The simplicity of creoles in a cross-linguistic perspective

Mikael Parkvall
Stockholms Universitet

This paper discusses the possibility of quantifying complexity in languages in general, and in creoles in particular. It argues that creoles are indeed different from non-creoles, primarily in being less complex. While this has been argued before, this is the first attempt to prove it through the use of an extensive typological database. It is noteworthy that the differing complexity is not related to the relative lack of morphology in creoles, since they are also simpler than analytical languages. Finally, the parallels between pidgins and creoles (and in particular the fact that languages sociologically intermediate between the two categories are also structurally intermediate) support the increasingly questioned belief that pidgins are born out of pidgins.

1. Introduction

A dogma in modern linguistics is that all languages are equally expressive. While they may have lexicalized different chunks of reality, they are thought to structurally possess the same expressive potential, and thus be equally suited to encode human experience. I do not intend to question this assumption. However, I do object to jumping to the conclusion that since all languages are equally expressive, they must by definition also be equally complex.

In various guises, such an assertion is found in e.g., Atchison (2001: 253), Crystal (1987: 6), Fromkin and Rodman (1974: 26), Hockett (1958), Smith (1999: 168) and countless other works.

In the following, I shall explore one possibility of ranking languages with regard to their structural complexity (or perhaps rather *one aspect* of complexity), and also address the issue of whether "the world's simplest grammars are creole grammars", as claimed by McWhorter (2001).[1]

---

1. In this paper, "complexity" is used in the same way as does McWhorter (2001). This means that an expression is more complex than another if it involves more rules, i.e., if it requires a lengthier description. No claims are made with regard to learning difficulties or other psycholinguistic aspects.

While linguists shun from claiming that a given language is more or less complex than another, they normally do not have any problem admitting that *constructions* vary with regard to complexity. A typical example would be the passive voice. A passive sentence in a language such as English contains more material than an active one, and many theoretical frameworks would have it that it is derived from its active counterpart through the application of a rule (or set of rules).

If we admit that different sentence structures vary in complexity within a single language, there is no reason to assume that this would not be possible in cross-linguistic comparisons. For instance, (1a) in Classical Chinese (from Norman 1988: 107) contains less material than its English translation in (1b).

(1a)	Classical Chinese:	*Zhì*	*zé*	*xíng*	*yǐ*
		arrive	CONJ	leave	PARTICLE
(1b)	English:	*When*	*he*	*arrive-d,*	*he*
		SUBJUNCTION	3SG.M.SUBJ	arrive-PAST	3SG.M.SUBJ
		had	*already*	*left*	
		AUX.PAST	already	leave.PAST	

Inversely, many would no doubt agree that the English translations in (2b) and (3b) are simpler than the corresponding Kathlamet Chinook (Mithun 1999: 386) or Temirgoi Adyghe (Hewitt 2005: 127) sentences:

(2a)	Chinook:	*ik-u-kua-qi.uit-x̣it*			
		IMMEDIATE.PAST-3PL.ABS-AGENTIVE-rest-PASSIVE			

(2b)	English:	*They*	*slept*
		3PL.SUBJ	sleep.PAST

(3a)	Adyghe:	*txəɫə-m*	*s-j-a-d͡ʒa-n-aw*
		book-DEF	OBL-1SG-3SG-INDIR.OBJ-read-FUT
		sə-qa-kʷʼa-ːʁ	
		ABS-1SG-hither-come	

(3b)	English:	*I*	*came*	*to*	*read*	*the*	*book*
		1SG.SUBJ	come.PAST	COMPL	read	DEF	book

Of course, Classical Chinese, Chinook and Adyghe are typologically rather different from English, but even related languages such as Russian and French are adorned with grammatical markings to extents differing from English:

(4a)	English:	*He*	*is*	*a*	*student*
(4b)	French:	*Il*	*est*	Ø	*étudiant*
(4c)	Russian:	*On*	Ø	Ø	*student*
		3SG.M	COP.3SG.PRES	INDEF	student

While the average linguist might agree that we are indeed dealing with differing degrees of complexity here, the standard objection is that complexity in one area of

grammar is compensated by relative simplicity in another. There is certainly some merit to this assumption – the English version of (4) is the most complex one, but we all know that French and Russian are equipped with intricacies quite unknown to English, such as rather arbitrary grammatical gender distinctions, extensive agreement, and pervasive case marking. And while Sinitic languages may look comfortably analytic (and at times even telegraphic) to a westerner, it is well known that they have, for instance, fairly complex tone systems and large sets of numeral classifiers. The communicative efficiency that the realities of human life require from languages would seem to be similar across societies, and given identical mental and articulatory equipments, it is indeed not unreasonable to expect languages all over the world to be equally complex.

However, human minds are clearly capable of handling more structure than is found in any one language. Many people grow up with two (or even more) mother tongues, and thus spend their childhood storing more linguistic data in their brains than did those of us who were raised in a monolingual environment. This appears not to pose any insurmountable obstacles to full acquisition, and a language with more complex structure than, say English, French or Russian would therefore not be incompatible with our language faculty – it would still be possible to acquire natively.

One might surmise that any excessive complexity would automatically wither away, as humans not only demand expressiveness from their languages, but also efficiency. Surely, the history of any language is replete with examples of assimilations and simplificatory reductions. But then, if all languages represent an optimal cost-benefit trade-off, all that disappears would have to somehow be replaced in order to maintain the precarious balance. To some extent this does hold true. Latin case endings were lost in its Romance offspring, but prepositions and a stricter word order compensate for it. For other features, though, it is not obvious how a loss was compensated for. As the only major European language, English has shed its former system of grammatical gender marking, and to my knowledge, no one has proposed that its job (whatever that would be!) has been taken over by any other mechanism. Indeed, in many of the case studies in Kusters (2003), it is quite obvious that reduction of a particular subsystem did *not* lead to expansion of another. And as Kusters (2003: 11) himself points out, for the conception that "complexities balance each other out" to hold, it would have to be shown that this is *always* the case, since no one denies that it sometimes is.

The very least that we must conclude, then, is that it is not scientifically responsible to *a priori* claim that all languages are equally complex. They might well be, but they need not.

