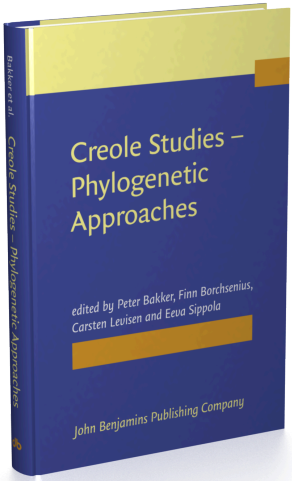
CHAPTER 4

Methods

On the use of networks in the study of language contact

Peter Bakker | Aarhus University
Eeva Sippola | University of Bremen
Finn Borchsenius | Aarhus University

https://doi.org/10.1075/z.211.04bak



Pages 59–78 of
Creole Studies – Phylogenetic Approaches
Edited by Peter Bakker, Finn Borchsenius, Carsten Levisen and Eeva M. Sippola
2017. x, 414 pp.

CHAPTER 4

Methods

On the use of networks in the study of language contact

Peter Bakker, Eeva Sippola and Finn Borchsenius
Aarhus University / University of Bremen / Aarhus University

This chapter provides an overview of the phylogenetic models used in this book. In the introduction, we present the aims and limitations of the chapter and clarify some basic concepts. After presenting the steps of linguistic phylogenetic analyses, we proceed to explain the different data types that are examined in this book, such as lexical and typological features, and describe how these are coded. In the final sections, we discuss the resulting network and tree models and how to interpret the graphic representations built using linguistic data. In short, this chapter enables the reader to interpret the graphs and to assess the validity of the results.

4.1 Introduction

This introductory chapter provides an overview of the methods applied in this book, focusing mainly on network models. Computational phylogenetic methods have been developed in evolutionary biology and have been applied for the past two decades in historical linguistic studies to track language developments, evolution and relationships (see also Chapter 3). We believe that they are also suitable for the comparative study of contact languages in general, and creoles in particular, and can respond to the challenges creoles present for models of language evolution in historical linguistics.

Phylogenies, or evolutionary trees, are basic structures that help to visualize differences between species and how they are related to one another. In phylogenetic studies today, computers assist researchers in systematizing massive data sets and in analysing the differences statistically (Felsenstein 2004: xix). Evolution of life forms and products of cultures, including language, can be visualized in trees of descent or other types of graphs. These trees map the phylogeny, i.e. the

DOI 10.1075/z.211.04bak

development of entities that have evolved from a common ancestor, and they map the degree of similarity between those entities.

The chapters in this book study creole languages and their interrelationships using mainly a method called splits networks. Depending on the aims, these are often applied in cases where the linguistic history and the links between varieties are unclear or ambiguous, as is often the case with creoles.

Although the field of phylogenetic linguistics has attracted a fair amount of attention among linguists in recent years (Bouckaert et al. 2012; Dunn et al. 2008; Dunn 2014; Gray et al. 2007; Wichmann & Good 2014), only a limited number of works include a comprehensive introduction to the methods most commonly used by linguists (e.g. McMahon & McMahon 2005; Nichols & Warnow 2008). This chapter is intended as a basic introduction to the computational phylogenetic methods relevant for this book, especially splits networks, from a linguist's point of view. For a more detailed explanation of the concepts and methods used in the field in general and recent developments, we refer the reader to Dunn (2014) and the references therein.

The remainder of the chapter is structured as follows. After presenting the basic steps required for performing linguistic phylogenetic analyses, we proceed to present the different data types that are examined in this book (such as lexical and typological features) and to describe how they are coded. In the last sections, we discuss the resulting network and tree models, how to read them and how to interpret the information they display.

4.2 Steps of analysis: Encoding, representation, and interpretation

The application of phylogenetic and computational methods to the study of languages can be divided into three stages (after Heggarty 2006: 184). These stages include the *encoding* step, the *representation* step and the *interpretation* step. In the *encoding* phase, the linguistic data are selected and transformed into an adequate input data format that will be used for phylogenetic analysis. For some of the studies in this book, the authors have collected, selected and encoded the linguistic properties they study. This can include data both from secondary sources such as reference grammars and findings from fieldwork. Other authors use data from typological databases compiled by others, such as the *Atlas of Pidgin and Creole Structures* (APiCS; Michaelis et al. 2013a, b) and the *World Atlas of Language Structures* (WALS; Dryer et al. 2013). Another important dataset is the one compiled in the volume edited by Holm and Patrick (2007), *Comparative Creole Syntax*, which includes a large grammatical dataset of 18 creole languages worldwide. These data attest the presence or absence of a range of typological

features in the selected creoles and can be summarized as numerical values. If a feature exists in the language, this will be encoded as '1' and if the feature is absent as '0' (e.g. presence of tone = 1, absence of tone = 0). Thus, the collected data are typically first converted into numerical values in a matrix, where the languages are arranged in rows and lexical or grammatical features in columns. These numbers in the matrix form the basis of the next step, which involves converting the input data to a graphical, more easily readable form. Figure 4.1. is an example of a matrix with binary values for 21 typological properties in 24 languages from the Amazon.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Cayubaba	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0	0	0
Baure	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	1	1	1	1	0	0
Mosetén	1	1	1	1	0	1	1	1	1	1	0	1	0	1	1	1	1	1	0	0	1
Lakondê	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	0	0	1	1	0	0
Kwaza	1	1	1	0	1	1	1	1	0	1	0	1	1	1	1	0	1	0	1	1	0
Chiquitano	1	1	?	1	1	1	1	1	0	1	1	1	1	1	?	1	1	0	0	?	0
Itonama	1	1	1	1	1	1	1	1	0	1	1	1	0	0	0	0	1	1	1	0	0
Gavião	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	0	0	0	0	0
Quechua	1	1	1	1	1	0	1	1	1	1	0	1	1	1	0	1	0	0	0	1	0
Aymara	1	1	1	1	1	0	1	1	1	1	0	1	1	1	0	1	0	0	0	1	0
Leko	1	1	1	1	1	1	1	1	0	0	0	0	1	1	0	1	1	1	0	0	0
Chácobo	1	0	1	1	0	1	1	1	1	0	1	1	0	1	0	1	0	1	0	1	0
Movima	1	1	1	1	1	1	0	0	0	1	1	1	0	1	0	0	0	0	1	0	1
Mekens	1	1	1	0	1	1	0	1	1	1	1	?	0	0	1	0	0	0	0	1	0
Aikanã	1	1	?	0	0	1	1	1	0	0	?	?	1	1	1	0	1	0	1	1	0
Kanoê	1	1	0	0	1	?	?	1	1	1	1	0	1	0	1	0	1	0	1	0	0
Wari'	1	0	1	1	1	1	?	0	0	1	1	?	1	0	0	0	1	1	0	0	1
Yuki	0	1	1	1	1	?	0	0	1	1	1	1	1	0	1	1	0	0	0	0	0
Cavineña	0	0	1	1	0	0	1	1	1	0	1	1	0	1	0	1	1	0	0	1	0
Uru	?	1	1	1	1	0	1	0	1	1	0	0	1	1	0	1	0	0	0	0	0
Arikapú	1	1	?	1	1	0	1	0	1	0	0	?	0	0	1	0	1	1	0	0	0
Yurakaré	1	1	0	1	1	1	0	1	1	0	0	0	1	0	0	1	0	0	0	0	0
Karo	1	0	1	0	0	1	0	0	1	1	1	0	0	0	1	0	0	0	1	0	0
Karitiana	0	1	1	1	0	1	0	0	1	1	1	?	0	0	1	0	0	0	0	0	0

