	5	4	8	1,04319E-09	0,0980392
Sp - Fr	46	5	2	10	1,04319E-09	0,0980392
It - Rum	45	5	7	6	1,88184E-09	0,1
It - Gri	45	5	7	6	1,88184E-09	0,1
Sal - Fr	45	5	5	8	1,88184E-09	0,1
OE - D	44	5	5	9	3,38731E-09	0,102041
D - Nor	41	5	6	11	1,94796E-08	0,108696
ClG - Grk	40	5	9	9	3,47245E-08	0,111111
E - D	40	5	6	12	3,47245E-08	0,111111
Grk - Got	39	5	9	10	6,17324E-08	0,113636
Got - OE	39	5	9	10	6,17324E-08	0,113636
Blg - SC	31	4	19	9	1,52388E-06	0,114286
Blg - Rus	31	4	20	8	1,52388E-06	0,114286
E - Nor	38	5	8	12	1,09435E-07	0,116279
Rum - Gri	45	6	6	6	7,99781E-09	0,117647
Gri - Grk	45	6	5	7	7,99781E-09	0,117647
Sal - Gri	44	6	8	5	1,41138E-08	0,12
Lat - NTG	29	4	17	13	4,76371E-06	0,121212
Lat - Got	29	4	11	19	4,76371E-06	0,121212
OE - Rus	36	5	13	9	3,40787E-07	0,121951
Ptg - Rum	43	6	8	6	2,48402E-08	0,122449
ClG - Got	34	5	12	12	1,0473E-06	0,128205
Sal - Lat	27	4	25	7	1,4652E-05	0,129032
It - E	38	6	11	8	4,01261E-07	0,136364
Ptg - E	38	6	10	9	4,01261E-07	0,136364
Fr - Lat	25	4	26	8	4,42397E-05	0,137931
Sp - Rum	43	7	7	6	8,87152E-08	0,14
Sal - Rum	42	7	10	4	1,5259E-07	0,142857
Ptg - Gri	42	7	8	6	1,5259E-07	0,142857
Lat - Blg	24	4	25	10	7,62753E-05	0,142857
Gri - Ir	36	6	15	6	1,19275E-06	0,142857
Got - D	36	6	10	11	1,19275E-06	0,142857
Rus - Hi	30	5	9	19	9,44803E-06	0,142857
Fr - Rum	41	7	9	6	2,61583E-07	0,145833
Fr - Gri	41	7	9	6	2,61583E-07	0,145833
Rum - Got	35	6	16	6	2,04472E-06	0,146341
SC - Hi	29	5	10	19	1,61966E-05	0,147059
It - D	40	7	9	7	4,46871E-07	0,148936
Ptg - D	40	7	8	8	4,46871E-07	0,148936
Heb - Ar	40	7	4	12	4,46871E-07	0,148936
OE - SC	34	6	14	9	3,49099E-06	0,15
Lat - Grk	28	5	20	10	2,76295E-05	0,151515
Sal - Sp	44	8	3	8	1,67097E-07	0,153846
Ptg - Blg	38	7	10	8	1,28977E-06	0,155556
NTG - Rus	32	6	16	9	1,00433E-05	0,157895
It - Nor	37	7	13	6	2,17827E-06	0,159091
Sp - ClG	37	7	11	8	2,17827E-06	0,159091

(…)

(…)

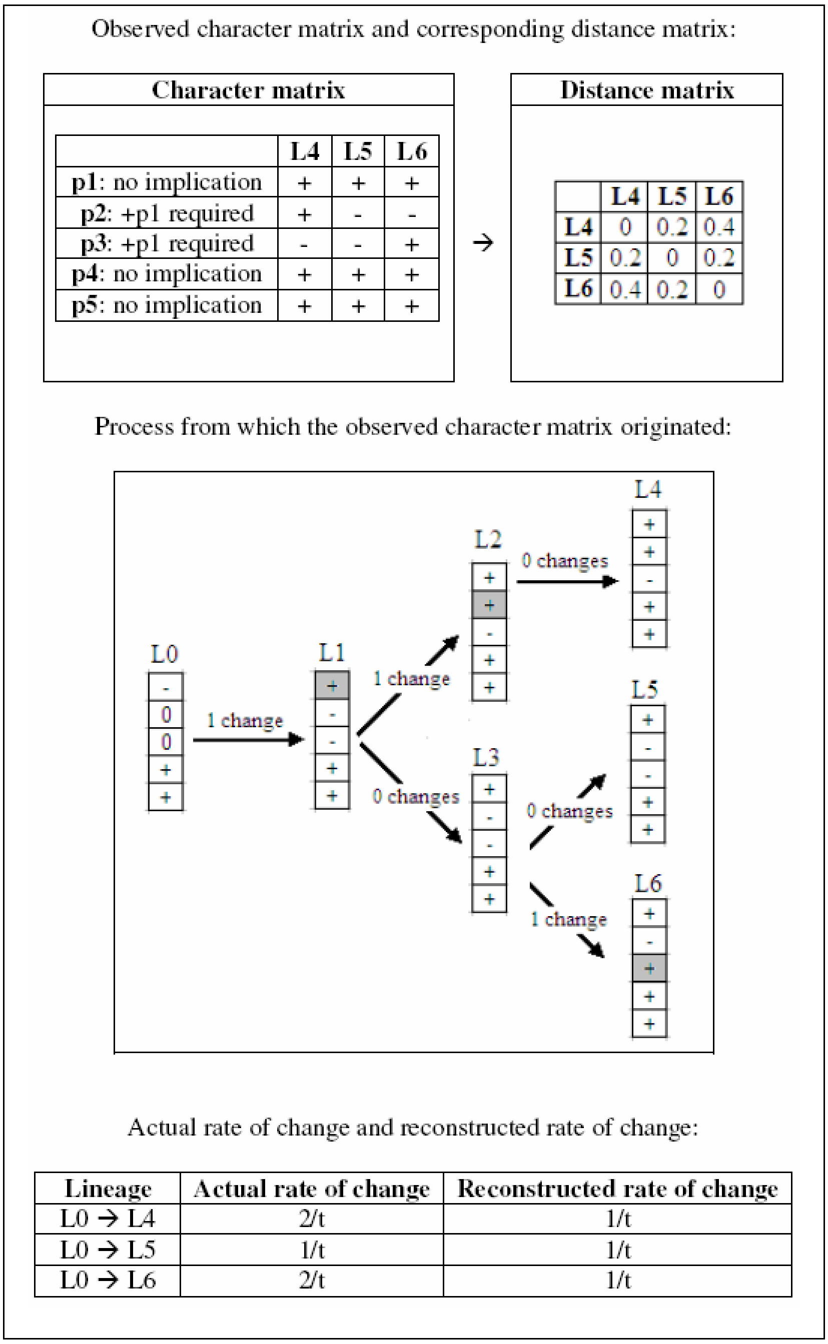
	i	d	0	00	Chance probability	Distance
NTG - E	27	14	14	8	0,016025366	0,341463
Grk - Fin	25	13	17	8	0,019699474	0,342105
Rus - Hu	25	13	13	12	0,019699474	0,342105
Ptg - Wo	21	11	25	6	0,030040853	0,34375
Wel - Fin	19	10	25	9	0,037308801	0,344828
Wo - Fin	17	9	25	12	0,046559423	0,346154
Ar - Hu	26	14	16	7	0,02110658	0,35
Ar - Fin	22	12	21	8	0,031918406	0,352941
It - Wo	21	12	24	6	0,041306172	0,363636
ClG - Fin	21	12	20	10	0,041306172	0,363636
OE - Ar	28	16	14	5	0,023687494	0,363636
OE - Bas	21	12	25	5	0,041306172	0,363636
Blg - Fin	21	12	22	8	0,041306172	0,363636
Heb - Hi	19	11	26	7	0,050875638	0,366667
Sal - Ar	29	17	11	6	0,024864662	0,369565
Sp - Fin	22	13	23	5	0,042967085	0,371429
Rus - Ar	22	13	21	7	0,042967085	0,371429
Fr - Ar	27	16	14	6	0,03014772	0,372093
Nor - Ar	25	15	17	6	0,036584738	0,375
Hu - Bas	20	12	21	10	0,052571492	0,375
Gri - Ar	28	17	13	5	0,031351095	0,377778
Wel - Wo	18	11	22	12	0,064442474	0,37931
NTG - Wo	21	13	19	10	0,054015764	0,382353
E - Fin	21	13	18	11	0,054015764	0,382353
Got - Wo	19	12	19	13	0,065714365	0,387097
Sp - Wo	20	13	24	6	0,066725356	0,393939
Rum - Wo	20	13	25	5	0,066725356	0,393939
E - Ar	24	16	15	8	0,057163653	0,4
Wel - Hi	15	10	30	8	0,097416639	0,4
D - Ar	25	17	15	6	0,057903418	0,404762
E - Bas	19	13	20	11	0,080879219	0,40625
Lat - Bas	13	9	27	14	0,11859417	0,409091
SC - Ar	20	14	22	7	0,081023646	0,411765
Rum - Bas	21	15	20	7	0,081023646	0,416667
Gri - Bas	21	15	20	7	0,081023646	0,416667
Grk - Wo	21	15	18	9	0,081023646	0,416667
Nor - Wo	17	13	25	8	0,111535052	0,433333
E - Wo	18	14	19	12	0,109764654	0,4375
Nor - Bas	18	14	22	9	0,109764654	0,4375
Ptg - Bas	19	15	22	7	0,108031528	0,441176
Got - Bas	15	12	28	8	0,12951985	0,444444
Lat - Wo	12	10	26	15	0,154172421	0,454545
D - Bas	18	15	22	8	0,12074112	0,454545
Sp - Bas	19	16	21	7	0,118159484	0,457143
SC - Bas	15	13	23	12	0,139482915	0,464286
Rus - Bas	15	13	24	11	0,139482915	0,464286
Wo - Bas	15	13	19	16	0,139482915	0,464286
Blg - Wo	17	15	21	10	0,131717585	0,46875
Grk - Bas	19	17	19	8	0,125110042	0,472222
Ar - Hi	16	15	24	8	0,139949934	0,483871
Heb - Bas	17	16	21	9	0,13583376	0,484848
It - Bas	18	17	21	7	0,1320606	0,485714
Sal - Bas	18	17	22	6	0,1320606	0,485714
ClG - Bas	15	15	24	9	0,144464448	0,5
Blg - Bas	16	16	22	9	0,139949934	0,5
Heb - Wo	16	16	22	9	0,139949934	0,5
Fr - Bas	17	18	19	9	0,1320606	0,514286
NTG - Bas	16	17	22	8	0,13583376	0,515152
ClG - Wo	15	16	21	11	0,139949934	0,516129
Ar - Bas	15	19	19	10	0,108031528	0,558824
Ar - Wo	14	19	20	10	0,095321937	0,575758
Ir - Bas	11	15	31	6	0,115128756	0,576923
Wel - Bas	9	16	31	7	0,0608854	0,64