2.   The relationship between expressiveness and complexity

The ideological roots of the equal-complexity dogma are not particularly difficult to identify. In the early 20th century, institutional racism gradually gave way to cultural

relativism. There was a need to convince people that the languages of technologically less advanced (from a western point of view) societies were by no means primitive. A classic illustration of the new line of thought was provided by Sapir (1921: 218-219), who wrote that "When it comes to linguistic form, Plato walks with the Macedonian swineherd, Confucius with the head-hunting savage of Assam".

These days, racism has ceased to be a salient feature of academic discourse. Yet, the concept that all languages are equally complex has become so rooted in the discipline that it is difficult to bring up for discussion, and is at best considered to be outside the scope of modern linguistics (Dixon 2003: 169; Kusters 2003: 4). As Dixon (2003: 171) observes, taking the next step forward "might be confused with returning to the first stage", i.e., again embracing Eurocentrism. There is a crucial difference, however, between Dixon's "first stage" and the proposed "third stage". In the bad old days, complexity was seen as intrinsically linked to expressive power – complex was good, and simple was bad, and that was that. Indeed, when certain languages have been claimed to be simpler than certain others, this has resulted in outcries of indignation (well exemplified in DeGraff 2001a) from readers who clearly perceive the claim to be that the purportedly simpler languages are less expressive and intrinsically less worth than the allegedly more complex ones.

However, as used by Mühlhäusler (1997: 128), Posner (1997: 122) and myself, the concept of simplification has nothing to do with reduction of expressive power (for which Mühlhäusler reserves the term *impoverishment*). Indeed, given that I here treat the equal expressiveness of languages as axiomatic, a simple language could rather be seen as a more "efficient" one, in the sense that it does the same job at a lesser cost.

A hypothetical language consisting only of the schwa phoneme would certainly be both less complex and less expressive than actually existing human languages. To take a real-life example, Busnel and Classe (1976: 87, 108–109) conclude that the whistled languages they examine are indeed simpler than ordinary speech, but they also show that this is at the expense of redundancy and preciseness, with frequent misinterpretations as a result. To some extent, thus, complexity and expressive power do go hand in hand. But at the level of human languages (which, after all, are so complex that no one in the history of mankind has ever managed to produce an ideally exhaustive reference grammar), this is not necessarily so. The simplest way of demonstrating this is through a parasitic cryptolect such as *Pig Latin*. Such a variety is arguably more complex (albeit only marginally) than English in that it contains one additional rule: "take the onset of each word's first syllable and add it to the end of the word, followed by *ay*". Yet, Pig Latin does not allow the speaker to express anything that English cannot. If it were the language of the ruling class, Pig Latin would perhaps be thought of as superior to English, but structurally speaking, it is only more complex without being more expressive. Inversely, we could imagine a variety of English which is identical to today's, only without any irregularities. This fictious English would be simpler, but again, neither more nor less expressive. In a sense, Esperanto, with its exceptionless grammar is such

a language, for it basically allows a finite number of rules to be applied without (in theory) any restrictions whatsoever. While Esperanto is an artificial language (and as such often scorned by professional linguists), it has proven capable of producing literary and scientific texts, and – more importantly, given the assumption that all natively spoken languages are expressively equal – it does have native speakers (Bergen 2001; Corsetti 1996; Corsetti et al. 2004; Versteegh 1993).

For the above reasons, therefore, it is perfectly possible to claim that all languages have the same expressive potential, while not necessarily being equally complex.

3.   Material and method

If, again, we take the expressivity part as axiomatic, is there any way we could investigate the degree of complexity? It is surely not possible to prove the differing-complexity hypothesis beyond any doubt, but, I claim, it is possible to adduce evidence that strongly supports it. The basic assumption behind my reasoning is simply that a complex language is a language with more complex constructions. If English has a passive, while Classical Chinese did not (Norman 1988: 101), then English is more complex *with regard to that specific construction*. This complexity may or may not be compensated for elsewhere, but if we studied a large number of constructions, and English turned out as the more complex in every single case, then, I propose, it would be justified to claim that English were more complex than Classical Chinese.

In 2005, The *World Atlas of Linguistic Structures* (Haspelmath et. al (eds) 2005; henceforth WALS) was published. The WALS database is the most extensive typological survey to date. If any material allows a large-scale assessment (for whatever purpose) of similarities and difference between languages, WALS certainly offers the best possibilities. It contains data on more than 2 500 languages, and discusses more than 140 linguistic features. The wealth is unevenly distributed, though. Almost half of all languages are represented by 15 features or less, and the ten best covered are all big languages, such as (in descending order of frequency) English, French, Finnish, Russian, Spanish, Turkish, Hungarian, Indonesian, Japanese and Mandarin. Yet, WALS includes close to 60 000 data points, which is unparalleled by any other database (at least any that is publicly available).

Among the features discussed in WALS, most are such that they do not have any obvious positive and negative value with regard to complexity. Typical representatives of this category are the word order features – it is not obvious that SOV is neither more nor less complex than SVO. Nor is it self-evident that the presence of an unusual feature (such as click phonemes or object-initial word order) is more complex than a common one in its place. Cross-linguistic frequency is a matter of statistics, and within a given system, fixed stress on the final syllable does not require a more complex specification than stressing of the initial one (which is more common). In many

cases, languages simply do things differently, without this having any bearing on the issue of complexity.[2]

A third category of features which I have made an effort to exclude is synthesis. A high degree of synthesis is certainly perceived as being complex by laymen, and so far as "gut feeling" is concerned, also by many of us linguists. It is certainly true that synthesis tends to *lead to* arguably complex structures, such as fusion. Yet, synthesis *in itself* is not complex by my use of the term. Danish *bogen* 'the book' (where *-en* is the non-neuter definite article) differs from Spanish *el libro* only by being synthetic, but otherwise requires the same grammatical machinery to produce. There is also a pedagogical reason to exclude synthesis – we already know that the languages of the world differ considerably with regard to this, and a complexity measure including it might do little but confirm the impression that puts Greenlandic and Vietnamese at opposite ends of the scale. That said, the material used here does not allow to circumnavigate the synthesis issue completely – for some of the features, WALS only indicates the presence of a bound form versus the complete absence. This, however, applies to a small minority of features, as indicated in the list below, and its effect seems marginal, as we shall see.

In short, while any feature could be subject to debate, my aim is to use those that are likely to cause the least controversy among linguists. For all the traits listed, I consider their presence (or the presence in larger numbers) to add to the overall complexity of a language.

