Complexity in numeral systems with an investigation into pidgins and creoles

Harald Hammarström
Chalmers University of Technology

This paper defines and surveys numeral systems from languages across the world. We define the complexity of a numeral system in some detail and give examples of varying complexity from different languages. The examples are chosen to illustrate the bounds on complexity that actually occur in natural languages and to delineate tricky issues of analysis. Then we contrast the complexity in numeral systems of pidgin/creole languages versus their lexifiers and versus languages generally in the world. It turns out that pidgins/creoles have slightly less complex numeral systems than their lexifiers, but probably still more complex than the world average. However, the conclusions in this respect are limited by gaps in documentation and unsystematic knowledge of the linguistic and social history of alleged pidgin/creole languages.

1. Numerals

1.1 What are numerals?

In this paper, I define numerals as follows.[1] They are

1. *spoken*,
2. *normed expressions* that are used to denote the
3. *exact number* of objects for an
4. *open class of objects* in an
5. *open class of social situations* with
6. *the whole speech community* in question.

With the first point I mean to disregard symbol combination systems, e.g., Roman numerals, that are confined to written communication, but of course most (actually all) of our primary data come from written representations of the spoken language.

---

1. Acknowledgements: The author has benefited much from working in a project to write numeral grammars in Grammatical Framework (GF) under Aarne Ranta. Mikael Parkvall and Peter Bakker have provided invaluable help with sorting out pidgins and creoles, but neither should be held accountable for any mischaracterizations in this paper.

The second point serves to exclude expressions that also denote exact numbers, but are not the normal or neutral way to say those numbers, e.g., 'eight-times-nine-and-another-two' for the normal 'seventy-four', but also to demarcate the area where the numeral system ends, which is, when there are no normed expressions.

As for the third point, languages usually have a rich set of expressions for inexact quantities, 'a lot', 'few', 'really many', 'about fifty' (but hardly *'about fifty-one') that have relatively high frequency in discourse. These are interesting in themselves but will not be included here because of their different fuzzy nature compared to exact number expressions.

Concerning the fourth point, some languages have special counting systems for a restricted class of objects (e.g., in Wavulu (Hafford 1999) for counting coconuts). These can be quite idiosyncratic and since all languages which have exact enumeration must have a means for counting an open class of objects it is better to study that.

The reason for the fifth point, the requirement on social situations, is to take a stand on so-called body-tally systems (cf. Laycock 1975; Lean 1992). A body-tally-system may be defined as follows. Assume a sequence of body parts beginning with the fingers of one hand continuing with some points along the lower and upper arm, reaching one or more points of the head, then ending with the corresponding body-parts on the opposite arm and finally hand. A number *n* is then denoted by the *n*th body-part-term in the sequence, e.g., 'nose' or 'elbow on the other side'. Typically, body-tally systems are only used in special circumstances, such as bridal price negotiations, and in other cases one would use a different numeral system or not use exact enumeration at all. The information on the social status of the body-tally numeral systems is very incomplete; for the vast majority we do not have such information, but for those in which we do, the social situation restriction applies. Body-tallying has to be done on a physically present person, and to understand what number is referred to the process must be watched, so, for instance, body-tallying numerals would be infelicitous when it is dark. For instance, de Vries (1998) found that body-tally numerals in a Bible translation could not be understood, i.e., were often mistranslated back to Indonesian by bilingual persons. Of course, there could be some other language(s), unknown to me at present, where body-tally numerals can be used in a fully open class of social situations; such a body-tally system would, accordingly, be included in the study.

Finally, regarding the sixth point, I am not interested in numeral systems which are particular to some small subsets of the speakers of the language in question (e.g., professional mathematicians) because such systems might not respond to the conditions and needs of the majority of a society.

1.2 Why study numerals?

It is true that many languages have very small numeral inventories, that is, words up to two or three and perhaps a possibility to express exact numbers up to at most ten using these and the word for hand (Hammarström, in preparation). But in languages which do not have small numeral inventories, numeral expressions form a system whose properties can be meaningfully studied in terms of complexity.

Numerals provide a good testing bed for patterns across languages given their comparatively clear semantics and modularity. As to numeral semantics, languages may differ as to which quantificational meanings they express/lexicalize, notably in approximate numeration and whether a counted set of objects constitute a group or not, but these matters are minor compared to differences languages show e.g., in verbal tense/aspect. Likewise, although not universally, numerals tend to have uniform, clearly identifiable, syntactic behaviour within a language. Also, if two languages have exact numeration for a certain range of numbers, one expects the two to give a similar functional load to these expressions, excluding possibilities such as numbers also being used for say colours or as metaphors significantly wider in one language or the other. This appears sound also in the light of the only corpus study of numeral frequencies in a language with a small numeral system (McGregor 2004: 204), which shows that 'one' and 'two' in Gooniyandi occur with comparable frequency to 'one' and 'two' in English.

Also, lots of data are available in one form or another for numerals. It seems that numerals together with pronouns, kinship terms, body part terms, and other basic vocabulary (sun, water, etc), and perhaps "sketchy" phonological inventory, are the parts of language where there exists empirical data for a really large subset of the world's known languages. One may legitimately ask just how large this subset is when it comes to numerals – for how many languages do we have data on numerals? Let us say we count about 7000 attested native spoken languages for the world. A definite lower bound is 2500, since I can produce a list of references to numeral data from 2500 definitely distinct languages. An upper bound is harder to give. I entertain the rather time-consuming methodology of trying to obtain every first-hand descriptive data reference found in any handbook or relevant publication whatsoever. I currently have about 5000 such items, some describing numeral systems of many languages in the same publication, but it is impossible to say at this point how many languages they account for since they include dialectal varieties, varieties from the same location but different centuries, partial data, data of varying quality, duplicated data, etc. I also have about a 1000 more references that I have not yet been able to obtain (which may contain further references).

2. Complexity

The following characterization and prevalence of complexity in numeral systems is based on inspection of the above mentioned set of data.

