# Paradigmatic complexity in pidgins and creoles


## Abstract

The last decade has seen increasing attention paid to questions of grammatical complexity, in particular regarding the extent to which some languages can be said to be more "complex" than others, whether globally or with respect to particular subsystems. Creoles have featured prominently in these debates, with various authors arguing that they are particularly simple when set against non-creoles, with an apparent lack of overt morphology in creoles often cited as one of the ways in which their grammars are especially simplified. This paper makes two contributions to this discussion. First, it develops metrics of grammatical complexity that derive directly from a well-known model of creole development, thus providing an explicit link between the sociohistorical circumstances in which creoles formed and grammatical outcomes. Second, it applies these metrics to the newly published dataset from the Atlas of Pidgin and Creole Structures, setting this data against that from the well-known World Atlas of Language Structures, allowing for a more comprehensive and rigorous quantitative comparison of complexity in contact and non-contact languages than has been previously been possible. It will be seen that there is good evidence that contact languages are simplified overall with respect to a class of complexities labelled *paradigmatic* here but that this general conclusion nevertheless masks significant underlying variation among them.



**Keywords:** creoles, complexity, typology, paradigmatic, syntagmatic

# Paradigmatic complexity in pidgins and creoles[1]

Jeff Good

jcgood@buffalo.edu

University at Buffalo

## 1 Introduction

The last decade has seen increasing attention paid to questions of grammatical complexity, in particular regarding the extent to which some languages can be said to be more "complex" than others, whether globally or with respect to particular subsystems (see, e.g., Miestamo et al. (2008) and Sampson et al. (2009) for relatively recent volumes on the topic and Sinnemäki (2011: 8–36) for a review). This issue has been of particular interest in studies of creole language typology, largely in the context of an overarching concern regarding whether or not creoles may be especially "simple" when set against non-creole languages. Much of the recent discussion has been sparked by the work of McWhorter (2001), though it has been continued by others, for instance Parkvall (2008), which extends results based largely on anecdotal observation by using quantitative metrics applied to the large-scale typological database of the the World Atlas of Language Structures (Dryer & Haspelmath 2013).[2]

Good (2012) contributes to this discussion by proposing that, if creoles are indeed "simple" in some sense, there is no reason to expect them to be globally simple. Rather, their patterns of simplification should directly reflect their sociohistory. Specifically, their origins as languages formed in sociolinguistic contexts characterized by a transmission "bottleneck" should differentially impact complexities defined in paradigmatic terms over syntagmatic ones. The discussion in that

[1] I would like to acknowledge Martin Haspelmath and Susanne Maria Michaelis for inviting me to speak the workshop *Creole and pidgin language structure in cross-linguistic perspective* where the main results of this paper were originally presented and to thank participants at that workshop for their input on this work. I would also like to thank Robert Forkel for providing me with a version of the APiCS data that facilitated the analysis presented below and an anonymous review for providing useful feedback.

[2] This line of research is hardly uncontroversial (see, e.g., DeGraff (2005)). I do not attempt to summarize these controversies here since they are amply discussed elsewhere. A good sense for their contours can be found in three recently published papers from the same issue of the *Journal of Pidgin and Creole Languages* (Mufwene 2014; McWhorter 2014; Bakker 2014).

paper was largely programmatic, with supporting data that was, in some sense, "cherry-picked". However, with the recent publication of the Atlas of Pidgin and Creole Structures online (APiCS; Michaelis et al. (2013d)), it is now possible to examine the paper's claims more systematically.[3] Of particular importance is the fact that a subset of the APiCS database is designed to be directly comparable to the content of the World Atlas of Language Structures (WALS) (Dryer & Haspelmath 2013). Thus, APiCS languages can be rigorously compared with other languages in assessing their typology.

The purpose of this paper is, therefore, to re-consider the claims of Good (2012) on the basis of this newly available data. It is part of a broader trend of work such as Parkvall (2008), Bakker et al. (2011), and Siegel et al. (2014), where quantitative metrics are used to explore aspects of creole typology in ways that would have been quite difficult until recently. While this paper does not only consider morphological aspects of creole typology, it should be emphasized that much of the work on simplicity and complexity in creoles has focused specifically on the structure of creole words, with the frequent claim that creoles are relatively simple in morphological terms (see, e.g., Siegel (2004); Plag (2008; 2009)). Therefore, any general examination of creole typology is also likely to be relevant to questions of creole morphology, and this will be made clear at varying points in the discussion below and, in particular, in Section 4.5.

In Section 2, I first give a summary of the claims made in Good (2012) in order to provide context for the later discussion. The APiCS and WALS data that will be employed to compare complexities across creoles and non-creoles is introduced in Section 3, with a particular emphasis on how it can be recoded to reflect a distinction between paradigmatic and syntagmatic complexities that will be of interest here. The recoded data is then quantitatively explored in Section 4, with a focus on whether or not there is an asymmetry in paradigmatic complexity and syntagmatic complexity in creoles, as compared to non-creoles, and consideration of the importance of mor-

[3] The online version of APiCS is also associated with printed volumes (e.g., (Michaelis et al. 2013a;b)). It should be emphasized here that APiCS online is an edited database with contributions from a large number of different authors, referred to as the *APiCS Consortium* (see Michaelis et al. (2013a: xxxii–xxxiii) for a list of its members as well as http://apics-online.info/contributors). Since the present work is making use of the database in general, their individual contributions are not specifically cited here except where this is relevant to a particular point of discussion, but it should be acknowledged that the present paper would not be possible without their significant efforts.

phological paradigms, in particular, in establishing any such asymmetry. A brief conclusion is given in Section 5.

## 2 Typologizing creole complexities

In this section, I summarize the main ideas presented in Good (2012), with a focus on its assumed model of creolization and its predictions regarding creole complexity. The basic approach of Good (2012) is largely deductive in nature: It considers the sorts of grammars that are likely to emerge from a process of creolization that assumes the existence of a "bottleneck" in the transmission of grammatical material from source languages into an emerging contact variety and then assesses the extent to which attested creole grammars match the predictions of that model. As mentioned above, however, the predictions were tested through the examination of exemplary case studies rather than against a systematically collected sample, which is the focus of the present paper.

In Figure 1, the model assumed by Good (2012: 5) is schematized. Crucially, it is not a model of an emerging contact variety in terms of the mental states of speakers, as has been more typical in studies of creolization. The Interlanguage Hypothesis (Plag 2008) presents one relatively recent example, and Bickerton's (1984) Language Bioprogram Hypothesis (see Veenstra (2008) for overview discussion) is, perhaps, the most well-known such model. Rather, it is intended to shed light on the ways in which a shared lexicogrammatical code develops in a community that would come to be associated with a creole.

Creolization is heuristically modeled in Figure 1 in terms of three stages: a jargon stage, a pidgin stage, and a creole stage. The terms *jargon*, *pidgin*, and *creole*, of course, have some flexibility in their usage. Of these terms, the sense of jargon is the most important to make explicit here since, as indicated in Figure 1, this is assumed to be the stage where the special patterns of simplification associated with creolization are taken to occur. By *jargon*, I refer to a stage "in which people experiment with forms and structure, before any norms establish" (Bakker 2003: 4).[4] By contrast, I use the term pidgin to refer to a stage where a contact variety has become largely

[4] This stage is recognized elsewhere, under different terms, such as the Stage 1 pidgin of Winford (2006: 296–298).

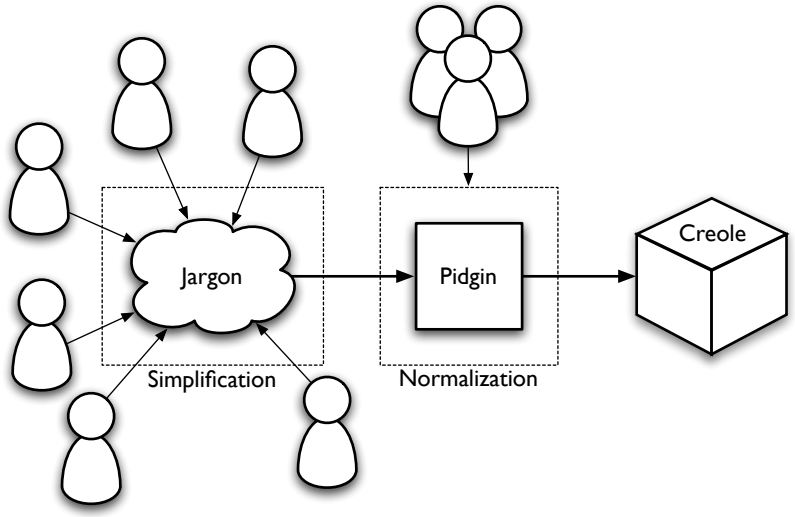


Figure 1: Schematization of structural stages of creolization

normalized, if not a full-fledged language. That is, unlike in a jargon, it is possible for an utterance to be correctly interpreted but still be considered, in some sense, to be the "wrong" way to say something by members of the community speaking the relevant pidgin. A creole in this context is understood as a lexicogrammatical code with the same level of expressivity as any other language, where canonical creoles are additionally the primary language of a well-defined community. The norming associated with a pidgin is depicted by placing it within a square and associating it with a defined group of speakers, as set against the jargon, which is less structured both linguistically and socially. The creole is depicted as a cube to reflect its richer expressivity when set against a pidgin.

The model presented in Figure 1 is not uncontroversial, as discussed in Good (2012: 3) (see also Mufwene (2014: 150)), and I do not mean to present it as such here. However, clearly, to the extent that testing it against a broader data set than was previously feasible may yield results that are consistent with it, it becomes more plausible. Since this is the primary goal of the present paper, I do not explore the model's controversial aspects in detail here, though I will return briefly to the issue of competing models in Section 5.

As discussed in Good (2012: 3–6), if we want to understand what special patterns of simplification, if any, that we might find in creoles, it is necessary to break down the process of creolization into subprocesses and determine whether any of these subprocesses might be prone to loss of

grammatical complexities. In this regard, the heuristic states depicted in Figure 1 are less important than the processes that produce them, with the claim of Good (2012) being that the process through which a jargon is formed, i.e., *jargonization*, results in a specific kind of grammatical simplification taking place. This type of simplification is referred to here as *paradigmatic simplification* and it builds on the distinction between paradigmatic and syntagmatic complexity given in (1), as found in Good (2012: 9), who, in turn, builds on Moravcsik & Wirth (1986: 7).[5]

(1) a. **Syntagmatic complexity:** Complexity deriving from the partonomic/meronymic structure of a given linguistic construction.

b. **Paradigmatic complexity:** Complexity deriving from the range of subdistinctions available within a particular, grammaticalized (in a broad sense) linguistic category.

The senses of *syntagmatic* and *paradigmatic*, as used here, should be understood broadly, applying to phonological, morphological, or syntactic structures. A phonological string, for instance, with phonemic representation /kæt/ (for 'cat') would have syntagmatic complexity based on the arrangement of its three phonemes in a fixed linear order. At the same time, a passive construction would necessarily involve a degree of syntagmatic complexity, but also paradigmatic complexity, since the presence of a passive presupposes the existence of an opposing active in the grammatical domain of voice (see also Good (2012: 9–11)).

The basic prediction regarding simplification in creoles (and pidgins) found in Good (2012) centers around an asymmetry in the ability of each class of complexity given in (1) to be *transferred* from a source language into a jargon. As such, the prediction is somewhat limited in scope, not relevant to any and all kinds of complexities but, rather, to those that are ultimately found in a creole due to transfer from the languages that provided its initial lexical and grammatical material. In logical terms, a syntagmatic complexity can be transferred into a jargon (and, thus, ultimately be found in a creole) via the successful use and comprehension of a single unit of lin-

[5] The characterization of the notion of syntagmatic complexity in (1) has been reformulated in the interest of clarity following the advice of an anonymous reviewer. This reformulation is not intended to be construed as the result in any change in the understanding of the original concept.

guistic replication—i.e., a *lingueme* in the sense of Croft (2000: 200–205)—while a paradigmatic complexity requires the transfer of multiple linguemes, each evincing a different subdistinction associated with the relevant paradigm.

Consider, for instance, what is required for the string of phonemes associated with *cat* to be transferred into a jargon from English. All that is needed is for /kæt/ to be used in by a single member of the jargon community and for the other members to understand what is being referred to, at which point they may re-use this string to refer to similar entities. By contrast, for the English singular/plural distinction to be transferred, logically speaking, two linguemes must be transferred, e.g., *cat* and *cats*, since it is only possible for plural marking to enter the new contact variety if the coding of plurality via an affix is discoverable on the basis of the forms being used by its speakers. For this to happen, at least one singular and one plural noun must be transferred.

This "logical" model just described is clearly a vast simplification from the real world. Plural marking via *-s* would, presumably, only be transferred if used on at least several nouns, not just two. Similarly, if /kæt/ is uttered only once, it is not likely to eventually make its way into a creole. Mufwene's (2001:4–6) notion of a feature pool in language contact is clearly relevant here in understanding what features may or may not survive during processes of transfer. Whether or not a complexity is actually transferred will clearly depend on a complex set of ecological factors, only some of which are covered by Figure 1.

Moreover, even if a core set of linguemes instantiating a complexity does find a way into a jargon and, ultimately, a creole, this does not necessarily tell us the way that complexity will become instantiated in the creole itself. For instance, Saramaccan shows a distinction between singular and plural definite articles, with *dí* being used for singular referents and *déé* for plural ones (see McWhorter & Good (2012: 76–78)). The etymological source of these elements is generally seen to be connected to the English definite/demonstrative/pronoun system (e.g., with *this* cited as a likely source for *dí* and *them* as a likely source for *déé*), but in Saramaccan the most salient characteristic of *déé* is that it is the only grammatically available means to encode a noun as plural, making it an apparent case of a plural word (see Dryer (1989)), a category not found

in English. The presence of this paradigmatic complexity in Saramaccan can ultimately be traced to a paradigmatic complexity in English. It can, therefore, be considered to be an instance of a transferred complexity, even if it is somewhat restructured from its original function.

However, what is important here are not the details of specific complexities but, rather, the inherent asymmetry between syntagmatic and paradigmatic complexities. The former require successful transfer of a single lingueme, while the latter require transfer of a *set* of linguemes. From a purely statistical perspective, this should make paradigmatic complexities less likely to be transferred than syntagmatic complexities. Moreover, the more complex the number of paradigmatic distinctions, the less likely the entirety of the distinctions will be transferred, simply because the set of linguemes required for successful transfer will be larger—thus, for instance, the transfer of the regular English plural-marking paradigm is predicted to be more likely than the transfer of, say, a complex Bantu noun class system (Good 2012: 15–22).

