Modeling the Evolution of Creoles

Fredrik Jansson, Mikael Parkvall and Pontus Strimling

Linköping University Post Print

N.B.: When citing this work, cite the original article.

Original Publication:

Fredrik Jansson, Mikael Parkvall and Pontus Strimling, Modeling the Evolution of Creoles, 2015, Language Dynamics and Change, (5), 1, 1-51.

<http://dx.doi.org/10.1163/22105832-00501005>

Copyright: Brill Academic Publishers

<http://www.brill.com/>

Postprint available at: Linköping University Electronic Press

<http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-115745>

Modeling the Evolution of Creoles

Fredrik Jansson

*Institute for Analytical Sociology, Linköping University
Centre for the Study of Cultural Evolution, Stockholm University*

Mikael Parkvall

Department of Linguistics, Stockholm University

Pontus Strimling

*Institute for Analytical Sociology, Linköping University
Centre for the Study of Cultural Evolution, Stockholm University*

Various theories have been proposed regarding the origin of creole languages. Describing a process where only the end result is documented involves several methodological difficulties. In this paper we try to address some of the issues by using a novel mathematical model together with detailed empirical data on the origin and structure of Mauritian Creole. Our main focus is on whether Mauritian Creole may have originated only from a mutual desire to communicate, without a target language or prestige bias. Our conclusions are affirmative. With a confirmation bias towards learning from successful communication, the model predicts Mauritian Creole better than any of the input languages, including the lexifier French, thus providing a compelling and specific hypothetical model of how creoles emerge. The results also show that it may be possible for a creole to develop quickly after first contact, and that it was created mostly from material found in the input languages, but without inheriting their morphology.

Keywords *Creoles, Pidgins, Cultural Evolution, Mathematical Modeling, Simulation*

1 Introduction

What happens when people speaking different languages are brought together and there is a need to communicate? In many cases people can resort to an existing lingua franca, but in exceptional cases, an entirely new lingua franca develops – a pidgin¹ or a creole.

Creoles are languages which typically derive their vocabulary from an existing language (most commonly a Western European one), but whose structure differs from this (opinions vary on just how different they are). In most cases, they emerged in the wake of European overseas colonization during the past few centuries, and they are often associated with slavery.

The emergence of creole languages has received a great deal of scholarly attention for quite some time, but many issues still remain unresolved. We will address some of the main questions that are being debated by establishing some possible and impossible processes underlying the development of a common language.

We will mainly focus on the question of whether creoles are the result of imperfect second language acquisition. While a creole has the lion's share of its vocabulary in common with another language, its grammar is rather different. The most common assumption is that the slaves aimed at learning the language of their masters – which is French for our case study of the emergence of Mauritian – but that the rapid demographic development made this impossible. According to the standard account², the first batches of slaves may well have acquired most of the structural properties of French, but the more numerous the slaves became, the less exposure they would have had to the language spoken natively only by the ruling minority. They would therefore not have been able to acquire the finer details of the prestige language, and the result would have been roughly what we see today: a language which has inherited the lexicon, but not most of its grammar from that of the slave-owning group. On the face of it, this is a compelling view, but it implicitly makes two specific assumptions: first, that one language is more prestigious than the others and therefore a desirable target for learning, and second, that lexicon, phonology, and syntax evolve differently, something which makes the hypothesis difficult to test empirically.

Other creolists (e.g. Baker 1990, 1995a; Smith 2006) rather picture a sit-

¹ A pidgin is an extremely basic contact language with a vocabulary and a structure far more limited than that of any natively spoken language. It has limited complexity and limited expressiveness, as opposed to a creole, which is considered having the same expressive power as any other human language.

²This view is common enough to be featured in introductory textbooks, but is particularly salient in the works of Chaudenson and Mufwene (see the reference list for some examples)

uation in which everyone involved, including masters as well as slaves, made efforts to communicate in whatever linguistic material was at hand. The goal would thus only have been successful communication, rather than the acquisition of an existing (prestigious) language.

The difference between these two approaches may not be obvious at first sight, but in our use of the term, second language acquisition implies that there is a language to acquire in the first place, and that some people make an attempt to do so. Obviously, pidgin creation in some sense includes acquisition, in that phonetic strings get to be used by speakers who did not know them previously. However, in our view, pidginization – as opposed to second language acquisition – does not (necessarily) imply that anybody was trying to learn a pre-existing language, nor does it contrast a group of learners/receivers with a group of teachers/transmitters. While the traditional creologenetic scenario associates these groups with non-Europeans and Europeans respectively, we would rather emphasize that both groups are interested in communication, above all. Or to put it another way, a European having invested plenty of money in a slave would be interested in having him work, while his victim would try to make the best of a miserable situation; and these desires, we suggest, would override the urge to teach or learn the correct forms of past subjunctives.³ Thus, we believe that non-Francophones on Mauritius were not striving to acquire French, but that everyone present on the island – including Francophones – was searching for a way of communicating with the others, regardless of the origin of the building blocks of the emerging language.

Both theories have received a great deal of verbal argumentation, but, to our knowledge, fully rigorous treatment that could settle the case has been scarce. The first theory is certainly a possible one, depending on the specific assumptions made for the evolution of lexicon, phonology, and grammar, as well as the aims of the people involved. However, the second one is simpler in that it makes fewer assumptions. In adherence with Occam's razor, we ought to favor the simpler theory if (but only if) it makes predictions which fit our observations as well as the more complex one does. Our main mission is thus to answer the following question: can a creole develop without assuming a target language or a prestige bias?⁴ If so, under what circumstances?

We will also touch upon the following related questions:

³Please note that we are not claiming that the two processes are entirely exclusive, and that the difference is a matter of focus.

⁴Please note that even if the answer is yes, the first theory cannot be refuted only on these grounds, but we would show that its assumptions are not necessary to produce known properties of creole languages.

1. How fast do creoles develop? Some stipulate an abrupt emergence (Adone, 1994; Corne, 1999: 164; Hancock, 1987: 265; Jourdan and Keesing, 1997: 403; Lefebvre, 1993: 256; 1997: 79, Mühlhäusler, 1997: 54; Munteanu, 1996: 43; Owens, 1996: 135; Roberts, 2000; Smith, 2006), whereas others propose a protracted genesis, sometimes even stretching over more than a century (Alleyne, 1971; Arends, 1989: 253, Arends, 1993: 376; Bartens, 1996: 138; Mufwene, 2006a).
2. Do creoles develop from pidgin languages? This was once virtually uncontroversial, but has recently been questioned – most vocally by Mufwene (2000, 2002, 2006b, 2007, 2008a,b), but also by Chaudenson (1995: 66, 2003: 140), DeGraff (2002: 377,378, 2003: 398,399, 2009: 916,922), Mather (2004), Neumann-Holzschuh (2006: 265), Valdman (2006), and Winford (2008: 44).
3. To what extent do creole structures derive from the languages in contact? Is everything derived from those languages, or can linguistic features be present or absent regardless of what is offered by the input languages? This question is particularly relevant in view of the so-called “pool theory” (Mufwene, 2001, 2008b; Aboh and Ansaldo, 2007; Aboh, 2009).

1.1 Modeling Language Evolution

In order to answer these questions, we develop a simple mathematical model of the process of merging populations speaking different languages. Modeling language evolution requires certain theoretical assumptions about the underlying processes. These assumptions do not necessarily represent actual processes, but will be analyzed with the model to determine what their consequences are. This forces the researcher to make explicit assumptions which might otherwise have been considered too self-evident to even merit discussion. With real-world data, we can compare the consequences with what is observed, and thus establish which assumptions are necessary and which processes underlying the phenomenon under study are possible and impossible.

For example, when investigating whether creoles are the result of imperfect second language acquisition, a point of departure would be to determine whether such an assumption *needs* to be made in order for a language to evolve that has the structure of a creole language, or if a model with fewer assumptions could in fact lead to the same result.

Models can be minimal or complex, depending on their purposes and the material at hand. If the main purpose is to make predictions (as in weather

forecasts), and parameter values can be based on solid calibration through empirical data, complexity can be preferable. However, complexity comes at the cost of transparency and necessitates many assumptions. Unless the assumptions are driven by available data, transparency is preferred if the purpose is to find parsimonious explanations to fundamental processes and test basic hypotheses.

This study deals with creole genesis in the 18th century, a process with few historical records. A sound modeling process then starts with highly simplified models that might have only rudimentary resemblance to what is being modeled, so that the model is mathematically tractable. By mathematical analysis, we can determine the exact consequences of the assumptions, and then, with this knowledge, increase the complexity until the model manages to make the right predictions. Specifically, assumptions should either be required by the model or based on data.

To our knowledge, little has been done in terms of modeling the cultural evolution of language together with deriving analytical results and testing the model empirically. Surveys on what we have learned from modeling language so far have been made by Jaeger et al. (2009) and Castellano et al. (2009). The most-studied model is called *the Naming Game*. It was introduced by Steels (1996, 1997) (and often used in a simplified version, suggested by Baronchelli et al. 2006). In this model, agents develop their own vocabulary to map words to meanings. The agents then communicate in pairwise interactions, taking on the roles of speaker and hearer. The speaker randomly selects a topic and encodes it with the word that has been most successful in previous interactions of the speaker concerning the present topic. Should the speaker lack words for encoding the topic, she will invent one. If the hearer does not understand the word, then he might include it in his inventory for future reference. Simulations have shown that a globally shared vocabulary can emerge under such circumstances. It has been shown analytically for a similar model that the population converges to a common vocabulary (De Vylder and Tuyls, 2006). See further Loreto et al. (2010) for a survey on different varieties of the game.

There are a few simulation models that have specifically addressed creolization. Nakamura et al. (2007) have modified a more general model dealing with the evolution of Universal Grammar (Nowak et al., 2001) to investigate creolization. The model is not tested on empirical data and is based on transmission between generations, thus requiring demographic data over long time periods. There are models using real-world data (Satterfield, 2001, 2008), but these have a different approach – they are complex models including a large set of free variables that require strong and

unverified assumptions, which makes them opaque. Moreover, there is evidence suggesting that the data used refer to a setting which is not the one where the bulk of the creole structures were born⁵.

In this paper we present a model that is related to the Naming Game. The model allows for fast convergence and is mathematically tractable, so we can not only show that it is possible for a vocabulary to converge into a set of commonly shared words within reasonable time, but also derive the circumstances for *when* it does so and when it does *not*. We then test this model on empirical data, first to see that agents converge on the right vocabulary, and then to make verifiable predictions on phonology and syntax.

1.2 Structure of the Paper

The rest of the paper is structured as follows: In the Materials and Methods section 2, we present our basic assumptions for a model of Creole genesis and what kind of data is used to test it empirically. In the Model section 3, we describe the model in detail with possible extensions; then we present the results from a mathematical analysis of the model in the Analysis section 4, with the minimal conditions for a Creole to emerge. The full analysis can be found in Appendix A.1. In 5, we give a background to Mauritian Creole and the data used in the case study. For the interested reader, we give a thorough discussion in Appendix A.2, and raw data in Appendix A.3. In the Empirical Results section 6, we put the model to work on that data to generate a prediction on what Mauritian Creole *should* look like, and then compare the results to the actual language and also to other languages. Finally, in the Conclusions section 7, we discuss regarding the research questions presented above.

2 Materials and Method

In this paper, we develop a model and determine analytically what assumptions are required for a common language to emerge. The validity of the model is tested by using empirical data of an instance of creole genesis. We run simulations incorporating data to predict what the language would look like based on certain assumptions, and compare this prediction

⁵ For instance, it is assumed that Sranan emerged in Surinam, where it is currently spoken. However, the topic is controversial, and there is good reason to assume that the foundations of Sranan were laid elsewhere (perhaps most likely on the Lesser Antilles) and that the language was imported to Surinam (Baker, 1999; Baker and Huber, 2001; McWhorter, 1995; Parkvall, 1999).

to the actual language in question. Vocabulary, phonology, syntax and—to a lesser extent—morphology will be considered.

The code was implemented in Java and can be obtained from the corresponding author upon request.

2.1 Basic Modeling Assumptions

Regarding the structure of creole languages, two claims are prevalent: first, creoles typically have a lexifier – a language that provides the basis for the majority of the vocabulary – while much of its morphosyntax is not obviously derived from this source, and often appears to originate elsewhere (be it from other languages or from the workings of the human mind). Second, creoles are typically highly analytic, that is, they tend to have rather limited amounts of morphology. These claims are both largely uncontroversial (though the works by some creolists, most notably Chaudenson and Mufwene, of which several are included in the reference list, tend to emphasize the lexifier contribution beyond the lexicon). The aim of the model is thus to correctly predict vocabulary, phonology, and syntax.