Basically I subscribe to the idea of measuring the complexity of a numeral system by the amount of information necessary to describe the forms. This clearly depends on the scope and flexibility of the means of description (the *description language*). From a computer science perspective a maximally expressive description language is a Turing machine, and it turns out that the minimum-length-description of an object can be meaningfully defined, which, up to a constant factor, is only a property of the object itself and not restricted by the poverty of the description language. The size of this minimal description of an arbitrary object, represented as a string of binary symbols, is called its *Kolmogorov*

Complexity (Vitányi and Li 1997). However, there are several reasons why we will not use Kolmogorov complexity here; in general, Kolmogorov complexity is not computable, that is, it can be proven that there is no one algorithm that will find the minimum-description for *all* possible objects (this does not contradict it being well-defined). Further, Kolmogorov complexity only gives valuable insights on asymptotic behaviour, which is not really relevant for particular finite natural language numeral systems and the level of formality required is completely foreign to traditional linguistic analysis.

Instead, I will look at complexity, i.e., minimum description length, on a level more familiar to computational linguistic analysis, namely where the description language is any standard phrase-structure-based grammar formalism and the objects to be described are strings of phonemes. I will freely ignore (morpho-)phonological alterations and other properties of the numeral forms that belong to the language as a whole rather than the numerals in particular. I will only sketch structural properties to hint how they must be described, and no exact computations of complexity will be given. The reason for this laxness is that I will really only be interested in relative complexity, e.g., if such and such creole language is less complex than such and such non-creole, so it does not really matter how I treat the details as long as I vow to treat them uniformly, but in an unspecified way.

2.1 Complexity as irregularity

We will begin with the following vague formulation of irregularity and make it more precise with examples:

> The form of a numeral is not systematically predictable from the forms of its mathematical parts and knowledge of the rest of the language.

For example, in Russian (Comrie 1992) shown in Table 1, 40 is irregular.

Table 1. The forms for the tens in Russian

10	ДЕСЯТЬ	desyat’	60	ШЕСТЬДЕСЯТЬ	šest’desyat’
20	ДВАДЦАТЬ	dvadcat’	70	СЕМЬДЕСЯТЬ	sem’desyat’
30	ТРИДЦАТЬ	tridcat’	80	ВОСЕМЬДЕСЯТЬ	vosem’desyat’
40	СОРОК	sorok	90	ДЕВЯНОСТО	devyanosto
50	ПЯТЬДЕСЯТЬ	pyat’desyat’	100	СТО	sto

Many languages have irregularities, specific to their numeral forms, that can be captured by a (sub-)rule. That is, we have several irregular forms that can be captured by one rule that is not subsumed by some greater rule. For example, many modern Indo-Aryan languages form all of 19, 29, . . ., 89 by subtraction as 1–20, 1–30, . . ., 1–90 rather than the usual additive pattern for the other numbers 11–99 (Berger 1992). Another example, Camus (a Maa Dialect in Kenya) (Heine 1980: 110–111) in Table 2, has idiosyncratic forms for the tens up to 50, after which formation can be captured

by a rule. In general, I will count complexity as proportional to the number of rules necessary, where one exception counts as much as one new rule.

Table 2. Formation of tens in Camus

1	-bô	10	tomon
2	-aré	20	tíkítam
3	-uní	30	osom
4	-oŋwán	40	ártam
5	ímíét	50	onom
6	Ile	60	n-tomon-i ile
7	sapa	70	n-tomon-i sapa
8	isiet	80	n-tomon-i isiet
9	saal	90	n-tomon-i saal

It follows that the most complex languages are those that have the largest number of irregularities. For example, in Panjabi (Shackle 2003: 602), as shown in Table 3, the numbers 1–99 (although easily etymologizable) show no consistent pattern in formation and have to be learned more or less by rote.

Table 3. Panjabi words 1–100

1	Ik	21	Ikki	41	Iktaḷi	61	Ikáṭh	81	Ikasi
2	do	22	bai	42	bətaḷi	62	báṭh	82	bIasi
3	tIn	23	tei	43	tətaḷi	63	təréṭh	83	tIrasi
4	car	24	cə́vvi	44	cUtaḷi	64	cə̃́ṭh	84	cUrasi
5	pəñj	25	pə́ñji	45	pəñjtaḷi	65	pɛ̃́ṭh	85	pəñjasi
6	che	26	chəbbi	46	chIaḷi	66	chIáṭh	86	chIasi
7	sətt	27	sətai	47	səntaḷi	67	sətáṭh	87	sətasi
8	əṭṭh	28	əṭhai	48	əṭhtaḷi	68	əṭháṭh	88	əṭhasi
9	nɔ	29	Unətti	49	Unəñja	69	Unəttər	89	Unanvẽ
10	dəs	30	tí	50	pə́ñjá	70	səttər	90	nabbe
11	gIaraã	31	Ikətti	51	Ikvəñjá	71	Ikə̀ttər	91	Ikanvẽ
12	barã	32	bətti	52	bəvəñja	72	Bə̀ttər	92	banvẽ
13	terã	33	tẽti	53	tərvəñja	73	tIə̀ttər	93	tIranvẽ
14	cədã	34	cə̃ti	54	cUrvəñja	74	cUə̀ttər	94	cUranvẽ
15	pəndrã	35	pẽti	55	pəcvəñja	75	pəñjə̀ttər	95	pəcanvẽ
16	soḷã	36	chətti	56	chIvəñja	76	chIə̀ttər	96	chIanvẽ
17	sətarã	37	sẽti	57	sətvəñja	77	sətə̀ttər	97	sətanvẽ
18	əṭharã	38	əṭhətti	58	əṭhvəñja	78	əṭhə̀ttər	98	əṭhanvẽ
19	Unni	39	Untaḷi	59	Unáṭh	79	Unasi	99	nəṛInvẽ
20	ví	40	caḷi	60	səṭṭh	80	əssi	100	sɔ

It is readily seen that the regularity of a numeral system is strongly connected to the concept of base. The set of bases of a natural language numeral system may be defined as follows.

The number *n* is a base iff

1. the next higher base (or the end of the normed expressions) is a multiple of *n*; and
2. a proper majority of the expressions for numbers between *n* and the next higher base are formed by (a single) addition or subtraction of *n* or a multiple of *n* with expressions for numbers smaller than *n*.