(1) Subordination through nominalization; (2) = Cross-reference; (3) = Evidentiality; (4) = Vowel harmony; (5) = Head marking; (6) = Verbal number; (7) = Directionals; (8) = Polysynthesis; (9) = Postpositions; (10) = Inclusive/exclusive distinction; (11) = Alienable/inalienable distinction; (12) = Verb classification; (13)= Strictly Nom-Acc alignment system; (14) = Asymmetrical morphology; (15) = Nasal harmony; (16) = Nominal number; (17) = Head-Modifier; (18) = Possession marked on possessum; (19) = Classifiers; (20) = Switch reference; (21) = Grammaticalized system.

Figure 4.1 Feature matrix for 21 properties of 24 Guaporé-Mamoré languages (Brazil, Bolivia), from Crevels & van der Voort (2008).

In the *representation* stage, these matrices are analyzed and transformed into a graph representing the relationships between the languages being studied. Depending on the assumptions underlying the researcher's hypothesis, a method for phylogenetic reconstruction will be selected. Most of the networks presented in this book are produced using the Neighbor-Net method (Bryant & Moulton 2004), which uses the Neighbor-Joining (NJ) algorithm (Saitou & Nei 1987) to construct phylogenetic trees. NJ is a distance-based algorithm that starts off by calculating the cumulative differences between each pair of languages in the matrix. These distances will then serve as the numerical basis for constructing a phylogenetic network. Figure 4.2 shows a splits network produced by Neighbor-Net on the basis of the matrix of Amazonian languages in Figure 4.1.

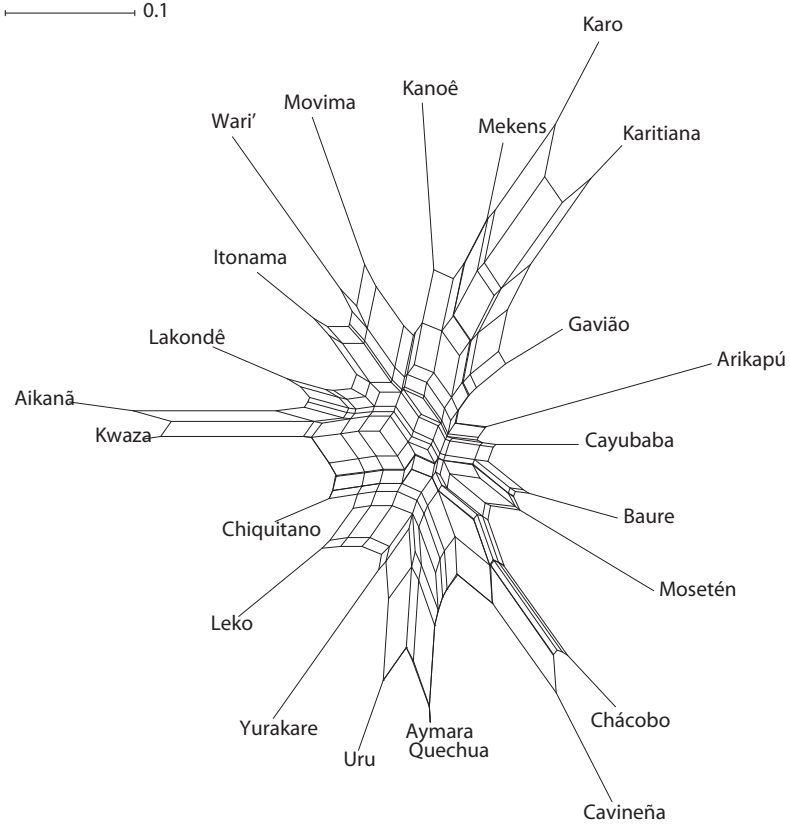


Figure 4.2 A splits network of 24 languages of Lowland Amazonia (based on data from Crevels & van der Voort 2008).

In this book, most authors chose to visualize their data as unrooted splits networks using the Neighbor-Net method as implemented in the software package SplitsTree4 (Huson & Bryant 2006). SplitsTree4 is relatively easy to use and makes it possible to produce different types of graphs and perform various statistical calculations.

In the third step, the *interpretation* stage, the results are used to draw inferences on the relationships between the selected languages. Are the visualized results in line with what we know about the languages included in the comparison? Are there languages that end up in unexpected clusters? Are there regional or genetic clusters? In our experience and based on reports from the literature on studies using phylogenetic software, almost all results conform to expectations based on earlier, non-computational studies, but there can be surprises as well.

The programs and network models can help us visualize the degree of similarity between languages in the data, but linguists with detailed knowledge of the languages in question are always needed to sort out the significance of the results (see also McMahon & McMahon 2005: 154). Unexpected results could, for example, be caused by coding errors or by incorrect preconceptions. Indeed, computer programs mechanically follow the selected algorithms on the data input, whereas humans are influenced by their ideas and expert knowledge. Hence, the data selection and encoding procedure are crucial steps towards achieving reliable results.

4.3 Data types

When working with linguistic data in phylogenetic analyses, two main types of data can be used: lexical and typological data.