It is necessary for the output of the model to display the following two properties:

1. The vocabulary converges to having a single lexifier.
2. All agents use the same phonology and syntax, but in some cases, this is not the phonology and syntax of the lexifier.

It is a reasonable assumption that all languages involved generally have different words for the same meaning (cf. the Saussurean concept of the “arbitrariness of the linguistic sign”). There is an infinite number of possible lexical representations of a semantic reference, while for phonology and syntax, the number of possible features is limited. For example, there is an anatomically defined limit to the number of possible phones (and thereby phonemic contrasts), and a given morphosyntactic device is typically either present or absent. Thus, for phonology and syntax, languages will form groups that may be different for each feature. This is what makes it possible for both properties above to be satisfied in our model.

2.2 Empirical Data

For the empirical test, we use demographic data from Mauritius in the 18th century, when the French colonized the island and imported slaves (and, to a lesser extent, free workers) of various origins. These immigrants are likely to have spoken at least six different languages: Malagasy, Tamil,

Bengali, Gbe, Wolof, and Manding⁶. An important aspect of this particular creologenic setting is that we can assume the creole to have emerged locally, rather than having been imported from elsewhere (as is the case for many other locations). Another is that we have exceptionally detailed demographic data for the early years of colonization (Baker and Corne, 1982). Other settings fulfilling the first assumption usually have only sketchy demographic data.

Nearly all the vocabulary of Mauritian has a French origin, while the phonology and the grammar differ significantly from those of the lexifier. In our analysis (Section 6), we present our computer simulations of the model and verify that the model correctly produces a French vocabulary using the demographic data. We also present simulations with data on phonology and syntax for the languages represented on the island as input, and compare the results to modern Mauritian.

3 The Model

Our model is similar to existing models such as the Naming Game, but its specifics are not based on or borrowed from any previous work. In our model, individuals, or agents, meet in pairwise interactions and try to communicate. The interaction results in both agents being slightly more likely to use the linguistic item that was used in this interaction in the future, but just how likely depends on the outcome of the interaction.

In more detail, every agent has a certain probability to speak each of the languages represented in the population. When agents enter a population, they only know their native language and thus have a 100% probability of using that language in their first interaction⁷.

In each round in our simulation, every agent has an interaction with another randomly selected agent from the population. The interaction can

⁶ More languages were present, but the historical records suggest that the ones listed here would have been the major ones. Somewhat later on, speakers of Mozambican Bantu languages (such as Makhuwa) became very numerous, but this was after the period under consideration here. Gbe is a cluster consisting of several rather closely related languages, chief among which are Fon and Éwé. For most of the features used here, these two share the same feature values; whenever we did not have the values for both, we used the value of the one language that we did know.

⁷ This is an obvious simplification, of course, since a large part of the world's population has always been bi- or multilingual. The simplification is warranted, however, by the extreme geographical diversity of the population. Whatever second languages they may have commanded would be of little use, as they would in most cases not have been understood by people of other origin. Some of the free Indians constitute a possible exception, in that they may have used an Indo-Portuguese creole language with the French in India, a language not unlikely to have been familiar also to some Frenchmen in Mauritius.

be thought of as an exchange of a word, a sound or a syntactic feature. Assume that there are native speakers of three languages present in a population: French, Wolof, and Malagasy. Should a native French speaker have previously communicated about *bread* with native Wolof speakers, then there is a chance he will use the word *mburu* next time instead of *pain*. His interlocutor, on the other hand, might choose between the words *pain* and Malagasy *mofo*, depending on previous interactions. The word they use will affect their counterpart, so that the latter will have a slightly higher propensity to use that word the next time he wants to talk about bread. Note that the model does not account for any differences in sociolinguistic status of the languages, but assumes a mutual interest in communicating using whatever material is available.

After an interaction, the probability of using the word (or linguistic feature) an agent hears increases proportionate to its complementary probability (that is, $\Delta p \propto 1 - p$). Thus, an agent who has previously used the word only rarely will be more affected than someone who already uses it almost exclusively. Before being added, the inverse probability is multiplied by a constant ‘learning’⁸ parameter, ℓ , $0 \leq \ell \leq 1$, representing how much the agent is affected by hearing a word. In order for the probabilities to sum to one, the probabilities for the other languages are subtracted proportionate to their present values. Thus, if there are n words representing the same meaning in the population, then an agent who uses the word w_i with probability p_i , for all $i \in \{1, 2, \dots, n\}$, will change his probabilities in the following way after hearing the word w_j for some $j \in \{1, 2, \dots, n\}$:

$$\begin{cases} p'_j = p_j + (1 - p_j)\ell \\ p'_i = p_i - p_i\ell, \quad i \neq j \end{cases}.$$

An alternative model is one where the probability of using a word increases by a constant. However, such a model needs to treat the limits (0% and 100%) as special cases and define by what proportions other words will decrease in use. Also, our proportionate model reflects the fact that, when little of the information you receive was previously known to you, then there is more information to extract from the message.

3.1 Extension 1: Conservatism

As we will see in the Analysis section, more assumptions will be needed to satisfy both conditions in Section 2.1.

⁸We use ‘learning’ as a collective term for all the processes that are involved in updating the usage of a linguistic item, including social factors.

One reasonable assumption of this kind would be that, after having been in the population for a while, people become more reluctant to change, while newcomers will adapt to their new situation and be keener on learning the language they encounter. We model this by multiplying ℓ by a discount factor δ^k , where $0 \leq \delta \leq 1$, and k is the number of rounds since the individual entered the population. (For positive $\delta < 1$, this discount factor decreases as k increases.)

Note that this extension of the model makes it more general, since $\delta = 1$ gives us the first model.

3.2 Extension 2: Coordination

People may react differently depending on whether or not communication was successful. If both individuals use the word *pain*, then the learning constant, ℓ , is likely to differ from when one of the individuals uses *pain*, and the other *mofō*. In the first case, communication was successful, while in the second, it was not. We will model this difference by splitting ℓ into two constants: ℓ_c , which is used when the couple are coordinated (that is, use the same form), and ℓ_u , which is used when they are uncoordinated.

An alternative formulation, with one speaker and one hearer, is that p_i represents the probability that hearer i will interpret the word correctly, or that it aligns with his current hypothesis about the linguistic item in question. (Thus, all the words that the agent has heard recently have nonzero probabilities of being interpreted correctly.) The inequalities $\ell_c > \ell_u$ and $\ell_c < \ell_u$ will then introduce a confirmation and an anti-confirmation bias, respectively.

This is also a generalization, since $\ell_c = \ell_u$ gives us the first model.

3.3 The Extended Model

Combining the two generalizations, with n languages in the population, an individual who has lived on the territory for k rounds and speaks language i with probability p_i will change his probabilities in the following way, after hearing language j :

$$\begin{cases} p'_j = p_j + (1 - p_j)\ell_{\bullet}\delta^k \\ p'_i = p_i - p_i\ell_{\bullet}\delta^k, \quad i \neq j \end{cases},$$

where $\bullet = c$ if the individual spoke language j , and $\bullet = u$ otherwise.

4 Analysis

The full analysis is presented in Appendix A.1. In the present section we will summarize the results. In the first model (without Extensions 1 and 2), the expected change for any language after a round of interactions is 0 (see Appendix A.1.1). Thus, the analysis tells us that we have a neutral process, without any biases toward selected agents and linguistic items, where changes in language frequencies are due only to immigration and random drift. In such a process, the language most likely to be the dominant contributor, if any (fixation is not guaranteed in a bounded time period), is the one with the largest number of native speakers.

Given the specific demography of our case (and indeed most other creologenic settings), we can confidently say that this model alone cannot explain the resulting Mauritian vocabulary, since, except for a short introductory period, French is a minority language with respect to native speakers. Thus, the model does not only fail to meet the first condition in Section 2.1 (vocabulary converges to a single lexifier), but specifically, it predicts a language with a minority of French words.

4.1 Extension 1: Conservatism

Our analysis (see Appendix A.1.2) shows that, by adding a discount factor δ to the first model, the process is no longer completely neutral. By adjusting the value of the parameter, the model can produce a vocabulary with a majority of French words (although less than 100%).

However, unless the discount function δ takes on different values for different linguistic items, the analysis also shows that whatever language becomes the lexifier, that language will also provide the full phonology and syntax, and thus violate condition 2 in Section 2.1 (some aspects of phonology and syntax come from other sources than the lexifier language).

Thus, in order to get a French vocabulary together with a phonology and syntax that often differs from French, it is necessary to use different values for the discount function δ for each trait. This may well mirror reality – different parts of language may evolve (and be acquired) at different pace – but such a model is untestable, since we do not know the values of δ and would require data from a large number of populations to estimate it. Therefore, the first extension will not be useful for our simulations.

4.2 Extension 2: Coordination

Our analysis (see Appendix A.1.3) shows that different values for ℓ_c and ℓ_u (individual updating when using the same or different words, respectively)

generate three different results depending on the relation between the two variables:

From the first model, we know that $\ell_c = \ell_u$ gives a neutral process, with random drift.

With $\ell_c < \ell_u$, we get a leveling process toward all languages having the same frequencies. All agents converge to using all the input languages equally often, thus having several synonyms and, in general, no rules for syntax.

Finally, $\ell_c > \ell_u$ gives a positive value of changed frequency for the language that already has the highest frequency in the population. Thus, well-established languages grow, while small ones decline.

For one language to dominate the vocabulary, then, it is necessary that $\ell_c > \ell_u$, in which case one language will always reach 100% after sufficiently long time. If the learning process is fast enough (large ℓ_c or many rounds), the dominating language will never change as long as there are not more immigrants at one single occasion than people already living in the territory.

With a population that starts off with a small number of speakers of one language, and where the population never doubles at one single occasion (both requirements being fulfilled in Mauritius), our model can produce a vocabulary that is dominated by that language, while potentially allowing for other coalitions of languages to dominate syntactic and phonological features. Thus, we will concentrate on this model for our simulations.

It can be noted that in a similar model, De Vylder and Tuyls (2006) show that fast convergence can occur when frequently used language items are amplified, such that they increase in usage disproportionately more than other language items. The model presented here suggests a process for how amplification may occur.

5 Case Study: Mauritian Creole

In order to test our model empirically, and to shed some light on the issues stated in the introduction, we devised simulations of the early years of settlement in Mauritius, a country where a creole language is spoken.

Appendix A.2 gives a detailed background on the conditions on the island during the period when the creole emerged and on the data used. For those unfamiliar with creoles, we also give examples of what Mauritian looks like in this Appendix, and then proceed with an account for the demographics that underpin the data. Even though the Appendix lays out more thorough arguments for the conclusions below, it is not fundamental for a coherent understanding of the paper, and we have not included it

here due to space constraints.

In the following, we provide a brief background on the demographic and linguistic data used.

5.1 Background

The island of Mauritius in the Indian Ocean received its first permanent human population in 1721, when it was annexed by France. It was turned into a plantation colony and followed the development of many islands in the Caribbean, where slaves from several different locations were brought in to toil for their European masters. One of many interesting aspects of this historical development was that it led to the emergence of an entirely new language, Mauritian Creole or *Morisyen* (Baker and Corne, 1982).

Mauritian is a French-lexicon creole language. This implies that its vocabulary is almost entirely derived from French, while its grammatical structures diverge radically from any dialect of the lexifier. Some of the more notable features which set it apart from French include the lack of grammatical gender, a lack of case distinctions in most pronouns, and tense/mood/aspect marking by means of free preverbal particles rather than suffixing. Since the precise nature of the differences between creoles and their lexifiers is vividly debated,⁹ the reader unfamiliar with Mauritian may refer to Appendix A.2.1 for a text sample.

5.2 The Data

5.2.1 Demographics

The initial peopling of Mauritius is documented in considerable detail in Baker and Corne (1982), and for most of those who arrived, a reasonable guess can be made with regard to their native language. The most relevant languages, and those used in the simulation, are French, Malagasy, Manding languages, Gbe languages, Tamil, and Bengali. A discussion of the proportions, together with a map illustrating the geographical origins, is found in Appendix A.2.2.

We only have detailed demographic data for the first fifteen years of settlement, and an obvious question is whether that is enough. One reason to assume that it is, as we will see, is the remarkable match between our simulation and the actual language. For independent indications that the

⁹While the grammatical description may cause moderate amounts of disagreement, the exact character of the lexifiers (which were, of course, not identical to modern-day standard varieties) makes assessments of the actual differences subject to discussion.

Figure 1: Demographic evolution (by language group) of Mauritius during its first 5,118 days of settlement. Please note that the data includes not only arrivals, but also departures.

language had roughly taken on its present form by the mid-18th century, see Appendix A.2.3.

The demography is presented in Fig. 1, with the number of native speakers of each language plotted against the number of days after the first settlement. The population of the island grew rapidly, and after only nine years, the French settlers – who were the first on the scene – were outnumbered by slaves and voluntary immigrants from outside Europe.

