Language change as an evolutionary process

Jyri Lehtinen

MA Thesis

Department of General Linguistics

University of Helsinki

Autumn 2009

Contents

1. Introduction

Recent times have seen the framework of evolutionary biology making its way into the study of linguistics. On one hand, the origin of human language has been approached as a product of biological evolution, with the capacity of language having developed on the foundation of primate cognition and physiology (e.g. Hurford, Studdert-Kennedy and Knight 1998). On the other hand, one line of study has seen languages themselves as products of another kind of evolution, and the mechanisms of language change as driving that evolutionary process. In this work, it will be investigated in what way the evolutionary framework can be used as complementing the traditional methods of historical linguistics, dealing with language change, variation and contact.

However, language is neither completely a biological phenomenon, nor life a linguistic one. Consequently, language change and biological evolution cannot be thought of as the same thing, except in a limited sense. This has important consequences for the nature of the evolutionary framework as used in linguistics. This conceptual tool kit must be able to take into account the properties that are characteristic of language change that differ from possible counterparts in biological evolution. It remains to be seen if these two kinds of evolution can be seen as instances of a general principle of evolving systems, or if the differences prove to be great enough for a principle like that to become meaningless.

1.1 Goals and structure

This study is aimed at investigating the differences and similarities between biological evolution and language change. It is discussed whether language change can be seen as an instance of a generalised principle of evolution in a meaningful way, and whether an approach of this kind can provide linguistics with new solutions in language change, variation and contact. This study also attempts to determine if the applicability of this approach allows the methodology of evolutionary biology to be used in the study of language change and variation.

In section 2, the history of the two lines of study, both historical linguistics and evolutionary biology, are briefly outlined. It is noted that at the time of their emergence, particularly in the nineteenth century, ideas from both were borrowed as analogies with the subject matter of the other discipline. This discussion provides a background to the investigations in recent decades of a possible theory of generalised evolution that would subsume both biological and cultural evolution. It is explored in what way the process of language change has been approached as an instance of this theory. It is evaluated in which way the requirements of a theory of this kind are realised by language change, and in what ways languages behave differently from what would be expected from a process analogous to biological evolution.

Also, evolutionary biology utilises powerful computational statistical tools for assistance in determining the divergence of contemporary lineages from a common ancestral population, and for the study of population structure. If languages can be seen to change in a way similar to biological evolution, it becomes a point of interest whether these computational tools can be used in the study of language change as well. It will be seen in section 3 that this line of approach has been taken by many researchers, mainly in the form of phylogenetic inference methods applied to investigate the history of divergence of language families. The methods and results of many of these studies are discussed, and it is further explored in what form the linguistic data used in these methods should be, and what they have to offer to historical linguistics.

In the conclusion, a summary is made about the applicability of the conceptualisation of language change as an evolutionary process on one hand, and about the success or failure of the phylogenetic methods in the study of language on the other. It is also discussed what, in general, the evolutionary approach has to offer to the understanding of language change, and linguistic theory and typology.

Some terminology that is frequently encountered in the literature of historical linguistics is avoided here because of potential confusion with similar terms in evolutionary biology. For instance, languages that have descended from a common ancestral language are often said to be genetically related. To a biologist, this could imply that the languages in question are carried by the genes of biologically related speakers. Accordingly, in cases like this, languages 'descended from a common ancestral language' or just 'related languages' are talked about. Also, in historical linguistics parallel developments in related but different languages are called instances of drift. That term will be avoided because in biology, genetic drift refers to the process of genetic change in a population in the absence of a selection pressure. When that process is mentioned here, it is called genetic drift or neutral evolution, and does not involve the connotations of linguistic drift.

2. Language change in an evolutionary framework

Recently many linguists have begun to approach the phenomenon of human language with concepts that have their origin in biology (Lass 1997, Croft 2000, Ritt 2004, Givón 2002). Ideas of a Darwinian view of cultural evolution on one hand (Dawkins 1976, Hull 1988) and the adoption of phylogenetic methods developed in statistics and biology to investigate the divergence of language families on the other have been central to this recent surge of cross-discipline interaction (see e.g. Gray and Atkinson 2003, Nakhleh et al. 2005b for Indo-European languages, Kitchen et al. 2009 for Semitic, Holden 2002 for Bantu, Greenhill et al. 2008 for Austronesian, Dunn et al. 2008 for languages of Melanesia). This is not the first time such conversation has taken place. In the following, it will be seen that historical linguistics and evolutionary biology have exchanged ideas ever since they were originally formulated, and the recently growing interest towards fruitful interaction is mainly a return to closer terms after a period of relative quiet in the relationship between these disciplines.

2.1. The emergence of historical linguistics and evolutionary biology

At the end of the eighteenth century, Europe was living the Age of Enlightenment. Advances in the natural sciences motivated the systematic comparative study of languages in different parts of the world, in part inspired by the concern of finding out patterns of kinship between different nations, thought to be apparent in the histories of their languages. Linguistic typology in its earliest forms was born from this scientific atmosphere, as exemplified by the categorising of the languages of the world into four different morphological types by the German scholar and statesman Wilhelm von Humboldt (1767-1835). Another venue of exploration was opened up by the discovery of a relationship between the chief classical languages of Europe, Latin and Greek, and the classical language of India, Sanskrit, which received most attention through the writings of a British colonial administrator in India, Sir William Jones (1746-94). This finding was instrumental in the birth of a whole new field of linguistic science, that of historical-comparative linguistics. One of the most prominent founders of this field was Friedrich Schlegel, who was the first to explicitly declare a program for systematically comparing the grammatical systems of the languages that we today group together in the Indo-European languages. Mirroring the scientific motivations of the emerging field, he announced that the study of what he called comparative grammar would advance the knowledge of language as much as the development of comparative anatomy and embryology starting in the second half of the eighteenth century had done in biology. (Alter 1999: 7-9.)

One of the concepts central to this new science was that for each group of related languages there had been a single *protolanguage*, a language that gradually had diverged into different dialects that in turn had become the new languages. It had been known for long that the Romance languages, for instance, had developed from dialects of Vulgar Latin (ibid.), but only through the first reconstructions of Proto-Indo-European and Proto-Finno-Ugric it was demonstrated that this process could operate on a larger scale bringing about whole language families and, at least in principle, could account for any given group of related languages. The first scholar to actually reconstruct original forms of Proto-Indo-European words and grammatical morphemes was August Schleicher. His other main original contribution was the application of tree diagrams to illustrate the divergence of the protolanguage into the descendant branches (id.: 10-11).

All this preceded the publication of Charles Darwin's *The origin of species* (1859). Though neither being the first to suggest that species of plants and animals could evolve and change, nor the first to propose that similar species could have their origin in a common ancestral species, the work left its mark in the history of science by convincingly demonstrating the mechanisms that bring about new species. Darwin laid out the principles by which naturally and non-teleologically acting mechanisms could bring about evolution of species and divergence of a given species into new ones, given enough time (Bowler 1984).

Already in his notebooks preceding the publication of *The origin of species*, Darwin envisioned many of the principles of evolution through analogies with language change and divergence of languages as discovered by the new field of historical linguistics. For instance, he explained the absence of intermediate forms between different species as being caused by a phenomenon similar to that encountered when we have two very different words, but both demonstrably coming from the same source, as is the case with English *bishop* and French *évêque*, both coming from the Latin word *episcopus* (Alter 1999: 20-22). He used this comparison to illustrate the possibility that intermediate forms could well be lost, leaving only widely divergent but still related forms. Darwin also used analogies with language change in *The origin of species*, but the clearest expression of his consciousness of the similarities between the different processes can be found in the later *The descent of man* (Darwin 1874: 48):

> The formation of different languages and of distinct species, and the proofs that both have been developed through a gradual process, are curiously parallel. … We find in distinct languages striking homologies due to community of descent, and analogies due to a similar process of formation.

After the publication of The Origin of Species, the work was introduced to Schleicher by one of the earliest proponents of Darwin's theory in continental Europe, Ernst Haeckel. In this he had in mind Schleicher's interest in botany and gardening, but what immediately caught Schleicher's attention were the analogies with the evolution of languages that he saw, prompting him to remark in a letter to Haeckel, 'What Darwin lays down of the animal creation in general, can equally be said of the organisms of speech', i.e. individual languages (Schleicher 1863: 15).

However, statements like these were to remain for long only as analogies supporting the emerging conceptual background of the individual sciences, and many of Schleicher's ideas about linguistic evolution did not prove as long lived as his methods of reconstruction and the tree diagram. Consequently, the development of evolutionary biology continued independent of the development of historical linguistics, with no serious attempt at devising a unified science of evolution of both life and language. It was not until the second half of the twentieth century that these two fields of study began exchanging theory and methods, and by then evolutionary biology had developed a new framework, that of the modern synthesis of evolution. As with Darwin, one of the thinkers in this paradigm, Richard Dawkins, turned his attention to cultural evolution. In *The selfish gene* (1976), he put forth a concept of cultural evolution as a mechanism parallel to its biological counterpart, with cultural replicators, *memes*, taking the place of genes. This idea rapidly caught the attention of social scientists, and eventually found its way into linguistics.

2.1.1. The modern synthesis of evolutionary biology[1]

[1] Sometimes the term *Neo-Darwinian synthesis* is used. However, this usage is seen misleading in biology, as Neo-Darwinism generally refers to an earlier scholarly position.

After its initial reception in biology, the Darwinian theory of evolution became less popular in the beginning of the twentieth century, with Lamarckian theories and orthogenesis becoming the favourite explanatory systems for field naturalists and paleontologists (Bowler 1984: 289). In contrast to Darwin's theory, Lamarckism held that behaviours and traits acquired by an organism during its lifetime were heritable. Orthogenesis instead denied the role of selection and adaptation altogether, maintaining that the evolution of species came from factors internal to that species, and continued in a straight line, as it were, irrespective of external forces. It took the combination of mathematical models of the newly emerged study of population genetics on one hand and the appreciation of geographical isolation for adaptation on the other to initiate the new synthesis during the time between 1930 and 1950 and to convince the mainstream of biologists of the validity of Darwinian selection (id: 289-296). This trend received one of its major early confirmations as DNA was recognised as the genetic material in the 1940s.

At the end of the century, what biologists and others meant when they spoke of ”evolution”, was evolution as pictured in the modern synthesis. Central to this new paradigm was the Darwinian view of natural selection, with environmental factors affecting the reproductive capability of organisms through their adaptiveness to those factors. Evolution according to the modern synthesis is based on 1) genetic variability in populations of organisms caused by both genetic recombination through sexual reproduction and random mutations and 2) forces of natural selection acting in favour of some combinations of genes more than other, making some genes and combinations of genes more likely to appear after time passes through their improving the survival and reproductive success of the organisms. (Kutchera and Niklas 2004). In addition, random genetic drift is seen a possible cause of change in the gene pools of small populations. The division of organisms into higher taxa can be regarded as the same process as that dividing populations of a single species into different varieties and races, only happening on a larger scale and during longer time periods (ibid.).

Every event where populations became divided into different groups with no interbreeding between them was an event of speciation. In addition to *allopatric speciation* of this type, speciation was also seen to happen *sympatrically*, i.e. without geographic isolation between populations. Also, if the geographic barrier between two populations were not sufficient to prevent gene flow altogether, but the populations would still diverge on both sides of the contact zone between them, this would be an instance of *parapatric speciation*. In any case of speciation, eventually the gene pools of the new populations would drift apart, preventing reproduction between members of the different populations. (ibid.)

The new synthesis also put an end to what is called the *essentialist* view of species. Prior to modern biology, the common assumption was that species were fixed entities, Platonic ideas of sorts, of which individual organisms were imperfect manifestations, the variation of which masked the archetypal ideal of the species. In the nineteenth century, this view was most inspired by Hegelian idealism, in whose political philosophy the individual human being was subordinated to the ideal of the state. In the same way, idealists viewed individual organisms as subordinate, secondary realisations of the ideal type of the species (Bowler 1984: 100-101).

Not until the modern synthesis was it clearly recognised that the species could not be understood as an abstract type. The species was understood as a distinct set of characteristics essential to the species, and it was seen that evolution could cause any of these characteristics to change, and some populations that were morphologically similar were found to be in reproductive isolation (called *sibling species*). The new synthesis brought along with it a *population* view of species, which held that a species consists of reproductively isolated populations of organisms (Mayr 1982: 271-272). No amount of morphological difference would make different organisms part of different species, as far as they belonged to the same interbreeding population. Similarly, organisms that could not be distinguished by appearance or careful examination, could be part of different species if their respective populations showed no signs of interbreeding. (Mayr 1982.)

In practice, reproductive isolation doesn't need to be categorically clear cut. In many instances, where the ranges of different but closely related species come into contact, there is a band of interbreeding. However, when gene flow between the populations can be shown to be restricted only to this narrow area, the two species can be thought of as separate (Hull 1988: 102-103). The population view of species is a radically different view from the essentialist position, and for its part made the new evolutionary science stand out from its predecessors.

2.1.2. Memes: from biology into culture

One of the influential works written in the paradigm of the modern evolutionary synthesis is *The selfish gene*, the major work of the British zoologist Richard Dawkins (1976). In that book, Dawkins sought to explain the existence of altruism while holding that the primary unit that natural selection acts on is the gene instead of organism, group of organisms or species. The role of the organism, compared to that of the gene, was downplayed by calling the organism a 'vehicle' for the more important 'replicators' (i.e. genes).

What makes an entity a replicator, according to Dawkins, is that it satisfies three conditions: 1) It has sufficient copying-fidelity. If the structure of the entity is radically modified each time it is copied, we would not call this a process of replication. It should be noted that some alterations in copying are necessary for selection to have some variation to act on, but these alterations just need to be not too radical. 2) The entity has sufficient fecundity. This means that it must be able to make copies of itself, and these copies themselves must be able to do the same. Only this way lineages of successive replicators can form. 3) The entity has sufficient longevity. It must survive long enough in a stable form to produce faithful, fecund copies of itself. (ibid.)

Also included in this work was a chapter on what Dawkins proposed to be the unit in cultural evolution corresponding to genes in biological evolution. For this concept he introduced the word *meme*, and this term has caught on in what has become known as the study of *memetics* (id.: chap. 11, Croft 2006). Meme, according to Dawkins, can be any thought, idea, ideology or other piece of cultural substance that is learned by people from each other in a mostly conserved form, which in effect makes it behave as a replicator in a way reminiscent of genetic transmission.

Dawkins also proposed that once cultural transmission became reality in some ancestral hominid population, the memes that started spreading could do so without conferring any adaptive value in the biological sense. Memes would have a transmission mechanism independent of biological genetics, which would make memes that spread efficiently do so even if they decreased the ability of their hosts (i.e. humans holding those ideas) to survive and to reproduce, as would be the case, in the latter respect, with a meme of celibacy (Dawkins 1976: 198-199).

The concept of the meme was a speculative proposal from Dawkins, but other writers, notably psychologist Susan Blackmore (1999) have developed the idea, and memetics has become a source of interest in human sciences, including linguistics.

2.1.3. A general analysis of evolutionary processes

David Hull is a major contributor in the modern philosophy of science, having followed close up the development of modern evolutionary biology in the twentieth century. He has joined the ranks of cultural evolution theorists along with the memeticists with his account of scientific progress as an evolutionary process (Hull 1988). In this work one of his concerns was to define an evolutionary process in a way that could be applicable both to biological evolution and to the development and spread of scientific ideas. This conceptual system is referred to as the generalised analysis of selection.

In Hull's account, the process of selection is characterised by two kinds of entities: *replicators* and *interactors* (see figure 1). He defines replicator as an 'entity that passes on its structure largely intact in successive replications' and interactor as an 'entity that interacts as a cohesive whole with its environment in such a way that this interaction *causes* replication to be differential' (id.: 408; emphasis original). Hull notes that the same entity can function both as a replicator and as an interactor, but since these two processes are affected by forces acting on different levels and it is not efficient for the same entity to carry out both tasks, these different functions tend to be assigned to different kinds of entities (id.: 409). In biological evolution, strips of DNA, i.e. genes, function as replicators, whereas interactors can be organisms, groups of organisms or parts of them.

