nature human behaviour

LETTERS

DOI: 10.1038/s41562-017-0192-4

Grammars are robustly transmitted even during the emergence of creole languages

Damián E. Blasi[1,2*], Susanne Maria Michaelis[2,3] and Martin Haspelmath[2,3]

Most languages of the world are taken to result from a combination of a vertical transmission process from older to younger generations of speakers or signers and (mostly) gradual changes that accumulate over time. In contrast, creole languages emerge within a few generations out of highly multilingual societies in situations where no common first language is available for communication (as, for instance, in plantations related to the Atlantic slave trade). Strikingly, creoles share a number of linguistic features (the 'creole profile'), which is at odds with the striking linguistic diversity displayed by non-creole languages[1–4]. These common features have been explained as reflecting a hardwired default state of the possible grammars that can be learned by humans[1], as straightforward solutions to cope with the pressure for efficient and successful communication[5] or as the byproduct of an impoverished transmission process[6]. Despite their differences, these proposals agree that creoles emerge from a very limited and basic communication system (a pidgin) that only later in time develops the characteristics of a natural language, potentially by innovating linguistic structure. Here we analyse 48 creole languages and 111 non-creole languages from all continents and conclude that the similarities (and differences) between creoles can be explained by genealogical and contact processes[7,8], as with non-creole languages, with the difference that creoles have more than one language in their ancestry. While a creole profile can be detected statistically, this stems from an over-representation of Western European and West African languages in their context of emergence. Our findings call into question the existence of a pidgin stage in creole development and of creole-specific innovations. In general, given their extreme conditions of emergence, they lend support to the idea that language learning and transmission are remarkably resilient processes.

The past 50 years of research on the languages of the world have revealed an impressive breadth of linguistic structures. A host of new linguistic features—such as the labiodental flap or object–subject–verb basic word order—have been described and exceptions to many patterns previously thought to be universal have been found[9]. Concurrently, a better grasp of linguistic diversity has brought a more precise understanding of the distributions and diachronic development of the over 7,000 extant languages. Everything else being equal, some linguistic features (or associations between linguistic features) are considerably more frequent than others, such as the overwhelming preference for languages with verb–object order to possess prepositions[10] or the bias towards certain sound–meaning associations as reflected in the vocabulary[11].

In this scenario of broad linguistic diversity with salient statistical tendencies, language structures are not randomly distributed, but they form more or less coherent groups. The most important source of similarity between languages is sharing a common ancestor: for instance, Persian, English, Russian, Hindi and Albanian share some features because they all descend from a language that was spoken between 5,000 and 9,000 years ago. Additionally, areal contact usually leads to similarities, as is the case with the languages spoken in Mesoamerica and South East Asia. In addition to these historical and areal factors, otherwise unrelated languages may resemble each other due to shared pressures acting on them. For instance, languages with larger populations tend to have simpler morphology, presumably due to the larger number of non-native speakers[12,13], and languages spoken in dry and cold areas are unlikely to develop or maintain tonal systems, which require a precise pitch production hindered by the effects of the environment on the larynx[14,15].

In general, the coincidence of specific linguistic features with extra-linguistic factors (such as shared history or area, demography or ecology) has served as both an empirical test and a discovery procedure for the forces that shape the distribution of language structures.

There is a peculiar set of languages scattered over all continents that originated under conditions which differ substantially from the regular processes of language transmission and learning: the creole languages. Normally, new generations acquire their first words and grammatical patterns at a young age from older generations, who acquired very similar words and grammars from the generations preceding them. In contrast, creole languages emerge within the course of a few generations—so that they may differ considerably from one generation to another—partially as a complex mixture of (potentially very diverse) languages, some of which may also have been learned natively at the same time. Crucially, a large number of studies have suggested that in parallel to these common sociocultural settings creole languages share structural properties as well.

Creole languages have emerged as the result of intense language contact situations, prototypically (but not exclusively) as the outcome of multilingual and multiethnic plantation societies following European colonial expansion since the 16th century, when slaves or indentured labourers indigenous to Africa, Asia or the Pacific worked for European colonists. The ancestry of creole languages has multiple sources and is traditionally divided between lexifiers and substrates. The lexifiers are the languages that contributed the bulk of their words and they are usually colonial dialectal varieties of Western European languages, such as Portuguese, French and English. The substrate languages are those that were spoken by the slaves or labourers and carried over—often from Africa and the Pacific—to the new overseas settlements. However, the ancestry of creoles can be more diverse, with instances of Arabic and Malay acting as lexifiers,

[1]Department of Comparative Linguistics, University of Zürich, Plattenstrasse 54, 8032 Zürich, Switzerland. [2]Department of Linguistic and Cultural Evolution, Max Planck Institute for the Science of Human History, Kahlaische Str. 10, 07745 Jena, Germany. [3]Department of English Studies, Leipzig University, Nikolaistrasse 8-10, 04109 Leipzig, Germany. *e-mail: damian.blasi@uzh.ch

and substrates originating from Australian Aboriginal languages or Austronesian languages from the South Pacific.

It has been claimed that creole languages share similar or identical linguistic features that distinguish them from non-creole languages: Daval-Markussen[16] reviewed over two dozen such features, including the presence of multiple negation and subject–verb–object (SVO) word order and the absence of relative pronouns, tones and gender systems. In contrast, others (for example, ref. [17]) argue that most studied creoles have very similar ancestry (mostly Western European and Macro-Sudan languages), which could explain the shared features[8]. The specific role of ancestry in creole genesis, however, is debatable.

