Modern Phylogenetics and Creole Evolution: Creole Family Values

Recent research into the evolution of creole languages ([1], [2], [3]) claims that novel phylogenetic tools that construct 'networks' rather than 'trees' (e.g, "Splitstree," [4]) conclusively establish that (1) creoles are typlogically simpler than other human languages, clustering into groups well apart from all other languages, using a method that is neutral with respect to the selection of features that characterize creoles as opposed to other languages; and (2) the historical contact events involved in creole formation can be recovered via this novel phylogenetic analysis. In contrast, we show that that these new tools for analyzing creole origins cannot reconstruct the evolutionary history about contact events, even in principle, let alone anything like the conventional phylogenetic evolutionary history of creoles or creole origins. Rather, the phylogenetic programs have been used simply in place of more conventional methods for cluster analysis. Here, the methods are in fact not neutral in their feature selection, being subject to ascertainment bias – a biased sampling that favors the prior selection of creole 'friendly' language features, even when drawn from just the WALS database, and thus unsurprisingly detects a 'creole typological signal' as a result. At least on phylogenetic grounds then, one can reject the view that creoles serve as "living fossils" yielding some special insight into human language origins. Given the increasing use of computational phylogenetic methods imported directly from biology to inform linguistics in this way, we conclude more generally that such methods and their results must be reviewed with great care, as is the case here. To remedy these defects, one must employ phylogenetic methods that explicitly model language contact events and incorporate the dynamics of language change within single language 'species', as in [5], quite unlike all current models of creole phylogenetic analysis.

The (evolutionary) origin of creole languages remains a well-known subject of debate, falling into one of two rough accounts: one asserting the continuity of creole languages with all other languages; and another asserting that creoles are typologically distinct from all other human languages, perhaps in virtue of a their evolutionary origins (e.g., from pidgins, as 'conventional inter-languages of an early stage'). Recent research has attempted to resolve this question via a relatively new method for linguistic phylogenetic analysis [4] that constructs reticulated networks rather than trees. Conventionally, phylogenetic language analysis uses lexical (semantically cognate) or structural language features drawn from a set of languages to construct a graphical representation of shared family traits for these languages. The end result of phylogenetic analysis is a branching tree whose tips denote traits of the (generally contemporary) observed language 'species' under analysis, whose topology indicates family relationships grounded on evolutionary history, whose branch lengths approximate time or more simply an estimated number of trait changes, and, importantly for the creole account, whose internal nodes stand for ancestral languages. Crucially, in a phylogenetic tree there is a *single* way to move from any ancestral node (including the root), down to a particular language at the tree's tips.

All phylogenetic models must adopt either explicitly or implicitly some model of evolutionary (linguistic) change so that one can correctly ascertain whether traits are common simply in virtue of common history, as opposed to independent invention. However, besides shared ancestry and independent invention, there is a third way that two linguistic 'species' might come to share a trait in common or not, and that is via *horizontal* trait transfer – as when two languages come into geographical contact, and one language borrows lexical items (or other typological, structural properties) from the other. In this case, evolution need no longer be treelike and ordinary phylogenetic models need not apply. Here, as Nichols & Warnow (2008) note, "when there is borrowing between languages, the proper graphical model will reflect that borrowing through the addition of contact edges. Such graphical models are called '*explicit* phylogenetic networks' since they represent an explicit evolutionary scenario." (2008:4-5) [our emph.]. As before, interior nodes in such a network denote ancestral languages, while the contact edges – horizontal lines – denote specific historical contact events. For example, one might posit that the Haitian Creole use of *o* is borrowing from the [xxxx] language form *au* [MdG please correct]. If so, an *explicit* phylogenetic network computer program that is purpose-built to recover horizontal contact events and legitimate ancestral states, like the one implemented by Warnow, Tandy, Steven N. Evans, Donald Ringe, and Luay Nakhleh 2006. "A stochastic model of language evolution that incorporates homoplasy and borrowing" *Phylogenetic Methods and the Prehistory of*

Languages ed. by Peter Forster and Colin Renfrew:75-90, MacDonald Institute for Archaeological Research] is at least able in principle to recover such facts from language data.