These two kinds of entities participate in a process of *selection* (figure 1), which Hull defines as a 'process in which the differential extinction and proliferation of interactors *causes* the differential perpetuation of the relevant replicators' and this process creates *lineages*, which are entities that persist 'indefinitely through time either in the same or an altered state as a result of replication' (id.: 409; emphasis original). Lineages are entities that are not replicated but instead change or stay the same through successive replications, and in biological evolution can be either chains of parent-offspring links or a passage through time of interbreeding populations (id.: 411-2).

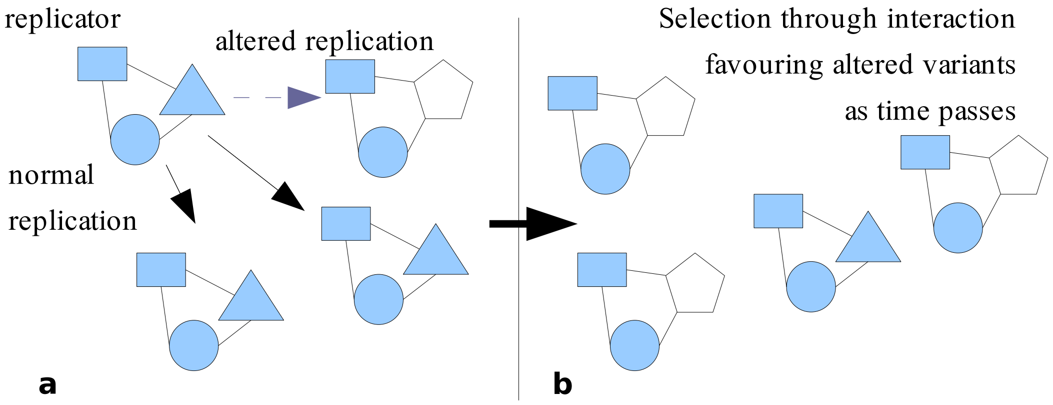

Figure 1. Replication and selection. a. One of the copies of the original replicator has been modified through altered replication. b. Through time, the frequency of the altered variants increases in the population as interaction with the environment favours their replication.

As a concluding note, it would be worth keeping in mind the remark of Hull that people 'tend to reject selection models in conceptual change out of hand because they have a simplistic understanding of biological evolution. Most objections to selection operating in conceptual change, if cogent, would count just as much decisively against selection models in biological evolution' (id.: 402-403). That is, many critics who oppose the notion that biological and cultural transmission can be seen as fundamentally very similar processes mostly have a mistaken view of biological evolution.

On closer inspection, what might be more surprising than the similarity of cultural transmission to biological evolution, is, vice versa, the resemblance of natural selection in biology to the more familiar social selection of cultural practices and ideas.

2.2. From biological to linguistic evolution

As we have seen, it took several decades for Darwin's theory to be finally included in a mature form in the paradigm theoretical framework of biology. Subsequently, this new framework became a source of inspiration for linguists looking for new directions. Even though at times historical linguists would speak of language evolution, until recently it tended to be only an alternative term for language change, with no implication that 'evolution' should be understood as a direct analogy with biological evolution (Andersen 2006). Historical linguistics had proceeded without troubling the minds of linguists with evolutionary theory, and synchronic description of languages was carried out largely without the notion that the systems that were described were products of historical development. This was largely due to the influential position of Ferdinand de Saussure (1857-1913) in modern linguistics. His dichotomy between synchrony and diachrony was justified by the observation that at any given moment the system used in a community seemed to be defined by the social agreement between the speakers rather than by historical development. Consequently, description without recourse into diachrony was seen to be not only possible, but desirable (Saussure 1959).

After August Schleicher (see 2.1.1 above), one of the first proposals in linguistics for a direct application of the biological evolutionary thought was by the historical linguist Roger Lass, who in his *Historical linguistics and language change* (1997) presented ideas for understanding a language as an evolutionarily evolving 'quasi-species' (a term borrowed from population biology of viruses), a 'population of variants moving through time, and subject to selection' (id.: 377). He remains agnostic about the selection pressures driving the evolution of languages. Instead, he goes to the heart of the problem of defining the boundaries of a given language by taking the view that it raises many problems to speak of a language as an abstract system when, for instance, we want to know when a certain sound change has 'happened in a language'. How many speakers of a language do we need to implement a particular change in order to say that the system of that language has changed? Lass attempts to tackle this problem by disposing of the view of languages as abstract, discrete systems (a view reminiscent of the essentialist view of species discussed above, see also section 2.3) and putting the idea of quasi-species in its place: it is not the system of language that changes, but the proportions of different variants in the population of linguistic units that is used in the community. He calls for a *medium-neutral* picture of evolution where the evolving object is not the subject of interest rather than the central characteristics of the process itself. In connection to this he refers Dawkins' idea of 'universal Darwinism' (i.e. evolutionary theory of both biological and memetic evolution; Dawkins 1983).

William Croft based his framework more on David Hull's discussion (1988, section 2.1.4 above) than on Dawkins. He set out to lay a conceptual foundation for viewing language change as an instantiation of Hull's generalised account of selection (GAS) in *Explaining language change* (Croft 2000). This approach will be more closely explored in the next section.

Taking up the task set up by Lass, Nikolaus Ritt explored the evolutionary perspective in the area of phonological change in his *Selfish sounds and linguistic evolution* (2004), inspired by his reading of *The selfish gene* (Dawkins 1976), as is evident already from the title. Ritt, too, is concerned with the concept of language, and, like Lass, views languages like Old English not as disembodied, abstract, eternal systems, but 'structured sets, “pools” or “populations” of individual competences' (Ritt 2004: 37). Ritt also notes that the pursuit of an evolutionary theory of language change is compatible with the research of Complex Adaptive Systems as defined by the Santa Fe Institute that studies such systems (Gell-Mann 1992). The term is meant to be a neutral one in place of 'Darwinism' and other concepts that are usually associated with biology, and refer to systems that evolve in terms of Dawkins' universal Darwinism or Hull's generalised account of selection (ibid., Ritt 2004: 91-93).

In trying to find linguistic entities corresponding to replicators in the sense of Dawkins (1976, see Hull's closely similar definitions in section 2.1.3), Ritt ends up with at least phonemes and morphemes because they stay relatively intact in linguistic replication, they are stable in the linguistic competence of a speaker, and they are easily passed on in replication. He also includes idiomatic phrases, syntactic categories and rules operating productively on different levels of a language (Ritt 2004: 122-152). In regard to the material existence of these replicators, Ritt suggests that we understand their structure as the mental patterns in which they are instantiated in speakers.

Interestingly, Ritt also attempts to apply the evolutionary view to sound change during the history of English. In the literature of historical linguistics regarding certain vowel lengthenings and shortenings that happened during the Middle English period, it has been noted that the changes in question have required complex formulas in order to account for them in neogrammarian terms, and even then many individual words have been left over to be assigned as cases of analogy. Ritt analyses these changes as expressions of selection pressures resulting from a 'memeplex' (a complex of linguistic memes) that favours certain kinds of metrical units in morphemes (id.: 240-288).

2.2.1. A theory of utterance selection in language change

One of the first and most thorough attempts at creating an evolutionary framework for language change in terms of concepts from the modern evolutionary synthesis was Croft (2000). His foremost source of concepts of cultural evolution is David Hull's (1988) generalised analysis of selection, outlined in section 2.1.3 above. What Croft proposes is a *theory of utterance selection*, setting utterances in the place of genes as replicators in linguistic evolution. By doing this he anchors his theory firmly in concrete language use. What matters in Croft's theory are utterances that occur in actual use and their linguistic structure as conceived by the speaker and the hearer.

Central to Croft's account of linguistic selection is to view the forms of a language as forming a population. This population is what he proposes us to consider a language. He refers to the modern view in biology of considering a species as an interbreeding population instead of an ideal, abstract type which individual organisms would exemplify (see 2.1.1), and compares this to the difference between his population view of language and the more prevalent idea that, in essence, languages are abstract types in the essentialist view. In the same way as natural selection in biological evolution acts on the variation that is internal to the species, Croft proposes that we consider language change as a selection process acting on the variation within a given language.

Croft uses the term *utterance* to denote an individual stretch of produced language including its grammatical structure, meaning and function in the social context as intended by the speaker and interpreted by the hearer. He stresses that an utterance needs to have not only structure, but it also needs to be grammatically possible, actually produced and structurally understood by the hearer. From this it follows, contrary to the formalist view of a language as the set of *possible* sentences generated by the rules of grammar, that a *language* should be understood as the population of actually occurring utterances used in a particular community.

He also proposes that the structure of utterances is made up of individual *linguemes* (a term Croft borrows from Martin Haspelmath), which are individually learned units of phonology, morphology, lexicon and syntax (i.e. phonemes, morphemes, lexemes and syntactic constructions). The participants in linguistic interaction incorporate and interpret these linguemes in utterances with the aid of their *grammars*. 'Grammar', following Croft's usage, is an actual, psychologically real cognitive system which the speakers have acquired through their learning of the language in question. Grammar in this context is distinguished from a descriptive grammar obtained through linguistic analysis, and from a postulated social fact or norm that regulates the linguistic behaviour of speakers. This definition of grammar follows from Hull's generalised analysis of selection, which requires that all participating entities are actually occurring spatio-temporal entities with internal structure. Of course, the physical forms of utterances are composed of sound waves, written marks on a paper and so on, and are not obviously analysable into linguemes by their physical properties alone. Only with the additional cognitive structure of grammar, the physical signal becomes overlaid with an interpreted linguistic structure. (Croft 2000: 26-7.)

Croft proceeds to find out what features in language exemplify the entities Hull (1988) names as essential to a process in order for it to be called selectionist in the same way biological evolution is. As the linguistic replicator Croft names the lingueme, and as the interactor, the speaker and their knowledge of grammar. Accordingly, the utterance, made up of linguemes, functions as a structured set of replicators; the production of utterances according to the conventions of the community functions as normal replication; the mechanisms for linguistic innovation function as bringing about altered replication. An utterance, then, as a structured set of linguemes, exhibits a hierarchical structure: as interpreted by a grammar of a speaker, it is composed by morphosyntactic constructions, which are in turn made up of morphemes (i.e. words, clitics and affixes), and they themselves are ultimately conceived as phonological structures. Like genes in a population construct a gene pool, linguemes used in a language form a *lingueme pool.* A gene pool contains different alleles, which are different versions of the same genes. In the same way, the lingueme pool contains all the different *variants* of the linguistic structures used in the language. He notes that in biological evolution the paradigm case of an interactor is the organism, and accordingly, in language change it is the speaker. The environment that in the case of language causes the differential success of linguemes he equates with the social-communicative context. (Croft 2000: 27-41.)

Croft notes that there have been two major currents in the study of the causes of language change. One is the child-based theory of language change, which tries to explain change with imperfect first language acquisition. Croft notes that theories of this kind were debated at the end of the nineteenth century, but are today most popular among generativists, from whose point of view it seems most natural to view change as modifications in the initial 'parameter-setting' of the language acquisition device. However, Croft points out that the changes that languages typically go through over time are different from errors made by children learning their first language. On the other hand, speakers modify their language during all of their childhood, approximating the linguistic practice of their peers at all ages beginning in preadolescence (id.: 44-53). Consequently, Croft sees that alternatives to the child-based theory should be sought.

Next, Croft considers different frameworks of study of language change that he sees as more useful to his theory of utterance selection. The sociohistorical theory, or variationist sociolinguistics, first explicated by Weinreich, Labov and Herzog (1968), studies the patterns of variation and changes in the distribution of linguistic variants in language communities. Croft finds that this research is important in regard to selection in his theory, which he sees as being based on sociolinguistic factors that are the subject of sociohistorical theory. However, he sees that as a theory of language change it is incomplete: it is concerned with the variable success of some variants over others, but does not explain how those different variants come into being; this problem is referred to as the *actuation problem* in sociohistorical study. Therefore, Croft regards sociohistorical theory as a theory of utterance selection without explanations of innovation (Croft 2000: 53-59).

Next, he turns to the invisible hand theory of language change, formulated in Keller (1994), which he sees as making an important distinction about the role of individuals' intentions in language change. Keller sees language change as what he calls a *phenomenon of the third kind*, i.e. an unintended causal effect of intended human social actions (id.: 57). Languages do not change in certain ways because speakers intend them to do so, but they change as a by-product of the speakers' intentions to attain socio-communicative goals with their language use, which sometimes involves using their language in novel ways. The phenomenon is said to be of the third kind to distinguish it from the products of intentional design (artifactual phenomena) and products of purely natural processes with no involvement of human intentions (natural phenomena; Croft 2000: 59-62).

Croft agrees with Keller that language change is not teleological in the sense that language change would follow the intentional design of speakers or norm authorities, but only in the sense that the speakers' teleological design is directed towards the socio-communicative effect of their utterances, and any changes in the distribution or number of linguistic variants in the language community are non-teleological effects of those communicative processes. Somewhat confusingly, Croft calls only the previous possibility 'teleological' and the latter 'intentional' (id.: 63-66). It could be argued that in both instances teleology and intentions are involved, but only directed at different outcomes. In any case, Croft argues that mechanisms for normal replication, for selection and for altered replication all need some other explanations than the teleological design of speakers. For example, he sees normal replication resulting from the intentional acts of speakers to try to speak in such a way that they are understood best and in a way that doesn't violate the conventions of the speech community, and from the non-intentional processes of cognitive entrenchment of the most frequently used linguistic structures in the mental grammar of the speaker (id.: 71-73).

Conventions in speech communities arise from coordination problems frequently encountered by the members of those communities. However, Croft points out that speakers will always face coordination problems for which there are no sufficient solutions to be found in the conventions, and the context-depended communicative function of an utterance is always ultimately negotiated between the participants of linguistic interaction. In those cases the speakers are forced to rely on different non-conventional coordination devices, for which function Croft invokes different semantic models of cognitive linguistics. In the end it is impossible in principle to draw a clear boundary between conventional and non-conventional coordination devices, as the meaning in terms of the communicative function of a linguistic structure is in practice a result of its history of use, i.e. its lineage, in different communicative situations. (id.: 99-115).

Utterances are formed by combining different linguistic structures, i.e. linguemes, into a structured set, and the meaning of an utterance is a function of these linguemes and their conventional meanings. Because meaning is never completely conventional, there is always the possibility for the appearance of novel form-function pairings, in other words for *form-function reanalysis*, which Croft sees as an important mechanism of altered replication. Other sources of innovation are *interference*, where competence in several languages causes structures to be affected by another language and *intraference*, where these kinds of changes are caused not by different languages but by different dialects, sociolinguistic variants or different structures of the same language. (id.: 117-165.)

As novel variants are brought into being by altered replication, they come under differing selection pressures, which according to Croft are social in nature, and affected by the structure of the community in which they are in use. As established in research of the diffusion of cultural innovations, linguistic variants, too follow in their spread the so-called S curve: the diffusion starts out slowly, but once it picks up, it gets exponentially accelerated, and once it moves well past midway to its final spread, it starts to slow down again, and slowly settles to some value. There are three important ways in which an innovation can become a new convention: either the new variant is given a function that is complementary to that of the older variants, the different variants get associated with different sociolinguistic groups, or one of the variants gets favoured more and more at the expense of the others so that their relative frequencies change in the community (id.:174-178). To explain how new variants are introduced into a community from another, Croft adopts the weak-tie theory of Milroy and Milroy (1985) and Milroy (1992). In contrast to earlier views of Labov (1980) that new variants enter communities through individuals that have strong ties within the community but also strong ties outside it, the Milroys argue that social and anthropological studies support the opposing idea that it is instead the individual with weak ties in the community and outside it who acts as the central introducer of a new variant. Strong-tie individuals, in contrast, tend to be at a greater pressure to maintain the conventions of their speech communities (Croft 2000: 178-180).

At a larger scale, it generally holds that peripheral, more communicatively isolated areas tend to maintain earlier conventions more faithfully compared to more central areas of a language community, while perhaps generating structures more complex than earlier as a sign of local identity, whereas in more central areas where much communication goes on between groups of different place of origin, the language tends to converge to a levelled, 'hybrid' variety of the speech forms of the area, called a koiné, often with simplification of the grammatical structure of the language (id.: 190-194).