Thus, this model should not be understood as arguing that creoles must lack all kinds of paradigmatic complexities. Rather, it suggests that, when compared to non-creoles, they should be paradigmatically *simplified*, since their jargonization stage would not have been conducive to transfer of paradigmatic complexities from a creole's source languages. Moreover, it suggests that we should not expect to see any special simplification with respect to syntagmatic complexities, since there is nothing particular about the process of creolization which would block their transfer (see Good (2012: 28–36)). A key question that remains, however, is how we can establish, in a rigorous manner, whether the model's predictions are actually satisfied.

In the next section, I introduce the WALS–APiCS dataset and describe how it was recoded to test the claims of Good (2012), with an emphasis on the claim that contact languages should be less complex in paradigmatic terms than other languages. We will see in the ensuing discussion that there are a number of difficulties in using this dataset. However, I use it here both for the pragmatic reason that is readily available and also to avoid any bias that might arise if I were to make use of data that I had collected myself, where my choices of features and coding distinctions might have been influenced by preconceived ideas of creole typology. Of course, this is not to say

the data is not colored by various biases (see, e.g., Kouwenberg (2010a;b) for relevant discussion). Rather, what is important is that the biases cannot be directly connected to the model of creole simplification developed in Good (2012).

# 3 The WALS–APiCS dataset

## 3.1 Using APiCS and WALS data in the present study

The Atlas of Pidgin and Creole Language Structures Online (APiCS; Michaelis et al. (2013d)) comprises a rich database of information on seventy-six pidgins, creoles, and other contact languages. Its orientation is largely typological. Therefore, for instance, for a given language in the database, one can get information on its basic word order, phonological inventory, predetermined lexical features, etc. In many respects, the information available in APiCS is richer than what is found in the better-known World Atlas of Language Structures Online (WALS; Dryer & Haspelmath (2013)). For instance, data underlying various typological classifications is regularly included, unlike WALS, and it is also possible to specify aspects of language variation in a relatively precise manner.

However, in the present context, if we want to verify a claim about creole simplicity in typological terms, we cannot look at creoles alone but must compare them with non-creole languages. In order to do this here, I will use a subset of the APiCS data that is specifically designed to be comparable to the WALS data (henceforth the WALS–APiCS dataset). This unfortunately means that many of the APiCS details will be lost, since the WALS data is simplified in comparison, but there is no straightforward way around this problem at present that I am aware of without, for instance, building a new version of the WALS database, which is well outside the scope of the present study. An important consequence of this choice is that the various contact languages to be looked at will be filtered through the general typological lens of WALS (see Kouwenberg (2010a) for a critique), though, as will be seen in Section 4, the typological comparison between the APiCS and WALS languages to be conducted here is not clearly suggestive of a bias against APiCS languages. For instance, they will, on the whole, seen to be paradigmatically simpler than

WALS languages, but with a lot of internal variation, and many individual WALS languages will come out as less paradigmatically complex than many APiCS languages.

Setting aside the limitations that the use of the WALS dataset imposes on the use of the APiCS dataset, an issue more specific to the present paper is that the existing values for typological features found in WALS and APiCS are not directly oriented in a way to test the predictions of the model outlined in Section 2. In particular, the values are of a traditional, descriptive type (e.g., a language may be specified as having SOV, SVO, VSO, etc. word order), rather than being designed for questions regarding the transferability of a grammatical complexity. In Section 3.2 I discuss how the WALS–APiCS values were recoded for their paradigmatic complexity, and in Section 3.3 I discuss this for syntagmatic complexity. At the end of this section, in Section 3.4, I briefly discuss the general utility of using theoretically-driven metrics of the sort seen here, whether or not the specific metrics or coding choices may turn out to be ideal.

The discussion immediately below is exemplary in nature. Full description of which WALS–APiCS features were used in this study, whether they were taken to provide information on paradigmatic or syntagmatic complexity, and the specific complexity scores assigned to their values is given in the Appendix to this paper. Altogether, this study is based on forty-five of the forty-eight WALS–APiCS features.[6] While a given typological domain can, of course, involve both paradigmatic and syntagmatic complexities, in practice, the WALS–APiCS features could generally only be easily associated with one or the other class (but see Section 3.3 for a counterexample). Thirty-one of the features were classified as describing paradigmatic complexities and fourteen as describing syntagmatic complexities.

### 3.2 Recoding a paradigmatic complexity

As just discussed, the WALS–APiCS dataset is not specifically designed to address the issue of whether there is an asymmetry in patterns of paradigmatic and syntagmatic complexity in contact

[6] Three WALS–APiCS features were not included in the present study because they did not allow for obvious interpretation in terms of a transfer-based complexity metric. Two of these involved cases of syntagmatic structures where it was not clear how to assign different complexity scores to the different values, or if they even would differ (Stassen 2013a;b). The third focused on a paralinguistic, rather than a linguistic, feature (Gil 2013b).

languages of the sort discussed in Section 2. In order to use it in this way, recoding is required so that its various typological features found can be associated with complexity scores. For example, consider WALS feature 53A covering ordinal numerals (Stolz & Veselinova 2013). Languages are assigned one of the values seen in (2).

(2) a. None
    b. One, two, three
    c. First, two, three
    d. One-th, two-th, three-th
    e. First/one-th, two-th, three-th
    f. First, two-th, three-th
    g. First, second, three-th
    h. Various

For the values in (2), a term like *one-th* is used for cases where a word meaning 'first' is formed using a regular morphological strategy, while the use of a term like *first* is used for cases where a word meaning 'first' is formed via a suppletive strategy. A term like *one* is used for cases where cardinal and ordinal numbers do not differ from each other. In this classificatory scheme, English is assigned the value *first, second, three-th* to reflect that the fact that the words *first* and *second* bear a suppletive relationship to their cardinal counterparts *one* and *two*. The word *third* is not a regular formation from *three*, but, at the same time, the relationship between the two is not "canonically" suppletive (see Corbett (2007)) since there is some formal correspondence between the words, thus *third* is treated as a word of the *-th* type.

There are various ways we could categorize the values in (2) with respect to some general notion of complexity. For instance, suppletion could be treated as more complex than a regular morphological strategy. From a production perspective, a value like *one, two, three* may be considered less complex than *one-th, two-th, three-th*, since it would require less morphological material, while, from a processing perspective, it may be more complex due to potential ambiguity between

cardinal and ordinal numeral senses. In this respect, it is well known that applying a monolithic notion "complexity" to linguistic data is not really possible (see also Karlsson et al. (2008)). However, for present purposes, the model discussed in (2) gives us a very specific heuristic for measuring complexity in terms of the minimal number of linguemes that would need to enter a jargon from a source language in order for a given paradigmatic complexity to be transferred.

The values in (2) encode aspects of both paradigmatic and syntagmatic complexity (with the latter being invoked in cases where a special morphological strategy is employed to derive ordinals). However, they are clearly primarily targeting an area of paradigmatic complexity in grammars. Does the morphological paradigm contain specific terms for ordinal numbers at all? If so, how many distinct rules for forming them are present? If we reframe such questions in terms of the minimal number of linguemes instantiating the relevant patterns, we can assign the complexity scores to the values in (2) as seen in (3), where a single complexity "point" is given for each lingueme which would have to be transferred.

(3) a. None: Zero linguemes

b. One, two, three: Zero linguemes (beyond numbers in general)

c. First, two, three: Two linguemes; *one* vs. *first*

d. One-th, two-th, three-th: Two linguemes; *one* vs. *one-th*

e. First/one-th, two-th, three-th: Three linguemes; *one* vs. *first*/*one-th*

f. First, two-th, three-th: Four linguemes; *one* vs. *first* and *two* vs. *two-th*

g. First, second, three-th: Six linguemes; *one* vs. *first* and *two* vs. *second* and *three* vs. *three-th*

h. Various: Six linguemes; choice based on upper bound of other strategies

In some cases, the complexity scores seen in (3) are fairly straightforward to understand, in other cases, they require elaboration. For instance, if a language simply lacks ordinal numbers (i.e., is in the category *none*), then, in terms of a transfer-oriented complexity metric based on lingueme counts, it seems clear that its complexity score should be zero: The lack of a paradigmatic

distinction means the absence of a paradigmatic complexity. The same score is also given to the category *one, two, three*, though, here, there is a conceptual complication. If a language's strategy for expressing ordinal functions is to use the same paradigm as found for cardinal functions, then this clearly does not require any transfer of a special ordinal morphological paradigm, which is why this value is scored as zero here. However, there is a clear sense in which the *one, two, three* category could be considered to be more complex than the *none* category: The former requires the presence of two distinct constructions involving numbers, whereas the latter does not. The presence of a cardinal/ordinal opposition at all could, thus, be seen to represent a constructional paradigmatic complexity, and this reveals the fact that a given WALS–APiCS feature may actually be encoding multiple kinds of complexities (see Section 3.3 for another example).

For the present study, each feature was recoded only once, focusing on the most salient complexity it described due to the difficulties involved in teasing apart all potential dimensions of complexity that might be "hidden" within a given feature. Nothing, in principle, however, prevents a feature from being recoded across multiple dimensions of complexity. In the case of (3), for instance, it could be coded both across morphological paradigmatic complexity, as done here, and across constructional complexity—i.e., whether a cardinal/ordinal opposition is present at all—in which case the *none* value would get a score of zero (since there is no relevant constructional distinction) and the others values would all get a score of two (representing the transfer of at least one lingueme evincing an cardinal construction and one evincing an ordinal construction).

Moving onto the other values in (3), *first, two, three* is assigned a score of two under the assumption that a *one* vs. *first* paradigmatic contrast can only be transferred into a contact variety if at least two linguemes are present each illustrating the different way the concept of 'one' is expressed in cardinal and ordinal contexts. The *one-th, two-th, three-th* value is also given a score of two since, even though it describes a fairly distinctive grammatical type from *first, two, three*, if we consider how a complexity of this kind could be transferred, it would still logically only require two linguemes, e.g., *four* and *fourth* used in cardinal and ordinal contexts respectively, a pairing

which could then, in principle, serve as the basis for an analogical extension across the number system into an emerging creole.

In a comparable fashion, the *first, two-th, three-th* value would require four transferred linguemes, one pair to instantiate the suppletion for the cardinal and ordinal forms of 'one' and one pair to instantiate the regular pattern. The *first, second, three-th* would require six linguemes, two pairs to instantiate the suppletive relationship for the forms of 'one' and 'two' respectively, and one more to instantiate the regular pattern. The final value of *various*, of course, does not allow for such a straightforward assignment of a metric. To determine the score used here, I considered the specific example of Vietnamese given in Stolz & Veselinova (2013) where there are three different word formation patterns, requiring six linguemes, as is the case for *first, second, three-th*, and the fact that the "mixed" nature of any wastebasket category like this will probably be relatively high in terms of complexity, and, therefore, assigned it the upper bound of the complexity scores for the other values (i.e., six). This is obviously somewhat unsatisfactory, but it is an inherent limitation in using the WALS–APiCS dataset. However, in many cases, such wastebasket categories do not contain many languages to start with (seven out of 321 in Stolz & Veselinova (2013) and four out of sixty four for the APiCS equivalent), meaning the actual impact of such decisions for typological comparison is mitigated (see also Section 4.6).

The heuristic nature of the metric for assigning the complexity scores seen in (3) must be emphasized here. I am not claiming that merely six linguemes are required for a system like *first, second, three-th* to actually be transferred. The metric is not intended to be employed as a predictive mechanism in such a specific way. Rather, it allows us to assess relative differences in paradigmatic complexity. The key claim is that transfer *first, second, three-th* system is more difficult than a *first, two-th, three-th* one which is in turn more difficult than a *one-th, two-th, three-th* one. Complexity scores using this sort of metric can then form the basis of a relative comparison of complexity among the APiCS and WALS languages, as will be done in Section 4.

Before moving on to an example of syntagmatic complexity in Section 3.3, it should be emphasized that a transfer-based metric of complexity is distinct from a transfer-based analysis of the presence of some kind of complexity. I will discuss this issue in more detail in Section 3.4.

### 3.3 Recoding a syntagmatic complexity

It is easy to imagine linguistic structures with high degrees of syntagmatic complexity—consider, for instance, the articulated constituency structures frequently attributed to clauses. However, the WALS–APiCS features that primarily target syntagmatic complexities are relatively abstract and schematic in nature and, thus, are not, on the whole, associated with high degrees of complexity. There are also comparatively fewer features relevant to syntagmatic complexity. Therefore, the conclusions of this paper regarding paradigmatic complexities in contact languages can almost certainly be considered stronger than those involving syntagmatic ones. Still, however, it will be interesting to look at asymmetries in paradigmatic and syntagmatic complexities in the dataset due to the different predictions regarding them made by the model outlined in Section 2.

As an example of the syntagmatic complexities found in the WALS–APiCS dataset, consider the values associated with a feature covering the expression of negative morphemes given in (4) (Dryer 2013b), the numbers after the value labels indicate the degree of syntagmatic complexity assigned to them here.

(4) a. Negative affix: One
    b. Negative particle: One
    c. Negative auxiliary verb: One
    d. Negative word, unclear if verb or particle: One
    e. Variation between negative word and affix: One
    f. Double negation: Two

As can be seen, the syntagmatic complexity scores are quite homogenous, all being of value one, to indicate that there is only one special marker in the clause coding negation, except for the

value of *double negation* where two morphemes are employed, hence giving it a value of two. This general strategy of considering how many "marks" there are of a given category is the metric used to determine syntagmatic complexity scores here. Obviously, the classification in (4) does not cover the scope of possible syntagmatic complexities for negative marking even when only one morpheme is involved. For instance, if a negative element was placed in a special position (e.g., second position), we may want to consider this to be more complex than it being placed next to the verb, depending on our theory of syntactic positions. However, as can be seen, the WALS values are not sufficiently detailed to allow for such fine-grained consideration.

An examination of the values in (4) reveals another complication. One of the values, *variation between negative word and affix* (4e), represents a paradigmatic complexity since variation between two strategies could only be detectable via two linguemes, each evincing one of the variants, while the other strategies only require one lingueme, following the transfer-based complexity metric developed in Section 2. In principle, we could thus doubly code the feature in (4) for both syntagmatic and paradigmatic complexity (see also Section 3.2 for another case where a feature appears to encode more than one complexity). In practice, each feature was treated as exclusively syntagmatic or paradigmatic here since there were not many features obviously associated with such a dual complexity type. As will be briefly discussed in Section 4.6, most of the major results to be reported on in this paper were relatively robust in the sense that changes in the data such as addition/removal of a small set of features or languages did not lead to different conclusions. So, it did not seem worthwhile to introduce dual coding for the present study, which must be viewed as only an initial foray into a new way to measure featural complexity in any event (see Section 5).