The relexification hypothesis suggests that many creole languages have substrate grammars coated with a vocabulary from the lexifier[18,19]. Speakers of substrate languages had very limited access to the lexifier and, as a consequence, adopted the words of the lexifier but kept the original grammars. Other less radical substrate models highlight the role of transfer in second language use of functional and grammatical categories of the dominant (substrate) languages spoken by the slaves or indentured labourers into the nascent creoles[20].

Some have argued that creoles are regular offspring of their lexifiers with some influence from their substrates, similar to the Romance languages in relation to Latin[21]. One such theory places weight on the disproportionate influence of the original population of speakers of the lexifier and substrate languages, which is referred to as the 'founder principle'[7,22]. Due to largely uninterrupted language transmission in the early phases of colonial societies, the non-standard varieties of European languages of the first settlers would have been learned by the first groups of slaves and thus preserved as the basis for the nascent creoles. The success of the plantations increased the need for manpower, and soon the Europeans (and the first slaves) were outnumbered. This complicated access to the colonial language varieties for newcomers and thus successive generations shaped the language according to the linguistic features they brought in from the substrates, as the founder population had shaped the bulk of the grammar. While the special circumstances of creoles might have accelerated the pace of language change, the general processes they undergo are common to a large number of (non-creole) languages.

In opposition to these ideas, some researchers hold that the features that distinguish creoles from non-creoles are not derived from their genealogy directly, but are instead the product of the extreme sociolinguistic conditions underlying their emergence—collectively, these proposals can be referred to as the creole profile hypothesis. The creole profile hypothesis postulates the existence of a transmission bottleneck, resulting in pidgins as precursor stages of the new languages. Pidgins are restricted codes of communication with an extremely limited lexicon and grammar that do not have any native speakers. As such, they represent a radical simplification of the ancestral languages.

Several competing theories have attempted to describe and explain the mechanisms that produce such profiles out of pidgins, focusing on different aspects of the transmission, acquisition process and/or cultural setup of creole emergence. Bickerton[1] suggested that pidgins, when passed on to newer generations, are enriched with the full expressive machinery of any other natural language by means of a genetic blueprint characteristic of our species, thus giving rise to the observed commonalities. The creole profile might correspond to some kind of 'default' configuration for languages, a presumed window into the dawn of language.

Other researchers have proposed that the creole profile might arise from pidgins as transparent, economical or optimal solutions to achieve efficient and successful communication[2–5,16]. The transmission bottleneck thus appears as an opportunity for languages to make do without unnecessary or irregular material.

Finally, others have pointed to the very nature of the transmission bottleneck as being responsible for the creole profile. Everything else being equal, a given aspect of a language is considered more complex than the counterpart of another if it involves more distinctions in a paradigm (for example, French uses different pronouns depending on whether the referent is singular or plural, whereas Burmese does not) or more elements (for example, Arabic has eleven basic colour terms, whereas Murrinh-Patha has three). Good[6,23] argues that the transmission bottleneck will lead to a paradigmatic rather than syntagmatic reduction in complexity: while the successful transmission of a paradigm—such as the different pronouns of a language or a tonal system—involves the individual transmission of each of its members or generating rules (which translates into a prolonged exposure to the ancestry), lexical items—or even some constructions—can, in principle, be passed on in a single instance.

However, these ideas and proposals on the existence and nature of the creole profile are limited in a number of ways.

First, most creoles emerged within a very short amount of time—a few generations—and their early stages are generally poorly documented. In practice this entails that, apart from a few exceptions[24], the properties and distribution of the ancestral languages that gave rise to a creole are not always well known, and when they are known, we often do not have access to appropriate descriptions of those language varieties. Slaves and indentured labourers involved in the creation of creoles had diverse geographical and linguistic origins: only in the West Indies could one find speakers of East Asian, Pacific, Western European and Native American languages. As for the European languages involved, the varieties that contributed to the genesis of creole languages were not standard varieties but non-standard dialects that sometimes diverged in important ways. This has led to what is referred to as the 'cafeteria principle'[25]: those who support the notion that creoles inherit their structure from their ancestry could, in principle, find a language or a specific language variety (among the many substrates and European non-standard ones involved) fitting the feature under discussion. This complex picture of the ancestry of creoles (and the diversity among those languages) calls into question evaluations of the creole profile based on identifying a single lexifier or substrate for each creole[26].

Second, researchers have used diverse and not necessarily comparable data to sustain their claims, which has led to the puzzling situation that, while many authors of contrasting theoretical stances agree on the existence of a distinctive creole profile, they base their arguments on non-overlapping sets of linguistic features and languages[16]. Importantly, the large majority of these comparisons involved only a relatively small number of languages and almost exclusively creoles coming from an extremely narrow set of ancestors, namely West African substrates and Western European lexifiers.

Third, given the many possible dimensions of linguistic description—as a reference, the *World Atlas of Language Structures* (WALS; ref. [27]) comprises 142 features—there has not been an attempt to evaluate statistically the likelihood of any arbitrary set of languages sharing a number of features by chance given a fixed number to choose from. Without this assessment, and given the fact that some of the claims rely on as few as three distinctive features, it is impossible to evaluate how strong the case is from a statistical point of view.