This assumes that for any expression the linguist can unambiguously analyse each numeral expression into its constituent parts (or analyse it as consisting of only one part). As an example, for Swedish we would begin by finding the biggest part of the highest normed expression, which according to my own knowledge is *miljard* (10^9). Thereafter we can find the next lower base by trying divisors x of 10^9 to see if the numbers between x and 10^9 are expressed in the required form. E.g., $x = 5 \times 10^8$ is not because we do not say **en-halv-miljard plus ett* 'half-a-billion plus one' or the like for $5 \times 10^8 + 1$ or any, let alone a majority, of the numbers between 5×10^8 and 5×10^9. However, *miljon* (10^6) fulfils the requirements and we can easily arrive at the conclusion that Swedish has $\{10, 10^2, 10^3, 10^6, 10^9\}$ as its set of bases.

The definition of base as stated gives unambiguous decisions for formations which are sometimes (and sometimes not) called base by other authors; systematic subtractions, special lexemes for base-multiples, or isolated cases of addition, e.g., only 7 = 6 + 1 but otherwise no additions involving 6. Examples of such cases and their systematic resolution with my definition are given in Table 4.

Once the bases are known we can compactly describe each numeral expression in the multiplicative-additive form as:

$$a_n b_n + a_{n-1} b_{n-1} + \ldots + a_1 b_1 + a_0$$

Where b_i make up the set of bases, $a_i < b_i$ for all i, and $b_i > b_j$ if $i > j$. The a_i: s can be called coefficients and are uniquely determined for each numeral expression given the set of bases. This essentially matches all natural language numeral systems, since a non-small natural language numeral system which does not use bases is not known (the Appendix shows that a numeral system without base is logically possible). Moreover, the bases are always expressed as overt morphemes, in particular, they are never coded as place-values (an alternative way to say 11–19, . . ., 90–99 in spoken Samoan comes the closest (Mosel and Hovdhaugen 1992)). No further generalization over the formation of the b_i :s is possible because no system is known that uses exponentiation.[2] Occasionally it is a non-forced analysis to claim that $a_n < b_n$ for the highest base b_n. Pluses, and sometimes minuses, are expressed overtly or covertly and often different markings are used for different pluses between different bases.

---

2. Occasionally one may see e.g., 100 as 'big-ten' or so which may be called exponentiation, but still no case is known where this extends beyond one single form. Also, the Greek-derived words for 10^{15} and above in a lot of European languages are not (yet) common to the whole speech community.

Table 4. Examples of formation types and outcomes of the definition of base.

	Lutuami (Klamath-Modoc) (Dixon and Kroeber 1907: 673)		Nyokon (Niger-Congo) (Richardson 1957: 30)		Kare (Niger-Congo) (Dijkmans 1974: 147)		Ainu (Isolate) (Refsing 1986: 110)	
	Analysis	Expression	Analysis	Expression	Analysis	Expression	Analysis	Expression
1	1	nas	1	ámɔ̀	1	emotí	1	sine
2	2	lap	2	àfɔ́ɔ̀	2	ibili	2	tu
3	3	ndan	3	átár	3	etotu	3	re
4	4	umit	4	ɲnnɨs	4	biu	4	ine
5	5	tunip	5	ɲtɔ́ɔ̀r	5	etano	5	asikne
6	5 + 1	nas-ksapt	6	átʃɲ́n	5 + 1	etano na emoti	10 – 4	iwan
7	5 + 2	lap-ksapt	6 + 1	ɲtʃɲ́n námɔ̀	5 + 2	etano na ibili	10 – 3	arwan
8	5 + 3	ndan-ksapt	?	ɲyáá nɨ̀ màn	5 + 3	etano na etotu	10 – 2	tupesan
9	10 – 1	nas-xept	8 + 1	ɲyáá nɨ̀ màn námɔ̀	5 + 4	etano na bînu	10 – 1	sinepesan
10	10	te-unip	10	àwát	10	la-ato	10	wan
11	10 + 1	taunep-anta nas	10 + 1	àwát ámɔ̀	10 + 1	laäto na emoti	10 + 1	sine ikasma wan
…	…	…	…	…	…	…	…	…
15	…	…	…	…	15	sanga	…	…
16	…	…	…	…	15 + 1	sanga-na-emoti	…	…
…	…	…	…	…	…	…	…	…
20	2 × 10	lap-eni taunep	20	nɨ̀tʃɲ́n	2 × 10	atumbili	20	hot
21	2 × 10 + 1	lap-eni taunep-anta nas	20 + 1	nɨ̀tʃɲ́n ámɔ̀	…	…	20 + 1	sine ikasma hot
…	…	…	…	…	…	…	…	…
30	3 × 10	nda-ni taunep	3 × 10	àwát átár	2 × 10 + 10	atumbili na laato	20+10	wan e tu hot
…	…	…	…	…	…	…	…	…
40	…	…	…	…	…	…	2 × 20	tu hot
Base		5–10		10		5–20		5–10–20

One often sees that languages have irregularities between b_1 and $2 \times b_1$, for example the Spanish teens as shown in Table 5. Perhaps more often one sees irregularities in $x \times b_1$, especially $2 \times b_1$, particularly in the languages of Eurasia (e.g., Turkish (Lewis 2000) as in Table 5). Spanish and related varieties are virtually unique in having an idiosyncracy as high as in 500 – *quinientos* rather than the regular **cincocientos*.

Table 5. Irregularities in Spanish teens (left) and Turkish tens (right)

1	uno	11	once	1	bir	10	on
2	dos	12	doce	2	iki	20	yirmi
3	tres	13	trece	3	üç	30	otuz
4	cuatro	14	catorce	4	dört	40	kırk
5	cinco	15	quince	5	beş	50	elli
6	seis	16	dieciséis	6	altı	60	altmış
7	siete	17	diecisiete	7	yedi	70	yetmiş
8	ocho	18	dieciocho	8	sekiz	80	seksen
9	nueve	19	diecinueve	9	dokuz	90	doksan
10	diez	20	veinte	10	on	100	yüz

Generally Indo-European languages tend to have more irregularities in their numeral system than the impressionistic world average, perhaps culminating in Modern Indo-Aryan languages, e.g., Panjabi as above. Welsh (King 1993), though there are varieties which have restructured and switched to base 10, is another widely known case of a numeral system with many idiosyncracies.