Of the thirteen features in the syntagmatic complexity class, nine involve word order (reflecting the fact that this is probably the most well-developed area of traditional linguistic typology). In general, for such features, if a language's word order was specified as fixed, this counted as a syntagmatic complexity. If the language was classified as having no dominant word order, this was treated as the lack of a complexity. Good (2012: 13–15) discusses the logic behind this coding

choice in detail, but the essential point is that fixed word order is treated as requiring special grammatical specification, while free word order is understood as requiring no such specification.

### 3.4 Recoding based on sociohistorically-derived metrics

As mentioned above, the full range of coding choices for the WALS–APiCS data used in the present study can be found in the Appendix to this paper, which includes some additional discussion of the coding principles. Inevitably, if these choices were closely scrutinized, there would be places where a different analyst may question various aspects of them, though I have tried, to the extent possible, to be internally consistent. I would, therefore, like to distinguish two distinct aspects of the use of the WALS–APiCS data here: The concrete details of the recoding vs. their conceptual motivation. The former can be easily altered if debates reveal them to be less than ideal and are not as significant as the latter, in my view.

What I believe is innovative here is the idea that the specific metrics used to assess complexity in the context of contact language typology should be grounded in a model of how an emerging contact variety forms in its social context. This model is, therefore, derived from an understanding of the sociohistorical circumstances under which creoles are often thought to arise—albeit in highly schematized form. While it is almost certainly the case that many details of this this study could be improved upon, I believe that a research program grounding metrics of complexity in a sociohistorically-motivated model is ultimately bound to result in a better understanding of creole typology than the use of more general metrics of the sort adopted by McWhorter (2001) or Parkvall (2008), and I will return to this issue briefly in Section 5.

In the interests of clarity, we should distinguish here between the use of metrics defined with respect to a model of transfer and whether a given complexity is present in a language because of transfer. Clearly, it is not the case that all instances of paradigmatic and syntagmatic complexity in contact languages are the results of transfer (see, e.g., the discussion of Haitian Creole determiner allomorphy in Good (2012: 27–28)). And, when we speak of non-contact languages, such an idea makes even less sense. However, if we define a transfer-oriented complexity metric with sufficient

generality (as is done here), there is no reason it cannot be applied to complexities that are not the result of transfer in a given language. The goal here is not to unravel the history of any particular complexity but to look for typological signals associated with a language's sociohistory. What we have here, then, are metrics specifically designed to detect languages with a particular sociohistorical profile. We can test their validity precisely by applying them to two sociohistorically distinct classes of languages, one associated with that profile and one not, and looking for asymmetries in resulting quantitative investigation.

Bearing all of the above points in mind—and in particular the limits imposed on us by the use of the WALS–APiCS dataset—in the next section, I will discuss the results of quantitative investigation of the WALS–APiCS dataset in terms of paradigmatic and syntagmatic complexity across the two sets of languages.

# 4 Quantitative comparisons between APiCS and WALS languages

## 4.1 Analytical procedure

Having introduced the model on which this study is based in Section 2 and discussed how the WALS–APiCS data can be adapted to test the model, we are now in a position to see whether or not the WALS–APiCS data supports the idea that simplification in creoles is especially likely to target paradigmatic complexities. In order to do this, the WALS–APiCS feature values were all recoded with complexity scores of the sort exemplified in Section 3.2 and Section 3.3. The complete set of the recodings is given in the Appendix, along with some additional notes on the recoding procedure. This recoding permits a quantitative investigation of differences in paradigmatic and syntagmatic complexity in languages from the APiCS and WALS datasets.[7]

[7] The precise data sources used for this study in terms of feature–value pairings were derived from the WALS data made available at http://wals.info/download and downloaded in July 2013, and the precise APiCS data was made available by Susanne Michaelis and Robert Forkel in July 2013 in the form of tables specifically suited to this study. The full APiCS dataset can be found at http://apics-online.info/download. The data was then processed and entered into a MySQL database which was used for the study described here. Copies of the database, associated data processing scripts, and various data reports used to produce this paper are available at http://github.com/jcgood/complexity.

As can be seen in an example like (3), the assigned complexity scores were reduced to integers (in the form of lingueme counts or syntagmatic markings as discussed in Section 3.2 and Section 3.3 respectively). For the purposes of the quantitative analysis, as will be seen below, complexity scores were normalized so that they fell between 0 and 1 by dividing a given value's associated complexity with the highest possible complexity for the relevant feature. Thus, the integer scores in (3) were transformed into the scores between 0 and 1 as seen in (5), with the normalized scores given in parentheses at the end of the value description.

(5) a. None: Zero linguemes (*score:* 0)

b. One, two, three: Zero linguemes (beyond numbers in general) (*score:* 0)

c. First, two, three: Two linguemes; *one* vs. *first* (*score:* 0.33)

d. One-th, two-th, three-th: Two linguemes; *one* vs. *one-th* (*score:* 0.33)

e. First/one-th, two-th, three-th: Three linguemes; *one* vs. *first*/*one-th* (*score:* 0.5)

f. First, two-th, three-th: Four linguemes; *one* vs. *first* and *two* vs. *two-th* (*score:* 0.67)

g. First, second, three-th: Six linguemes; *one* vs. *first* and *two* vs. *second* and *three* vs. *three-th* (*score:* 1)

h. Various: Six linguemes; choice based on upper bound of other strategies (*score:* 1)

This normalization procedure creates clear disparities among features where, for instance, a feature whose maximum complexity is 2, can only take on values 0, 0.5, and 1, while a feature like the one in (5) has a wider range of values. This would clearly be a problem if we were comparing features directly. However, in the present study, we are not trying to determine, for instance, which feature shows greatest complexity in the world's languages but, rather, how complexity in APiCS languages compares to complexity in WALS languages. Since languages of the two groups were not treated differently in the scoring and since the same set of features was compared across all of them, this aspect of coding should not impede our ability to use the normalized scores to compare the two groups to each other.

The quantitative comparisons discussed below were done between all of the APiCS languages and all of the WALS languages, excepting those specifically identified as creoles and pidgins in the

genealogical information provided at http://wals.info/languoid/genealogy.[8] Beyond this, no judgment was made regarding whether or not a given language may or may not qualify as a "pidgin" or "creole", etc.[9] In the APiCS dataset, for instance, as we will see in Section 4.3, there are languages belonging to the conventional set of contact languages but which are generally put into a third category of mixed languages, and one of these has strikingly different paradigmatic complexity scores by the metrics employed here than the conventional "creoles".

While justification could be made for the removal of some of the languages found in the APiCS dataset based on the sociohistorical circumstances of their development, this would run the risk of circularity: That I had defined a set of languages that I already believed to be "creoles" based on their properties and then found that they shared some property. Therefore, I preferred instead to leave the whole APiCS dataset intact, in which case it is not my own judgment, but, rather, the judgment of the APiCS editors regarding which languages belonged to the set of "pidgins and creoles", even if some of the APiCS languages do not clearly fit that classification (see Michaelis et al. (2013a: xxxii–xxxvi) for further discussion). Below, I will sometimes rhetorically treat APiCS languages as a stand-in for "pidgins and creoles" or "creoles" despite the inclusion of a few contact languages of different types in the APiCS language set. Furthermore, while Good (2012) focused on creoles rather than pidgins, it is clear that its predictions regarding simplification should be taken as applying to pidgins as well (see Figure 1).

In the following sections, I will consider three questions: (i) whether the average complexity scores for the features considered here show significant differences across the APiCS and WALS datasets (see Section 4.2), (ii) whether the average paradigmatic complexity scores across

[8] Reference to WALS languages from here onwards should be understood as excluding those creoles and pidgins found in the WALS language set.

[9] The mappings between the APiCS features and their WALS equivalents was done as part of the larger APiCS project, and I did not alter that mapping here. It should be noted, however, that there are some discrepancies between the APiCS variants of WALS feature values and the original WALS feature values. For instance, for the APiCS feature relating to numeral classifiers (see Maurer & The APiCS Consortium (2013)), only two values were used for the presence vs. absence of classifiers, while, for WALS three values were possible, with an additional value for optional classifiers (see Gil (2013a)). Rather than impose my own judgment regarding whether the recodings were done appropriately in all cases, I simply used the ones that were found in the WALS–APiCS mapping itself, following the general principle here of only manipulating the original datasets to the minimal extent required, in order to avoid imposing personal biases on the results. I do not expect that any issues arising from such discrepancies would alter the most important conclusions of this paper.

the APiCS and WALS languages show significant differences (see Section 4.3), (iii) whether the average syntagmatic complexity scores across the APiCS and WALS languages show significant differences (see Section 4.4), and (iv) which of the paradigmatic features under consideration here have the greatest predictive power with respect to placing a given language into the APiCS and WALS dataset (and, by extension, which seem most typically "creole" or "non-creole").[10] The first three questions can be considered to fall out more or less directly from the predictions of the model in Figure 1, along with distinction between paradigmatic and syntagmatic complexities developed in Section 2. The fourth question was considered for examination after it was found that the APiCS languages were, overall, simpler in paradigmatic terms than the WALS languages, and it was chosen as a way to explore this result in more detail.

As with any study of this kind, a wide range of caveats apply. Most importantly, the results are most proximately about the WALS–APiCS data and only indirectly about "creoles" and "non-creoles". In Section 4.6, I will address this concern, and related ones, in more detail, though on the whole I leave open the question of the extent to which the WALS–APiCS data is a valid data set for a study like this one simply because there is no better dataset available.

### 4.2 Complexity by feature across APiCS and WALS

In Table 1, the average normalized complexity scores are presented across the WALS–APiCS features examined for this study. The table specifically gives information about (i) the feature in terms of its WALS identifier along with an abbreviated description (for purposes of presentation), (ii) an indication as to whether or not the feature was treated as describing a type of paradigmatic or syntagmatic complexity (via the abbreviations P or S in the column entitled T for type), (iii) the average normalized complexity score for the feature across the languages in the the APiCS and WALS datasets, (iv) the results of a statistical comparison between the complexity scores across

[10] All statistical tests described here were performed using R (R Core Team 2013), with specific additional packages cited below where appropriate. The data itself was processed using various Python scripts, and these can be found at the GitHub repository at https://github.com/jcgood/complexity. These scripts are not part of the formally reviewed materials for this paper and are not designed for general use, but have been made accessible for those interested in examining them, and the author will provide further information on how to use them if requested. The scripts are designed to process data stored in a custom MySQL database (see fn. 7).

the two sets of languages given in the form of an approximate *p*-value for statistical significance (using a $p \leq 0.05$ threshold), and (iv) based on this statistical test, an indication as to whether the APiCS language set was more complex than the WALS set, less complex, or statistically the same.[11] The features are ordered first according to whether or not the APiCS score was higher than WALS, about the same as WALS, or lower than WALS, and then by their p-value, from lowest to highest (i.e., from whether the difference between APiCS and WALS is more or less statistically significant).

Various generalizations emerge from Table 1. Perhaps the most important here is the lack of a consistent partitioning of features where paradigmatic features are always more complex for WALS than APiCS. For some features, such as those involving the presence of articles and how they relate to other words (i.e., WALS 38A and WALS 37A), APiCS languages are, on the whole, more paradigmatically complex than WALS languages. For others, such as whether or not there is a distinctive class of nasal vowels (WALS 10A), the two groups are not distinct in statistical terms. Furthermore, one finds syntagmatic features differing across these sets as well in a similar fashion.

Thus, it is clear that we cannot suggest there is broad homogeneity with respect to featural complexity across either the APiCS or WALS datasets. Of course, we would not necessarily expect this, but, if we were take seriously claims such as that embodied in the title of McWhorter (2001), that the "worlds' simplest grammars are creole grammars", this is not clearly borne out by the metrics and statistical tests employed here.

Still, however, this does not mean there are no relevant patterns in the data. The most striking of these is surely connected to those cases where WALS langages are scored as more complex than APiCS languages, given in the third section of the table. Out of fourteen such features, thirteen are, in fact, paradigmatic. A number of these deal with grammatical domains classically associated with morphological paradigms, such as case marking on noun phrases.[12] Others related to phenomena

[11] The particular statistical test employed was Welch's two-sample *t* test as implemented in R (R Core Team 2013). The set of complexity scores across each feature for APiCS languages was treated as the first sample, and the set of complexity scores across WALS languages (excepting those classified as creoles as discussed in Section 4.1) was treated as the second sample.