One thing these data do not take into account (at least not in a consistent fashion) is births and deaths. The Baker and Corne figures yield a cumulative import of 1,490 slaves in 1735, whereas Vaughan quotes the figure of 648 (2005:47), which seems to suggest a high mortality. We investigated the possible effects by computing the additional death rate that Vaughan’s figures imply (0.0437% per day compared to the Baker and Corne figures) and applying this to the agents in the simulation, which produced essentially the same results. (The main difference was that there were more random fluctuations between simulations.) As for births, birth rates are typically low in slave societies (Debien, 1974; Williams, 1991), and the input that the children received during their early upbringing would in any case have been determined by the languages spoken by the adults present. Whatever other impact the few children may have had, they are therefore unlikely to have tipped the scales as far as the relative contributions from the input languages are concerned.

5.2.2 Linguistics

The basis for the linguistic data is the *World Atlas of Language Structures* (WALS) (Haspelmath et al., 2005), combined with the structural features of Grant and Baker (2007: 24–27) and the *UCLA Phonological Segment Inventory Database* (UPSID). These sources all deal with linguistic features in terms of whether a given trait is present in a given language.

Many of the languages we are interested in (including Mauritian itself) are covered by these sources only in a patchy manner, so we consulted a variety of reference grammars to fill in the blanks. A certain degree of convenience must be admitted here, as some features are considerably more difficult to extract from a reference grammar than others. For instance, the phoneme inventory of a given language is explicitly tabulated in most language descriptions, while figuring out the behavior of subordinate clauses under specific circumstances requires a great deal more work. In essence, the linguistic data collection was simply halted when we felt that the cost began to outweigh the benefits.

In the end, we used only those features which were available for all languages involved, and after deleting doublets, the linguistic database finally consisted of 128 features (of which 89 pertain to phonology, 30 to syntax, and 9 to morphology). All features with their respective values can be found in Appendix A.3.

6 Empirical Results

We now present the results from simulation runs of the model, followed by a discussion of which traits cannot be predicted by the model, and why. Finally, we compare Mauritian to all of the input languages, and the languages of the world in general.

6.1 The Simulation

Agents represent actual individuals and are added (or subtracted, as the case may be) in the chronological order following the demographic data, starting with 17 agents having a 100% probability of speaking French (the first Frenchmen, who arrived on the island on Christmas Eve of 1721). In each round, every agent selects another agent at random to interact with. In the standard setting, there is one round of interactions for each day. This number is arbitrarily chosen, and it is likely that in the real world different language items are communicated with different frequencies. However, apart from discretization effects, modifying the number of interactions amounts to modifying the learning rate; and this can also be

achieved by adjusting the variables ℓ_c (learning rate when we agree on a linguistic item) and ℓ_u (learning rate when we use different items), so the analysis is restricted to varying these variables. According to our analytical results, for French to become the lexifier, we have to assume that $\ell_c > \ell_u$.

An interaction amounts to each agent transmitting a linguistic item (a word, a sound or a morphosyntactic feature), chosen randomly according to their language preferences/probability distribution, to the other agent. Both agents will then update their preferences according to the model equation. The simulation stops after 5,218 days, 100 days after the last recorded demographic data.

6.1.1 Vocabulary

We know from the analytical results that $\ell_c > \ell_u$ is a necessary assumption for the French vocabulary to gain ground throughout the population. Whether it is also a *sufficient* assumption depends on the demography. If, at any time, the population doubles from a non-French monolingual population moving in, then the language of that population will become dominant. In our data, there is no such event, but there is a vast immigration of people speaking Tamil, Wolof, and Malagasy (in this order) in a short time. The question is thus: what is a sufficiently high learning rate for these to be assimilated before they outnumber those who have adopted a French vocabulary?

The results are presented in Figs 2 and 3, and, indeed, the outcome is robust. Fig 2 presents the results for three different size ranges of ℓ_c and ℓ_u , with the amount of black representing the amount of French. The figures look very similar. Irrespective of the value of ℓ_c , the simulations result in a close to 100% French vocabulary with the condition that $\ell_c > \ell_u$ (the part above the diagonal is black), except for the special case of $\ell_u = 0$, for which no updating takes place (since agents start out with a single language and do not update when someone uses a word from another language). Thus, the emerging language will have a French vocabulary if and only if agents increase their usage of a certain word more when the counterpart agrees on using that word.

Fig 3 shows the evolution of the vocabulary. There are some steep increases of the proportion of non-French lexical items, due to immigration, but the increase is never large enough to take over the vocabulary entirely.

Figure 2: Proportion of the vocabulary that has a French origin, given update rates from 0 up to 1 (100%), 0.1 (10%), and 0.01 (1%), respectively.

6.1.2 Phonology and Syntax

We ran a simulation for each of the 119 phonological and syntactic traits (that is, morphology is excluded for the time being, but is discussed later) for which we have complete data. The difference from the vocabulary simulations is that languages cluster, so for a binary trait, for instance, our population consists of two groups only, rather than as many as there are languages (that is, seven). There are unlimited possibilities for lexical representations, but not for structure. A linguistic label for ‘apple’ might be *apple*, *pomme*, *Apfel*, *yabloko*, *manzana*, or any number of other possibilities. A structural feature such as the order of noun and adjective, on the other hand, yields but three possibilities: ADJECTIVE+NOUN, NOUN+ADJECTIVE, or both. This implies that, when a number of unrelated languages are in contact, chances that two of them are going to share a lexical item are slim; for a structural feature with three parameter settings, however, some value is bound to be shared within any group of more than three languages. The languages in contact can therefore reinforce one another within syntax, but not within the lexicon¹⁰.

For each feature, we compared the “winning” trait with that actually attested in (present-day) Mauritian Creole. Summing up the total number of correct predictions, we get a total similarity ratio for all of the traits. French is the language among the seven that has the highest similarity to Mauritian, with a ratio of 84%, so we use this number as a benchmark for the predictive value of our model. As established previously, we look at parametric values where $1 > \ell_c > \ell_u > 0$.

When $\ell_c \gg \ell_u \gg 0$ (ℓ_u is considerably smaller than ℓ_c , but considerably larger than zero), there is a bias towards the language of early arrivals, in our case French. With $\ell_c \gg \ell_u$, all agents are reluctant to change other

¹⁰ Within lexical semantics, the possibility exists, but only exceptionally when it comes to the assignment of phonetic strings to various concepts.

Figure 3: Etymological composition of the vocabulary of the emerging lingua franca plotted against number of days since first settlement. Steep increases of the proportion of non-French lexical items are due to immigration. Here, $\ell_c = 0.01$ and $\ell_u = 0.005$.

than increasing the use of their majority language. With $\ell_u \gg 0$, agents may still learn at a sufficiently fast pace to assimilate their language to what the majority of the population is speaking before new immigrants arrive. Indeed, these simulations not only produce a French-dominated vocabulary, but also French syntax and phonology for these values. We will therefore focus on values where $\ell_u \approx \ell_c$ or $\ell_u \approx 0$ (ℓ_u is close to ℓ_c or zero), but which still produce a French vocabulary.

The results for $\ell_u \approx \ell_c$ are given in Fig. 4, with relatively large (0.1–0.5) and small (0.01–0.03) values of ℓ_c . Each value is represented by a curve, for which the similarity to Mauritian is plotted against values of the difference $\ell_c - \ell_u$, from 0.003 (the limit where the majority of simulations still generate a French vocabulary) to 0.009. Simulations that do not produce a French vocabulary are excluded. The lines have similar trends, except for $\ell_c = 0.01$, for which $\ell_u \approx 0$ in the right end of the figure, which brings it to the case discussed below and explains the rise at the end. For small values of $\ell_c - \ell_u$, the similarity to Mauritian Creole attains at most 92%, while it converges to the benchmark value as the difference between the learning parameters increases. Where it attains the benchmark value, the model predicts a French syntax and phonology.

The results for $\ell_u \approx 0$ are given in Fig. 5. Two curves are given, with $\ell_u = 0.0001$ and 0.001, respectively. (For smaller values, the vocabulary

Figure 4: Mean values for the similarity to Mauritian Creole generated by simulations with respect to the difference $\ell_c - \ell_u$ for different values of ℓ_c . The simulation was run 25 times for each parameter setting, and those that did not generate a French vocabulary (a minority of cases: none for $\ell_c \leq 0.1$, but increasing for larger values of ℓ_c and smaller differences) were excluded. Standard errors are within 0.01.

will not be French.¹¹) For the former, the simulations give a similarity to Mauritian between 90% and 92% for all values of ℓ_c .

To conclude, the simulations show that the model is always at least as good a predictor of Mauritian Creole as is French.¹² For small nonzero ℓ_u (Fig 5), or if both ℓ_c and $\ell_c - \ell_u$ are small and nonzero (Fig 4), the similarity to Mauritian reaches 92%. In both these cases, conservatism is relatively high, and the alliances that form between languages for particular structural features may have the time to accumulate and be large enough to outcompete the one that French represents. In particular, if ℓ_u is small but nonzero, then the model consistently produces highly accurate predictions, for any value of ℓ_c . These are the cases where successful communication leads to learning (since ℓ_c is not necessarily small), while only little learning ensues if I do not understand what you are saying (but still

¹¹In these cases, the emerging language will not converge to a single lexifier. Which language will have the largest share depends on the parameters, and this language will in general not be in the majority.

¹²We have here excluded extreme values of ℓ_c close to 1, in which case this statement does not apply.

Figure 5: Mean values for the similarity to Mauritian Creole generated by simulations with respect to ℓ_c for two different values of ℓ_u . The simulation was run 25 times for each parameter. Standard errors are within 0.005.

more than nothing, since ℓ_u is small but nonzero).

6.1.3 An Alternative Scenario

Did the slaves interact with their enslavers? And if so, did the enslavers use their own language or the (emerging) pidgin? Available documentation from other situations suggests the latter. For instance, communication between Romance-speaking slaves and Arabic-speaking enslavers in North Africa is well documented to have taken place in a pidgin lexically based not on Arabic, but on Romance (Schuchardt, 1980: 77). On Fiji, the Indians who were slaves in all but name, were addressed not in English, the prestige language, but in a pidginized version of Hindi (Burton 1910: 288f; Siegel 1990: 185f). Similarly, in Surinam, the ruling Dutch learned the slave language Sranan in order to communicate with their slaves (Holm, 1989: 435).

In any case, our simulations show that the assumption of symmetric updates for all agents can be relaxed. Modeling the extreme case where the French do not acquire any item from any other language still produces a language that is 89%–90% similar to Mauritian for a vast array of parameter values; see Fig. 6.¹³

¹³A slight difference is that, with this assumption, the vocabulary can be French also when

Figure 6: Mean values for the similarity to Mauritian Creole generated by simulations when French enslavers do not acquire a new language. The solid line represents $\ell_c = 0.1$ and is plotted with respect to different values of ℓ_u (bottom axis). The dashed line represents $\ell_u = 0.001$ and is plotted with respect to different values of ℓ_c (top axis).

6.2 Failed Predictions

The best prediction our model can make for Mauritian Creole bears 92% similarity to the language, excluding morphology, and 88% if we do include the nine morphological traits for which we have full data. For fifteen of the features (morphology now included), our model fails to predict the attested outcome¹⁴. As it happens, nine of these can be explained by Mauritian Creole having had a pidgin past. The most well-known and least controversial feature of pidgins is their complete or near-complete lack of morphology and dearth of grammatical elements in general.¹⁵ Because of this, we would not expect proto-Mauritian (if this was indeed a pidgin) to have had, for example, grammatical gender or a morphological imperative,

$\ell_c = \ell_u$, which is why we included these cases in the figure. For these extreme values, the predictive value of the model drops.

¹⁴ These fifteen features are: definiteness marking mainly postposed, Nominal and Locational Predication, Number of Genders, Predicative adjectives=nonverbal, Predicative adjectives=verbal, Preverbal negation, Suffixed nominal plural, Sufficing inflexional morphology, The Morphological Imperative, /ɔ/, /ɛ/, /ʏ/, /h/, /ɲ/, and /ʃ/.

¹⁵ Bakker, 2003 and Roberts and Bresnan, 2008 are often cited in order to illustrate that pidgins need not be perfectly analytic. This is true, and questioned by no one. Still, even those varieties discussed there (and we would not consider them all pidgins) have very small affix inventories in comparison to most of the world's languages.

regardless of the preferences among the input languages. Such features are simply not carried over into pidgins, irrespective of the languages involved (there are exceptions, of course, but this is the normal case)^{16,17}. Some morphological categories are indeed present in modern Mauritian, but represent later developments – that is, they emerged long after the period that concerns us here (Baker, 2009; Guillemin, 2009), and thus outside the scope of our model. At that time, Mauritian had native speakers, and the later developments may thus be unrelated to the languages present in the initial contact situation (cf. Roberts 2004 for evidence of the important role native speakers play in such a situation).