4.3.1 Lexical data

Lexical data can be coded according to formal properties (e.g. phonological form) or cognate relationships (are forms with the same meaning derivable from one original form?) or semantic properties (e.g. does the word meaning "father" include "father's brother", or is there a separate word like "uncle"?). The most commonly used type of data in phylogenetic studies is comparative lexical data based on cognate relationships of the lexical items. Lexical similarity methods, based on formal properties of words, are commonly used in investigations of dialectal data, because the presumption of cognacy in dialects is well-justified (Prokić 2010; Dunn 2014). In general, these kinds of lexical similarity measures work best at relatively shallow time depths (Greenhill 2011). Although it might first seem that

formal lexical similarity measures would be suitable for investigating creoles (if they have the same lexifier) due to their shallow time depth, this is not directly the case. Any exclusively lexical measure would make creoles look more similar to their lexifier than they are, as the lexicon of a creole originates predominantly in one language (except a handful of creoles which have a mixed lexicon). However, the grammatical system of a creole differs in many respects from the structures observed in the lexifier.

Due to synonymy and to the closeness of lexifier and creole forms in the lexicon, it can sometimes be difficult to determine which items should be selected when meanings are compared (see also McMahon & McMahon 2005: 156 and Dunn 2014: 193 ff. for a similar discussion). To address the issue, different levels of granularity can be used. As coding in itself is a form of analysis, as is the selection of word meanings, it is always necessary to carefully describe the procedure followed. In this book, word forms are compared in some of the chapters, but obviously only in groups of languages with the same lexifier (Dutch, Spanish, French, Portuguese, English). In addition, some chapters also use lexical semantics rather than mere word forms as data input (Levisen & Bøegh this volume, Chapter 13–15).

4.3.2 Typological data

Obviously, the exclusive use of lexical evidence in the case of creoles, be it based on phonological forms or cognacy, does not tell the whole story about the history and origin of these languages. Structural and functional features are therefore used prominently in this book. Such grammatical properties have also been used by others in the study of the evolution and interrelationships of both contact (Bakker et al. 2011) and non-contact languages (Dunn et al. 2005, 2008; Cysouw & Comrie 2009; Longobardi et al. 2013; Bøegh et al. 2016). This type of data is particularly useful for the analysis of grammatical contact and admixture (Dunn 2014: 193). Besides, it should also be kept in mind that the set of structural features at play in a language is subject to evolution, just like the lexicon of a language. In this sense, computer programs should also be able to detect linguistic evolution on a typological as well as a lexical level.

Today, the existence of typological databases makes this task much easier. In 2013, a new database of contact languages with large amounts of data on typological features with copious examples was published in the *Atlas of Pidgin and Creole Language Structures* (APiCS, Michaelis et al. 2013a) and the accompanying online version (Michaelis et al. 2013b). APiCS includes comparable synchronic data on the grammatical and lexical structures of 76 pidgins, creoles, and mixed languages. It provides a great resource for large-scale comparisons of these languages. The atlas was compiled by a team of experts who filled in the feature values and provided

example sentences for each language. The accompanying examples make it possible to evaluate the feature value selections and thus assess the data. This procedure can be contrasted with previous typological enterprises, where experts coded many different languages for a few specific features. This was done when putting together WALS (Dryer & Haspelmath 2013), the next database we will present.

The *World Atlas of Language Structures* (Dryer & Haspelmath 2013) describes the distribution of 144 predominantly morphosyntactic features in 2880 languages, including all known language families and covering all continents. Due to the data collection procedure followed for producing this work (a single author filled out many data points for many languages for a few features), the degree of completeness of the database varies enormously (from every single data point to none). This work has been widely exploited and used for a variety of purposes, including phylogenetic studies (see for example Wichmann & Saunders 2007; Bakker et al. 2011).

Such databases can be exploited and expanded for conducting phylogenetic analyses. Although studying linguistic similarities between unrelated languages with typological features is rather uncontroversial (Dunn 2014), there are some challenges regarding the use of typological data in phylogenetic studies. Typological features evolve in a limited design space, which result in a higher probability of chance homology. In other words, features can have the same value due to chance, simply because there are only few possibilities available. Whereas words with the same meaning can be expressed quite differently between languages, both in the length and nature of the phonemes, features such as verbs and subjects combine in a limited number of patterns (e.g. VS, SV or variable). The likelihood that two words with the same meaning in two unrelated languages accidentally have the same form is quite small compared to accidental similarities in structures. Thus, if one encounters two languages from different parts of the world (and with no known contact history) with a VS order, these orders will most likely be the result of independent developments, rather than inheritance. But an independent development is less likely in cases where two languages call a dog *hund*.

As mentioned above, most of the chapters in this book use typological data to study the evolution of creole (and other contact) languages and their relationships to other varieties. These typological data may be of a grammatical or lexical-typological nature. Several contributors to this book work almost exclusively with structural data, especially when they are interested in general properties of creole languages.

Since both structural and lexical properties of languages evolve, both can therefore be used for detecting a phylogenetic signal. The study of lexical evolution is much older than the study of typological evolution, and the latter is as yet less well-understood. However, some historical linguists have claimed that the most stable typological properties of languages would reach a deeper time depth than would be achieved with the lexicon only (Nichols 1992, 1994; Wichmann 2015).

The reason that the use of typological data is much less common, may be that most of the input data in the studies in general come from established studies in historical linguistics, especially involving cognate data of common words which are quite readily available, such as Swadesh lists for several languages and families. Stable words are the most common source in standard procedures in establishing genetic relationships between languages. They are sometimes combined with grammatical structures, notably grammatical morphemes. However, it is not uncommon to also use data involving phonological similarity or typological features when traditional historical linguistic input in the form of word lists is unavailable or unfeasible (Dunn 2014). In addition, it has been shown that the most accurate classifications of languages into traditional language families are obtained through a combination of lexical and typological data (Bakker et al. 2009).

4.4 Data coding

As mentioned above, when gathering input data for a phylogenetic analysis, a data matrix is generally constructed, where rows represent the given taxa, i.e. languages, and the columns represent different linguistic features that are being compared (Nichols & Warnow 2008: 764). The features can be words, speech sounds, grammatical constructions, parameters, etc., which will be reduced to a single value, a simple number. It should be emphasized that more often than not, such a number summarizes a complex reality in one numerical value. Romance languages, for instance, are notorious for displaying some highly-frequent adjectives which appear before the noun, whereas all other adjectives follow the noun. Scoring the result for this feature with a value standing for ADJ-N, N-ADJ or "free" could all be considered correct and wrong at the same time, which makes it important to be transparent and consistent about the coding principles.

The coding of features can be done either in a binary manner (e.g. 0 and 1) or with multi-state values (e.g. 1, 2, 3, 4). Binary features attest, for instance, the presence or absence of a word, phoneme, semantic distinction, grammatical distinction or syntactic construction, or they can cover two opposites (e.g. verb-initial or not). Multi-state features allow the presence of various values (e.g. the six possible orders of S, V and O, perhaps with an additional feature for "free order"). Naturally, multi-state features can be transformed into binary features by collapsing or separating some values, depending on the overall design of the feature set.