[12] It should be noted, however, that the relevant WALS sense of *case* encompasses both morphological case and adpositionally-coded case (Comrie 2013).

| FEATURE | DESCRIPTION | T | APICS | WALS | ≈P | COMP |
|---|---|---|---|---|---|---|
| WALS 93A | Position of interrogative phrases | S | 0.55 | 0.51 | 0.00 | APiCS > WALS |
| WALS 38A | Indefinite articles | P | 0.66 | 0.36 | 0.00 | APiCS > WALS |
| WALS 122A | Subject relative clauses | P | 0.15 | 0.03 | 0.00 | APiCS > WALS |
| WALS 116A | Polar questions | S | 0.62 | 0.51 | 0.00 | APiCS > WALS |
| WALS 124A | 'Want' complement subjects | S | 0.59 | 0.52 | 0.00 | APiCS > WALS |
| WALS 90A | Order of relative clause and noun | S | 0.53 | 0.47 | 0.00 | APiCS > WALS |
| WALS 37A | Definite articles | P | 0.67 | 0.52 | 0.01 | APiCS > WALS |
| WALS 115A | Negation and indefinite pronouns | S | 0.71 | 0.67 | 0.01 | APiCS > WALS |
| WALS 42A | Pronominal/adnominal demonstratives | P | 0.43 | 0.28 | 0.02 | APiCS > WALS |
| WALS 99A | Case marking of personal pronouns | P | 0.47 | 0.36 | 0.02 | APiCS > WALS |
| WALS 106A | Reciprocal constructions | P | 0.58 | 0.47 | 0.02 | APiCS > WALS |
| WALS 89A | Order of cardinal numeral and noun | S | 1.00 | 0.94 | 0.04 | APiCS > WALS |
| WALS 120A | Predicative noun phrases | P | 0.57 | 0.46 | 0.07 | APiCS ≈ WALS |
| WALS 91A | Order of degree word and adjective | S | 0.60 | 0.56 | 0.09 | APiCS ≈ WALS |
| WALS 88A | Order of demonstrative and noun | S | 0.45 | 0.48 | 0.10 | APiCS ≈ WALS |
| WALS 33A | Expression of nominal plural meaning | P | 0.52 | 0.48 | 0.11 | APiCS ≈ WALS |
| WALS 81A | Order of subject, object, and verb | S | 0.92 | 0.86 | 0.15 | APiCS ≈ WALS |
| WALS 86A | Order of possessor and possessum | S | 0.56 | 0.54 | 0.17 | APiCS ≈ WALS |
| WALS 87A | Order of adjective and noun | S | 0.88 | 0.92 | 0.21 | APiCS ≈ WALS |
| WALS 53A | Ordinal numerals | P | 0.63 | 0.58 | 0.22 | APiCS ≈ WALS |
| WALS 47A | Intensifiers and reflexive pronouns | P | 0.36 | 0.44 | 0.25 | APiCS ≈ WALS |
| WALS 64A | Nominal and verbal conjunction | P | 0.34 | 0.42 | 0.26 | APiCS ≈ WALS |
| WALS 112A | Negative morpheme types | S | 0.53 | 0.55 | 0.29 | APiCS ≈ WALS |
| WALS 46A | Indefinite pronouns | P | 0.17 | 0.12 | 0.29 | APiCS ≈ WALS |
| WALS 105A | Ditransitive constructions with 'give' | P | 0.71 | 0.70 | 0.33 | APiCS ≈ WALS |
| WALS 34A | Occurrence of nominal plural markers | P | 0.55 | 0.53 | 0.51 | APiCS ≈ WALS |
| WALS 45A | Politeness distinctions | P | 0.23 | 0.26 | 0.53 | APiCS ≈ WALS |
| WALS 10A | Nasal vowels | P | 0.30 | 0.26 | 0.57 | APiCS ≈ WALS |
| WALS 85A | Order of adposition and noun phrase | S | 0.52 | 0.51 | 0.61 | APiCS ≈ WALS |
| WALS 63A | NP conjunction and comitative | P | 0.53 | 0.56 | 0.62 | APiCS ≈ WALS |
| WALS 79A | Suppletion for tense and aspect | P | 0.24 | 0.24 | 0.86 | APiCS ≈ WALS |
| WALS 98A | Case marking of full noun phrases | P | 0.12 | 0.33 | 0.00 | WALS > APiCS |
| WALS 101A | Expression of pronominal subjects | P | 0.32 | 0.63 | 0.00 | WALS > APiCS |
| WALS 129A | 'Hand' and 'arm' | P | 0.41 | 0.63 | 0.00 | WALS > APiCS |
| WALS 52A | Comitatives and instrumentals | P | 0.26 | 0.76 | 0.00 | WALS > APiCS |
| WALS 54A | Adnominal distributive numerals | P | 0.19 | 0.53 | 0.00 | WALS > APiCS |
| WALS 71A | The prohibitive | P | 0.16 | 0.54 | 0.00 | WALS > APiCS |
| WALS 119A | Predicative noun and locative phrases | P | 0.48 | 0.69 | 0.00 | WALS > APiCS |
| WALS 24A | Marking of possessor noun phrases | S | 0.36 | 0.48 | 0.00 | WALS > APiCS |
| WALS 109A | Applicative constructions | P | 0.04 | 0.37 | 0.00 | WALS > APiCS |
| WALS 55A | Sortal numeral classifiers | P | 0.04 | 0.29 | 0.00 | WALS > APiCS |
| WALS 41A | Distance contrasts in demonstratives | P | 0.35 | 0.49 | 0.00 | WALS > APiCS |
| WALS 39A | Inclusive/exclusive distinction | P | 0.11 | 0.34 | 0.00 | WALS > APiCS |
| WALS 13A | Tone | P | 0.21 | 0.33 | 0.01 | WALS > APiCS |
| WALS 44A | Gender distinctions in pronouns | P | 0.14 | 0.25 | 0.02 | WALS > APiCS |

Table 1: Average complexity by feature

involving particular inflectional or quasi-inflectional morphological distinctions, such as in the use of numeral classifiers or inclusive and exclusive pronouns—i.e., they belong to domains which can be considered broadly morphological in nature.[13] It seems reasonable to infer from this that some degree of paradigmatic simplification is a true creole property, especially in morphological domains.

When we set this count against the one for cases where APiCS languages were found to be more complex than WALS languages, we find a distribution where there are six syntagmatic and six paradigmatic features. And, when we look at cases where there was no statistically significant difference, we see a distribution of twelve paradigmatic features against seven syntagmatic ones. Given that only around one third of the total features examined here were classified as syntagmatic (fourteen out of forty-five), the second set of features in Table 1, where seven out of nineteen features showed similar complexity across APiCS and WALS is in line with what we would expect if the paradigmatic/syntagmatic complexity distinction were not a relevant factor with respect to the typology of creoles, while the first set of features, where APiCS languages are more complex than WALS languages appears somewhat biased towards involving syntagmatic features, though the figures involved are too small for any robust statistical generalization to be made in this latter regard.

The results in Table 1, therefore, seem to be in line with the suggestion of Good (2012) that creoles should be simpler in paradigmatic terms than syntagmatic terms when set against non-creoles. At the same time, this result is only appears when we look at these two classes of languages across many features, not just a few. Though I will not examine the issue statistically here, perhaps what we are seeing is that creoles show a certain set of biases in their featural specifications in line with what we expect of families or areas, rather than a sharp divide from other languages.

[13] At the same time, there are also cases where a morphologically-oriented feature shows more complexity in the APiCS languages than the WALS ones, with the most noticeable being case marking on pronouns, which patterns differently from case marking on noun phrases. Explanations for the results for each individual feature are outside the scope of this paper. However, cases like this do reveal the extent to which the simple vs. complex dichotomy fails to account for the full range of the data. See also Section 5.

In the next two sections, I will change the focus from features to languages, first considering the paradigmatic domain and then the syntagmatic one.

### 4.3 Paradigmatic complexity by language across APiCS and WALS

In Table 2, I present a list of all the languages in APiCS and WALS which were specified for twenty-six or more of the thirty-one paradigmatic features presented in the Appendix.[14] The choice of twenty six was somewhat arbitrary, intended to ensure that languages only specified for a small number of features did not skew the results and to provide a set of comparison languages of about equal size. This resulted in a selection of seventy-three languages from APiCS and sixty-two from WALS.[15] A language's paradigmatic complexity score was calculated by summing its normalized paradigmatic complexity scores across all features for which it was specified and dividing this by that same number of features. Languages are presented in order from those scored least paradigmatically complex to most paradigmatically complex. APiCS languages are italicized in the list.

Various generalizations emerge from Table 2. The most significant here is probably the extent of the mixing of APiCS and WALS languages, where many APiCS languages are more complex than many WALS languages. Indeed, the two most complex languages, in paradigmatic terms, are APiCS languages, Michif and Sri Lanka Portuguese. That Michif comes out as distinctive is not surprising. It is generally classified as a mixed language, rather than a creole, and the sociohistorical circumstances of its creation were quite distinct from the model depicted in Figure 1 (see Bakker (1997; 2013)). Other languages in the APiCS set typically considered to be mixed, namely Media Lengua (Muysken 2013), Gurindji Kriol (Meakins 2013), and Mixed Ma'a/Mbugu (Mous 2013), do not score especially high, however. So, while Michif's mixed status probably can explain its outlier position, "mixing" alone cannot be considered the sole factor in this.

[14] For purposes of presentation, the names of the various creoles associated with Cape Verde are abbreviated in Table 2 and Table 3, beginning with *Cape Verd. Cr.* for *Cape Verdean Creole*.

[15] Though there are thousands of languages in the WALS database, against only seventy six in the APiCS database, the nature of the APiCS data collection process has meant that many more APiCS languages are specified for a large number of features than WALS languages, which is why the two language sets come out as around equal in this case.

| | | | | | |
|---|---|---|---|---|---|
| *Pidgin Hindustani* | 0.17 | *Chinese Pidgin Russian* | 0.36 | French | 0.42 |
| *Chinuk Wawa* | 0.20 | Chukchi | 0.36 | Hausa | 0.42 |
| *Palenquero* | 0.22 | *Guyanais* | 0.36 | Latvian | 0.42 |
| *Berbice Dutch* | 0.24 | *Haitian Creole* | 0.36 | Malagasy | 0.42 |
| *Martinican Creole* | 0.24 | *Jamaican* | 0.36 | *Mauritian Creole* | 0.42 |
| *Media Lengua* | 0.24 | *Lingala* | 0.36 | *Diu Indo-Portuguese* | 0.43 |
| *Negerhollands* | 0.24 | *Nigerian Pidgin* | 0.36 | Fijian | 0.43 |
| *Juba Arabic* | 0.25 | *Singlish* | 0.37 | *Guinea-Bissau Kriyol* | 0.43 |
| *Kikongo-Kituba* | 0.27 | *Sri Lankan Malay* | 0.37 | Hebrew (Modern) | 0.43 |
| *Krio* | 0.27 | Swahili | 0.37 | Lango | 0.43 |
| *Nicaraguan Creole English* | 0.27 | *Angolar* | 0.38 | Yoruba | 0.43 |
| *Singapore Bazaar Malay* | 0.27 | *Cape Verd. Cr. of São Vicente* | 0.38 | Lakhota | 0.44 |
| *Tok Pisin* | 0.27 | Chamorro | 0.38 | Oromo (Harar) | 0.44 |
| Yimas | 0.27 | *Gurindji Kriol* | 0.38 | Thai | 0.44 |
| *Belizean Creole* | 0.28 | Hixkaryana | 0.38 | Tukang Besi | 0.44 |
| *Early Sranan* | 0.29 | *Kriol* | 0.38 | Vietnamese | 0.44 |
| *Guadeloupean Creole* | 0.29 | *Louisiana Creole* | 0.38 | *Chinese Pidgin English* | 0.45 |
| Kobon | 0.29 | Mapudungun | 0.38 | *Ghanaian Pidgin English* | 0.45 |
| Greenlandic (West) | 0.30 | *Principense* | 0.38 | *Hawai'i Creole* | 0.45 |
| *Kinubi* | 0.30 | *Reunion Creole* | 0.38 | *Korlai* | 0.45 |
| *Ambon Malay* | 0.31 | Turkish | 0.38 | *Norf'k* | 0.46 |
| *Cameroon Pidgin English* | 0.31 | *Fa d'Ambô* | 0.39 | Basque | 0.47 |
| English | 0.31 | *Gullah* | 0.39 | Burushaski | 0.47 |
| Evenki | 0.32 | Russian | 0.39 | Supyire | 0.47 |
| Jakaltek | 0.32 | Tiwi | 0.39 | *Ternate Chabacano* | 0.47 |
| *Sranan* | 0.32 | Yukaghir (Kolyma) | 0.39 | Korean | 0.48 |
| *Fanakalo* | 0.33 | Zulu | 0.39 | Yaqui | 0.48 |
| *Nengee* | 0.33 | *Cape Verd. Cr. of Brava* | 0.40 | Greek (Modern) | 0.49 |
| *Papi* | 0.33 | *Cape Verd. Cr. of Santiago* | 0.40 | Hungarian | 0.49 |
| *Pichi* | 0.33 | Imonda | 0.40 | Mandarin | 0.49 |
| *Pidgin Hawaiian* | 0.33 | Indonesian | 0.40 | Meithei | 0.49 |
| *Seychelles Creole* | 0.33 | Iraqw | 0.40 | Japanese | 0.50 |
| *Tayo* | 0.33 | Krongo | 0.40 | Tagalog | 0.50 |
| *Trinidad English Creole* | 0.33 | *San Andres Creole English* | 0.40 | *Cavite Chabacano* | 0.51 |
| *African American English* | 0.34 | *Santome* | 0.40 | Ainu | 0.52 |
| Arabic (Egyptian) | 0.34 | *Saramaccan* | 0.40 | *Zamboanga Chabacano* | 0.52 |
| *Bislama* | 0.34 | *Bahamian Creole* | 0.41 | Maori | 0.53 |
| *Papiamentu* | 0.34 | Georgian | 0.41 | Abkhaz | 0.54 |
| Amele | 0.35 | German | 0.41 | Spanish | 0.54 |
| *Casamancese Creole* | 0.35 | Hindi | 0.41 | Kannada | 0.57 |
| Finnish | 0.35 | Khalkha | 0.41 | Guaraní | 0.58 |
| *Mixed Ma'a/Mbugu* | 0.35 | Lezgian | 0.41 | *Michif* | 0.62 |
| Ngiyambaa | 0.35 | Slave | 0.41 | *Sri Lanka Portuguese* | 0.64 |
| Persian | 0.35 | *Afrikaans* | 0.42 | | |
| *Sango* | 0.35 | Alamblak | 0.42 | | |
| *Vincentian Creole* | 0.35 | *Creolese* | 0.42 | | |

Table 2: Average paradigmatic complexity by language, sorted from lowest to highest

Mixing does not appear to be relevant to the Sri Lanka Portuguese case, and its complexity score is presumably at least partly connected to the effects of contact between it and Tamil, resulting in a noteworthy amount of grammatical convergence between the two (Smith 1977:3; 2013).[16]

It is outside of the scope of this paper to consider the precise sociohistories and grammars of all the languages in Table 2 in order to understand their relative rankings. Suffice it to say, there is no evidence that creoles are *uniformly* simple paradigmatically in their grammars. However, an impressionistic examination of Table 2 does suggests they may be simpler overall. For instance, the eight languages with the lowest complexity scores are all from the APiCS set (as indicated by italics). Five of the six languages at the next ranking (with a rounded score of 0.27) are also from APiCS. Furthermore, as one moves to languages with higher scores, the APiCS languages start to be less prominent.

The impressionistic sense that APiCS languages are, on the whole, less paradigmatically complex following the metric used here can be further verified by statistical examination. In Figure 2, kernel density plots are given estimating a distribution of paradigmatic complexity scores for both the APiCS and WALS languages, as provided in Table 2. Scaling these so that they can be superimposed on each other gives an intuitive visual presentation of the differences in the paradigmatic complexity scores across the two sets of languages.[17] The difference between these sets of scores is statistically significant. The mean paradigmatic complexity score for the APiCS languages is 0.36 ($\pm$0.08), and the mean for the WALS languages is 0.42 ($\pm$0.07), with p $\approx$ 0 by Welch's two-sample *t*-test as implemented in R (R Core Team 2013).

We can, therefore, infer from these results that creoles and pidgins are on average, paradigmatically simpler than other languages, even if they are not uniformly "simple". Notably, we arrived at this result without removing obvious outliers such as Michif, suggesting it is relatively robust (see also Section 4.6). In the next section, I will consider the issue of syntagmatic complexity where, it

[16] While Tamil is not a language in the WALS–APiCS dataset with twenty-six or more paradigmatic features specified, another Dravidian language, Kannada, is in this group and, as can be seen, in Table 2, it has a fairly high complexity score, coming in as the fourth most paradigmatically complex language, lending support to the idea that Tamil influence on Sri Lanka Portuguese could have raised the latter's complexity score considerably.