2.3. Summary of evolutionary approaches to language change

The evolutionary perspective provides new possibilities for determining what varieties are different languages and what are dialects of a single language. If it is agreed that a language is to be understood as a quasispecies forming a population with its own lingueme pool, it can be considered if the population view offers a way to define borders between languages. In modern linguistics, languages have been mainly defined, following the situation in Europe in the nineteenth and twentieth centuries, on the basis of standard languages. As there exist standardised, codified and therefore easily accessible versions of 'English', 'Dutch', 'German', 'Danish' and 'Swedish', for instance, the linguist has been satisfied in focusing on those standardised linguistic systems, and has been able to account for all variation in linguistic practice in the areas where those standards are valid as dialects or sociolinguistic varieties of English and so one, i.e. as varieties of those standard languages.

This practice of viewing all linguistic practice as representing, more or less faithfully, some standard language, probably contributes to the distinctions made by several influential linguists in the twentieth century between the normative system of a language and the actual linguistic practice of the speakers, and viewing the latter as being a realisation of the former. For instance, Saussure (1959) spoke of *langue*, the norm of the speech community that governed *parole*, the linguistic practice in the community, Hjelmslev (1942) posited a hierarchy with the abstract *system* or *schema* in the higher end and *observable manifestation* in speech activity in the lower, and Coşeriu (1970) made the tripartition between *System*, *Norm* and *Sprechen*, where Norm was one of the possible realisations of the abstract System, and Sprechen constituted individual realisations of the socially shared Norm (Bartsch 1987).

However, when the actual speech of the areas under different standard languages are studied, one may notice that there are no clear boundaries between what on the basis of normative standard systems seem different languages. For instance, research in dialectology has shown that most of Europe can be divided into areas where just five different *dialect continua* are spoken, corresponding to many more standard languages: Continental Scandinavia, for instance, where at least five standards of Swedish, Norwegian and Danish exist, belongs to the Scandinavian dialect continuum. In no locality situated in this area is there such a break that people living at a distance from each other couldn't understand each other, even though people from opposite ends of the region would probably have trouble. (Chambers and Trudgill 1999: 5-7.)

In this case, it could be said that speech in this area forms just one linguistic population that exhibits geographical gradation. In other cases where the term dialect continuum is applied, things may not be so simple: another European dialect continuum involves the continental West Germanic languages like Dutch and German, but the transition from Dutch dialects to German ones is not as gradual as in the previous example. Instead, it is fairly discontinuous, and intermediate varieties in the areas are better described as having arisen as contact varieties between dialects of previously diverged languages (Bakker and Muysken 1995). This would be analogous to a hybrid species coming into existence as a result of hybridisation between different but related and geographically proximate species in biology.

Mühlhäusler (1996) points out in his work about the linguistic diversity in the Pacific, "What most observers fail to realize is that the identification of languages and their subsequent naming is far from being an act of objective description, and it can constitute a very serious trespass on the linguistic ecology of an area. … In many parts of the Pacific, … we find long chains of interrelated dialects and languages with no clear internal boundaries" (id.: 5). He presents a case from the archipelago of Micronesia, where in linguistic literature commonly 10-20 languages are counted, but where all contiguous dialects are understood by the closest neighbours, but not necessarily the ones further apart. He also notes how the different varieties in a dialect continuum like this can become marginalised and threatened by decay as one variety is chosen as a language of administration and economics, as has been the case with

Trukese in Micronesia (ibid.). Only by becoming a standardised and well described *lingua franca* of the larger region in this way does the variety receive the hallmarks of a distinct standard language, which allows us to disregard the messy, geographically gradual variation that lies in the background.

The linguist R.M.W. Dixon (1997) discussed the birth of new languages from single parent languages. He notes that when two populations, initially speaking the same language, get separated geographically, their language varieties start to drift apart from each other through lack of contact between the populations, slowly becoming divergent enough to warrant calling them different languages. On the other hand, when two regional dialects develop into different languages while maintaining geographical proximity, Dixon argues that this most often happens quite suddenly, and the degree of mutual intelligibility between the varieties drops quickly (id.: 58-63). Divergence in geographical proximity is, according to him, most often an expression of political motivation to gain distance to other dialects. Dixon makes an analogue to the evolutionary biological concept of punctuated equilibria (Eldredge and Gould 1972). This view proposes that the morphological and physiological characteristics of species usually change only little, but equilibria of this kind are disrupted by relatively sudden events of speciation, whereby these characteristics can change quickly, only to return to a new equilibrium phase as a new species (ibid.).

The invention and spread of new technologies or cultural practices are seen by Dixon to be an important cause for punctuations in linguistic equilibria that can otherwise last thousands of years. He suggests that many of the most widespread language families like Indo-European and Uralic have had their beginning in a wide area of linguistic equilibrium, from which the respective ancestral languages started to spread and diverge following a punctuation, following a pattern that later could be illustrated by a tree model, contrary to the earlier equilibrium situation (id.: 97-102). Dixon maintains that punctuations have a finite length, and in the end always give rise to a new period of linguistic equilibrium (id.: 67-96).

This means that during most of the time human languages have been spoken, they may have not followed a tree-like, divergent phylogeny in contrast to the evolutionary history of most animal clades, for instance. Instead, there may have been long periods with widespread 'hybridisation', i.e. pervasive contact influence across linguistically diverse areas. In a case like this, internal classification has only limited feasibility, because historical lineages do not stay within distinct languages and hence do not form distinct branches of a phylogenetic tree. A situation like this is seen to apply to the linguistic diversity of aboriginal Australia, among other cases (id.: 87-93). Also, this possibility has important consequences for the application of computational phylogenetic methods, which are discussed in section 3.2.

Mufwene explored the concept of language as a species in his work about linguistic ecology in the birth of contact languages (2001). He points out that in earlier times, a language was analogised with an organism (e.g. Bopp 1833), and argues that this view has prevented historical linguistics to identify the causes that operate both in internally motivated change, i.e. change both initiated and carried out in a limited community, and externally motivated change that is change brought about through contact between different languages. Mufwene considers the analogue with organisms to prevent understanding the boundaries of both languages and dialects as permeable, which causes linguists to view the processes of internal diffusion and diffusion from other communities as separate mechanisms (Mufwene 2001: 15). Moreover, he points out that individuals in any language community have some variation from each other, which is something not expected of entities comprising a single organism. Instead, he uses the terminology of Hanski (1996) in calling languages *metapopulations*, which in ecology are populations that "consist of '*habitat patches*' connected by '*dispersing individuals*'" (Mufwene 2001: 16, emphases in the original).

Mufwene calls language a parasitic species, because languages can only exist through their hosts, i.e. speakers. He notes that '[many] of the ecological factors that affect a language are not physical features of its speakers but features of other parasitic systems that are hosted by the same individuals, such as culture – which brings along notions such as status, gender, and power – and other language varieties' (id.: 152). However, parasitism is usually considered to be a relationship between organisms that is harmful to the host organism. Clearly, language is not generally harmful to its speakers, but is acquired for its communicative advantages. It could be said instead that languages are in symbiotic relationships with their human hosts. Another thing is whether this terminology should be used of 'species' of very different kinds: an entity consisting of cultural replicators residing in a biological host organism.

2.3.1. Linguistic replication

Linguistic replicators, or linguemes to use the term utilised by Croft (2000), exist in the form of cognitive structure in the mind of a member of a speech community, i.e. as their knowledge of and their capacity to recognise and produce the phonemes, morphemes and syntactic structures used in socio-communicative interaction around them. However, the person who uses the linguemes he has at his disposal to produce certain communicative outcomes is not analogous to the biological organism whose genes function to bring about certain kinds of phenotypic effects. The organism, for one thing, is a phenotypic expression of its genes itself. It is created through the interaction between the genes and the environment in which the organism develops to maximise the dispersal of its genes (Dawkins 1976). In contrary, a person is not formed by the linguemes he has, but instead functions as a host for them, and lets them function for his own benefit. In this they are like viruses, but instead of having pathogenic effects, they are symbiotic. Mufwene (2001, see previous section), among others, calls languages parasites residing in human hosts, but instead of parasitism, the relationship is commensal: the human hosts achieve communicative goals and the linguemes get spread at the same time.

However, even in such simple lifeforms as viruses, one can distinguish between different organisms that all have their own, separate DNA. It is doubtful whether the same can be said of linguemes. A person can know several variants of the linguemes he knows, whereas in biological organisms, there is only ever one allele of a certain gene. As noted, the speaker cannot be compared to a biological organism, but even if we take the utterance as an analogue for DNA as Croft (2000) does, we see that nothing prevents us from using different variants of the same linguemes in the same utterance. It seems that there is in language nothing analogous to the genome of an organism. Linguemes form systems, but those systems are clearly delineated and integrated structures only through in their standardised, normative forms. In linguistic practice, they resemble more 'free-floating', solitary replicators that form temporary alliances with other replicators and even variants of themselves in the grammars of speakers and the utterances they produce.

Also, the replication process of linguistic replicators differs from their biological counterparts. In DNA replication, the double strands of DNA come physically apart, after which new counterpart nucleotides are produced by DNA polymerases. In linguistic replication, however, there is no direct physical contact involved. The hearer of a lingueme who has not heard it before replicates its form through analysing the perceptual properties of the utterance in which they hear it, and infers its communicative function from linguistic and communicative context. This is not a very reliable way to replicate, but successive use of the linguemes in question ensure that the speakers converge more or less to the same form and function.

2.3.2. Linguistic selection

Hull (1988, see 2.1.3 above) noted that a process that involves replicators is not necessarily a selection process. In a population of biological organisms, the environmental conditions may favour all diversity in the gene pool equally, in which case there is no natural selection acting on the diversity. Different alleles may still spread in the population at the expense of the others, but this is not due to differential success of interactors. Instead, all changes is the frequency of alleles result from genetic drift, and this is a case of neutral evolution (id.: 410).

What can be said of language change in this respect? Are there entities that function as interactors in the sense of Hull (1988), whose success and survival account for the direction of change, or is all change in relative frequencies of linguistic variants the product of neutral evolution? Croft (2000, see 2.2.1 above) treats speakers as the paradigm case of interactors in language change and the selection process as essentially a social one, but it is clear that speakers do not fulfil Hull's (1988) literal requirements for being interactors: in his generalised analysis of selection he requires that selection is carried out by the differential extinction and proliferation of interactors. Of course, the fate of linguemes is not determined mainly by the reproductive success of the speakers in whose minds they are entrenched, because linguemes are not tied to individual speakers as closely as are genes to their organisms. Instead, a person may stop using a linguistic form, or start to use a new one, at any point of his life. It would seem that if we want to ascribe the function of interactor to an entity in language change, it would have to be very close to or the same as the replicator, i.e. the lingueme. It is mainly their own 'extinction and proliferation' that causes their differential success. Hull (ibid.) does not require that the replicator and the interactor were different entities, but notes that these functions merely tend to get ascribed to different entities over time[2], as the processes of selection and interaction are different kinds of mechanisms, happening at different scales. Accordingly, it could be suggested that linguemes, as well as other kinds of cultural replicators, are only in the beginning of this process, having yet to develop a way to delegate the different processes to different entities.

Also, it is not necessary that language change is a selection process even if it is determined that in principle it involves replicators that can function as interactors as well. There might not be environmental conditions that would cause some linguemes to be selected more than others. Indeed, Reali and Griffiths (2009) demonstrated that a model representing language learners with Bayesian inference could account for characteristics observed in language change in the same way as the Wright-Fisher model of genetic drift describes changes in a population of biological organisms in the absence of differential selection pressures. Their analysis could account for the S curve (see 2.2.1) apparent in the progress of a linguistic change, among others, with a

[2] In biological evolution, it is seen that during the history of life on Earth, genes have produced ever more complex organisms to function as interactors as effectively as possible, leaving the genes themselves being able to specialise in effective replication.

learning mechanism that imitated the effects of genetic drift, assuming no external selection acting on linguistic variants.

Some writers (Haspelmath 1999, Ritt 2004) have focussed on cognitive factors as well as the effects of other linguemes in their attempts to find adaptivity or selection pressures in language. However, neither is clearly a case where the external environment effects a selection pressure. Functional explanations having to do with cognitive predispositions in effect regard the physical structure of the replicators, the neural structures in which they are entrenched in speakers, not their environment, which should probably seen as the effects on and demands of the socio-communicative context in which they function. Similarly, when different linguemes exert selection pressures on each other, it is questionable whether this constitutes an effect of the environment or just internal stability of the entrenched cognitive structures. This depends in part of whether linguemes should be thought as forming integrated, maximally cooperative structures, or as more solitary entities only forming temporary alliances with each other, a question considered in the previous section.

Croft (2000) sees social factors as causing differential selection pressures on different linguemes. This is not due to any inherent properties of the linguemes themselves, but due to their being in use by groups that are simultaneously associated with diverse social values. Consequently, some linguistic variants get associated with more positive social values than others, and these valuations in turn vary between social groups. In the same way, Mühlhäusler (1996) notes that despite the dizzying diversity of languages in the Pacific has been viewed as a problem by European administrators, it does have a function for the native inhabitants of the area, namely social identification (id.: 14). The diversity of languages is used 'to maintain social groupings at a small and manageable level – and, conversely, to keep other groups at a distance' (Laycock 1982:35). In this way, propagation of innovative features can be seen as adaptive: they function as social markers, allowing members of interacting communities to recognise the membership of each other in these communities and to act accordingly as required by the social situation.

3. Traditional and computational methods for the study of language change

Since the birth of historical linguistics, the main method of exploring and demonstrating the linguistic histories of language families has been the comparative method. The other important, complementary method has been internal reconstruction. Below these methods will be briefly outlined before moving on to quantitative methods that began to be noted in historical linguistics during the second half of the twentieth century.

3.1. Traditional methods in historical linguistics

When the linguist sets out to show that a set of languages have a common origin, their work cannot get off ground without first finding potential cognates in the languages, words or grammatical affixes that seem to be inherited from a common ancestral language, called a *protolanguage*. These shared forms are usually first looked for in what is considered most basic vocabulary or most common inflectional forms, though it is not possible to define 'basic vocabulary' in such a way that would ensure the words have not been borrowed from a language to another. For instance, words for small numbers, which are among the most stable in Indo-European languages (Pagel et al. 2007), are in most East and South-East Asian languages borrowed from Chinese (Rankin 2003: 187-188).

Once a number of potential cognates have been established, they are searched for correspondences between individual phonemes that make up the forms in each language. These correspondence sets are usually arranged in terms of articulatory features of the phonemes in question. As an example, in comparing the Romance languages, one would arrange side by side forms that have the phoneme /ʃ/ in French but /k/ in other languages, for instance French *chèvre* /ʃɛvr/ vs. Italian *capra* /kapra/ and Spanish *cabra* /kabra/. For each of these correspondence sets, a preliminary reconstruction of the proto-phonemes, i.e. the form of these corresponding phonemes in the common ancestral language, can then be performed. When this has been done to all phonemes making up the cognate forms, the words and affixes themselves can be given a reconstructed form in the protolanguage. For the example with Romance languages, we would have many instances with French /ʃ/ vs. /k/ in other languages Based on knowledge of common directionality in sound change, we would conclude that the /ʃ/ is derived from earlier /k/ and not vice versa, which is supported by the presence of /k/ in all languages but French (Campbell 2004: 129-131). Doing this for other phonemes, we reconstruct the original form of /kapra/, which Italian seems to continue. This form is confirmed by Classical Latin *capra* of the same meaning, as (Vulgar) Latin is seen as closely corresponding to Proto-Romance. However, these reconstructed forms are mainly convenient illustrations of shared ancestry and representations of the cognacy of the forms of different languages. In some cases the linguist might be convinced that they have obtained the accurate phonetic form of an ancestral word, but usually the reconstructions are not intended as accurate pronunciation guides (Rankin 2003: 195).