It follows from the multiplicative-additive form that the most regular system is one where any plus is always expressed the same way and there are no other irregularities. Such a system is evidenced in e.g., Yamba (a Grassfields language of Cameroon) (Lucia 2001).

2.2 Complexity as global ordering constraints

If one assumes a description language using phrase-structure rules, which many people do, then not all rules are equally complex. Although completely regular, some phenomena require a more elaborate phrase structure grammar or even some more expressive description language. These complexities may be characterized as follows:

> The formation rule of a subpart of a numeral expression depends on the rest of the numeral expression.

Recall the multiplicative-additive expression decomposition from Section 2.1. We are now going to discuss the kind of complexity where a particular plus or a particular multiplication is expressed differently depending on the context in which it appears.

There are rare examples, e.g., Guajiro (Zubiri and Jusayú 1986), Kikongo (Söderberg and Widman 1966) and Breton (Press 1986), of languages with discontinuous numerals, i.e., when used attributively, they surround the noun that is quantified. This may also lead to cases where a numeral expression appears discontinuously within a larger numeral expression, such as the Breton 64 in (1).

(1) *pevar mil ha tri-ugent*
four thousand on three-twenty
'64 000'

Some languages, e.g., Erromangan (Crowley 1998), that have N Num order while still having big-before-small order between additive constituents get a potential ambiguity in phrases like 1000 8 which can then mean either 8000 or 1008.[3] Perhaps from pressure to resolve the ambiguity more clearly than by intonation alone, some languages have evolved reordering possibilities that depend on the diagnosed ambiguity of the whole resulting expression. Let's look at an example from Swahili which has N Num order generally. When a composite base multiplier and a rest are present, one can unambiguously use the order $b_i\, a_i + r$ (2).

(2) *elfu sabini na nane tatu*
thousand seventy and/with eight three
'78 003'

When the rest is not present the remaining expression may be ambiguous (3).

(3) *elfu sabini na nane*
thousand seventy and/with eight
'1 078' or '78 000'

One way to force the base-multiplier reading to put it before the base (4).

(4) *sabini na nane elfu*
seventy and/with eight thousand
'78 000'

This section shows that numeral systems can exhibit forms that are perfectly regular but require a longer description if a phrase-structure grammar is used as the description language.

2.3 Complexity vs. economy

Perhaps in response to a sudden demand for an expanded numeral system, there are languages which have evolved very morpheme-promiscuous number expressions e.g., Murinypata (Walsh 1976) in (5) or Maipure (Zamponi 2003) in (6).[4]

(5) *maŋenumimaŋenumimenumimenumimaŋenuminumi*
'hand-one-hand-one-foot-one-foot-one-hand-one-one'
'26'

3. N Num order is quite common (Dryer1989) whereas Malagasy (Parker 1883) seems to be the only modern language with consistent small-before-big ordering.

4. The abbreviations used in the gloss: CL classifier, NPOSS non-possessed.

(6) *papeta* *janà* *pauria* *capi-ti* *purenà*
one.CL.HUMAN take/follow(?) other hand-NPOSS relative
'one takes one relative from the other hand'
'6'

The forms have norms and are highly regular yet some grammar writers tend to label them "cumbersome" in comparison to the more familiar expressions in European languages. Presumably the alleged cumbersomeness of an expression lies in the duration, number of syllables or number of morphemes. In either case, we may speak of the *economy* of a numeral system as the average length (in duration, syllables or morphemes – I would opt for syllables) of its expressions. In particular, if two numeral systems have the same domain and one has longer expressions for every numeral in the domain, it is less economical. This label economical also preserves the intuition that cumbersomeness is something one wishes to spare. Most of the time grammars do not investigate the cumbersomeness in detail, but there are at least cases, e.g., in Murinypata, where "one native speaker with practically no formal education readily produced the number term for 'one hundred' which consists of seventy syllables" (Walsh 1976: 198).

Less economic numerals like these are uncommon, whereas numerals with about the same level of economy as English are standard. When we see cases where uneconomic numeral expressions are replaced with more economic numerals of another language, it is always the case that the language with the economic numerals is also socio-economically dominant. It is then sufficient to explain this numeral replacement within the general situation, i.e., take-over of the socio-economically dominant language, so a causal link between cumbersomeness (or the like) and abandonment is not necessary to explain these cases.

There appears to be no reason to link uneconomical numeral systems with complex ones. Indeed, the only difference between those and regular ones is the number of phonemes/syllables/morphemes. From a description length perspective, long forms are trivial to compress by code-booking as long as they are regular, so the addition in description length resulting for uneconomical numeral systems is negligible. Therefore, they will not be matter for further discussion.

2.4 Overall complexity

Probably the most complex numeral systems I have observed are those in Modern Indo-Aryan languages, like Panjabi. Breton does not have as many irregular items but probably has the most structural idiosyncracies; it is vigesimal up to about 200 then switches to decimal, it forms 50 and 150 with halves (half-hundred and hundred-half respectively), and has discontinuous formations.

Not counted are some numeral systems with enormous amounts of quite idiosyncratic forms when the difficulty invariably comes from the classifier fusing with the numeral, e.g., Ket (a Yeniseic language of Siberia) (Dul'zon 1968), rather than irregularity in forming the numerals themselves.

As for a "world average", there is neither time nor space to back it up with detailed evidence, but I may give an idea impressionistically. For these considerations, to make the objects comparable, imagine that, for all languages, we cut away all numerals over a hundred, and we leave out entirely languages that don't have numerals up to a hundred. Looking at all languages with a numeral systems up to a hundred this way, the average complexity of these would be similar to that of Camus above, i.e., around 15 forms to be learned by rote, one main formation rule and one formation sub-rule (no global ordering constraints). However, an average obtained from that language set will be less interesting because due to borrowing and inheritance the languages, and thus numeral systems, in question are not independent. If instead we look at the complexity of the subset of independent numeral systems – this time even more impressionistically – the average will come closer to he simplest possible, with around 11 rote forms and only one formation rule, on the same level as Nyokon above.[5]

3. Complexity in numeral systems of pidgin and creole languages

If the characterization of complexity makes sense, it would be highly interesting to see to what extent pidgins, pidgincreoles and creoles (definitions, following Bakker (2006), in Table 6) have the same amount or less complexity than their respective lexifier language. This question is immediately pertinent to the ongoing debate on the "simplicity" of creole languages (see e.g., McWhorter (2001) [also commentaries, pp. 167–412 in the same issue of the journal]). Since the alleged simplicity of creoles is held to be due to an earlier pidgin stage, the investigation of the numeral systems of all varieties of pidgins, pidgincreoles and creoles are relevant to the debate.