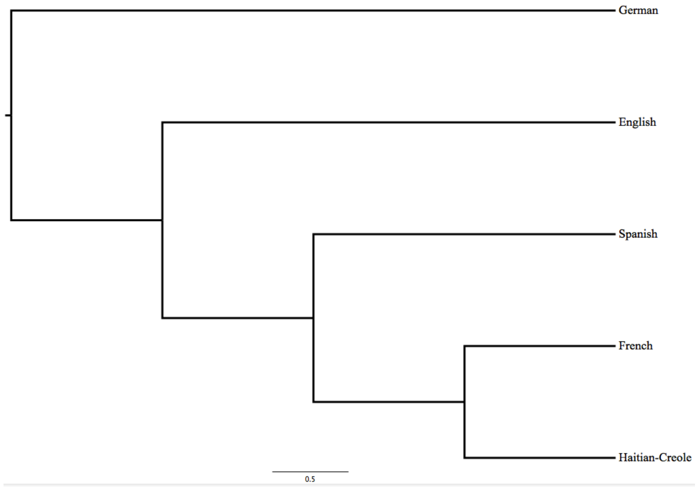
However, the authors of [1]-[4] do not use such a program. Instead, they adopt the phylogenetic program 'SplitsTree' or 'NeighborNet' [ref 7, 8 Huson & Bryant], and implementation of what is called 'split decomposition' phylogenetics, with the claim that such methods "can account for horizontal relationships, i.e., contact phenomena" [Bakker, D-M, Parkvall, 2011:11], and that these networks are therefore "ideal for application to creole languages, as both inheritance and contact played an important role in the formation of development of creoles" [Ibid, 14]. As Nicholls and Warnow (2008) note though, the output from the SplitsTree program "does not explicitly indicate any evolutionary scenario, and instead represents graphically how the input data (distances or characters) do not fit a tree exactly. Thus, the graph represents a combination of tree-like signal and the noise in the data. In particular, the internal nodes of this graph do not represent ancestors of the given languages, but are introduced in order to make possible the representation of the conflict between the different splits that are produced in the data analysis" (2008, 764-765). As a result, contrary to what is claimed in [2011 Bakker et al.] the SplitsTree method they use *cannot* reconstruct networks displaying ancestral states possibly implicated in creole language evolution; nor do the 'horizontal lines' that the program outputs as a network, as in Fig. 1 below, denote putative contact events. Perhaps in response to this fact, a later paper by the authors [D-M & Bakker, 2012, Explorations in creole research] weaken their previous claim, stating only that the method is "used in order to *visualize* the impact of various languages present in the contact situation on the new vernaculars" (2012:89, our emph.) or that the method may be "used to shed light on the relationships between creoles by presenting the results in a ...visually appealing manner" [Ibid: 90]. This too is inaccurate; all that the network can actually display are potentially alternative groupings – clusters – for languages.

To see why this is so, it is important to describe in general how the SplitsTree program computes its network output. Consider the (artificial) binary feature data for five languages as shown below, where 0/1 denotes the presence/absence of 8 different character traits:

Language	Trait values
German	01010101
English	01100110
Spanish	10101010
French	10101011
Haitian-Creole	10101001

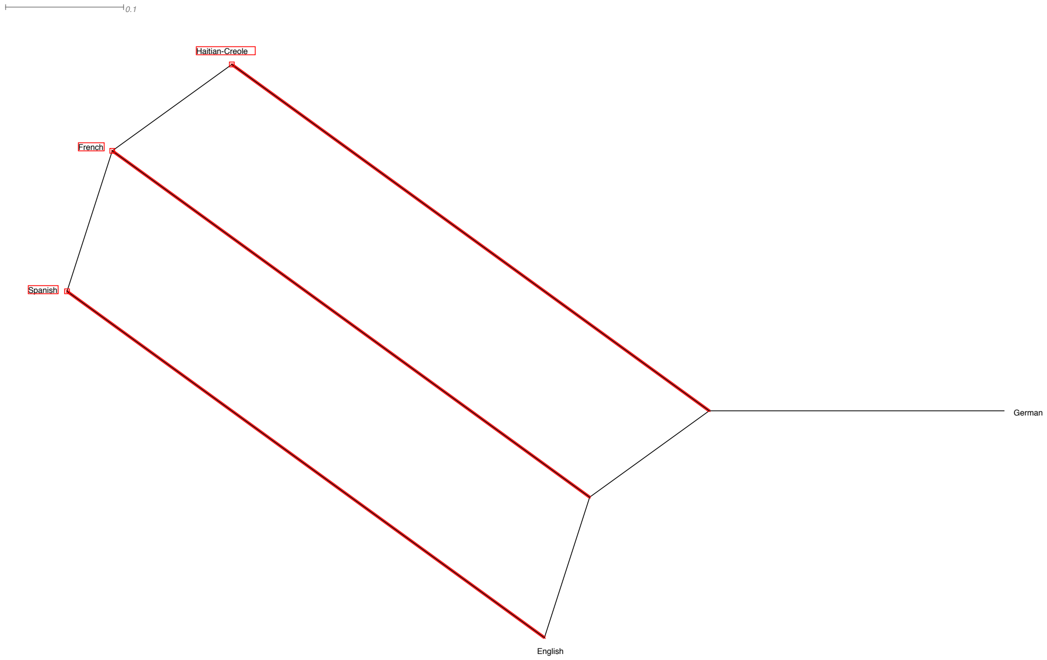
By inspection, looking down the trait columns at each language, it seems that French [Spanish? Check this Mdg] differs from Haitian Creole by exactly 1 trait feature, the last one; Spanish differs by two (the last two); and so forth, with English and then German being the most 'distant' from Haitian Creole. The phylogenetic tree one obtains by ordinary methods (using the so-called *neighbor joining* method), reflects these character differences transparently.