In addition to phonological form, morphemes like words and affixes have a meaning as well. Semantic change, however, does not follow such clearly defined paths as usually does phonological change, and semantic reconstruction of the original meaning is in many cases difficult. Usually the most obvious option is to give as the reconstructed meaning one that is vague enough to encompass most of the meanings in the descendant languages, but in some cases knowledge of history may help. For instance, a widespread cognate word in Siouan languages of North America means 'shoot' in most languages of the family, and 'throw' in just a few of them. However, the Siouan family is known to have started to diverge before the introduction of either modern firearms or earlier bow and arrow, and must have been previously used in the context of spear-throwers, called atlatls. Thus, the reconstructed meaning 'throw' is solid, and the meaning 'shoot' must have been a later development, becoming prevalent in most of the languages (id.: 196-197).

Morphology and syntax can also be reconstructed. For individual grammatical morphemes, comparative method works the same way as for lexical elements, namely through establishing correspondence sets from which original morphemes can be reconstructed. However, grammatical forms are often involved in paradigmatic alternations that are subject to analogical change, which must be taken into account when reconstructing inflectional morphology. In addition to analogy, reanalysis is a complicating issue in syntactic change as well as in morphology. However, in most cases many aspects of the morphological and syntactic constructions in the protolanguage can be established fairly firmly (Campbell 2004).

In contrast to the comparative method, which is used with material with several languages in comparison with each other, the complementary method of internal reconstruction works with patterns in the synchronic system of a single language. As a simplification, internal reconstruction can be used to recover patterns in an ancestral language that show up as unproductive alternations in the forms of a descendant language. In many cases, old alternations may completely disappear during the history of a language, so that the comparative method, which can detect patterns of this kind in related languages, is always the method of choice for reconstruction. (Ringe 2003: 244-245.)

However, in many cases internal reconstruction is useful. One of these cases is caused by the conditioned merger of phonemes, which is accompanied by a complementary split of one of the phonemes. An often-used case to illustrate this is the devoicing of word-final obstruents in Standard German, whereby voiceless and voiced stops and fricatives are merged word-finally. This causes alternations in the singular and plural of certain words, so that the words /taːt/ (<Tat>, 'deed') and /graːt/ (<Grat>, 'edge, ridge'), for instance, have plurals in /taːtən/ and /graːtə/, with /t/, whereas words like /pfaːt/ (<Pfad>, 'path') and /graːt/ (<Grad>, 'degree, rank') have as their plurals /pfaːdə/ and /graːdə/, with /d/. Because alternations of this kind occur in all morphological classes, it is improbable that this pattern is a product of morphological change, and the pattern occurs widely in the most basic vocabulary so that it is unlikely to be a product of borrowing. The most obvious explanation, therefore is that the pattern is a product of a sound change that has caused all word-final original voiced obstruents (in this case /d/) to be unvoiced (in this case, to change into /t/). We can deduce, accordingly, that the original singular and plural forms for these words have been, in the order presented above, *taːt (sg.) / *taːtən (pl.), *graːt / *graːtə, *pfaːd / *pfaːdə, *graːd / *graːdə[3]. (id.: 245-246.)

Internal reconstruction can recover broader patterns, as well. Based on verb forms in the Germanic languages of Gothic, Old Norse, Old English and Old High German, the following partial paradigms for different verbs can be reconstructed:

Present infinitive	*Preterite 3sg.*	*Preterite 3pl.*
*bi:tanã 'to bite'	*bait 's/he bit'	*bitun 'they bit'
*beudanã 'to order'	*baud 's/he ordered'	*budun 'they ordered'

[3] In historical linguistics, an asterisk is used to denote a reconstructed form.

*bindanã 'to tie'	*band 's/he tied'	*bundun 'they tied'
*werpanã 'to throw'	*warp 's/he threw'	*wurpun 'they threw'

If only the alternating parts are considered, the following correspondences show up: /i: ai i/ vs. /eu au u/ vs. /in an un/ vs. /er ar ur/. A generalisation can be made that in the infinite forms, there has been an original /i/ or /e/, in 3rd singular of the preterite an original /a/ and in in 3rd plural of the preterite originally nothing, with *bndun and *wrpun developing into *bundun and *wurpun. This is confirmed by the comparative method, which points to an original alternation in Proto-Indo-European (IE) of /e/, /o/, and nothing. The Proto-Indo-European stem forms corresponding to the third row above would have been *b^{h}end^h-, b^{h}ond^h-, b^{h}nd^h-. (id.: 257-259.)

The main function of the comparative method and its auxiliary methods is to demonstrate that a given set of languages exhibits an amount of forms with the same origin because of shared ancestry, and not because of borrowing and other contact influence. However, once a given language family is in this way established, the next question of interest is the internal classification of that family. Usually, a set of languages can be seen to be so closely related to each other that they clearly form a discrete subgroup within the larger family. Examples of these are Germanic languages within the Indo-European family, which share a large number of developments independent of other Indo-European subgroups, indicating an early divergence from other Indo-European languages that survive today, and the Romance languages of the same family, known to have diverged from an ancestral language close to Vulgar Latin. (Fortson 2004: 8-11.)

After delineating several subgroups of this kind, the question becomes that of determining whether these subgroups have diverged from each other in some discernible chronological order, or whether they spread virtually simultaneously apart from the protolanguage during a single phase of divergence. In the investigation of the Indo-European family, both of these views have had support. While some have maintained that the patterns in which subgroups share signs of early changes can be used to define the order of divergence, others have viewed these patterns as reflecting dialectal variation already present in the Indo-European protolanguage (Proto-IE) and argued that the subsequent divergence of subgroups has not necessarily followed the dialectal borders (Mallory and Adams 1997: 550-556). This possibility makes determining the branching order of a language family uncertain using the traditional methodology without complementary archaeological evidence, but it is rarely possible to connect cultures or migrations apparent in the archaeological record with a given linguistic classificatory group. However, in section 3.3.4 below, it is considered if this is in practice attainable regarding the Indo-European subgroups.

3.1.1. The rise and fall of lexicostatistics

Until the 1950s, only comparative method and internal reconstruction were used to investigate genetic relationships between languages. The comparative method could uncover original forms from which one could deduce what kind of changes had taken place after the break-up of the protolanguage in the histories of the individual languages, and posit subgroupings based on shared changes. However, this required much philological work, and often it was not clear whether some changes were independent and parallel, and so not useful for subgrouping, or dialectal features existing prior to the break-up of the ancestral language. Divergence may have not happened along these dialectal borders, in which case pre-existing dialectal features would be markers not of subgroups but of patterns internal to the protolanguage existing before its divergence.

After Morris Swadesh (1952, 1955) introduced the quantitative methods of lexicostatistics for finding out the pattern of branching into subgroups of a protolanguage, and glottochronology for dating those branchings, however, an enthusiasm for these alternatives for traditional methods soon started. Lexicostatistics was based on collecting for data lists of words for each language to be investigated, nowadays usually called Swadesh lists, which have 100 or 200 meanings for which the most basic word from each language are to be provided. Next, it is determined which of these words are cognate with each other between the languages, and a percentage of cognates between all pairs of languages is calculated. Subgroupings are then determined based on the percentages of shared vocabulary. Glottochronology involved an additional formula for calculating the time of divergence of subgroups, which is sometimes called a 'glotto-clock'. It was based on assuming an universal, constant rate of lexical retention, so that the time since the divergence of two languages (in millennia) was calculated as $t = (\log C) / (2 \log r)$, where C is the percentage of cognates shared and r the assumed universal rate of retention during 1,000 years, for which value 81% was usually used (i.e. on average, 19% of the words on the list were thought to be replaced during a thousand years after the divergence of two languages).

In the following decades after Swadesh's introduction of the methods, they were used quite widely, but among historical linguistics critical voices soon emerged, and after it became understood that the methods could not consistently produce satisfactory results and that there were several critical methodological issues that could not be disregarded, the majority view was established that the methods are not helpful in the study of language history. Among the reasons why lexicostatistics and glottochronology were seen to be inadequate was that the reduction of shared vocabulary to a general percentage loses information about each individual word pairs, which reduces the power of the methods to reconstruct tree topology. Also, it is in these days seen that the assumption of a constant rate of change produces incorrect results in distance-based methods to which lexicostatistics belong, and moreover rates of lexical replacement in reality vary so much in different but closely related branches that this assumption alone causes in many cases discrepancies between the results and actual branching order (Bergsland and Vogt 1962, Atkinson et al. 2005; see also 3.4 for further discussion about the Swadesh lists in particular).

3.2. New quantitative models for language change: adoption of modern biological phylogenetic methods

Around the turn of the millennium, new computational methods started to be applied in investigating the branching order of language families. This time, the methods came from evolutionary biology, and they were backed by powerful statistical inference algorithms designed in finding a *phylogenetic signal* in the data, i.e. the pattern of divergence of the branches in an evolutionary tree.

3.2.1. Biological phylogenetics and computational phylogenetic inference

In biology, the understanding of the phylogenetic patterns leading to present species has been important since the emergence of the theory of evolution. Since then it has been understood that the differences between species has come about through evolution from a common ancestors, and the patterns of evolution cannot be understood without finding out the phylogenetic relationship of the species (Henning 1965).

Phylogenetics is concerned with a form of internal classification of a set of related species. This classification is based on the degree of evolutionary relatedness between them, so that species that have most recently diverged from a common ancestral species are grouped most closely together. The phylogenetic relationships between species under investigation are usually presented in a *phylogenetic* or *evolutionary tree*. This tree can be either rooted or unrooted. If it is unrooted, it only shows the relatedness between species and does not indicate at which point the earliest divergence happened, i.e. where in the tree the common ancestor of all included species would be located. Unrooted trees are most often graphically represented by distributing the investigated species in a circle and drawing the branches of the phylogenetic tree inside this circle. If the tree is rooted, it includes a point from which the first branches diverge, and this point represents the common ancestral species. Often rooted trees are represented with the root at one edge of the diagram with the branches diverging away towards the endpoints that represent the species that provide the data for the tree. A schematic tree illustrating the concepts introduced in this section is given in figure 2 below.

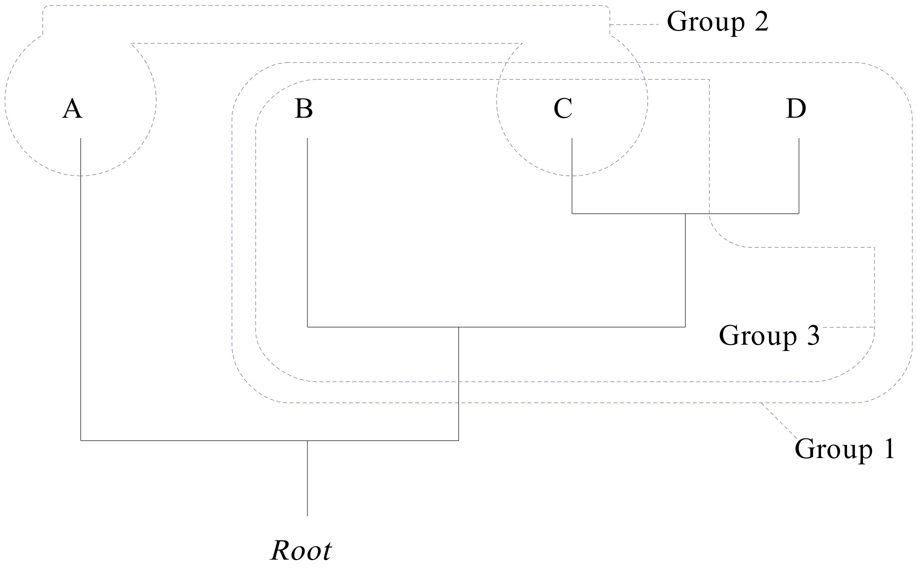

Figure 2. A rooted phylogenetic tree for species A, B, C and D. Species are included varyingly in three groups, of which group 1 is monophyletic, group 2 polyphyletic and group 3 paraphyletic. The terms are explained in the text.

In this figure there is a simple rooted tree, which initially diverges to two branches: that containing species A, and another containing the common ancestor of B, C and D. The latter branch then diverges into one containing B, and another representing the common ancestor of C and D, which then splits into the branches for those individual species. In this case, it could be said that B, C and D form a *monophyletic* group, which includes all species descended from a common ancestor, as well as the ancestor itself (indicated as group 1 in the figure). A grouping that included only A and C (group 2) would be called *polyphyletic*, as it would include only some species from distantly related monophyletic groups and leave out others descending from their common ancestor. If the group were defined as B and C and their ancestors, but not D, it would be a case of a *paraphyletic* group (group 3), with all but one branch of a monophyletic group included. (Henning 1965.)

The tree presented above could be inferred from morphological (anatomical, etc.) characteristics of the species in question, or from molecular data obtained from them in the form of proteins encoded by DNA or the DNA nucleotides themselves. Divergences are deduced from what are determined to be innovative characteristics in an intermediate node in the tree, which are inherited by the descendants of the species that node represents. Shared innovations of this kind can define monophyletic groups and are termed *synapomorphies*. If the other branches retain the ancestral state in respect to these characteristics, these are termed *symplesiomorphies*. They can neither be used to determine splits in the tree nor therefore monophyletic groups, but grouping species according to them can lead to paraphyletic groups. In figure 2 above, suppose that species D has innovated in some respect, and B and C have retained the corresponding ancestral state. Group 3 could have been determined in terms of this symplesiomorphy, which makes it paraphyletic. Finally, characteristics that are similar or identical in different species but have arisen through independent, parallel development instead of shared ancestry are termed *homoplasies* (*convergences* in Henning 1965). Groups defined in this way are most likely polyphyletic. An example of this would be a group including warm-blooded animals, to which would belong mammals and birds, phylogenetically distantly related but having evolved the characteristics of warm-bloodedness independently. (ibid.)

As the first computers became available as tools for scientific study in the 1950s, they were soon adopted for computing algorithms for what became known as numerical taxonomy. The first phylogeny inferred by computational methods was by Michener and Sokal (1957) and involved a clustering method using morphological characteristics of different bees (Felstenstein 2004). Later, foundations were laid for three of the most widely used phylogenetic inference methods in a short time by the pair of colleagues Anthony Edwards and Luigi Luca Cavalli-Sforza. They started developing phylogenetic inference methods for constructing phylogenetic trees of human populations around the World, trying to determine their patterns of migration. As a consequence, they came to present for the first time both parsimony and likelihood models (Edwards and Cavalli-Sforza 1963) and later the distance matrix method of least squares (Cavalli-Sforza and Edwards 1967), all three of which were to become important alternative methods for subsequent development. In the next section it will be seen that all three also found their way into the study of language.

3.2.2. Phylogenetic methods in language history

During each decade, the statistics and mathematics behind the phylogenetic inference algorithms became more sophisticated, and besides analyses of different biological species, the methods were used with increasing accuracy to construct phylogenetic trees, among others for the dispersal of human populations around the World, starting with Edwards and Cavalli-Sforza (1963), as discussed above. This led biologists and linguists to consider if these same methods, working so well with biological data, could be used with linguistic data, as well. For these methods it does not matter what the data is in its nature, so long as it evolves and splits into different lineages. As discussed in section 2.1., these characteristics have long been thought to apply in many ways to language change in a way similar to biological evolution.

Accordingly, many linguists started to notice the potential applicability of these new tools for their own subject matter, the divergence of languages. Many felt compelled to apply the phylogenetic methods to the most thoroughly investigated and most historically important language family, the Indo-European languages (e.g. Ringe et al. 2002, Rexová et al. 2003, Gray and Atkinson 2003, Nakhleh et al. 2005b), in order to see if these methods could give new answers to old, unsolved problems of the field. These attempts are evaluated in the following section, but first we go through the various phylogenetic methods as used in linguistics and their applicability to model the divergence of languages.

