A COMPUTATIONAL SIMULATION OF THE GENESIS AND SPREAD OF LEXICAL ITEMS IN SITUATIONS OF ABRUPT LANGUAGE CONTACT

Ching-Yat Cheung [*1,2], Kofi Yakpo [1], Christophe Coupé [1,3]

[*]Corresponding Author: chingyat@connect.hku.hk
[1]Department of Linguistics, The University of Hong Kong, Hong Kong SAR
[2] LEADS Group, Max Planck Institute for Psycholinguistics, The Netherlands
[3] Laboratoire Dynamique Du Langage, Centre National de la Recherche Scientifique, France

The current study presents an agent-based model which simulates the innovation and competition among lexical items in cases of language contact. It is inspired by relatively recent historical cases in which the linguistic ecology and sociohistorical context are highly complex. Pidgin and creole genesis offers an opportunity to obtain linguistic facts, social dynamics, and historical demography in a highly segregated society. This provides a solid ground for researching the interaction of populations with different pre-existing language systems, and how different factors contribute to the genesis of the lexicon of a newly generated mixed language. We take into consideration the population dynamics and structures, as well as a distribution of word frequencies related to language use, in order to study how social factors may affect the developmental trajectory of languages. Focusing on the case of Sranan in Suriname, our study shows that it is possible to account for the composition of its core lexicon in relation to different social groups, contact patterns, and large population movements.

1. Introduction

A consequence of colonialism and population movements is the genesis of a multilingual and sociologically complex ecology, in which people of diverse language backgrounds are brought into contact, creating the need for a new shared means for communication. In some of these cases, a pidgin or a creole, which is characterized by the mixture of lexical and grammatical features from different languages in the language ecology, emerges to answer communicative needs. Such recent emergence of languages may shed light on the overall evolutionary process of human languages (Mufwene, 2008). Specifically, population movements have been argued as one of the factors that motivate the evolution of contact languages (Mufwene, 2007).

The present study presents a computational model which serves as a complement to empirical studies when investigating how population movements and social structures affect the emergence of a language and specifically the developmental trajectory of its lexical inventory. The paper expands on previous studies by considering social structures, population movements, communicative needs, and the functional frequency of the lexical items. Our model is specifically inspired by the historical case of Suriname, a former colony in South America that was ruled by English and Dutch speakers. The Dutch took over the colony from the English, which led to an exodus of English native speakers and influenced the development of Sranan, an English-lexified creole spoken in the area (Arends, 2017). Like other contact languages, it is characterized by a mixture of various languages spoken in the language ecology. We thus constructed a model to investigate how the sudden departure of a dominant social group may have an influence on the collective convergence towards a linguistic form when there is a language competition. Essentially, the agent-based model allows us to experiment with several sociolinguistic variables and observe different population movements can affect language evolution. As reported below, the model demonstrates the possibility for agents to develop a multilingual - rather than monolingual - lexical inventory. The predictions also align quantitatively with the lexical data of Sranan.

2. Previous Studies

There have been recent attempts to construct computational models to better understand contact languages. Jansson et al. (2015), inspired by the Naming Game (Baronchelli et al., 2006), simulated the evolution of the lexicon, phonology and syntax of Mauritian Creole using data on the demography and population movements in Mauritius in the 18th century. Specifically for the lexicon, their model showed how agents can converge to one single lexifier (i.e., French) solely based on communicative needs. Tria et al., (2015) also employed a computational model to study the possible emergence of creoles. Their model – a modified version of the Naming Game with non-trivial interaction rules based on the observation of communicative pressures in highly segregated societies – revealed that such an emergence can be accurately predicted on the sole basis of historical demographic data, more precisely the proportion of different social/ethnic groups in the multilingual ecology. Additionally, Furman & Nitschke (2020) investigated the role of external factors which affect the evolution of creole languages, such as the population size of different interacting groups and the lexical similarity between these groups, in the convergence of lexical items with an iterative agent-based naming game.

Recently, Cheung (2022) further showed that computational social networks with different density and connectivity patterns within the same social group and

across different social groups – which can differ due to their economic structures (such as 'sparse' cotton plantations and dense urban areas/ports) – could affect the likelihood and rate of creole emergence. These results, which are specific to scenarios of recent abrupt language contact, can be related to more generic studies that attempt to model language competition, change and death in a multilingual society (e.g. Abrams & Strogatz, 2003; Minett & Wang, 2008), especially when different social structures (Ke et al., 2008; Loureiro-Porto & Miguel, 2017) are found to be relevant.