In some cases, as discussed earlier, we could expect an absent feature to be compensated for by the presence of something not studied by WALS, but I have made some effort to exclude such traits.[3] Of course, a language not having a given feature is normally *able* to express it, but it does not *have to*, and may not even have a conventionalized strategy to unambiguously do so. It is at the very least difficult to identify what in Standard Average European would correspond to the duals, the several degrees of "pastness", the antipassives, the two dozen noun classes, the numeral classifiers, the inclusivity, evidentiality and alienability distinctions and the distinction between various copula types found in certain more "exotic" languages. Again, while a language such as English is able to indicate the distinctions made by evidentiality affixes in Amazonia by circumlocutions (*It is said that. . .*, etc.), these are by no means obligatory, and do not

---

2.   Sometimes, crosslinguistic frequency and complexity are even opposed to one another: the vast majority of languages require their speakers to make a distance contrast in demonstratives (WALS, Map 42), while a demonstrative system without this elaboration (everything else being equal) would clearly require a shorter description.

3.   A typical example is that of case affix inventories, for which WALS provides data. There is none, however, on the size of the adposition arsenal, whose members in case-less languages tend to correspond to (at least locative) cases.

form a central part of the grammar of English. Evidentiality in English is purely a matter of pragmatics, rather than distinctions imposed on the speaker by her language.

Importantly, if the correspondences between a feature in language A and a compensatory feature in language B cannot be made, the only reason to assume that it is at all compensated for is the idea that differences *must* somehow balance each other out – and this clearly makes the reasoning circular. In cases such as that of evidentiality, we can at least find a periphrastic correspondence in English, but in the case of, say, gender, there is not even a way of translating the gender tag that every NP in certain other languages carries with it. French *la table* consists of a definite article, the noun 'table' and a gender marker, of which only the first two correspond to anything whatsoever in English. There is simply no way the gender marker could be expressed in English without making reference to French (especially not since gender is language-specific, with 'table' being masculine in e.g., German and Russian).

It could be argued (and rightly so) that such distinctions make a contribution to the overall redundancy of a language, in which case gender might as well correspond to, say, a larger phoneme inventory or longer morphemes. But then again, this presupposes that all languages have a similar (and constant!) degree of redundancy, which again makes the argument circular.

Having set the basis of my reasoning, here follow the data from WALS chosen for further consideration:

Table 1. WALS features included in the study.

	Feature[a]	WALS map no.
F01	Size of consonant inventories	1
F02	Size of vowel quality inventories	2
F03	Phonemic vowel nasalization	10
F04	Complexity of syllable structure	12
F05	Tone	13
F06	Overt marking of direct object	23
F07	Double marking of direct object	23
F08	Possession by double marking	24
F09	Overt possession marking	24
F10	Reduplication	27
F11	Gender	30–32
F12	Number of genders	30
F13	Non-semantic gender assignment	32
F14	Grammaticalized nominal plural	33–34
F15	Definite articles	37
F16	Indefinite articles	38
F17	Inclusivity (in either pronouns or verb morphology)	39–40
F18	Distance contrast in demonstratives	41

(*Continued*)

Table 1. Continued.

	Feature[a]	WALS map no.
F19	Gender in pronouns	44
F20	Politeness in pronouns	45
F21	Person marking on adpositions	48
F22	Comitative ≠ instrumental	52
F23	Ordinals exist as a separate class beyond 'first'	53
F24	Suppletive ordinals beyond 'first'	53
F25	Obligatory numeral classifiers	55
F26	Possessive classification	59
F27	Conjunction 'and' ≠ adposition 'with'	63
F28	Difference between nominal and verbal conjunction	64
F29	Grammaticalized perfective/imperfective	65
F30	Grammaticalized past/non-past	66
F31	Remoteness distinctions of past	66
F32	Morphological future	67
F33	Grammaticalized perfect	68
F34	Morphological imperative	70
F35	Morphological optative	73
F36	Grammaticalized evidentiality distinctions	76
F37	Both indirect and direct evidentials	77
F38	Non-neutral marking of full NPs	98
F39	Non-neutral marking of pronouns	99
F40	Subject marking as both free word and agreement	101–102
F41	Passive	107
F42	Antipassive	108
F43	Applicative	109
F44	Obligatorily double negation	112
F45	Asymetric negation[b]	113
F46	Equative copula ≠ Locative copula	119
F47	Obligatorily overt equative copula	120
F48	Demonstratives marked for number	–
F49	Demonstratives marked for gender	–
F50	Demonstratives marked for case	–
F51	Total amount of verbal suppletion	–
F52	Alienability distinctions	–
F53	Number of pronominal numbers	–

[a] It is not possible to provide detailed descriptions of each feature here, so for precise definitions, please refer to WALS.
[b] This basically means that the negated sentence differs from its positive counterpart in other ways than the mere addition of a negator morpheme.

A few features not present in the WALS database were added. F48-50 were taken from Diessel (1999: 171–173), while F51 was provided by Ljuba Veselinova (p.c.) – these data are the "leftovers", as it were, from her contributions to WALS and from Veselinova (2003). F52 was taken from Nichols (1992: 294-301), and F53 from Harley and Ritter (2002). The additions of these (but not other) data were done simply because I happened to have access to them. With the exception of Harley and Ritter, all the authors concerned are also contributors to WALS, and the added material largely consists of elaborations of their respective WALS articles.

In some cases, we could wish for more data that might reveal if the absence of a feature is indeed compensated for elsewhere in the grammar. However, in one sense, there is a point in using a predefined data source. WALS is produced by the world's most renowned typologists, whose combined expertise is second to none. They have chosen their features purely out of an interest in typology, and without the hidden agenda that I could, would, and have been accused of having.

Most features have binary yes/no values (occasionally with an intermediate value) which easily translate into the numbers 1 and 0 (or 0.5, as the case may be). Those which do not are F01, F02, F04, F05, F12, F18, F20, F26, F51 and F53. Some of these are numerically coded, and in those cases I compressed the values to fit the zero-to-one range. Thus, the pronominal numbers (F53), which range from zero to four, were straightforwardly converted into the format 0, 0.25, 0.5, 0.75 and 1. In the remaining cases, the values were expressed in prose in WALS. Syllable structures (F04), for instance, were classified in the source as "simple", "moderately complex" and "complex". In my grid, these classifications were converted into 0, 0.5 and 1. For all features, thus, the possible range is from zero (minimally complex) to one (maximally complex).[4]

As mentioned earlier, there are plenty of gaps in the WALS database. I therefore decided to restrict myself to the languages with values for at least 30 out of 53 features considered. This results in 155 languages and a total coverage above 80% (i.e., blank cells amounting to less than 20% in a 155 × 53 grid). The total score for each language is divided by the number of features included for that language, and thus, Ladakhi (with 31 attestations) stands a chance to get an average score on par with Basque, for which all 53 cells are filled in.