The remaining six features all relate to the presence or absence of specific phonemes. These can be arranged into the following groups:

- Presence incorrectly predicted by our model: /ʃ/, /ɲ/, /ɛ/, and /ɔ/.
- Absence incorrectly predicted by our model: /h/ and /ɣ/.

The members of the first group are all found in French, but did not make their way into Mauritian: the mid vowels /ɛ/ and /ɔ/, the fricative /ʃ/, and the nasal /ɲ/. Here, Mauritian aligns itself perfectly with Malagasy, but not with any other relevant language. In fact, the set of phonemes *common to French and Malagasy* (a French phoneme inventory filtered through the phonology of Malagasy, as it were) is considerably more similar to Mauritian than is any individual language.

Malagasies made up the majority of the non-white population for the first seven years of settlement, and one possible implication of this is that the phoneme inventory of Mauritian was fixed before other aspects of the language were (cf. the suggestion of Parkvall 2000: 156–157, that the phonology of a creole crystallizes before its syntax does).

The second group of incorrectly predicted phonemes consists of /h/ and /ɣ/, the latter being the Mauritian counterpart of French /ʁ/. It should be remembered, however, that the two are articulatorily and acoustically close, and /ʁ/ has a wealth of allophones within and outside France. In western France, from where most of the settlers destined for Mauritius

¹⁶ Similarly, the feature “verbal encoding of predicative adjectives” follows from the lack of morphology in combination with the absence of an equative copula (again common in pidgins). This automatically yields a situation in which adjectives (*You Ø sick*) are indistinguishable (or virtually so) from verbs (*You run*). We have also included the use of a simple preverbal negator in this category – while not a byproduct of analyticity, it does represent a strategy typical of pidgins.

¹⁷ Several of these features are of course commonly attested in second language acquisition (e.g. Klein et al., 1993), and, taken individually, they are in principle compatible with an SLA-based scenario

originated, [x] is a frequently encountered realization (Walter, 1982), and [ɣ] (voiced velar) might be thought of as a compromise between [ʁ] (voiced uvular) and [x] (unvoiced velar).

As for /h/, it is a truly marginal phoneme in Mauritian, occurring only in very few words of French origin (*dohor* 'outside', ← *dehors*, *hale* 'to pull', ← *haler*). Either it could be considered absent from Mauritian, or it could be considered present in French, since the words just mentioned clearly indicate – despite the state of modern Standard French – that it did exist in the 18th-century French dialects which are relevant in this context. Regardless of which option one chooses, Mauritian has followed French on this point, and if one considers /h/ to be absent, then it also follows Malagasy.

Thus, the features which our model failed to predict correctly can be explained if we assume that: a) Mauritian started its life as a pidgin (entirely or virtually) devoid of morphology, and b) the phoneme inventory crystallized at a very early date, when speakers of Malagasy still dominated the non-white population of the island. However, after a period of demographic decline, Malagasies again dominated the slave population between about 1740 and 1765 (Grant and Baker, 2007: 202), which opens up the possibility of the phonology gelling at this point in time. The failed predictions are summarized and categorized in Table 1.

In sum, most of the failed predictions are failures only in a scenario where Mauritian was not born from a pidgin, and where it did not develop after its crystallization. The latter is obviously false, since no language escapes change over time. We also believe that Mauritian did indeed go through a pidgin stage, not least because this would make almost all the pieces of the puzzle fall into place. The only features remaining are four phonemes, whose absence could indicate that the phoneme inventory stabilized in a period of Malagasy numerical dominance.

6.3 Mauritian Compared to Other Languages

The similarities between Mauritian, all of the input languages and the best predictions of our simulations (with the least favorable outcome being when the simulations predicted all traits to originate from French) are presented in Table 2.

It is notable that French is the language closest to Mauritian not only with a close to 100% resemblance in the vocabulary (due to the early arrival of the French), but also with an 87% resemblance for syntax, while the phonology is closest to Malagasy. In all cases, our model is as good or a better predictor of Mauritian than any of the input languages.

One of our research questions is to what extent creole structures derive

Feature	Explanation
/y/, /h/	Not really wrong
/ɔ/, /ɛ/, /ɲ/, /ʃ/	Adaptation to Malagasy phonology
Number of genders, Suffixed nominal plural, Suffixing inflexional morphology, Morphological imperative	Morphology, and loss is therefore expected due to pidginization
Nominal and locational predication, Nonverbal adjectives, Preverbal negation	Other expected consequences of pidginization
Postposed definiteness article	Later development (c. 1820, according to Guillemin 2009).

Table 1: Failed predictions and explanations.

from the languages in contact. It could be that some features could be ascribed to universals rather than to any of the input languages, and if so, a comparison with the languages in the world as a whole could shed some light in the issue. We therefore computed the similarities for all languages for which there is a sufficient amount of data in the WALS database, from which we collected 28 of the 119 features analyzed here. For these features, the simulation consistently predicts French to align with Mauritian, resulting in 93% accuracy. All languages with data for more than half of the traits are less similar. Looking at languages where *all* features have been classified (there are only four other such languages), we get the list presented in Table 3.

If we allow for 20% of the data to be missing, then we have 68 languages to compare to. Ranking languages from highest similarity (lowest rank, 1) to lowest (highest rank, 68), the 7 input languages have a mean rank of 22.86, while the 61 remaining have a mean rank of 35.84. The difference between these two samples is statistically significant (Mann-Whitney $U = 132$, $p < .05$, one-tailed).

If instead we compare featurewise, by computing how many languages in the world share the setting of Mauritian for each feature, then we get an average of 51% of languages per feature. On average, 16% of languages

All	119	Phonology	89	Syntax	30
<i>Simulation</i>	92%	<i>Simulation</i>	93%	<i>Simulation</i>	87%
French	84%	Malagasy	87%	French	87%
Malagasy	82%	French	83%	Wolof	80%
Wolof	77%	Bengali	81%	Malagasy	70%
Bengali	72%	Manding	81%	Gbe	67%
Gbe	71%	Wolof	76%	Bengali	47%
Manding	71%	Tamil	73%	Manding	43%
Tamil	63%	Gbe	73%	Tamil	33%

Table 2: Similarities to modern Mauritian Creole.

Input	Sim.	Other	Sim.
French	93%	English	75%
Wolof	82%	Spanish	75%
Malagasy	71%	Slave	50%
Gbe	57%	Japanese	43%
Bengali	54%		
Manding	46%		
Tamil	39%		

Table 3: Similarities to modern Mauritian Creole with respect to the 28 WALS features.

align with regard to the least common setting for each feature, and 61% with regard to the most common setting for each feature. The range of possible averages is thus 16%–61%. Stretching this interval to 0%–100%, the average for Mauritian would be 77%.

Should we construct a language that always uses the most common setting of all the languages in the world, we would get a language that shares 55% of its syntax with Mauritian.

7 Summary and Discussion

We have developed a model to predict the outcome of a situation when speakers of different languages meet and try to communicate. The model assumes that people increase their use of a particular language when they encounter that language in an interaction. We have shown that, for all speakers to converge on speaking the same language, they need to have a bias towards successful communication, a confirmation bias: that is, they

must increase the use of the language they heard in the interaction more often when they used the same language in that interaction – or, alternatively, interpreted the word correctly – than when they did not. Thus, we have excluded the opposite model, where people learn more from failed interactions or do not make a difference between the two types of interactions, from the set of possible explanations how languages emerge in the present context.

We then used the model to derive predictions on the evolution of Mauritian Creole that could be compared to the known outcome, thus providing an empirical test. As a benchmark value, we would predict all features to originate from French, the most similar language (reaching 84%). The model predicted a language at least as similar to Mauritian as is French, with up to 92% accuracy for syntax and phonology when people have a very strong confirmation bias, where learning takes place in successful interactions, while unsuccessful ones have very little (but nonzero) impact on the language. We now have material to address four issues regarding creole genesis:

Are creoles the result of failed second language acquisition? We cannot claim to have proven that targeted language learning¹⁸ did not play a role in the genesis of Mauritian, but our model generates a language highly similar to modern Mauritian without assuming the lexifier as a target language, and would not have produced better results with the addition of a motivational component. Such a component would have made the individuals in our simulation more prone to accept input from native speakers of French than from others, and it would, in fact, have rendered the results less accurate, insofar as more structural (as opposed to lexical) features from French would have been included in the creole – features which *de facto* are not there. Except for a few cases where the learning rate is set at exceptionally high levels, our model produces its worst results when the predicted language is identical to French, so any increased similarity to French would imply a result more divergent from the creole as it is spoken today.

How fast do creoles develop? Mauritian may have formed rather rapidly. The model must be set to work at a sufficiently high pace in order to match the demographic developments, and this applies to any setting where the speakers of the lexifier language were once in the majority, but other larger

¹⁸ In the sense that non-Europeans tried to acquire French, as opposed to an inclination to simply communicate by means of whatever linguistic material might be understood by the interlocutor.

populations arrived within short time spans. The historical data cited in section A.2.3 suggests that the language had largely crystallized within half a century, but it would seem from our simulations that the first fifteen years could well suffice (this being the only period that we considered, while yet making predictions with high accuracy). It could of course be argued that the process was more protracted if one assumes a less direct relation between demographics and creolization, but the very fact that a simulation based entirely on demographic evidence yields results of the kind we obtained (a near-perfect match with modern Mauritian, if one takes prior pidginization into account) casts doubt on such an assumption.

Do creoles develop from pidgin languages? In our simulations, a language quickly evolved that, with respect to vocabulary, syntax, and phonemes, highly resembles modern Mauritian Creole. The main linguistic domain where our model failed to make correct predictions is that of morphology, of which Mauritian has very little.

While this is not central to our investigation, the near-total loss of French morphology¹⁹ is hardly compatible with anything other than prior pidginization. The fact that morphology in the contact situation behaves differently from syntax (almost complete loss in the former case versus a compromise between what was offered by the input languages in the latter) is precisely what is observed in stages of observed pidginization followed by nativization (such as some English-lexifier varieties in the Pacific), but starkly different from what occurs in other kinds of language contact.

To what extent do creole structures derive from the languages in contact? The “pool theory”, according to which creole structures are derived from the input languages, and only from these (Mufwene, 2001, 2008b; Aboh and Ansaldi, 2007; Aboh, 2009), cannot be upheld. On the contrary, it is clear that creoles contain features not found in these languages, but more importantly, that they often lack traits that were indeed shared by the creole creators (Plag, 2011; McWhorter, 2012; Parkvall and Goyette, forthcoming). However, while much of the counterevidence is related to morphology (the absence of which is, we believe, related to

¹⁹ Note that this does not imply that current Mauritian is *completely* analytical. However, much of the morphology that does exist is not inherited from French. Reduplication, for instance, is not a feature of French in the first place. The diminutive prefix *ti-* is derived from French *petit* ‘small’ but is not an affix in the lexifier, and thus, grammaticalization seems to have taken place within Mauritian itself. The distinction between long and short verb forms bears some (albeit limited) similarity to French verbal morphology (and may well have been influenced by it), but in all likelihood represents a post-formative development (e.g. Corne 1999:167).

pidginization), credit must be given where credit is due – the *syntactic* features included in our simulation *are* in fact predicted by our simple model. We have also seen that the input languages do resemble Mauritian more than languages do in general, which strengthens the hypothesis that the input determines the syntax of a contact language. It could thus be that a “pool” approach does work as a predictor of creolization, if (but only if!) one takes into account the limited amount of structure that is to be expected within a pidgin in the first place.

The similarities between the input components and the output would no doubt be expected by most people. However, those doubting the pidgin past of creoles would presumably not expect the discrepancies to mainly consist of features which are explicable precisely through recourse to pidginization.

Conclusions To conclude, then, we have established some necessary and impossible assumptions for a simple model of the emergence of a common language. We tested the model on data for Mauritian Creole, with a twofold purpose: to see how well the model performs on real-world data, and to gain insights into four much-debated issues on creole structure and genesis. With reference to Mauritian, at least, our conclusions are:

- It is not necessary to assume a bias towards the lexifier language (that is, targeted learning).
- Creoles may develop quickly.
- Creoles may develop from pidgin languages.
- The input languages to a great extent determine the phonological and syntactic make-up of the new language, but have little or no influence on its morphology.

It might be objected that our model is overly simple. It is indeed simple, but we argue that this is a strength rather than a weakness: since the simpler the model, the fewer controversial assumptions need to be made, and the more transparent the assumptions and their consequences are.

For future research, more creole languages should be investigated, to test the universal applicability of the model and determine whether our results on creole structure and genesis can be generalized to universal principles. However, finding and applying such data is a difficult task, since two necessary prerequisites are fulfilled by few other creologenic

settings: 1) the certainty that the creole emerged locally, rather than having been imported from elsewhere, and 2) the highly detailed demographic data that is available for Mauritius.