The previous attempts have all inspired the present study.

3. Modelling the lexical propagation of historical creoles

Our model is inspired by the historical case of Suriname, a former colony in South America. Historical data (Migge, 1998, based on Postma, 1990 and Voorhoeve & Lichtveld, 1975) show the demographic evolution of each social group/ethnicity of the region through time. There was a significant drop of the English-speaking population from 1668 due to the colony being ceded by the British Empire to the Netherlands under the Treaty of Breda in 1667 (Arends, 2017). There were approximately 2000 English speakers in 1666, but the number dropped to 820 in 1668, and there were only 38 still remained in 1680 (Voorhoeve & Lichtveld, 1975). The number of other non-English Europeans, mainly Dutch, increased slightly, while the number of Blacks also increased. The exodus of English native speakers reportedly influenced the development of Sranan, an English-lexified creole spoken in the area (Arends, 2017).

Our computational model employs an agent-based naming game that simulates repeated interactions between agents. Each agent initially has their own lexicon of 300 basic words in their own language. Each lexical referent is assigned with a probability drawn from the Zipfian distribution, which indicates the chance of being picked out and used during an interaction, such that we can partially account for language use.

Agents can be assigned with one of three fixed roles: Blacks, European group 1 (E1) (the English) or European group 2 (E2) (the Dutch). The population of Blacks is further divided into five groups of rural plantation workers (R1-R5) and an urban group (U). Members of each rural group are sparsely distributed and related according to a scale-free social network structure (the average degree of each node is ~1.93), while each of the European and Black Urban groups is a dense small-world network (rewiring probability of 0.05 with an average degree of 8). Random links across the different groups are added in proportion to the total population of these groups. This hypothetical network setup is based on other studies of models and empirical cases (c.f. Cheung, 2022).

Given the situation, an agent may know up to four words for a given referent: one in English (E1), one in Dutch (E2), one in an African language (A), and one in an emerging creole (C). We modify the interaction rules presented in Tria et al. (2015), where all the agents could interact with each other. In our model, for each interaction, an existing edge between two agents is randomly picked and these agents play a naming game with a random assignation of the roles of speaker and listener. If the communication is successful (i.e., the listener knows the speaker's word , in the language they chose, for the chosen referent), any other word – in another language – for the referent is discarded. If the communication fails, the listener picks up the word from the speaker. While this widely used paradigm may not be the most accurate reflection of the underlying cognitive mechanisms of word learning, it offers a plausible model for the switch of belief between different lexical representations for a particular referent in order to choose which one is most suitable to foster communication success.

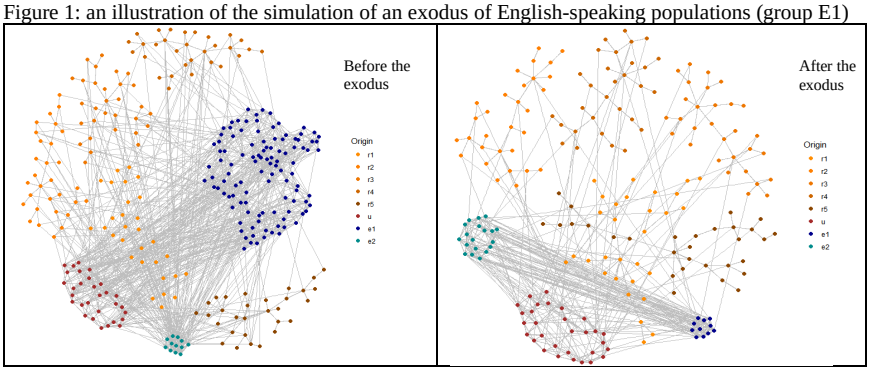
The interaction is governed by several constraints reflecting the segregated social dynamics. Even if both speaker and listener have a word in an African language for a particular referent, they will only be able to communicate successfully with a probability δ. This parameter reflects the actual diversity of African languages spoken by Blacks. The Europeans (*E1*, *E2*) only have a probability β to learn a word in an African language successfully (this includes European-to-European communication), and they further only have a probability ε to learn from the Blacks (*R/U*), due to differences in socioeconomic status. The chance of creole innovation, if the listener possesses both a European and an African language for a lexical referent, is given by γ, and the chance of a listener adopting this creole is given by α, an adjustment made to represent the failure of learning due to potential difficulties in interpreting the mixing and the preference to adapt items from the superstratum. Given the case of Suriname and our focus on the effect of population movements, we assumed a reasonable fixed value for each of the parameters across all conditions: $\alpha = 0.8$, $\beta = 0.2$, $\delta = 0.25$, $\varepsilon = 0.25$ and $\gamma = 0.8$ (the parameters could of course be modified to explore significantly different language ecologies)