Table 6. Definitions of pidgins, pidgincreoles, and creoles

	Pidgins	**Pidgincreoles***	**Creoles**
has norms	yes	yes	yes
reduced from other language(s)	yes	yes	yes
ethnic or political group language	no	no	yes
native language	no	yes/no	yes
main language of speech community	no	no/yes	yes

*A pidgincreole should have a "yes" in at least one of the last two rows.

5. To be more precise, independent means that borrowing and inheritance can be ruled out. An independent set of numeral systems may be reached as follows. Exclude all systems which have morpheme(s) that have cognate(s) in some other family. Of the remaining systems, if two systems share at least one cognate put both of them in the same *chain*. Each chain is now independent from all other chains under the assumption that structural borrowing (metatypy) does not take place without simultaneous borrowing of a morpheme and that the etymology of numerals is generally known. To get an average one may take the average within each chain and then the average of the average of each chain.

In this study I have tried to gather data on numeral systems for all known pidgins, pidgincreoles and creoles, excepting only those whose status as such is doubted. However, a lot of descriptive accounts of pidgin/creole languages do not feature full data on numerals, and in a few more cases the publications containing the numeral data was not accessible to the author (and in an unknown number of cases a publication containing the sought information was simply not known to the author). The resulting set of pidgin/creole numeral systems are given in Table 7 together with a classification into pidgin (p), pidgincreole (pc) or creole (c) – acknowledging, however, that the sociohistorical data available to back up such a classification is quite uneven. To show whether restructuring of the numeral system of the lexifier language has taken place, I have grouped on the language that supplied the numerals. The language that supplied the numerals was usually the same as the lexifier for general vocabulary, but not always; e.g., in Fanakalo the numerals come from English but the general lexifiers are Nguni Bantu languages.

Most lexifiers had complexities in the numerals, allowing a test for the possibility of a simplification in the daughter pidgin/creole. A **C** in the teens/tens column indicates that there is some complexity in the formation, whereas an S indicates transparent formation, i.e., as $10 + x$ or $x \times 10$. No daughter language in our sample invented more complexity than the lexifier, except Kinubi which, unlike Arabic, has multiplier-base order for 10 and 100 but base-multiplier order for multiplications of 1000.

Table 7. Complexity in pidgin, pidgincreole and creole numeral systems

Language	Clf.*	Teens	Tens	Source
French		**C**	**C**	
Tayo	c	C	C	(Ehrhart 1993: 135)
Pointe Coupée	c	C	C	(Klingler 2003: 197–198)
Breaux Bridge	c	C	C	(Neumann 1985: 124)
Seychellois	c	C	C	(Bollée 1977: 39 + App. D)
Mauritian Creole	c	C	C	(Baker 1972: 144–145)
Haitian Creole	c	C	C	(Hall 1953: 29)
St. Lucian Creole	c	C	C	(Carrington 1984: 77–79)
Karipuna do Amapá	c	C	?	(Tobler 1987)
Tay Bôi	p	s	s	(Reinecke 1971: 52)
Arabic		**S**	**C**	
Nubi	c	S	C	(Luffin 2005: 157–167)
Turku	p	S	C	(Prokosch 1986: 100–101)
Portuguese		**C**	**C**	
Sãotomense	c	S	S	(Lorenzino 1998: 107–109)
Principense	c	C	C	(Günther 1973: 63–64)
Papia Kristang	c	C	C	(Baxter 1988: 48–49)
Crioulo Guiné	c	S	C	(Wilson 1962: 16–17)
Guinea-Bissau	c	S/C	C	(Honório do Couto 1994: 99–100)
Kap-Verde	c	C	C	(Veiga 1995: 172–173)
Annobonese	c	?	S	(Schuchardt 1888: 22)

(*Continued*)

Table 7. Continued.

Language	Clf.*	Teens	Tens	Source
Spanish		**C**	**C**	
Papiamentu	c	S	C	(van Name 1870: 154) (Munteanu 1996: 319–320)
Dutch		**C**	**C**	
Sranan	c	S	S	(Braun 2005: 295–307)
Negerhollands	c	C	C	(Oldendorp 1996 [1767–1768]: 58–154)
Ngbandi		**S**	**S**	(Kondangba 1991)
Sango	pc	S	S	(Giraud 1908: 266)
Russian		**C**	**C**	
Russenorsk	p	C	C	(Broch & Jahr 1981: 122)
Assamese		**S**	**S**	(Babakaev 1961)
Naga Pidgin	pc	S	S	(Bhattacharjya 2001: 143–146)
Chinook		**S**	**S**	(Ross 1849: 322–323)
Chinook Jargon	p	S	S	(Holton 2004: 69–70)
Choctaw		**S**	**S**	(Byington 1915)
Mobilian Jargon	p	S	S	(Drechsel 1997: 107)
MacKenzie Inuit		**S**	**S**	(Stefansson 1909: 232)
Eskimo Trade Jargon	p	S	S	(Stefansson 1909: 232)
English		**C**	**C**	
Broken (Torres Str.)	c	C	C	(Shnukal 1988: 252–253)
Saramaccan	c	C	C	(Wullschlägel 1965 [1854]: 14)
Coastal NG Pidgin	pc	S	S	(Laycock 1970: 1)
Highlands NG Pidgin	pc	S/C	S/C	(Wurm 1971: 81–82)
Tok Pisin	pc	S/C	S/C	(Verhaar 1995) (Mosel 1980: 61–63)
Bislama	pc	C	C	(Guy 1975: 26–28)
Solomons Pijin	pc	C	C	(Jourdan 2002)
Fanakalo	p	C	C	(Kaltenbrunner 1996: 92)
Samoan Plantation	p	S	S	(Mühlhäusler 1978: 100–101)
Krio	c	C	C	(Fyle & Jones 1980)
Carriacou	c	C	C	(Kephart 2000: 174–198)
Nigerian Pidgin	pc	C	C	(Faraclas 1996: 231)
Ndyuka	c	C	S	(Huttar & Huttar 1994: 532)
Cantonese Pidgin	p	C	?	(Hall 1944: 97)
Jamaican Creole	c	C	C	(Own Knowledge)
Cameroon Pidgin	pc	S	S	(Parkvall 2000: 107)
Hiri Motu	pc	C	C	(Wurm & Harris 1963)

*Clf. is short for classification; p = pidgin, c = creole, pc = pidgincreole.