Fourth, considerable attention has been paid to the alleged lack of inflectional marking as one ingredient feature of the typological profile of creoles, which presumably makes them a distinct class of languages synchronically[2,4,16,28]. Critically, such claims rely on the notion of word as an orthographic unit in deciding whether a morpheme is a separate word or an affix[16,29,30], but conventions on orthographic words do not entail any necessary morphosyntactic

or phonological properties[31,32]. Therefore, any meaningful comparison between creoles and non-creoles should take these reservations into account.

All of these circumstances compromise the validity of general claims on the creole profile. Here, we make explicit the minimum requirements for such a test to be empirically sound and we formalize the proposals of the creole profile in a more statistically explicit manner.

First, it is crucial to distinguish two arguments in the creole profile hypothesis: (1) there are structural commonalities that distinguish creole languages from the rest of the languages in the world and (2) those commonalities are not due to regular genealogical or contact transmission from their ancestral languages. It should be noted that most language families and areas naturally satisfy the first requirement. Most European languages, for instance, have a number of features that are cross-linguistically rare—such as the use of case-bearing pronouns to mark subject relative clauses[33]—but their widespread presence within the continent does not require an explanation beyond regular language transmission.

As for the empirical test of the creole profile hypothesis, creole data should be as extensive as possible and readily comparable with non-creole languages. Ideally, the features should reflect variables of wide typological interest to avoid or reduce the bias of features being pre-selected due to their perceived similarity across creoles[34]. Given the goal of detecting creole-wide properties regardless of their ancestry, features for which a case for a genealogical origin can be made should be removed from the evaluation.

Regarding the actual model of the creole profile, it is possible to distinguish two instantiations in the literature: the rule-based profile and the probabilistic profile. The rule-based profile consists of a fixed template of a few feature values (for example, "no tense-aspect inflection + indefinite article equal to numeral one + negation expressed through particle + predicative possession expressed with a have verb"[16]) that is associated with all or most creoles but is unattested or vanishingly rare in non-creole languages. The probabilistic profile is instead expressed as non-deterministic biases of creoles towards certain feature values (in contrast with non-creoles), usually involving a large number of features. Under this view, creoles stand out from the other languages as having general tendencies towards certain properties (for example, they tend to display a smaller number of distinctions in multiple domains, such as morphology and phonology[2,3,6,23]).

Considering these requirements, we used the largest published dataset on the structure of creole languages, the *Atlas of Pidgin and Creole Language Structures* (APiCS; refs [35,36]) for the present study. This database shares a number of typological features that can be compared with another large database of languages, the WALS[27]. This set of common features comprises mostly properties of nominal and verbal phrases, word order patterns and clausal syntax.

As mentioned above, in most cases, comparing creoles with their direct lexifiers or substrates is difficult or even impossible. However, for each creole it is possible to determine the broad sources of its ancestry. The largest ancestry groups contributing to the creoles in our sample were the Romance and Germanic subfamilies (for the lexifiers) and the Macro-Sudan belt (within which many grammatical features are shared across several different language families[37]) and the Austronesian family (for the substrates). As an example, while looking at the relative order of the possessor (Gen) with respect to its possessed noun (N) in relation to the lexifier genealogy (Table 1), we found a clear tendency: creoles lexified by Romance languages are prone to exhibit NGen order, whereas those lexified by Germanic languages display GenN or no dominance of a particular order (which in this case coincides with the characteristics of the lexifier languages as well).

Table 1 | Number of creole languages for the order of possessor (Gen) and possessed noun (N) according to the broad ancestry group of their lexifier

	Germanic lexifier	Romance lexifier	Other
GenN	13	3	1
NGen	3	19	5
Both orders occur and none is dominant	8	1	0

Using this information, we evaluated whether each of the features in the APiCS is associated with substrate or lexifier ancestry groups (see Methods and Supplementary Table 4). A conservative independence test (which minimizes false positives) was used to find those features for which our data showed the clearest cases of dependency between creoles and their ancestry. Complementarily, an anti-conservative test (which minimizes false negatives) was performed for the sake of detecting features for which our data did not reveal any dependency between them.

Applying these evaluations, approximately half of all features were found to be associated with ancestry through the conservative test and a similar number of features were found not to be associated with ancestry in the anti-conservative test (see Supplementary Table 4).

Testing the creole profile requires a comparison group. We chose a balanced sample from the WALS that was intended to approximate the world's linguistic diversity. None of the languages of that sample had been shown to be creoles—although there had been arguments for Chamorro[38] and Hmong Njua[39]—and they all belonged to different linguistic genera (groups of languages that descend from a common ancestor of roughly comparable time depth[27]) and varied linguistic areas. Considering only the features and languages with enough coverage in both the APiCS and the WALS, we ended up with a set of 48 creole languages (with less than 2% of missing data), a balanced sample of 111 non-creole languages (with an average of 14% of missing data) and 41 shared features (see Supplementary Fig. 1 and Supplementary Materials for details).

The rule-based profile was evaluated using an efficient data mining algorithm[40]. Concretely, for rules comprising one to four features, we mined the best rules according to their classification efficiency by means of F scores, which combined precision (that is, the fraction of times that a rule was associated with a creole rather than non-creole languages) and recall (that is, the fraction of creoles that complied with the rule) in a single measure. This method required the imputation of missing data (see Methods).