Figure 1.



If we analyze the same data using the *Splitstree* program advocated by Bakker *et al*, one obtains the following graphical output (using either the 'neighbor-net' version of this method or the 'split-decomposition' approach):

Figure 2.



As before, the figure above depicts Haitian Creole as 'closer' to French and Spanish than either English or German, but in addition, 3 parallel lines have been drawn in the graph – it is no longer a tree. These lines mark the 'split' between, on the one hand, German and English; and on the other hand, Spanish, French, and Haitian Creole: if one makes a single 'cut' through the three red lines, the graph would fall apart – decompose - into two completely connected, separate components – hence the term for this method, 'split decomposition.' Note that this is *all* this method computes: the various decompositions, given the trait data. In particular, as should be apparent from the Figure, the three red lines do not in any way denote possible horizontal, historically grounded transfer events. Rather, all the three red lines tell us is that features 1, 2, 5 and 6 have different values for Spanish, French, and Haitian Creole (1, 0, 1, 0) than English and German (0, 1, 0, 1).[1] The red lines simply denote those features that happen to split the graph into two separate parts,

[1] To be sure, this fact does not *preclude* one from using such graphical depictions of language clusters to suggest where one might look for evidence of historical contact. It is in this sense that Bakker *et al*. are correct in saying that split-decomposition or neighbor-net representations "may" represent contact events – it is still logically possible that they might, but this must be

without any regard for evolutionary history. Indeed, the authors of the method note, these parallel lines do *not* represent horizontal/contact events, but rather alternative possible tree branches (Huson and Bryan) – simply a way to visualize alternative species clusters. While there is nothing problematic in using visualization as some heuristic device, it is problematic if one wants to claim that this method reveals historical events. Underscoring this fact, notice that the red lines connecting French and Haitian-Creole back to the German-English portion of the graph end at nodes that have no labels whatsoever: these nodes do not denote a possible reconstructed ancestral language.

As a concrete example of how this affects the conclusions drawn in [1]-[3], consider the claim in [D-M & Bakker 2012] that this phylogenetic method can be used to adjudicate between four theories of creole formation: (1) superstratist (lexifier); (2) substratist; (3) feature pool; and (4), universalist. D-M & B's basic assumption is that each of these theories predict that non-creole languages drawn from each of these 4 types ought to cluster in an intermingled fashion with creoles. For example, they note that if the superstratist (lexifier) account is correct, then one would expect to find the creoles clustering in an intermingled fashion with their lexifier sources, e.g., French, Spanish, and so forth, so these Indo-European languages ought to be found within their associated creole clusters, e.g., Haitian creole near to French. To do this, they use the 97 morphosyntactic features from the 18 creoles in Holm & Patrick's *Comparative Creole Syntax* (CCS) [2007], along with four different groups of languages, one group for each of the four possible accounts (a superstratist group that includes 7 presumptive Indo-European lexifiers; a substratist group based on 19 presumptive West African substrates; and so forth), outputting four different SplitsTree networks. They draw two conclusions based on their four SplitsTree analyses: first, that "creoles form a relatively homogenous group of languages, in that the creoles are clearly visible and easily distinguishable from other languages" [Ibid:93]; second, that their results reject theories (1), (2), and (3), in favor of the universalist theory (4), because in every case but (4), the presumptive 'influencing' languages do not cluster with the creoles. For example, they display in their Figure 2 a SplitsTree network where Atlantic creoles stand apart from their Indo-European lexifiers. (We re-analyze this particular example in our Fig. 3 below.)