Table 4: Table D including the coefficients <*i*, *d*>.



	Ir	Wel	Rum	It	Fr	Sp	Ptg	D	Nor	E	Rus	Blg	SC	Hi	Grk
Ir	0	0.65	0.84	0.8	0.81	0.81	0.82	0.81	0.84	0.82	0.78	0.82	0.8	0.88	0.86
Wel	0.65	0	0.81	0.79	0.81	0.81	0.8	0.82	0.85	0.84	0.82	0.84	0.82	0.88	0.87
Rum	0.84	0.81	0	0.34	0.42	0.41	0.37	0.75	0.79	0.77	0.78	0.8	0.78	0.83	0.84
It	0.8	0.79	0.34	0	0.2	0.21	0.23	0.74	0.75	0.75	0.76	0.77	0.76	0.82	0.82
Fr	0.81	0.81	0.42	0.2	0	0.27	0.29	0.76	0.77	0.76	0.78	0.79	0.77	0.82	0.84
Sp	0.81	0.81	0.41	0.21	0.27	0	0.13	0.75	0.76	0.76	0.77	0.78	0.77	0.82	0.83
Ptg	0.82	0.8	0.37	0.23	0.29	0.13	0	0.75	0.76	0.76	0.77	0.78	0.77	0.81	0.83
D	0.81	0.82	0.75	0.74	0.76	0.75	0.75	0	0.37	0.42	0.76	0.77	0.76	0.85	0.81
Nor	0.84	0.85	0.79	0.75	0.77	0.76	0.76	0.37	0	0.45	0.76	0.77	0.77	0.85	0.82
E	0.82	0.84	0.77	0.75	0.76	0.76	0.76	0.42	0.45	0	0.76	0.77	0.77	0.85	0.84
Rus	0.78	0.82	0.78	0.76	0.78	0.77	0.77	0.76	0.76	0.76	0	0.37	0.33	0.8	0.83
Blg	0.82	0.84	0.8	0.77	0.79	0.78	0.78	0.77	0.77	0.77	0.37	0	0.29	0.8	0.81
SC	0.8	0.82	0.78	0.76	0.77	0.77	0.77	0.76	0.77	0.77	0.33	0.29	0	0.81	0.82
Hi	0.88	0.88	0.83	0.82	0.82	0.82	0.81	0.85	0.85	0.85	0.8	0.8	0.81	0	0.87
Grk	0.86	0.87	0.84	0.82	0.84	0.83	0.83	0.81	0.82	0.84	0.83	0.81	0.82	0.87	0

Table 5: Matrix of lexical distance (data of Dyen *et alii* 1992).



Observed character matrix and corresponding distance matrix:

Character matrix

	L4	L5	L6
p1: no implication	+	+	+
p2: +p1 required	+	-	-
p3: +p1 required	-	-	+
p4: no implication	+	+	+
p5: no implication	+	+	+

→

Distance matrix

	L4	**L5**	**L6**
L4	0	0.2	0.4
L5	0.2	0	0.2
L6	0.4	0.2	0

Process from which the observed character matrix originated:

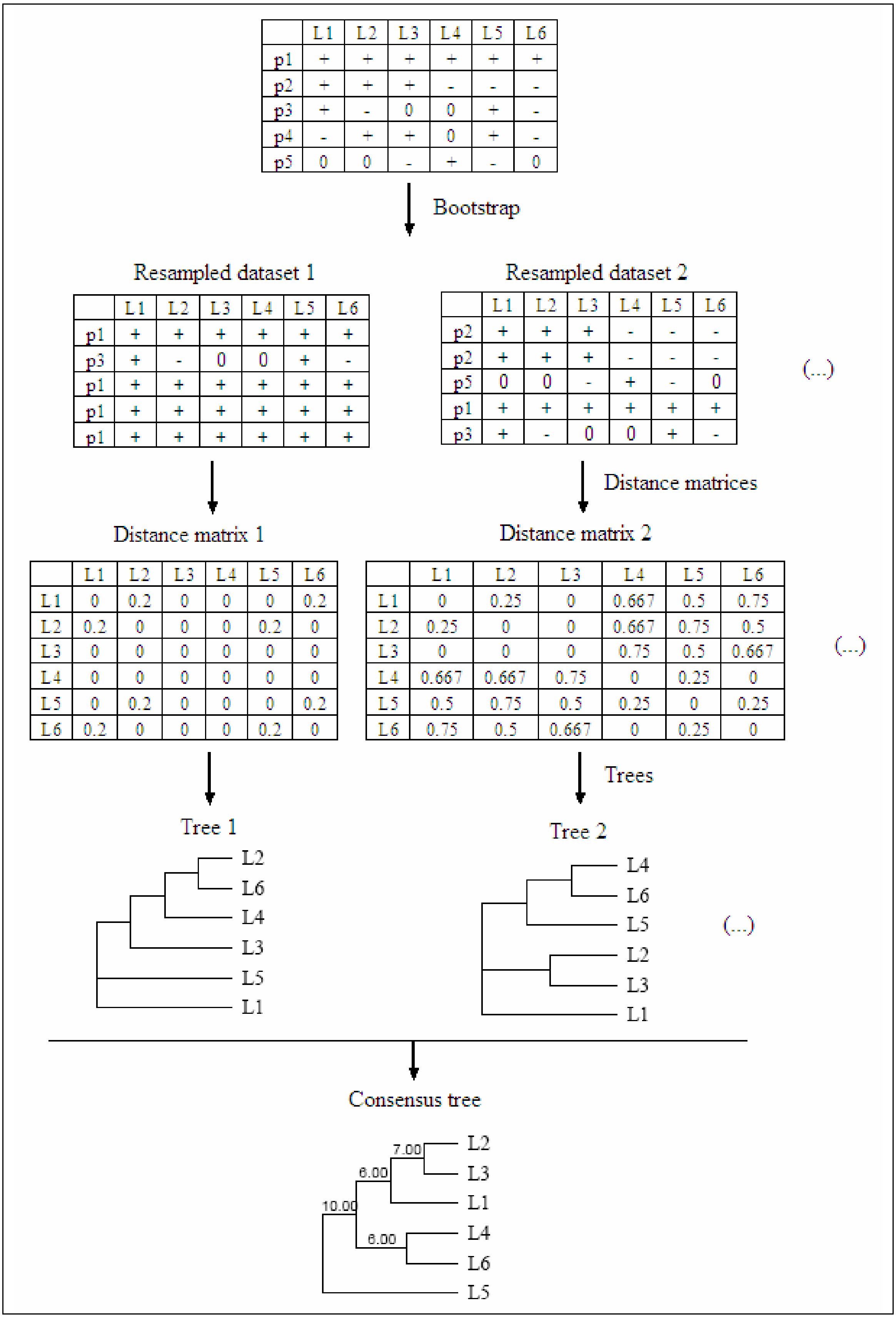
Actual rate of change and reconstructed rate of change:

Lineage	**Actual rate of change**	**Reconstructed rate of change**
L0 → L4	2/t	1/t
L0 → L5	1/t	1/t
L0 → L6	2/t	1/t

Sketch 2: Problematic estimates of rates of change dealing with structured data in the short period.

Method family	**Method**	**Program**	**Underlying hypothesis**	**Abbreviation**
Distance based – least squares method	Fitch-Margoliash	Kitsch (PHYLIP)	Molecular-clock	FMC
	Fitch-Margoliash	Fitch (PHYLIP)	-	FM
Distance based – Clustering method	UPGMA	Neighbor (PHYLIP)	Molecular-clock	UPGMA
	Neighbor-Joining	Neighbor (PHYLIP)	-	NJ
Character-based	Maximum parsimony	PAUP	Parsimonious evolution	MP

Table 6: The phylogenetic methods used in the experimentation.



	L1	L2	L3	L4	L5	L6
p1	+	+	+	+	+	+
p2	+	+	+	-	-	-
p3	+	-	0	0	+	-
p4	-	+	+	0	+	-
p5	0	0	-	+	-	0

Resampled dataset 1

	L1	L2	L3	L4	L5	L6
p1	+	+	+	+	+	+
p3	+	-	0	0	+	-
p1	+	+	+	+	+	+
p1	+	+	+	+	+	+
p1	+	+	+	+	+	+

Resampled dataset 2

	L1	L2	L3	L4	L5	L6
p2	+	+	+	-	-	-
p2	+	+	+	-	-	-
p5	0	0	-	+	-	0
p1	+	+	+	+	+	+
p3	+	-	0	0	+	-

Distance matrix 1

	L1	L2	L3	L4	L5	L6
L1	0	0.2	0	0	0	0.2
L2	0.2	0	0	0	0.2	0
L3	0	0	0	0	0	0
L4	0	0	0	0	0	0
L5	0	0.2	0	0	0	0.2
L6	0.2	0	0	0	0.2	0

Distance matrix 2

	L1	L2	L3	L4	L5	L6
L1	0	0.25	0	0.667	0.5	0.75
L2	0.25	0	0	0.667	0.75	0.5
L3	0	0	0	0.75	0.5	0.667
L4	0.667	0.667	0.75	0	0.25	0
L5	0.5	0.75	0.5	0.25	0	0.25
L6	0.75	0.5	0.667	0	0.25	0