To determine whether these results were statistically distinguishable from finding features that separated any two arbitrary sets of languages in our data, we compared the output of the previous algorithm against a baseline resulting from permuting the creole versus non-creole labels while keeping the feature values constant. Informally, this corresponded to dividing the data into two arbitrary groups of the same size as the creole and the non-creole groups in the original data and running the same analysis as before. We repeated this procedure 50 times for each of 30 imputations of the original data.

In addition to testing the creole profile with the full feature set ('full dataset') for both the rule-based and probabilistic profile, we repeated the analyses using only those features for which no ancestry dependency could be established through the anti-conservative independence test ('reduced dataset').

With these specifications, the rule-based profile model was able to discriminate efficiently between creoles and non-creoles with both the full and reduced dataset, with only a slight decrease

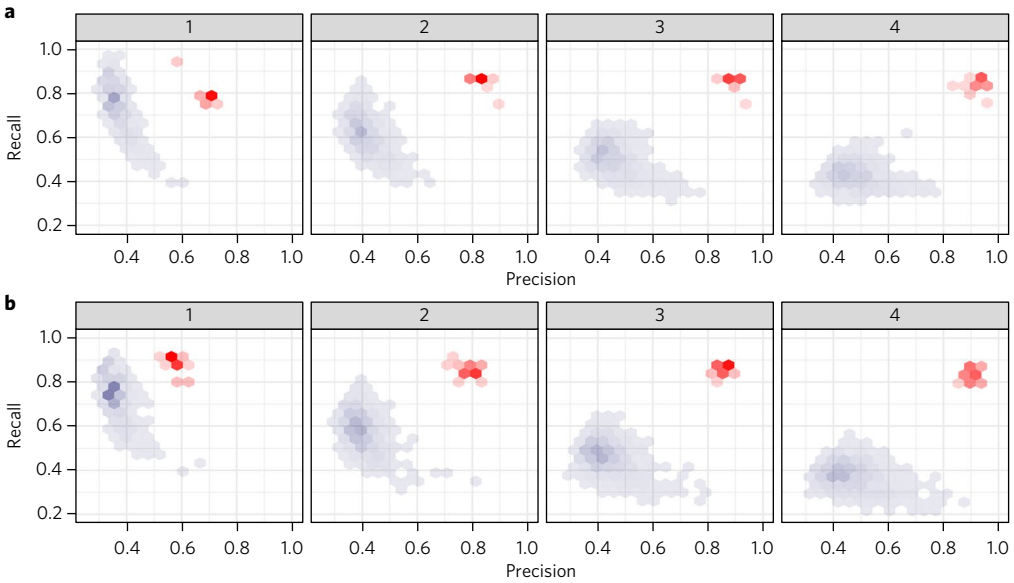

Fig. 1 | Classification under the rule-based creole profile for the full and reduced datasets, with rules chosen by best F_1 score. a, Full. **b**, Reduced. Distributions of the empirical (red) and randomized (blue) precision and recall for rule lengths involving between one and four variables. For rules of a length larger than 1, the majority of the rules obtained in the empirical data had better precision and recall values than the randomized controls, which supports the notion that creoles can be distinguished from non-creole languages.

of classification quality in the reduced dataset (see Fig. 1 and Supplementary Table 3).

A closer inspection of the best rules inferred in the rule-based profile revealed some coincidence with previous claims (such as the SVO word order, expression of negation through a negative particle and predicate possession through a 'have' verb), but in general they constituted fairly common typological properties (see Table 2).

The probabilistic creole profile was approached through a standard machine-learning technique based on ensembles of conditional inference trees (see Methods). Conditional inference trees begin by finding the feature value that is better associated with creole languages, such that languages that have it are more likely to be creoles than those that do not (in a statistically significant way). Then, the data are partitioned in two halves (languages that have the feature value and languages that lack it) and each half is analysed in the same way until the algorithm does not find any variable that is associated with creole languages.

The same baseline comparison as with the rule-based profile was applied: the labels were randomized 500 times and the results compared.

As expected—given the flexibility of the probabilistic creole profile—the results revealed an even sharper discrimination between creoles and non-creoles (see Table 3), with randomizations of the dataset mostly producing no relevant features associated with the creole versus non-creole distinction, leading the algorithm to adopt the simplest classification strategy (classification of all languages as non-creoles, since this is the most frequent group).

Detecting dependency between feature values and ancestry relies on the existence of variation structured according to the lexifier and substrate groups, which implies that features without variation or with considerable variation within the groups yield false negatives. For instance, the overwhelming majority of lexifiers relevant to our sample of creoles were prepositional languages, so even if this characteristic were faithfully transmitted from lexifiers to creoles, our test would not have been able to detect a statistical association between the ancestry groups and the corresponding creoles.

Heuristically, we found this to be true for most of the features in the reduced dataset (see Supplementary Table 6 for a summary of the distribution of these features). This opened the door to the

Table 2 | Best rule-based creole profile rules according to F_1 involving one to four variables for both full and reduced sets of features

	Full	Reduced
One variable	Obligatory pronoun in subject position (0.6)	Have-possessive (0.67)
Two variables	Negative particle	Prepositions
	No applicative constructions (0.37)	Have-possessive (0.9)
Three variables	SVO order	Have-possessive
	Have-possessive	Negative particle
	Negative particle (0.5)	No applicative constructions (0.57)
Four variables	SVO order	Noun-relative order
	Have-possessive	Have-possessive
	Negative particle	Negative particle
	No applicative constructions (0.7)	No applicative constructions (0.63)

The numbers in brackets indicate the fractions of imputed datasets where the rule was inferred.