---

4.   As pointed out by Matti Miestamo (p.c.), the use of 0.5 for intermediate values has its problems. For instance a language that has both symmetric and asymmetric negation (F45) may have much more complex asymmetry than a language where negation is always asymmetric. While this is clearly something that requires future revision, I am convinced that the overall effect of this on the following calculations is minimal.

4.   Results

The results of the calculations are tabulated in Table 2 below:

Table 2. Complexity ranking of WALS languages.

Rank	Language	Complexity score	Rank	Language	Complexity score
1	Burushaski	0.62	79	Fijian	0.41
2	Copainalá Zoque	0.57	80	Amele	0.41
3	Khoekhoe	0.57	81	Khalkha	0.41
4	Beja	0.56	82	Oneida	0.41
5	Koasati	0.55	83	Chamorro	0.41
6	Kannada	0.55	84	Maung	0.41
7	Ladakhi	0.54	85	Malagasy	0.40
8	Abkhaz	0.54	86	Kewa	0.40
9	Hindi	0.54	87	Ainu	0.39
10	Latvian	0.53	88	Mezquital Otomí	0.39
11	Spanish	0.53	89	Mandarin	0.39
12	Greek	0.53	90	Arapesh	0.39
13	Grebo	0.52	91	Plains Cree	0.39
14	Basque	0.52	92	Semelai	0.39
15	Ingush	0.52	93	Nez Perce	0.39
16	French	0.51	94	Egyptian Arabic	0.39
17	Kanuri	0.51	95	Ju|'hoan	0.39
18	Supyire	0.50	96	Nivkh	0.39
19	Maricopa	0.50	97	Hungarian	0.38
20	Passamaquoddy-Maliseet	0.50	98	Maori	0.38
21	Evenki	0.50	99	Kayardild	0.38
22	Comanche	0.50	100	Igbo	0.38
23	Yaqui	0.50	101	Yurok	0.38
24	Mundari	0.49	102	Urubú-Kaapor	0.38
25	Alamblak	0.49	103	Burmese	0.38
26	Dyirbal	0.49	104	Zuni	0.37
27	German	0.49	105	Acoma	0.37
28	Meithei	0.49	106	Asmat	0.37
29	Korean	0.49	107	Ngiyambaa	0.37
30	Slave	0.49	108	Imbabura Quechua	0.37
31	Lezgian	0.49	109	Wari'	0.37
32	Georgian	0.49	110	Yoruba	0.37
33	Jakaltek	0.49	111	Bagirmi	0.37
34	Hunzib	0.48	112	Kolyma Yukaghir	0.36
35	Zulu	0.48	113	Martuthunira	0.36
36	Karok	0.48	114	Gooniyandi	0.36

(*Continued*)

Table 2. Continued.

Rank	Language	Complexity score	Rank	Language	Complexity score
37	Hebrew	0.48	115	Rama	0.36
38	Paiwan	0.48	116	Iraqw	0.36
39	Nenets	0.48	117	Suena	0.35
40	Tagalog	0.47	118	Chukchi	0.35
41	Swedish	0.47	119	Ket	0.35
42	Kunama	0.47	120	Brahui	0.35
43	Mangarrayi	0.47	121	Lango	0.34
44	Diola-Fogny	0.47	122	Yagua	0.34
45	Luvale	0.47	123	Awa Pit	0.33
46	Japanese	0.46	124	Tiwi	0.33
47	Middle Atlas Berber	0.46	125	Canela-Krahô	0.33
48	Eastern Armenian	0.46	126	Rapanui	0.33
49	Swahili	0.46	127	Ewe	0.32
50	Persian	0.46	128	Haida	0.32
51	Lower Grand Valley Dani	0.46	129	Koyraboro Senni	0.32
52	Hausa	0.46	130	Khmer	0.32
53	Krongo	0.46	131	Wichita	0.31
54	Kiowa	0.46	132	Thai	0.30
55	Finnish	0.45	133	Taba	0.30
56	Barasano	0.45	134	Kilivila	0.30
57	West Greenlandic	0.45	135	Ika	0.30
58	Coos	0.45	136	Tetelcingo Nahuatl	0.30
59	Turkish	0.45	137	Warao	0.30
60	Wardaman	0.45	138	Mapudungun	0.30
61	Harar Oromo	0.44	139	Cayuvava	0.29
62	Cahuilla	0.44	140	Kutenai	0.29
63	Khasi	0.44	141	Shipibo-Konibo	0.28
64	Lahu	0.44	142	Tukang Besi	0.28
65	Guaraní	0.44	143	Yidiny	0.27
66	Lavukaleve	0.44	144	Apurinã	0.27
67	Imonda	0.44	145	Chalcatongo Mixtec	0.27
68	Epena Pedee	0.44	146	Indonesian	0.26
69	Lakhota	0.43	147	Vietnamese	0.26
70	Nunggubuyu	0.43	148	Daga	0.25
71	Sanuma	0.43	149	Usan	0.23
72	Hixkaryana	0.43	150	Ndyuka	0.22
73	Russian	0.43	151	Maybrat	0.22
74	Wichí	0.43	152	Kobon	0.20
75	Southern Sierra Miwok	0.42	153	Hmong Njua	0.20
76	Yimas	0.42	154	Pirahã	0.18
77	Paumarí	0.42	155	Sango	0.15
78	English	0.42			

A striking – and, to me, unexpected – characteristic of these results is that the Indo-European languages are so well-represented at the top of the list. I am uncertain why that is so. Either this is a fair description of the state of affairs, or perhaps it illustrates subconscious Eurocentrism on behalf of me or the WALS contributors (or both).

With this exception, however, the spontaneous reaction from most of those with some grasp of typology to whom I have shown the list is that it is largely compatible with their intuitions.