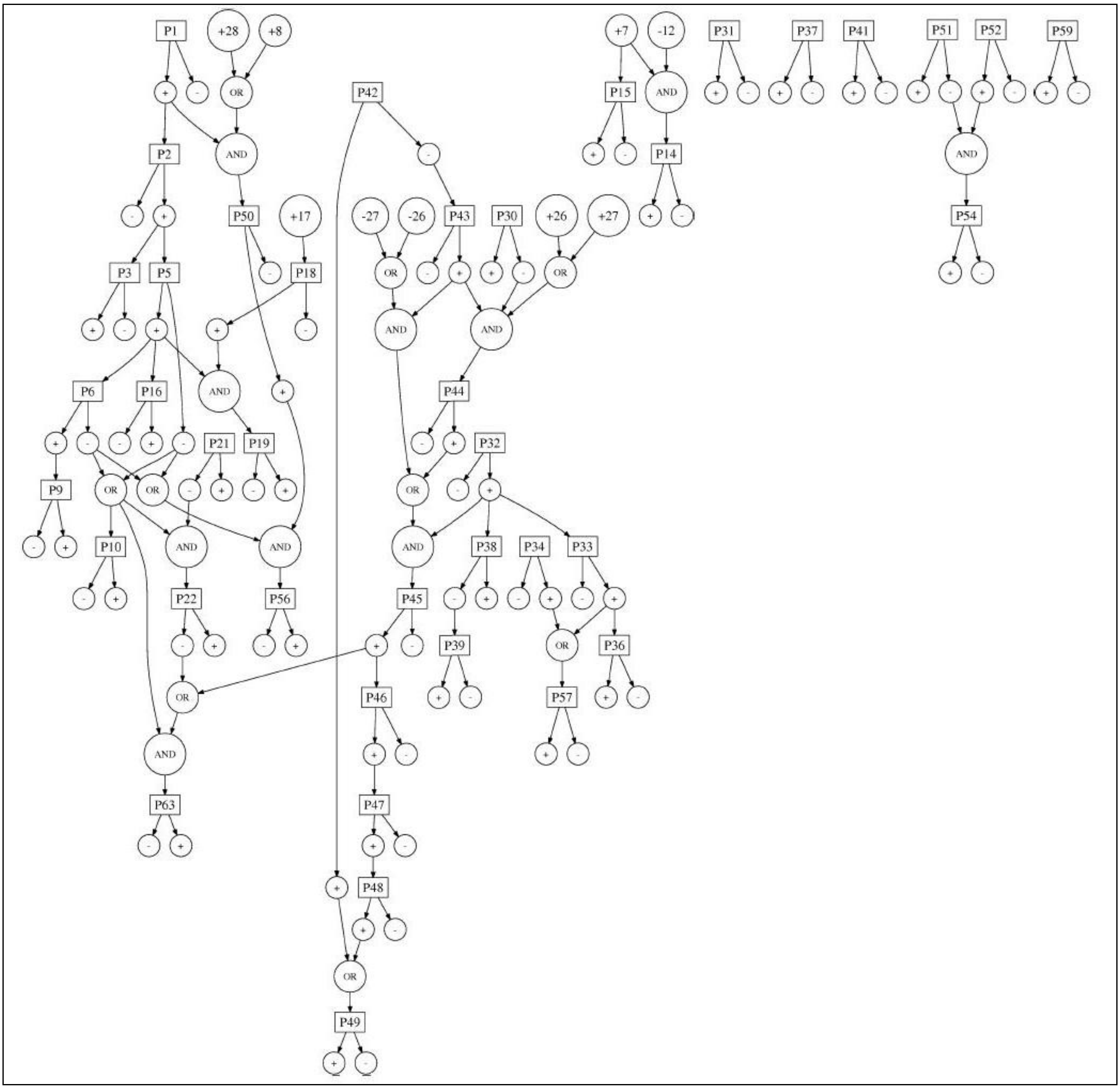
Sketch 3: Simplified account of the bootstrap-consensus procedure applied to the parametric dataset.

P	Parameter	Split
p1	± gramm. person	-
p2	± gramm. number +1	-
p3	± gramm. gender +2	(Ur, Bas) / others
p5	± feature spread to N +2	(Wo, Bas) / others
p6	± numb. on N (BNs) +5	Fr / others
p9	± free null partitive Q +6	(It, Sal, Sp, Ptg, Rum, Lat, ClG, NTG, Gr, Got, OE, Germ, Slav, Celt, Sem, Hu, Hi) / Fin
p10	± gramm. dist. art. -5 or -6 or +7	(Wo, Bas) / (Rom, ClG, NTG, Gr, Germ, Blg, Celt, Sem, Hu)
p14	± def on attrib. +7, -12	(It, Sal, Sp, Fr, Ptg, Gri, OE, E, D, Celt, Wo, Hu) / (ClG, NTG, Grk, Got, Sem)
p15	± def on relatives +7	(It, Sal, Sp, Fr, Ptg, Rom, ClG. NTG, Gr, Got. OE, Germ, Blg, Celt, Heb, Hu) / (Ar, Wo)
p16	± D-controlled infl. on N +5	(Rom, Lat, ClG, NTG, Gr, Got, OE, Germ, Slav, Celt, Sem, Ur, Hi) / Ar
p18	± gramm. cardinal adjectives +17	(SC, Wel, Heb) / (Rus, Ar, Fin)
p19	± plural spread from cardinals +5, -17 or +18	(Rom, Lat, ClG, NTG, Gr, Got, OE, Germ, Rus, Ar, Ur, Hi) / (Hi, Fin)
p21	± N-to-predicate incorporation	(Wo, Hu, Hi) / others
p22	± gramm. partial count -5 or -6 or +7, -21	(Rom, ClG, NTG, Gr, OE, Germ, Blg, Celt, Sem, Bas) / (Celt, Sem)
p30	± Gen over DemP	(Rom, Lat, ClG, NTG, Gr, Got, OE, Germ, Slav, Celt, Sem, Wo, Ur) / Hi
p31	± poss checking N	(Sem, Wo, Ur) / others
p32	± structured APs	Wo / others
p33	± feature spread to structured APs +32	(Rom, Lat, ClG, NTG, Gr, Got, OE, D, Nor, Slav, Ir, Sem, Fin, Hi) / (E, Hu, Bas)
p34	± feature spread to pred. APs	(Rom, Lat, ClG, NTG, Gr, Got, OE, Nor, Slav, Ir, Sem, Wo, Ur, Hi, Bas) / (E, D)
p36	± D-controlled infl. on A +33	(Rom, Lat, ClG, NTG, Gr, Slav, Ir, Sem, Fin, Hi) / (Got, OE, D, Nor)
p37	± DemP over relatives	(Rom, NTG, Gr, Got, OE, Germ, Slav, Celt, Sem, Wo, Ur) / Bas
p38	± free APs in Mod +32	(It, Sal, Sp, Fr, Ptg, Lat, ClG, NTG, Grk, Got, Ar) / (Rum, Gri, OE, Germ, Slav, Celt, Heb, Ur, Hi, Bas)
p39	± APs in Mod -38	(Rum, Gri, D, Nor, Ir, We, Ur) / (OE, E, Slav, Heb)
p41	± adjectival Gen	(Rom, Lat, ClG, NTG, Gr, Got, OE, Germ, Celt, Sem, Wo, Ur, Bas) / (Slav)
p42	± N-raising with pied-piping	(Rom, Lat, ClG, NTG, Gr, Got, OE, Germ, Slav, Celt, Wo, Ur, Hi) / (Sem, Bas)
p43	± N over ext.arg. -42	(Rom, Lat, ClG, NTG, Gr, Got, OE, D, Slav, Celt, Wo, Ur, Hi)/ (E, Nor)
p44	± N over GenO +26 or +27, -30 ,+43	(NTG, Gr, Got, OE, D, SC, Rus, Celt, Wo) / (Lat, ClG, Fin)
p45	± N over As +32, (-26 or -27, +43) or +44	(Rom, Gri, Celt) / (NTG, Grk, Got, OE, D, Slav, Hu, Hi)
p46	± N over M2 As +45	(Rom, Gri, Celt)
p47	± N over M1 As +46	(It, Sp, Fr, Ptg, Rum) / (Sal, Celt)
p48	± N over high As +47	Sal / (Celt)
p49	± N over cardinals +42 or +48	(Sal, Bas) / (Sem)
p50	± strong D (person) +1, +8 or +28	(Rom, ClG, NTG, Gr, Blg, Sem, Bas) / (OE, Germ, Celt, Wo)
p51	± NP over D	(Wo, Bas) / others
p52	± strong deixis	(Celt) / others
p54	± DP over Dem -51, +52	(Rom, Lat, ClG, NTG, Gr, Got, OE, Germ, Slav, Ur, Hi) / Heb
p56	± D-checking poss -5 or -6 or +8, +50 or -28	(It, Sal, Rum, ClG, NTG, Gr, Blg, Sem, Bas) / (Sp, Fr, Celt, Wo)
p57	± feature spread on poss +35 or +34 or +33	(Rom, Lat, ClG, NTG, Got, OE, D, Nor, Slav, Wo, Hi) / (Gr, Ir, Sem, Ur, Bas)
p59	± enclitic possessives	(It, Sal, Fr, Rum, ClG, NTG, Gri) / (Sp, Ptg, Blg, Sem, Hu, Bas)
p63	± gramm. geogr. art. -5 or-6 or +7, -22 or -23 or +45	(It, Sal, Fr, Rum, ClG, NTG, Gri) / (Sp, Ptg, Blg, Sem, Hu, Bas)

Table 7: Selection of 40 genealogically relevant parameters[1].