Both Nakhleh et al. (2005a) and Holm (2007) have discussed the characteristics, advantages and weaknesses of several phylogenetic methods in terms of their suitability for inferring patterns in Indo-European language history, with Nakhleh et al. testing the methods on a dataset comprising lexical items from various IE languages. Both come to the conclusion that the so-called distance-based methods, namely UPGMA and NJ ('neighbour-joining'), are not the most powerful methods at finding a phylogenetic signal in linguistic data. Holm explains that is due to the fact that these methods are phenetic ones and measure perceived distance, which easily leads to what he calls a *proportionality trap*: if only the number of shared characters between different languages are taken into account, and not the historical patterns that might have brought them about, languages could get placed together by a phenetic method if they happen to share a large number of retentions from their common protolanguage, disregarding any actual intervening branchings if they have lost a larger amount of cognates, for whatever reason (Holm 2007: 184-186). Therefore, it may be best to approach with caution using linguistic data with distance-based methods. One of these is UPGMA, which Nakhleh et al. (2005) dropped from their comparison right after the initial results, which were immediately seen to contradict with some of the most well established historical relationships among the IE languages, in contrast to all other methods they tested.

Other phylogenetic methods used in linguistics are based on character states. In biology, these methods use as their data either morphological characteristics, DNA sequences or protein data. In the case of DNA sequences, a position in a sequence corresponds to a character, and the nucleotide in that position corresponds to the state of that character. In applications to language history, these characters are usually lexical items whose states are assigned based on their being cognates (i.e. words inherited from the same ancestral language without being borrowed). In practice this means that when languages A and B share a cognate and C has an individual replacement for that lexical item, the corresponding character has the same state for languages A and B, and a different one for C (see e.g. Gray and Atkinson 2003). In a data matrix, languages A and B might have the state '1' for this character, and language C '2'. Other languages could have states '3' etc., and if the algorithm permits, a character can have several states if no single cognate is seen as the primary word for a certain meaning in a given language.

Optionally, the same data can be arranged in a binary matrix that requires all character states to be either '0' or '1', for 'absent' and 'present', respectively. This means that if there are three cognate sets to which the words for a given meaning belong in a set of languages, three different characters must be utilised. Thus both languages A and B might receive as states for three characters '1 0 0' if they only have the first cognate, and language C '0 1 0' if it has only the second. In a binary matrix, it is simple to encode the presence of several equal cognates: if a language has members of both the first and the second cognate sets as equal forms for a given meaning, this can be encoded simply as '1 1 0', with no multiple states for a character needed. There are different types of character state methods, but they are generally seen to be better able to infer evolutionary phylogenetic patterns in the data than distance-based methods.

One simple type of character state methods is Maximum Parsimony (MP), which calculates phylogenetic trees that would result from the minimum amount of replacements in the lineages. Holm (2007: 192) argues that this makes it not applicable to linguistic change as historical events and varying tendencies can cause change to be very rapid in some branches, which causes MP to place them towards the periphery of the tree from their actual position, a phenomenon known as *long branch attraction.* Nakhleh et al. (2005a), however, note that in their comparison MP produced trees very similar to the results of their unweighted Maximum Compatibility method.

Maximum Compatibility (MC) is another character state method, which works on the criterion of finding trees that have the least amount of incompatible characters, i.e. the algorithm tries to find a tree where different character states are found in different branches with the least possible amount of splitting of different states to different branches. Holm criticises the method for not being able to take into account significant borrowing and areal influence, and for not distinguishing between shared innovations in the posited branch and shared retentions from a common ancestral language (2007: 193). However, Nakhleh et al. (2005b) included in their model, which they call perfect phylogenetic networks, the possibility of borrowing, which allowed the production of trees with no incompatible characters (see next section). Nakhleh et al. (2005a) used both an unweighted and a weighted MC method. In the latter case, some characters were assigned larger weights so that the model would try to optimise these character at the expense of the lesser-weighted characters. This way, the method did not attempt to find a tree with the smallest amount of incompatible characters, but the tree with the smallest weight. The 'perfect phylogenetic' algorithm employed by Nakhleh et al. (2005b; see next section) was of this type.

Finally, most linguists who have applied a computational phylogenetic method to language history have come to the conclusion that the one that produces the best results is a type of a Maximum Likelihood (ML) method. Most often, a heuristic algorithm called Monte Carlo Markov Chain (MCMC) Bayesian inference is used to maximise the likelihood function (see e.g. Dunn et al. 2008). This is computationally a very complicated method, and requires a lot of processing time. However, it is often seen as modelling evolution most realistically. Holm (2007: 192) dismisses it on the grounds that it assumes a rate of change, which he sees as inapplicable for language change. However, rates of change can and do vary in different lineages and in different genetic loci in biological evolution, as well, and MCMC Bayesian inference methods must be able to model widely differing rates of change. Atkinson and Gray (2006) and Pagel (2009) illustrate how Bayesian inference models can be made to take into account the possibility that in some branches characters are replaced much faster than in others. However, in the next section we will discuss the application of this approach to Indo-European languages and raise some doubts over its results.

3.3. Phylogenetic analyses of Indo-European languages

As mentioned, phylogenetic methods have been most widely applied to the investigation of the branching order of Indo-European (IE) languages, as this is the most thoroughly researched language family with most historical data accumulated during the time of historical-comparative linguistics. Consequently, phylogenetic analyses have been able to use detailed philological research as guides for arranging their data. The methods are being constantly improved both in respect to their statistical basis and to their modelling of linguistic data. The computational phylogenetic analyses of IE languages agree in many respects, but have significant differences depending on model and data used. In figures 3 to 5, some trees have been reproduced that have been presented as results of these analyses, whose main methods and results are discussed next.

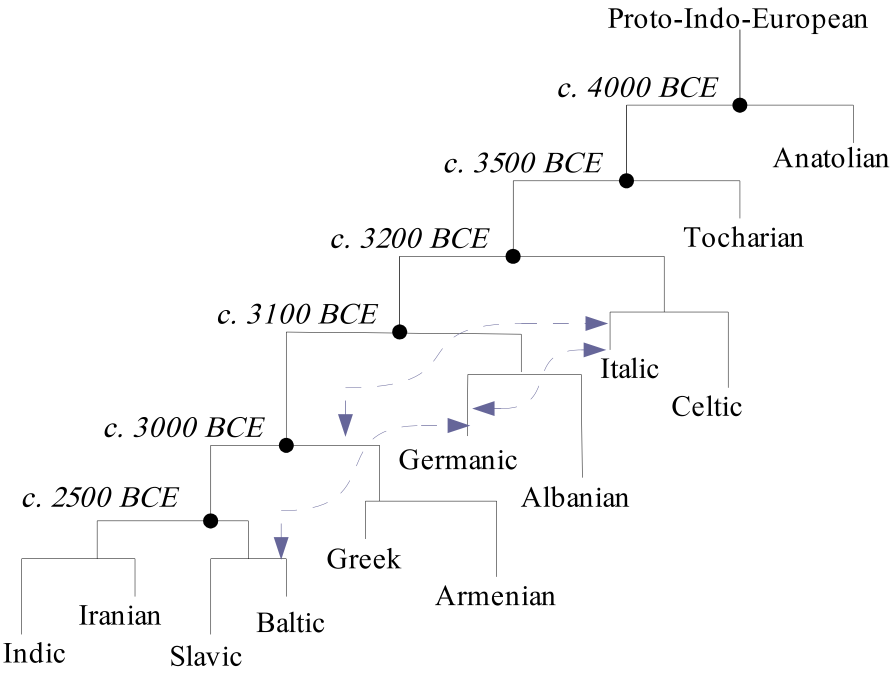

Figure 3. One Indo-European tree generated by the weighted maximum compatibility-type Perfect Phylogeny algorithm of Nakhleh et al. (2005b; discussed in section 3.3.1 below). Reproduced by this author from Tree A in their figure 12 (id.: 403), with only major branches included for simplicity and comparability. Dates for splits are from their diagram, and estimated by the authors from archaeological evidence instead of being results of their computational method. One solution with three contact edges (see text) is indicated with dashed lines between branches.

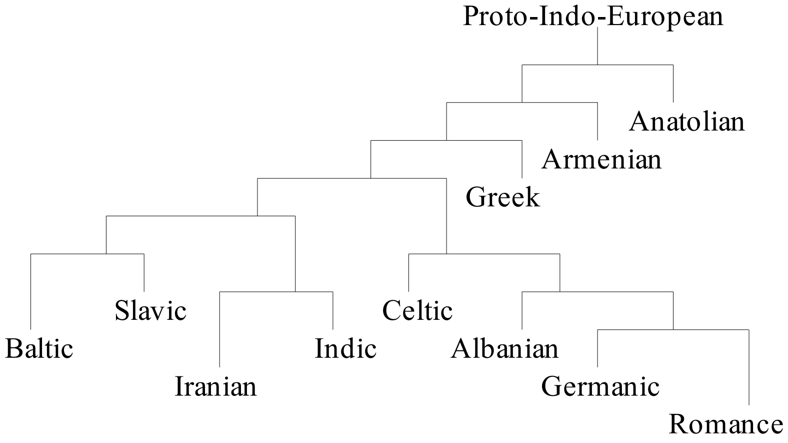

Figure 4. The consensus tree from a sample of trees generated by a maximum parsimony algorithm using the standard multistate matrix of Rexová et al. (2003; discussed in section 3.3.2 below). Reproduced by this author from their figure 1a (id.: 123), using 'Anatolian' and 'Slavic' for their 'Hittite' and 'Slavonic' for consistency. Splits were not dated.

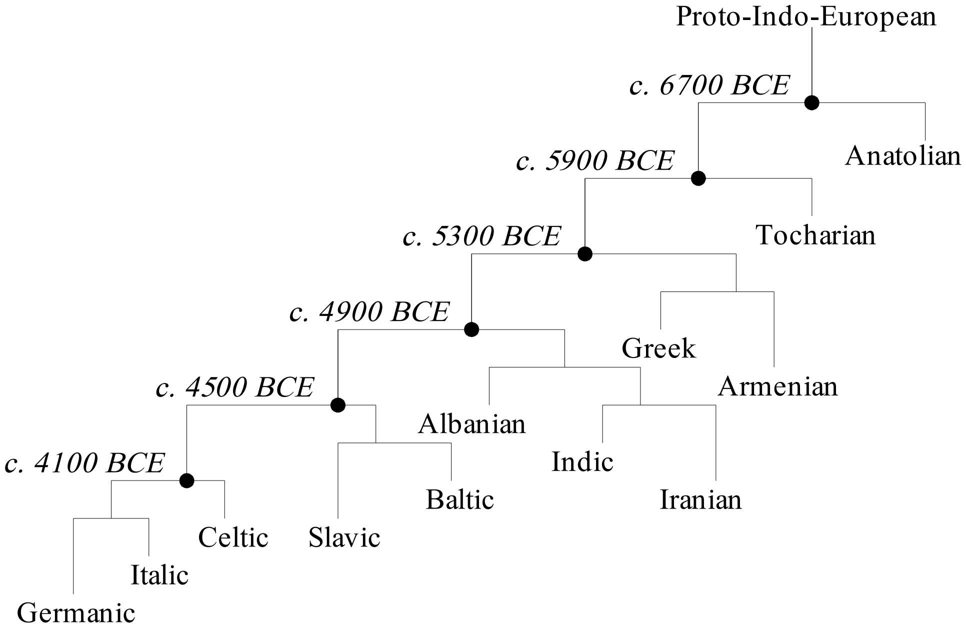

Figure 5. Consensus tree from an initial sample generated by a Bayesian MCMC phylogenetic algorithm in the study of Gray and Atkinson (2003), with divergence times estimated by the algorithm. Reproduced by this author from their figure 1, retaining major branches and their divergence times, converted from years before present to years Before the Common Era.

3.3.1. Case 1. Maximum compatibility and Perfect Phylogeny

The team led by Don Ringe has been developing a weighted maximum compatibility method for modelling of language divergence. In their original article (Ringe et al. 2002), they started by setting out basic guidelines for a phylogenetic method to be able to accurately model patterns of language divergence. They note that if one chooses to include phonological characters in the analysis, there is a risk that similarities between different branches have developed independent of each other, i.e. they are homoplasies instead of synapomorphies that would be useful for establishing a subgroup in a tree. On the other hand, morphological characters such as specific inflectional or derivational endings are unlikely to develop independently, but when one has different forms in different branches with no knowledge of which was the one in the ancestral language, it is difficult to ascertain which are symplesiomorphies (i.e. shared retentions from the ultimate ancestor) and which are synapomorphies (i.e. shared innovations originating in an earlier subgroup). They proceed to note that the possibility of back-mutations of linguistic characters should be excluded as all linguistic data that is useful for phylogeny is sufficiently complex to make back-mutation infeasible, so that if all borrowings and other products of language contact as well as parallel developments (i.e. homoplasies) could be excluded, a phylogenetic tree of the type they call Perfect Phylogeny could be produced. They note that since all characters in the data would be required to be compatible with the tree for it to fulfil perfect phylogeny, a maximum compatibility method would be preferable to other kinds of phylogenetic methods like maximum parsimony.

Their test run on a set of Indo-European languages was made utilising a set of 370 characters, of which 22 were phonological, 15 morphological and 333 lexical. Keeping in mind the caveats mentioned above, they chose the phonological characters in a way that was most helpful for subgrouping, for example by avoiding too simple and commonplace changes. For lexical characters, they noted that in some cases, they had to include polymorphic characters, so that for some languages there were characters that received more than one different state. For instance, English has two different words, *little* and *small*, corresponding to single lexemes in most languages. Accordingly, for this character English received two different states for each cognate set that these words belonged in. Their algorithm was required to be able to handle polymorphic characters without treating them about a priori assumptions about tree topology. The test run did not result in a perfect phylogeny, but the best tree had 18 incompatible characters. All of these that were incompatible with the positions of the primary IE subgroups in the tree involved the Germanic subgroup, i.e. these characters would have had to develop the same state independently in Germanic and in another primary subgroup. Noting this, another try was run without including Germanic languages, which resulted in a best tree with only 4 incompatible characters, all lexical.

In the next article, (Nakhleh et al. 2005), the team put forward their results obtained through use of a model that could produce perfect phylogenetic trees where no characters remained incompatible. This was attempted by making the model produce networks instead of pure trees. First, a maximum compatibility tree was generated. Second, the model was required to add a minimum possible amount of contacts between branches that were involved in character incompatibilities in the best tree, mirroring possible contact influences between the languages in question. As this modification had the effect of making the resultant diagram topologically a network rather than a tree, they accordingly modified their name for the approach to Perfect Phylogenetic Networks (PPN). They then proceeded to analyse the Indo-European languages using their weighted maximum compatibility-type PPN algorithm, employing similar data than the earlier attempt in Ringe et al. (2002), but this time using 22 phonological characters, 13 morphological and 259 lexical ones with a total of 294 characters. Their dataset included several ancient languages that are only fragmentarily attested, leading to their having as much as 85 % of characters unknown for those languages. However, it was seen that this should pose no problem because their model was able to work with fragmentary data, and being character-based instead of distance-based, didn't require the languages to be at the same time depth.

Their model, applied to the IE dataset, initially produced treelike maximum compatibility diagrams in the same way than in Ringe et al. (2002). Then, contact edges were added between branches of the best trees to produce PPNs, for which a minimum of three contact edges was needed. For instance, the algorithm returned 16 solutions with three contact edges for their Tree A, which had the least amount of incompatible characters, of which the authors disqualified all but three based on historical considerations of what languages could have been in close contact during the time required. In figure 3 above, Tree A of Nakhleh et al. (2005b) is reproduced, having the least amount of incompatible characters. It had three plausible network solutions with three contact edges, one of which is indicated in the figure.

This method, then, was seen to be able to produce tree-like phylogenies, and additionally to suggest events of language contact to explain patterns not compatible with tree-like inheritance. However, the algorithm returned a large amount of possible solutions for these language contacts, most of which could be seen by the authors to be implausible considering the history of the groups speaking these languages. Accordingly, most plausible contact scenarios had to be hand-picked. The authors drew their trees in a diagram that indicated the dates of the splits, but these were not obtained through the method, but inferred from historical and archaeological record. Their Tree A, mentioned above, exhibits the largest degree of similarity of the phylogenetic trees discussed here to the tree in figure 6 below, inferred from linguistic palaeontological and archaeological evidence. However, it is questionable to what extent this is a result of the method used by the authors. As discussed, many aspects of their results, like the dates in their diagrams, were not obtained through their computational methods, but also from linguistic palaeontology.