5.   The special case of creoles

Disregarding the position of European languages, I find the bottom ranks to be the most interesting part of the list. Positions 150 and 155 are occupied by the only two creoles in the sample – Ndyuka and Sango. Creoles have frequently been suggested to have simpler grammars than other languages, most recently by McWhorter (2001). The suggestion has been met with scepticism by many linguists (see for instance the reactions in *Linguistic Typology* 5 (2–3), 2001), and within creolistics itself by the vast majority. Nevertheless, the subject has been considered so important that even the most prestigious journal of the discipline, *Language*, took the rare step of publishing an opinion piece on the matter (DeGraff 2003). Without a single linguistic example (let alone comparative evidence), that paper dismisses ideas about the relative simplicity of creoles as 'myths', 'prejudices', 'illusions', and 'fantasies'.

In this light, it is indeed striking that the two creoles here rank among the least complex languages, both surpassed by 98% of the sample. Nevertheless, four languages turn out to be simpler than at least one of the creoles. It thus seems that, at least in the present metric, individual non-creoles can indeed be less complex than individual creoles. These four, however, do not form a class (areal, genetic or otherwise), a fact to which we shall return shortly.

6.   What are creoles?

Before continuing, a few background details are necessary.

During the decades that creolistics has existed as a semi-autonomous subdiscipline, opinions on what creoles *really* are have varied considerably. Two traditional views are perhaps better known to the general linguist than others. First, some creolists have seen creoles more or less as relexifications of their substrates, i.e., old (typically African or Melanesian) grammars equipped with new (typically European) lexica. Representatives of this school are e.g., Keesing (1988) and Lefebvre (1998). Then, according to Derek Bickerton (1981, 1984), creoles offer a unique window on the human language faculty by representing overt manifestations of the unmarked parameter settings. Known as the *Language Bioprogram Hypothesis*, this view had an impact outside of

creolistics, and is occasionally still quoted as representing the creolistic state of the art. Within the subdiscipline itself, the reception was less cordial, and the Bioprogram is now dead as the dodo.

Meanwhile, the dominant tendency among Francophone creolists (prime examples being Chaudenson 1979, 1992, 1995) was to consider creoles direct descendants of their lexifiers. The relationship between, say, Haitian and French was claimed to be largely similar to that between Latin and French.

After having been viewed with suspicion by outsiders, the 1990s saw this view make advances among creolists writing in English. Propagated by in particular Salikoko Mufwene and Michel DeGraff, it has now gained a considerable following, and may perhaps represent the closest thing to a consensus that there is in creole studies. Following DeGraff (2003), I shall refer to this school of thought as *uniformitarian.*

Two of the most central claims made by the uniformitarian school are 1) that creoles do not derive from pidgins, but instead fit equally well into a family tree model as other languages do (Chaudenson 1995: 66; Mufwene 2000, 2001: 152, 153; DeGraff 2002: 377, 378, 2003: 398, 399), and 2) that the very label "creole" has no structural/typological correlate, and is solely defined by its history (Chaudenson 2001: 145; DeGraff 2003: 391; Mufwene 2001: 10, 2002, 2003: 277–278). A language is/should be/can be classified as a creole exclusively on the basis of its past. Mufwene (2000) defines creoles as "a group of vernaculars whose developments are similar especially in their temporal and geographical positions, viz., in tropical colonies settled by Europeans practicing slave-based economy from the 17th to the 19th centuries. The lexifiers are typically non-standard varieties of European languages" (similar definitions are found in Mufwene 2003 and DeGraff 2003: 391).

The results from the present study do not confirm the view on creoles supported by uniformitarians.

7.   Adding more contact languages

The presence of only two creoles in the sample (Ndyuka and Sango) is somewhat problematic, however, so I entered as much information as I could from another 30 pidgins and creoles, resulting in the following values.[5]

---

5.   Reference grammars are available for only some of these languages, so much of the data were taken from sources too numerous to mention here. The most important sources, however, are Bailey (1966), Baker (1972), Barrena (1957), Baxter (1988), Bollée (1977), Broch and Jahr (1984), Carrington (1984), Corne (1977), Damoiseau (1984), Ehrhart (1993), Ferraz (1979), Günther (1973), Heine (1982), Kihm (1994), Kouwenberg and Murray (1994), Munteanu (1996), Scantamburlo (1981) and Verhaar (1995) (details available on request). In cases where several sociolects have been identified, I am referring here to the basilect, i.e., the variety the furthest removed from the lexifier. The same 30-feature limit as used before applies here as well.

Table 3. Complexity values of pidgins and creoles added to the sample.

Language	Main lexifier	Area	Type[a]	Score
Annobonese	Portuguese	West Africa	creole	0.22
Australian Creole	English	Pacific	creole	0.16
Bislama	English	Pacific	expanded pidgin	0.13
Chinook Jargon	Chinook	North America	(expanded) pidgin	0.15
Dominican	French	Caribbean	creole	0.21
Fanakalo	Zulu	South Africa	pidgin	0.20
Guadeloupean	French	Caribbean	creole	0.25
Guinea Bissau Creole	Portuguese	West Africa	creole	0.27
Haitian	French	Caribbean	creole	0.29
Jamaican	English	Caribbean	creole	0.27
Kinubi	Arabic	East Africa	creole	0.24
Krio	English	West Africa	creole	0.28
Lingua Franca	Italian/Romance	Mediterranian	pidgin	0.06
Martinican	French	Caribbean	creole	0.20
Mauritian	French	Indian Ocean	creole	0.20
Negerhollands	Dutch	Caribbean	creole	0.20
Nigerian Pidgin	English	West Africa	expanded pidgin	0.25
Palenquero	Spanish	South America	creole	0.27
Papia Kristang	Portuguese	South-East Asia	creole	0.18
Papiamentu	Spanish	Caribbean	creole	0.33
Principense	Portuguese	West Africa	creole	0.23
Russenorsk	Norwegian & Russian	Arctic Ocean	pidgin	0.00
Sãotomense	Portuguese	West Africa	creole	0.31
Saramaccan	English	South America	creole	0.25
Seychellois	French	Indian Ocean	creole	0.25
Sranan	English	South America	creole	0.24
St. Lucian	French	Caribbean	creole	0.22
Tayo	French	Pacific	creole	0.22
Tok Pisin	English	Pacific	expanded pidgin	0.22

[a] The labels "pidgin" and "creole" are used here in the traditional way, i.e., for non-native and native varieties respectively. "Expanded pidgin" refers to a variety which is mainly used as an L2, but which serves as the primary (and in some cases even native) language for some of its users.