Reaching beyond creolization, the model presented here should be potentially applicable to any situation where different cultures merge into a common one. Within linguistics, one such area of interest could be the merging of dialects, and there are likely several more.

Acknowledgments

We would like to thank Philip Baker and three anonymous referees for their comments on a previous version of this paper. We would also like to thank Jonas Sjöstrand for his comments on a previous version of the mathematical analysis appendix. This research was supported by the Swedish Research Council, Riksbankens Jubileumsfond, and the European Research Council under the European Union's Seventh Framework Programme (FP7/2007–2013) / ERC grant agreement no 324233.

References

- Aboh, Enoch. 2009. Competition and selection: that's all! In Enoch Aboh and Norval Smith (eds.), *Complex processes in new languages*, 317–344. Amsterdam: John Benjamins.
- Aboh, Enoch and Umberto Ansaldi. 2007. The role of typology in language creation: a descriptive take. In Umberto Ansaldi, Stephen Matthews, and Lisa Lim (eds.), *Deconstructing creole*, 39–66. Amsterdam: John Benjamins.
- Adone, Dany. 1994. Creolization and language change in Mauritian Creole. In Ingo Plag and Dany Adone (eds.), *Creolization and language change*, 23–43. Tübingen: Niemeyer.
- Alleyne, Mervyn. 1971. Acculturation and the cultural matrix of creolisation. In Dell Hymes (ed.), *Pidginization and creolization of languages*, 169–186. Cambridge: Cambridge University Press.
- Arends, Jacques. 1989. *Syntactic Developments in Sranan: Creolization as a Gradual Process*. Ph.D. thesis, Katholieke Universiteit Nijmegen.
- Arends, Jacques. 1993. Towards a gradualist model of creolization. In Francis Byrne and John Holm (eds.), *Atlantic meets Pacific: A Global view of Creolization*, 371–380. Amsterdam: John Benjamins Publishing Company.
- Arno, Toni and Claude Orian. 1986. *Île Maurice – une société multiraciale*. Paris: l'Harmattan.
- Baker, Philip. 1976. *Towards a social history of Mauritian Creole*. B. phil., University of York.
- Baker, Philip. 1990. Off target? *Journal of Pidgin and Creole Languages* 5: 107–119.
- Baker, Philip. 1995a. Motivation in creole genesis. In Philip Baker (ed.), *From Contact to Creole and beyond*, 3–15. London: University of Westminster Press.

- Baker, Philip. 1995b. Some Developmental Inferences from Historical Studies of Pidgins and Creoles. In Jacques Arends (ed.), *The Early Stages of Creolization*, 1–24. Amsterdam: John Benjamins.
- Baker, Philip. 1999. Investigating the origin and diffusion of shared features among the Atlantic English Creoles. In Philip Baker and Adrienne Bruyn (eds.), *St. Kitts and the Atlantic Creoles*, 315–364. London: Westminster University Press.
- Baker, Philip. 2007. Elements for a sociolinguistic history of Mauritius and its Creole (to 1968). In Philip Baker and Guillaume Fon Sing (eds.), *The making of Mauritian Creole*, 307–333. London: Battlebridge Publications.
- Baker, Philip. 2009. Productive bimorphemic structures and the concept of gradual creolisation. In Rachel Selbach, Hugo Cardoso, and Margot van den Berg (eds.), *Gradual creolization*, 27–53. Amsterdam: John Benjamins.
- Baker, Philip and Chris Corne. 1982. *Isle de France Creoles: Affinities and origins*. Ann Arbor: Karoma Publishers.
- Baker, Philip and Magnus Huber. 2001. Atlantic, Pacific, and world-wide features in English-lexicon contact languages. *English World-Wide* 22(2): 157–208.
- Baker, Philip and Anand Syea. 1991. On the copula in Mauritian Creole, past and present. In Francis Byrne and Thom Huebner (eds.), *Development and structures of creole languages*, 159–175. Amsterdam: John Benjamins.
- Bakker, Peter. 2003. Pidgin inflectional morphology and its implications for creole morphology. In Geert Booij and Jaap van Marle (eds.), *Yearbook of Morphology 2002*, 3–33. New York: Kluwer.
- Baronchelli, Andrea, Maddalena Felici, Vittorio Loreto, Emanuele Caglioti, and Luc Steels. 2006. Sharp transition towards shared vocabularies in multi-agent systems. *Journal of Statistical Mechanics* 2006: P06014.
- Bartens, Angela. 1996. *Der Kreolische Raum*. Geschichte und Gegenwart. Helsinki: Suomalainen Tiedekatemia.
- Beaton, Patrick. 1859. *Creoles and Coolies; or Five years in Mauritius*. London: James Nisbet & Co.
- Bollée, Annegrét. 1977. *Le créole français des Seychelles: Esquisse d'une grammaire, textes, vocabulaire*. Tübingen: Max Niemeyer Verlag.
- Burton, John Wear. 1910. *The Fiji of to-day*. London: Charles H. Kelly.
- Castellano, Claudio, Santo Fortunato, and Vittorio Loreto. 2009. Statistical physics of social dynamics. *Reviews of Modern Physics* 81: 591–646.
- Chaudenson, Robert. 1979. À propos de la genèse du créole mauricien : le peuplement de l'Île de France de 1721 à 1735. *Études Créoles* 1: 43–57.
- Chaudenson, Robert. 1983. Où l'on reparle de la genèse et des structures des créoles de l'océan indien. *Études créoles* 6(2): 157–237.
- Chaudenson, Robert. 1988. Où l'on reparle (mais pour la dernière fois) de la genèse des créoles de l'Océan Indien. *Études Créoles* 11(2).
- Chaudenson, Robert. 1995. *Les créoles*. Paris: Presses Universitaires de France.
- Chaudenson, Robert. 2003. Creolistics and sociolinguistic theories. *International Journal of the Society of Language* 160: 123–146.
- Corne, Chris. 1999. *From French to Creole*. London: University of Westminster Press.

- De Vylder, Bart and Karl Tuyls. 2006. How to reach linguistic consensus: A proof of convergence for the naming game. *Journal of Theoretical Biology* 242: 818–831.
- Debien, Gabriel. 1974. *L’esclavage aux Antilles françaises*. Basse-Terre & Fort-de-France: Société d’histoire de la Guadeloupe/Société d’histoire de la Martinique.
- DeGraff, Michel. 2002. Relexification: a reevaluation. *Anthropological Linguistics* 44(4): 321–414.
- DeGraff, Michel. 2003. Against Creole exceptionalism. *Language* 79(2): 391–410.
- DeGraff, Michel. 2009. Language Acquisition in Creolization and, Thus, Language Change: Some Cartesian-Uniformitarian Boundary Conditions. *Language and Linguistics Compass* 3/4: 888–971.
- Grant, Anthony and Philip Baker. 2007. Comparative Creole typology and the search for the sources of Mauritian Creole features. In Philip Baker and Guillaume Fon Sing (eds.), *The making of Mauritian Creole*, 197–216. London: Battlebridge Publications.
- Guillemin, Diana. 2009. The Mauritian Creole determiner system: A historical overview. In Enoch Aboh and Norval Smith (eds.), *Complex Processes in New Languages*, 173–200. Amsterdam: John Benjamins.
- Hancock, Ian. 1987. A Preliminary Classification of the Anglophone Atlantic Creoles, with Syntactic Data from Thirty-Three Representative Dialects. In Gilbert Glenn (ed.): *Pidgin and Creole Languages: Essays in Memory of John E. Reinecke*: 264–333.
- Haspelmath, Martin, Matthew Dryer, David Gil, and Bernard Comrie (eds.). 2005. *The World Atlas of Language Structures*. Oxford: Oxford University Press.
- Holm, John. 1989. *Pidgins and creoles*. Cambridge: Cambridge University Press.
- Jaeger, Herbert, Luc Steels, Andrea Baronchelli, Ted Briscoe, Morten H. Christiansen, Thomas Griffiths, Gerhard Jäger, Simon Kirby, Natalia L. Komarova, Peter J. Richerson, and Jochen Triesch. 2009. What Can Mathematical, Computational and Robotic Models Tell Us about the Origins of Syntax? In Derek Bickerton and Eörs Szathmáry (eds.), *Biological Foundations and Origin of Syntax*, 385–410. Cambridge, Massachusetts: MIT Press.
- Jourdan, Christine and Roger Keesing. 1997. From Fisin to Pijin: Creolization in process in the Solomon Islands. *Language in Society* 26(3): 401–420.
- Klein, Wolfgang, Rainer Dietrich, and Colette Noyau. 1993. The acquisition of temporality. In Clive Perdue (ed.), *Adult language acquisition: cross-linguistic perspectives, volume II: The results*, 73–118. Cambridge: Cambridge University Press.
- Lefebvre, Claire. 1993. The Role of Relexification and Syntactic Reanalysis in Haitian Creole: Methodological Aspects of a Research Program. In Salikoko Mufwene (ed.), *Africanisms in Afro-American Language Varieties*, 254–279. Athens: University of Georgia Press.
- Lefebvre, Claire. 1997. On the cognitive process of relexification. In Julia Horvath and Paul Wexler (eds.), *Relexification in Creole and Non-Creole Languages*, 72–99. Wiesbaden: Harrassowitz Verlag.
- Loreto, Vittorio, Andrea Baronchelli, and Andrea Puglisi. 2010. Mathematical Modeling of Language Games. In Stefano Nolfi and Marco Mirolli (eds.), *Evolution of Communication and Language in Embodied Agents*, 263–281. Berlin Heidelberg: Springer-Verlag.
- Mather, Patrick-André. 2004. Les créoles à base lexicale européenne et le mythe des pidgins. Manuscript.

- McWhorter, John. 1995. Sisters under the skin: A case for genetic relationship between the Atlantic English-based Creoles. *Journal of Pidgin and Creole Languages* 10(1): 289–333.
- McWhorter, John. 2012. Case closed? Testing the feature pool hypothesis. *Journal of Pidgin and Creole Languages* 27(1): 171–182.
- Mufwene, Salikoko. 2000. Creolization is a social, not a structural, process. In Ingrid Neumann-Holzschuh and Edgar Schneider (eds.), *Degrees of restructuring In creole languages*, 65–84. Amsterdam: John Benjamins.
- Mufwene, Salikoko. 2001. *Ecology of Language Evolution*. Cambridge: Cambridge University Press.
- Mufwene, Salikoko. 2002. Review of "From French to Creole" by Chris Corne. *Journal of Pidgin and Creole Languages* 17(1): 121–150.
- Mufwene, Salikoko. 2006a. Albert Valdman on the development of creoles. In Clancy Clements, Thomas Klingler, Deborah Piston-Hatlen, and Kevin J. Rottet (eds.), *History, Society and Variation*, 203–223. Amsterdam: John Benjamins.
- Mufwene, Salikoko. 2006b. The comparability of new-dialect formation and creole development. *World Englishes* 25(1): 177–186.
- Mufwene, Salikoko. 2007. Population movements and contacts in language evolution. *Journal of language contact* 1: 63–91.
- Mufwene, Salikoko. 2008a. From Genetic Creolistics to Genetic Linguistics: Lessons We Should Not Miss! Paper presentation at the 34th Annual Meeting of the Berkeley Linguistics Society.
- Mufwene, Salikoko. 2008b. *Language evolution: contact, competition and change*. London: Continuum.
- Mühlhäusler, Peter. 1997. *Pidgin and Creole Linguistics. Expanded and revised edition. Westminster Creolistics Series 3*. London: University of Westminster Press.
- Munteanu, Dan. 1996. *El papiamentu, lengua criolla hispánica*. Madrid: Editorial Gredos.
- Nakamura, Makoto, Takashi Hashimoto, and Satoshi Tojo. 2007. Exposure Dependent Creolization in Language Dynamics Equation. In Kôiti Hasida Sakurai, Akito and Katsumi Nitta (eds.), *JSAI 2003/2004, LNAI 3609*, 295–304. Berlin, Germany: Springer-Verlag.
- Neumann-Holzschuh, Ingrid. 2006. Gender in French creoles: The story of a loser. In Clancy Clements, Thomas Klingler, Deborah Piston-Hatlen, and Kevin J. Rottet (eds.), *History, Society and Variation*, 251–272. Amsterdam: John Benjamins.
- Nowak, Martin A., Natalia L. Komarova, and Partha Niyogi. 2001. Evolution of Universal Grammar. *Science* 291: 114–118.
- Owens, Jonathan. 1996. Arabic-based Pidgins and Creoles. In Sarah Thomason (ed.), *Contact Languages. A wider perspective*, 125–172. Amsterdam: John Benjamins.
- Papen, Robert. 1978. *The French-based creoles of the Indian Ocean: an analysis and comparison*. Ph.D. thesis, University of California at San Diego.
- Parkvall, Mikael. 1999. Feature selection and genetic relationships among Atlantic Creoles. In Magnus Huber and Mikael Parkvall (eds.), *Spreading the Word*, 29–66. London: Westminster University Press.
- Parkvall, Mikael. 2000. Reassessing the role of demographics in language restructuring. In Ingrid Neumann-Holzschuh and Edgar Schneider (eds.), *Degrees of Restructuring in Creole Languages*, 195–213. Amsterdam: John Benjamins.
- Parkvall, Mikael and Stéphane Goyette. forthcoming. Principia Creolica.

