The Creolization of Pidgin: A Connectionist Exploration

Javier Marín Serrano (jms@um.es)
Department of Experimental Psychology, Campus de Espinardo – Univ. of Murcia
30100 Murcia (Spain)
Francisco Calvo Garzón (fjcalvo@um.es)
Department of Philosophy, Campus de Espinardo – Univ. of Murcia
30100 Murcia (Spain)
Javier Valenzuela Manzanares (jvalen@um.es)
Department of English Philology, Campus de la Merced – Univ. of Murcia
30001 Murcia (Spain)

Abstract

According to Derek Bickerton's (1984) Language Bioprogram Hypothesis (LBH), creole genesis, the process by which a Pidgin language develops into a Creole, can only be explained by appealing to Chomskian nativism. *Contra* Bickerton, substratists contend that creole genesis is influenced, crucially, by *substratum* languages. We propose to review the nativist/substratist debate in creolistics under the light of connectionist theory. An analysis of simple recurrent networks exposed to different pidgin-cum-substrate environments may shed light, we suggest, upon the issue of whether a Universal Grammar (UG) is required.

Introduction

Pidgins and *creoles* are "contact languages"—i.e., languages that arise from the need to establish a linguistic contact among different ethnic groups. A pidgin is an auxiliary made-up language that lacks many features of natural languages. The inflections and grammatical morphemes that we find in pidgins across the world, for example, are minimal; complex clauses are not employed; word order can vary drastically; etc., etc. Pidgin languages, by definition, have no native speakers. A creole, on the other hand, is the native language of children grown up in a pidgin environment. Creoles are full-fledged languages that have many typological features in common. Nearly all creoles have, for instance, an unmarked SVO word order; they mark plural animate nouns with what would correspond to the pronoun "they" in English; they use serial verb constructions, where at least two VPs may be concatenated without (morphological) marking; etc.[1]

Most creoles draw their lexicon from European (colonial) languages, known as *superstrates* or *lexifiers*. According to superstratists (Chaudenson, 1995), contact with a lexifier is what explains the formation of creole languages. On the other hand, substratists (Lefebvre, 1998) contend that creole genesis is crucially influenced by *substratum* languages (mutually unintelligible languages spoken by the ancestors of creole speakers). A third route would attemp to explain creole genesis by means of a genetically-driven language faculty (Bickerton, 1984). The following diagram shows the potential routes of influence considered in the literature.

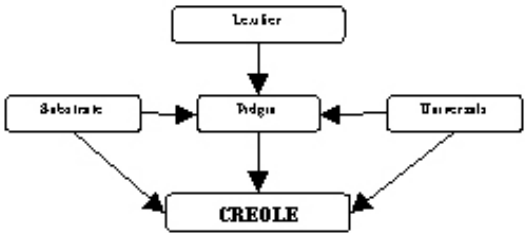

Figure 1: Potential influences in creole genesis (adapted from Cole, 1990)

Derek Bickerton (1984) argues that substratum influence cannot account for creole genesis: "[E]ven if the presence of appropriate languages could be demonstrated, the substratum case would remain incomplete. It would still be necessary to provide plausible mechanisms by which rules could have passed from substratum to creole speakers" (Bickerton, 1984, p. 183). Bickerton combines a modular (Chomskian) nativism and creolistics to argue for a Language Bioprogram Hypothesis (LBH). In his opinion, creole genesis, and the fact that different creoles share many of their features, can only be explained by appealing to a LBH. According to him, we're born with a ready blueprint for grammar—i.e., a bioprogram for language—that can account genetically for those aspects of creolization not explainable in terms of linguistic environmental input. Even though creole speakers are exposed to an

[1] For a detailed study of the main typological features shared among different creoles, see Cole (1990).

impoverished (pidgin) input,[2] a "LBH grammar" allows them to compute all rules required to acquire grammatical competence.

In this work, we shall adopt the substratist hypothesis, according to which creole grammars are transmitted from pre-existing substratum languages, and argue that creole genesis follows quite naturally from connectionist (empiricist) assumptions.[3]

Network, task, and stimuli.

