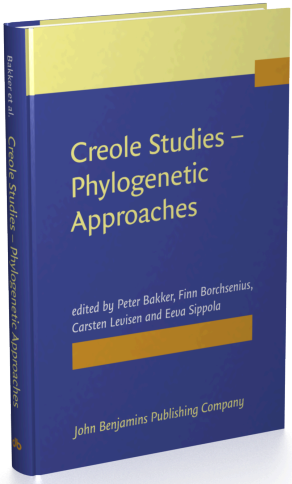
CHAPTER 3

Phylogenetics in biology and linguistics

Finn Borchsenius | Aarhus University
Aymeric Daval-Markussen | Aarhus University
Peter Bakker | Aarhus University

https://doi.org/10.1075/z.211.03bor



Pages 35–58 of
Creole Studies – Phylogenetic Approaches
Edited by Peter Bakker, Finn Borchsenius, Carsten Levisen and Eeva M. Sippola
2017. x, 414 pp.

CHAPTER 3

Phylogenetics in biology and linguistics

Finn Borchsenius, Aymeric Daval-Markussen and Peter Bakker
Aarhus University

The main goal of this chapter is to introduce the reader to the parallels and commonalities that exist between the fields of biology and linguistics. Researchers from both fields faced similar problems when seeking to account for the descent and diversification of related entities (species, languages). Therefore, they often sought mutual inspiration and opted for similar solutions. This has resulted in a convergence of models and methods in both fields. This chapter is divided into two parts. Firstly, we review some of the methodological and conceptual developments that have occurred in biology since the emergence of the field of evolutionary biology. There will be an emphasis on the last decade, where a variety of computer-based analyses have been developed. To illustrate the benefits of these tools, phylogenetic methods are applied in the second part of the chapter to a group of high-contact languages (creoles), which have long defied attempts at classification due to their multiple ancestry.

3.1 Origin of phylogenetics in biology and linguistics

Phylogenesis can be defined as the evolutionary development and diversification of a group of organisms. Phylogenetics is the study of that process, aiming at drawing up phylogenies, usually in the form of trees showing the evolutionary relationships between the members of a group of living organisms, languages, or other entities subject to evolution through descent with modification.

Evolution as a theory explaining the diversity of living nature dates back to biologists such as Charles Darwin and Alfred Russel Wallace in the middle of the 19th century. Though Darwin founded evolutionary theory, he was, however, not particularly concerned with phylogeny; rather, he was interested in the process of change leading to the emergence of new species. In fact, just a single sketch of a phylogenetic tree appears in Darwin's ground-breaking book *On the Origin of Species by Natural Selection* (1859: 116–117). Nevertheless, phylogenetic thinking was the logical outcome of Darwin's evolutionary theory.

DOI 10.1075/z.211.03bor

It is noteworthy to realize that linguists adopted phylogenetic thinking at a very early stage. In fact, some evolutionary concepts were used in linguistics before they were used in biology. By the end of the 18th century, linguists already used the concept of homology (i.e. similarities in form between words may point to a common origin) and had realized that diversity could be explained via descent with modification (e.g. Schleicher 1853). Inspired by Darwin's publications, they quickly moved to making trees representing linguistic descent. In 1863, the German comparative linguist Schleicher published a paper depicting an Indo-European language tree (Schleicher 1863). Interestingly, he wrote the paper as an open letter to his friend, the evolutionary biologist Ernst Haeckel, who had introduced him to *the Origin of Species* earlier the same year. The concept of language trees depicting evolutionary processes and the diversification of languages from a common ancestral form, however, predates Darwinian thinking. The idea of a genealogical language tree has been accredited to the early work of Friedrich Schlegel, who introduced a language tree, or Stammbaum (family tree) approach in a publication as early as 1808 (Schlegel 1808). The first published manuscript depicting a phylogeny was published by Carl Johan Schlyter in 1827 (Collin & Schlyter 1827). Schleicher had also used language trees with branches six years before the publication of Darwin's book (Schleicher 1853).

In their review of the early history of cross-fertilization and borrowing of ideas in the establishment of phylogenetic thinking in linguistics and biology, Atkinson and Gray (2005: 517) conclude that "Darwinian ideas of descent with modification were probably less revolutionary in linguistics than they were in biology. Phylogenetic thinking and methodology in linguistics had already developed rapidly before Darwin, and this continued throughout the 19th century". Particularly noteworthy is the fact that linguists in the late part of the 19th century recognized and understood the importance of distinguishing between innovations (i.e. new characters not present in the ancestral form) and retentions (characters inherited from a common ancestor). Biologists today make a similar distinction, but they call them synapomorphies (shared innovations) and symplesiomorphies (shared retentions). In biology, this distinction was widely accepted until 70 years later (Hennig 1950), when it led to a revolution of systematic biology, setting the principles for algorithmic reconstruction of phylogenetic trees.

Following Hennig's (1950) publication of *Grundzüge einer Theorie der Phylogenetischen Systematik*, the development of methods for reconstructing trees based on the principle of shared derived character states gained much interest. With access to computers becoming easier and more widespread, particularly from the 1980s, and due to the rapid increase of processor speeds, biologists soon engaged in developing computer-based algorithms for reconstructing phylogenetic trees (see Felsenstein 2004 for an extensive overview).

The fundamental principle for tree construction was parsimony, i.e. searching for the shortest possible tree with the minimum number of character changes across the entire tree that could also explain the variation in character states observed in the set of entities being analysed. By the 1980s, this type of phylogenetic analysis had become widespread and generally accepted as the basis of systematic biology. It was also proliferating into the field of biogeography, i.e. the study of the geographical distribution of living organisms across the Earth as a way of reconstructing the ancestral distribution areas of groups of organisms and the series of geological events that have determined their current distribution (e.g. Wiley 1988). With improved computer capacity, alternative principles to parsimony also began to gain terrain, most notably maximum likelihood, a principle derived from statistical theory, where one searches for the model that gives the highest probability of observing the actual data. Maximum likelihood estimation of phylogenetic trees is computationally more demanding, but it offers the possibility of implementing a more sophisticated, and hence, a more realistic model for evolution.

Access to ever-increasing quantities of DNA sequence data (see e.g. Marx 2013), in conjunction with faster computers and improved algorithms, has resulted in a move away from parsimony towards likelihood-based methods in biology. At the same time, there are fewer and fewer phylogenetic analyses based on the morphological characteristics of the studied organisms (e.g. whether a species has developed limbs or not).

While biologists from the very beginning of the silicon age incorporated computer-based methods in phylogenetic analysis, linguists have been more reluctant to do so. Early attempts to introduce numerical approaches to linguistic analysis based on distance matrices (e.g. lexicostatistics by Swadesh 1952, 1955) were heavily criticized (e.g. Bergsland & Vogt 1962) and are now largely discredited by most linguists – but perhaps not always for the right reasons (see for instance the papers in Wichmann & Grant 2012 for a rehabilitation). Atkinson and Gray (2005) discuss some of the problems and criticisms of distance-based methods in linguistics. This discussion can be seen as a parallel to the massive critique that was initially raised against numerical taxonomic methods and distance-based tree estimation methods in evolutionary biology. A fundamental problem with these methods in such evolutionary analyses is that symplesiomorphies (retentions) that do not carry any true signal of common descent are as important as synapomorphies (innovations). They are therefore likely to group unrelated old lineages together, although similarities between them are simply a result of the absence of derived features present in more recently-formed groups.

In linguistics, one of the points of critique has been that the conversion of lexical character data such as Swadesh lists to distance scores results in loss of information, thus reducing the capacity of the method to reconstruct evolutionary

history accurately (Steel et al. 1988). In contrast, tree-building methods based on parsimony or maximum likelihood incorporate character changes as an implicit part of the model. This allows the underlying assumptions concerning the precise evolution of characters associated with the hypothesized tree to be traced and evaluated. However, over the past ten years or so, the transfer of computer-based methods from evolutionary biology to linguistic studies has accelerated and is now diversifying rapidly. This has been termed the "new synthesis" of biology and linguistics (McMahon & McMahon 2003: 18–21) and incorporates a variety of methods and subfields in both evolutionary biology and linguistics. In what follows, we will provide an overview of the recent transfer of methods and discuss current trends in computational biology that could have future applications in linguistics.