Finally, another important issue related to the data collection procedure is how many data points are necessary in order to conduct meaningful analyses (i.e. filled-in values in your matrix). There is no general rule of thumb, but in principle, the more, the better, and a few dozen features, if chosen carefully, may be indicative of

classificatory patterns. One can also have as a purpose to isolate as few features as possible that would give certain results, e.g. creoles in one cluster, and non-creoles in another (Daval-Markussen 2011, 2013). Whereas it has been argued that 40 of the most stable words are sufficient to map most language families of the world accurately (Brown et al. 2008; Holman et al. 2008), an ideal or minimal number of typological features has not been subject to discussion.

4.5 Networks and trees

Trees are well-known from traditional historical linguistics studies in the form of graphic representations of language families, where an earlier protolanguage splits off into new languages. However, such trees showing bifurcating splits are only one type of graph among several. Huson and Bryant (2006: 254) define the cover term phylogenetic network as any graphic illustration that depicts evolutionary relationships between a set of taxa (languages in our case). In terms of graph theory, phylogenetic trees form a subset of phylogenetic networks. Thus, this definition encompasses both phylogenetic trees (e.g. language family trees) and splits networks (see Figure 4.3 for an overview of various tree representations).

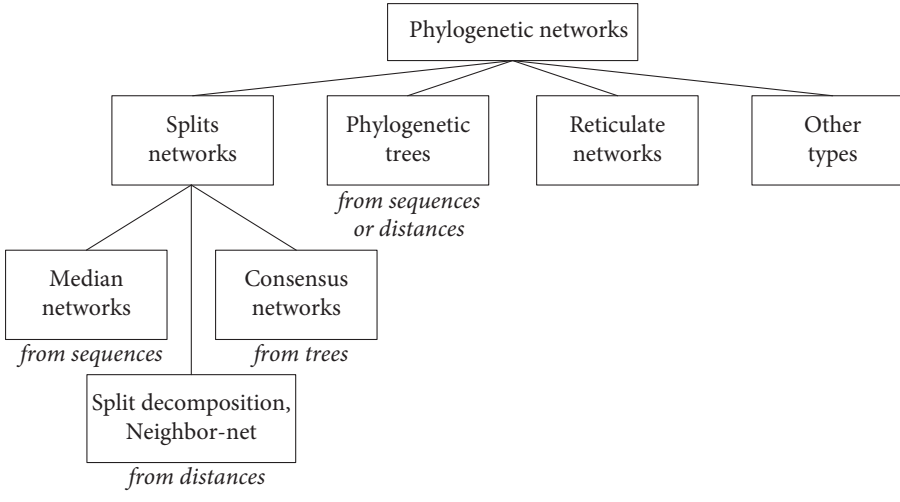


Figure 4.3 Diagram of different network types (based on Huson & Bryant 2006: 255).

There are explicit and implicit networks. In explicit networks, an actual evolutionary pathway is specified based on the assumption that the languages included in the sample are related. In implicit or abstract networks, there is no assumption of an evolutionary pathway and the network merely represents ambiguities or

conflict in the data (Huson et al. 2010: 70–71). The networks in this book are almost without exception implicit networks. An evolutionary inference is not necessarily made from the results.

Trees are graphs consisting of a set of 'nodes' (splits in a tree) and a set of 'edges' or 'branches' (see Figure 4.4 for a terminological overview), each of which connects a pair of branches (subfamilies) or taxa (languages). Two nodes share only one path, which means that the tree is acyclic (Nichols & Warnow 2008: 761). Figure 4.4 shows a rooted tree in which the bottom node (labelled "root") represents the linguistic ancestor (or protolanguage) of the selected languages. The figure shows ancestral nodes for the groups *abcd*, *abc* and *bc*. The tree can also be reversed, with the root on top, or sideways, with the root either on the left or on the right. In a rooted tree, an ancestor to all the sampled languages is assumed, whereas an unrooted tree simply maps similarities without assuming a common ancestor for the branches or languages. An unrooted tree is depicted in Figure 4.5.

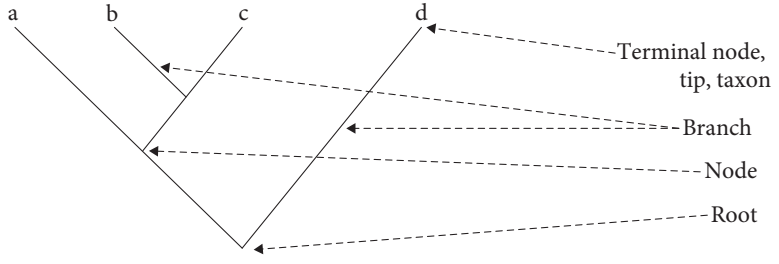


Figure 4.4 Example of a rooted tree (adapted from Nichols & Warnow 2008: 761).

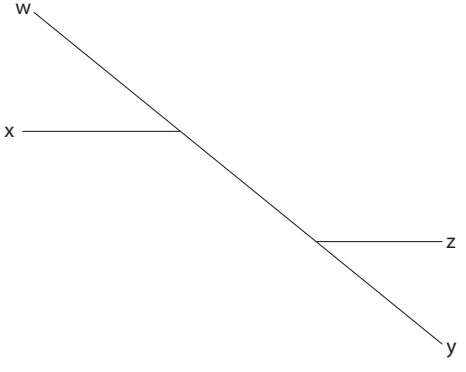


Figure 4.5 Example of an unrooted tree, where w and x are closer to each other than w and z, or w and y (based on Nichols & Warnow 2008: 761).

These rooted and unrooted trees are unambiguous in the sense that they present a single solution for a classification or a visual representation of similarities.

In a splits network, on the other hand, there is more than one way to get from one terminal node (tip, or taxon) to another. A splits network is thus able to depict conflicts and ambiguity in the data, which can be, for example, due to reticulation events, missing data, etc. Reticulations could be interpreted as resulting from, for instance, borrowings, like the lexical borrowings from French into English, which link the Germanic language English with the Romance language French. Reticulations are visible in the box-shaped webbing that can be seen in networks with conflicting information.

Figure 4.6 (based on Lehtinen et al. 2014: 201) shows a Neighbor-Net network where some of the splits are marked with dashed lines. The graph also includes bootstrap values for the splits (see Section 4.6 for an explanation of the bootstrapping procedure). Languages *a, b, c, d*, and *e* are at the tips of the independent lines emerging from the central part of the web. The parallel lines marked with a dashed line (Split 1) separate languages *a* and *b* from the others. Also, the connection between *a* and *c* (Split 2) conflicts with the branch containing *a* and *b* and weakens their cluster.