[17] The kernel density estimation on which Figure 2 is based was calculated using R (R Core Team 2013), and the visualization itself was produced using the **ggplot2** package (see Wickham (2009)).

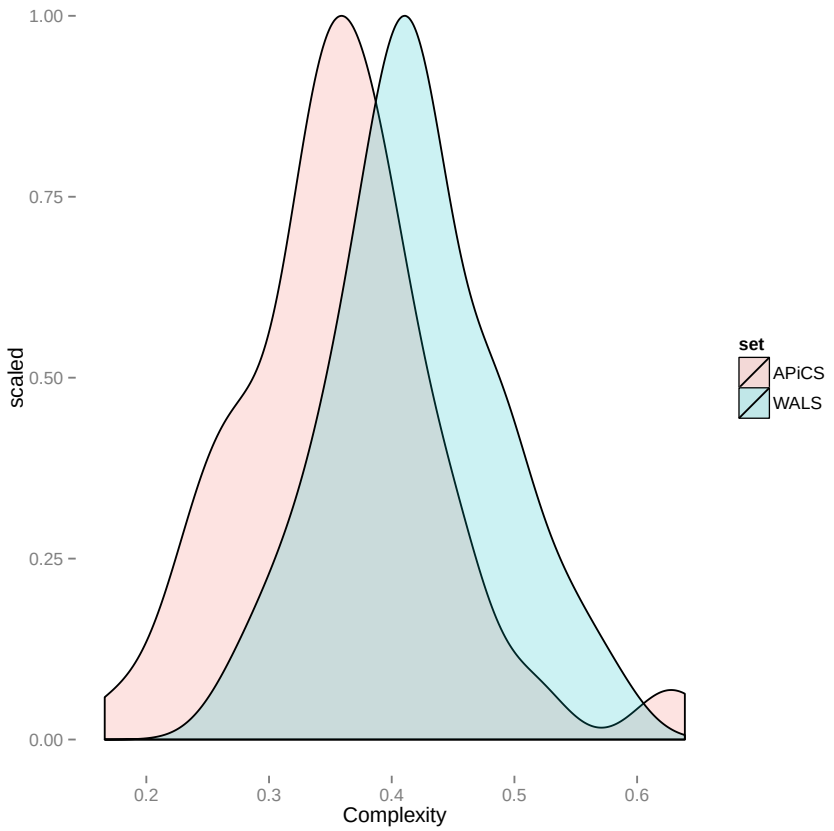


Figure 2: Density plot of paradigmatic complexity distribution across APiCS and WALS languages

will be seen, the APiCS languages are not found to be simpler than the WALS languages following the metric used here.

### 4.4 Syntagmatic complexity

As discussed in Section 2, the model presented in Good (2012) specifically predicts that creoles (and pidgins) will, on the whole, be simpler in paradigmatic terms than other languages. It does not as strongly predict any particular trend towards simplification in the syntagmatic domain. To the extent it predicts anything in this regard, it would be that there should be "averaging" over the syntagmatic complexities found in the languages contributing to the development of a contact language as a result of effects connected to second-language acquisition in general (see Good (2012: 29–36)).

The WALS–APiCS dataset is not as well-suited for examining differences in the syntagmatic complexity of APiCS and WALS languages as it is for paradigmatic complexities. There are fewer total features that characterize syntagmatic complexities (fourteen against thirty one), and there is less diversity in the grammatical domains that these features target with eight out of the fourteen covering aspects of word order. Nevertheless, the WALS–APiCS data still gives us an initial means to test the predictions of Good (2012) on a large-scale dataset in a more rigorous way.

The Appendix lists the various syntagmatic features looked at in this study as well as how they were coded in terms of their complexity scores. Following the same basic procedure as for the paradigmatic complexity data presented in Table 2, in Table 3, the normalized average syntagmatic complexity score is given by language for those languages which were coded for at least thirteen of the fourteen syntagmatic features. This ensured that only well-covered languages were considered together and produced a roughly similar number of languages from the APiCS and WALS sets (seventy-four APiCS languages against eighty-three WALS languages). As was the case for Table 2, the APiCS language names in Table 3 are italicized.

For present purposes, the most striking feature of the language distributions in Table 3, when set against the results for average paradigmatic complexity in Table 2, is the lack of any obvious clustering of the APiCS languages towards the simpler end of the list. We can visualize the syntagmatic complexities of APiCS languages versus WALS languages via kernel density plots as in Figure 3, as was done for paradigmatic complexity averages in Figure 2. The APiCS and WALS languages do not show the same distributional shape, with the somewhat "roller-coaster" shape distribution for APiCS languages reflecting the fact that they cluster together at various points across the syntagmatic complexity spectrum rather than being more evenly distributed. However, there is no skewing of the APiCS languages towards being simpler in terms of syntagmatic distinctions and, in fact, they turn out, if anything, to be slightly more syntagmatically complex than WALS languages according to the syntagmatic complexity metric used here. The mean syntagmatic complexity for APiCS languages is 0.63 ($\pm$0.05), and the mean syntagmatic complexity for APiCS

| | | | | | |
|---|---|---|---|---|---|
| Bininj Gun-Wok | 0.40 | *Guadeloupean Creole* | 0.62 | *Batavia Creole* | 0.65 |
| Tzutujil | 0.45 | *Haitian Creole* | 0.62 | Burmese | 0.65 |
| Chukchi | 0.47 | Hausa | 0.62 | *Creolese* | 0.65 |
| Kutenai | 0.50 | Hmong Njua | 0.62 | Finnish | 0.65 |
| Arabic (Egyptian) | 0.51 | Indonesian | 0.62 | *Gurindji Kriol* | 0.65 |
| *Guyanais* | 0.51 | Jakaltek | 0.62 | Hebrew (Modern) | 0.65 |
| Hungarian | 0.51 | Japanese | 0.62 | Hindi | 0.65 |
| Macushi | 0.51 | Khoekhoe | 0.62 | *Jamaican* | 0.65 |
| *Seychelles Creole* | 0.51 | *Kinubi* | 0.62 | *Juba Arabic* | 0.65 |
| German | 0.52 | Korean | 0.62 | *Lingala* | 0.65 |
| Yagua | 0.54 | Lezgian | 0.62 | Mandarin | 0.65 |
| Abkhaz | 0.55 | *Mixed Ma'a/Mbugu* | 0.62 | Maybrat | 0.65 |
| *African American English* | 0.55 | *Negerhollands* | 0.62 | *Palenquero* | 0.65 |
| *Cape Verd. Cr. of Santiago* | 0.55 | Otomí (Mezquital) | 0.62 | *Papi* | 0.65 |
| Kobon | 0.55 | *Papiamentu* | 0.62 | *Principense* | 0.65 |
| *Louisiana Creole* | 0.55 | Persian | 0.62 | *Sango* | 0.65 |
| *Mauritian Creole* | 0.55 | Quechua (Imbabura) | 0.62 | *Singapore Bazaar Malay* | 0.65 |
| *Pidgin Hawaiian* | 0.55 | Russian | 0.62 | *Ternate Chabacano* | 0.65 |
| *Reunion Creole* | 0.55 | Sango | 0.62 | Turkish | 0.65 |
| Shipibo-Konibo | 0.55 | *Sri Lankan Malay* | 0.62 | Abun | 0.67 |
| Mixtec (Chalcatongo) | 0.56 | *Tayo* | 0.62 | Awa Pit | 0.67 |
| *Michif* | 0.57 | Thai | 0.62 | Bulgarian | 0.67 |
| Greek (Modern) | 0.58 | *Tok Pisin* | 0.62 | Cantonese | 0.67 |
| Kannada | 0.58 | Acehnese | 0.63 | Mapudungun | 0.67 |
| *Korlai* | 0.58 | Ainu | 0.63 | Quechua (Huallaga) | 0.67 |
| Maori | 0.58 | *Bahamian Creole* | 0.63 | English | 0.68 |
| *Martinican Creole* | 0.58 | Basque | 0.63 | *Ghanaian Pidgin English* | 0.68 |
| *Nicaraguan Creole English* | 0.58 | Buduma | 0.63 | *Nengee* | 0.68 |
| *Pidgin Hindustani* | 0.58 | Chamorro | 0.63 | *Nigerian Pidgin* | 0.68 |
| *San Andres Creole English* | 0.58 | *Chinese Pidgin Russian* | 0.63 | Yaqui | 0.68 |
| *Zamboanga Chabacano* | 0.58 | *Fanakalo* | 0.63 | *Angolar* | 0.69 |
| *Afrikaans* | 0.59 | Greenlandic (West) | 0.63 | *Chinuk Wawa* | 0.69 |
| Bagirmi | 0.59 | *Gullah* | 0.63 | Evenki | 0.69 |
| Bawm | 0.59 | *Hawai'i Creole* | 0.63 | Georgian | 0.69 |
| Berber (Middle Atlas) | 0.59 | Kanuri | 0.63 | *Guinea-Bissau Kriyol* | 0.69 |
| *Chinese Pidgin English* | 0.59 | *Kikongo-Kituba* | 0.63 | *Kriol* | 0.69 |
| Coos (Hanis) | 0.59 | *Krio* | 0.63 | *Media Lengua* | 0.69 |
| Epena Pedee | 0.59 | Lakhota | 0.63 | Oromo (Harar) | 0.69 |
| Ma'di | 0.59 | Lango | 0.63 | *Santome* | 0.69 |
| Romanian | 0.59 | Latvian | 0.63 | *Sranan* | 0.69 |
| Slave | 0.59 | Navajo | 0.63 | Taba | 0.69 |
| *Trinidad English Creole* | 0.59 | Nkore-Kiga | 0.63 | *Belizean Creole* | 0.71 |
| Vietnamese | 0.59 | O'odham | 0.63 | Haida | 0.71 |
| Yukaghir (Kolyma) | 0.59 | Rawang | 0.63 | *Norf'k* | 0.71 |
| Nahuatl (Tetelcingo) | 0.60 | *Sri Lanka Portuguese* | 0.63 | Somali | 0.71 |
| Amele | 0.62 | Tetun | 0.63 | Supyire | 0.71 |
| *Berbice Dutch* | 0.62 | Tidore | 0.63 | *Vincentian Creole* | 0.71 |
| *Bislama* | 0.62 | Wichí | 0.63 | *Saramaccan* | 0.74 |
| *Cape Verd. Cr. of Brava* | 0.62 | Wolof | 0.63 | *Cameroon Pidgin English* | 0.75 |
| *Cape Verd. Cr. of São Vicente* | 0.62 | *Early Sranan* | 0.64 | *Pichi* | 0.76 |
| *Casamancese Creole* | 0.62 | French | 0.64 | | |
| *Cavite Chabacano* | 0.62 | *Singlish* | 0.64 | | |
| *Diu Indo-Portuguese* | 0.62 | Spanish | 0.64 | | |
| *Fa d'Ambô* | 0.62 | *Ambon Malay* | 0.65 | | |

Table 3: Average syntagmatic complexity by language, sorted from lowest to highest

languages is 0.61 ($\pm$0.06), with p $\approx$ 0.04 by Welch's two-sample *t*-test as implemented in R (R Core Team 2013) across the complexity scores for the APiCS and WALS language sets.[18]

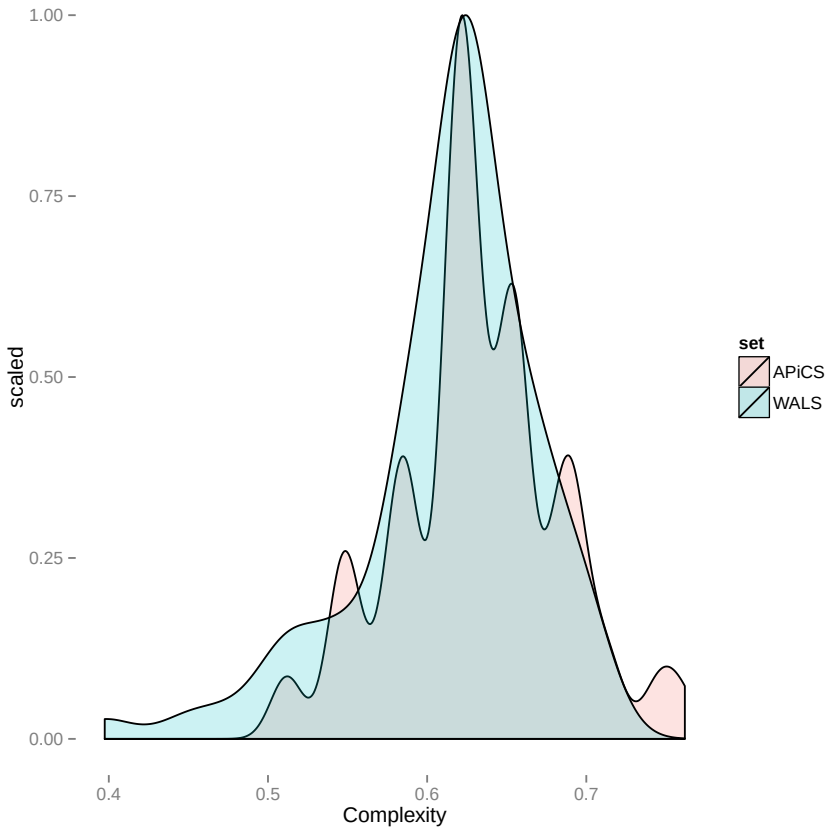


Figure 3: Density plot of syntagmatic complexity distribution across APiCS and WALS languages

I have no particular account for the slightly higher mean syntagmatic complexity scores for the APiCS languages than the WALS languages, and given the limitations on the datasets examined here with respect the capturing of a language's patterns of syntagmatic complexity, I hesitate to read much into this result beyond the important fact that syntagmatic complexity appears to pattern quite differently from paradigmatic complexity. Recall that, for paradigmatic complexity, as discussed in Section 4.3, the APiCS scores were not merely lower on average than the WALS languages, but

[18] It should be noted here that, despite the fact that this difference in averages crosses over into the standard range of statistical significance, this is the only major result reported on here which was not robust during various aspects of recoding and reprocessing that took place to ensure consistency across examined features and to correct for errors. While the APiCS languages consistently were scored higher in terms of syntagmatic complexity than WALS languages, earlier variations of the coding resulted in distributions that fell above above the standard p $\leq$ 0.05 significance threshold. Therefore, it seems inappropriate to read much into this specific result merely due to its statistical significance.

this was also a very robust statistical result, with a probability of the two sets of scores reflecting some underlying equivalence being effectively of zero. On a broad level, then, the prediction of a complexity asymmetry as proposed in Good (2012) seems to be verified, though this does not mean we can say that we fully understand the entire picture (see Section 5 for further discussion). Some indication of the features that are most important in the result that APiCS languages are scored slightly more syntagmatically complex than WALS languages can be found in the first section of Table 1, where those features which were more complex for APiCS languages than WALS languages are listed, six of which are treated here as syntagmatic in nature.