3.2 Phylogenetic studies in linguistics

The initial wave of papers in the new synthesis of linguistics and biology focused on understanding the history of human populations using linguistic data and phylogenetic methods, both at the local level (e.g. Forster et al. 1998 for dialects of Alpine Romance languages) and for widespread language families such as Austronesian (e.g. Gray & Jordan 2000). Remarkably, some of these early linguistic studies were initiated by non-linguists (anthropologists, psychologists, geneticists). Initially, most early studies employed maximum parsimony for the estimation of phylogenetic trees (e.g. Gray & Jordan 2000; Holden 2002; Rexová et al. 2003; Dunn et al. 2005; 2008).

Similar to what is happening in evolutionary biology, we are seeing a strong tendency of moving towards likelihood-based methods for reconstructing phylogenies in linguistics. In particular, Bayesian analyses based on the Markov Chain Monte Carlo algorithms (Huelsenbeck & Ronquist 2001) have become increasingly popular. These methods use an iterative approach to sample a large number of almost equally likely models that can be summarized statistically afterwards. These have become a standard tool in many areas of computational biology. Papers employing such methods include, among others, Gray & Atkinson (2003), Dunn et al. (2008), Dunn (2009), Gray et al. (2009), Fortunato & Jordan (2010), Dediu (2011), Bowern & Atkinson (2012), Nurbakova et al. (2013) and Maurits & Griffiths (2014). Interestingly, however, a recent simulation study reported that maximum parsimony did a better job of finding correct trees than both maximum likelihood and distance-based methods (Barbançon et al. 2013).

3.3 Dated language phylogenies

Another significant development in biology has been the application of a "relaxed clock". Over the last decades, evolutionary biologists have increasingly applied molecular clock approaches to date splits in phylogenies and to make inferences about the timing of evolutionary events, such as the emergence of major evolutionary lineages relative to, e.g. geological history (Rutschmann 2006). The underlying principle is that mutations happen stochastically at a certain mean rate ("strict clock", the biological parallel to the "glottoclock" in linguistics). In this case, the number of mutations separating two evolutionary lineages would be directly proportional to the time passed since they diverged from their common ancestor. Biologists, however, soon realized that this assumption rarely holds true. Mutation rates vary greatly among different groups of organisms and across evolutionary time. This led to various methods of "rate smoothing", where substitution rates are allowed to vary across the branches of the phylogenetic tree (the "relaxed clock", Drummond et al. 2006).

Dating of phylogenies has been applied to linguistic data, which assumes that cognates, like bases in a DNA sequence, evolve at a relatively constant frequency over time. Morris Swadesh was a pioneer in this area (e.g. 1952; 1972). A problem with all dating methods is, however, how to estimate the overall mean substitution rate and hence to time-calibrate the clock. In addition, social and historical differences such as relative isolation, community size (Nettle 1999) and social upheaval (Thomason & Kaufman 1988; Dixon 1997) may influence rates of change. Biologists will typically use the age of fossils of ancestral species to calibrate their clock or, on shallow timescales, an average mutation rate that has been determined experimentally. In parallel to this, Gray and Atkinson (2003) put a time-constraint on the points (nodes) of initial divergence in all of the major languages based on known historical information on earlier stages of the languages. The results indicated that the expansions of the early Indo-Europeans were linked to the development of agriculture approximately 8000 years ago, a suggestion made by Colin Renfrew a decade and a half earlier (Renfrew 1987).

A similar approach was used by Gray et al. (2009) to test the pulse-and-pause hypothesis accounting for the expansion and dispersal of Austronesian settlers from Taiwan throughout Insular Southeast Asia and the Pacific. In this study, they calibrated the tree with archaeological date estimates and known settlement times. Their results are consistent with the idea of migration pulses followed by pauses obtained through archaeological and DNA studies.

Recently, a number of investigations have used a Bayesian framework for dating linguistic trees (BEAST; Drummond et al. 2012) to test hypotheses concerning the origin and spread of language groups and potential correlates in terms of

cultural events (Lee & Hasegawa 2011) or climatic changes (Honkola et al. 2013). Altschuler et al. (2013) used a linguistic clock model to date the writing of the Homeric epics. Phylogenetic dating has also played a role in the nascent field of extraction of stable typological features for the purpose of studying deep-ancestry relationships (Dunn et al. 2005; Greenhill et al. 2010). Stable features are those that can be shown to change only very slowly over time and they are assumed to be less prone to borrowing. For example, Pagel et al. (2013) used a combination of statistical modelling and a dated phylogenetic tree to infer the relationship between common usage of words and stability over a 15,000-year timeframe.

The relative stability of structural properties of languages has long been a controversial issue and, apart from a few serious attempts at making generalizations on the basis of empirical data encompassing a large number of languages from various families (e.g. Nichols 1992), no conclusive evidence demonstrating the evolutionary variability of structural features of language has been put forward. With the advent of quantitative investigations taking large amounts of linguistic data into account, the issue has received renewed interest over the last decade. The results of several recent studies based on the features described in WALS strongly suggest that some grammatical features evolve more rapidly than others (e.g. Parkvall 2008; Wichmann & Kamholz 2008; Wichmann & Holman 2009; Dediu 2011; Ellison 2014), although the exact ranking of the various features identified by these authors does not match entirely (see Dediu & Cysouw 2013 for a review).

In order to pinpoint those features that are thought to evolve at a different pace, the following logic was adopted. In cases where the various values attested for an individual WALS feature are relatively equally distributed across members of a language family (i.e. only a few differences can be observed across a single family), the low level of variation in the descent from ancestor to present-day languages suggests a relative stability for that feature. In other cases, however, the distribution of the various feature values is strongly uneven, which indicates that the rates at which the possibilities for that feature have evolved through time are different. The more variation we observe, the more likely this variation has its roots in independent innovations or contact effects, thus suggesting a considerable instability for that feature (see e.g. Wichmann & Holman 2009: 12).

Phylogenetic dating has been somewhat controversial in biology, in part because of the uncertainty linked to the use of fossils as calibration points. A fossil can, in principle, only provide a minimum age for a given lineage because the lineage in theory could be much older than the fossil record. Another problem using fossils is how to align a fossil of an extinct species to modern taxa, resulting in uncertainty as to where in the phylogeny the fossil, and hence, the calibration point should be placed. Finally, criticism has arisen from the frequent disagreement between the diversification time estimates based on respectively molecular

dating and fossil records, the so-called "rock-clock" problem. Nevertheless, dated phylogenies have become a standard tool in analyses of biogeographic patterns and community ecology at a range of spatial scales. The use of dated phylogenies has also raised discussion in linguistics (McMahon & McMahon 2006; Heggarty 2006), but the concept has nevertheless gained considerable currency.

3.4 Is linguistic evolution tree-like?

In linguistics, the idea that language diversification rarely proceeds through clean speciation events has long been acknowledged (i.e. lateral transfer between languages is known to occur pervasively at all levels of linguistic description in the process of descent with modification). Most notoriously, a preliminary step in historical linguistic analyses following the Comparative Method[1] explicitly stipulates that known borrowings should be removed (e.g. Campbell 1998: 128). Therefore, the whole enterprise of reconstructing language families and depicting their interrelationships completely ignores an important aspect of language evolution: horizontal influences. Reticulate (i.e. non-tree-like) evolution in language can be attributed to contact-induced innovations. However, only a few alternative models to represent these non-tree-like developments have been proposed and today, linguists still widely favor the tree model to represent the evolutionary histories of languages, following the Stammbaum model (Schleicher 1863). Schmidt's (1871) 'wave theory' stands out as a notable exception, because in that model, linguistic evolution is represented with concentric circles, thus reflecting the overlapping spread of linguistic features as a gradual process involving both vertical and lateral influences (illustrated in Figure 3.1).

Therefore, the validity of the traditional Stammbaum model used in historical linguistics has been questioned by sociolinguists (e.g. Labov 2007), and especially so by creolists (e.g. Mühlhäusler 1980: 34; Hancock 1987: 265–266; Thomason & Kaufman 1988; McWhorter & Parkvall 2002: 179–180). As Nichols & Warnow (2008: 762) point out, "[t]rees are often reasonable models of evolution, but sometimes a network model is more appropriate. For example, when creolization or language mixture occurs the correct graphical model will contain additional edges between branches in order to indicate the dual parentage".

1. The standard tool for historical linguistic analysis, the Comparative Method, is a technique that enables the researcher to describe and reconstruct the evolution and development of sibling languages from a proto-ancestor. For a detailed account, the reader is referred to e.g. Campbell (1998) or any other introductory textbook.