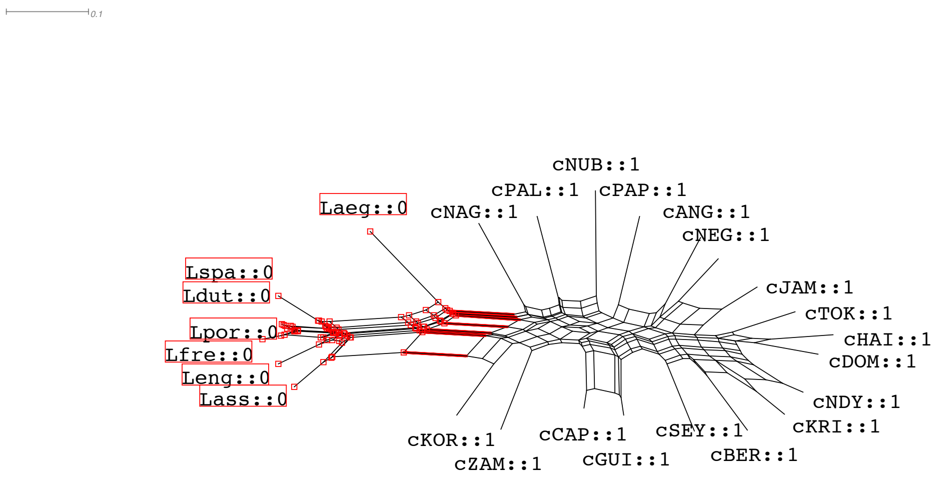
On closer inspection though, it appears that in each case D-M & Bakker 2012 have simply recovered the familiar fact that creoles differ from non-creoles with respect to how tense is marked on verbs, using pre- and post-verbal markers rather than inflections. Take their rejection of account (1), the superstratist view, where 7 I-E languages are analyzed along with 18 creoles. D-M & Bakker [2012] claim that from their SplitsTree analysis, "since creoles do not group with their respective lexifiers…this strongly suggests that superstrates have had a rather limited influence on the grammatical makeup of the incipient creoles" [Ibid: 91]. But this conclusion is misleading. First, as we have seen, the 'horizontal' webbing in a SplitsTree display does not in general indicate anything about evolutionary history. Second, the phylogenetic display's partition might not be as informative as it seems. Using the SplitsTree program, one can highlight particular features that 'force' the network to be split into different subcomponents. This is shown in our Figure 3 below, our replication of D-M & Bakker's [2012] figure using 18 superstratist and 7 I-E languages. In this figure, the red lines reveal that a single feature forces a split between the 18 creoles and 7 lexifier languages. We have also indicated alongside each language the actual value for the feature in question, either 1 or 0; one can see that all the I-E languages have a '0' for this feature, and all the creoles have a '1'; e.g., LFre, French, has value 0 for this feature, and cHai, Haitian creole, has value 1 for this same feature. Indeed, this is exactly why the program indicates that the network can be 'split' into two parts using this feature. But most importantly, what is this feature that splits the I-E languages from the Atlantic creoles? It is the first CSS feature from Holms and Patrick, their feature 1.1: whether a language has unmarked (i.e., uninflected) verbs and statives with non-past reference (value 1) or does not have unmarked verbs with non-past reference (value 0). So this phylogenetic network analysis tells us simply what was already known, namely that I-E

somehow determined in another way, and other alternatives must be ruled out, as noted by Nichols & Warnow (2008). In some places, the authors seem to be aware that the methodology is best regarded as a heuristic device, e.g., "Even though the trees and networks have been designed for mapping evolution, we use them for **"fi**nding similarities in languages that came into being independently (in most cases) from one another, and that are not in areal contact" (D-m, thesis p. 14).

languages differ with respect to their creoles in terms of verbal inflection, used in I-E languages, but absent in creoles. Given this difference, it is completely unsurprisingly that "creoles do not group with their respective lexifiers" – this was a foregone conclusion. The network display has in this sense led us astray; it has 'amplified' this single bit difference between creoles and their lexifiers, and simply returned an answer that was already known. Put another way, in this case the SplitsTree analysis is too sensitive to already-known differences between Atlantic creoles and their respective I-E lexifiers; it cannot reveal any detail about the 'influence' of the grammatical makeup of lexifiers on their creoles, as D-M & Bakker seem to imply

.