In order to test the substratist hypothesis, we trained three simple recurrent networks (Elman type) on a prediction task. A simple recurrent network (SRN) is a standard feedforward network supplemented with a feedbackward pathway. The recurrent architecture brings into play a short-term memory. The information in state space at any given step of processing is fed back into the hidden layer of the network *along with* the input pattern being fed at the subsequent step of processing. SRNs can in this way process contextualized sequential information. In our simulation, all the SRNs had 31 input and output units, and 60 units in both the hidden and context layers (figure 2 shows the architecture of the SRNs).

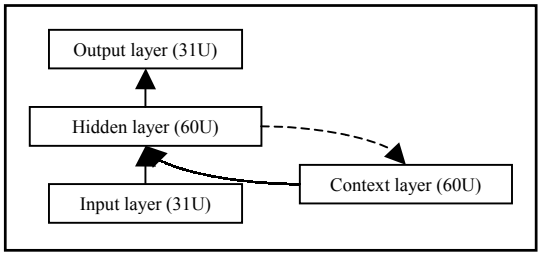

Figure 2: Architecture of SRN used to discriminate grammatically correct sentences (the dashed line represents a copy connection).

Based on Elman (1990), we created three "substratum" toy grammars—Sbs1, Sbs2, and Sbs3. We focused on three creole features that may be present or not in Sbs1, Sbs2, and Sbs3: namely, SVO word order, Plural marking (*Plural*), and Verb Serialization (*VS*)—see table 1.

Table 1: Creole features in substrate languages

	Sbs1	Sbs2	Sbs3
SVO word order	✓	✓	×
Plural marking	✓	×	✓
Verb serialization	×	✓	✓

In our toy grammars, verb serialization indicates that a single verb can play a double role: it can function as a transitive verb and, in its serial, grammaticalized form, as a preposition, preceding certain noun phrases. We introduced the category *VS* in six of the grammatical templates to indicate that the serial verb constructions *(Verb-Tran) + like* + NP*, and *like* + NP + (Verb-Tran)* were present in Sbs2 and Sbs3, respectively. In this way, for example, a template of the form <Noun-Anim Verb-Tran *VS* Noun-Anim> may generate the sentence "Cat chase like* dog", where "*" serves to differentiate the prepositional use of "like" from its transitive-verb form.[4]

Sbs1:	**Word 1**	**Word 2**	**Word 3**	**Word 4**
	Boy	Eat	Cookie	
	Cat	Move		
	Dragon	Break	Plate	
	They	Girl	See	Rock
	Lion	Eat	***They***	Woman
Sbs2	**Word 1**	**Word 2**	**Word 3**	**Word 4**
	Man	Destroy	Glass	
	Mouse	Eat	Bread	
	Monster	Eat	Man	
	Dog	Chase	***Like****	Cat
Sbs3:	**Word 1**	**Word 2**	**Word 3**	**Word 4**
	Girl	Car	See	
	Cat	Mouse	Chase	
	They	Woman	Sandwich	Eat
	Boy	***They***	Girl	Like
	Woman	***Like****	Man	Chase

Figure 3: Possible utterances generated by the artificial toy grammars Sbs1, Sbs2, and Sbs3.

[2] It is somewhat problematic to give a clearcut definition of "pidgin." When Bickerton uses the term in relation to Hawaiian Creole, for instance, he's refering to an unstable pre-pidgin stage which we may think of as a "jargon." Pidgin languages, however, are not completely unstable, and do follow certain norms, albeit less so than natively spoken languages. For present purposes, we shall interpret the term "pidgin" in the latter sense (see discussion, below). Many thanks to Mikael Parkvall for bringing this point to our attention.

[3] For the purposes of this paper, we shall focus on innatist (universal-based) *versus* substratist approaches to creolistics (bottom arrows in figure 1). Although see fn. 8 for some caveats with regard to lexifier influence on creole via pidgin (figure 1, top arrow).

[4] Templates <Noun-Anim Verb-Tran *VS* Noun-Anim> and <Noun-Anim *VS* Noun-Anim Verb-Tran> in Sbs2 and Sbs3, respectively, may thus generate the sentences "Cat like like* dog", and "Cat like* dog like".

We constructed three sets of two-, three-, and four-word grammatical sentences (figure 3). We then created a grammatically inconsistent pidgin corpus by mixing up a 33% of each substratum. In order to impoverish the pidgin input signal, we removed all templates that contained serial verb constructions.