The discussion to this point has focused on tests of the recoded WALS–APiCS directly prompted by the predictions of Good (2012). In the next section, I will look at the data in a post hoc fashion to see which of the various paradigmatic features used in this study may most effectively predict whether a language is a member of the APiCS or WALS dataset as a means for detecting potentially interesting subpatterns for further investigation of patterns of complexity in contact languages.

### 4.5 The most predictive paradigmatic features

Good (2012) made an overall prediction regarding two classes of complexities, paradigmatic and syntagmatic. However, it is clearly not the case that the full range of complexity patterns in creoles and pidgins can be explained with a simple two-way division. One question of interest is whether or not some of the WALS–APiCS complexity values seen here may somehow be more typically "creole" or "non-creole" than others.[19] Knowing this would be helpful in allowing us to further refine our models of creole typology. In order to examine this issue, a generalized linear model was constructed where the complexity scores across a given feature were treated as potential predictors of a language's membership in the APiCS or WALS dataset.[20] Only features associated with

[19] Daval-Markussen (2014) addresses a roughly similar concern to the discussion in this section in trying to find typological features that uniquely identify creoles from non-creoles. His focus is determining whether there is a set of features that uniquely identifies creoles, while here I am merely interested in knowing which of the features examined are most effective for determining whether or not a language may have been creolized.

[20] The analysis described here was conducted using the built-in functionality of R (R Core Team 2013) for generalized linear regression assuming a binomial distribution for the value being predicted—i.e., here, the language set the language belongs to. A drawback of this choice is that it does not account for the fact that there may be within-language correlations across some feature-value pairings. However, it was not possible to include language as a predictive fac-

| FEATURE | DESCRIPTION | MAX | COMPLEXITY | EST | SE | $\approx$ p |
|---|---|---|---|---|---|---|
| WALS 109A | Applicative constructions | 4 | APiCS < WALS | 2.62 | 0.67 | 0.00 |
| WALS 98A | Alignment of case marking in nouns | 3 | APiCS < WALS | 1.94 | 0.49 | 0.00 |
| WALS 71A | The prohibitive | 4 | APiCS < WALS | 1.90 | 0.41 | 0.00 |
| WALS 55A | Sortal numeral classifiers | 3 | APiCS < WALS | 1.43 | 0.63 | 0.02 |
| WALS 44A | Gender distinctions in pronouns | 4 | APiCS < WALS | 1.27 | 0.45 | 0.01 |
| WALS 54A | Adnominal distributive numerals | 3 | APiCS < WALS | 1.24 | 0.42 | 0.00 |
| WALS 52A | Comitatives and instrumentals | 2 | APiCS < WALS | 1.18 | 0.27 | 0.00 |
| WALS 101A | Expression of pronominal subjects | 3 | APiCS < WALS | 1.05 | 0.29 | 0.00 |
| WALS 39A | Inclusive/exclusive distinction | 2 | APiCS < WALS | 0.94 | 0.42 | 0.02 |
| WALS 45A | Politeness distinctions | 3 | APiCS $\approx$ WALS | 0.93 | 0.39 | 0.02 |
| WALS 41A | Distance contrasts in demonstratives | 5 | APiCS < WALS | 0.87 | 0.40 | 0.03 |
| WALS 38A | Indefinite articles | 3 | APiCS > WALS | -0.66 | 0.30 | 0.03 |
| WALS 63A | Noun phrase conjunction and comitative | 2 | APiCS $\approx$ WALS | 0.61 | 0.25 | 0.01 |

Table 4: Significant paradigmatic features for predicting a language's category

paradigmatic complexities were considered, and only languages with twenty-six or more feature values for these features (i.e., the same set of languages was considered as found in Table 2).

Results of this analysis are presented in Table 4. Only those features whose effect on categorization was determined to be significant based on a $p \leq 0.05$ threshold are included.[21] In addition to their approximate significance score, the table indicates the coefficient assigned to the feature (which, here, can be understood as an indicator of the strength of that feature for predicting whether a language is part of the APiCS or WALS set), as well as the standard errors of the coefficients. For purposes of reference, the table also gives the maximum possible complexity score assigned to a given feature (see the Appendix for further details) and whether that feature was found to differ in complexity across the APiCS and WALS languages as presented in Table 1 in Section 4.2. I should stress here that, to the best of my knowledge, the use of a generalized linear model in typological analysis of the sort undertaken here has not been done before, meaning that there is no standard way to interpret its results. I will therefore focus on the patterns that seem most robust and, therefore, less likely to disappear on the basis of small changes in the analytical procedure.

In examining the features in Table 4, two noteworthy generalizations emerge. First, consistent with the general claims of Good (2012), the features that are most diagnostic of a language's

tor or to include it as a random variable in a Generalized Linear Mixed Model (see Jaeger (2008) for discussion in a linguistic context) because of the fact that a language's membership in the WALS or APiCS dataset is completely predetermined, rather than being an independent "observation" for this dataset.

[21] Significance was determined via a Wald test as implemented in the R function **glm()**.

status as being part of the APiCS class or the WALS class strongly favor cases where the APiCS languages were found to be less complex overall than the WALS languages. Ten of the thirteen features in Table 4 follow this pattern, and only one is directly contrary to it (relating to the presence of indefinite articles) with two others being among the features where the quantitative test results summarized in Table 1 showed no significant difference.[22] Moreover, if we consider the features with the higher coefficients (and which, therefore, seem most likely to be truly informative rather than being artifacts of the construction of the model), these all involve cases where the APiCS languages have lower complexity scores than the WALS languages. These results can, therefore, be taken as evidence that a distinctive feature of creoles is specifically that they are paradigmatically simple (rather, than, for instance, being simpler in some areas and more complex than others).

The second generalization that emerges from Table 4 is that the most informative features of a language's membership in the APiCS and WALS set strongly tend to be morphological in nature. While it is the case that many of the paradigmatic features examined in this study are morphologically oriented (see Table 5 in the Appendix), phonological and syntactic paradigmatic features were also examined. Nevertheless, all of the features in Table 4, except for one (relating to the expression of pronominal subjects (Dryer 2013a)), either target canonically morphosyntactic domains (such as applicatives or case marking) or involve contrasts within a morphological paradigm (such as politeness distinctions or demonstratives for different distances). This suggests that there may be special pressures on morphological paradigmatic complexity in the formation of pidgins and creoles, which is perhaps not all that surprising considering that morphology is the domain of grammar which probably allows for the highest degree of paradigmatic complexity in the first place (see also Section 4.6). Of course, as discussed in Section 1, relatively reduced morphology

[22] The presence of features which were not found to be significantly distinctive in earlier tests being treated as significant factors in the model that was produced in this section suggest that there may be some degree of overfitting—that is, the model is too dependent on idiosyncrasies of this dataset rather than reflecting actual distributions. This means that some of the features deemed to be significant in Table 4 may not be truly predictive. I do not view this as especially problematic here since the goal of this analysis is not to actually develop a predictive model of whether a language should or should not be classified as a "creole" but, rather, to get a general sense of which of the paradigmatic WALS–APiCS features are most informative in this regard, and I assume that those with relatively high coefficients in Table 4 are likely to represent genuinely informative features.

has long been considered a feature of creoles. So, this should not be viewed as a new result but, rather, a verification of earlier ideas using new methods and a comparatively large dataset.

The construction of a generalized linear model here, therefore, both corroborates the earlier results that creoles appear to be simpler in paradigmatic terms than non-creoles and further suggests that the simplification may not be equally distributed in grammatical terms, instead targeting morphological patterns more strongly than phonological or syntactic ones. However, this latter conclusion should probably be considered tentative, given that the WALS–APiCS dataset is not specifically designed to be balanced across grammatical domains.

### 4.6 Robustness and validity

I would like to conclude this discussion of the quantitative results with some brief comments regarding the likely robustness of the findings and the extent to which the WALS–APiCS data may or may not be providing a clear window onto actual patterns of paradigmatic and syntagmatic complexity in contact and non-contact languages.

First, if we consider the narrow question of how robust the results reported in the previous section are with respect to the specific data considered, on the whole, they do seem to be fairly robust. For instance, while not treated as part of the formal analytical process, at various points in the investigation, the complexity scores for a given feature value were recoded in order to ensure that the coding criteria were consistently applied to different features or to correct obvious mistakes. Such recoding, however, never changed the major results reported above, only affecting the quantitative figures quite marginally. Similarly, changes to the details of the sets of languages considered did not change the results in major ways. The two clear complexity outliers, Sri Lanka Portuguese and Michif (see Section 4.3) were removed from the APiCS dataset at one point to see how much they affected the overall quantitative complexity patterns, but this, too, had no noteworthy effect.

The only case where a change of this kind did seem to have an impact worth specifically noting involved whether or not the languages classified as pidgins and creoles in the WALS data set were included or excluded from the comparison. In this case, removing them shifted the higher

degree of syntagmatic complexity found for APiCS languages just into the realm of conventional statistical significance (see Section 4.4), whereas before the p-value was slightly over 0.1. While this difference in syntagmatic complexity is interesting and not directly predicted by Good (2012), it is still quite consistent with its main arguments that paradigmatic and syntagmatic complexities should behave differently and there should be simplification in the paradigmatic domain.

Therefore, with the exception of the results relating to syntagmatic complexity, I believe the results discussed here are a fairly accurate reflection of complexity patterns in the WALS–APiCS dataset. However, there is the more difficult question as to whether or not the WALS–APiCS dataset itself properly represents the underlying complexity of non-contact and contact languages. It is always possible to find fault with any typological sample. In the APiCS case, for instance, there is the problem—a persistent one in studies of contact languages—that there is a strong bias in the language set towards European-lexifier creoles and, especially, English-based creoles (see also Michaelis et al. (2013a: xxxxviii)), which is clearly not ideal for a study such as this one.[23] Another issue that arises is that the WALS classifications were done by those examining many languages at a time for one typological feature, while the APiCS classifications were done by language specialists examining many features at a time. We can anticipate coding discrepancies from this, but I am not aware of any clear means for controlling for them.

A more particularized problem to the present study relates to the fact that the "binning" of feature values in the WALS–APiCS dataset (largely a carryover from WALS itself) leads to a much less nuanced view of paradigmatic complexity, in particular, than would be ideal. For instance, the results shown in Table 1 suggest that an important distinguishing feature of creoles is the comparative lack of use of tone as compared to non-creoles. The feature values only allow for a three-way distinction for languages having no tone, a "simple" tone system (i.e., generally two tones), or a "complex" tone system (i.e., more than two) (Maddieson 2013). We can assume that, if the typological classifications were more fine-grained for the WALS–APiCS datasets, APiCS

[23] Future investigations could, perhaps, examine paradigmatic complexities within various language groupings (e.g., genealogical, areal, etc.) in APiCS and WALS to see if any significant generalizations arise, but this is outside the scope of the present study.

languages would still be less complex with respect to tone, but, perhaps the differences in the complexity scores would have been even starker.

Moreover, certain WALS features that strongly target paradigmatic contrasts, such as one for the number of genders in a language (Corbett 2013), are missing from the WALS–APiCS dataset. In the end, what we are lacking is a database whose features are specifically designed to target paradigmatic and syntagmatic complexities. This is not a criticism of APiCS or WALS, but, rather, just an issue which limits the strength of the conclusions we can reach from this study. If anything, however, the differences in paradigmatic complexities between creoles and non-creoles are probably lower here than they would be if a more targeted dataset were available due to the fact that some of the most elaborated types of morphological paradigms are not attested in creoles, which one would certainly want to code directly if one were building a new database to come to a better understanding of creole typology in general, and creole morphological typology in particular.

## 5 Testing creole typology with APiCS

We have seen in this study that, on the whole, the data available in WALS and APiCS is consistent with the claim of Good (2012) that, if creoles are "simple", they should be simple in a specific way—that is, with respect to paradigmatic complexity rather than syntagmatic complexity. At the same time, when looked at in detail, the results raise additional questions not considered by Good (2012). Are morphological paradigmatic complexities even more prone to being simple in contact languages than other kinds of paradigmatic complexities (see Section 4.5)? Could contact languages actually be syntagmatically more complex than non-contact languages and, if so, why (see Section 4.4)? And, how can we account for the fact that paradigmatic features are not uniformly simple, if simpler on the whole (see Section 4.2)?

Of course, some of the more specific results may not hold here up under closer scrutiny or if more appropriate data were available. I would, therefore, like to end this paper on a methodological point. To the extent that there is a particular original contribution to this paper in the domain of creole studies, I believe it is in the design and application of metrics of complexity emanating

from a specific, well-known (if controversial) model of creolization (see Section 2). This makes it different from previous similar work such as Parkvall (2008) or McWhorter (2001) where the complexity metrics were more general in nature (e.g., presence of a coding distinction being more complex than absence or a larger inventory of elements in some class being more complex than a smaller one). Of course, there is some overlap between those metrics and the ones used here, but, as a general point, if one wants to investigate whether the typology of a sociohistorically defined set of languages is distinct from another class of (implicitly or explicitly) sociohistorically defined languages, the conclusions will be more open to theoretical scrutiny, and useful for further analytical reference, to the extent that our system of "measuring" languages can be linked to the sociohistorical circumstances of interest.

This last point is true, of course, whether or not one agrees with the model summarized in Section 2, the specific metrics devised from that model, or the application of those metrics to the features of interest here. Fortunately, once a dataset like the WALS–APiCS one is made available, recoding it according to new metrics is a relatively small amount of work—indeed, the work involved is trivial compared to the work required to assemble the data in the first place. Further studies of the WALS–APiCS dataset testing whether the predictions of different models are a better fit for the data than the one used here, are, fortunately, now within easy reach.

Finally, we have seen in this paper that if we want to understand creole morphology, we may need to look past complexity in a general sense and think about different types of complexities. In so doing, we can move away from statements about the relative lack of "morphology", for instance, in creoles and, instead, think about what specific sorts of morphology are missing. In this case, it seems to be that morphology that can only be transmitted via paradigms may, in particular, be lacking. This observation can help us understand results such as those of Braun & Plag (2003) where more syntagmatic morphological strategies, such as compounding and reduplication, were relatively robustly attested in a creole, and it further opens the door to more nuanced views of morphological complexity in creoles than has been found in much of the previous literature, which has often assumed a more monolithic kind of simplicity/complexity.