Table 3 | Precision and recall values for the probabilistic creole profile model for both datasets (full or reduced), divided between the estimates for the actual data and the mean and s.d. obtained in the permuted datasets (rounded to the nearest tenth)

Dataset	Values	Precision	Recall
Full	Empirical	0.96	0.92
	Permutations	0 ± 0.06	0 ± 0.01
Reduced	Empirical	0.76	0.85
	Permutations	0.01 ± 0.08	0 ± 0.02

possibility that even the features we deemed independent were carrying an ancestry signal as well. This seemed to be coherent with some of the misclassified cases: some of the languages that were incorrectly identified as creoles were either among the set of ancestral languages or were similar to them (such as English or Yoruba), whereas some of the creoles with typological properties different from the majority of lexifiers and substrates (such as Angolar and Tayo) were classified as non-creoles (see Supplementary Table 5). In addition, the features of the reduced dataset did not constitute a consistent bundle in terms of any evident function or linguistic nature (see Supplementary Table 4), which could have warranted a genuine explanation in terms of creolization. Since the ancestry of creole languages usually comes from a few groups, distinguishing them from a balanced sample of non-creole languages is as unsurprising as the fact that a set of languages related through genealogy or area could be distinguished from other unrelated or less related languages.

Even under the limitations of our method for the assessment of ancestry dependency, some generalizations could be made about the regular ancestral source of some of the creole features (see Supplementary Table 3). Word order in creoles overwhelmingly patterns according to the lexifier[20,41]. It is possible to detect this as Romance and Germanic languages differ in some of these, such as the order of the genitive and adjective in relation to the noun. When there is no variation, creoles overwhelmingly follow the dominant pattern in the lexifiers (see Supplementary Table 6). Knowing the lexifier group improves the classification accuracy between 7 and 20% (in relation to the baseline of choosing the most frequent order).

Substrate ancestry was seen to be related to a number of diverse features in our sample, most prominently the verbal domain, such as tense and aspect marking and argument marking (see Supplementary Table 3). For a number of features, a dependency with both lexifiers and substrates was detected, which might suggest that (at least in some domains) creoles can continue the linguistic structure from any of their ancestral languages. Importantly, the number of features that turned out to be associated with either side of the ancestry cannot be taken as proportional to their importance for the process of creole formation, given that the limitations in our tests might have been different for each group. This constraint might explain why other associations that are well-known in the literature (such as that of pronominal systems and substrate languages) did not emerge patently.

The strongest case for a truly innovated creole feature would come from a feature value that is homogenous across creoles but different across the board in their ancestry, but no feature satisfies this scenario. This does not necessarily preclude the existence of other features that are indeed the result of creolization, perhaps in areas not well covered by our data such as morphology[2,16]. However, our findings indicate that for the overwhelming majority of features analysed there is always a way of showing that they vary with ancestry, that both ancestry and creoles overwhelmingly share one particular feature value (for example, prepositions) or that the ancestral groups are too diverse for our test to yield any definitive conclusion. Thus, the majority of creole grammars have been transmitted—as in any other natural language—from their ancestry, either from their lexifiers and substrates or through later contact (since some of the creoles have coexisted with some of their ancestral languages; for example, Korlai[42]).

Complementarily, our analyses did not rule out the possibility that in specific cases creoles could develop a new structure away from their ancestry (as could happen in any other language), although this does not necessarily reflect any creole-specific tendency. This is particularly important when one or a few innovated instances are put in relation to patterns that are otherwise well represented in the ancestry. The best-known case is the basic word order SVO (which is overwhelmingly present in creole lexifiers), which it has been argued emerged in Berbice Dutch in spite of its OV ancestral languages[43]. Since the transmission-from-ancestry account is sufficient to explain SVO in almost all attested creoles, a single well-established exceptional case (which incidentally might not be the case of Berbice Dutch[44]) provides no immediate support for the creoleness of the feature. Hence, claims that there is a special association between SVO and creole languages[45–47] are unsupported.

These results call into question the idea of a transmission bottleneck and a pidgin phase in the history of the development of creoles that explains commonalities across creole languages. If, as we have shown, a substantial number of features are passed along from the ancestry to the creoles, positing an intermediate pidgin stage that would have considerably reduced and simplified the ancestry features does not seem to be plausible, since it would remain to be explained why creoles faithfully continue grammatical patterns—such as word orders, ditransitive constructions, subject relative clauses and indefinite pronoun patterns—from their ancestral languages.

In general, why such a complex human behaviour can be successfully transmitted even in the typical (intricate and multilingual) contact situations of creoles is still unclear. The remarkable efficiency of the human species to learn and transmit language has been explained in a number of ways, such as the progressive adaptation of language to the human brain[48], an innate biological machinery specialized for learning language[49], sophisticated statistical and social learning strategies[50] and the very nature of cultural evolution in which language is embedded[51], among others. Either way, our results reflect the astonishing resilience of language transmission.

Methods

Data. Data on creole languages were extracted from the APiCS Online database[35], which contains information on 76 contact languages (pidgins, creoles and mixed languages) and 130 structural features.