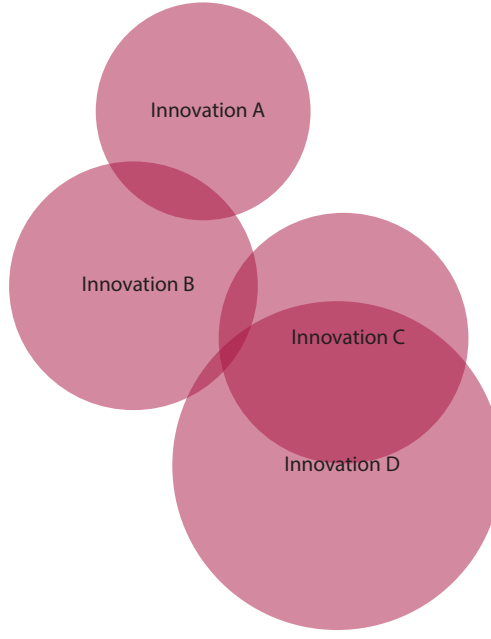


Figure 3.1 The wave model of language change and propagation according to Schmidt (1871).

The question of whether tree-like structures are an appropriate model of language evolution continues to be debated (e.g. Geisler & List 2013) and some authors argue that language evolution should essentially be regarded as reticulate rather than tree-like because of extensive borrowing between languages (e.g. List et al. 2014a; 2014b). Nevertheless, theoretical studies based on simulated data have shown that tree-like representations are quite robust in relation to lexical borrowing (Greenhill et al. 2009). Likewise, an empirical study found no evidence for the common claim that borrowing would be particularly high in prehistoric languages, such as those spoken in hunter-gatherer communities, which would compromise hierarchical relationships in the deep nodes of linguistic trees (Bowern et al. 2011).

The potential problems in applying tree-like evolution to linguistic data has led to the widespread use of biological computational methods designed to visualize conflicting signals in the data such as split networks or reticulate networks (Bryant & Moulton 2004; Huson & Bryant 2006). In particular, split networks have been applied to the study of creole languages (Bakker et al. 2011; Daval-Markussen 2011; 2013). The issue has also been subject to theoretical work (Nakhleh et al. 2005; Erdem & Ringe 2006; Kanj et al. 2008; Towner et al. 2012). However, it would perhaps seem that the choice is not between trees or other types of networks, but rather that the methods may complement each other in exploring the structure

and phylogenetic signal in the data (Gray et al. 2010). This viewpoint is in line with Huson and Bryant (2006), who suggest using split networks and associated tests for tree-like evolution as an initial data exploration procedure that can be used to determine the best path for further action, e.g. the construction of phylogenetic trees or perhaps reticulate networks.

3.5 Other lateral influences between biology and linguistics

Other recent applications of biological methods to linguistic analysis include phylogeographic and population genetics models. Phylogeography is a recent field of biology currently in explosive development (Avise 2000). It aims to study the movement of genes and populations across geographic scales. It thus deals with individuals and populations assumed to actively exchange genes through mating or migration. Nevertheless, it relies extensively on the use of phylogenetic methods including trees and other types of networks. In linguistics, phylogeographic methods have been applied to problems such as the origin of Native Americans (Sicoli & Holton 2014), the expansion of the Arawak language family in lowland South America (Walker & Ribeiro 2011) and the location of the Indo-European homeland (Bouckaert et al. 2012).

Methods from population biology originally developed to discover limits between distinct populations on the basis of recombining genetic markers have been applied to linguistic problems, such as explaining the language diversity of the Sahul region (between Australia and Papua, Reesink et al. 2009). Likewise, methods for studying population bottlenecks, such as those associated with founder effects when small groups of individuals colonize a new isolated area, have been applied. These have been used to test whether the decay in phoneme inventories in human languages following migration from Africa across Asia was the product of a serial founder effect, such as has been proposed for human genetic diversity (Atkinson 2011; but see also the critical reactions in *Linguistic Typology* 15 issue 2, and elsewhere).

Another interesting recent approach to the study of language evolution involves direct modelling (see Gavin et al. 2013). Although not strictly a phylogenetic method, this approach offers an interesting complement for testing some of the assumptions underlying language evolution and phylogenesis. Jansson et al. (2015) applied a mathematical modelling approach to test whether Mauritian creole may have been created only from a mutual desire to communicate (see also Parkvall et al. 2013). Their results strongly support this assumption and further indicate that it may be possible for a creole to develop quickly after first contact. The issue of modelling language evolution is increasingly gaining interest, albeit

controversially. The subject was recently reviewed by Gong et al. (2014a), provoking numerous comments and discussion (Gong et al. 2014b).

Lateral influences between the fields of biology and linguistics are not unidirectional. Recent years have witnessed a steady increase in the amount of biological studies incorporating linguistic data. These include a large number of studies mapping linguistic or other cultural traits on molecular phylogenies in order to study correlations between genetic and linguistic groupings, or to test directly if language barriers have resulted in genetic diversification in human populations. Two studies specifically addressing this question, however, both failed to establish language as a direct driver of human population divergence (Donohue et al. 2012; Zhang et al. 2014). In the latter case, language diversification correlated instead with geographic distance. Other examples of biological studies using linguistic data include three phylogeographic studies of maize (Mir et al. 2013), chili (Kraft et al. 2014) and banana (Perrier 2011), where cognate words related to these plants were used together with molecular and archaeological data to establish the area of origin and subsequent spread of domestication. The general value of linguistic data for solving problems in other fields of science was discussed by Pagel (2009), who concluded that for many comparative questions of anthropology and human behavioral ecology, historical processes estimated from linguistic phylogenies may be more relevant than those estimated from genes.

While computational phylogenetic analysis and related methods of evolutionary biology are by now well-established tools in linguistic studies, other recent developments of potential importance for understanding the diversity of human language, in particular in ecological science are just beginning to be applied (Gavin et al. 2013). Ecologists are increasingly studying biological diversity across spatial scales using a conjunction of data including species distributions, traits and phylogenetic position, a field known as macroecology or ecoinformatics. Methods from these fields were applied by Amano et al. (2014) to analyse the geographical patterns and drivers of extinction risk in languages worldwide. The authors found that both small geographical extension of the language and small speaker population size were associated with rapid declines in speaker numbers, causing 25% of existing languages to be threatened with extinction. In another article, Turvey and Pettorelli (2014) used methods from macroecology to assess the geographical congruence between the occurrence of languages and mammals threatened with extinction, suggesting that similar drivers may be affecting both entities. Lee and Hasegawa (2011) used methods for studying biological diversity patterns to demonstrate that geographical proximity and, more importantly, isolation by a surrounding ocean independently explain a significant proportion of lexical variation across Japonic languages.

3.6 Creoles, stable features and their substrates and lexifiers

The effects of language contact can be paralleled to lateral gene transfer in biological evolution, a phenomenon whereby one species receives genetic material from another species that is prevalent in bacteria and related organisms lacking a cell nucleus (Enfield e.g. 2003 even pushed the analogy so far as to refer to language contact as 'linguistic epidemiology'). Thus, the effect of borrowing might compromise genealogical language trees. In extreme cases, the degree of contact is such that whole linguistic systems are disrupted following a catastrophic event (also referred to as a 'break in transmission', see Thomason & Kaufman 1988). The creolization process is unique in the following senses: On the conceptual level, it is unlike what can be observed in normal evolutionary processes, namely that as in the case of creoles, several parent languages spawn a single daughter language. In addition, linguistically, it results in languages with a number of structural similarities. These are the result of the analogous circumstances under which they emerged, and this then implies that creoles share a typological profile. The question of the genetic affiliation of creoles and other high-contact varieties and their typological status are still very much debated (e.g. Thomason & Kaufman 1988; Owens 1991; Winford 2005).

The recent development of quantitative typological databases such as WALS (Dryer & Haspelmath 2013) has facilitated large-scale grammatical comparisons of languages from a variety of families around the world. This has helped address a number of questions which have been central concerns for linguists for a long time and has led to novel ways of tackling these problems. One issue that is particularly relevant in the context of creole studies concerns the existence of stable morphosyntactic features.