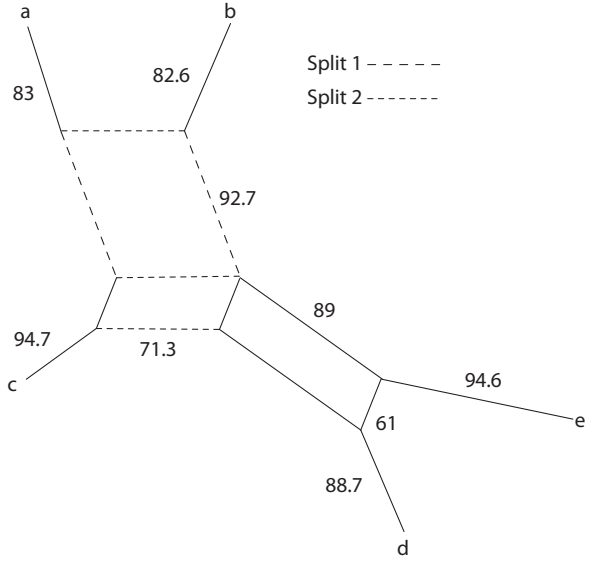


Figure 4.6 Neighbor-Net network with bootstrap values and two splits indicated with dashed lines (based on Lehtinen et al. 2014: 201).

In Figure 4.7 we present a graph based on real linguistic data from Dutch-based creole varieties. The resulting network shows the interrelationships between Dutch and several of its creole varieties included in the comparison.

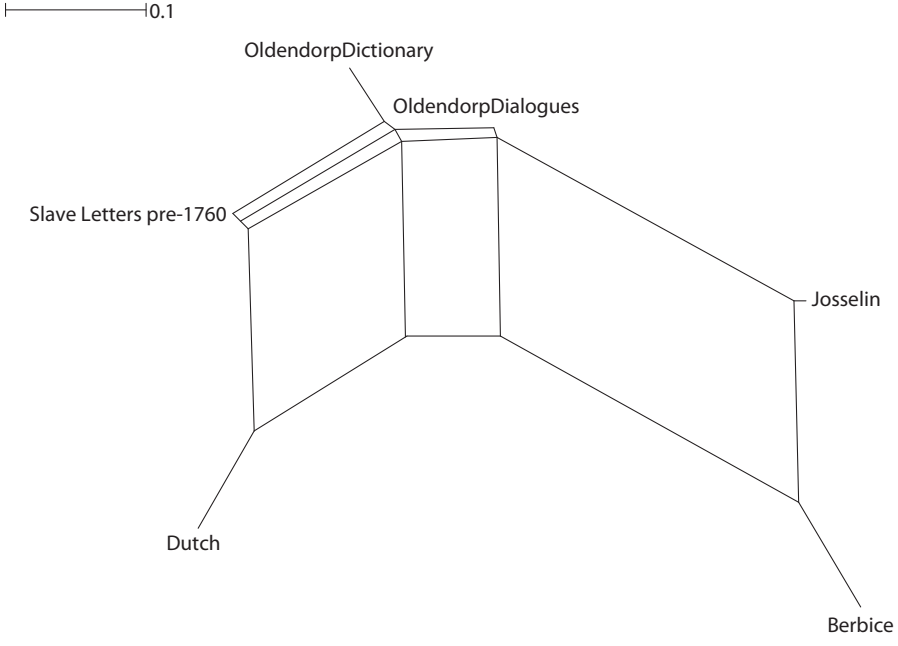


Figure 4.7 A splits network of Dutch, Berbice Creole, and three varieties of Virgin Islands Creole Dutch (see Bakker, this volume, Chapter 10, for a more detailed description of the varieties included in the comparison).

The main type of networks used in this book are splits networks. As already mentioned, splits networks include representations of incompatible and ambiguous signals in the data. This is also useful when working with high-contact varieties such as creoles, where multiple influences can be observed in the genesis, but normal change in their subsequent development. In splits networks, parallel edges, rather than single branches, represent the weighted splits that are calculated from the data. The nodes in these networks do not necessarily represent ancestral states of evolution, but they rather represent ambiguous signals in the data.[1] Splits networks provide therefore an implicit representation of several possible evolutionary

1. This is due to the fact that even if every splits network represents a unique collection of splits, the uniqueness does not hold in the other direction: a given collection of splits can have many different splits network representations (Huson & Bryant 2006: 256).

pathways (Huson & Bryant 2006: 255). In this way, they can depict a combination of tree-like signal and 'noise' in the data (Nichols & Warnow 2008: 764).

The historical development of a language, including contact events, can be interpreted from these graphs. However, it is important to note that the parallel lines (splits) that produce the webbing do not necessarily imply contact with a substrate or adstrate language in the case of creoles. They can also indicate homoplasy i.e. independent innovations, where people creating creoles invent parallel solutions to communicative challenges, for instance in the creation of a system of articles. The parallel lines may also be indicative of deviation due to insufficient or misrepresentative feature coding. It can therefore be necessary to go back into the unprocessed data in order to explain the patterns which can be observed in a split network, especially the unexpected ones. This does not mean that unexpected results are wrong. Indeed, one's working hypothesis might be incorrect from the outset or it may be the case that the researcher's expectations were misconceived before starting the phylogenetic analysis proper.

4.6 Interpreting the results

Network models have many advantages when studying contact languages. For example, the Neighbor-Net algorithm can represent conflicting signals in the data and provide a detailed snapshot of the selected languages for the properties considered. It gives an overall representation of the structure of the data and works as a guide for further analysis. The interpretation of the resulting networks is not always straightforward (Bryant & Moulton 2004) and depends on exactly how the splits were constructed (Huson & Bryant 2006: 256).

Let us illustrate this further by means of a number of graphs. Figure 4.1, given above, showed a matrix with binary data for 21 structural properties in 24 indigenous languages spoken in the border area of Bolivia and Brazil, which is extremely diverse linguistically. Zeros and ones indicate respectively absence and presence of a property, while question marks are used to represent uncertainties or unknown data points. By just inspecting the features, it is not easy to perceive which languages are more closely connected to others, either via common descent or through contact. This is where the computational approach can prove valuable.