---

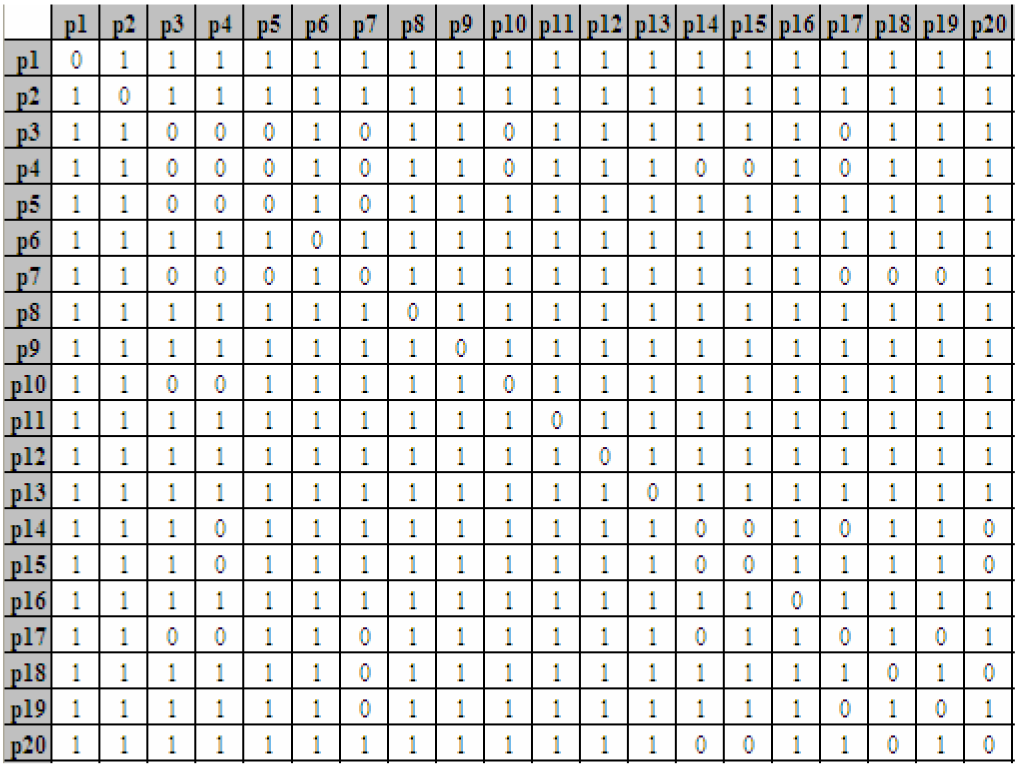
[1] In the table the following abbreviations are used: Rom (It, Sal, Sp, F, Ptg, Rum), Gr (Gri, Grk), Germ (E, D, Nor), Slav (Blg, SC, Rus), Celt (Ir, Wel), Sem (Heb, Ar) and Ur (Hu, Fin).



Sketch 4: Implicational structure(s) of the selection in table 7.

N	Parameter	Selection	Clique	CI	RI
p1	± gramm. person	1	1	1	1
p2	± gramm. number +1	1	1	1	1
p3	± gramm. gender +2	1	0	0.5	0.5
p4	± variable person on D +1	0	0	0.25	0
p5	± feature spread to N +2	1	1	1	1
p6	± numb. on N (BNs) +5	1	1	1	1
p7	± gramm. partial def	0	0	0.25	0.25
p8	± gramm. def +7	0	1	1	1
p9	± free null partitive Q +6	1	1	1	1
p10	± gramm. dist. art. -5 or -6 or +7	1	1	1	1
p11	± gramm. top. art. -10	0	1	1	1
p12	± def checking N +7	0	0	0.333	0
p13	± def spread to N +12	0	1	1	1
p14	± def on attrib. +7, -12	1	0	0.5	0.5
p15	± def on relatives +7	1	0	0.5	0
p16	± D-controlled infl. on N +5	1	1	1	1
p17	± gramm. cardinal nouns	0	0	0.25	0.4
p18	± gramm. cardinal adjectives +17	1	0	0.333	0
p19	± plural spread from cardinals +5, -17 or +18	1	1	1	1
p20	± gramm. mass-to-count	0	1	0.5	0
p21	± N-to-predicate incorporation	1	0	0.333	0
p22	± gramm. partial count -5 or -6 or +7, -21	1	1	0.5	0.667
p23	± gramm. count +22	0	0	1	1
p24	± count checking N +21 or +22	0	1	1	1
p25	± prepositional Gen	0	0	0.2	0.556
p26	± free inflected Gen -25	0	1	1	1
p27	± GenO +25 or -26	0	0	0.2	0.6
p28	± GenS +25 or -26	0	0	0.2	0.556
p29	± postpositional Gen +27 or +28	0	0	0.25	0
p30	± Gen over DemP	1	1	1	1
p31	± poss checking N	1	0	0.5	0.75
p32	± structured APs	1	1	1	1
p33	± feature spread to structured APs +32	1	0	0.333	0
p34	± feature spread to pred. APs	1	0	1	1
p35	± numb. on A +6, +33 or +34	0	1	1	1
p36	± D-controlled infl. on A +33	1	1	1	1
p37	± DemP over relatives	1	1	1	1
p38	± free APs in Mod +32	1	0	0.333	0.667
p39	± APs in Mod -38	1	0	0.333	0.5
p40	± overt Mod° -32 or +38 or +39	0	1	0.5	0
p41	± adjectival Gen	1	1	1	1
p42	± N-raising with pied-piping	1	0	0.5	0.5
p43	± N over ext.arg. -42	1	1	0.5	0
p44	± N over GenO +26 or +27, -30 ,+43	1	1	1	1
p45	± N over As +32, (-26 or -27, +43) or +44	1	0	0.333	0.667
p46	± N over M2 As +45	1	1	1	1
p47	± N over M1 As +46	1	1	0.5	0.5
p48	± N over high As +47	1	1	1	1
p49	± N over cardinals +42 or +48	1	1	1	1
p50	± strong D (person) +1, +8 or +28	1	0	0.333	0.6
p51	± NP over D	1	1	1	1
p52	± strong deixis	1	1	1	1
p53	± strong anaphoricity +52	0	0	0.2	0
p54	± DP over Dem -51, +52	1	1	1	1
p55	± D checking Dem -5 or -6 or +7, +52	0	0	0.25	0
p56	± D-checking poss -5 or -6 or +8, +50 or -28	1	0	0.25	0.25
p57	± feature spread on poss +35 or +34 or +33	1	0	0.333	0.714
p58	± feature spread on postp. Gen +29, +57	0	1	1	1
p59	± enclitic possessives	1	1	1	1
p60	± Consisten. Princ. +51 or (-44 or … or -47, +A-Compl)	0	1	0.5	0.667
p61	± null-N-licensing art -5 or-6 or -12, +50 or +51	0	1	0.5	0.5
p62	± obl. def. inh. +7,-22, (-25, +26) or +27,+42 or +45 or -50	0	0	0.5	0
p63	± gramm. geogr. art. -5 or-6 or +7, -22 or -23 or +45	1	1	0.5	0.75