Figure 1. Re-analysis of D-M & Bakker 2012, figure 2. 18 Creoles, 7 Lexifiers, with split for CSS feature #1.1 highlighted in red.



Methodological issues

The assumptions behind the phylogenetic analysis also pose some general difficulties. First, in all their analyses aside from one analysis with multi-valued WALS data, [1]-[3] assume that each typological feature change – from a 0 to a 1 or the reverse – counts as one 'evolutionary time step,' equally for all characters, and further, that each such feature value is independent of another. This is required to establish a valid distance metrc. Neither of these assumptions seems warranted in general; contact events can cause clusters of feature changes, and feature values can be correlated with one another [reference needed from MdG?]. So for example, an Atlantic creole language that has the values 1111 for its first four CCS feature values – that is, features having to do with 'unmarked verbs' – is by implication assumed to be 4 evolutionary 'clock ticks' away from any language that has values 0000 for these features –values that are more typical, as we have seen, for a possible Indo-European language source. But in fact it is more reasonable to state that there is just a single difference between an Atlantic creole and likely I-E sources, namely, the difference between inflecting verbs or not: if verbs are uninflected, then the first four values of the CCS I-E features will typically be 0 and the creoles, value 1 (as shown in Figure 1 above). This seems more consonant with positing a single evolutionary step, where all feature values changed en bloc, though of course one cannot be certain of this without examining the historical record. If this is so, then the ascription of evolutionary distances in terms of feature value units is incorrect; no such historical inferences can be made. Further, phylogenetic 'clustering' using non-independent features will lead to over-weighting of some features (those counted multiple times) as opposed to others. For example, if the set of four tense aspect marker features correlate as a group, as seems likely, then this will have the tendency of 'counting' such features more heavily in analysis, which seems to be the case as we have seen in Figure 1. While a full feature independence analysis has not yet been carried out on all the datasets some of the features used in Bakker et al. [2011] are correlated in this way; for example, this work