As with the sample already discussed, there are gaps in the additional creole data, and it would be surprising if there were no errors or debatable classifications among these almost 1 200 additional data points. In case the suspicion should arise that these additional values were tailor-made to suit my preconceptions, it should be noted that these additional languages score slightly *higher* than Sango and Ndyuka, the two creoles for which data were provided by WALS. While the addition did not so much alter

the position of creoles, it does indicate that Sango and Ndyuka are rather representative in this respect.

8.   Creoles and other groups compared

Now, recall that the uniformitarian school denies that creoles constitute a typological class, or that they are in any way identifiable on structural grounds. Therefore, any resemblances that would make them *look like* a group would be due to chance. It should thus be possible to come up with other groupings that are superficially equally tenable.

I experimented with 90 different ways of lumping languages together. First, on the basis of geography. Secondly, I grouped languages according to family. Third, on typological grounds, using 26 randomly selected WALS features not already involved in my calculations. Fourth, 16 groups were formed on the basis of a number of sociolinguistic factors. These four groups are referred to in the table below as GEO, GEN, TYPO and SOC respectively.

It could be suspected that these factors would have some bearing on complexity, e.g., because of areal pressure, through inheritance of complexity (or simplicity), by means of factors such as typological "harmony" or sociological correlations. Finally, however, I tested a number of *intentionally nonsensical* features that are not normally believed to have anything to do with linguistic structure (indicated as SILLY in the table below).

The point here is this: if there are individual languages which are less complex than individual creoles, are there also groups of languages which are less complex than creoles? If not, such a result would further support the notion of creoles as a typological class.

The following table demonstrates the clustering of various groupings with regard to the degree of complexity:[6]

Table 4. Complexity values of various types of groupings of the WALS languages.

	Complexity score 0.30–0.39	Complexity score 0.40–0.49
GEO	Languages of Oceania, of South America, of South-East Asia, of the southern hemisphere	Languages of Africa, of Asia, of Europe, of North America, of the eastern hemisphere, of the northern hemisphere, of the western hemisphere
GEN	Austro-Asiatic, Austronesian, Pama-Nyungan, Trans-New Guinea	Afro-Asiatic, Algic, Gunwingguan, Indo-European, Isolates, Niger-Congo, Nilo-Saharan, Penutian, Sino-Tibetan, Uralic, Uto-Aztecan lgs

(*Continued*)

---

6.   Groups with less than three representatives in the sample are excluded.

Table 4. Continued.

	Complexity score 0.30–0.39	Complexity score 0.40–0.49
TYPO[a]	Head marking languages, SVO languages, predominantly prefixing languages, languages encoding ditranstitives as double-object constructions, with 'exceed'-comparatives, without laterals, without morphological case marking	Dependent marking languages, lgs differentiating 'hand' and 'finger', in which "adjectives" are morpho-syntactically "noun-like", in which "adjectives" are morphosyntactically "verb-like", not differentiating 'hand' and 'finger', requiring overt subject pronoun, with /ŋ/, with decimal numeral systems, with glottalized consonants, with laterals, with less than average degree of synthesis, with more than average degree of synthesis, with morphological case marking, with non-decimal numeral systems, without /ŋ/, without 'exceed'-comparatives, predominantly suffixing lgs, pro-drop lgs, SOV lgs
SOC[b]	Languages spoken mainly in virtually monolingual nations, spoken mainly in countries where Spanish is official	Languages spoken in former European colonies, without official status in any country, with less than 1000 speakers, with official status in more than one country, with official status in at least on e country, with more than 1M speakers, spoken in less literate (=below world average) societies, spoken in highly literate (=above average) societies, spoken mainly in Commonwealth countries, spoken mainly in *Francophonie* countries, spoken mainly in very multilingual nations, majority languages (in their respective countries), minority languages (ditto)
SILLY[c]	Languages using Roman script, spoken by yellow peoples, by red peoples, mainly in countries whose name begin with the letter C, mainly in countries whose name begin with the letter P	Languages mentioned in Chomsky (1966), lgs spoken by black peoples, by white peoples, by Christians, by Moslems, in areas free of malaria, in malaria-infested areas, in countries with an above average mobile phone ownership, in countries with a below average mobile phone ownership, in democracies, in dictatorships, using Arabic script, using Cyrillic script, whose names begin with the letter A, whose names begin with the letter K, whose names begin with the letter M, whose names begin with the letter S

[a] The typological information in this category was taken from the following WALS maps: 7, 8, 9, 25, 26, 49, 81, 101, 105, 118, 121, 130 and 131. The "average degree of synthesis" refers to an index (produced in the same way as the complexity index) based on boundness of tense/aspect marking, number marking and negation, person marking on verbs, two measures of inflexional morphology in general and two indices of fusion (maps 20, 21, 22, 26, 33, 69, 100, 101 and 112 in WALS).

[b] "Virtually monolingual" are those with a diversity index of 0.1 or less according to the *Ethnologue* database (15th edition). "Very multilingual" nations are those with a diversity index of at least 0.9 according to the same source (for the definition of this index, see Greenberg 1956). Numbers of speakers were also derived from the *Ethnologue*.

[c] Literacy figures and data on mobile phone ownership are from the CIA World Factbook. Democracies vs. dictatorships correspond to the labels "free" and "not free" respectively, as used by the organization Freedom House. Scripts and religions were checked with the Ethnologue. Spread of malaria was defined by the World Health Organization.

The absence of a column for values below 0.30 has a simple explanation – no group, neither sensible nor nonsensical, falls into this range. One single (genetic) group, namely North Caucasian, produced an average slightly above 0.50, and as can be seen, all groups hover around 0.40.

The average value for all non-creole (and non-pidgin) languages involved here is 0.41, and deviations from this value are relatively small. It is interesting, then, to compare this to the pidgin and creole values:

Table 5. Complexity values of the contact languages and their subgroups.

Language type	Average complexity
Creoles (excluding expanded pidgins)	0.24
Creoles (including expanded pidgins)	0.23
Sango & Ndyuka only	0.19
Expanded pidgins only	0.19
All pidgins	0.14

Given the intermediate status of the "expanded pidgins", they are here included both in the pidgin and the creole categories, as well as being given separately. Sango and Ndyuka are indicated separately because they are included in the WALS sample, and so their values cannot be due to my supplying suspect data.