Table 8: Summary of the weights applied to the parameters of Table A.



	p1	p2	p3	p4	p5	p6	p7	p8	p9	p10	p11	p12	p13	p14	p15	p16	p17	p18	p19	p20
p1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
p2	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
p3	1	1	0	0	0	1	0	1	1	0	1	1	1	1	1	1	0	1	1	1
p4	1	1	0	0	0	1	0	1	1	0	1	1	1	0	0	1	0	1	1	1
p5	1	1	0	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1
p6	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
p7	1	1	0	0	0	1	0	1	1	1	1	1	1	1	1	1	0	0	0	1
p8	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1
p9	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1
p10	1	1	0	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1
p11	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1
p12	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1
p13	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1
p14	1	1	1	0	1	1	1	1	1	1	1	1	1	0	0	1	0	1	1	0
p15	1	1	1	0	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	0
p16	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1
p17	1	1	0	0	1	1	0	1	1	1	1	1	1	0	1	1	0	1	0	1
p18	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	0
p19	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0	1	0	1
p20	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	0	1	0

Table 9: Sample of the compatibility matrix drawn from Table A.

		Average CI	**Average RI**
	Full Table A	0.671428524	0.652255016
→	(*p1 – p6*): Person, Number and Gender	0.791667	0.75
	(*p7 – p16*): Definiteness	0.758333	0.675
	(*p17 – p24*): Countability and related concepts	0.614583	0.508333
	(*p25 – p31*): Genitive case	0.478571	0.637302
→	(*p32 – p41*): Adjectives and relative modifications	0.7499999	0.7166667
→	(*p42 – p50*): N-raising	0.685185	0.696296
	(*p51 – p55* and *p60 – p63*): Demonstratives and other determiners	0.605556	0.546296
→	(*p56 – p59*): Possessive pronouns	0.645833	0.741072

Table 10: Average values of the weights in different subsets of the data.

Method	**References**	**Field of study**	**Applicable here**	**Motivation**
STRATIFICATION TECHNIQUE	Hübschmann (1875), Wen (1940), Sagart and Xu (2001)	Historical linguistics	No	Dating problem, insufficient data
DISTILLATION PROCEDURE	Wang (2004)	Historical linguistics	No	Dating problem, insufficient data
GEOGRAPHIC NEIGHBORHOOD	Sankoff (1972), Embleton (1981; 1986)	Lexicostatistics	No	Not useful
ERROR-MATRIX TEST	Wang (1989)	Lexicostatistics	No	Not proved
SKEWING TEST	Heine (1974), Hinnebusch (1999), Wang and Minett (2005)	Lexicostatstics	No	Estimate problem for syntax, data not available
NEGATIVE-BRANCH-LENGTHS TEST	Minett and Wang (2003)	Lexicostatistics	No	Not valid
BOOTSTRAPPING TEST	Cavalli-Sforza *et alii* (1997)	Population genetics	No	No distinction between borrowing, back-mutation and parallel evolution
SPLIT DECOMPOSITION (SPLITSTREEE AND NETWORK)	Bandelt and Dress (1992), Baret *et alii* (1995), Huson (1998),	Quantitative phylogenetics	No	Only internal evidence, no distinction between borrowing, back-mutation and parallel evolution
NEIGHBORNET	Bryant e Moulton (2004), Ben Hamed (2005), Baret *et alii* (2006)	Quantitative phylogenetics	No	Only internal evidence, no distinction between borrowing, back-mutation and parallel evolution
PARSIMONY TEST	Minett and Wang (2003)	Quantitative phylogenetics	No	Not proved
COMPATIBILITY TEST	Nakhleh *et alii* (2005a), Erdem *et alii* (2003), Brooks *et alii* (2005)	Quantitative phylogenetics	Yes	Sufficiently verifiable with external evidence

Table 11: Overview of the quantitative-oriented methods used to distinguish between vertical and horizontal transmission in language evolution.

P	**Split**
p1	-
p2	-
p5	(Wo, Bas) / others
p6	Fr / others
p9	(It, Sal, Sp, Ptg, Rum, Gri, Grk, E, D, Nor, Blg, SC, Rus, Ir, Wel, Heb, Ar, Hu, Hi) / Fin
p10	(Wo, Bas) / (It, Sal, Sp, Fr, Ptg, Rum, Gr, E, D, Nor, Blg, Ir, Wel, Heb, Ar, Hu)
p16	(It, Sal, Sp, Fr, Ptg, Rum, Gri, Grk, E, D, Nor, Blg, SC, Rus, Ir, Wel, Heb, Ar, Hu, Fin, Hi) / Ar
p19	(It, Sal, Sp, Fr, Ptg, Rum, Gri, Grk, E, D, Nor, Rus, Ar, Hu, Fin, Hi) / (Hi, Fin)
p30	(It, Sal, Sp, Fr, Ptg, Rum, Gri, Grk, E, D, Nor, Blg, SC, Rus, Ir, Wel, Heb, Ar, Wo, Hu, Fin) / Hi
p32	Wo / others
p36	(It, Sal, Sp, Fr, Ptg, Rum, Gri, Grk, Blg, SC, Rus, Ir, Heb, Ar, Fin, Hi) / (D, Nor)
p37	(It, Sal, Sp, Fr, Ptg, Rum, Gri, Grk, E, D, Nor, Blg, SC, Rus, Ir, Wel, Heb, Ar, Wo, Hu, Fin) / Bas
p41	(It, Sal, Sp, Fr, Ptg, Rum, Gri, Grk, E, D, Nor, Ir, Wel, Heb, Ar, Wo, Hu, Fin, Bas) / (Blg, SC, Rus,)
p44	(Gri, Grk, D, SC, Rus, Ir, Wel, Wo) / Fin
p46	(It, Sal, Sp, Fr, Ptg, Rum, Gri, Ir, Wel)
p48	Sal / (Ir, Wel)
p49	(Sal, Bas) / (Heb, Ar)
p51	(Wo, Bas) / others
p52	(Ir, Wel) / others
p54	(It, Sal, Sp, Fr, Ptg, Rum, Gri, Grk, E, D, Nor, Blg, SC, Rus, Hu, Fin, Hi) / Heb
p59	(It, Sal, Sp, Fr, Ptg, Rum, E, D, Nor, SC, Rus, Ir, Wel, Heb, Ar, Hu, Fin, Hi, Bas) / (Gri, Grk)
	Tot = 21 /40 (52.50%) PS-informative parameters: 9 /21

Table 12: List of compatible parameters (grey coloured cells indicate constant of PS-uninformative parameters).