3.3.2. Case 2. Maximum parsimony analysis of Indo-European

The analysis of Rexová et al. (2003) used the method of maximum parsimony, which has been criticised above on the grounds that its basic assumptions are not applicable to language change, namely because the model builds trees that require the least amount of change to have happened. When some branches change faster than others, this results in their being placed further out from their actual position in the tree. However, instead of dismissing a method on theoretical grounds, it is useful to see how it performs in practice. Rexová et al. used as their data a Swadesh-type list (see next section) of 200 lexical characters, compiled by Dyen et al. (1992) for lexicostatistical research. They arranged the data into three different datasets, for which phylogenetic trees were computed separately. They had two versions of a multistate matrix, where many words could be assigned for each meaning. The third was a binary dataset for which all cognate sets received their own characters, with states coded as present/absent ('1'/'0') accordingly for each language.

The algorithm, using the standard multistate matrix, produced a consensus tree that included all of the uncontroversial IE branches, as well as on higher levels the branches Balto-Slavic, Indo-Iranian and, with somewhat lower support, 'Romano-Germanic' (i.e. Italic grouped with Germanic). The tree also combined the Balto-Slavic and Indo-Iranian into a still larger subgroup, as well as included Albanian and Celtic in a separate subgroup with the smaller one comprising of Italic and Germanic, however with a very weak support. This tree is reproduced in figure 4 above. The altered matrix produced a tree with a very similar topology with only minor differences, whereas using a binary matrix reordered the tree significantly. Albanian, the language with the smallest amount of inherited IE vocabulary, was shifted near the base of the tree to diverge after Anatolian, and Iranian, with similarly high amount of replacements, received a new position as splitting before Indic. Using a binary matrix with the maximum parsimony method, then, seemed to accentuate the problem of long branch attraction. For the multistate coded matrices, the results were in many respects in line with the other approaches discussed here.

3.3.3. Case 3. Controversial dating of splits using Bayesian inference

Only one team of researchers have defied the bad reputation of glottochronology by including in their phylogenetic studies of the Indo-European languages estimates of times for the splitting of languages, namely that of Russell Gray (Gray and Atkinson 2003, Atkinson et al. 2005, Atkinson and Gray 2006), which has done more work on phylogenetic analyses of the Austronesian language family as part of their project Austronesian Basic Vocabulary Database (e.g. Greenhill et al. 2008). What was significant in their analysis of the Indo-European family was their conclusion that the analysis supported the so called Anatolian hypothesis of Indo-European origins, which is perhaps the minority view in Indo-European linguistics compared to the alternative of the Pontic-Caspian Steppe homeland hypothesis (see next section).

Gray and Atkinson (2003) used a maximum likelihood-related Bayesian MCMC inference method (see 3.2.2) that allowed differing rates of change, modelled by the gamma function, and used as their data binary matrices that coded the presence or absence of a cognate word in each language. Their method found all the principal IE branches and included them as distinct subgroups in the resulting consensus tree (i.e. there were no branches with languages that would be known to belong to another branch), and also found several sub-groupings of the IE languages that are supported by many traditional IE scholars, such as Balto-Slavic, Graeco-Armenian, but not Italo-Celtic. In their consensus tree, the first branch to diverge from Proto-IE was Hittite (Anatolian), followed by Tocharian. What led Gray and Atkinson to the conclusion that their study backed the Anatolian hypothesis was the result that the Anatolian branch separated c. 8700 years before present (BP), i.e. 6700 BCE, and Tocharian 7900 BP (5900 BCE). The consensus tree obtained through their analysis is reproduced in figure 5 above.

3.3.4. Linguistic palaeontology vs. computational phylogenetics in early Indo-European history

As noted in the previous section, Gray and Atkinson (2003) came to support the Anatolian hypothesis of Proto-Indo-European origins, which is one of the two principal hypotheses regarding the place and time of the language community from which all Indo-European languages descend from. Most widely supported among linguists is the hypothesis that Proto-Indo-European was spoken around 4000-3500 BCE on the steppes of what is today Ukraine and southern Russia, north of the Black Sea and the Caspian Sea, from which the area receives the appellation of Pontic-Caspian Steppe (for this hypothesis see e.g. Mallory 1989). This conclusion is based on applying what is sometimes called *linguistic palaeontology* to the reconstructed language and associated archaeological evidence. For instance, there are forms belonging to the reconstructed Proto-IE vocabulary that probably have been used to denote agricultural practices and concepts of horsemanship, which, connected to the observation of the early Indo-Europeans as expansionist invaders points to the archaeological cultures of the Pontic-Caspian Steppe, among whom the widespread use of the wheeled chariot contributed in their expansion in Eastern Europe and Central Asia (Anthony 2007).

In contrast, the Anatolian hypothesis, formulated by archaeologist Colin Renfrew as part of a wider framework that seeks to explain the spread of several language families with the concurrent spread of agriculture (Renfrew 2002), argued for a homeland in Anatolia (modern Turkey) and the Aegean c. 7000-6500 BCE, much earlier than the steppe hypothesis (Renfrew 2001). Renfrew noted that the archaic nature of Hittite and other languages of the ancient Anatolian branch supports the notion that it split from the other IE languages before any other branch, which is agreed to by many Indo-European scholars. He used this as evidence in favour of his view that some of the first farmers in Anatolia, those of Çatalhöyük c. 7000 BCE, spoke Proto-Indo-European or a language closely related to it, and the IE languages spread from this area into Europe with the people that also brought farming along with them, with the possible intermediate phase of the Pontic-Caspian Steppe in 4000-3500 BCE, allowing the inclusion of the competing hypothesis (Renfrew 2001).

What any study supporting the Anatolian hypothesis must explain is the appearance of cognate words for horse and wagon technology in almost all IE branches (Mallory and Adams 2006). The first archaeological evidence for wheeled vehicles dates to 3300-3200 BCE, so words for wheeled horse-powered transport could not have originated much earlier than this (Fortson 2004). Atkinson and Gray's (2006) answer to this is that either these terms could have been independently coined using the same roots (for instance, the reconstructed form $*k^wek^wlo$- for 'wheel' is agreed to be formed from the Proto-IE root $*k^wel$- 'to turn', and Atkinson and Gray argue that it could have been formed at three different times in different branches according to their model), or the terms are borrowings from one IE branch into the others. The first argument becomes implausible when one considers the amount of these cognates and the required fortuitous coincidences leading to a large amount of chance agreements between the branches, but the second receives some support from traditional Indo-European studies as Mallory (1989) notes that words similar to the IE wheeled vehicle terms appear in several non-Indo-European languages like Sumerian, Semitic and Kartvelian languages, which indicate that they have been widely borrowed in ancient times. Also, Mallory and Adams (2006) note that convincing cognates seem to be lacking from Anatolian, which means that if it has indeed been the first IE branch to diverge, this could have happened already before the invention of wheeled vehicles.

However, as Anthony (2007) notes, the wheel vocabulary seems to have been inherited in the non-Anatolian languages from an ancestral language, showing the same regular sound correspondences than other inherited PIE words. This makes the early dates of Gray and Atkinson (2003) for the divergence of the later branches problematic, and pushing them forward in time would leave a gap of several thousands of years between the split of Anatolian and any further splits, which would in turn make it hard to account for the very slow overall rate of change of the ancestor of all non-Anatolian languages and the absence of any documented branching during that period. To support the Pontic-Caspian Steppe hypothesis of the Proto-Indo-European homeland, Anthony

(2007) draws together an impressive account of the migrations and dispersals of probable Indo-European-speaking people from the archaeological evidence regarding the prehistory of Eastern Europe and Western and Central Asia, citing also early contacts with (Proto-)Uralic speaking populations as deductible from ancient borrowings. Like many others (see Mallory 1989), he connects the Proto-Indo-Europeans with the Sredni Stog culture of the Western Pontic-Caspian Steppe c. 4500-3500 BCE and the later Yamna culture of the same region. He sees the migration of the Suvorovo-Novodanilovka Complex, possibly a ruling elite among the Sredni Stog, to the lower Danube valley as the event that separated the Pre-Anatolian language from the other ancestral IE dialects, not earlier than 4200-4000 BCE (2007: 239-58).

Next, he dates the separation of Tocharian as happening 3700-3500 BCE, as the eastern Afanasievo culture, often connected with the speakers of Tocharian, developed from the Yamna-related Repin culture in the south of the Urals. Afterwards, 3100-3000 BCE, the Yamna culture extended up the Danube valley in the west, which Anthony connects with the separation of Proto-Italo-Celtic, and he argues that Germanic diverged with the imposing of a Yamna ruling elite on the indigenous Corded Ware Culture, 2800-2600 BCE. Finally, he finds a correlation to the south-east spread of the Indo-Iranian branch in the Sintashta culture in the south of the Urals 2200-2000 BCE, from which they would have spread southwards towards the Iranian Plateau and India (Anthony 2007: 305-6). Of course, there is no way to tell for certain what language these archaeological cultures used, and the Sintashta culture, for instance, might have preceded the actual speakers of Pre- or Proto-Indo-Iranian, or it might have succeeded their expansion, and have been subsequently run over by later Indo-Iranian expansion. However, Anthony's account is the most thorough and convincing attempt at correlating the spread of archaeological cultures with the divergence of the IE branches. A dated tree incorporating this scenario is presented in figure 6 (compare with figures 3 to 5 above).

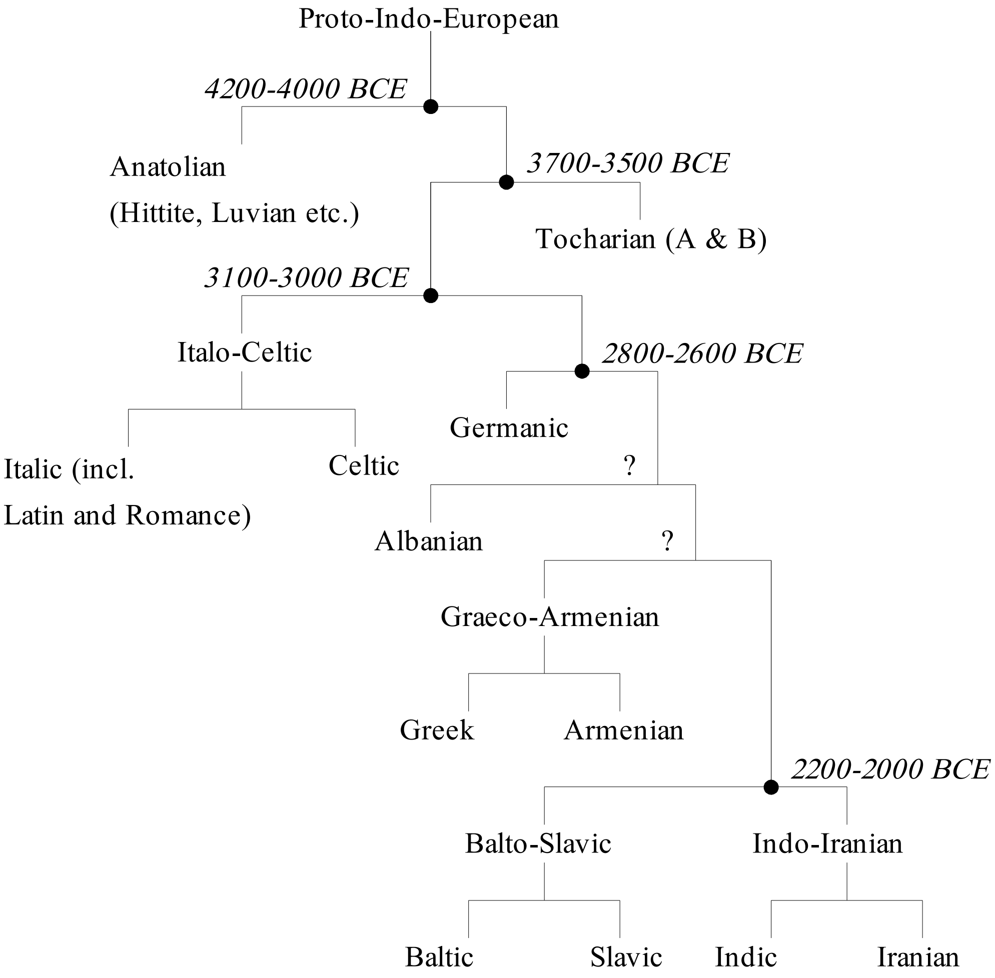

Figure 6. Tree of primary Indo-European branches based on linguistic palaeontology and archaeology, especially Anthony (2007), from which dates for some important splits.

In light of this, it seems that if the conclusions of Gray and Atkinson (2003) were to be upheld, much additional work is needed to account for the incongruence between linguistic and archaeological evidence on one hand, and on phylogenetic computational methods on the other. Traditional methods in historical linguistics and archaeology have been tested by time and thorough investigations, whereas phylogenetic methods have only started to be applied to modelling language prehistory. Different methods return different results with different kinds of data, and it remains to be evaluated which results are closer to the truth. Dates for divergence times of different branches, especially, need to be viewed with caution, but with improving methods and understanding of their applicability to language history, attempts at dating should not be abandoned, but constantly improved instead. Also, using lexical items coded as states in binary character matrices might not be the best data format for detecting a phylogenetic signal. Ringe et al. (2002) and Nakhleh et al. (2005a,b) were able to generate trees with very few incompatible characters using both lexical and structural (complex phonological and morphological) data, and all of the incompatibilities were lexical. Furthermore, the results seem to line up well with archaeological evidence. Bayesian inference methods can be the most powerful ones, but in their application the model used must be carefully configured to be applicable to linguistic evolution, and the data should preferably include structural features of the kind that are resistant to independent development and borrowing. These questions will be more closely discussed in the next section.

3.4. Linguistic data in phylogenetic methods

Phylogenetic methods have been designed to be used with data comprising, for instance, sequences of nucleotides in the DNA that is examined. Thus, in a biological phylogenetic analysis there might be two different sequences 'AATCGTACAGG' and 'AAGCGT.GAGG' in the data to be compared, where '.' indicates the absence of a nucleotide in the second sequence corresponding to the seventh nucleotide in the first one. For these two sequences, we can calculate a figure called the Hamming distance, which is a simple measure of observed distance between sequences. In this case, this would be the number of pairwise differences divided by the number of loci, e.g. 3/11, yielding a Hamming distance of 0.2727. Transferred most directly to linguistic comparison, this approach would require us to find cognate words, e.g. English /hɛd/ 'head' and German /haʊpt/ 'main' and treat them similarly as sequences: 'h ɛ . d' vs. 'h aʊ p t'. We could give a Hamming distance of 0.75 for this phoneme sequence, because they only have /h/ in common, with three of the four characters different. (Holm 2007.)

However, the relationship between cognate words in their accurate phonetic form and the relationship between related nucleotide sequences in DNA is different. The presence of cognates in different languages in the first place tells more about their genetic relationship than does the more mundane appearance of related DNA sequences in related biological species. The amount of shared cognates between languages is already a indication of their relatedness, whereas we can expect to find similar sequences of DNA in even distantly related species, albeit having probably widely differing nucleotides in those sequences. Furthermore, the difference in phonological form in cognate words tells us little about their relatedness. Even dialects used in the same language community can have divergent pronunciations, whereas distantly related languages might exhibit comparatively little phonological divergence but only have a small amount of shared cognates. Holm (2007) points out that it has been established that simple phonetic correspondences can too easily arise independently of each other, among other reasons because sound changes often follow similar paths cross-linguistically, which makes them of little use in attempts to unravel patterns of shared ancestry. For instance, the reconstructed 'voiced aspirates' of Proto-IE ($*b^h, d^h, \hat{g}^h, g^h, g^{wh}$) have changed into ordinary voiced stops in at least Balto-Slavic, Celtic and Albanian branches, seemingly independent of each other.

This difference between biological and linguistic data has practical consequences for the form of the data used in linguistic applications of phylogenetic methods. In what Holm (2007: 181-2) calls the onomasiological point of view, every character is connected with a given meaning. The state of the character is not the phonetic form of an ancestral phoneme, but is instead determined by the cognacy of the word having that meaning in the given language, so that the English form for 'head', i.e. *head*, would receive the same state as Latin *caput* (both from Proto-IE **kaput-*), but the German form *Kopf* a different one (as seen above, German has a cognate of the English and Latin forms, but in a different meaning). If this kind of data is incorporated in a matrix where each character receives a state on the basis of which cognate set the word belongs to, in the present example English and Latin would receive for the character 'head' the state '0', for instance, German '1', and another language having a word not being a cognate with either '2', and so on.