For non-creole languages, we used the WALS[27]. There are 48 shared features between the APiCS and the WALS, but some APiCS features can display, instead of a single feature value, a probability distribution over several values (for example, in Seychelles Creole, adjectives can precede or follow nouns with comparable frequency). To permit a comparison between the two databases, we mapped (for each creole) these probability distributions to their most frequent value whenever its probability equalled or exceeded two-thirds. Otherwise the value was marked as 'mixed'.

A few languages from the APiCS database (Guadeloupean Creole, Batavia Creole, Cavite Chabacano, Cape Verdean Creole of Brava, San Andres Creole English, Cape Verdean Creole of Sao Vicente, Kikongo-Kituba, Zamboanga Chabacano, Mauritian Creole and Santome) were removed as there was a closely related variety or dialect already present, to avoid over-counting essentially the same language. After removing both languages and features with low coverage, we ended up with a set with 48 creole languages and a balanced sample of 111 non-creole languages (see Supplementary Tables 1 and 2). Of the 48 shared features between the APiCS and the WALS, 41 had more than a 75% coverage in our sample and thus were retained for analysis. About 10% of the relevant feature values were missing in the aggregated data. Most of these come from the WALS set (which exhibited 14% of missing entries), whereas creoles from the APiCS had an almost perfect coverage, with less than 2% of missing data. Data imputation (required for the analysis of the rule-based creole profile) was performed through a non-parametric Bayesian model specialized for categorical data[52].

We chose the ancestry groups, in such a way that they (1) did not require strong assumptions about the specific ancestry of the language, (2) covered a sizeable number of creoles and (3) tended to have properties in common (due to genealogical or areal dependencies). The groups we chose were the Germanic and Romance subfamilies (for the lexifiers) and the Macro-Sudan area and the Austronesian family (for the substrates). To these, we added the group 'other' for the creoles for which their lexifiers or substrates did not belong to the previously defined groups. All the decisions on ancestry were based on the APiCS chapters on those languages[35].

Association tests. As association tests between ancestry groups and the values of the features under consideration we used simple Fisher's exact tests. First, we obtained the P values of the relevant contingency tables comparing ancestry groups with feature values (as in Table 1) by approximating the baseline distribution through $B = 10{,}000$ Monte Carlo simulations. In the anti-conservative test, we used a conventional level of $\alpha = 0.05$ for determining the statistical significance based on these P values. For the conservative test, to take into account the inflation of false positives due to multiple comparisons, we also considered the local false discovery rate of the previously obtained P values[53] and we used $\alpha = 0.05$ for the significance of these corrected values. The first tests were likely to overestimate the dependency between ancestry and feature values (due to the absence of a multiple comparisons correction), whereas the second tests not only controlled for that circumstance, but were known to be more conservative than the (tail) false discovery rate[53].

Classification metrics. Recall is the fraction of all instances of the target group that have been correctly classified—in this case, the fraction of creole languages that are classified correctly as such. Precision, complementarity is the fraction of all instances of the target group that have been correctly classified over the total number of instances classified in that group—the fraction of creoles correctly classified over the total number of languages classified as creoles. F scores combine precision and recall through the parameter β in such a way that the larger β is, the smaller the relative importance of precision is. As a reference, $\beta = 1$ is the case in which both precision and recall are equally weighted and $F_1 = (\text{precision} \times \text{recall})/(\text{precision} + \text{recall})$. Choosing an appropriate β depends on the problem at hand: for instance, when considering a classification method that detects a lethal disease, the consequences of poor recall (that is, not diagnosing someone who has the disease) are far more critical than the consequences of poor precision (that is, wrongly diagnosing someone as having the disease). For this reason, while we centre our analysis on F_1, we also provide the results for F scores that count precision as being more important ($\beta = 0.5$) or less important ($\beta = 2$) than recall (see Supplementary Figs. 2 and 3).

Conditional inference forests. The fundamental idea behind conditional random trees (which are the building blocks of conditional random forests[54]) is that feature values that better and significantly discriminate between creoles and non-creoles are used to divide the data into two in a recursive fashion, thus partitioning the languages into smaller subsets that are subsequently more homogeneously creoles or non-creoles. By the end of this process, each of the languages belongs to one of these subsets and its probability of being classified as a creole is simply the ratio of creoles over the total number of languages within that subset. This is repeated for a number of subsets of the languages and features, and the individual results—the conditional inference trees—are aggregated to produce a unique assignment of each possible combination of features (that is, each possible language) to either the creole or the non-creole group, which constitute the conditional inference forest.

The specific implementation of the conditional inference forest used here followed Strobl and colleagues[55] to guarantee that the number of levels a variable had did not influence its likelihood of being chosen; this also determined that about 60% of the data points (languages in our case) were sampled without replacement for every tree. We used 1,000 trees per forest. Apart from the main empirical classification, the comparison baseline was produced by randomizing the creole versus non-creole labels 500 times and evaluating the classificatory properties of random forests on each of the sets.

Code availability. The code that supports the findings of this study is available upon reasonable request from the corresponding author.

Data availability. The data that support the findings of this study are available upon reasonable request from the corresponding author.