A few remarks are in order. An entry like S/C means that the source gives parallel forms. In a few other cases there is a difference between an early and a late set of numerals, i.e., there has been borrowing (rarely internal restructuring) overlaying numerals attested earlier, in which case Table 7 shows only the earliest attested set. In another few more cases one may *suspect* borrowing, often from the lexifier language anew, but where I have found

no attestation of an earlier set, I analysed the attested forms – be they original or not. A couple of languages are curious in that they have numerals from two source languages. It was always possible to decide on a "main" lexifier for Table 7, but I could not discern whether the multi-source situations were original or the result after some borrowing.

3.1 Discussion

Whereas some pidgins/creoles do tend to analyticity, a majority do not. Parkvall (2000) shows similar findings for African creoles specifically. If it were necessary that pidgins, and/or languages descended from pidgins, have a maximally simple structure, we would have seen quite different empirical results. The prime example is Naga Pidgin which has a numeral system of the same complexity as Panjabi. However, it is also true that pidgins/creoles have slightly less complex numerals relative to their lexifiers. The same lexifier needs not produce the same result in its daughter pidgins/creoles. It appears that we can not predict where restructuring is more likely to take place if it takes place at all; for example Papiamentu restructures 500 to a regular formation, whereas the parallel Kap Verde and Guinea Bissau cases do not.

Impressionistically, the pidgin/creole numeral systems are on the average more complex than the world average, both if we count all systems or only independent systems. So it is not possible to look only at a numeral system and say whether it is from a pidgin/creole language or not. This appears to be easily explained by the fact that well-documented pidgins/creoles have a set of lexifiers which is non-representative of the (documented) languages of the world as a whole.

Furthermore, unless they borrow, languages which change from having only a few lexical numerals to a combinatorial system of numerals, universally do this by forming a 5-10-20 system with transparent formations. Pidgins do not follow this path even though they undoubtedly have the means sufficient to do so, i.e., juxtaposition and words for 'and', 'hand', 'foot' and 'man'.

Why is the prediction that pidgin/creole languages should be (maximally?) simple not borne out as regards numeral systems? The answer is obscure to me as lack of sociohistorical data prevent the mechanisms behind the prediction from being fully scrutinized.

References

Babakaev, V.D. 1961. *Assamskij Jazyk* [Jazyki Zarubezhnogo Vostoka i Afriki]. Moskva: Akademia Nauk SSSR.

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

Bakker, P. 2006. Pidgins versus other contact languages. In *Handbook of Pidgins and Creoles*, S. Kouwenberg & J.V. Singler (eds). Oxford: Blackwell.

Baxter, A.N. 1988. *A Grammar of Kristang (Malacca Creole Portuguese)* [Pacific Linguistics B 95]. Canberra: Australian National University.

Berger, H. 1992. Modern Indo-Aryan. In *Indo-European Numerals* [Trends in Linguistics. Studies and Monographs 57], J. Gvozdanovic (ed.), 243–288. Berlin: Mouton de Gruyter.

Bhattacharjya, D. 2001. *The Genesis and Development of Nagamese: Its Social History and Linguistic Structure*. PhD Dissertation, City University of New York.

Bollée, A. 1977. *Le créole français des Seychelles: Esquisse d'une grammaire - textes - vocabulaire* [Beihefte zur Zeitschrift für romanische Philologie 159]. Tübingen: Niemeyer.

Braun, M. 2005. *Word-Formation and Creolisation: The Case of Early Sranan*. PhD Dissertation, Universität Siegen.

Broch, I. & Jahr, E. 1981. *Russenorsk – et Pidginspråk i Norge* [Tromsø-Studier i Språkvitenskap III]. Oslo: Novus.

Byington, C. 1915. *A Dictionary of the Choctaw Language* [Bureau of American Ethnology Bulletin 46]. Washington: Smithsonian Institution.

Carrington, L.D. 1984. *St. Lucian Creole: A Descriptive Analysis of its Phonology and Morpho-Syntax* [Kreolische Bibliothek 6]. Hamburg: Helmut Buske Verlag.

Comrie, B. 1992. Balto-Slavonic. In *Indo-European Numerals* [Trends in Linguistics. Studies and Monographs 57], J. Gvozdanović (ed.), 717–833. Berlin: Mouton de Gruyter.

Crowley, T. 1998. *An Erromangan (Sye) Grammar* [Oceanic Linguistics Special Publication 27]. Honolulu: University of Hawaii Press.

de Vries, L.J. 1998. Body part tally counting and bible translation in Papua-New Guinea and Irian Jaya. *The Bible Translator (Practical Papers)* 49(4): 409–415.

Dijkmans, J.J.M. 1974. *Kare-taal: Lijst van woorden gangbaar bij het restvolk Kare opgenomen in de jaren 1927–1947*. Sankt Augustin: Anthropos-Institut – Haus Völker und Culturen.

Dixon, R.B. & Kroeber, A.L. 1907. Numeral systems of the languages of california. *American Anthropologist* 9(4): 663–689.

Drechsel, E.J. 1997. *Mobilian Jargon: Linguistic and Sociohistorical Aspects of a Native American Pidgin* [Oxford Studies in Language Contact]. Oxford: Clarendon.

Dryer, M.S. 1989. Large linguistic areas and language sampling. *Studies in Language* 13(2): 257–292.

Dul'zon, A.P. 1968. *Ketskij Jazyk*. Tomsk: Izdatel'stvo Tomskogo Universiteta.

Ehrhart, S. 1993. *Le Créole Français de St-Louis (Le Tayo) en Nouvelle Calédonie* [Kreolische Bibliothek 10]. Hamburg: Helmut Buske Verlag.