The onomasiological approach is the most often used option in linguistic phylogenetic analyses. In practice, this usually takes place in the form of using a so-called Swadesh list, after linguist Morris Swadesh, who devised it for use with his original glottochronological method (Swadesh 1952, see 3.1.1 above). Aside from critique directed against other aspects of glottochronological methodology, the use of Swadesh lists itself has been brought into question. There are many versions of the list, with most often 100 or 200 meanings. The basic assumption, however, is always the same: the meanings chosen should reflect universal concepts, present in all languages, and sufficiently independent of culture so that borrowing of words for these meanings is rare. This is seen as sufficient to ensure that languages replace words for these meanings relatively rarely, and that historical events affecting the linguistic communities have relatively little effect on the replacement rates for the meanings in the list.

However, all of these assumptions have been shown to be incorrect. Not all languages have separate words for all meanings on different Swadesh lists. For example, the 200 word lists often have colour terms like 'green', 'red' and 'yellow', but the famous research by Berlin and Kay (1969) showed that in some languages the only colour terms were what could be translated as 'black' and 'white', or 'dark colour' and 'light colour'. Also, some meanings in the lists correspond to many forms in some languages, none of which can be shown to be most neutral or in other ways primary. For instance, the lists usually include the meaning 'we'. However, many languages distinguish between inclusive vs. exclusive we ('me and you, and perhaps others' vs. 'me and others, but not you'), and it is not clear which we should use as representing this item in the list. Also, even with such basic and seemingly culture-independent meanings, replacement rates can and do vary. Since the beginning of historical attestation, Icelandic has lost very little of its core vocabulary, much less than the glottochronological assumptions maintain, because of the long-term isolation of its speakers. At the same time, English has lost far in excess of the expected amount of basic words due to intensive contacts with Scandinavian and French-speaking people. (Campbell 2004: 201-10.)

It would be interesting to investigate whether the universal semantic primitives uncovered during the semanticist Anna Wierzbicka's research (see e.g. Wierzbicka 1996) would be an improvement to the Swadesh lists, but there have been no attempts to use them as data in phylogenetic analyses. In any case, Swadesh lists are a useful and sufficiently simple way of using lexical data in phylogenetic analyses, if any items for which there are problems of the kinds discussed above are removed for all languages. Thus, in case of wide-spread borrowing of words for some items in the Swadesh list in some of the languages under analysis, or in case of ambiguity of choice between equally neutral words or single words corresponding to many items in the list, those items should be excluded from use as characters.

Concerning the optimal size of the Swadesh list in terms of number of meanings, Embleton (1986: 89-93) notes that in an analysis using simulated data, increasing the meanings from 200 to 500 had little impact on the accuracy of the analysis. In contrast, however, accuracy decreased significantly when a 100 word list was used, compared to the list of 200 words. McMahon and McMahon (2003) note that these results are based on the glottochronological assumption that all words on the list are replaced at approximately the same rate, which in practice is not true. However, this is a well known problem in biology as well, and phylogenetic methods are designed to take it into account. Also, it is probable that variable rates of change do not change the basic conclusion about the lengths of the lists. Therefore, a list of 200 meanings should probably by used when using lexical form-function pairings as characters in phylogenetic methods. When the size of the list is increased, the advantages of having a larger dataset are offset by difficulties in finding so many words belonging to basic vocabulary that would be relatively resistant to borrowing.

Holm (2007: 182) points out that another option to the onomasiological point of view, choosing meanings as characters and their membership in a cognate set as states, is what he calls an etymological point of view. This would involve taking as the starting point a number of word forms reconstructed for the protolanguage from which the languages under investigation are known to have descended, and coding for each language the presence or absence of a form descending from that ancestral form, whatever its meaning. If we were to use the Proto-IE form **kaput-* as a character in this way, English, German and Latin would all receive the state 'present', despite German having a different meaning. Holm argues that using forms rather than meanings as a basis would increase the reliability of the method used, because meanings change much more rapidly than forms are lost. It is not clear what kind of difference this would make in a phylogenetic analysis, as to the knowledge of the present author this kind of approach has not been taken with phylogenetic methods.

However, words are only part of the machinery of languages inherited from a protolanguage by its branches. Establishing sound correspondences is an important part of the comparative method, but above we have touched on the issue of why phonological characters may not be optimal data for phylogenetic approaches. Morphological correspondences like inflectional paradigms, especially suppletive ones[4], have been traditionally seen as perhaps the optimum kind of evidence for demonstrating membership in the same language family, because they are some of the patterns most resistant to borrowing (Campbell 2003). Syntactic patterns have generally been in a lesser role in historical linguistics, and it is thought that they are more easily influenced by areal convergence than other kinds of structures (Campbell 2004: 338-43). In any case, a reasonable assumption until otherwise shown would be that data incorporating structural features in addition to lexical items would be optimal for finding a phylogenetic signal compared to pure lexical data.

[4] As examples of a suppletive paradigm, where some inflectional forms are derived historically from different word roots, English paradigms with *be* vs. *is*, *go* vs. *went* and *good* vs. *better* could be mentioned. The fact that in German, a direct cognate paradigm for the latter one can be found (i.e. *gut* vs. *besser*), provides already by itself strong evidence that these two languages are related trough inheritance from a common ancestral language.

This was acknowledged by Ringe et al. (2002) and Nakhleh et al. (2005b), whose phylogenetic analyses of Indo-European have been discussed in the previous section. In addition to lexical items, their data included both phonological and morphological characters, selected for their complexity and uniqueness which was seen to guarantee that their appearance in different languages was due to shared retention from a protolanguage and not a product of independent innovations. As their results seem to correspond most closely with historical, linguistic and archaeological data, this data format seems promising. However, in the same way as using cognate classes as character states, this approach requires much preceding work in reconstructing sound changes and morphemes and detecting probable contact influences.

Dunn et al. (2008), on the other hand, performed a phylogenetic analysis using only structural data, with no lexical items arranged in cognate sets and with no prior work in terms of traditional historical-comparative linguistics. Their objective was to attempt to compute a phylogenetic tree for the so-called Papuan languages of the Melanesian Archipelago, off the eastern coast of New Guinea, which share a very low amount of vocabulary and so have been difficult to classify genetically. In order to test the applicability of their approach for finding a phylogenetic signal, they perform the same analysis with the Austronesian languages of the Oceanic subfamily spoken in the area, whose classification is on more solid foundation already as a result of of the application of the traditional methodology of historical linguistics. Among the characters used by the authors, instead of meanings corresponding to individual form-function pairings in individual languages, typological grammatical features like 'Are there prenasalised stops?', 'Are more than two degrees of distance morphologically marked in demonstratives?' and 'Are there complement clauses?' (id.: 750-753). They note that it has been argued that typological features should not be used for establishing genetic classification (e.g. Croft 2004), one of the reasons being that in linguistic convergence areas anything in principle can be borrowed (Thomason and Kaufman 1988), but defend their approach by arguing that the use of over a hundred features regarding all areas of grammar minimises the possibility of contact influence masking any phylogenetic signal present in the languages (Dunn et al. 2008: 714-717). On the other hand, they use both distance-based and Bayesian phylogenetic methods for producing network diagrams from the data instead of tree-like models to visualise phases of contact influence among the languages (id.: 722-725).

The authors used both a maximum parsimony method and a Bayesian MCMC phylogenetic inference method for investigating the phylogeny of the languages, in addition to a NeighborNet method for producing distance-based networks. They note that the Bayesian method was superior in being able to detect a phylogenetic signal to maximum parsimony, supporting observations referred to previous sections. Specifically, they were able to approximate the tree produced for the Oceanic languages in the area with the comparative method, not with NeighborNet nor with maximum parsimony, but fairly accurately with Bayesian phylogenetic inference (id.: 732-734). The authors proceeded to produce a Bayesian consensus network for the Papuan languages. They note that the results agree to some extent with some preliminary linguistic classifications, but remind that the method forces all languages to attach to the network at some point, even if some languages were not even distantly related to the others (id.: 734-737). Their study is notable also in that the authors applied a method for discerning the population structure of the linguistic data. As a result, they find that the model that best explains the data involves two or three linguistic populations, with the populations more or less corresponding to Oceanic languages and Papuan languages in the first case, and the Meso-Melanesian linkage among the Oceanic languages, rest of the Oceanic languages and the Papuan languages in the second (id.: 745-747). As a consequence, the study shows that other kinds of computational methods used in biology, apart from phylogenetic ones, can be fruitfully used in historical linguistic research.

3.5. Summary

As has been seen, phylogenetic methods generally succeed in finding subgroups that are generally agreed upon in historical linguistics, and additionally suggest more unconventional groupings that in many cases agree with tentative views in linguistics as well as with the results of other phylogenetic analyses. For instance, in Indo-European linguistics the subgroups of Indo-Iranian and Balto-Slavic are generally accepted, but several phylogenetic analyses additionally group these two groups into another, even more inclusive group. Dating of the divergence of protolanguages, on the other hand, is still not something that can be relied upon in producing clear directions for answers. The analysis of Gray and Atkinson (2003; section 3.3.3) produced a phylogeny consistent with an expansion from the Eurasian steppes (see section 3.3.4), but resulted in dates agreeing with the hypothesis of earlier Anatolian origins. Along with new hypotheses, it is hoped that taking these methods along as assistance in historical linguistics will help provide new answers.

One question does not seem to have been considered in phylogenetic analyses of language families. Namely, what is it that diverges when there is a split in a phylogenetic tree generated by the algorithm? The assumption seems to be that at those points a language diverges into two languages. As discussed in section 2.3, it could be said that this entails the breaking of a dialect continuum into two separate continua, with the consequent 'speciation' of the lingueme population, to use terminology influenced by population genetics.

It is of interest to this consideration that Holm (2007) criticises the Indo-European phylogenetic analysis of Gray and Atkinson (2003; discussed in section 3.3.3) in part because the phylogenetic tree obtained in that analysis has English diverge from the rest of the West Germanic languages before the other languages of that group. Holm's criticism stems from the view that the closest relative of English is Frisian, which in Gray and Atkinson is located on a branch diverging after English. The 'Anglo-Frisian' group containing English and Frisian is determined mostly from phonological and syntactic developments. However, Davis (2006: 154) notes that “the languages of the Germanic group in the Old period are much closer than has previously been noted. Indeed it would not be inappropriate to regard them as dialects of one language.” This would include the hypothesised Anglo-Frisian group along with the rest of the West Germanic languages at least until the time that English became the language spoken in the British Isles. Accordingly, the Anglo-Frisian group along with its linguistic developments has probably been a dialect cluster in a West Germanic dialect continuum, until English diverged from it on the other side of the English Channel. This would mean that the method used by Gray and Atkinson (2003) managed to recover something not accomplished with earlier methods, instead of being defective.

Could the divergence of languages be indeed be recoverable using lexical data in a phylogenetic method? This could be so. According to Dixon (1997), when a language diverges into two while maintaining close geographical proximity, the extent of mutual intelligibility drops quite suddenly,[5] with the corresponding decline in shared vocabulary. If it can be demonstrated that this effect can be seen in basic vocabulary, it is probable that a phylogenetic analysis performed with a dataset of the Swadesh list type can detect these splits. However, this does not apply in the case of divergence through geographical separation. Here, the languages will drift apart gradually. Future analyses could consider whether the data used can detect divergence of languages. Some linguistic patterns may diverge already between the dialects of a single language, and some may diverge only after a long time since the break-up to separate language communities.

Phylogenetic methods cannot replace traditional methodology in historical linguistics. They cannot be used to demonstrate that a given set of languages is related. With any kind of data, phylogenetic methods will join the languages together, whether the data has information that can prove their relatedness or not. Accordingly, when a given set of languages has been chosen to be included in a phylogenetic analysis, the assumption has been made that they are related. Of the methods in existence, only the comparative method can demonstrate that a given set of languages have descended from a common ancestral language. The phylogenetic methods can only be used to investigate the history of divergence of languages in a language family already established by the comparative method. However, approaches like that of Dunn et al. (2008), presented in the previous section, are showing signs of being able to assist in explorations of languages with unknown descent. It remains to be seen to what extent they can be helpful in this respect.

[5] Dixon (1997) estimates that in most cases, it is found that closely related languages are around 10% mutually intelligible, whereas speakers of (contiguous) dialects of the same language can understand 70% or more of each other. In case of divergence in geographical proximity, this degree of intelligibility quickly drops from the latter number to the former.

Also, the phylogenetic methods that have been most often used in the study of language history require much prior traditional work. The assignment of lexical forms into cognate sets, as required by several approaches, requires that it has been demonstrated for every such form that they have descended from a corresponding form in the protolanguage through well understood developments, and that borrowing and other contact influence has been factored out. The same applies to reconstructed morphological constructions and rare phonological developments as employed by Ringe et al. (2002; see section 3.3.1).

In this respect, as well, the analysis of Dunn et al. (2008) is unique. This analysis involved only typological features obtained from contemporary descriptive grammars of the languages included. If their approach proves to be fruitful in more general application and not susceptible to contact influence, it would entail more expeditious analysis of the history of the World's languages than ever before. Competent descriptive grammars would become more central to this area of linguistic study, as well, and languages that have been until now relatively marginal in the time-consuming study of historical linguistics could wait longer for confirmation done with traditional methods while receiving preliminary results with phylogenetic methods.

4. Conclusion

As noted since the beginning of both evolutionary biology and historical linguistics, the mechanisms that drive biological evolution and those that affect languages over time are in many respects similar. In section 2, it was discussed that in many respects languages can be seen as populations formed of linguistic replicators, bringing about their effects through their socio-communicative functions. Either languages change through neutral evolution in these populations or through social selection, but the basic mechanisms are similar enough to those affecting populations of biological organisms that many methods used in population genetics and biological phylogenetics can be seen as fruitful to the study of language history.

In section 3, it was seen that methods of computational phylogenetic inference, used in biology since the 1960s, can be indeed used to investigate language history. Mostly this has been done with the aid of vocabulary lists, which require words to be assigned into cognates based on prior work in historical linguistics with the comparative method. Consequently, phylogenetic methods as used in these cases cannot supplant the traditional methods, but they can be of great help in determining the order of divergence of the subgroups of a language family, which in turn helps in determining patterns of language change affecting those descendant groups. For a well-researched language family like the Indo-European, it has been seen that the results can be very promising, and to be corroborated with expert views in historical linguistics and linguistic palaeontological and archaeological information. Also, there is promise that the phylogenetic methods can achieve preliminary internal classifications for groups of languages with no or little prior work done with the traditional methods.

4.1. Consequences for linguistic theory

Since the 1960s, variationist sociolinguistics has focussed on variation found in languages, its social causes and effects and its central position in the mechanisms of language change. However, this research has mostly been concentrated on a relatively small scale, and patterns that have an effect over different but related languages have been left to historical linguists working with traditional methodology. In evolutionary biology, it has long been recognised that small-scale variations lie behind all subsequent developments, and that in order to understand patterns in evolution over long stretches of time, one must understand the mechanisms that create small-scale variation in population and those that act on this variation to make this variation more significant. In the terminology of the generalised analysis of selection outlined in section 2.1.3, the mechanisms of replication and selection must be understood to explain patterns between different lineages. Similarly, to understand and to be able to investigate more fully the relationships between related languages, it could be fruitful to approach large-scale language change as coming about through variation internal to an ancestral language community in the first place. Research done in the tradition of variationist sociolinguistics is indispensable in this regard, but it should be preferably complemented with fuller understanding of the mechanisms of altered replication, i.e. the emergence of new variants, which is referred to as the actuation problem (see section 2.2.1) in that research tradition.

The study of language contact is important as a link between small-scale variation and large-scale change. The divergence of a language family into daughter branches is often characterised by contact between different languages of that family and also with other, unrelated languages, so that different branches of the family often show signs of different internal and external influences. On the other hand, language contact is often an important factor in variation internal to a language community. Accordingly, the study of language contact encounters and is informed by both small-scale variation and large-scale patterns of phylogeny. This makes it a potential area of fruitful insights into the relationship between variation and change.