- Plag, Ingo. 2011. Creolization and admixture: Typology, feature pools, and second language acquisition. *Journal of Pidgin and Creole Languages* 26(1): 89–110.
- Roberts, Sarah. 2000. Nativization and the genesis of Hawaiian Creole. In John McWhorter (ed.), *Language change and language contact*, 257–300. Amsterdam: John Benjamins.
- Roberts, Sarah. 2004. *The emergence of Hawaii Creole English in the early 20th century: the sociohistorical context of creole genesis*. Ph.D. thesis, Stanford University.
- Roberts, Sarah and Joan Bresnan. 2008. Retained inflectional morphology in pidgins: A typological study. *Linguistic Typology* 12: 269–302.
- Satterfield, Teresa. 2001. Toward a socio-genetic solution: examining language formation processes through SWARM modeling. *Social Science Computing Review* 19(3): 281–295.
- Satterfield, Teresa. 2008. Back to Nature or Nurture: Using computer models in creole genesis research. In G. Jäger Eckardt, Regine and Tonjes Veenstra (eds.), *Variation, Selection, Development: Probing the Evolutionary Model of Language Change*, 143–178. Amsterdam: Mouton de Gruyter.
- Schuchardt, Hugo. 1980. *Pidgin and Creole Languages*. Cambridge: Cambridge University Press.
- Siegel, Jeff. 1990. Pidgin Hindustani in Fiji. In Jeremy H. C. S. Davidson (ed.), *Pacific Island languages*, 173–196. London: SOAS.
- Smith, Norval. 2006. Very rapid creolization in the framework of the restricted motivation hypothesis. *Language Acquisition and Language Disorders* 42: 49–65.
- Steels, Luc. 1996. A self-organizing spatial vocabulary. *Artificial Life* 2: 319–332.
- Steels, Luc. 1997. Language Learning and Language Contact. In Antal van den Bosch Daelemans, Walter and Ton Weijters (eds.), *Proceedings of the workshop on Empirical Approaches to Language Acquisition*, 11–24. Prague: ECML.
- Toussaint, Auguste. 1971. *Histoire de l'île Maurice*. Paris: Presses Universitaires de France.
- Valdman, Albert. 2006. Response to Parkvall. *Studies in Second Language Acquisition* 28(3): 516–517.
- Vaughan, Megan. 2005. *Creating the creole island*. Durham: Duke University Press.
- Walter, Henriette. 1982. *Enquête phonologique et variétés régionales du français*. Paris: Presses universitaires de France.
- Williams, Eric. 1991. Capitalism and Slavery. In Hilary Beckles and Verene Shepherd (eds.), *Caribbean Slave Society and Economy*, 120–129. New York: The New Press.
- Winford, Don. 2008. Atlantic Creole Syntax. In Silvia Kouwenberg and John Singler (eds.), *Handbook of Pidgin and Creole Studies*, 19–47. Chichester: Blackwell.

A Appendix

A.1 Mathematical Analysis

A.1.1 Basic Model

We have a population $A = \{a_1, a_2, \dots, a_N\}$ of N agents and a set $W = \{w_1, w_2, \dots, w_n\}$ of n words (or any transmissible entities). For each agent a_i in A , there is a probability vector $\mathbf{p}_i = (p_{i1}, p_{i2}, \dots, p_{in})$, where p_{ik} denotes the probability that a_i will use the word w_k in W in his next interaction, and $\sum_{k=1}^n p_{ik} = 1$.

In a round of interactions, for each agent a_i another agent a_j , $j \neq i$, is chosen at random to interact with. Thus, each agent will be part of at least one interaction, and on average two. In the interaction, a_i says the word w_{k_1} with probability p_{ik_1} , and a_j responds with w_{k_2} with probability p_{jk_2} . Both agents will then update their probability vectors to $\mathbf{p}'_i$ and $\mathbf{p}'_j$, respectively, such that

$$\begin{cases} p'_{ik_2} &= p_{ik_2} + (1 - p_{ik_2})\ell \\ p'_{ik} &= p_{ik} - p_{ik}\ell, \quad k \neq k_2 \end{cases} \quad (1)$$

and

$$\begin{cases} p'_{jk_1} &= p_{jk_1} + (1 - p_{jk_1})\ell \\ p'_{jk} &= p_{jk} - p_{jk}\ell, \quad k \neq k_1 \end{cases},$$

where ℓ is a real-valued constant (denoting the amount of learning an agent makes in an interaction). Note that $\sum_{k=1}^n p'_{ik} = \sum_{k=1}^n p_{ik} = 1$. The population may change between rounds.

Lemma 1. *Assume that the agents $a_i, a_j \in A$, $i \neq j$, have an interaction. For each word $w_k \in W$, the expected value of the change in usage after the interaction is*

$$\mathbb{E}[(p'_{ik} + p'_{jk}) - (p_{ik} + p_{jk})] = 0.$$

Proof. Choose a word $w_k \in W$ arbitrarily. The expected change of p_i after the interaction is

$$\mathbb{E}[p'_{ik} - p_{ik}] = p_{jk}(1 - p_{ik})\ell - (1 - p_{jk})p_{ik}\ell = (p_{jk} - p_{ik})\ell,$$

and similarly for p_j . Thus, the expected change of the frequency of the word in the population is

$$\mathbb{E}[(p'_{ik} - p_{ik}) + (p'_{jk} - p_{jk})] = (p_{jk} - p_{ik})\ell + (p_{ik} - p_{jk})\ell = 0.$$

□

Thus, the process is neutral in the sense that changes in frequencies between the words are subject to randomness exclusively, with no bias towards any of the words. The frequencies converge eventually to any of the n equilibria where only one word is present, but adding agents may shift the equilibrium, also after fixation.

A.1.2 Conservatism

We will extend the model by adding conservatism. Let t_i denote the number of rounds of interactions after $a_i \in A$ entered the population. Let $\delta : \mathbb{N} \rightarrow [0, 1]$ be a decreasing function from the natural numbers to the unit interval. By multiplying ℓ by δ , agents will change their vocabulary less and less with time, and Equations 1 will change into

$$\begin{cases} p'_{ik_2} &= p_{ik_2} + (1 - p_{ik_2})\ell\delta(t_i) \\ p'_{ik} &= p_{ik} - p_{ik}\ell\delta(t_i), \quad k \neq k_2 \end{cases} \quad (2)$$

and accordingly for p'_{jk} .

With time, agents will be more and more static. If δ tends to zero, then agents will converge to being completely static, that is, they will not update their preference vector. We have thus added a bias towards words that are frequent early in time.

Suppose that the agents would interpret two different words w_r and w_s as the same word. For example, if the words are instead grammatical features, then two languages may share the same feature. Without loss of generality, let us assume that both words are interpreted as w_r . Let $\mathbf{q}_i = (q_{i1}, q_{i2}, \dots, q_{in})$ denote the probability vector for a_i , where initially, at time $t_i = 0$, $q_{ir} = p_{ir} + p_{is}$, $q_{is} = 0$ and $q_{ik} = p_{ik}$, $k \notin \{r, s\}$, and q'_{ik} is defined as p'_{ik} in Equations 2 for all k .

Proposition. *For any agent $a_i \in A$ and words $w_r, w_s \in W$, at any stage in time,*

$$q_{ir} = p_{ir} + p_{is}.$$

Proof. Assume that for an arbitrarily chosen stage in time, $q_{ir} = p_{ir} + p_{is}$. We know this equality to hold when $t_i = 0$. The value of q_{ir} after an

interaction with an agent a_j at this stage in time is then

$$\begin{aligned}
q'_{ir} &= \begin{cases} q_{ir} + (1 - q_{ir})\ell\delta(t_i) & \text{with probability } q_{jr} \\ q_{ir} - q_{ir}\ell\delta(t_i) & \text{with probability } 1 - q_{jr} \end{cases} \\
&= \begin{cases} p_{ir} + p_{is} + (1 - p_{ir} - p_{is})\ell\delta(t_i) & \text{with probability } p_{jr} + p_{js} \\ p_{ir} + p_{is} - (p_{ir} + p_{is})\ell\delta(t_i) & \text{with probability } 1 - p_{jr} - p_{js} \end{cases} \\
&= \begin{cases} p_{ir} + (1 - p_{ir})\ell\delta(t_i) + p_{is} - p_{is}\ell\delta(t_i) & \text{with probability } p_{jr} \\ p_{ir} - p_{ir}\ell\delta(t_i) + p_{is} + (1 - p_{is})\ell\delta(t_i) & \text{with probability } p_{js} \\ p_{ir} - p_{ir}\ell\delta(t_i) + p_{is} - p_{is}\ell\delta(t_i) & \text{with probability } 1 - p_{jr} - p_{js} \end{cases} \\
&= p'_{ir} + p'_{is}.
\end{aligned}$$

□

Thus, if for a word $w_k \in W$, $p_{ik} > \frac{1}{2}$, then $q_{ir} < \frac{1}{2}$ for all $r \neq k$. This means that no matter how we group words together, the group containing the word w_k that dominates without groupings will still dominate. For our applications, this means that a language that dominates the vocabulary of a mixed language must also dominate the syntax and phonology, since other languages cannot overcome this dominance by joint forces.

A.1.3 Coordination

Given the results in the previous section, we need to break the additivity of the update function with respect to merging words. This can be achieved by allowing for different learning rates ℓ_c and ℓ_u depending on whether an agent used the same word as he heard in an interaction (whether communication was coordinated/successful or uncoordinated/unsuccessful). If we modify Equations 1 accordingly, then, after an interaction in which an agent a_i has uttered the word w_{k_1} and a_j the word w_{k_2} , a_i will update his probability vector so that

$$\begin{cases} p'_{ik_2} = p_{ik_2} + (1 - p_{ik_2})\ell_c \\ p'_{ik} = p_{ik} - p_{ik}\ell_c, k \neq k_2 \\ p'_{ik_2} = p_{ik_2} + (1 - p_{ik_2})\ell_u \\ p'_{ik} = p_{ik} - p_{ik}\ell_u, k \neq k_2 \end{cases} \quad , \text{ if } k_1 = k_2, \text{ and} \quad (3)$$

and similarly for a_j .

Let $\bar{p}_{\cdot k} = \sum_{i=1}^N p_{ik}/N$ be the average frequency of word w_k being used in the population A . We are interested in the behaviour of $\bar{p}'_{\cdot k} - \bar{p}_{\cdot k}$, that is, the change of the frequency after an interaction. For convenience, we will let $R_k\ell_c$ denote the expected change of $\bar{p}_k$ due to an interaction where both agents used the word w_k , $S_k\ell_c$ the expected change from both agents

using a word $w_{k'}$, $k' \neq k$, and $T_k \ell_u$ from the agents using different words. Now we can write the expected change of the use of w_k after a random interaction as

$$\mathbb{E}[\bar{p}'_{\cdot k} - \bar{p}_{\cdot k}] = (R_k - S_k) \ell_c + T_k \ell_u. \quad (4)$$

Assume, without loss of generality, that we have ordered the words such that, at a given point in time, $\bar{p}_{\cdot 1} > \bar{p}_{\cdot k}$ for all $k \neq 1$. We have the following lemma.