Figure 4.2 showed a network based on these data. A substantial amount of internal webbing can be observed. This indicates conflicts in the signal, which may be due to contact, or independent developments – in both cases, linguistic traits that are not the result of direct descent with modification as is typical of inheritance. On the other hand, some independent lines (edges) are also quite long, as for example in the case of Arikapú and Wari', which indicates that there are also considerable

differences between these two languages and the rest of the sample. In the bottom, we can see one node with Quechua/Aymara. Looking at the matrix, it appears that both Quechua and Aymara have exactly the same scores for all the scored features in this dataset, which places them at the same terminal node. The language closest to these two is Uru. Again, this language belongs to a different family than the others, but it is spoken in an area near Quechua and Aymara and has been influenced by the former two (Muysken 2000). This historical development is visible in the webbing and the nearby positions of these languages in the graph. On the left, we can see Aikanã and Kwaza with a lot of webbing. The two languages are isolates, but they are spoken in the same region and most Kwazas also speak Aikanã, so the languages have influenced one another, which again is visible in the form of dense webbing.

We can see that some edges (lines) are longer than others, either measured from an imaginary centre or measured from the splits. On the right side, for instance, the line for Arikapú is much longer than the one for Cayubaba and Gavião. This may indicate two things: either fewer features are known for the language with the shorter line than for the one with the longer line, or it may mean that the features that are deviant for Arikapú, are more often different from all the other included languages, i.e. Arikapú is more divergent than other languages in the sample.

Let us now look at a creole example. Based on the 97 morphosyntactic binary features described in Holm and Patrick (2007), we have data on 18 creoles based on Arabic, Dutch, English, French, Nagamese, Portuguese and Spanish and from all parts of the world. In Figure 4.8, we see less webbing and longer edges, which means that fewer properties included in the comparison are shared between these creoles than between the languages in the Amazonian example before. The long edges indicate high diversity, more than in the Amazonian case. The only two languages that are closely connected are Guinea-Bissau creole and Cape Verdean creole, forming a cluster in the upper right of the graph. These two languages are both historically related and areally connected, their position in a cluster together comes therefore as no surprise. The other languages that show webbing are often not related at all, for instance Nagamese (spoken in India), Nubi (KiNubi) (spoken in Eastern Africa), in the upper left, and Palenquero (South America). This shows that the adjacent languages in the network share more structural features than languages further away, but that does not necessarily imply that these languages share a common origin.

After a number of non-creoles were added to the data set by looking up feature values in reference grammars, we get the result presented in Figure 4.9. The non-creoles (indicated in CAPITALS) chosen were either the most analytic language in their phylum, or they were among the non-complex languages of the world (see Bakker et al. 2011 for details). The isolating and non-related (to each other) languages Mina (Afro-Asiatic), Koyra Chiini (Songhay or Nilo-Saharan), and

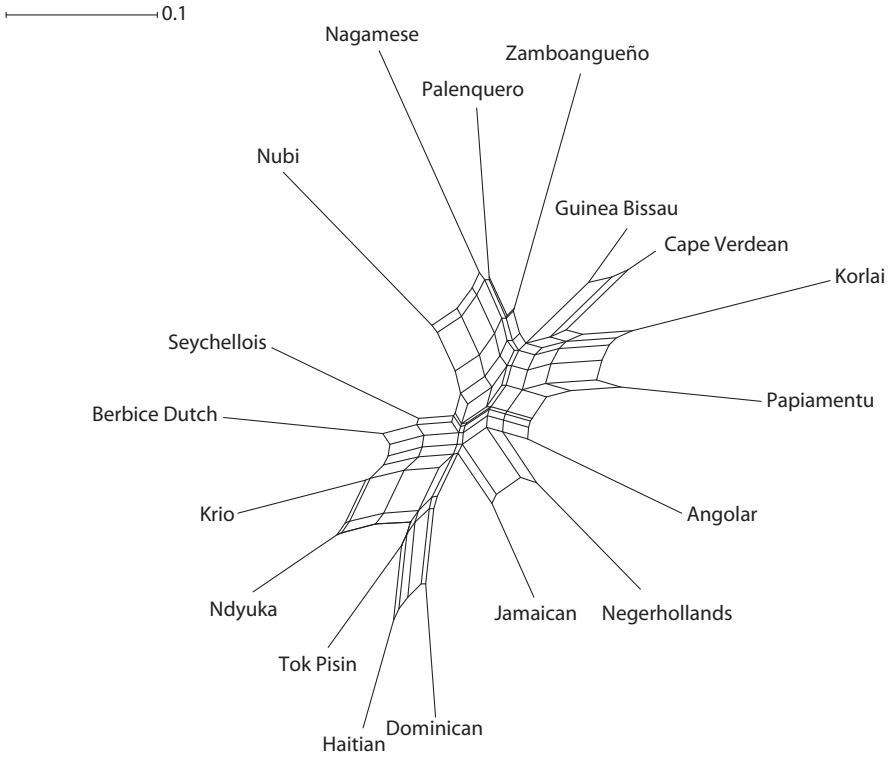


Figure 4.8 A network of 18 creole languages based on the morphosyntactic features described in Holm & Patrick (2007).

Mandarin (Sino-Tibetan), as well as relatively non-complex Pirahã (Mura-Pirahã or Amerind), Brahui (Dravidian) and Indonesian (Austronesian) are all located in the upper part of the graph, clearly separated from the creoles. Note also that the inclusion of these languages has a minor, albeit visible, effect on the internal distribution of the 18 creoles relative to one another (compare for instance the position of Angolar in Figs. 4.8 and 4.9). The 18 creole languages cluster together against the non-creoles, despite the diversity of creoles as a group.

A network can also be rooted, that is, one can suggest a certain language to be (close to) the ancestor of the sampled languages. In the case of creoles derived from a single lexifier, rooting the network with the lexifier is justified, based on the reasonable assumption that e.g. all French creoles derive from French. An example of a rooted phylogenetic network can be found in Figure 4.10.