## References


Bakker, Peter. 1997. *A language of our own: The genesis of Michif, the mixed Cree-French language of the Canadian Métis*. Oxford: OUP.

Bakker, Peter. 2003. Pidgin inflectional morphology and its implications for creole morphology. In Geert Booij & Jaap van Marle (eds.), *Yearbook of Morphology 2002*, 3–33. Dordrecht: Kluwer.

Bakker, Peter. 2013. Michif. In Michaelis et al. (2013c), 158–165.

Bakker, Peter. 2014. Creolistics: Back to square one? *Journal of Pidgin and Creole Languages* 29. 177–194.

Bakker, Peter, Aymeric Daval-Markussen, Mikael Parkvall & Ingo Plag. 2011. Creoles are typologically distinct from non-creoles. *Journal of Pidgin and Creole Languages* 26. 5–42.

Bickerton, Derek. 1984. The language bioprogam hypothesis. *Behavioral and Brain Sciences* 7. 173–188.

Braun, Maria & Ingo Plag. 2003. How transparent is creole morphology? A study of Early Sranan word-formation. In Geert Booij & Jaap van Marle (eds.), *Yearbook of Morphology 2002*, 81–104. Dordrecht: Kluwer.

Comrie, Bernard. 2013. *Alignment of case marking of full noun phrases*. In (Dryer & Haspelmath 2013). http://wals.info/chapter/98.

Corbett, Greville G. 2007. Canonical typology, suppletion, and possible words. *Language* 83. 8–42.

Corbett, Greville G. 2013. Number of genders. In Dryer & Haspelmath (2013). http://wals.info/chapter/30.

Croft, William. 2000. *Explaining language change: An evolutionary approach*. Harlow, England: Longman.

Daval-Markussen, Aymeric. 2014. First steps towards a typological profile of creoles. *Acta Linguistica Hafniensia* 46. 1–22.

DeGraff, Michel. 2005. Linguists' most dangerous myth: The fallacy of Creole Exceptionalism. *Language in Society* 34. 533–591.

Dryer, Matthew S. 1989. Plural words. *Linguistics* 27. 865–895.

Dryer, Matthew S. 2013a. Expression of pronominal subjects. In Dryer & Haspelmath (2013). http://wals.info/chapter/101.

Dryer, Matthew S. 2013b. Negative morphemes. In Dryer & Haspelmath (2013). http://wals.info/chapter/112.

Dryer, Matthew S. & Martin Haspelmath (eds.). 2013. *The World Atlas of Language Structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wals.info/.

Gil, David. 2013a. Numeral classifiers. In Dryer & Haspelmath (2013). http://wals.info/chapter/55.

Gil, David. 2013b. Para-linguistic usages of clicks. In Dryer & Haspelmath (2013). http://wals.info/chapter/142.

Good, Jeff. 2012. Typologizing grammatical complexities *or* why creoles may be paradigmatically simple but syntagmatically average. *Journal of Pidgin and Creole Languages* 27. 1–47.

Jaeger, T. Florian. 2008. Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language* 59. 434–446.

Karlsson, Fred, Matti Miestamo & Kaius Sinnemäki. 2008. Introduction: The problem of language complexity. In Miestamo et al. (2008), vii–xiv.

Kouwenberg, Silvia. 2010a. Creole studies and linguistic typology: Part 1. *Journal of Pidgin and Creole Languages* 25. 173–186.

Kouwenberg, Silvia. 2010b. Creole studies and linguistic typology: Part 2. *Journal of Pidgin and Creole Languages* 25. 359–380.

Maddieson, Ian. 2013. Tone. In Dryer & Haspelmath (2013). http://wals.info/chapter/13.

Maurer, Philippe & The APiCS Consortium. 2013. Sortal numeral classifiers. In Michaelis et al. (2013a), 138–139.

McWhorter, John H. 2001. The world's simplest grammars are creole grammars. *Linguistic Typology* 5. 125–166.

McWhorter, John H. 2014. A response to Mufwene. *Journal of Pidgin and Creole Languages* 29. 172–176.

McWhorter, John H. & Jeff Good. 2012. *A grammar of Saramaccan Creole*. Berlin: De Gruyter Mouton.

Meakins, Felicity. 2013. Gurindji Kriol. In Michaelis et al. (2013c), 131–139.

Michaelis, Susanne Maria, Philippe Maurer, Martin Haspelmath & Magnus Huber (eds.). 2013a. *The Atlas of Pidgin and Creole Language Structures*. Oxford: OUP.

Michaelis, Susanne Maria, Philippe Maurer, Martin Haspelmath & Magnus Huber (eds.). 2013b. *The Survey of Pidgin and Creole Language Structures* (three volumes). Oxford: OUP.

Michaelis, Susanne Maria, Philippe Maurer, Martin Haspelmath & Magnus Huber (eds.). 2013c. *The Survey of Pidgin and Creole Language Structures: Volume III: Contact languages based on languages from Africa, Asia, Australia, and the Americas*. Oxford: OUP.

Michaelis, Susanne Maria, Philippe Maurer, Martin Haspelmath & Magnus Huber (eds.). 2013d. *The Atlas of Pidgin and Creole Language Structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://apics-online.info/.

Miestamo, Matti, Kaius Sinnemäki & Fred Karlsson (eds.). 2008. *Language complexity*. Amsterdam: Benjamins.

Moravcsik, Edith A. & Jessica R. Wirth. 1986. Markedness: An overview. In Fred R. Eckman, Edith A. Moravcsik & Jessica R. Wirth (eds.), *Markedness*, 1–11. New York:Plenum.

Mous, Maarten. 2013. Mixed Ma'a/Mbugu. In Michaelis et al. (2013c), 42–49.

Mufwene, Salikoko S. 2001. *The ecology of language evolution*. Cambridge: CUP.

Mufwene, Salikoko S. 2014. The case was never closed: McWhorter misinterprets the ecological approach to the emergence of creoles. *Journal of Pidgin and Creole Languages* 29. 157–171.

Muysken, Pieter. 2013. Media Lengua. In Michaelis et al. (2013c), 144–148.

Parkvall, Mikael. 2008. The simplicity of creoles in a cross-linguistic perspective. In Miestamo et al. (2008), 265–285.

Plag, Ingo. 2008. Creoles as interlanguages: Inflectional morphology. *Journal of Pidgin and Creole Languages* 23. 114–135.

Plag, Ingo. 2009. Creoles as interlanguages: Word-formation. *Journal of Pidgin and Creole Languages* 24. 339–362.

R Core Team. 2013. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. http://www.R-project.org/.

Sampson, Geoffrey, David Gil & Peter Trudgill (eds.). 2009. *Language complexity as an evolving variable*. Oxford: OUP.

Siegel, Jeff. 2004. Morphological simplicity in pidgins and creoles. *Journal of Pidgin and Creole Languages* 19. 139–162.

Siegel, Jeff, Benedikt Szmrecsanyi & Bernd Kortmann. 2014. Measuring analyticity and syntheticity in creoles. *Journal of Pidgin and Creole Languages* 29. 49–85.

Sinnemäki, Kaius. 2011. *Language universals and linguistic complexity: Three case studies in core argument marking*. University of Helsinki doctoral dissertation.

Smith, Ian. 1977. *Sri Lanka Creole Portuguese phonology*. Ithaca, NY: Cornell University Ph.D. dissertation.

Smith, Ian. 2013. Sri Lanka Portuguese. In Susanne Maria Michaelis, Philippe Maurer, Martin Haspelmath & Magnus Huber (eds.), *The Survey of Pidgin and Creole Language Structures: Volume III: Portuguese-based, Spanish-based, and French-based languages*, 111–121. Oxford: OUP.

Stassen, Leon. 2013a. Comparative constructions. In Dryer & Haspelmath (2013). http://wals.info/chapter/121.

Stassen, Leon. 2013b. Predicative possession. In Dryer & Haspelmath (2013). http://wals.info/chapter/117.

Stolz, Thomas & Ljuba N. Veselinova. 2013. Ordinal numerals. In Dryer & Haspelmath (2013). http://wals.info/chapter/53.

Veenstra, Tonjes. 2008. Creole genesis: The impact of the Language Bioprogram Hypothesis. In Silvia Kouwenberg & John Victor Singler (eds.), *The handbook of pidgin and creole studies*, 219–241. Chichester, West Sussex: Blackwell.

Wickham, Hadley. 2009. *ggplot2: Elegant graphics for data analysis*. New York: Springer.

Winford, Donald. 2006. Reduced syntax in (prototypical) pidgins. In Ljiljana Progovac, Kate Paesani, Eugenia Casielles & Ellen Barton (eds.), *The syntax of nonsententials: Multidisciplinary perspectives*, 283–307. Amsterdam: Benjamins.

## Appendix: Complexity scores for feature values

In Table 5 and Table 6 below I give the complexity codings assumed for the values of the paradigmatic and syntagmatic features used in this study. Full discussion of the features can be found in Dryer & Haspelmath (2013) and Michaelis et al. (2013d). I have, in some cases, shortened the names of the features and values for purposes of presentation. All feature identifiers are similarly shortened to a single number rather than including the prefix *WALS* or the following *A* identifier.

Each table lists the WALS identifiers, the feature names, the feature values, the complexity scores assigned to those values, and brief justifications of those assignments for the paradigmatic features (Table 5) and the syntagmatic feature (Table 6). The abbreviation LGM refers to *lingueme* (see Section 2 for use of the term in this context). The abbreviation IHRC refers to *internally headed relative clause*. The compound *benefactive-plus* for feature WALS 109A in Table 5 refers to "benefactive plus other functions".

The complexity scores assigned to a given value represent the minimum possible score for any language which can be assigned that value even if a higher score might apply to specific languages. The reference to an "optionality lingueme" refers to the fact that, for any pattern to be established as optional, there must be at least one lingueme instantiating a pattern in a given context and another not showing the pattern, hence increasing the paradigmatic complexity by one. In some cases, the choice of how to code a value's complexity was based on an examination of the examples found in the relevant WALS chapter (see, e.g., feature WALS 24A in Table 6).

The complexity encodings relate only to the complexity described by the relevant feature even if some other complexity is implied in some way. Thus, for instance, for the feature WALS 106A in Table 5, regarding reflexives and reciprocals, if the reciprocal is identical to the reflexive, this is treated as having zero paradigmatic complexity, since there is no distinction in that domain, even though there is presumably paradigmatic complexity elsewhere in the system to establish the reflexive/non-reflexive distinction. Furthermore, a zero for a paradigmatic complexity score does not mean that no transfer was required, but rather that no paradigmatic complexity was specifically transferred.

Finally, the complexity scores here do not take into account the possibility that negative evidence might be required for a certain pattern to be transferred (e.g., if only a benefactive applicative is found, this may be because no instrumental applicative was ever attested), reflecting this paper's emphasis on language as a communal object rather than a cognitive one. A cognitively-oriented study would require a different system of metrics.

| ID | FEATURE NAME | VALUE NAME | N | JUSTIFICATION |
|---|---|---|---|---|
| 10 | Nasal vowels | Contrast absent | 0 | No distinction, no transfer |
| 10 | | Contrast present | 2 | LGM pair evincing the contrast |
| 13 | Tone | No tones | 0 | No distinction, no transfer |
| 13 | | Simple tone system | 2 | One LGM for each tone |
| 13 | | Complex tone system | 3 | One LGM for each tone |
| 33 | Expression of nominal plural meaning | No plural | 0 | No distinction, no transfer |
| 33 | | Plural clitic | 2 | LGM pair for singular/plural |
| 33 | | Plural complete reduplication | 2 | LGM pair for singular/plural |
| 33 | | Plural prefix | 2 | LGM pair for singular/plural |
| 33 | | Plural stem change | 2 | LGM pair for singular/plural |
| 33 | | Plural suffix | 2 | LGM pair for singular/plural |
| 33 | | Plural tone | 2 | LGM pair for singular/plural |
| 33 | | Plural word | 2 | LGM pair for singular/plural |
| 33 | | Mixed morphological plural | 4 | Two LGM pairs for mixed pattern |
| 34 | Occurrence of nominal plural markers | No nominal plural | 0 | No distinction, no transfer |
| 34 | | Obligatory | 2 | LGM pair for plurality |
| 34 | | Always optional | 3 | LGM pair for plurality + optionality LGM |
| 34 | | Only on human nouns | 4 | Two LGM pairs for animacy |
| 34 | | Optional on human nouns | 5 | Two LGM pairs for animacy + optionality LGM |
| 34 | | Optional on inanimates | 5 | Two LGM pairs for animacy + optionality LGM |
| 37 | Definite articles | No definite or indefinite | 0 | No distinction, no transfer |
| 37 | | No definite, but indefinite | 0 | No distinction, no transfer |
| 37 | | Definite affix | 2 | LGM pair for definite marking |
| 37 | | Demonstrative as definite | 2 | LGM pair for definite marking |
| 37 | | Distinct definite | 3 | LGM pair for definite + LGM for demonstrative |
| 38 | Indefinite articles | No definite or indefinite | 0 | No distinction, no transfer |
| 38 | | No indefinite, but definite | 0 | No distinction, no transfer |
| 38 | | Indefinite affix | 2 | LGM pair for indefinite marking |
| 38 | | Indefinite same as 'one' | 2 | LGM pair for indefinite marking |
| 38 | | Distinct indefinite | 3 | LGM pair for indefinite + LGM for 'one' |
| 39 | Inclusive/exclusive distinction | No inclusive/exclusive | 0 | No distinction, no transfer |
| 39 | | No 'we' | 0 | No distinction, no transfer |
| 39 | | 'We' the same as 'I' | 0 | No distinction, no transfer |
| 39 | | Inclusive/exclusive | 2 | LGM pair for clusivity distinction |
| 39 | | Only inclusive | 2 | LGM pair for inclusivity distinction |
| 41 | Distance contrasts in demonstratives | No distance contrast | 0 | No distinction, no transfer |
| 41 | | Two-way contrast | 2 | Two LGMs for two-way contrast |
| 41 | | Three-way contrast | 3 | Three LGMs for three-way contrast |
| 41 | | Four-way contrast | 4 | Four LGMs for four-way contrast |
| 41 | | Five (or more)-way contrast | 5 | Five LGMs for five-way contrast |