Faraclas, N. 1996. *Nigerian Pidgin* [Descriptive Grammars Series]. London New York: Routledge.

Fyle, C.N. & Jones, E.D. 1980. *A Krio-English Dictionary*. Oxford: Oxford University Press.

Giraud, G. 1908. Vocabulaire des dialectes sango, balkongo, et a-zandé. *Révue Coloniale, Nouvelle Serie* 58: 263–291, 332–354.

Günther, W. 1973. *Das portugiesische Kreolisch der Ilha do Príncipe* [Marburger Studien zur Afrika- und Asienkunde: Serie A 2]. Marburg an der Lahn.

Guy, J.B.M. 1975. *Handbook of Bichelamar/Manuel de Bichelamar* [Pacific Linguistics C 34]. Canberra: Australian National University.

Hafford, J.A. 1999. Elements of Wuvulu grammar. MA Thesis, University of Texas at Arlington.

Hall, Jr., R.A. 1944. Chinese Pidgin English grammar and texts. *Journal of the American Oriental Society* 64(3): 95–113.

Hall, Jr., R.A. 1953. *Haitian Creole: Grammar, Texts, Vocabulary* [Memoirs of the American Folklore Society 43]. The American Anthropological Association.

Hammarström, H. In preparation. Small numeral systems. Ms.

Heine, B. 1980. *The Non-Bantu Languages of Kenya* [Language and Dialect Atlas of Kenya II]. Berlin: Verlag von Dietrich Reimer.

Holton, J. 2004. *Chinook Jargon: The Hidden Language of the Pacific Northwest*. San Leandro, California: Wawa Press.

Honório do Couto, H. 1994. *O Crioulo Português da Guiné-Bissau* [Kreolische Bibliothek 14]. Hamburg: Helmut Buske Verlag.

Huttar, G.L. & Huttar, M.L. 1994. *Ndyuka* [Descriptive Grammars Series]. London New York: Routledge.

Jourdan, C. 2002. *Pijin: A Trilingual Cultural Dictionary* [Pacific Linguistics 526]. Canberra: Australian National University.

Kaltenbrunner, S. 1996. *Fanakalo: Dokumentation einer Pidginsprache* [Beiträge zur Afrikanistik 53 / Veröffentlichungen der Institute für Afrikanistik und ägyptologie der Universität Wien 72]. Vienna: Afro-Pub.

Kephart, R.F. 2000. *"Broken English": The Creole Language of Carriacou* [Studies in Ethnolinguistics 6]. New York: Peter Lang.

King, G. 1993. *Modern Welsh: A Comprehensive Grammar*. London: Routledge.

Klingler, T.A. 2003. *If I Could Turn My Tongue Like That: The Creole Language of Pointe Coupée Parish, Louisiana*. Baton Rouge: Louisiana State University Press.

Kondangba, Y. 1991. Structure des numeraux en bantu (lingɔmbɛ) et en non-bantu (ngbaka, minagende, ngbandi, ngbundu, mɔnɔ, mbanza). *Annales Æquatoria* 12: 307–319.

Laycock, D.C. 1970. *Materials in New Guinea Pidgin (Coastal and Lowlands)* [5*Pacific Linguistics: Series D*]. Canberra: Australian National University.

Laycock, D.C. 1975. Observations on number systems and semantics. In *New Guinea area languages and language study, vol 1: Papuan Languages and the New Guinea linguistic scene* [Pacific Linguistics C 38], S.A. Wurm (ed.), 219–233. Canberra: Australian National University.

Lean, G.A. 1992. *Counting Systems of Papua New Guinea and Oceania*. PhD Dissertation, Papua New Guinea University of Technology.

Lewis, G.L. 2000. *Turkish Grammar* (2nd ed.). Oxford: Oxford University Press.

Lorenzino, G.A. 1998. *The Angolar Creole Portuguese of São Tomé: Its Grammar and Sociolinguistic History* [LINCOM Studies in Pidgin & Creole Languages 1]. München: Lincom Gmbh.

Lucia, N.N. 2001. Yamba: A morphosyntactic study of the basic sentence. MA Thesis, The University of Yaoundé I.

Luffin, X. 2005. *Un créole arabe: Le kinubi de Mombasa, Kenya* [LINCOM Studies in Pidgin & Creole Linguistics 07]. München: Lincom GmbH.

McGregor, W.B. 2004. *The Languages of the Kimberley, Western Australia*. London New York: Routledge.

McWhorter, J. 2001. The world's simplest grammars are creole grammars. *Linguistic Typology* 5(3/4): 125–166.

Mosel, U. 1980. *Tolai and Tok Pisin: The Influence of the Substratum on the Development of New Guinea Pidgin* [Pacific Linguistics B 73]. Canberra: Australian National University.

Mosel, U. & Hovdhaugen, E. 1992. *Samoan Reference Grammar* [Instituttet for sammenlignende kulturforskning 85]. Oslo: Scandinavian University Press.

Mühlhäusler, P. 1978. Samoan Plantation Pidgin English and the origin of New Guinea Pidgin. In *Papers in Pidgin and Creole Linguistics 1* [Pacific Linguistics A 54], 67–120. Canberra: Australian National University.

Munteanu, D. 1996. *El Papiamento, lengua criolla Hispánica* [Biblioteca Románica Hispánica: Tratados y Monografías 17]. Madrid: Editorial Gredos.