Received: 21 December 2016; Accepted: 31 July 2017;
Published online: 04 September 2017

References

1. Bickerton, D. *Roots of Language* (Karoma, Ann Arbor, MI, 1981).
2. McWhorter, J. H. The world's simplest grammars are creole grammars. *Ling. Typol.* **5**, 125–166 (2001).
3. Parkvall, M. in *Language Complexity: Typology, Contact, Change* (eds Miestamo, M., Sinnemäki, K. & Karlsson, F.) 265–285 (Benjamins, Amsterdam, the Netherlands, 2008).
4. Bakker, P., Daval-Markussen, A., Parkvall, M. & Plag, I. Creoles are typologically distinct from non-creoles. *J. Pidgin Creole Lang.* **26**, 5–42 (2011).
5. Daval-Markussen, A. & Bakker, P. in *Cambridge Handbook of Linguistic Typology* 254–286 (Cambridge Univ. Press, Cambridge, 2017).
6. Good, J. Paradigmatic complexity in pidgins and creoles. *Word Struct.* **8**, 184–227 (2015).
7. Mufwene, S. S. The founder principle in creole genesis. *Diachronica* **13**, 83–134 (1996).
8. Aboh, E. O. & Ansaldo, U. in *Deconstructing Creole* (eds Ansaldo, U., Matthews, S. & Lim, L.) 39–66 (Benjamins, Amsterdam, the Netherlands, 2007).
9. Evans, N. & Levinson, S. C. The myth of language universals: language diversity and its importance for cognitive science. *Behav. Brain Sci.* **32**, 429–448 (2009).
10. Greenberg, J. H. Some universals of grammar with particular reference to the order of meaningful elements. *Univ. Lang.* **2**, 73–113 (1963).
11. Blasi, D. E., Wichmann, S., Hammarström, H., Stadler, P. & Christiansen, M. H. Sound-meaning association biases evidenced across thousands of languages. *Proc. Natl Acad. Sci. USA* **113**, 10818–10823 (2016).
12. Lupyan, G. & Dale, R. Language structure is partly determined by social structure. *PLoS ONE* **5**, e8559 (2010).
13. Bentz, C. & Winter, B. Languages with more second language learners tend to lose nominal case. *Lang. Dyn. Change* **3**, 1–27 (2013).
14. Everett, C., Blasi, D. E. & Roberts, S. G. Climate, vocal folds, and tonal languages: connecting the physiological and geographic dots. *Proc. Natl Acad. Sci. USA* **112**, 1322–1327 (2015).
15. Everett, C., Blasi, D. E. & Roberts, S. G. Language evolution and climate: the case of desiccation and tone. *J. Lang. Evol.* **1**, 33–46 (2016).
16. Daval-Markussen, A. First steps towards a typological profile of creoles. *Acta Linguistica Hafn.* **46**, 1–22 (2013).
17. Michaelis, S. M. World-wide comparative evidence for calquing of valency patterns in creoles. Preprint at https://zenodo.org/record/844616 (2017).
18. Lefebvre, C. *Creole Genesis and the Acquisition of Grammar* (Cambridge Univ. Press, Cambridge, UK, 1998).
19. Lefebvre, C. *Functional Categories in Three Atlantic Creoles: Saramaccan, Haitian and Papiamentu* (Benjamins, Amsterdam, the Netherlands, 2015).
20. Siegel, J. *The Emergence of Pidgin and Creole Languages* (Oxford Univ. Press, Oxford, UK, 2008).
21. Chaudenson, R. *Des Îles, des Hommes, des Langues* (L'Harmattan, Paris, France, 1992).
22. Mufwene, S. *The Ecology of Language Evolution* (Cambridge Univ. Press, Cambridge, UK, 2001).
23. Good, J. Typologizing grammatical complexities or why creoles may be paradigmatically simple but syntagmatically average. *J. Pidgin Creole Lang.* **27**, 1–47 (2012).
24. McWhorter, J. & Good, J. *A Grammar of Saramaccan Creole* Vol 56. (Walter de Gruyter, Berlin, Germany, 2012).
25. Dillard, J. L. Principles in the history of American English: paradox, virginity, cafeteria. *Florida FL Reporter* **7**, 32–33 (1970).
26. Murawaki, Y. *Statistical Modeling of Creole Genesis* in *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics* (2016).
27. Dryer, M. S. & Haspelmath, M. The World Atlas of Language Structures Online (2013); http://wals.info
28. McWhorter, J. H. *Defining Creole* (Oxford Univ. Press, Oxford, UK, 2005).
29. Dryer, M. S. in *The World Atlas of Language Structures* (eds Haspelmath, M., Dryer, M. S., Gil, D. & Comrie, B.) 282–285 (Oxford Univ. Press, Oxford, UK, 2005).
30. Siegel, J., Szmrecsanyi, B. & Kortmann, B. Measuring analyticity and syntheticity in creoles. *J. Pidgin Creole Lang.* **29**, 49–85 (2014).
31. Haspelmath, M. The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Ling.* **45**, 31–80 (2011).
32. Haspelmath, M. & Michaelis, S. M. in *Language Variation—European Perspectives VI: Selected Papers from the 8th International Conference on Language Variation in Europe (ICLaVE 8), Leipzig 2015* (eds Buchstaller, I. & Siebenhaar, B) 3–22 (Benjamins, Amsterdam, the Netherlands, 2017).
33. Haspelmath, M. in *Language Typology and Language Universals* 1492–1510 (De Gruyter Mouton, Berlin, Germany, 2001).
34. Aboh, E. O. Creole distinctiveness. *J. Pidgin Creole Lang.* **31**, 400–418 (2016).
35. Michaelis, S. M., Maurer, P., Haspelmath, M. & Huber, M. *The Atlas of Pidgin and Creole Language Structures Online* (Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany, 2013a); http://apics-online.info/
36. Michaelis, S. M., Maurer, P., Haspelmath, M. & Huber, M. *The Atlas of Pidgin and Creole Language Structures* (Oxford Univ. Press, Oxford, UK, 2013).
37. Güldemann, T. in *Language and Space: An International Handbook of Linguistic Variation* Vol. 2 (eds Lameli, A., Kehrein, R. & Rabanus, S.) 561–585 (De Gruyter Mouton, Berlin, Germany, 2010).