The process of creolization entails a strong reduction of the lexicon and the morphosyntactic systems of the languages present in the contact situation (as a consequence of going through a pidgin phase). The implication that linguistic features evolve at differential rates in the context of creole studies is that the features which are temporally most stable, and hence more resistant to borrowing and change, are precisely those that will be expected to be retained in creoles. Hall (1959) already demonstrated that lexical-phonological changes observed in Tok Pisin occurred much faster than in non-creole settings.

Creolists have proposed several theories in order to explain the presence of shared structural properties in creole languages, such as a relative absence of inflections or systems virtually devoid of gender agreement. Some argue that the lexifier language provided both the bulk of the lexicon and a good portion of the grammatical structures observed in creoles. According to this view, creole

languages are considered to be continuations of the lexifier and should therefore be treated as dialects of those. Hence, according to Mufwene (e.g. 2007), Germanic and Romance-based creoles worldwide should simply be regarded as the latest members of the Indo-European family. Others, on the other hand, emphasize the role of the substrate languages in creologenetic settings (e.g. Lefebvre 1998). The proponents of this view claim that creole structures can best be explained in light of the grammars of West African languages. A third view argues that similarities in creoles reflect universal cognitive mechanisms underlying heavy linguistic restructuring processes. Most creolists take an intermediate position.

These questions have been central to creole studies since their inception and we will suggest potential answers by using phylogenetic networks to represent the degree of similarity between the languages included in our samples. We will discuss several analyses comparing various languages involved in the formation of a sample of Atlantic and Indian Ocean creoles on the basis of a selection of stable features from WALS identified by Wichmann and Holman (2009). The samples that were used consist of creole languages, to which additional data on the lexifiers and substrates were encoded in order to shed light on the degree of similarity between the languages involved in the creation of creoles. If creoles are more closely related to their lexifier, then we should observe that the creole languages cluster according to their respective lexifier. This would suggest that the lexifiers have played a particularly important role in the creation of creoles. Conversely, if the creoles cluster closely with the substrates included in our sample, this would indicate that the original languages of the subjugated populations provided an important part of the grammatical structure of creoles. And if the creoles group together apart from both lexifiers and substrates, this would indicate that creoles are a relatively homogeneous group of languages from a typological perspective and have more commonalities between them than between any other group of languages considered – which would also emphasize the importance of the creolization process as responsible for the similarities between creoles, and not the types of languages that happened to be involved in the process.

A sample consisting of 26 creoles and their six lexifiers produced the network in Figure 3.2 following the Neighbor-Net algorithm (Saitou & Nei 1987) implemented in SplitsTree v. 4.1.13 (Huson & Bryant 2006). The lexifiers all appear in a single group on the left of the graph, with the Germanic languages in the upper part of the cluster and the Romance languages in the middle. The creoles are scattered across the rest of the graph, with no apparent patterns according to geography, lexifier or substrate languages.

The graph in Figure 3.3 was built using data on the 26 creoles and 48 substrate languages representing the different branches of the Niger-Congo family that

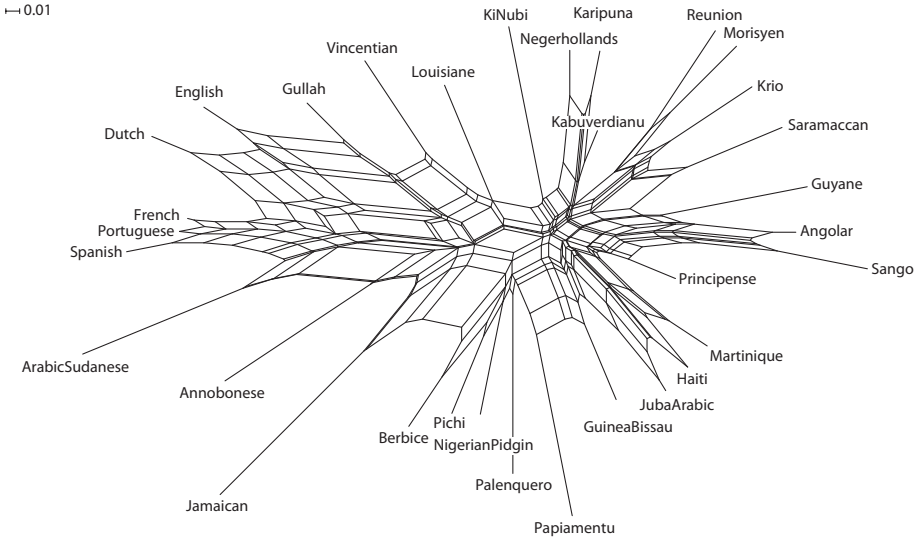


Figure 3.2 A network of 26 creoles and their six lexifiers.

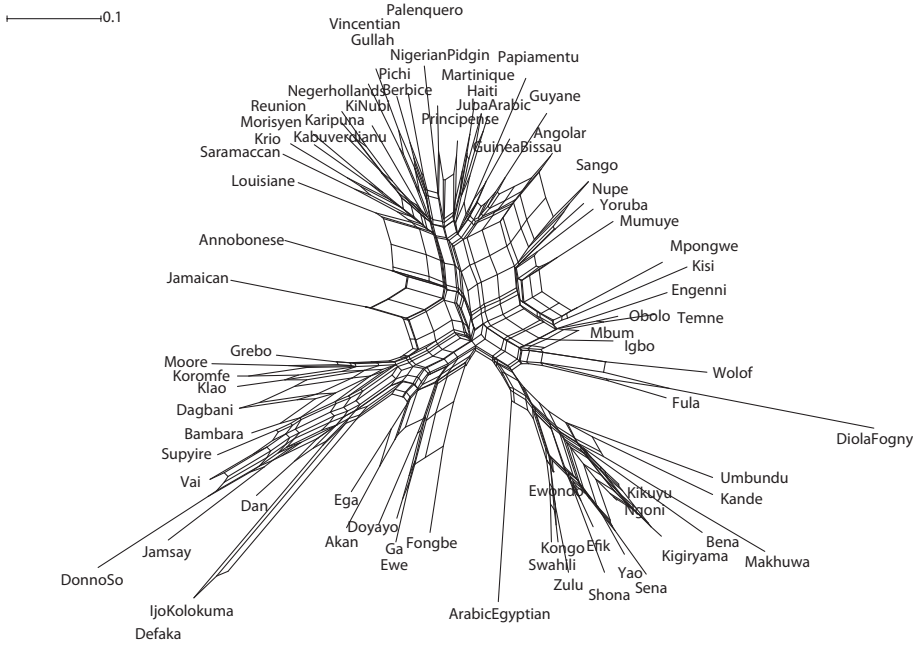


Figure 3.3 Network of 26 creoles and 48 Niger-Congo substrate languages.

were involved in the creation of Atlantic creoles (principally Adamawa-Ubangi, Atlantic, Bantu, Ijoid, Kru, Kwa and Mande). All the creoles are grouped in a single cluster at the top of the graph, again without showing any groupings which could be consistently explained either by geography or by their lexifier. Only Sango, a creole spoken in the Central African Republic and based on Adamawa-Ubangi languages, appears at the periphery of the creole cluster – but removed from the three other Adamawa-Ubangi languages present in the sample (Doyayo, Mbum and Mumuye). The genealogy of the included substrate languages is partially reflected in the various clusters.

These results suggest that the typological make-up of creoles is rather similar and that the selection of stable features adequately reflects the grammatical parts of the linguistic systems that were not greatly affected by the creolization process. Hence, the selected stable features which were also present in the lexifier and substrate languages were not continued in the creolization process.

3.7 Creoles and genetic affiliation: *Stammbaum*, convergence, contact

The genetic classification of creoles has been a long-debated issue, especially after the publication of Thomason & Kaufman (1988), who posited a non-genetic origin of creoles and other high-contact varieties due to a break in intergenerational transmission, resulting in languages that inherit their grammar and lexicon from different sources (Thomason & Kaufman 1988: 200). A number of papers addressing the question have been produced since (see e.g. Owens 1991; Dimmendaal 1995; Clements 2002; Cardoso 2008; Lefebvre 2011 among many others), but no consensus has been reached and the question of how best to classify creoles and represent their interrelationships remains unsettled. Hence, the application of phylogenetic networks to incorporate lateral transfer in contact varieties is a highly-appealing method for representing the reticulation events that characterize this type of languages. Evidently, this is also true for more conservative languages, since no language can be claimed to be completely immune to contact-induced change (see for instance Friðriksson 2008 for a study of stability and change in Icelandic).