Table 5: Complexity scores assigned to values for the paradigmatic features in the WALS–APiCS data (continued)

| ID | FEATURE NAME | VALUE NAME | N | JUSTIFICATION |
|---|---|---|---|---|
| 42 | Pronominal/adnominal demonstratives | Identical | 0 | No distinction, no transfer |
| 42 | | Different inflection | 2 | At least two LGMs for distinction |
| 42 | | Different stem | 2 | At least two LGMs for distinction |
| 44 | Gender distinctions in pronouns | No gender distinctions | 0 | No distinction, no transfer |
| 44 | | 1st/2nd person but not 3rd | 2 | LGM pair for minimal gender distinction |
| 44 | | 3rd person non-singular | 2 | LGM pair for minimal gender distinction |
| 44 | | 3rd person singular | 2 | LGM pair for minimal gender distinction |
| 44 | | 3rd person + 1st/2nd | 4 | Two LGM pairs, 3rd person/other person |
| 44 | | 3rd person, singular/plural | 4 | Two LGM pairs, 3rd person singular/plural |
| 45 | Politeness distinctions | No distinction | 0 | No distinction, no transfer |
| 45 | | Binary distinction | 2 | Two LGMs for two-way contrast |
| 45 | | Multiple distinctions | 3 | Minimum three LGMs for non-binary contrast |
| 45 | | Pronouns avoided | 3 | Pronouns replaced, assuming at least three nouns |
| 46 | Indefinite pronouns | Generic-noun-based | 0 | No distinction, no transfer |
| 46 | | Interrogative-based | 0 | No distinction, no transfer |
| 46 | | Existential construction | 2 | Minimal syntactic paradigm |
| 46 | | Special | 3 | Three LGMs for indef./generic/interr. distinction |
| 46 | | Mixed | 4 | Two LGM pairs for mixed pattern |
| 47 | Intensifiers and reflexive pronouns | Identical | 0 | No distinction, no transfer |
| 47 | | Differentiated | 2 | LGM pair for basic distinction |
| 52 | Comitatives and instrumentals | Identity | 0 | No distinction, no transfer |
| 52 | | Differentiation | 2 | LGM pair for basic distinction |
| 52 | | Mixed | 2 | LGM pair for basic distinction; mixing functional |
| 53 | Ordinal numerals | None | 0 | No distinction, no transfer |
| 53 | | One, two, three | 0 | No distinction, no transfer |
| 53 | | First, two, three | 2 | LGM pair for 'one'/'first' |
| 53 | | One-th, two-th, three-th | 2 | LGM pair for basic two-way distinction |
| 53 | | First/one-th, two-th, three-th | 3 | LGM pair for basic distinction + 'first' |
| 53 | | First, two-th, three-th | 4 | Two LGM pairs for basic pattern + 'one'/'first' |
| 53 | | First, second, three-th | 6 | Three LGM pairs for basic + 'two' + 'one' |
| 53 | | Various | 6 | Mixed group; assigning it upper bound of others |
| 54 | Adnominal distributive numerals | No distributive numerals | 0 | No distinction, no transfer |
| 54 | | Following word | 2 | LGM pair for basic distinction |
| 54 | | Preceding word | 2 | LGM pair for basic distinction |
| 54 | | Prefix | 2 | LGM pair for basic distinction |
| 54 | | Reduplication | 2 | LGM pair for basic distinction |
| 54 | | Suffix | 2 | LGM pair for basic distinction |
| 54 | | Mixed or other strategies | 3 | Mixed group; assuming at least one extra LGM |
| 55 | Sortal numeral classifiers | Absent | 0 | No distinction, no transfer |
| 55 | | Obligatory | 2 | LGM pair for minimal system |
| 55 | | Optional | 3 | LGM pair for minimal system + optionality LGM |
| 63 | NP conjunction and comitative | 'And' identical to 'with' | 0 | No distinction, no transfer |
| 63 | | 'And' different from 'with' | 2 | LGM pair for basic distinction |
| 64 | Nominal and verbal conjunction | Identity | 0 | No distinction, no transfer |
| 64 | | Juxtaposition | 0 | No distinction, no transfer |
| 64 | | Differentiation | 2 | LGM pair for basic distinction |
| 71 | The prohibitive | Normal imperative/negative | 0 | No distinction, no transfer |
| 71 | | Special imperative | 2 | LGM pair for special imperative |
| 71 | | Special negative | 2 | LGM pair for special negative |
| 71 | | Special imperative/negative | 4 | Two LGM pairs; for imperative and negative |

Table 5: Complexity scores assigned to values for the paradigmatic features in the WALS–APiCS data (continued)

| ID | FEATURE NAME | VALUE NAME | N | JUSTIFICATION |
|---|---|---|---|---|
| 79 | Suppletion for tense and aspect | None | 0 | No distinction, no transfer |
| 79 | | Aspect | 2 | LGM pair for aspect |
| 79 | | Tense | 2 | LGM pair for tense |
| 79 | | Tense and aspect | 4 | Two LGM pairs; for tense and aspect |
| 98 | Case marking of full noun phrases | Neutral | 0 | No distinction, no transfer |
| 98 | | Ergative–absolutive | 2 | LGM pair for intransitive/transitive pattern |
| 98 | | Marked nominative | 2 | LGM pair for intransitive/transitive pattern |
| 98 | | Nominative–accusative | 2 | LGM pair for intransitive/transitive pattern |
| 98 | | Tripartite | 2 | LGM pair for intransitive/transitive pattern |
| 98 | | Active-inactive | 3 | LGM pair for intrans. split + trans. LGM |
| 99 | Case marking of personal pronouns | Neutral | 0 | No distinction, no transfer |
| 99 | | None | 0 | This is difficult since it |
| 99 | | Ergative–absolutive | 2 | LGM pair for intransitive/transitive pattern |
| 99 | | Marked nominative | 2 | LGM pair for intransitive/transitive pattern |
| 99 | | Nominative–accusative | 2 | LGM pair for intransitive/transitive pattern |
| 99 | | Tripartite | 2 | LGM pair for intransitive/transitive pattern |
| 99 | | Active–inactive | 3 | LGM pair for intrans. split + trans. LGM |
| 101 | Expression of pronominal subjects | Obligatory | 0 | No distinction, no transfer |
| 101 | | Special position | 2 | LGM pair for basic syntactic distinction |
| 101 | | Subject affixes | 2 | LGM pair for basic syntactic distinction |
| 101 | | Subject clitics | 2 | LGM pair for basic syntactic distinction |
| 101 | | Mixed | 3 | Minimum of three LGMs for mixed pattern |
| 101 | | Optional | 3 | LGM pair for presence + optionality LGM |
| 105 | Ditransitive constructions with ‘give’ | Double object | 2 | LGM pair for basic transitive/ditransitive |
| 105 | | Indirect object | 2 | LGM pair for basic transitive/ditransitive |
| 105 | | Secondary object | 2 | LGM pair for basic transitive/ditransitive |
| 105 | | Mixed | 3 | Minimum of three LGMs for mixed pattern |
| 106 | Reciprocal constructions | Identical to reflexive | 0 | No distinction, no transfer |
| 106 | | No reciprocals | 0 | No distinction, no transfer |
| 106 | | Distinct from reflexive | 2 | LGM pair for basic distinction |
| 106 | | Mixed | 3 | Minimum of three LGMs for mixed pattern |
| 109 | Applicative constructions | No applicative construction | 0 | No distinction, no transfer |
| 109 | | Benefactive | 2 | LGM pair for presence/absence of applicative |
| 109 | | Benefactive in trans. | 2 | LGM pair for presence/absence of applicative |
| 109 | | Non-benefactive | 2 | LGM pair for presence/absence of applicative |
| 109 | | Non-benefactive in intrans. | 2 | LGM pair for presence/absence of applicative |
| 109 | | Non-benefactive in trans. | 2 | LGM pair for presence/absence of applicative |
| 109 | | Benefactive-plus | 4 | Two LGM pairs for two applicative functions |
| 109 | | Benefactive-plus in trans. | 4 | Two LGM pairs for two applicative functions |
| 119 | Predicative noun and locative phrases | Identical | 0 | No distinction, no transfer |
| 119 | | Different | 2 | LGM pair for basic distinction |
| 120 | Predicative noun phrases | Impossible | 0 | No distinction, no transfer |
| 120 | | Possible | 2 | LGM pair for basic distinction |
| 122 | Subject relative clauses | Gap | 0 | No distinction, no transfer |
| 122 | | Non-reduction | 0 | No distinction, no transfer |
| 122 | | Relative pronoun | 0 | No distinction, no transfer |
| 122 | | Pronoun-retention | 2 | LGM pair for relative/main clause distinction |
| 129 | ‘Hand’ and ‘arm’ | Identical | 0 | No distinction, no transfer |
| 129 | | Different | 2 | LGM pair for basic distinction |

Table 5: Complexity scores assigned to values for the paradigmatic features in the WALS–APiCS data

| ID | FEATURE NAME | VALUE NAME | N | JUSTIFICATION |
|---|---|---|---|---|
| 24 | Marking of possessor noun phrases | No marking | 0 | Zero dedicated morphemes |
| 24 | | Dependent marking | 1 | One coding device for syntagm |
| 24 | | Head marking | 1 | One coding device for syntagm |
| 24 | | Other | 1 | Examples show one coding device |
| 24 | | Double marking | 2 | Two coding devices for syntagm |
| 81 | Order of subject, object, and verb | No dominant order | 0 | Generally free order; no transfer |
| 81 | | OSV | 1 | One LGM to transfer one pattern |
| 81 | | OVS | 1 | One LGM to transfer one pattern |
| 81 | | SOV | 1 | One LGM to transfer one pattern |
| 81 | | SVO | 1 | One LGM to transfer one pattern |
| 81 | | VOS | 1 | One LGM to transfer one pattern |
| 81 | | VSO | 1 | One LGM to transfer one pattern |
| 85 | Order of adposition and noun phrase | No adpositions | 0 | No distinction, no transfer |
| 85 | | Inpositions | 1 | One LGM to transfer one pattern |
| 85 | | Postpositions | 1 | One LGM to transfer one pattern |
| 85 | | Prepositions | 1 | One LGM to transfer one pattern |
| 85 | | No dominant order | 2 | Generally split order; two patterns transferred |
| 86 | Order of possessor and possessum | Genitive-Noun | 1 | One LGM to transfer one pattern |
| 86 | | Noun-Genitive | 1 | One LGM to transfer one pattern |
| 86 | | No dominant order | 2 | Generally split order; two patterns transferred |
| 87 | Order of adjective and noun | No dominant order | 0 | Split or free order; coding as lower complexity |
| 87 | | Adjective-Noun | 1 | One LGM to transfer one pattern |
| 87 | | Internally-headed rel. clause | 1 | Classification unclear; coding as fixed pattern |
| 87 | | Noun-Adjective | 1 | One LGM to transfer one pattern |
| 88 | Order of demonstrative and noun | Mixed | 0 | Generally free order; no transfer |
| 88 | | Demonstrative prefix | 1 | One LGM to transfer one pattern |
| 88 | | Demonstrative suffix | 1 | One LGM to transfer one pattern |
| 88 | | Demonstrative-Noun | 1 | One LGM to transfer one pattern |
| 88 | | Noun-Demonstrative | 1 | One LGM to transfer one pattern |
| 88 | | Before and after Noun | 2 | Two coding devices for syntagm |
| 89 | Order of cardinal numeral and noun | No dominant order | 0 | Generally free order; no transfer |
| 89 | | Noun-Numeral | 1 | One LGM to transfer one pattern |
| 89 | | Numeral only modifies verb | 1 | Classification unclear; coding as fixed pattern |
| 89 | | Numeral-Noun | 1 | One LGM to transfer one pattern |
| 90 | Order of relative clause and noun | Mixed | 0 | Generally free order; no transfer |
| 90 | | Internally headed | 1 | One LGM to transfer one pattern |
| 90 | | Noun-Relative clause | 1 | One LGM to transfer one pattern |
| 90 | | Relative clause-Noun | 1 | One LGM to transfer one pattern |
| 90 | | Adjoined | 2 | Same as IHRC with additional complication |
| 90 | | Correlative | 2 | Same as IHRC with additional complication |
| 90 | | Doubly headed | 2 | Two coding devices for syntagm |
| 91 | Order of degree word and adjective | Adjective-Degree word | 1 | One LGM to transfer one pattern |
| 91 | | Degree word-Adjective | 1 | One LGM to transfer one pattern |
| 91 | | No dominant order | 2 | Generally split order; two patterns transferred |
| 93 | Position of interrogative phrases | Initial | 1 | One LGM to transfer one pattern |
| 93 | | Not initial | 1 | One LGM to transfer one pattern |
| 93 | | Mixed | 2 | Two LGMs to transfer a split |

Table 6: Complexity scores assigned to values for the syntagmatic features in the WALS–APiCS data (continued)

| ID | FEATURE NAME | VALUE NAME | N | JUSTIFICATION |
|---|---|---|---|---|
| 112 | Negative morpheme types | Negative affix | 1 | One coding device for syntagm |
| 112 | | Negative auxiliary verb | 1 | One coding device for syntagm |
| 112 | | Negative particle | 1 | One coding device for syntagm |
| 112 | | Negative word | 1 | One coding device for syntagm |
| 112 | | Negative word or affix | 1 | One coding device for syntagm |
| 112 | | Double negation | 2 | Two coding devices for syntagm |
| 115 | Negation and indefinite pronouns | No predicate negation | 1 | One coding device for syntagm |
| 115 | | Negative existential | 2 | Coding same as double to indicate extra complexity |
| 115 | | Predicate negation present | 2 | Two coding devices for syntagm |
| 115 | | Mixed behaviour | 3 | Two patterns; one with two coding devices |
| 116 | Polar questions | No distinction | 0 | No distinction, no transfer |
| 116 | | Intonation | 1 | One coding device for syntagm |
| 116 | | Lake of declarative coding | 1 | One coding device for syntagm |
| 116 | | Morphology | 1 | One coding device for syntagm |
| 116 | | Particle | 1 | One coding device for syntagm |
| 116 | | Word order | 1 | One coding device for syntagm |
| 116 | | Mixture of two types | 2 | Two coding devices for syntagm |
| 124 | 'Want' complement subjects | Desiderative particle | 1 | One coding device for syntagm |
| 124 | | Desiderative verbal affix | 1 | One coding device for syntagm |
| 124 | | Subject is expressed overtly | 1 | One coding device for syntagm |
| 124 | | Subject is left implicit | 1 | One coding device for syntagm |
| 124 | | Both construction types exist | 2 | Two patterns transferred |

Table 6: Complexity scores assigned to values for the syntagmatic features in the WALS–APiCS data