38. Salamanca, R. R. P. *Del Español al Chamorro: Lenguas en Contacto en el Pacífico* (ed. Gondo, E.) (Ediciones Gondo, Madrid, Spain, 2009).
39. Daval-Markussen, A. in *Workshop on Non-Indo-European Lexifier, Non-West African Pidgin and Creole Languages* 10–11 (Newcastle University, Newcastle, UK, 2010).
40. Agrawal, R. & Srikant, R. *Fast algorithms for mining association rules* in *Proceedings of the 20th International Conference on Very Large Data Bases* 487–499 (1994).
41. Velupillai, V. *Pidgins, Creoles and Mixed Languages: An Introduction*. (Benjamins, Amsterdam, the Netherlands, 2015).
42. Clements, J. C. *The Genesis of a Language: the Formation and Development of Korlai Portuguese* Vol 16 (Benjamins, Amsterdam, the Netherlands, 1996).
43. Kouwenberg, S. From OV to VO linguistic negotiation in the development of Berbice Dutch creole. *Lingua* **88**, 263–299 (1992).
44. Zeijlstra, H. & Goddard, D. On Berbice Dutch VO status. *Lang. Sci.* **60**, 120–132 (2017).
45. Bickerton, D. The language bioprogram hypothesis. *Behav. Brain Sci.* **7**, 173–188 (1984).
46. Hall, M. L., Mayberry, R. I. & Ferreira, V. S. Cognitive constraints on constituent order: evidence from elicited pantomime. *Cognition* **129**, 1–17 (2013).
47. Langus, A. & Nespor, M. in *Representing Structure in Phonology and Syntax* (eds van Oostendorp, M. & van Riemsdijk, H.) (De Gruyter Mouton, Berlin, Germany, 2015).
48. Christiansen, M. H. & Chater, N. Language as shaped by the brain. *Behav. Brain Sci.* **31**, 489–509 (2008).
49. Berwick, R. C., Pietroski, P., Yankama, B. & Chomsky, N. Poverty of the stimulus revisited. *Cogn. Sci.* **35**, 1207–1242 (2011).
50. Ambridge, B. & Lieven, E. V. *Child Language Acquisition: Contrasting Theoretical Approaches* (Cambridge Univ. Press, Cambridge, UK, 2011).
51. Smith, K. & Kirby, S. Cultural evolution: implications for understanding the human language faculty and its evolution. *Phil. Trans. R. Soc. Lond. B Biol. Sci.* **363**, 3591–3603 (2008).
52. Si, Y. & Reiter, J. P. Nonparametric bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys. *J. Edu. Behav. Stat.* **38**, 499–521 (2013).
53. Strimmer, K. A unified approach to false discovery rate estimation. *BMC Bioinformatics* **9**, 303 (2008).
54. Hothorn, T., Hornik, K. & Zeileis, A. Unbiased recursive partitioning: a conditional inference framework. *J. Comp. Graph. Stat.* **15**, 651–674 (2006).
55. Strobl, C., Boulesteix, A. L., Zeileis, A. & Hothorn, T. Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics* **8**, 1 (2007).

Acknowledgements

We thank J. Good, A. Daval-Markussen and P. Bakker for useful comments and discussions. The support of the European Research Council (ERC Advanced Grant 670985, Grammatical Universals) is acknowledged. No funders had any role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Authors contributions

All authors designed the research. D.E.B. designed and conducted the statistical analyses. S.M.M. and M.H. curated the data used for the analyses. D.E.B. and S.M.M. drafted the manuscript. All authors discussed the results and contributed to the final version of the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at doi:10.1038/s41562-017-0192-4.

Correspondence and requests for materials should be addressed to D.E.B.

nature research

Corresponding author(s): Damian E. Blasi

☐ Initial submission ☐ Revised version ☒ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

Describe how sample size was determined.

Sample sizes for this observational study were not determined beforehand

2. Data exclusions

Describe any data exclusions.

Some of the data in the databases employed in this study were excluded by criteria discussed in the manuscript

3. Replication

Describe whether the experimental findings were reliably reproduced.

This research is observational and based on statistical analyses that can be replicated with the code available from the corresponding author

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

Does not apply

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

The determination of creole/non-creole labels and the set of features used for the analyses occurred before the analysis of the results

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ A statement indicating how many times each experiment was replicated
☐	☒ The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☐	☒ A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒ The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
☐	☒ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒ Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.	We used several versions of R for this project - the latest version is R 3.3.2

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.	Materials are available both from online repositories and directly from the corresponding author

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	Does not apply

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	Does not apply
b. Describe the method of cell line authentication used.	Does not apply
c. Report whether the cell lines were tested for mycoplasma contamination.	Does not apply
d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	Does not apply

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.	Does not apply

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	Does not apply