As the results presented in the previous section suggest, the multiple parentage of creoles is not obvious when applying a selection of stable features. In order to assess the degree of similarity between creoles, we have produced a new graph showing the relationships between the 26 sampled creoles and presented in Figure 3.4.

The creoles are spread throughout the graph and do not pattern according to either lexifier, substrate or geography, apart from a few cases where diffusion is known to have played a role (e.g. the Indian Ocean creoles from Mauritius and

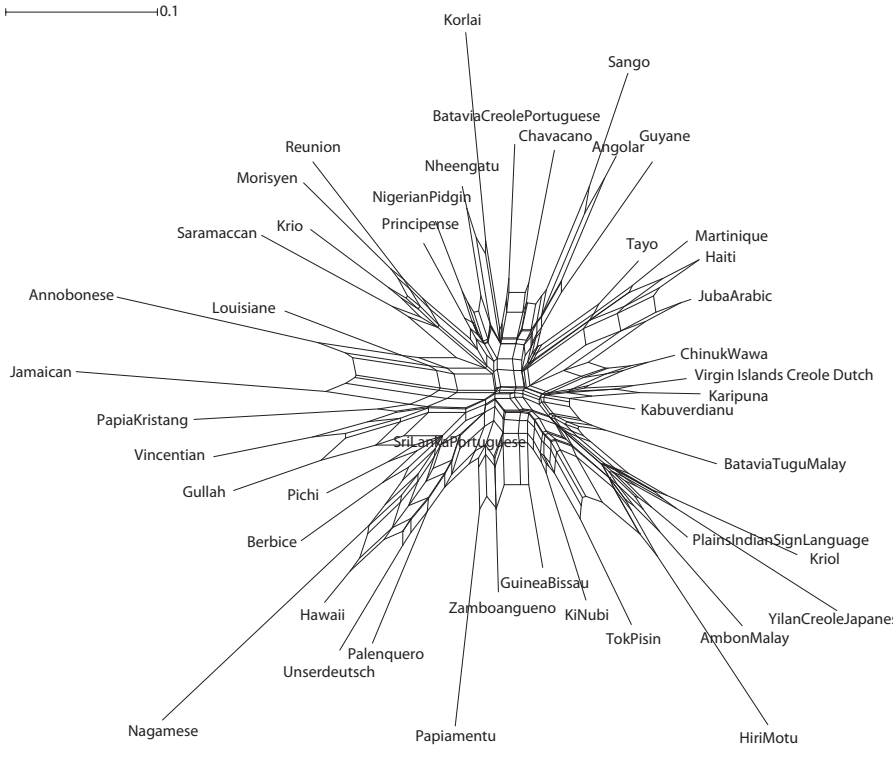


Figure 3.4 26 creoles in a network.

Réunion in the top left, or those spoken in Haiti and Martinique, on the upper right). Again, these results suggest that the classification of creoles does not depend on the genealogical background of the languages involved in their creation, but rather, that they should be considered a phylogenetic unit in their own right, albeit on different grounds than traditional language families (i.e. not due to inherited traits but rather because of common grammatical structures resulting from the similarity of circumstances that led to their emergence).

3.8 A cognitive account of creole genesis

In order to explain the similarities that can be observed between creoles, regardless of their lexifier or of the substrate languages involved in their creation, as shown above (see also Bakker et al. 2011), we have to understand the conditions surrounding the emergence and development of creole languages. Several factors have to be considered, but the most relevant factor is the pidgin ancestry of creole languages (Parkvall & Goyette forthcoming). Much of the grammatical apparatus that many

older languages display, was filtered through a bottleneck, which resulted in the disappearance of irregularities and other 'ornamental' features such as gender agreement, thus facilitating effective communication. Hence, the pidgin past of creoles is reflected linguistically in their synchronic structural make-up, in that several properties found in older languages are absent in creoles.

New insights on the exact nature of these features were recently gained from quantitative studies based on WALS (Daval-Markussen 2011; 2013). In this section, we will discuss the five features that were extracted as characteristic of creoles and provide a cognitively-oriented explanation as to their widespread occurrence in creoles worldwide.

Feature 38A (Indefinite Articles) in WALS describes the occurrence of indefinite articles and allows for five feature values (Dryer 2013a). Many creoles use the same word for the indefinite article as for the numeral 'one', corresponding to the second most-common value attested in the available sample of WALS languages (112 out of the 534 languages included, or 20%). Therefore, the occurrence of this feature value is relatively common in the languages of the world. Its presence in a vast majority of creoles can be explained as a direct outcome of the creolization process followed by the grammaticalization of the numeral 'one' for marking indefiniteness or non-specificity. Studies in grammaticalization have shown that languages that have an indefinite article equivalent to the numeral 'one' are believed to have gone through a similar developmental path (Givón 1981; 1984: 410–411; Heine 1997: 65–82; Heine & Kuteva 2007: 45–46).

The presence (obligatory or optional) or absence of numeral classifiers is covered by WALS feature 55A (Gil 2013). Only three feature values are possible and these are attested in 400 languages. The most frequent possibility in the sampled WALS languages is an absence of numeral classifiers and creoles align with this preference, as classifiers are not attested in creoles. This can be explained by the relative youth of creoles, since classifier systems tend to develop slowly over time (see Aikhenvald 2000) and are primarily found as an areal feature in languages of Meso- and South America and Southeast Asia (Aikhenvald 2000: 101).

WALS feature 69A describes the position of tense and aspect markers in a large sample of languages of the world (Dryer 2013c). One feature value relates to the absence of inflection, while the other values describe the various ways of marking tense and aspect distinctions morphologically. Of the 1131 attested WALS languages, only 152 (or 13%) do not display tense-aspect inflection altogether (two of which, incidentally, are the creoles Ndyuka and Sango). Languages that lack inflection are typically located in Southeast Asia and Africa. Due to the scarcity of languages worldwide displaying this preference and to its widespread occurrence in creoles, this phenomenon can be directly linked to the process of creolization. Indeed, it is a symptom of the morphological reduction alluded to above, in

connection with an initial phase of pidginization in the development of creoles, which together result in extremely analytic languages. In turn, in order to express tense-aspect distinctions, the most obvious choice would be to use lexical items (primarily verbs, adverbs and auxiliaries) carrying the semantics of the various aspectual nuances and to use them as verbal markers, following a grammaticalization path similar to the one that can be observed in standard non-creolizing language change.

With regard to WALS feature 112A, which looks at the distribution of negative morphemes in a sample of 1157 languages (Dryer 2013b), almost half of the languages (502, or 43%) exhibit a preference for a negative particle (the African creole Sango being one of them) rather than being marked morphologically or by double negation. Hence, this feature is not at all uncommon in the languages of the world and therefore cannot be entirely explained as being a unique result of the creolization process.

Finally, feature 117A in WALS relates to predicative possession and shows the distribution of five different strategies in 240 languages (Stassen 2013). A subset of the languages in the sample (63, or 26%) display a preference for marking possession with a verb meaning 'to have'. These languages are predominantly located in Europe, and therefore this strategy is also found in most lexifiers. However, the limited size of the sample does not permit further generalizations that would rule out an influence from substrate languages.

These five properties are almost universally present in the creoles of the world and are rarely attested together in other languages of the world (out of the 228 languages in WALS for which at least four feature values are known, only Rapanui follows the same pattern as creoles). The presence of some of these features can be taken as a direct reflection of the pidgin past of creole languages (e.g. lack of inflection and absence of complex tone), the linguistic consequences of which can still be detected synchronically. Other properties seem to surface in grammaticalization processes linked to creolization. These structural properties can therefore characterize creoles on linguistic grounds only. Bakker (2014) used these features to identify creoles with non-European lexifiers, with some success.

3.9 Conclusions

We reviewed the most important conceptual and methodological advances in both disciplines, especially during the past decades, which witnessed an explosion of studies taking advantage of computers for drawing phylogenies. In linguistics, phylogenetic analyses are becoming increasingly popular and diverse in the scope of issues. The theory-driven part was followed by a series of phylogenetic analyses

assessing the status of creoles among the languages of the world and especially among the languages that were involved in their creation.