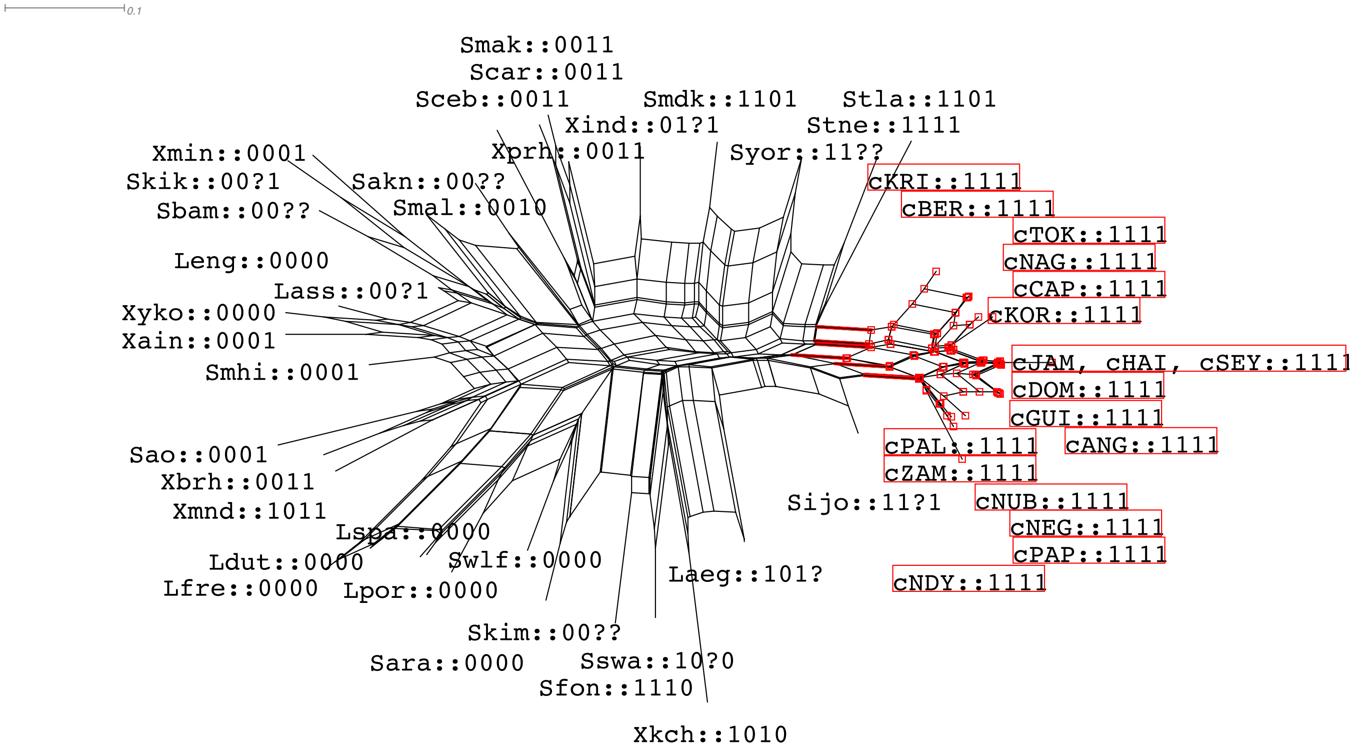
Second, D-M and Bakker correctly observe that using the CCS features poses a potential problem owing to the possibility of sample biasing (what is sometimes called *ascertainment bias*). This extends beyond the case of the CCS dataset. Each one of the three datasets examined in [1]-[3], the Hancock Atlantic creole features, the CCS features, and the Parkvall [2008] features, were developed in the first place as a way to describe creole languages. Therefore, from the outset these analyses focused on creoles themselves, isolating properties parochial to creoles, and so less likely to be found in non-creole languages. It is if one had already clustered languages into two groups: creoles and non-creoles. Any clustering analysis using these features is then more likely to simply to return a result confirming that these creole-derived features indeed distinguish creoles from non-creoles – that is, it simply confirms the initial partitioning that drove the feature selection process in the first place. From this perspective, nearly all of the results in [1]-[3] (Bakker et al. 2011, and the majority of results in D-M and Bakker 2012) are unsurprising, amounting to training a classifier on a dataset and then testing the classifier on the same training data. Near the end of the D-M & Bakker [2012] paper, this problem is acknowledged: "the validity of these results is somewhat undermined by the fact that the data which allowed the authors [e.g., Bakker et al. 2011] to reach [their conclusions] were specifically selected on the basis of creole properties" (D-M & Bakker, 2012:94).

To overcome this difficulty, D-M & Bakker carry out two analyses that they suggest resolve this issue, one using the CCS features, and one using only WALS features. However, neither of these proposed solutions actually resolve this serious problem.

Consider first the analysis using CCS features. Here, D-M & Bakker [2012] recode the 97 CCS features into 18 binary features "by selecting for each category [out of the 20 CCS categories] the feature(s) that were shared by most creoles" (Ibid, 94). For example, the 5 tense-aspect marker features (features 1.1-1.4) were marked as '1' if a language used some means other than inflection to mark tense, e.g., a typically creole value, and 0 otherwise. They then carried out a neighbor-joining analysis (rather than a Splitstree analysis) on 52 languages, grouping together the 18 creoles and 7 lexifier languages as described earlier, along with 19 substrate languages and 8 other non-creole languages (their published Figure 6 incorrectly cites 50 languages), to determine whether creoles cluster together, apart from non-creoles.

Using these lumped categories, they then carry out a neighbor-joining phylogenetic analysis, as opposed to a SplitsTree analysis. Neighbor-joining amounts to a clustering algorithm based on minimizing a 'least squares' distance, where the distances are computed using the feature values – with binary features, each difference amounts to a unit distance of 1, so that two languages that differ in two feature values are 2 units apart; while languages that differ in 3 features values are 3 units apart, and so forth. They find that just as with the finer-grained CCS feature analysis, creoles cluster apart from non-creoles. It is unclear however just how this approach resolves the bias problem. The features are still drawn from the CCS categorization, albeit lumped; so they are inherently creole-biased, which is also enhanced by "selecting for each category features shared by most creoles," as opposed to, say, selecting a feature for each category features shared by most lexifier or substrate languages. That is, it would have been more statistically valid to select features from languages in the WALS database without regard to their status as creoles, rather than starting from features represented in the CCS data. To be sure, this approach runs aground on the fact that for many WALS features, corresponding creole data might not be available. But that simply means that there is more data collection to undertake, not that one should abandon random sampling. Further, while the neighbor-joining tree analysis obtains the (unsurprising) result that creoles group together. In this regard it is instructive to note that a SplitsTree analysis, shown below in Figure 3, once again reveals that just a few features – here the 'lumped' features 1, 2, and 3, 4, corresponding roughly to CCS categories 1, 2, and 3 (unmarked verb, anterior past tense, progressive aspect, habitual marker, and completive aspect – that is, the verbal system as a group), perfectly split the creoles from the non-creoles. The simpler explanation here is that this is once again a reflection of the CCS's selection of features to begin with.