We then ran two sets of simulations.[5] Lexical items of the aforementioned combined lexicon were randomly assigned a thirty-one bit (localist) vector.[6] The input set consisted of the successive concatenation of all the sentences in the pool of data formed out of the stream of these vectors. The networks' task was to make correct predictions of subsequent words in the corpus of sentences. Being fed with a sequence of words from the input stream, the network had to predict the subsequent word. Using backpropagation, weights were adjusted to the desired output performance.

In an initial phase of creolization (**CR1**), we trained three SRNs on a corpus where pidgin sentences constituted 75% of the environment, and the remaining 25% was composed by sentences generated by the Sbs1-, Sbs2-, and Sbs3-templates, respectively. The SRNs were fed with an input stream of 10,000 sentences by concatenating the corresponding 31-bit localist word vectors. The networks were trained for six epochs to predict word order in the substrate-cum-pidgin input stream.

In a second phase of creolization (**CR2**), a 'creole' corpus (see discussion, below) was created by following the algorithm shown in table 2. We then exposed all three networks to an environment where this corpus constituted 70% of the sentences, and pidgin sentences made up the remaining 30%. The networks were trained for six more epochs. Probabilities of occurrence for all possibly correct predictions were determined by generating the likelihood vectors for every word in the corpus.

Table 2. Creole production algorithm

1. *Select noun randomly for Word-1 position*
2. *Feed network with selected word*
3. *Generate output response*
4. *Select next word, probabilistically, based on output activations*
5. *Go to step 2*

[5] The SRNs were simulated with PDP++ (O'Reilly, Dawson, and McClelland), and trained with a learning rate of 0.1.

[6] "like" and "like*" were assigned the same coding vector.

We tested performance on all three networks for phases CR1 and CR2. The error measures of probability-based predictions against the likelihood vectors are shown below:

Table 3: Error measures of probability-based predictions against the likelihood vectors.

	*CR1:*mean-error (SD)	*CR2:*mean-error (SD)
SRN-1	**.18 (.28)**	**.12 (.17)**
SRN-2	**.13 (.20)**	**.12 (.17)**
SRN-3	**.17 (.26)**	**.11 (.16)**

Discussion

In the simulations reported here, the networks exhibit appropriate sensitivity to the syntactical dependencies found in the grammatical structures of the limited number of sentences of our toy languages. To study the networks' grammatical competence, we performed cluster analyses on the trained SRNs (after 6 and 12 epochs) by recording hidden activations in response to a theoretical corpus containing all features in table 1 (i.e., SVO word order, plural marking, and verb serialization).

The networks create hidden (abstract) representations that capture the syntactical dependencies that exist in the pools of data, reducing thus their overall performance error. Creole genesis, we believe, can be approached *statistically* and studied in an incremental manner by looking at the hidden partitions generated in phases CR1 and CR2.

In CR1, training spaces are heterogeneous, grammatically speaking, since they are formed by an inconsistent pidgin corpus that has been supplemented with sentences that belong to the respective substrates. In this way, hidden clusterings reflect the dominant, most frequent, grammatical subregularities of the *combined* corpora. That would constitute an initial phase of 'substratum-based' creolization. This initial phase may be illustrated by paying attention to the behaviour of the lexical items "like" and "like*":[7]

In CR1, SRNs#2 and #3 build up two different hidden representations of this input vector, reflecting the two (functional) predictive roles it can play (cf. table 1). *It is their respective substratum-languages* what permits this divergence to take place. Since Sbs1 lacks verb

[7] Recall that "like" as a transitive verb and "like*" as a serial verb construction are coded with the same input vector.

serialization, this functional differentation does not take place in CR1Sbs1 (figure 4).

We can further focus on verb serialization to illustrate the process of enrichment that takes place in the emergence of a full-fledged creole. In CR2, the SRNs are exposed to a single corpus, a significant part of which is composed by (CR1) creole productions (70%). This means that all three networks will become competent with verb serialization since this construction is present in their shared environment. Thus, "like" and "like*" are differentiated in *all* CR2 clusters (see figure 4).