Lemma 2. *Using the notation above, and assuming $\bar{p}_{\cdot 1} < 1$,*

$$R_1 - S_1 > 0.$$

Proof. The expected update after a randomly chosen agent a_i has met a randomly chosen agent a_j is

$$\begin{aligned} R_1 - S_1 &= \frac{1}{N} \frac{1}{N-1} \sum_i \sum_{j \neq i} \left(p_{i1} p_{j1} (1 - p_{i1} + 1 - p_{j1}) - \sum_{k=2}^n p_{ik} p_{jk} (p_{i1} + p_{j1}) \right) \\ &= \{\text{include encountering self in sum}\} \\ &= \frac{1}{N^2} \sum_i \left(\sum_j \left(p_{i1} p_{j1} (1 - p_{i1} + 1 - p_{j1}) - \sum_{k=2}^n p_{ik} p_{jk} (p_{i1} + p_{j1}) \right) \right. \\ &\quad \left. - \left(p_{i1} (1 - p_{i1}) - \sum_{k=2}^n p_{ik} p_{i1} \right) \right) \\ &= \frac{1}{N^2} \sum_i \left(\sum_j \left(2p_{i1} p_{j1} - \sum_k p_{ik} p_{jk} (p_{i1} + p_{j1}) \right) - \left(p_{i1} - \sum_k p_{ik} p_{i1} \right) \right) \\ &= \frac{1}{N^2} \sum_j \left(2p_{i1} p_{j1} - \sum_k p_{ik} p_{jk} (p_{i1} + p_{j1}) \right) \\ &= \{\text{use symmetry between } p_i \text{ and } p_j\} \\ &= 2\bar{p}_{\cdot 1}^2 - \frac{2}{N^2} \sum_i \sum_j \sum_k p_{ik} p_{jk} p_{i1} \\ &= 2\bar{p}_{\cdot 1}^2 - \frac{2}{N^2} \sum_i \left(p_{i1} \sum_k \left(p_{ik} \sum_j p_{jk} \right) \right) \\ &= 2\bar{p}_{\cdot 1}^2 - \frac{2}{N} \sum_i \left(p_{i1} \sum_k \bar{p}_{\cdot k} p_{ik} \right) \\ &> 2\bar{p}_{\cdot 1}^2 - \frac{2}{N} \sum_i p_{i1} \sum_k \bar{p}_{\cdot 1} p_{ik} \\ &= 0. \end{aligned}$$

□

We can now prove our main theorem.

Theorem. *Using the notation above, and assuming $\bar{p}_{\cdot 1} < 1$,*

$$\text{sign}(\mathbb{E}[\bar{p}'_{\cdot 1} - \bar{p}_{\cdot 1}]) = \text{sign}(\ell_c - \ell_u).$$

Proof. Using the notation from Equation 4, we have that

$$\mathbb{E}[\bar{p}'_{\cdot 1} - \bar{p}_{\cdot 1}] = (R_1 - S_1)\ell_c + T_1\ell_u.$$

If $\ell_c = \ell_u$, then Equations 3 are the same as Equations 1, so by Lemma 1 we have that

$$((R_1 - S_1) + T_1)\ell_c = 0 \Leftrightarrow T_1 = S_1 - R_1,$$

from which follows that

$$(R_1 - S_1)\ell_c + T_1\ell_u = (R_1 - S_1)\ell_c - (R_1 - S_1)\ell_u = (R_1 - S_1)(\ell_c - \ell_u).$$

Finally, by Lemma 2,

$$\text{sign}((R_1 - S_1)(\ell_c - \ell_u)) = \text{sign}(\ell_c - \ell_u).$$

□

Thus, if $\ell_c = \ell_u$, then word frequencies will not change on average, but due to random fluctuations, usage will converge very slowly to one word as other words randomly get forgotten.

If $\ell_c < \ell_u$, then we will have convergence to all words being used equally often on average.

Finally, if $\ell_c > \ell_u$, then usage will converge to a single word, ignoring random fluctuations, being the word with the highest average usage at any given point in time. Only immigration and major random fluctuations may shift which word is used the most.

A.2 The Origins of Mauritian Creole

A.2.1 What Mauritian Creole Looks Like

The language is lexically based on French, that is, the overwhelming majority of the lexicon is drawn from that language. However, the grammar is very unlike French as we know it. To give the reader a taste of the differences, Table 4 presents a short piece of text chosen at random from the leftist monthly *Revi lalit deklas* (“The class struggle”²⁰), dealing with the 2011 nuclear disaster in Fukushima (elements commented on below are *emphasized*).

The reader can easily see that the bulk of the lexicon is French (e.g. *tramblemandeter* ← *tremblement de terre* ‘earthquake’, *travayer* ← *travailleur* ‘worker’, etc.), even though some words might be slightly difficult to analyze. Many have had an erstwhile article attached to them, which no longer functions as such (e.g. *larzan* ← *l’argent* means ‘money’, not, as one might expect *‘the money’, and *lavi* ← *la vie* is ‘life’, not *‘the life’). Definiteness is instead encoded by *sa...la*. Other words are French in origin even though they, for various reasons, do not immediately appear familiar to a French speaker (e. g. *zot* ← *eux autres* ‘they’, *lor* ← *là-haut* ‘(up)on’, *bann* ← *bande* ‘plural marker’, *ek* ← *avec* ‘and’). No less saliently, many distinctions made in French are absent from Mauritian, especially in the domain of morphology. There is no grammatical gender, and no inflections for person, number, tense, mood or aspect. The latter two are instead indicated by free preverbal markers, such as *pe* ‘progressive aspect’ and *ti* ‘past’ (compare *Zot ti pe gayne* ‘They had been receiving’ below with the corresponding French *Ils recevaient*, where the suffix marks person, number, tense, and aspect). Note also the lack of grammatical distinction above between forms such as ‘they’ and ‘their’ (French *ils* vs. *leur*), both being *zot* in Mauritian.

In short, Mauritian is a language rather unlike French (and indeed not mutually intelligible with it), despite its vocabulary. The origins of Mauritian grammar are less obvious, and rather controversial. What everyone would agree on is that the descendants of slaves in Mauritius (as well as many post-emancipation immigrants) no longer speak the languages of their ancestors. Instead, they speak something which has a clearly French

²⁰ As in so many other countries, writing in the language actually spoken by the people, is itself still considered thoroughly radical. Only after having chosen this text did we realize that it had been written by a non-native speaker, and contained a few errors. We therefore made a few changes in accordance with suggestions from native speakers.

²Our translation.

³We wish to thank Patrick Ostellari and Aymeric Daval-Markussen for providing a French version.

Mauritian	English²	French³
<p>Apré <i>tranblemandeter ek</i> tsunami dan Zapon, nu <i>pe</i> truv sa nuvel la byin trakasan <i>lor</i> problem otur reakter nikleer Fukushima preske tulezur <i>lor</i> TV.</p>	<p>After the <i>earthquake and</i> the tsunami in Japan, we were <i>seeing</i> truly bothersome news <i>about</i> the problems with the nuclear reactor <i>in</i> Fukushima almost daily on television.</p>	<p>Après le <i>tremblement de terre et</i> le tsunami au Japon, nous <i>voyions</i> appris des nouvelles bien préoccupantes <i>concernant</i> les problèmes autour du réacteur nucléaire de Fukushima presque quotidiennement à la télévision.</p>
<p>Anmemtan <i>travayer</i> ki finn sibir buku radyasyon dan aksidan nikleer Tchernobyl, 25 an desela, <i>pe</i> manifeste</p>	<p>At the same time, the <i>workers</i> who were exposed to plenty of radiation from the nuclear accident in Chernobyl 25 years ago, were <i>demonstrating</i>, because the Ukrainian government said it had no more <i>money</i> for medical treatment <i>and</i> material support that had until then been given to the workers</p>	<p>Au même moment, les <i>travailleurs</i> qui avaient été exposés à de fortes radiations lors de l'accident nucléaire de Tchernobyl, il y a 25 ans, <i>ont</i> manifesté parce que le gouvernement ukrainien <i>avait</i> annoncé qu'il n'avait plus d'<i>argent</i> pour assurer le soutien médical <i>et</i> matériel dont bénéficiaient jusque-là les <i>travailleurs</i></p>
<p>parski Guvernman Ukrenn <i>pe</i> dir nepli ena <i>larzan</i> pu donn sa <i>bann</i> <i>travayer</i> la sutyin medikal <i>la ek</i> materyel ki <i>zot ti pe</i> gayne</p>	<p>who risked <i>their</i> health <i>and</i> even lives to get the Chernobyl reactor under control.</p>	<p>qui avaient risqué <i>leur</i> santé et même <i>leur</i> <i>vie</i> pour reprendre le contrôle du réacteur de Tchernobyl.</p>

Table 4: Text example.

lexicon, but a structural profile rather different from that of French – a structure whose origins are less clear. Some sort of language shift has thus taken place, but is the resulting language the outcome of an *attempt* to shift to French, or is it simply a successful attempt at bridging a communication gap where communication was the only goal (rather than the acquisition of a specific language)? In other words, would the slogan of the original language creators have been “Let’s try to learn French”, or would it rather have been “Let’s try to communicate by whatever means available to us”?

A.2.2 Demographics and the Development of Society and Language in Early Mauritius

The development of creole languages has attracted a good deal of scholarly attention during the past half a century, and attempts at understanding the process often involves historical and demographical studies, intended to shed light on the nature of the language contact situation. In this context, Mauritius has (as already mentioned) two immense advantages as a field of investigation, compared to other countries where creoles are spoken. First, we can be reasonably certain that the language emerged *in situ*, rather than having been brought in from the outside, which is the case for some other creoles (e.g. Baker 1999; Baker and Huber 2001; McWhorter 1995; Parkvall 1999)²³. Secondly, the demographics are exceptionally well documented for the first fifteen years of settlement – Baker and Corne (1982) offers details on arrivals and departures between 1721 and 1735 which are unparalleled for any other creole speaking country. For most of these people, a mother tongue can be inferred on the basis of geographical provenance (or, occasionally, other known facts).

The initial settlers – free and enslaved – in Mauritius included people from the following locations, for which we have assumed the native languages presented in Table 5 and Fig. 7.

²³ One creolist, Robert Chaudenson (e.g. 1979; 1983; 1988), maintains that Mauritian was imported from the neighboring island of Réunion, but this account by and large lacks support in other literature.

Origin	Assumed language
Europe	French ⁵
Réunion	French or Bourbonnais ⁶
Madagascar	Malagasy
Senegal	6/7 Wolof, 1/7 Manding languages ⁷
French India	1/2 Tamil, 1/2 Bengali
Benin	Gbe languages
Mozambique	Makhuwa and related Bantu languages

Table 5: Native languages from different origins.

The Mozambican Bantu languages, however, only enter the scene after 1735, that is, right after the end of our simulation. They are therefore of no relevance in the following, but are mentioned here since they were present (in large numbers) in the early history of Mauritian.

The development of this colony was exceptionally rapid. In the history of a typical plantation colony, whites usually form a majority in the founding years – after all, they are needed to develop the infrastructure before massive slave importation can take place. As the plantation economy develops into a more large-scale venture, the influx of slaves increases, until they form a majority of the population. The initial period of white demographic dominance is sometimes referred to in the creolistic literature as the *homestead phase* (or *société d'habitation*), which is followed by the *plantation phase* (or *société de plantation*). The transition between these took a mere nine years in Mauritius, as compared to a couple of decades in most Caribbean colonies, and 64 years in neighboring Réunion (Parkvall, 2000). After a difficult start in 1721–5, Mauritius surpassed its

⁵Some (in particular the Swiss soldiers) might have spoken German, and the presence of a small number of Bretons can be inferred from a handful of Breton words in Mauritian. Being in French service, however, all Europeans can be assumed to have known and used primarily French in the colony, and we have therefore ignored other European languages. The Swiss did in any case not remain for long in Mauritius.

⁶The island of Réunion was known at the time as Île Bourbon, and Bourbonnais refers to an early variety of Réunionnais Creole. This is also a French-lexicon variety, but very different from Mauritian, and not mutually intelligible with it. Its most characteristic feature is that it is considerably closer to French than Mauritian is. As already mentioned, Bourbonnais is what Chaudenson considers the ancestor of Mauritian. Apart from the fact that this, in our view, has been resoundingly disproved (primarily through Baker and Corne 1982), it is unclear to what extent Bourbonnais actually differed from French at the time. Because of this (and there not being a description of Bourbonnais in the first place), we wound up treating the Réunionnais as speakers of French.

⁷These proportions are due to Philip Baker (p. c.).

Figure 7: Approximate geographical location of the languages relevant to the initial peopling of Mauritius. Also indicated are the other islands and archipelagos of the Indian Ocean where a French creole is (or, in the case of the Chagos group, was) spoken.

erstwhile “big brother” Réunion in economic importance already in 1739 (Baker, 2007: 309). Figure 8 summarizes this demographic development up to the year 1800.

This demographic development is typical of slavery-dependent plantation colonies established by Europeans in the tropics, and great significance is usually attached to it in the creolistic literature. Indeed, the numbers do suggest a limited access of the slave majority to native-speaking role models.

A.2.3 When Did Mauritian Become a Full Language?

There is no complete consensus on how long it took for Mauritian to emerge – no language ever stops developing, and there is therefore no point in time where Mauritian could be considered “finished” in the sense that it would longer change. The relevant question is rather whether a point in time could be identified after which the users of the language (or language-to-be) would share a fair amount of common conventions, and whether these conventions would be reasonably similar to the grammar of modern Mauritian. Clearly, we could not expect to be able to identify any

Figure 8: Demographic development of Mauritius by five-year intervals. Sources: Multiple, in particular Arno and Orian (1986: 171), Baker and Corne (1982), Beaton (1859: 59), Baker (1976), Toussaint (1971: 50), and Vaughan (2005: 47). The figures also include some interpolation in order to produce similarly-sized intervals.

absolute date. Baker (2009) is able to date certain key developments with some degree of accuracy, but primarily those refinements which postdate the moment at which we might want to identify Mauritian as a code separate from French, or for that matter “broken French.”