A *t*-test shows that the difference between creoles (and expanded pidgins) and non-contact languages is statistically significant: a P value of < 0.001 means that there is less than a one in a thousand possibility that the difference is due to mere chance.[7]

The most interesting observation here is that it is difficult to come up with *or even to invent* a category whose members behave typologically like creoles do. If we choose – as I have done – to talk about the crucial difference between creoles and non-creoles in terms of "complexity" or not, could even be considered a matter of taste. What is paramount is that creoles are different vis-à-vis non-creoles – that alone contradicts the uniformitarian position. Of some importance with regard to the birth of creoles is that the group of languages which most closely resembles them is the category of pidgins. And among these languages, the so-called expanded pidgins are not only sociologically, but also structurally intermediate between the pidgin and the creole groups. This conforms with the idea that pidgins expand into creoles – again something that the uniformitarian school denies.

Some of the parameters in Table 4 are particularly interesting. For instance, few people – including those who are convinced about the equicomplexity among

---

7. The standard deviation among the creoles (0.05) is also slightly less, suggesting that the group is more homogenous with regard to structural complexity than are languages in general (0.09), something that again emphasizes that they do constitute a valid typological group.

languages – would deny that creoles are rather analytic. It has sometimes been suggested that the perceived simplicity is first and foremost a product of the analytical makeup. If nothing else, however, I have apparently succeeded in avoiding such a bias by excluding most synthesis from my complexity measure – above-average synthetic languages are hardly more complex (0.42) than the members of the below-average group (0.40). In particular, the latter do not in any way approach the creoles (0.22). It is in other words perfectly possible to be analytic, and yet be as complex as the average language – it is just that this is not what creoles do. Among the typological features, several are common in creoles (for instance SVO word order, 'exceed'-comparatives, lack of morphological case marking and "verb-like" adjectives) – but again, other languages with these features are more complex than creoles.

Given the political concerns of DeGraff (2001b: 2, 2001c: 49; Saint-Fort 2002), it is notable that "languages spoken in former European colonies", languages "without official status in any country", and "languages spoken in less literate societies" do not behave differently from others.[8] Finally, even some of the SILLY features are pertinent in this context. DeGraff has repeatedly claimed that the very idea that creoles are different from other languages is rooted in racism (since speakers of the best known creoles, including the one that he discusses, are mostly black).[9] Therefore, it is worth noting that "languages spoken by white peoples" and "languages spoken by black peoples", as well as languages spoken by humans of other pigmental endowment are about equally complex. It is creoles that stand out, not the languages of black (or red, yellow or white) peoples.

It is of course possible that an unproportionate share of a given language's complexity has been tucked away in corners of the grammar that WALS did not reach. But while this is possible, and even likely, for individual languages, things are different for entire groups. Recall that it has been claimed that the label "creole" has no typological correlate whatsoever. Should it indeed be the case that the group which turns out to be simplest in my metric is in fact more complex in other areas, then *that in itself would set creoles apart from other languages* – they would then not be "the worlds simplest languages", but instead "the languages whose complexity resides in unusual places". Whichever option you choose, it is difficult to deny that there is such a thing as a "creole typological profile".

---

8.   The latter is also interesting because of Chaudenson's (e.g., 2001: 159) notion that at least some aspects of complexity are upheld because of normative pressure, something that virtually presupposes literacy.

9.   Views such as those presented in this paper are said to represent "neocolonial intellectual imperialism" (DeGraff 2001c: 45) and an "anti-egalitarian approach" (DeGraff 2001c: 46–47), whereby I and others uphold "mythical hierarchies among human languages, with Creoles consistently ranked at the bottom" (DeGraff 2001c: 49). Moreover this line of thinking is not only "race-based" (DeGraff 2003: 391), but even has "race theory as cornerstone" (DeGraff 2001d).

9. Conclusion

Typologically speaking, creoles stand out from languages in general, and the most salient difference is that they present a lower structural complexity. This does not necessarily have any bearing on issues regarding psycholinguistic complexity, however, and certainly not on their expressive potential.

A comparison of the complexity found within pidgins, expanded pidgins, creoles and other languages shows that the *complexity of a language correlates with its age*. This is precisely what is predicted by a scenario (the most well-known being that set out in McWhorter 2001) in which creoles emerge through broken transmission, and where complexity accretes over time.

Abbreviations

1	first person
3	third person
ABS	absolutive
AUX	auxiliary
COMPL	complementizer
DEF	definite
FUT	future
INDIR	indirect
M	masculine
OBJ	object
OBL	oblique
PL	plural
SG	singular
SUBJ	subject

References

Aitchison, J. 2001. *Language Change: Progress or Decay?* (3rd ed.). Cambridge: Cambridge University Press.

Bailey, B. 1966. *Jamaican Creole Syntax*. Cambridge: Cambridge University Press.

Baker, P. 1972. *Kreol: A Description of Mauritian Creole*. London: C. Hurst & Co.

Barrena, N. 1957. *Gramática Annobonesa*. Madrid: Instituto de Estudos Africanos/Consejo Superior de Investigaciones Científicas.

Baxter, A. 1988. *A Grammar of Kristang (Malacca Creole Portuguese)*. Canberra: Australian National University.

Bergen, B. 2001. Nativization processes in L1 Esperanto. *Journal of Child Language* 28(3): 575–595.

Bickerton, D. 1981. *Roots of Language*. New York: Karoma Publishers.

Bickerton, D. 1984. The language bioprogram hypothesis. *The Behavioral and Brain Sciences* 7: 173–188.

Bollée, A. 1977. *Le créole français des Seychelles: Esquisse d'une grammaire, textes, vocabulaire.* Tübingen: Max Niemeyer.

Broch, I. & Jahr, E.H. 1984. *Russenorsk – et pidginspråk i Norge.* Oslo: Novus.

Busnel, R.-G. & Classe, A. 1976. *Whistled languages.* Berlin: Springer-Verlag.

Carrington, L. 1984. *Saint Lucian Creole: A Descriptive Analysis of its Phonology and Morpho-Syntax.* Hamburg: Helmut Buske Verlag.

Chaudenson, R. 1979. *Les créoles français.* Paris: Nathan.

Chaudenson, R. 1992. *Des îles, des hommes, des langues. Essai sur la créolisation linguistique et culturelle.* Paris: l'Harmattan.

Chaudenson, R. 1995. *Les créoles.* Paris: Presses Universitaires de France.