How reliable are the positions of different languages and clusters in the different trees? A common procedure in phylogenetic studies is to calculate a statistical support measure for the various splits in a graph. The most commonly used

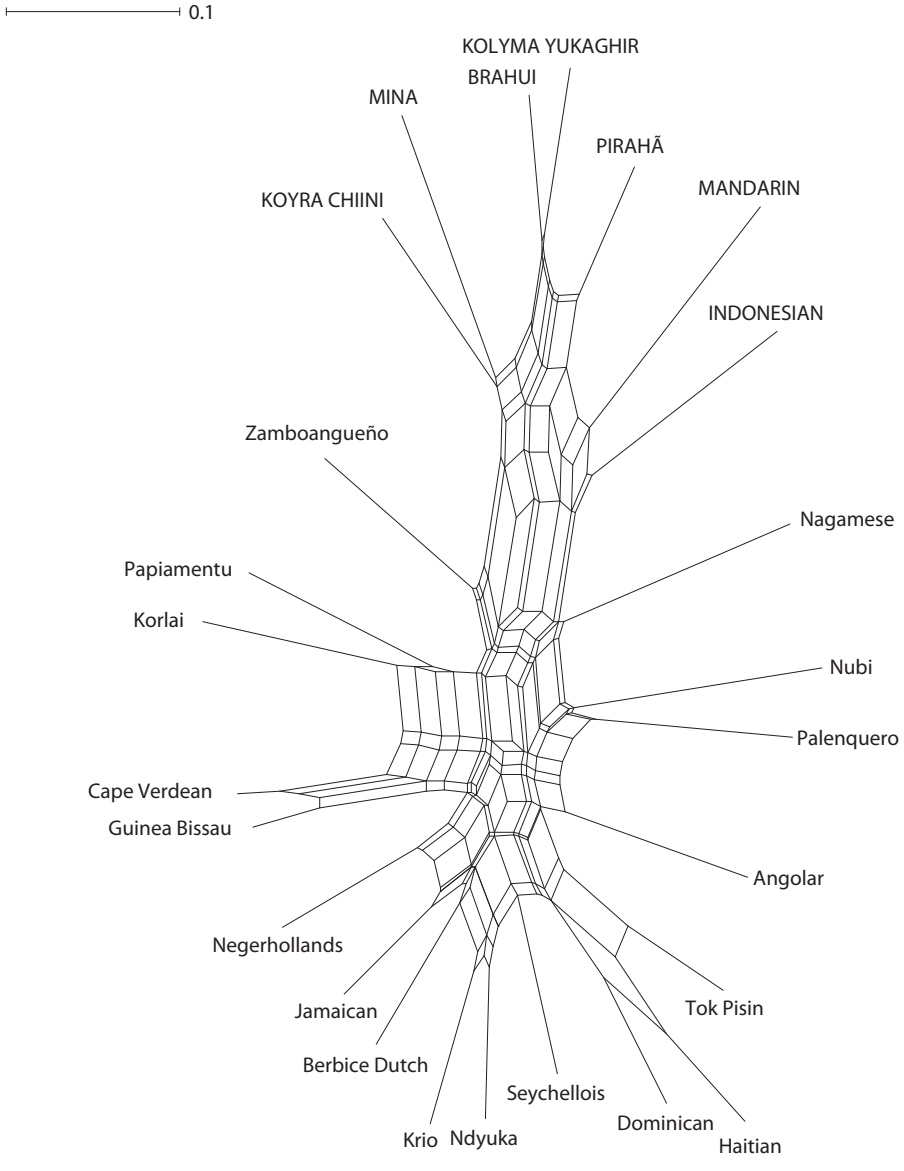


Figure 4.9 Neighbor-Net network with 18 creoles and six unrelated non-creoles.

method to calculate support values is called bootstrapping. An example of support values calculated using this method can be seen in Figure 4.6. Bootstrapping is a resampling procedure where new datasets are constructed by randomly resampling the characters in original dataset. The new dataset is then analysed using the same method as for the original dataset. Typically, this operation will be

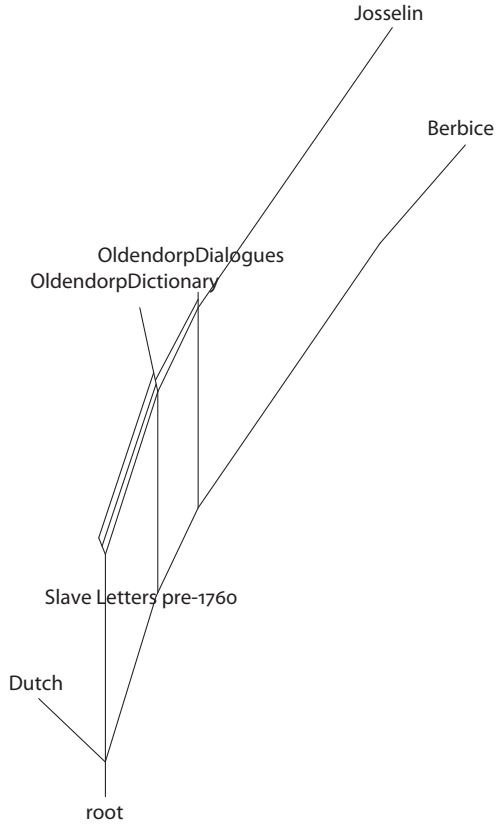


Figure 4.10 A rooted Neighbor-Net network, with Dutch, placed closest to the root, and two creoles, Berbice Creole and Virgin Islands Creole Dutch, based on four different textual sources and three time periods (see Section 10.3 for a more detailed explanation of the varieties and the periods).

repeated at least 1000 times (depending on the size of the dataset). For each split in the original graph, it is then calculated which proportion of the graphs produced by resampling, contain that particular split. A high value means that the split is supported by many characters in the dataset, whereas a low value means that the split depends on just a few critical characters and therefore might disappear if for example the dataset is enlarged by adding new features. Hence, the support value gives a statistical measure of the robustness of the result. In practice, a support value of at least 70% is considered necessary for robust results (Huson et al. 2010: 44). A rooted network of French-based creoles with bootstrap values is presented in Figure 4.11. All the edges and splits have a numerical value between 6.7 (very low) and 100 (very high).