The many parallels between biology and linguistics resulted in an extensive cross-fertilization between the two fields and this is best reflected in the number of linguistic studies using tools developed by bioinformaticians to track down evolutionary processes. Adopting methods originally designed to study biological evolution with linguistic data has become more and more widespread in historical linguistic studies. Phylogenetics has now come of age and the implementation of a wide range of algorithms has helped to advance our understanding of linguistic evolution as a whole, addressing new questions and sometimes finding unexpected answers and new challenges. The increasing number of interdisciplinary papers involving biologists, computer scientists and linguists, among others, reflects a growing interest in studies integrating insights from their respective disciplines so as to explain the patterns of human migration and diversity that can be observed worldwide today. The time now seems ripe for narrowing down on contact varieties so as to gain a better understanding of the processes and mechanisms underlying language change and that are at work in language contact situations.

References

Aikhenvald, A. Y. 2000. *Classifiers: A Typology of Noun Categorization Devices*. Cambridge: CUP.

Altschuler, E. L., Calude, A. S., Meade, A. & Pagel, M. 2013. Linguistic evidence supports date for Homeric epics. *BioEssays* 35(5): 417–420. doi: 10.1002/bies.201200165

Amano T., Sandel B., Eager H., Bulteau E., Svenning J.-C., Dalsgaard B., Rahbek C., Davies R. G. & Sutherland W. J. 2014. Global distribution and drivers of language extinction risk. *Proc. R. Soc. B* 281: 20141574.

Atkinson, Q. D. 2011. Phonemic diversity supports a serial founder effect model of language expansion from Africa. *Science* 332: 346. doi: 10.1126/science.1199295

Atkinson, Q. D. & Gray, R. D. 2005. Curious parallels and curious connections – Phylogenetic thinking in biology and historical linguistics. *Systematic Biology* 54(4): 513–526. doi: 10.1080/10635150590950317

Avise, J. C. 2000. *Phylogeography: The History and Formation of Species*. Cambridge MA: Harvard University Press.

Bakker, P. 2014. Creoles and typology: Problems of sampling and definition. *Journal of Pidgin and Creole Languages* 29(2): 437–455. doi: 10.1075/jpcl.29.2.09bak

Bakker, P., Daval-Markussen, A., Parkvall, M. & Plag, I. 2011. Creoles are typologically distinct from non-creoles. In P. Bhatt & T. Veenstra (eds), *Journal of Pidgin and Creole Languages* 26(1): 5–42. Republished in P. Bhatt & T. Veenstra (eds). 2013. *Creole Languages and Linguistic Typology* [Benjamins Current Topics 57], 9–45. Amsterdam: John Benjamins.

Barbançon, F., Evans S. N., Nakhleh, L., Ringe, D. & Warnow, T. 2013. An experimental study comparing linguistic phylogenetic reconstruction methods. *Diachronica* 30(2): 143–170. doi: 10.1075/dia.30.2.01bar

Bergsland, K. & Vogt, H. 1962. On the validity of glottochronology. *Current Anthropology* 3(2): 115–153. doi:10.1086/200264

Bouckaert, R., Lemey, P., Dunn, M., Greenhill, S. J., Alekseyenko, A. V., Drummond, A. J., Gray, R. D., Suchard, M. A. & Atkinson., Q. D. 2012. Mapping the origins and expansion of the Indo-European language family. *Science* 337(6097): 957–960. doi:10.1126/science.1219669

Bowern, C., Epps, P., Gray, R. D., Hill, J., Hunley, K., McConvell, P. & Zentz, J. 2011. Does lateral transmission obscure inheritance in hunter-gatherer languages? *PLoS ONE* 6(9): e25195. doi:10.1371/journal.pone.0025195

Bowern, C. & Atkinson, Q. 2012. Computational phylogenetics and the internal structure of Pama-Nyungan. *Language* 88(4): 817–845. doi:10.1353/lan.2012.0081

Bryant, D. & Moulton, V. 2004. NeighborNet: An agglomerative method for the construction of phylogenetic networks. *Molecular Biology and Evolution* 21: 255–265. doi:10.1093/molbev/msh018

Campbell, L. 1998. *Historical Linguistics: An Introduction*. Cambridge MA: The MIT Press.

Cardoso, H. 2008. The meaning of "European". The challenge of high-contact varieties for linguistic taxonomy. *Sophia Journal of European Studies* 1: 31–54.

Clements, J. C. 2002. On classifying language-contact varieties. In *Selected Proceedings of the First Workshop on Spanish Sociolinguistics*, L. Sayahi (ed.), 1–10. Somerville MA: Cascadilla Proceedings Project.

Collin, H. S. & Schlyter, C. J. 1827. *Corpus iuris Sueo-Gotorum antiqui. Samling af Sweriges gamla lagar*, Vol. 1: *Westgötalagen*. Stockholm: Häggström.

Darwin, C. 1859. *On the Origin of Species by Means of Natural Selection*. London: John Murray.

Daval-Markussen, A. 2011. Of Networks and Trees in Contact Linguistics: New Light on the Typology of Creoles. MA thesis, Aarhus Universitet.

Daval-Markussen, A. 2013. First steps towards a typological profile of creoles. *Acta Linguistica Hafniensia* 45(2): 274–295. doi:10.1080/03740463.2014.880606

Dediu, D. 2011. A Bayesian phylogenetic approach to estimating the stability of linguistic features and the genetic biasing of tone. *Proceedings of the Royal Society of London B* 278(1704): 474–479. doi:10.1098/rspb.2010.1595

Dediu, D. & Cysouw, M. 2013. Some structural aspects of language are more stable than others: A comparison of seven methods. *PLoS ONE* 8(1): e55009. doi:10.1371/journal.pone.0055009

Dimmendaal, G. J. 1995. Do some languages have a multi-genetic or non-genetic origin? An exercise in taxonomy. In *Actes du Cinquième Colloque de Linguistique Nilo-Saharienne / Proceedings of the Fifth Nilo-Saharan Linguistics Colloquium, Nice, 24–29 August 1992 Nilo-Saharan Language Studies* [Nilo-Saharan: Linguistic Analyses and Documentation 10], R. Nicolaï & F. Rottland (eds), 357–372. Cologne: Rüdiger Köppe.

Dixon, R. M. W. 1997. *The Rise and Fall of Languages*. Cambridge: CUP. doi:10.1017/CBO9780511612060

Donohue, M., Denham, T. & Oppenheimer, S. 2012. New methodologies for historical linguistics? Calibrating a lexicon-based methodology for diffusion vs. subgrouping. *Diachronica* 29(4): 505–522. doi:10.1075/dia.29.4.04don

Drummond, A. J., Ho, S. Y. W., Phillips, M. J. & Rambaut, A. 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biology* 4(5): e88 doi:10.1371/journal.pbio.0040088

Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution* 29: 1969–1973. doi:10.1093/molbev/mss075

Dryer, M. S. & Haspelmath, M. (eds). 2013. *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. <http://wals.info>

Dryer, M. S. 2013a. Indefinite articles. In Dryer & Haspelmath (eds). <http://wals.info/chapter/38> (11 November 2014).

Dryer, M. S. 2013b. Negative morphemes. In Dryer & Haspelmath (eds). <http://wals.info/chapter/112> (11 November 2014).

Dryer, M. S. 2013c. Position of tense-aspect affixes. In Dryer & Haspelmath (eds). <http://wals.info/chapter/69> (11 November 2014).

Dunn, M. 2009. Contact and phylogeny in Island Melanesia. *Lingua* 119(11): 1664–1678. doi: 10.1016/j.lingua.2007.10.026

Dunn, M., Terrill, A., Reesink, G., Foley, R. A., & Levinson, S. C. 2005. Structural phylogenetics and the reconstruction of ancient language history. *Science* 309(5743): 2072–2075. doi: 10.1126/science.1114615

Dunn, M., Levinson, S. C., Lindström, E., Reesink, G. & Terrill, A. 2008. Structural phylogeny in historical linguistics: Methodological explorations applied in island Melanesia. *Language* 84: 710–759. doi: 10.1353/lan.0.0069

Ellison, T. M. 2014. Identifying typological features likely to show synchronic variation. Paper presented at the Australian Linguistic Society Annual Conference, The University of Newcastle, Australia 10–12 December.

Erdem, E., Lifschitz, V. & Ringe, D. 2006. Temporal phylogenetic networks and logic programming. *Theory and Practice of Logic Programming* 6: 539–558. doi: 10.1017/S1471068406002729

Enfield, N. J. 2003. *Linguistic Epidemiology: Semantics and Grammar of Language Contact in Mainland Southeast Asia*. London: Routledge Curzon.

Felsenstein, J. 2004. *Inferring Phylogenies*. Sunderland MA: Sinauer Associates.