Chaudenson, R. 2001. *Creolization of Language and Culture.* London: Routledge.

Chomsky, N. 1966. *Cartesian Linguistics: A Chapter in the History of Rationalist Thought.* New York: Harper & Row.

Corne, C. 1977. *Seychelles Creole grammar: Elements for Indian Ocean Proto-Creole Reconstruction.* Tübingen: Verlag Günther Narr.

Corsetti, R. 1996. A mother tongue spoken mainly by fathers. *Language Problems and Language Planning* 20(3): 263–273.

Corsetti, R., Pinto, M.A. & Tolomeo, M. 2004. Regularizing the regular: The phenomenon of overregularization in Esperanto-speaking children. *Language Problems and Language Planning* 28(3): 261–282.

Crystal, D. 1997. *English as a Global Language.* Cambridge: Cambridge University Press.

Damoiseau, R. 1984. *Éléments de grammaire du créole martiniquais.* Fort-de-France: Hatier-Antilles.

DeGraff, M. 2001a. On the origin of Creoles: A Cartesian critique of Neo-Darwinian linguistics. *Linguistic Typology* 5(2–3): 213–310.

DeGraff, M. 2001b. The mismeasure of the Creole speaker. Paper presented at the University of Rochester, 27 April 2001.

DeGraff, M. 2001c. Morphology in Creole Genesis: Linguistics and Ideology. In *Ken Hale: A Life in Language*, M. Kenstowicz (ed.), 53–121. Cambridge: MIT Press. Pagination here refers to the MS version.

DeGraff, M. 2001d. Salikoko Mufwene – from the Congo to Chicago. *Carrier Pidgin* 29.

DeGraff, M. 2002. Relexification: A reevaluation. *Anthropological Linguistics* 44(4): 321–414.

DeGraff, M. 2003. Against Creole exceptionalism. *Language* 79(2): 391–410.

Diessel, H. 1999. *Demonstratives. Form, function and grammaticalization.* Philadelphia: John Benjamins.

Dixon, R. 2003. A program for linguistics. *Turkic Languages* 7: 157–180.

Ehrhart, S. 1993. *Le créole français de St-Louis (le tayo) en Nouvelle-Calédonie.* Hamburg: Helmut Buske.

Ferraz, L.I. 1979. *The Creole of São Tomé.* Johannesburg: Witwatersrand University Press.

Fromkin, V. & Rodman, R. 1974. *An Introduction to Language.* New York: Holt, Rinehart & Winston.

Greenberg, J. 1956. The Measurement of Linguistic Diversity. *Language* 32 (1): 109–115.

Günther, W. 1973. *Das portugiesische Kreolisch der Ihla do Príncipe.* Marburg: Marburger Studien zur Afrika- und Asienkunde.

Harley, H. & Ritter, E. 2002. A feature-geometric analysis of person and number. *Language* 78 (3), 482–526.

Haspelmath, M., Dryer, M, Gil, D. & Comrie, B. (eds). 2005. *The World Atlas of Language Structures*. Oxford: Oxford University Press.

Heine, B. 1982. *The Nubi Language of Kibera – An Arabic Creole*. Berlin: Dietrich Reimer Verlag.

Hewitt, G. 2005. North West Caucasian. *Lingua* 115(1–2): 91–145.

Hockett, C. 1958. *A Course in Modern Linguistics*. New York: Macmillan.

Keesing, R. 1988. *Melanesian Pidgin and the Oceanic Substrate*. Stanford: Stanford University Press.

Kihm, A. 1994. *Kriyol Syntax. The Portuguese-Based Creole Language of Guinea-Bissau*. Amsterdam: John Benjamins.

Kouwenberg, S. & Murray, E. 1994. *Papiamentu*. Munich and Newcastle: Lincom Europa.

Kusters, W. 2003. *Linguistic Complexity: The Influence of Social Change on Verbal Inflection*. Utrecht: LOT.

Lefebvre, C. 1998. *Creole Genesis and the Acquisition of Grammar*. Cambridge: Cambridge University Press.

McWhorter, J. 2001. The world's simplest grammars are creole grammars. *Linguistic Typology* 5(2–3): 125–166.

Mithun, M. 1999. *The Languages of Native North America*. Cambridge: Cambridge University Press.

Mufwene, S. 2000. Creolization is a social, not a structural, process. In *Degrees of Restructuring in Creole Languages*, E. Schneider & I. Neumann-Holzschuh (eds), 65–84. Amsterdam: John Benjamins.

Mufwene, S. 2001. *Ecology of Language Evolution*.: Cambridge University Press.

Mufwene, S. 2002. What do creoles and pidgins tell us about the evolution of language? Paper presented at the Conference on Language Origins, Paris, September 26–27, 2002.

Mufwene, S. 2003. Genetic linguistics and genetic creolistics: A response to Sarah G. Thomason's "Creoles and genetic relationships". *Journal of Pidgin and Creole Languages* 18(2): 273–288.

Mühlhäusler, P. 1997. *Pidgin and Creole Linguistics. Expanded and revised edition*. London: University of Westminster Press.

Munteanu, D. 1996. *El papiamentu, lengua criolla hispánica*. Madrid: Editorial Gredos.

Nichols, J. 1992. *Linguistic Diversity in Space and Time*. Chicago: University of Chicago Press.

Norman, J. 1988. *Chinese*. Cambridge: Cambridge University Press.

Posner, R. 1997. *Linguistic Change in French*. Oxford: Clarendon Press.

Saint-Fort, H. 2002. Qu'est-ce qu'une Langue Créole? (deuxième partie). *The Haitian Times* 4(31): 2.

Sapir, E. 1921. *Language: An Introduction to the Study of Speech*. New York: Harcourt & Brace.

Scantamburlo, L. 1981. *Gramática e dicionário da língua criol da Guiné-Bissau (GCr)*. Bologna: Editrice Missionaria Italiana.

Smith, N. 1999. *Noam Chomsky: Ideas & Ideals*. Port Chester: Cambridge University Press.

Verhaar, J. 1995. *Toward a Reference Grammar of Tok Pisin*. Honolulu: University of Hawai'i Press.

Versteegh, K. 1993. Esperanto as a first language: Language acquisition with a restricted input. *Linguistics* 31-33(325): 539–555.

Veselinova, L. 2003. Suppletion in verb paradigms: Bits and pieces of a puzzle. PhD Dissertation, Department of Linguistics, Stockholm University.