Were we to feed networks #1, #2, and #3 with shared creole productions once again in successive phases of creolization (CR3, CR4, …, CRn), weight spaces would accordingly reflect an increasing amount of common co-relational information. In this way, a linguistic environment that initially consisted exclusively of substratum and pidgin utterances, becomes dynamically replaced by the *lingua franca* of the "community"; that is, the "emergent creole". Having an environment where enriched creole productions with an increasingly internal consistency are more frequent, all networks will eventually induce this statistical tendency towards a convergent full-fledged creole.

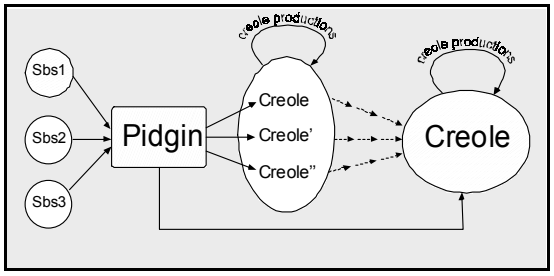

Figure. 5: Dynamic enrichment of environment and trend towards Creole convergence

Conclusion

Chomskian nativism has traditionally found support in Bickerton's LBH approach to creolistics. A Universal Grammar is seen as the only element that can explain creole genesis. Bickerton exploits a version of the "poverty of stimulus" argument and argues that a creole cannot arise from mere exposition to a pidgin. He further claims that substratum influence cannot account for creole genesis either: "*even if the presence of appropriate languages could be demonstrated, the substratum case would remain incomplete. It would still be necessary to provide plausible mechanisms by which rules could have passed from substratum to creole speakers*" (Bickerton, 1984, p. 183).

Contra Bickerton, the friend of substratism contends that the process of creolization is crucially influenced by substratum-languages.[8] Creole languages are not *ab ovo* creations. In this work, we have tried to show that the substratist, anti-Bickertonian, position can be backed up empirically by connectionist theory. Connectionist theory, we contend, furnishes us with a (statistical) alternative to Bickerton's required (rule-governed) mechanism. The process by which a pidgin develops into a creole can be modelled by an SRN exposed to a dynamic (substratum-based) environment. In this way, empiricism suffices itself to account for creole grammar as a by-product of general-purpose mechanisms: the ball is now on the nativist's quarters.

Acknowledgments

This work was supported in part by a *Ramón y Cajal* research contract to the second author (Spanish Ministry of Science and Technology), and by research grant PL/3FF/00 from *Fundación Séneca* to all three authors. A version of this material was presented at the 8th Annual Conference on Architectures and Mechanisms for Language Processing (AMLaP-2002) in Tenerife, Spain.

References

Bickerton, D. (1984). The Language Bioprogram Hypothesis. *The Behavioral and Brain Sciences*, *7*, 173-221.

[8] As mentioned earlier (fn. 3 above), for the purposes of this study we focused on a substratist line of response to Bickerton's LBH. The careful reader, however, will have noticed that we have omitted a critical part of the story: namely, the adoption of a lexicon that usually takes place in pidgins/creoles by borrowing it from a *lexifier* language. We coded lexical items that belong to substrates Sbs1, Sbs2, and Sbs3, and lexical items that belong to the pidgin and the creole corpora with the same vectors. To be accurate, we would have had to augment the input representational space in order to code differently (in a localist fashion) the several different lexicons, and then to pre-train the networks to master the existing lexical correspondences. For reasons of computational economy, we omitted this initial stage, and focused directly on the substratist hypothesis, since our interest in the simulations was on the acquisition of grammatical competence, rather than on the lexical influence that other languages may exhibit—see Calvo Garzón et al. (submitted) for further ellaboration on this issue.

Calvin, W.H. & Bickerton, D. (2000). *Lingua ex Machina: Reconciling Darwin and Chomsky with the Human Brain*. London: MIT Press.

Calvo Garzón, F., Marín Serrano, J. & Valenzuela Manzanares, J. (submitted). Nativism, Empiricism and Neural Networks: The Case for Substratism in Creole Genesis.

Chaundeson, R. (1995). *Les Créoles*. Presses Universitaires de France: PUF

Cole, J. (1990). A Look at Some Aspects of Creole Genesis. *Cognitive Science Research Reports 169*, University of Sussex.

Elman, J. (1990). Finding Structure in Time. *Cognitive Science*, *14*, 179-211.

Lefebvre, C. (1998). *Creole Genesis and the Acquisition of Grammar: The Case of Haitian Creole*. Cambridge, England: Cambridge University Press.