P	**Split**
p3	(Hu, Fin, Bas) / others
p14	(It, Sal, Sp, Fr, Ptg, Gri, E, D, Ir, Wel, Wo, Hu) / (Grk, Heb, Ar)
p15	(It, Sal, Sp, Fr, Ptg, Rom, Gri, Grk, Germ, Blg, Ir, Wel, Heb, Hu) / (Ar, Wo)
p18	(SC, Wel, Heb) / (Rus, Ar, Fin)
p21	(Wo, Hu, Hi) / others
p22	(It, Sal, Sp, Fr, Ptg, Rum, Gri, Grk, E, D, Nor, Blg, Ir, Wel, Heb, Ar, Bas) / (Ir, Wel, Heb, Ar)
p31	(Heb, Ar, Wo, Ur) / others
p33	(It, Sal, Sp, Fr, Ptg, Rum, Gri, Grk, D, Nor, Blg, SC, Rus, Ir, Heb, Ar, Fin, Hi) / (E, Hu, Bas)
p34	(It, Sal, Sp, Fr, Ptg, Rum, Gri, Grk, Nor, Blg, SC, Rus, Ir, Heb, Ar, Wo, Hu, Fin, Hi, Bas) / (E, D)
p38	(It, Sal, Sp, Fr, Ptg, Grk, Ar) / (Rum, Gri, E, D, Nor, Blg, Sc, Rus, Ir, Wel, Heb, Hu, Fin, Hi, Bas)
p39	(Rum, Gri, D, Nor, Ir, We, Hu, Fin) / (E, Blg, SC, Rus, Heb)
p42	(It, Sal, Sp, Fr, Ptg, Rum, Gri, Grk, E, D, Nor, Blg, SC, Rus, Ir, Wel, Wo, Hu, Fin, Hi) / (Heb, Ar, Bas)
p43	(It, Sal, Sp, Fr, Ptg, Rum, Gri, Grk, D, Blg, SC, Rus, Ir, Wel, Wo, Hu, Fin, Hi)/ (E, Nor)
p45	(It, Sal, Sp, Fr, Ptg, Rum, Gri, Ir, Wel) / (Grk, D, Blg, SC, Rus, Hu, Hi)
p47	(It, Sp, Fr, Ptg, Rum) / (Sal, Ir, Wel)
p50	(It, Sal, Sp, Fr, Ptg, Rum, Gri, Grk, Blg, Heb, Ar, Bas) / (E, D, Nor, Ir, Wel, Wo)
p56	(It, Sal, Rum, Gri, Grk, Blg, Heb, Ar, Bas) / (Sp, Fr, Ir, Wel, Wo)
p57	(It, Sal, Sp, Fr, Ptg, Rum, D, Nor, Blg, SC, Rus, Wo, Hi) / (Gri, Grk, Ir, Heb, Ar, Hu, Fin, Bas)
p63	(It, Sal, Fr, Rum, Gri) / (Sp, Ptg, Blg, Heb, Ar, Hu, Bas)
	Tot = 19 /21 (47.50%)

Table 13: List of incompatible parameters.

	n1	**n2**	**n3**	**n4**	**n5**	**n6**	**n7**	**n8**	**n9**	**n10**	**n11**	**n12**	**n13**	**n14**	**n15**	**n16**	Lat	ClG	NTG	Got	OE
p1	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
p2	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
p3	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
p5	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
p6	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
p9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
p10	-	-	-	-	-	-	-	-	-	0	0	0	-	-	-	-	0	-	-	-	-
p14	-	-	-	-	-	-	-	-	-	0	0	0	-	-	-	-	0	+	+	+	-
p15	-	-	-	-	-	-	-	-	-	0	0	0	-	-	-	-	0	-	-	-	-
p16	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
p18	0	0	0	0	0	0	-	0	0	-	-	0	0	0	0	0	0	0	0	0	+
p19	0	0	0	0	0	0	+	0	0	+	+	0	0	0	0	0	+	+	+	+	+
p21	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
p22	+	+	+	+	+	+	-	+	+	0	0	0	+	+	+	+	0	-	-	0	+
p30	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
p31	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
p32	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
p33	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
p34	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+	+	+	+
p36	-	-	-	-	-	-	-	-	-	-	-	-	0	+	-	-	-	-	-	+	+
p37	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	#	#	+	+	+
p38	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	-
p39	0	0	0	0	-	-	-	-	-	+	+	+	-	-	-	-	0	0	0	0	+
p41	-	-	-	-	-	-	-	-	-	+	+	+	-	-	-	-	-	-	-	-	-
p42	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
p43	+	+	+	+	+	+	+	+	+	+	+	+	-	-	+	+	+	+	+	+	+
p44	0	0	0	0	0	+	+	+	+	+	+	+	0	0	+	+	-	-	+	+	+
p45	+	+	+	+	+	+	+	+	+	-	-	-	0	0	-	+	0	0	-	-	-
p46	+	+	+	+	+	+	+	+	+	0	0	0	0	0	0	+	0	0	0	0	0
p47	-	-	-	-	-	+	+	+	+	0	0	0	0	0	0	+	0	0	0	0	0
p48	0	0	0	0	0	-	-	-	+	0	0	0	0	0	0	+	0	0	0	0	0
p49	0	0	0	0	0	0	0	0	#	0	0	0	0	0	0	#	0	0	0	0	0
p50	+	+	+	+	+	+	-	+	+	0	0	+	-	-	+	+	0	+	+	0	-
p51	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
p52	+	+	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+
p54	-	-	-	-	-	-	0	-	-	-	-	-	-	-	-	-	-	-	-	-	-
p56	+	+	-	-	-	-	+	-	-	-	-	-	0	0	-	-	0	-	-	0	0
p57	+	+	+	+	+	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+
p59	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-
p63	-	+	+	+	+	+	+	+	+	0	0	0	0	0	0	+	0	+	+	0	0

Table 14: Ancestral parameter states reconstructed for the internal nodes of the tree in figure 15.