Neumann, I. 1985. *Le Créole de Breaux Bridge, Louisiane* [Kreolische Bibliothek 7]. Hamburg: Helmut Buske Verlag.
Oldendorp, C.G.A. 1996 [1767–1768]. *Criolisches Wörterbuch* [Lexicographica: Series Maior]. Tübingen: Max Niemeyer.
Parker, G.W. 1883. *Concise Grammar of the Malagasy Language*. London: Trübner & Co.
Parkvall, M. 2000. *Out of Africa: African Influences in Atlantic Creoles*. London: Battlebridge Publications.
Press, I. 1986. *A Grammar of Modern Breton* [Mouton Grammar Library 2]. Berlin: Mouton de Gruyter.
Prokosch, E. 1986. *Arabische Kontaktsprachen (Pidgin- und Kreolsprachen) in Afrika* [Grazer Linguistische Monographien 2]. Graz: Institut für Sprachwissenschaft der Universität Graz.
Refsing, K. 1986. *The Ainu Language: The Morphology and Syntax of the Shizunai Dialect*. Aarhus: Aarhus University Press.
Reinecke, J.E. 1971. Tay bôi: Notes on the Pidgin French of Vietnam. In *Pidginization and Creolization of Languages: Proceedings of a Conference Held at the University of the West Indies, Mona, Jamaica, April 1968*, D. Hymes (ed.), 47–56. Cambridge: Cambridge University Press.
Richardson, I. 1957. *Linguistic Survey of the Northern Bantu Borderland* [Linguistic Survey of the Northern Bantu Borderland 2]. Oxford: Oxford University Press.
Ross, A. 1849. *Adventures of the First Settlers on the Oregon or Columbia River*. London: Smith, Elder & Co.
Schuchardt, H. 1888. *Ueber das Negerportugiesische von Annobom* [Kreolische Studien VII]. Wien: Carl Gerold's Sohn.
Shackle, C. 2003. Panjabi. In *The Indo-Aryan Languages* [Routledge Language Family Series], G. Cardona and D. Jain (eds), 581–621. London/New York: Routledge.
Shnukal, A. 1988. *Broken: An Introduction to the Creole Language of Torres Strait* [Pacific Linguistics C 107]. Canberra: Australian National University.
Söderberg, B. & Widman, R. 1966. *Kikongo*. Stockholm: Svenska Bokförlaget Bonniers.
Stefansson, V. 1909. The Eskimo trade jargon of Herschel Island. *American Anthropologist* 11(2): 217–232.
Tobler, A.W. 1987. *Dicionário crioulo karipúna/português, português/crioulo karípúna*. Brasília: Summer Institute of Linguistics.
van Name, A. 1869–1870. Contribution to creole grammar. *Transactions of the American Philological Association* 1: 123–167.
Veiga, M. 1995. *O Crioulo de Cabo Verde: Introdução à Gramática* (2nd ed.). Praia: Instituto Caboverdiano do Livro e do Disco, Instituto Nacional da Cultura.
Verhaar, J.W.M. (ed.) 1995. *Toward a Reference Grammar of Tok Pisin: An Experiment in Corpus Linguistics* [Oceanic Linguistics Special Publication 26]. Honolulu: Univeristy of Hawai'i Press.
Vitányi, P. & Li, M. 1997. *Kolmogorov Complexity.* (2nd ed.). Berlin: Springer-Verlag.
Walsh, M.J. 1976. *The Murinypata Language of North–West Australia*. PhD Dissertation, Australian National University.
Wilson, W.A.A. 1962. *The Crioulo of Guiné*. Johannesburg: Witwatersrand University Press.
Wullschlägel, H.R. 1965 [1854]. *Kurzgefasste Neger-Englische Grammatik*. Amsterdam: S. Emmering.
Wurm, S.A. 1971. *New Guinea Highlands Pidgin: Course Materials* [Pacific Linguistics D 3]. Canberra: Australian National University.

Wurm, S.A. & Harris, J.B. 1963. *Police Motu: An Introduction to the Trade Language of Papua (New Guinea) for Anthropologists and Other Fieldworkers* [Pacific Linguistics B 1]. Canberra: Linguistic Circle of Canberra Publications.

Zamponi, R. 2003. *Maipure* [Languages of the World/Materials 192]. München: Lincom GmbH.

Zubiri, J.O. & Jusayú, M.A. 1986. *Gramática de la Lengua Guajira (Morphosintaxis)*. San Cristóbal: Universidad Católica del Tachira.

Appendix: Logically possible numeral systems without base

By a numeral system I mean a finite set of atoms used combinatorially to denote each member of a serially ordered target set. If the set of atoms has cardinality n, each combinatorial expression may not be longer than $2n$, and the target set must have cardinality of at least 2^n.

There are several ways in which one can have a numeral system without a base (as defined in section 4). I will sketch a few examples here.

Example 1. Three Alterating Bases: Let the set of atoms be A = {1,2,3,4,5,6,7,8,9, 10,11,12,100,110,120,10000,11000,12000}. Given a number n to be expressed, decompose $n = pq + r$ such that $p < q$ and $q \epsilon A$ is maximal. Then express n as "p × q + r" with possible recursion until $p \epsilon A$. So e.g., 31 as 3 × 10 + 1, 34 as 3 × 11 + 1, 121 as 120 + 1, 778 as 7 × 110+8,132012001 as 11000 + 1 × 12000 + 1 and so on.

Example 2. Decomposition into Primes: The fundamental theorem of arithmetic says that every number n can be written as a product of primes $p_1^{e_1} \cdots p_n^{e_n}$. Thus we can have the primes as our set of atoms and express any $n > 1$ as: E(n) = $p_1(E(e_1))p_2(E(e_2))\ldots p_n(E(e_n))$. Of course, with E(1) = 1.

Example 3. Increasing Gaps: Instead of letting counting begin anew at uniform intervals we can have some more complex evolution of intervals. For example, instead of re-counting at 10, 20, . . ., 100 etc we can increase the gap size at each step, e.g., 10, 21, 33, and so on.

Example 4. Permutations: For a completely non-transparent counting system we might use permutations. With the set of atoms 1–9 we can form the set of permutations of the numbers 1–9. This set can be ordered using the number obtained by reading the permutation as a place-value number expression. So e.g., 123456789 is the smallest, and would be used to represent 1, 123456798 is 2 and so on.

Example 5. Subsets: Also using the atoms 1–9 we can denote numbers by subsets of 1–9. An ordering of the subsets that does not yield the existence of bases is the following. Each pair of subsets of the same cardinality can be compared in terms of what I shall call "smallness" – the sum of its members, and if that is a tie the set with the smallest member that is not present in the other. Now, to map subsets to numbers, first take the smallest one-member subset, then the smallest two-member subset, . . ., nine-member subset, the next smallest one-member subset and iterating so on until all the sizes of subsets are exhausted. This will yield, in increasing order, {1} {1,2} {1,2,3} {1,2,3,4} . . . {1,2,. . . ,9} {2} {1,3} {1,2,4} {1,2,3,5} {1,2,3,4,6} . . . .