O'Reilly, R. & Munakata, Y. (2000). *Computational Explorations in Cognitive Neuroscience: Understanding the Mind by Simulating the Brain*, Cambridge, Mass.: MIT Press.

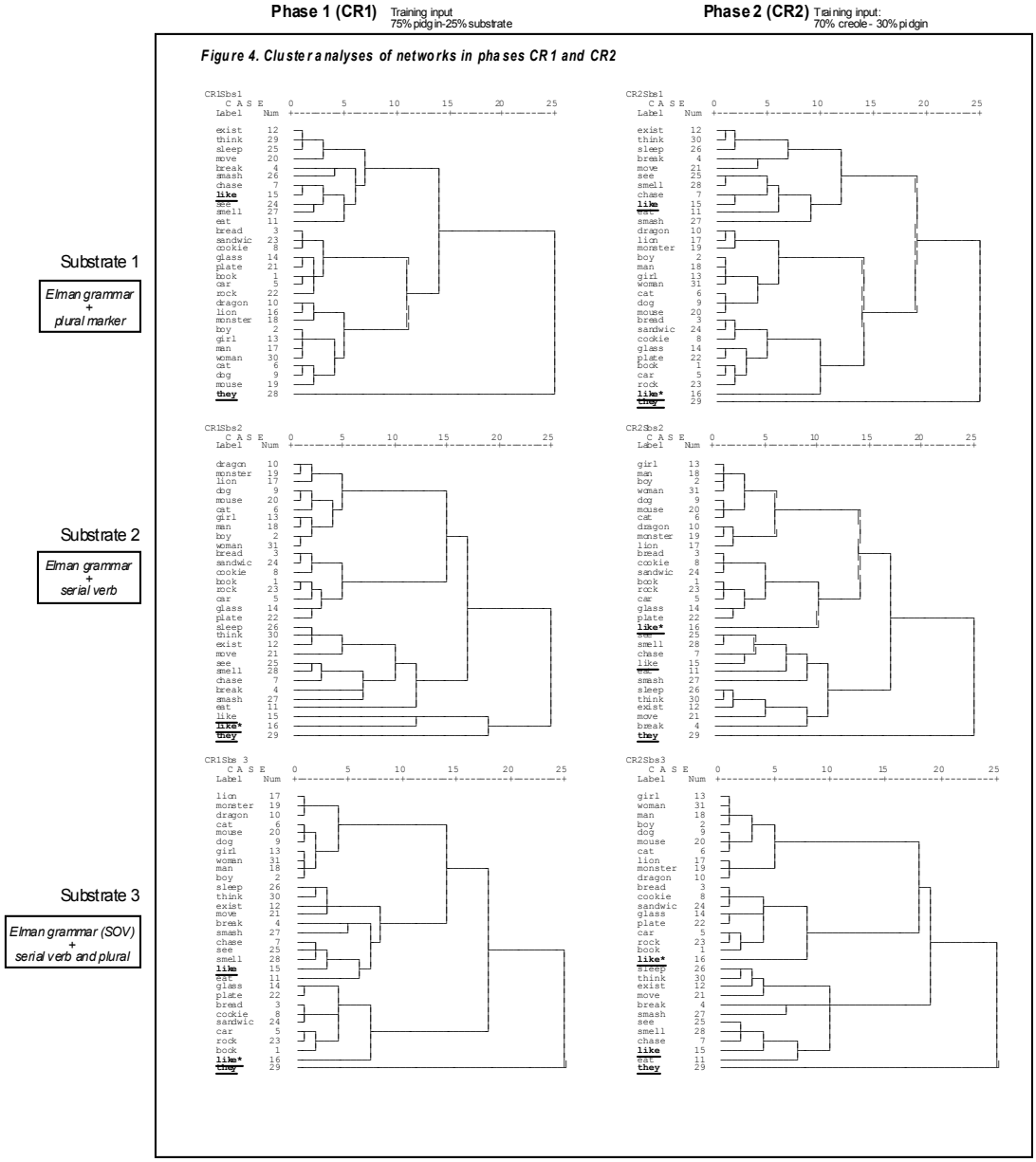

Figure 4: Hierarchical clusterings of hidden unit activations from the prediction task (phases CR1 and CR2).