	Distance	i	d	0s	0d
n1 - Lat	0	23	0	13	4
n2 - Lat	0	23	0	13	4
n3 - Lat	0	23	0	13	4
n4 - Lat	0	23	0	13	4
ClG - NTG	0.03125	31	1	2	6
Lat - ClG	0.04	24	1	7	8
n5 - Lat	0.0434783	22	1	14	3
n14 - OE	0.0666667	28	2	4	6
NTG - Got	0.0666667	28	2	4	6
Got - OE	0.0666667	28	2	4	6
Lat - NTG	0.08	23	2	9	6
Lat - Got	0.08	23	2	5	10
n9 - Lat	0.0833333	22	2	14	2
n15 - Lat	0.0833333	22	2	10	6
n16 - Lat	0.0833333	22	2	14	2
n15 - OE	0.09375	29	3	3	5
n3 - ClG	0.1	27	3	6	4
n4 - ClG	0.1	27	3	6	4
n10 - OE	0.103448	26	3	6	5
n11 - OE	0.103448	26	3	6	5
n13 - OE	0.103448	26	3	5	6
n15 - Got	0.103448	26	3	5	6
n1 - Got	0.107143	25	3	8	4
n2 - Got	0.107143	25	3	8	4
n3 - Got	0.107143	25	3	8	4
n4 - Got	0.107143	25	3	8	4
n10 - NTG	0.107143	25	3	8	4
n11 - NTG	0.107143	25	3	8	4
n12 - NTG	0.107143	25	3	7	5
n12 - OE	0.107143	25	3	7	5
ClG - Got	0.107143	25	3	6	6
n10 - Got	0.111111	24	3	6	7
n11 - Got	0.111111	24	3	6	7
n14 - Got	0.111111	24	3	6	7
n12 - Got	0.115385	23	3	7	7
n10 - Lat	0.12	22	3	5	10
n11 - Lat	0.12	22	3	5	10
Lat - OE	0.12	22	3	9	6
n3 - NTG	0.125	28	4	4	4
n4 - NTG	0.125	28	4	4	4
n8 - Lat	0.125	21	3	14	2
n9 - OE	0.125	28	4	7	1
n12 - Lat	0.125	21	3	6	10
n15 - NTG	0.125	28	4	3	5
n16 - OE	0.125	28	4	7	1
n5 - OE	0.129032	27	4	7	2
n14 - Lat	0.130435	20	3	9	8
n1 - OE	0.133333	26	4	8	2
n2 - ClG	0.133333	26	4	6	4
n2 - OE	0.133333	26	4	8	2
n3 - OE	0.133333	26	4	8	2
n4 - OE	0.133333	26	4	8	2
n5 - ClG	0.133333	26	4	7	3
n13 - Lat	0.136364	19	3	10	8
n9 - Got	0.137931	25	4	9	2
n16 - Got	0.137931	25	4	9	2
n5 - Got	0.142857	24	4	9	3
n6 - NTG	0.151515	28	5	5	2
n9 - NTG	0.151515	28	5	5	2
n16 - NTG	0.151515	28	5	5	2
n10 - ClG	0.153846	22	4	10	4
n11 - ClG	0.153846	22	4	10	4
n12 - ClG	0.153846	22	4	9	5
n13 - Got	0.153846	22	4	7	7
n2 - NTG	0.15625	27	5	4	4
n5 - NTG	0.15625	27	5	5	3
n8 - OE	0.15625	27	5	7	1
n6 - ClG	0.16129	26	5	7	2
n9 - ClG	0.16129	26	5	7	2
n16 - ClG	0.16129	26	5	7	2
n1 - ClG	0.166667	25	5	6	4
n6 - Lat	0.166667	20	4	14	2
n7 - Lat	0.166667	20	4	15	1
n15 - ClG	0.166667	25	5	5	5
n8 - Got	0.172414	24	5	9	2
n8 - NTG	0.181818	27	6	5	2
n1 - NTG	0.1875	26	6	4	4
n6 - OE	0.1875	26	6	7	1
NTG - OE	0.1875	26	6	4	4
n8 - ClG	0.193548	25	6	7	2
n6 - Got	0.206897	23	6	9	2
n7 - Got	0.206897	23	6	10	1
n7 - OE	0.212121	26	7	6	1
ClG - OE	0.233333	23	7	6	4
n14 - NTG	0.241379	22	7	6	5
n7 - NTG	0.242424	25	8	6	1
n13 - NTG	0.25	21	7	7	5
n14 - ClG	0.25	21	7	6	6
n7 - ClG	0.258065	23	8	8	1
n13 - ClG	0.259259	20	7	7	6

Table 15: Table D comprising the pairs internal node – ancient language derived from the character matrix in table 14.

Pair	Syntactic distance	Separation time	Rate
It - Lat	0.0968	2	0.0484
Sal - Lat	0.129	2	0.0645
Sp - Lat	0.0645	2	0.03225
Fr - Lat	0.138	2	0.069
Rum - Lat	0.0968	2	0.0484
ClG - NTG	0.0444	0.5	0.0888
ClG - Gri	0.222	2.4	0.0925
ClG - Grk	0.111	2.4	0.04625
NTG - Gri	0.163	1.9	0.0857895
NTG - Grk	0.0612	1.9	0.0322105
OE - E	0.178	1	0.178

Table 16: Syntactic distances and corresponding separation times.

Language	Location	Latitude[2]	Longitude
It	Rome	42	12.48333
Sal	Lecce	40.352011	18.169139
Sp	Madrid	40.396	-3.7167
Fr	Paris	48.85	2.333333
Ptg	Lisbon	38.7	-9.183333
Rum	Bucharest	44.416667	26.1
Gri	Corigliano d'Otranto	40.166667	18.25
Grk	Athens	37.966667	23.716667
E	London	51.5	-0.133333
D	Frankfurt	50.116667	8.683333
Nor	Oslo	59.949444	10.756389
Blg	Sofia	42.7	23.333333
SC	Osijek	45.55	18.683333
Rus	Moscow	55.752222	37.598889
Ir	Dublin	53.3472	-6.2592
Wel	Cardiff	51.485278	-3.186667
Hu	Budapest	47.471944	19.050278
Fin	Helsinki	60.170833	24.9375
Bas	Donostia-San Sebastián	43.321389	-1.985556

Table 17: Data used to calculate geographic distances.

---

[2] Coordinates are recorded in decimal degrees.

	It	Sal	Sp	Fr	Ptg	Rum	Gri	Grk	E	D	Nor	Blg	SC	Rus	Ir	Wel	Hu	Fin	Bas
It	0	510.4	1367	1098	1869	1136	525	1057	1426	949.7	2002	895.4	635.7	2369	1879	1589	799.9	2194	1192
Sal	510.4	0	1851	1567	2346	793.1	21.75	547.4	1874	1315	2241	503.4	580.2	2227	2336	2054	795.7	2255	1700
Sp	1367	1851	0	1055	505.7	2477	1861	2373	1266	1451	2397	2258	1905	3445	1454	1235	1978	2954	355.9
Fr	1098	1567	1055	0	1457	1874	1586	2100	343.4	480.2	1348	1761	1287	2489	781	490.7	1248	1912	699.7
Ptg	1869	2346	505.7	1457	0	2983	2355	2858	1590	1897	2748	2763	2410	3914	1646	1498	2477	3368	793.3
Rum	1136	793.1	2477	1874	2983	0	800.5	745.2	2094	1456	2011	293.8	597.2	1501	2541	2302	642.6	1755	2246
Gri	525	21.75	1861	1586	2355	800.5	0	532	1894	1336	2263	509.3	600.3	2240	2356	2073	815.7	2275	1713
Grk	1057	547.4	2373	2100	2858	745.2	532	0	2395	1803	2613	527.9	941.5	2235	2859	2582	1124	2473	2241
E	1426	1874	1266	343.4	1590	2094	1894	2395	0	638.6	1159	2017	1531	2503	463.7	211.6	1453	1824	921
D	949.7	1315	1451	480.2	1897	1456	1336	1803	638.6	0	1102	1391	902.6	2022	1089	847.9	814.4	1516	1109
Nor	2002	2241	2397	1348	2748	2011	2263	2613	1159	1102	0	2102	1687	1644	1269	1280	1490	786.8	2042
Blg	895.4	503.4	2258	1761	2763	293.8	509.3	527.9	2017	1391	2102	0	488.4	1778	2477	2219	628.7	1948	2054
SC	635.7	580.2	1905	1287	2410	597.2	600.3	941.5	1531	902.6	1687	488.4	0	1743	1989	1734	215.8	1679	1657
Rus	2369	2227	3445	2489	3914	1501	2240	2235	2503	2022	1644	1778	1743	0	2797	2694	1571	892.7	3121
Ir	1879	2336	1454	781	1646	2541	2356	2859	463.7	1089	1269	2477	1989	2797	0	294	1899	2026	1159
Wel	1589	2054	1235	490.7	1498	2302	2073	2582	211.6	847.9	1280	2219	1734	2694	294	0	1662	1985	913.3
Hu	799.9	795.7	1978	1248	2477	642.6	815.7	1124	1453	814.4	1490	628.7	215.8	1571	1899	1662	0	1464	1702
Fin	2194	2255	2954	1912	3368	1755	2275	2473	1824	1516	786.8	1948	1679	892.7	2026	1985	1464	0	2603
Bas	1192	1700	355.9	699.7	793.3	2246	1713	2241	921	1109	2042	2054	1657	3121	1159	913.3	1702	2603	0

Table 18: Matrix of the geographic distances.