Forster, P., Toth, A. & Bandelt, H.-J. 1998. Evolutionary networks of word lists: Visualising the relationships between Alpine Romance languages. *Journal of Quantitative Linguistics* 5: 174–187. doi: 10.1080/09296179808590125

Fortunato, L. & Jordan, F. 2010. Your place or mine? A phylogenetic comparative analysis of marital residence in Indo-European and Austronesian societies. *Philosophical Transactions of the Royal Society B* 365: 3913–3922. doi: 10.1098/rstb.2010.0017

Friðriksson, F. 2008. Language Change vs. Stability in Conservative Language Communities: A Case Study of Icelandic. PhD dissertation, University of Gothenburg.

Gavin, M. C., Botero, C. A., Bowern, C., Colwell, R. K., Dunn, M., Dunn, R. R., Gray, R. D., Kirby, K. R., McCarter, J., Powell, A., Rangel, T. F., Stepp, J. R., Trautwein, M., Verdolin, J. L. & Yanega, G. 2013. Toward a mechanistic understanding of linguistic diversity. *BioScience* 63(7): 524–535. doi: 10.1525/bio.2013.63.7.6

Geisler, H. & List, J.-M. 2013. Do languages grow on trees? The tree metaphor in the history of linguistics. In *Classification and Evolution in Biology, Linguistics and the History of Science*, H. Fangerau, H. Geisler, T. Halling & W. Martin (eds), 111–124. Stuttgart: Franz Steiner.

Gil, D. 2013. Numeral classifiers. In Dryer & Haspelmath (eds). <http://wals.info/chapter/55> (11 November 2014).

Givón, T. 1981. On the development of the numeral 'one' as an indefinite marker. *Folia Linguistica Historica* 2(1): 35–53. doi: 10.1515/flih.1981.2.1.35

Givón, T. 1984. *Syntax: A Functional-typological Introduction*, Vol. 1. Amsterdam: John Benjamins. doi: 10.1075/z.17

Gong, T., Lan, S. & Zhang, M. 2014a. Modelling language evolution: Examples and predictions. *Physics of Life Reviews* 11: 280–302. doi: 10.1016/j.plrev.2013.11.009

Gong, T., Lan, S. & Zhang, M. 2014b. Key issues for the prosperity of modelling research of language evolution. Reply to comments on "Modelling language evolution: Examples and predictions". *Physics of Life Reviews* 11: 324–328. doi:10.1016/j.plrev.2014.04.001

Gray, R. D. & Jordan, F. M. 2000. Language trees support the express-train sequence of Austronesian expansion. *Nature* 405(6790): 1052–1055. doi:10.1038/35016575

Gray, R. D. & Atkinson, Q. D. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* 426: 435–439. doi:10.1038/nature02029

Gray, R. D., Drummond, A. J. & Greenhill, S. J. 2009. Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* 323: 479–483. doi:10.1126/science.1166858

Gray, R. D., Bryant, D. & Greenhill, S. J. 2010. On the shape and fabric of human history. *Philosophical Transactions of the Royal Society* 365: 3923–3933. doi:10.1098/rstb.2010.0162

Greenhill, S. J., Currie, T. E. & Gray, R. D. 2009. Does horizontal transmission invalidate cultural phylogenies? *Proceedings of the Royal Society of London B* 276: 2299–2306. doi:10.1098/rspb.2008.1944

Greenhill, S. J., Atkinson, Q. D., Meade, A. & Gray, R. D. 2010. The shape and tempo of language evolution. *Proceedings of the Royal Society of London B* 277: 2443–2450. doi:10.1098/rspb.2010.0051

Hall, Robert A. Jr. 1959. Neo-Melanesian and glottochronology. *International Journal of American Linguistics* 25(4): 265–267. doi:10.1086/464542

Hancock, I. F. 1987. A preliminary classification of the anglophone Atlantic creoles with syntactic data from thirty-three representative dialects. In *Pidgin and Creole Languages. Essays in Memory of John E. Reinecke*, G. G. Gilbert (ed.), 264–333. Honolulu HI: University of Hawaii Press.

Heggarty, P. 2006. Interdisciplinary indiscipline? Can phylogenetic methods meaningfully be applied to language data – and to dating language? In *Phylogenetic Methods and the Prehistory of Languages*, P. Forster & C. Renfrew (eds), 183–194. Cambridge: McDonald Institute for Archaeological Research.

Heine, B. 1997. *Cognitive Foundations of Grammar*. Oxford: OUP.

Heine, B. & Kuteva, T. 2007. *The Genesis of Grammar. A Reconstruction* [Studies in the Evolution of Language 9]. Oxford: OUP.

Hennig, W. 1950. *Grundzüge einer Theorie der Phylogenetischen Systematik*. Berlin: Deutscher zentralverlag.

Holden, C. J. 2002. Bantu language trees reflect the spread of farming across sub-Saharan Africa: A maximum-parsimony analysis. *Proceedings of the Royal Society of London Series B* 269: 793–799. doi:10.1098/rspb.2002.1955

Honkola, T., Vesakoski, O., Korhonen, K., Lehtinen, J., Syrjanen, K. & Wahlberg N. 2013. Cultural and climatic changes shape the evolutionary history of the Uralic languages. *Journal of Evolutionary Biology* 26(6): 1244–1253. doi:10.1111/jeb.12107

Huelsenbeck, J. P. & Ronquist, F. 2001. MRBAYES: Bayesian inference of phylogeny. *Bioinformatics* 17: 754–755. doi:10.1093/bioinformatics/17.8.754

Huson, D. H. & Bryant, D. 2006. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* 23: 254–267. doi:10.1093/molbev/msj030

Jansson, F., Parkvall, M. & Strimling, P. 2015 Modeling the evolution of creoles. *Language Dynamics and Change* 5: 1–51. doi:10.1163/22105832-00501005

Kanj, I. A., Nakhleh, L. & Xia, G. 2008. The compatibility of binary characters on phylogenetic networks: Complexity and parameterized algorithms. *Algorithmica* 51(2): 99–128. doi:10.1007/s00453-007-9046-1

Kraft, K. H., Brown, C. H., Nabhan, G. P., Luedeling, E., Ruiz, J. D. L., d'Eeckenbrugge, G. C., Hijmans, R. J. & Gepts, P. 2014. Multiple lines of evidence for the origin of domesticated chili pepper, *Capsicum annuum, in Mexico. Proceedings of The National Academy of Sciences of The United States Of America* 111: 6165–6170. doi: 10.1073/pnas.1308933111

Labov, W. 2007. Transmission and diffusion. *Language* 83: 344–387. doi: 10.1353/lan.2007.0082

Lee, S. & Hasegawa, T. 2011. Bayesian phylogenetic analysis supports an agricultural origin of Japonic languages. *Proceedings of the Royal Society B* 278(1725): 3662–3669. doi: 10.1098/rspb.2011.0518

Lefebvre, C. 1998. *Creole Genesis and the Acquisition of Grammar: The Case of Haitian Creole*. Cambridge: CUP.

Lefebvre, C. 2011. The problem of the typological classification of creoles. In *Creoles, their Substrates, and Language Typology* [Typological Studies in Language 95], C. Lefebvre (ed.), 3–33. Amsterdam: John Benjamins. doi: 10.1075/tsl.95.03lef

List, J.-M., Nelson-Sathi, S., Geisler, H. & Martin, W. 2014a. Networks of lexical borrowing and lateral gene transfer in language and genome evolution. *BioEssays* 36: 141–150. doi: 10.1002/bies.201300096

List, J.-M., Nelson-Sathi, S., Martin W. & Geisler, H. 2014b. Using phylogenetic networks to model Chinese dialect history. *Language Dynamics and Change* 4(2): 222–252. doi: 10.1163/22105832-00402008

Marx, V. 2013. Biology: The big challenges of big data. *Nature* 498: 255–260. doi: 10.1038/498255a

Maurits, L. & Griffiths, T. L. 2014. Tracing the roots of syntax with Bayesian phylogenetics. *Proceedings of the National Academy of Sciences* 111(37): 13576–13581. doi: 10.1073/pnas.1319042111

McMahon, A. & McMahon, R. 2003. Finding families: Quantitative methods in language classification. *Transactions of the Philological Society* 101(1): 7–55. doi: 10.1111/1467-968X.00108

McMahon, A. & McMahon, R. 2006. Why linguists don't do dates: Evidence from Indo-European and Australian languages. In *Phylogenetic Methods and the Prehistory of Languages*, P. Forster & C. Renfrew (eds), 153–160. Cambridge: McDonald Institute for Archaeological Research.