Figure 3. 18 creoles, 7 lexifiers, 19 substrate, 8 non-creoles, 18 binary features (52 languages total), replication of D-M & Bakker (2012) neighbor-joining analysis but by SplitsTree, showing creole vs. non-creole split on features 1 through 4.

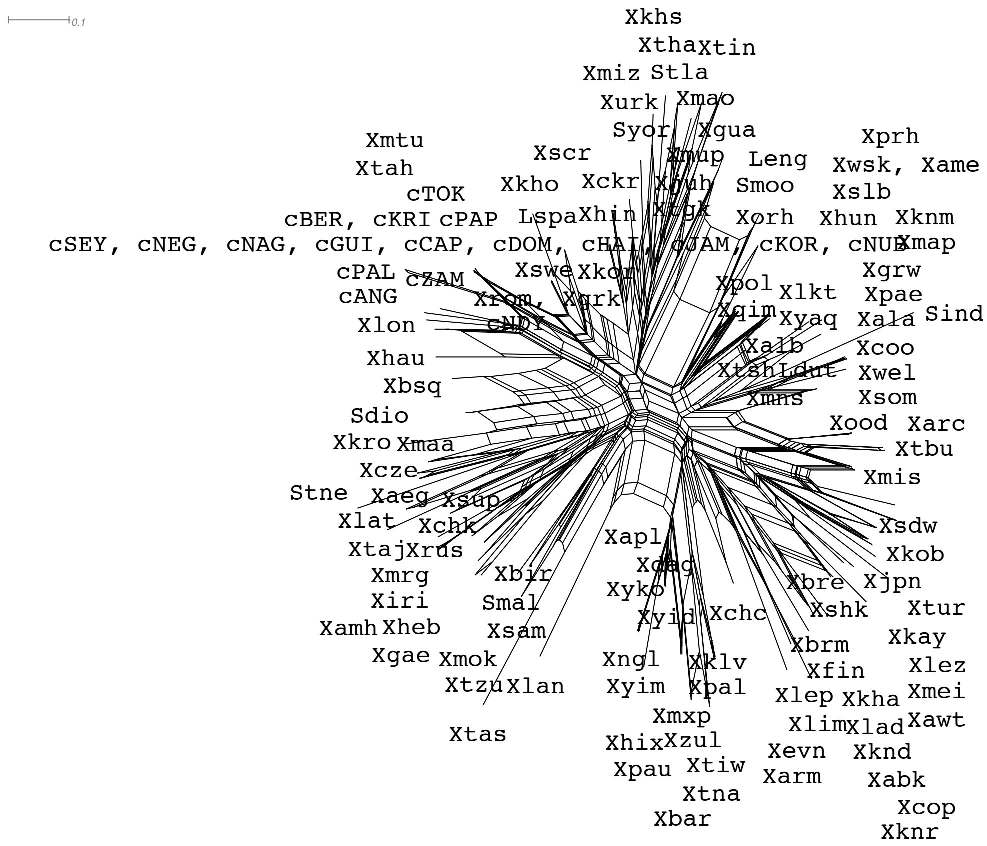


In addition, there is one other aspect of this analysis that further deserves comment since it diminishes the credibility of the neighbor-joining result. Note that there is a significant amount of cross-linked webbing in the SplitsTree result. As mentioned earlier, that is indicative of noise in the phylogenetic feature 'signal.' One can partially confirm that the signal is indeed noisy by attempting to run a phylogenetic analysis that is not 'greedy' like neighbor joining. For example, instead of assuming that the features denote (binary valued) distance values, one can assume that features are either simply present or absent, and carry out a so-called *parsimony analysis* that attempts to minimize the total number of feature changes from the root of a tree down to the languages at its leaves. In this case, standard parsimony analysis does not converge. Taken together with the SplitsTree display above, this suggests – though only tentatively – that using only 18 features to group 52 languages is not warranted, particularly if some of the binary features are still correlated (as the verbal features in sections 1–4 might well be). In any case, this does not remove a pro-creole bias.