Although contact between languages and varieties of the same language is an important factor in much change and many patterns in the phylogenies of major language families (Thomason and Kaufman 1988), at least some change starts out in the communities that it affects and is not diffused from another community. This means that the link between internal variation and long-term change apparent in divergence to different languages could also be formed in the study of internal change. This requires distinguishing the boundaries of languages in terms of making up linguistic populations as dialect continua, as discussed in section 2.3, in order to determine what variation is dialectal or sociolectal and part of the internal 'population structure' of a language, and what are characteristics of different languages and possible markers of shared ancestry with related languages.

Computational models for statistical inference of phylogenetic patterns, as explored in section 3.2, can be useful heuristic tools for historical linguistics. The comparative method is indispensable for demonstrating descent from a common protolanguage, but traditional methods do not offer much for gaining a consensus about the patterns of divergence internal to the language family. The statistical and mathematical basis of computational phylogenetics is developing all the time. In order to be of greatest possible use for historical linguistics, additional work can be done to determine more accurate models of language change to be used with the algorithms, to investigate the effects of contact influence on language divergence, and to develop the format of linguistic data as used in these methods. Determining the phylogenetic patterns of language families can bring about more detailed understanding about the mechanisms of change and of language contact if it becomes possible to determine which instances of shared forms and patterns are a result of shared ancestry and which have transmitted horizontally.

Also, understanding the historical developments of languages and the mechanisms that drive them over time can help in finding causes for observed universal properties of human languages. For instance, Givón (2002: 203-222) notes that typological patterns observed in grammatical constructions in the World's languages can only be fully understood by considering the grammaticalisation pathways that bring them about from other kinds of constructions. Hence, the study of synchronic variation in respect to grammatical constructions and the study of diachronic change are inseparable (id.: 217). Consequently, universal properties apparent in typological patterns can only be understood once the diachronic mechanisms are understood.

References

Alter, Stephen G. 1999. *Darwinism and the linguistic image: Language, race and natural theology in the nineteenth century.* Baltimore: The Johns Hopkins University Press.

Andersen, Henning 2006. Synchrony, diachrony, and evolution. In Ole Nedergaard Thomsen (ed.), *Competing models of linguistic change: Evolution and beyond* 59-90. Amsterdam: Benjamins.

Anthony, David W. 2007. *The horse, the wheel, and language: how Bronze age riders from the Eurasian steppes shaped the modern world*. Princeton, NJ: Princeton University Press.

Atkinson, Quentin, Geoff Nicholls, David Welch and Russell Gray 2005. From words to dates: Water into wine, mathemagic or phylogenetic inference? *Transactions of the Philological Society* 102(2): 193-219.

Atkinson, Quentin D. and Russell D. Gray 2006. How old is the Indo-European language family? Progress or more moths to the flame? In Peter Forster and Colin Renfrew (eds.), *Phylogenetic Methods and the Prehistory of Languages* 91-109. Cambridge: McDonald Institute for Archaeological Research.

Bakker, Peter and Pieter Muysken 1995. Mixed languages and language intertwining. In Jacques Arends, Pieter Muysken and Norval Smith (eds.), *Pidgins and creoles: An introduction* 41-52. Amsterdam: Benjamins.

Bartsch, Renate 1987. *Norms of language*. London: Longman.

Bergsland, Knut and Hans Vogt 1962. On the validity of glottochronology. *Current Anthropology* 3: 115-53.

Berlin, Brent and Paul Kay 1969. *Basic color terms: Their universality and evolution.* Berkeley: University of California Press.

Bopp, Franz 1833. *Vergleichende Grammatik des Sanskrit, Zend, Griechischen, Lateinischen, Gothischen und Deutschen*. Berlin.

Blackmore, Susan 1999. *The meme machine*. Oxford: Oxford University Press.

Bowler, Peter J. 1984. *Evolution: The history of an idea.* Berkeley: University of California Press.

Campbell, Lyle 2003. How to show languages are related: Methods for distant genetic relationship. In Brian D. Joseph and Richard D. Janda (eds.), *The handbook of historical linguistics* 262-82. Oxford: Blackwell.

___ 2004. *Historical linguistics: An introduction*. 2nd edition. Edinburgh: Edinburgh University Press.

Cavalli-Sforza, L. L. and A. W. F. Edwards 1967. Phylogenetic analysis: Models and estimation procedures. *American Journal of Human Genetics* 19: 233-257. *Evolution* 21: 550-570.

Chambers, J. K. and Peter Trudgill 1999. *Dialectology* (2nd edition). Cambridge: Cambridge University Press.

Coşeriu, Eugenio 1970. Sprache, Strukturen, Funktionen. Darin: "System, Norm und Rede" und "Synchronie, Diachronie, Typologie". *Tübinger Beiträge zur Linguistik*. Tübingen: G. Narr Verlag.

Croft, William 2000. *Explaining language change: An evolutionary approach*. Harlow: Longman.

___ 2003. The relevance of an evolutionary model to historical linguistics. In Ole Nedergaard Thomsen (ed.), *Competing models of linguistic change: evolution and beyond* 91-132. Current issues in linguistic theory. Amsterdam: John Benjamins.

___ 2004. Typological traits and genetic linguistics. Albuquerque: University of New Mexico, MS. Online: http://www.unm.edu/~wcroft/Papers/Typ-Gen.pdf (accessed October 12 2009).

___ 2006. The relevance of an evolutionary model to historical linguistics. In Ole Nedergaard Thomsen (ed.), *Competing models of linguistic change: Evolution and beyond* 91-132. Amsterdam: Benjamins.

Darwin, Charles 1859 [1964]. *On the origin of species by means of natural selection: or the preservation of favoured races in the struggle for life.* London, reprinted with an introduction by Ernst Mayr. Cambridge: Harvard University Press.

___ 1874. *The descent of man and selection in relation to sex.* 2nd edition. London: Murray. Online: http://www.munseys.com/diskone/darwindescent.pdf (accessed October 26 2009).

Davis, Graeme 2006. *Comparative syntax of Old English and Old Icelandic: Linguistic, literary and historical implications*. Bern: Peter Lang.

Dawkins, Richard 1976. *The selfish gene*. Oxford: Oxford University Press.

___ 1983. Universal Darwinism. In Derek S. Bendall (ed.), *Evolution from molecules to men* 403-28. Cambridge: Cambridge University Press.

Dixon, R. M. W. 1997. *The rise and fall of languages*. Cambridge: Cambridge University Press.

Dunn, Michael, Stephen C. Levinson, Eva Lindström, Ger Reesink and Angela Terrill 2008. Structural phylogeny in historical linguistics: Methodological explorations applied in Island Melanesia. *Language* 84(4): 710-59.

Dyen, Isidore, Joseph B. Kruskal and Paul Black 1992. An Indoeuropean classification: A lexicostatistical experiment. *Transactions of the American Philosophical Society* 82(5): 1-132.

Edwards, A. W. F. and L. L. Cavalli-Sforza 1964. Reconstruction of evolutionary trees. In V. H. Heywood and J. McNeill (eds.), *Phenetic and phylogenetic classification*. Systematics Association Publ. No. 6, London.

Embleton, Sheila M. 1986. *Statistics in historical linguistics*. Quantitative linguistics 30. Bochum: Studienverlag Dr. N. Brockmeyer.

Eldredge, Niles and Stephen Jay Gould 1972. Punctuated equilibria: An alternative to phyletic gradualism. In T. J. M. Schopf (ed.), *Models in Paleobiology* 82-115. San Francisco: Freeman Cooper.

Felsenstein, Joseph 2004. *Inferring phylogenies*. Sunderland: Sinauer Associates, Inc.

Fortson, Benjamin W. 2004. *Indo-European language and culture*. Padstow: Blackwell Publishing.

Gell-Mann, Murray 1992. Complexity and complex adaptive systems. In John A. Hawkins and Murray Gell-Mann (eds.), *The evolution of human languages* 3-18. New York: Addison-Wesley.

Givón, Talmy 2002. *Bio-linguistics. The Santa Barbara lectures*. Amsterdam: Benjamins.

Gray, Russell D. and Quentin D. Atkinson 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* 426: 435-9.

Greenhill, Simon J., Robert Blust and Russell D. Gray 2008. The Austronesian Basic Vocabulary Database: From bioinformatics to lexomics. *Evolutionary Bioinformatics* 4: 271-83.

Hanski, Ilkka 1996. Metapopulation ecology. In Olin E. Rhodes, Jr., Ronald K. Chesser, Michael H. Smith (eds.), *Population dynamics in ecological space and time* 13-43. Chicago: University of Chicago Press.

Haspelmath, Martin 1999. Optimality and diachronic adaptation. *Zeitschrift für Sprachwissenschaft* 18: 180-205.

Haugen, Einar 1972. *The ecology of language: Essays by Einar Haugen*. Stanford: Stanford University Press.

Henning, Willi 1965. Phylogenetic systematics. *Annual Review of Entomology* 10: 97-116.

Hjelmslev, Louis 1942. Langue et Parole. Cahiers Ferdinand de Saussure, 3. German translation: Langue und Parole. In Hjelmslev, Louis (1974) *Aufsätze zur Sprachwissenschaft* 44-55. Stuttgart: Klett Verlag.

Holden, Clare J. 2002. Bantu language trees reflect the spread of farming across sub-Saharan Africa: A maximum-parsimony analysis. *Proc. R. Soc. Lond.* B 269: 793-9.

Holm, Hans J. 2007. The new arboretum of Indo-European "trees". Can new algorithms reveal the phylogeny and even prehistory of Indo-European? *Journal of Quantitative Linguistics* 14(2): 167-214.

Hull, David 1988. *Science as a process: An evolutionary account of the social and conceptual development of science*. Chicago: University of Chicago Press.

Hurford, James R., Michael Studdert-Kennedy and Chris Knight (eds.) 1998. *Approaches to the evolution of language: Social and cognitive bases*. Cambridge: Cambridge University Press.

Keller, Rudi 1994. *On language change: The invisible hand in language*. London: Routledge.

Kitchen, Andrew, Christopher Ehret, Shiferaw Assefa and Connie J. Mulligan 2009. Bayesian phylogenetic analysis of Semitic languages identifies an Early Bronze Age origin of Semitic in the Near East. *Proc. R. Soc.* B 276: 2703-10.

Koerner, Konrad (ed.) 1983. *Linguistics and evolutionary theory. Three essays by August Schleicher, Ernst Haeckel, and Wilhelm Bleek.* Amsterdam: John Benjamins Publishing Company.

Kutschera, Ulrich and Karl J. Niklas 2004. The modern theory of biological evolution: An expanded synthesis. *Naturwissenschaften* 91: 255-76.

Labov, William (ed.) 1980. *Locating language in time and space*. New York: Academic Press.

Lass, Roger 1997. *Historical linguistics and language change*. Cambridge studies in linguistics. Cambridge: Cambridge University Press.

Laycock, Donald C. 1982. Melanesian linguistic diversity: A Melanesian choice? In Ron J. May and Hank Nelson (eds.), *Melanesia: Beyond diversity* 33-38. Canberra: Research School of Pacific Studies.

Mallory, J. P. 1989. *In search of the Indo-Europeans*. London: Thames and Hudson.

Mallory, J. P. and Douglas Q. Adams 1997. *Encyclopedia of Indo-European culture*. London: Fitzroy Dearborn.

Mallory, J. P. and Douglas Q. Adams 2006. *The Oxford introduction to Proto-Indo-European and the Proto-Indo-European world.* Oxford: Oxford University Press.

Mayr, Ernst 1982. *The growth of biological thought: Diversity, evolution and inheritance*. Cambridge, Ma: Belknap Press.

McMahon, April and Robert McMahon 2003. Finding families: Quantitative methods in language classification. *Transactions of the Philological Society* 101(1): 7-55.

Michener, Charles D. and Robert R. Sokal 1957. A quantitative approach to a problem in classification. *Evolution* 11: 130-62.

Milroy, James 1992. *Linguistic variation and change*. Oxford: Blackwell.

Milroy, James and Lesley Milroy 1985. Linguistic change, social network and speaker innovation. *Journal of Linguistics* 21: 339-84.

Mufwene, Salikoko S. 2001. *The ecology of language evolution*. Cambridge: Cambridge University Press.

Mühlhäusler, Peter 1996. *Linguistic ecology: Language change and linguistic imperialism in the Pacific region*. London: Routledge.

Nakhleh, Luay, Tandy Warnow, Don Ringe and Steven N. Evans 2005a. A comparison of phylogenetic reconstruction methods on an Indo-European dataset. *Transactions of the Philological Society* 103(2): 171–92.

Nakhleh, Luay, Don Ringe and Tandy Warnow 2005b. Perfect phylogenetic networks: A new methodology for reconstructing the evolutionary history of natural languages. *Language* 81(2): 382-420.

Pagel, Mark 2009. Human language as a culturally transmitted replicator. *Nature Reviews Genetics* 10: 405-15.

Pagel, Mark, Quentin D. Atkinson and Andrew Meade 2007. Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature* 449: 717-21.

Rankin, Robert L. 2003. The comparative method. In Brian D. Joseph and Richard D. Janda (eds.), *The handbook of historical linguistics* 183-212. Oxford: Blackwell Publishing.

Reali, Florencia and Thomas L. Griffiths 2009. Words as alleles: connecting language evolution with Bayesian learners to models of genetic drift. *Proc. R. Soc.* B. Online:

http://rspb.royalsocietypublishing.org/content/early/2009/10/06/rspb.2009.1513 (accessed October 8 2009).

Renfrew, Colin 2001. The Anatolian origins of Proto-Indo-European and the autochthony of the Hittites. In Robert Drews (ed.), *Greater Anatolia and the Indo-Hittite language family* 36-63. Journal of Indo-European Studies Monograph 38. Washington, D.C.: Institute for the Study of Man.

___ 2002. The emerging synthesis: The archaeogenetics of farming/language dispersals and other spread zones. In Peter Bellwood and Colin Renfrew (eds.), *Examining the farming/language dispersal hypothesis* 3-16. Cambridge: McDonald Institute for Archaeological Research.

Rexová, Kateřina, Daniel Frynta and Jan Zrzavý 2003. Cladistic analysis of languages: Indo-European classification based on lexicostatistical data. *Cladistics* 19: 120-7.

Ringe, Don 2003. Internal reconstruction. In Brian D. Joseph and Richard D. Janda (eds.), *The handbook of historical linguistics* 244-61. Oxford: Blackwell.

Ringe, Don, Tandy Warnow and Ann Taylor 2002. Indo-European and computational cladistics. *Transactions of the Philological Society* 100(1): 59-129.

Ritt, Nikolaus 2004. *Selfish sounds and linguistic evolution: A Darwinian approach to language change*. Cambridge: Cambridge University Press.

Saussure, Ferdinand de 1959. *Course in general linguistics* (transl. by W. Baskin). New York: Philosophical Library.

Schleicher, August 1863. The Darwinian theory and the science of language. (Transl. by Alexander V. W. Bikkers.) In Konrad Koerner (ed.; 1983), *Linguistics and evolutionary theory. Three essays by August Schleicher, Ernst Haeckel, and Wilhelm Bleek* 1-73. Amsterdam: John Benjamins Publishing Company.

Swadesh, Morris 1952. Lexico-statistic dating of prehistoric ethnic contacts. *American Philosophical Society, Proceedings* 96: 453-63.

___ 1955. Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics* 21: 121-37.

Steiner, George 1975. *After Babel*. Oxford: Oxford University Press.

Thomason, Sarah G. and Terrence Kaufman 1988. *Language contact, creolization and genetic linguistics.* Berkeley: University of California Press.

Weinreich, Uriel, William Labov and Marvin I. Herzog 1968. Empirical foundations for a theory of language change. Winfrid P. Lehmann and Yakov Malkiel (eds.), *Directions for historical linguistics* 95-195. Austin: University of Texas Press.

Wierzbicka, Anna 1996. *Semantics: Primes and universals*. New York: Oxford University Press.