The first known explicit mention of Mauritian as a language separate from French is from 1773 and is found in an advertisement concerning the whereabouts of a runaway slave boy, where the boy is said not to speak “the Creole language” (Baker and Corne, 1982: 248).

Additional clues are provided by the linguistic offshoots exported to neighboring islands. Varieties (currently or until recently) spoken in the Seychelles, Rodrigues, and the Chagos Archipelago are such examples, and while Seychellois is normally considered a separate language, this is more due to the political separation of the two countries than to linguistic differences between them – mutual comprehension between the two “presents no problems whatsoever”²⁷ (Bollée, 1977: 7).

These four varieties could therefore be considered dialects of one and the same language²⁸. The settlements were all established in the second half of the 18th century (and there is nothing to suggest that the offshoots changed dramatically after the implantation of the first Mauritians), and

²⁷ “L’intercompréhension entre locuteurs des deux idiomes ne présente [...] aucune difficulté”.

²⁸ A detailed comparison can be found in Papen (1978).

indicate that Mauritian had reached *roughly* its modern form no later than that.

At least four key features of Mauritian²⁹ are attested before 1750 (Baker, 1995b: 6–7), and given the rarity of writings in Creole, it is of course quite likely that many features were present a good deal of time before they are attested in written records.

It is thus difficult to *prove* the existence of Mauritian before the 1770s or so, but there are good reasons to assume that the formation of this and other creole languages was a rather rapid affair (Smith, 2006; Parkvall and Goyette, forthcoming).

Our key reference when it comes to the demographic situation (Baker and Corne, 1982), only covers the 1721–1735 period so far as migration history goes. It would be welcome to have equally detailed data for another couple of decades, but we do believe that the main aspects of the language were in place early. More importantly, later immigrants are, as we shall see, in any case likely to have had a far less decisive impact on the language than did early arrivals.

²⁹ Zero copula, a definite article derived from a demonstrative, the use of an oblique pronoun form in subject position, and a preverbal completive marker. In fact, the very first attested Mauritian sentence is very close to the modern language. In 1749, a slave is reported to have uttered "Ça blanc là li beaucoup malin; li couri beaucoup dans la mer là-haut; mais Madagascar li là." The translation suggested by (Baker and Syea, 1991: 165) is 'That white man is very cunning, he changed directions many times at sea over there; but Madagascar is there'. (Baker and Syea, 1991: 173) note that the only two differences consist in a) *beau-**coup* as an adjective intensifier (modern Mauritian would use a reflex of French *bien*), and b) and *couri* ('wander about', from French *courir*), which is indeed subsequently attested, but which became obsolete in the 19th century.

A.3 Linguistic data

The linguistic data used in this paper is presented in the table below. The data has been compiled from three sources, and the first column depicts which source and, if applicable, an identifying label. For labels with the first letter F, the source is Wals, for G it is Grant and Baker (2007), and for fono it is UCLA Phonological Segment Inventory Database.

The second column describes the linguistic item in question.

The remaining columns give the values for each language in the following order: Mauritian, French, Malagasy, Tamil, Bengali, Gbe, Wolof and Manding. Typically, variables that take on only binary values of 1 and 0 indicate existence and non-existence, while other values depict more complex categories. It is not always possible to make clear-cut distinctions for certain languages, giving two values in some positions. For these, we have assumed that the agents in question in the simulation initially have a 50% preference for each of the two values.

#	Description	Mn	Fh	My	Tl	Bi	Ge	Wf	Mg
<i>Phonology</i>									
F12b	Complex syllable structure	1	1	0	0	1	0	1	0
F13	Tone	1	1	1	1	1	2	1	2
fono	/a/	0	1	0	0	0	0	0	0
fono	/c/	0	0	0	0	0	0	1	0
fono	/ɔ/	0	1	0	0	1	1	1	1
fono	/d/	0	0	0	0	1	0	0	0
fono	/ɖ/	0	0	0	0	1	1	0	1 0
fono	/ɗ/	0	0	1	0	0	1 0	0	0
fono	/ẽ/	0	0	0	0	1	1	0	1
fono	/ɛ/	0	1	0	0	1	1	1	1
fono	/ə/	0	1	0	0	0	0	0	0
fono	/gb/	0	0	0	0	0	1	0	0
fono	/ɣ/	0	1	0	0	0	0	0	0
fono	/ĩ/	0	0	0	0	1	1	0	1
fono	/ʝ/	0	0	0	0	0	0	1	0
fono	/kp/	0	0	0	0	0	1	0	0
fono	/l/	0	0	0	1	0	0	0	0
fono	/mb/	0	0	1	0	0	0	1 0	0
fono	/mp/	0	0	0	0	0	0	0	0
fono	/ɲ/	0	1	0	1	0	1	1	1
fono	/ŋ/	0	0	0	1	0	0	0	0
fono	/nd/	0	0	1	0	0	0	1 0	0

fono	/ndz/	0	0	1	0	0	0	0	0
fono	/ɲɟ/	0	0	0	0	0	0	0	0
fono	/nt/	0	0	0	0	0	0	0	0
fono	/ɲ/	0	0	0	1	1	1 0	1	1 0
fono	/ɲg/	0	0	1	0	0	0	0	0
fono	/õ/	0	0	0	0	1	1	0	1
fono	/ø/	0	1	0	0	0	0	0	0
fono	/œ/	0	1	0	0	0	0	0	0
fono	/œ̃/	0	1	0	0	0	0	0	0
fono	/ɸ/	0	0	0	0	0	1 0	0	0
fono	/q/	0	0	0	0	0	0	1 0	0
fono	/r/	0	0	0	1	1	1 0	1 0	1 0
fono	/ʀ/	0	0	0	0	1	0	0	0
fono	/ʁ/	0	1	0	0	0	0	0	0
fono	/ɹ/	0	0	0	1	0	0	0	0
fono	/ɽ/	0	0	1	1	0	0	1 0	0
fono	/ʃ/	0	1	0	0	1	0	0	1 0
fono	/t/	0	0	0	1	1	0	0	0
fono	/tʃ/	0	0	0	1	1	1	0	1 0
fono	/ts/	0	0	1	0	0	1 0	0	0
fono	/ũ/	0	0	0	0	1	1	0	1
fono	/v/	0	0	0	1	0	0	0	0
fono	/x/	0	0	0	0	0	1	1 0	0
fono	/y/	0	1	0	0	0	0	0	0
fono	/ɜ/	0	1	0	0	0	0	0	0
fono	/ʎ/	0	0	0	0	0	1 0	0	0
fono	/ʐ/	0	0	0	0	0	0	1 0	0
fono	/β/	0	0	0	0	0	1 0	0	0
fono	/χ/	0	0	0	0	0	0	1 0	0
fono	/b ^{fi} /	0	0	0	0	0	0	0	0
fono	/d/	0	0	0	0	0	0	0	0
fono	/d ^{fi} /	0	0	0	0	0	0	0	0
fono	/d ^{fi} /	0	0	0	0	0	0	0	0
fono	/ɖ ^{fi} /	0	0	0	0	0	0	0	0
fono	/g ^{fi} /	0	0	0	0	0	0	0	0
fono	/k ^h /	0	0	0	0	0	0	0	0
fono	/p ^h /	0	0	0	0	0	0	0	0
fono	/t ^{fi} /	0	0	0	0	0	0	0	0
fono	/t ^h /	0	0	0	0	0	0	0	0
fono	/t ^h /	0	0	0	0	0	0	0	0
fono	/tʰ ^h /	0	0	0	0	0	0	0	0

fono	Tones	0	0	0	0	0	1	0	1
fono	Vowel harmony	0	0	0	0	1	1	1	0
fono	/a/	1	1	1	1	1	1	1	1
fono	/ã/	1	1	0	0	1	1	0	1
fono	/b/	1	1	1	0	1	1	1	1
fono	/õ/	1	1	0	0	1	1	0	1
fono	/d/	1	1	1	0	1	1	1	1
fono	/e/	1	1	1	1	1	1	1	1
fono	/ê/	1	1	0	0	1	1	0	1
fono	/f/	1	1	1	0	1	1	1	1
fono	/g/	1	1	1	0	1	1	1	1
fono	/ɣ/	1	0	0	0	0	1 0	0	0
fono	/h/	1	0	1	0	1	1 0	1 0	1
fono	/i/	1	1	1	1	1	1	1	1
fono	/j/	1	1	1	1	1	1	1	1
fono	/k/	1	1	1	1	1	1	1	1
fono	/m/	1	1	1	1	1	1	1	1
fono	/n/	1	1	1	1	1	1	1	1
fono	/o/	1	1	1	1	1	1	1	1
fono	/p/	1	1	1	1	1	0	1	1
fono	/s/	1	1	1	0	1	1	1	1
fono	/t/	1	1	1	1	1	1	1	1
fono	/u/	1	1	1	1	1	1	1	1
fono	/v/	1	1	1	0	0	1	0	0
fono	/w/	1	1	1	0	1	1	1	1
fono	/z/	1	1	1	0	1	1	0	1 0

Lexicon

F134b	Green=blue	0	0	0	0	0	0	0	0
F134c	Yellow=green	0	0	0	0	0	0	0	0
F134d	Black=blue	0	0	0	0	0	0	0	0
F135b	Red=yellow	0	0	0	0	0	0	1	0
F136	M-T Pronouns	2	2	1	1	2	1	1	1
F137	N-M Pronouns	1	1	1	1	1	1	1	1

Simple Clauses

F116	Polar Questions	1	1	1	2	1	1	1	1
F118b	Predicative adjectives = verbal	1	0	1	0	0	1	1	1
F118c	Predicative adjectives = nonverbal	0	1	0	1	1	1	0	1

F119	Nominal and Locational Predication	2	2	1	2	2	1	1	1
F120	Zero Copula for Predicate Nominals	1	1	1	2	2	1	1	1
<i>Word Order</i>									
F81	Order of Subject, Object and Verb	2	2	4	1	1	2	2	1
F82	Order of Subject and Verb	1	1	2	1	1	1	1	1
F83	Order of Object and Verb	2	2	2	1	1	2	2	1
F84	Order of Object, Oblique and Verb	1	1	1	3	3	1	1	5
F85b	Postpositions	0	0	0	1	1	1	0	1
F85c	Prepositions	1	1	1	0	0	1	1	1
F87	Order of Adjective and Noun	2	2	2	1	1	2	2	2
F88b	Dem_N	1	1	1	1	1	0	0	1 0
F88c	N_Dem	0	0	1	0	0	1	1	1 0
F89	Order of Numeral and Noun	1	1	2	1	1	2	1	2
F92	Position of Polar Question Particles	1	1	5	6	2	2	1	2
F93	Position of Interrogative Phrases in Content Questions	1	1	1	2	2	2	1	2
F94	Order of Adverbial Subordinator and Clause	1	1	1	4	1	1	1	1 2
F95	Relationship between the Order of Object and Verb and the Order of Adposition and Noun Phrase	4	4	4	1	1	3 5	4	1
F96	Relationship between the Order of Object and Verb and the Order of Relative Clause and Noun	4	4	4	1	5	4	4	5 2

F97	Relationship between the Order of Object and Verb and the Order of Adjective and Noun	4	4	4	1	1	4	4	2
GB09	Pluralisation marker precedes the noun stem	1	1	0	0	0	1	1	0
GB10	Definiteness marking mainly postposed	1	0	0	0	1	1	1	0
GB22	Preverbal negation	1	0	1	0	0	1	0	0
<i>Morphology</i>									
F26c	Suffixing inflexional morphology	0	1	0	1	1	0	0	1
F26d	Prefixing inflexional morphology	0	0	0	0	0	0	0	0
<i>Nominal Categories</i>									
F30	Number of Genders	1	2	1	3	1	1	5	1
F33b	Prefixed nominal plural	0	0	0	0	1	0	0	0
F33c	Suffixed nominal plural	0	1	0	1	1	0	0	1
F37	Definite Articles	1	1	1	5	3	2	1	1
F51	Position of Case Affixes	9	9	7	1	1	9	9	1
F53	Ordinal Numerals	6	7	6	6	6	6	6	6 5
<i>Verbal Categories</i>									
F70	The Morphological Imperative	5	2	4	1	4	5	1	1 2

Sources: Fxx: Wals, GBxx: Grant & Baker (2007), fono: UCLA Phonological Segment Inventory Database