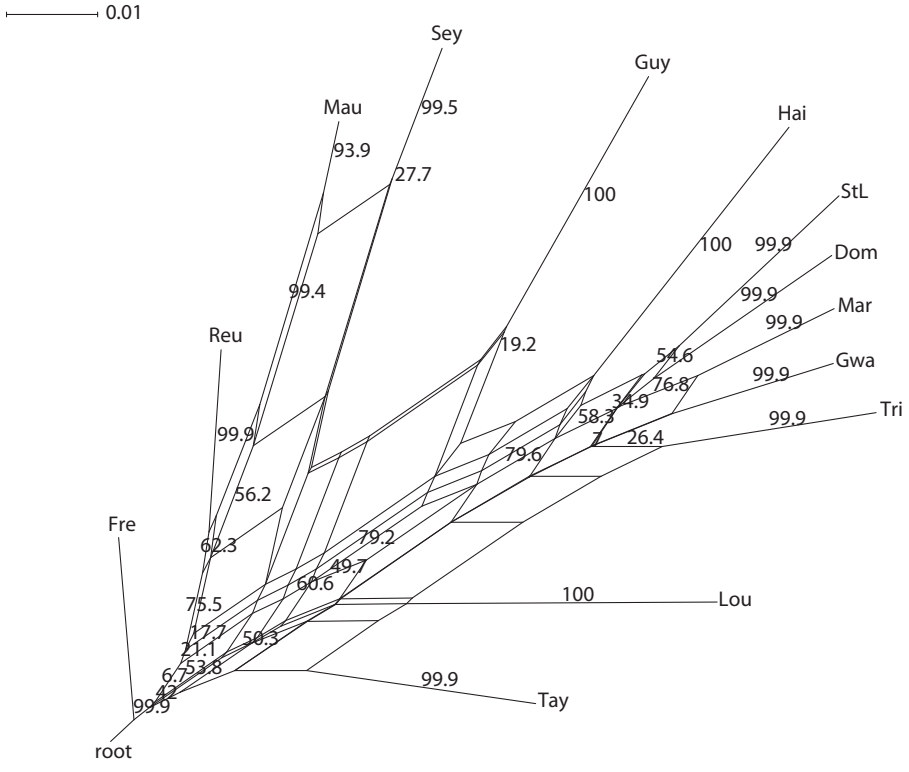


Figure 4.11 Rooted tree with French and a sample of French-based creoles with confidence values (for more details, see chapter 5).

We hope that these examples will help the readers in the study of this book by allowing them to assess the validity of the results presented throughout.

Notes

The datasets for this chapter can be found here: https://phylogenetic-creole-studies.blogspot.com

References

Bakker, D., Müller, A., Velupillai, V., Wichmann, S., Brown, C. H., Brown, P., Egorov, D., Mailhammer, R., Grant, A. & Holman, E. W. 2009. Adding typology to lexicostatistics: A combined approach to language classification. *Linguistic Typology* 13(1): 169–181. doi: 10.1515/LITY.2009.009

Bøegh, K. F., Daval-Markussen, A. & Bakker, P. 2016. A phylogenetic analysis of stable structural features in West African languages. *Studies in African Linguistics* 45(1–2): 61–94.

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: 957–60. doi: 10.1126/science.1219669

Brown, C. H., Holman, E. W., Wichmann, S. & Velupillai, V. 2008. Automated classification of the world's languages: A description of the method and preliminary results. *STUF – Language Typology and Universals* 61(4): 285–308. doi: 10.1524/stuf.2008.0026

Bryant, D. & Moulton, V. 2004. Neighbor-net: An agglomerative method for the construction of phylogenetic networks. *Molecular Biology and Evolution* 21(2): 255–265. doi: 10.1093/molbev/msh018

Crevels, M. & van der Voort, H. 2008. The Guaporé-Mamoré region as a linguistic area. In *From Linguistic Areas to Areal Linguistics* [Studies in Language Companion Series 90], P. Muysken (ed.), 151–179. Amsterdam: John Benjamins. doi: 10.1075/slcs.90.04cre

Cysouw, M. & Comrie, B. 2009. How varied typologically are the languages of Africa? In *The Cradle of Language*, P. Botha & C. Knight (eds), 189–203. Oxford: OUP.

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

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

Dryer, M. S. & Haspelmath, M. (eds). 2013. *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. <http://wals.info> (29 December 2015)

Dunn, M. 2014. Language phylogenies. In *The Routledge Handbook of Historical Linguistics*, C. Bowern & B. Evans (eds), 190–211, London: Routledge.

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

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

Gray, R. D., Greenhill, S. J. & Ross, R. M. 2007. The pleasures and perils of Darwinizing culture (with phylogenies). *Biological Theory* 2(4): 360–375. doi: 10.1162/biot.2007.2.4.360

Greenhill, S. J. 2011. Levenshtein distances fail to identify language relationships accurately. *Computational Linguistics* 37(4): 689–698. doi: 10.1162/COLI_a_00073

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.

Holm, J. & Patrick, P. L. 2007. *Comparative Creole Syntax*. London: Battlebridge.

Holman, E. W., Wichmann, S., Brown, C. H., Velupillai, V., Müller, A. & Bakker, D. 2008. Explorations in automated language classification. *Folia Linguistica* 42(2): 331–354.

Huson, D. H. & Bryant, D. 2006. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* 23(2): 254–267. doi: 10.1093/molbev/msj030

Huson, D. H., Rupp, R. & Scornavacca, C. 2010. *Phylogenetic Networks: Concepts, Algorithms and Applications*. Cambridge: CUP. doi: 10.1017/CBO9780511974076

Lehtinen, J., Honkola, T., Korhonen, K., Syrjänen, K., Wahlberg, N. & Vesakoski, O. 2014. Behind family trees: Secondary connections in Uralic language networks. *Language Dynamics and Change* 4: 189–221. doi: 10.1163/22105832-00402007

Longobardi, G., Guardiano, C., Silvestri, G., Boattini, A. & Ceolin, A. 2013. Toward a syntactic phylogeny of modern Indo-European languages. *The Journal of Historical Linguistics* 3(1): 122–152. doi: 10.1075/jhl.3.1.07lon

McMahon, A. & McMahon, R. 2005. *Language Classification by Numbers*. Oxford: OUP.

Michaelis, S. M., Maurer, P., Haspelmath, M. & Huber, M. (eds). 2013a. *Atlas of Pidgin and Creole Language Structures*. Oxford: OUP.

Michaelis, S. M., Maurer, P., Haspelmath, M. & Huber, M. (eds). 2013b. *Atlas of Pidgin and Creole Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. <http://apics-online.info> (29 December 2015).

Muysken, P. 2000. Drawn into the Aymara Mold? Notes on Uru grammar. In *Indigenous Languages of Lowland South America* [Indigenous Languages of Latin America (ILLA) Vol. 1; CNWS Publications 90], H. van der Voort & S. van de Kerke (eds), 99–109 Leiden: Research School of Asian, African, and Amerindian Studies (CNWS).

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

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

Nichols, J. 1994. The spread of language around the Pacific Rim. *Evolutionary Anthropology* 3: 206–215. doi: 10.1002/evan.1360030607

Prokić, J. 2010. *Families and Resemblances*. Groningen: Groningen Dissertations in Linguistics.

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

Wichmann, S. 2015. Diachronic stability and typology. In *Handbook of Historical Linguistics*, C. Bowern & B. Evans (eds), 212–224. London: Routledge.

Wichmann, S. & Saunders, A. 2007. How to use typological databases in historical linguistic research. *Diachronica* 24(2): 373–404. doi: 10.1075/dia.24.2.06wic

Wichmann, S. & Good, J. (eds). 2014. *Quantifying Language Dynamics: On the Cutting Edge of Areal and Phylogenetic Linguistics*. Leiden: Brill. doi: 10.1163/9789004281523