McWhorter, J. & Parkvall, M. 2002. Pas tout à fait du français: Une étude créole. *Études Créoles* XXV(1): 179–231.

Mir, C., Zerjal, T., Combes, V., Dumas, F., Madur, D., Bedoya, C., Dreisigacker, S., Franco, J., Grudloyma, P., Hao, P. X., Hearne, S., Jampatong, C., Laloe, D., Muthamia, Z., Nguyen, T., Prasanna, B. M., Taba, S., Xie, C. X., Yunus, M., Zhang, S., Warburton, M. L. & Charcosset, A. 2013. Out of America: tracing the genetic footprints of the global diffusion of maize. *Theoretical and Applied Genetics* 126: 2671–2682. doi: 10.1007/s00122-013-2164-z

Mufwene, S. S. 2007. Les créoles: De nouvelles variétés indo- européennes désavouées? In *Actes du colloque: Créolisation linguistique et sciences humaines*, 59–70. Paris: Les Presses universitaires Haitiano-Antillaises.

Mühlhäusler, P. 1980. Structural expansion and the process of creolization. In *Theoretical Orientations in Creole Studies*, A. Valdman & A. Highfield, (eds), 19– 55. New York NY: Academic Press.

Nakhleh, L., Warnow, T., Ringe, D. & Evans, S. N. 2005. A comparison of phylogenetic reconstruction methods on an IE dataset. *Transactions of the Philological Society* 3(2): 171–192. doi: 10.1111/j.1467-968X.2005.00149.x

Nettle, D. 1999. Is the rate of linguistic change constant? *Lingua* 108: 119–136 doi: 10.1016/S0024-3841(98)00047-3

Nichols, J. 1992. *Linguistic Diversity in Space and Time*. Chicago IL: University of Chicago Press. doi:10.7208/chicago/9780226580593.001.0001

Nichols, J. & Warnow, T. 2008. Tutorial on computational linguistic phylogeny. *Language and Linguistics Compass* 2(5): 760–820. doi:10.1111/j.1749-818X.2008.00082.x

Nurbakova, D., Rusakov, S. & Alexandrov, V. 2013. Quantifying uncertainty in phylogenetic studies of the Slavonic languages. Procedia Computer Science. 2013 International Conference on Computational Science 18: 2269–2277. doi:10.1016/j.procs.2013.05.398.

Owens, J. 1991. Nubi, genetic linguistics, and language classification. *Anthropological Linguistics* 33(1): 1–30.

Pagel, M. 2009. Human language as a culturally transmitted replicator. *Nature Reviews Genetics* 10: 405–415. doi:10.1038/nrg2560

Pagel, M., Atkinson, Q. D., Calude, A. S. & Meade, A. 2013. Ultraconserved words point to deep language ancestry across Eurasia. *Proceedings of the National Academy of Sciences* 110(21): 8471–8476. doi:10.1073/pnas.1218726110

Parkvall, M. 2008. Which parts of language are the most stable? *STUF* 61(3): 234–250.

Parkvall, M. & Goyette, S. Forthcoming. *Principia Creolica*.

Parkvall, M., Jansson, F. & Strimling, P. 2013. Simulating the genesis of Mauritian. *Acta Linguistica Hafniensia: International Journal of Linguistics* 45(2): 265–273. doi:10.1080/03740463.2013.900998

Perrier, X., De Langhe, E., Donohue, M., Lentfer, C., Vrydaghs, L., Bakry, F., Carreel, F., Hippolyte, I., Horry, J.-P., Jenny, C., Lebot, V., Risterucci, A.-M., Tomekpe, K., Doutrelepont, H., Ball, T., Manwaring, J., de Maret, P. & Denham, T. 2011. Multidisciplinary perspectives on banana (Musa spp.) domestication. *Proceedings of the National Academy of Sciences of the United States of America* 108(28): 11311–11318. doi:10.1073/pnas.1102001108

Reesink, G., Singer, R., & Dunn, M. 2009. Explaining the linguistic diversity of Sahul using population models. *PLoS Biology* 7(11), e1000241. doi:10.1371/journal.pbio.1000241

Renfrew, C. 1987. *Archaeology and Language: The Puzzle of Indo-European Origins*. Cambridge: CUP.

Rexová, K., Frynta, D. & Zrzavy, J. 2003. Cladistic analysis of languages: Indo-European classification based on lexicostatistical data. *Cladistics* 19: 120–127. doi:10.1111/j.1096-0031.2003.tb00299.x

Rutschmann, F. 2006. Molecular dating of phylogenetic trees: A brief review of current methods that estimate divergence times. *Diversity and Distributions* 12: 35–48. doi:10.1111/j.1366–9516.2006.00210.x

Saitou, N. & Nei, M. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4(4): 406–425.

Schlegel, F. 1808. *Ueber die sprache und weisheit der Indier: Ein beitrag zur begrundung der alterthumskunde. Heidelberg*: Mohr und Zimmer.

Schleicher, A. 1853. Die ersten Spaltungen des indogermanischen Urvolkes. *Allgemeine Zeitung für Wissenschaft und Literatur* (August 1853): 786–787.

Schleicher, A. 1863. *Die Darwinsche Theorie und die Sprachwissenschaft*. Weimar: Hermann Böhlau.

Schmidt, J. 1871. *Zur Geschichte des indogermanischen Vocalismus*, Vol. 1. Weimar: H. Böhlau.

Sicoli, M. A. & Holton, G. 2014. Linguistic phylogenies support back-migration from Beringia to Asia. *PLoS ONE* 9(3): e91722. doi:10.1371/journal.pone.0091722

Stassen, L. 2013. Predicative possession. In Dryer & Haspelmath (eds). <http://wals.info/chapter/117> (11 November 2014).

Steel, M. A., Penny, D. & Hendy, M. D. 1988. Loss of information in genetic distances. *Nature* 336(6195): 118. doi:10.1038/336118a0

Swadesh, M. 1952. Lexico-statistic dating of prehistoric ethnic contacts. *Proceedings of the American Philosophical Society* 96: 453–463.

Swadesh, M. 1955. Toward greater accuracy in lexicostatistical dating. *International Journal of American Linguistics* 21: 121–137. doi:10.1086/464321

Swadesh, M. 1972. *The Origin and Diversification of Languages*. London: Routledge & Kegan Paul.

Thomason, S. G. & Kaufman, T. 1988. *Language Contact, Creolization, and Genetic Linguistics*. Berkeley CA: University of California Press.

Towner, M. C., Grote, M. N., Venti, J. & Borgerhoff Mulder, M. 2012. Cultural macroevolution on neighbor graphs. *Human Nature* 23: 283–305. doi:10.1007/s12110-012-9142-z

Turvey, S. & Pettorelli, N. 2014. Spatial congruence in language and species richness but not threat in the world's top linguistic hotspot. *Proceedings of the Royal Society B* 218: 20141644.

Walker, R. S. & Ribeiro, L. A. 2011. Bayesian phylogeography of the Arawak expansion in lowland South America. *Proceedings of the Royal Society B* 278(1718): 2562–2567. doi:10.1098/rspb.2010.2579

Wichmann, S. & Kamholz, D. 2008. A stability metric for typological features. *STUF* 61(3): 251–262.

Wichmann, S. & Holman, E. W. 2009. *Temporal Stability of Linguistic Typological Features*. Munich: Lincom.

Wichmann, S. & Grant, A. P. (eds). 2012. *Quantitative Approaches to Linguistic Diversity: Commemorating the Centenary of the Birth of Morris Swadesh* [Benjamins Current Topics 46]. Amsterdam: John Benjamins. doi:10.1075/bct.46

Wiley, E. O. 1988. Vicariance biogeography. *Annual Review of Ecology, Evolution, and Systematics* 19: 513–542. doi:10.1146/annurev.es.19.110188.002501

Winford, D. 2005. Contact-induced changes. Classification and processes. *Diachronica* 22(2): 373–427. doi:10.1075/dia.22.2.05win

Zhang, Z., Wei, S. G., Gui, H. S., Yuan, Z. Y. & Li, S. B. 2014. The contribution of genetic diversity to subdivide populations living in the silk road of China. *PLoS ONE* 9(5): e97344. doi:10.1371/journal.pone.0097344