D-M and Bakker's remaining three analyses designed to eliminate pro-creole feature bias rely on using multi-coded WALS data, which they assert will "settle the matter" (2012, 94). Specifically, the aim is to select a non-creole biased sample of features in the WALS database – which were presumably *not* specifically selected with the aim of classifying creoles, but rather, human languages generally. While this is indeed the correct aim of unbiased sampling, a truly unbiased feature sample must be selected randomly that is, without any prior knowledge of the state of creole language features used to select features. However, that is not how D-M & Bakker proceeded. Rather, for each of the three analyses, they selected 9 WALS features "shared by at least 60% of the CCS languages"(Ibid, 95), (yielding 18 creoles and 43 non-creoles that had values for all of these features); a reduction of these to 6 features (yielding 18 creoles and 58 non-creoles); and third and finally, a reduction of these 6 features to just 4, "shared by at least 80% of the CCS creoles" (yielding the same 18 creoles and 116 non-creoles). Taking this last example as illustrative of their best case, they note (see their Figure 9) that WALS features 38A (indefinite articles, 5 possible values); 69A (position of tense-aspect affixes, 5 possible values); 112A (negative morphemes, 6 possible values); and 117A (predicative possession, 5 possible values) strongly separate creoles from non-creoles.

However, there are again two problems, again one regarding bias, and one regarding the applicability of their neighbor-joining method. First, once again it may be simply not possible to carry out the right analysis given the data limitations to select WALS features randomly, because there is a large chance that many creoles will fail to be coded for that value, even if one restricts oneself to features that are present in, say, 80% of the WALS languages. Second, since the WALS features are multivalued, then it is not certain that the data values constitute proper distance metric, which is what neighbor-joining (and many other phylogenetic methods) rely on. For example, consider feature 38A, which can take on values 1 (for an indefinite word distinct from 'one'), 2 (indefinite word same as 'one'); 3 (indefinite affix); 4 (no indefinite but definite article), and 5 (for no definite or indefinite article). One can now ask: is a language that has no indefinite but definite article (value 4, as in Scots Gaelic) twice as far away from a language that has an indefinite word the same as 'one' (value 2, as in all 18 creoles in the sample), as opposed to four times as far away from a language that has an indefinite word distinct from 'one' (value 1, as in Dutch or English for D-M&Bakker's 2012 data)? Such a computation does not in fact make sense. Finally, the data are exceptionally conflicting with respect to any phylogenetic 'signal,' as might be expected with only 4 features and 134 languages. The problem is that with so many languages, there is a high probability that more than one language will have exactly the same feature values, and this is true in many places for this data: e.g., five langauges, Albanian, Dutch, English, Lakota, and Yaqui (Mexican) all share the same feature values, 1,1,2,5. It is therefore impossible for any phylogenetic program to distinguish them; in such a case, one says that the data fails to resolve the phylogeny. Given the 'overloading' of just a few features, it is not surprising that a parsimony analysis fails to converge with such a dataset, and a SplitsTree analysis is extremely noisy, as shown below in Figure 4. (The general rule of thumb is that *n* features can accurately resolve *log n* distinct species, so for *n*=4, 2^4=16, and resolutions limits have already been exceeded; see Felsenstein, 2004 for additional discussion on this point.) More important than all of these caveats, however, is simply that by selecting 4 features that are present in 80% of the creoles in the CCS, one has again biased sampling in favor of creoles; in short, the methods advanced by D-M and Bakker [2012] do not solve this bias problem. The only proper solution would be to work in the other direction: to start with features found on, say, 80 or 90% of the WALS languages, and then, iteratively, for those features carry out the research needed to determine the values of those features for the CCS creoles. When that process has been taken as far as it can go, one could then sample randomly from a single feature pool, in an unbiased fashion.

Figure 4. SplitsTree analysis with 4 WALS features, 134 languages, using D-M & Bakker 2012 data



References

[1] Bakker et al. 2011. Creoles are typologically distinct.

[2] Daval-Markussen, Aymeric. 2012. M.A. thesis.

[3] Daval-Markussen, A. & Bakker, P. 2012. Explorations in creole research with phylogenetic tools.

Holms & Patrick

[5] Niyogi, P. & Berwick, R. 2009. The proper treatment of language acquisition and change in a population setting. *PNAS USA*.