In order to investigate the consequences of the sudden exit of the English speakers, a large proportion (over 60%) of the members of group 1 (E1) is removed after a number of interactions/steps – with two possible departure times (Conditions 1 and 2). Figure 1 illustrates how such an exodus transforms the overall social network. Furthermore, we consider the condition in which the graph remains unchanged throughout the whole simulation as a baseline (Condition 3). We finally model a condition (Condition 4), a setup based on condition 1, in which both groups of Europeans are removed after a number of interactions to simulate the development of a creole among the Blacks themselves – a situation typical of creole development. Table 1 illustrates the 4 conditions of the simulations.

Table 1: the different conditions that were used in the simulations

1	E1 departs E2 increases						End
2			E1 departs E2 increases				End
3							End
4	E1 departs E2 increases				Both E1 & E2 depart		End
Steps	100,000	...	1,000,000	.....	2,000,000	.....	5,000,000

Figure 1: an illustration of the simulation of an exodus of English-speaking populations (group E1)



3. Results

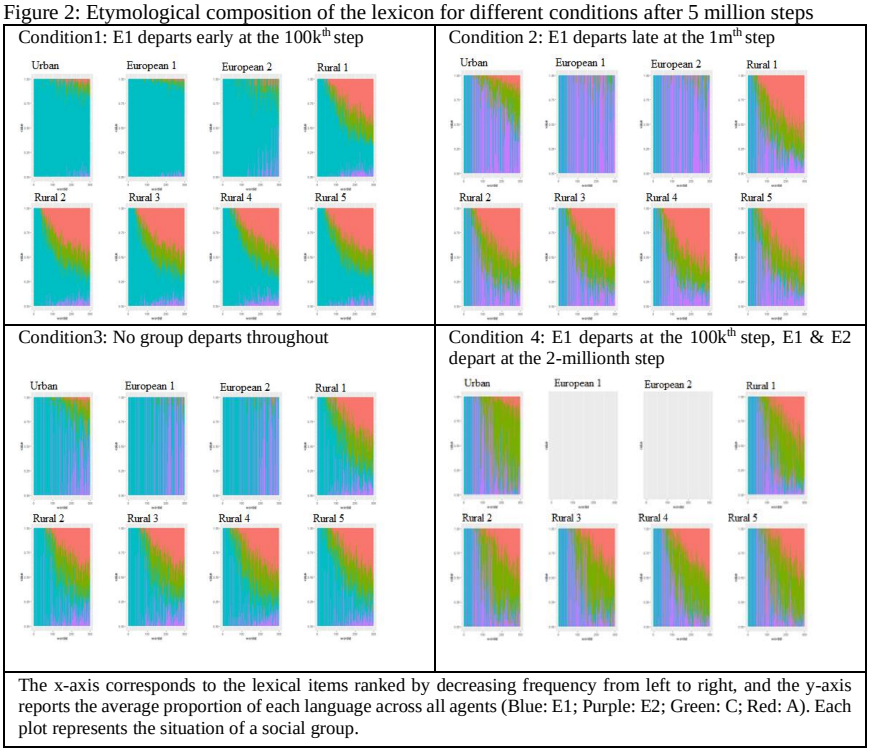
The resulting lexicons are assessed for the different social groups. We find that the departure of the native speakers of European languages affects the final stage of the process of linguistic convergence. The speakers remain under communicative pressure due to multilingualism after the exit of the speakers of the superstrate languages, and they still need to develop a lingua franca, causing more creole innovation and language diversification. This is in alignment with previous observations in the literature (Mufwene, 1996, 2007).

Figure 2 shows the results after 5 million steps of interaction in our different conditions. Across all conditions, the agents are under communicative pressure and must converge to communicate with each other. An observation that can be made from the simulation under Condition 3 is that if the native speakers of the main lexifier, English, remain throughout the simulation, all agents tend to converge to this single language, due to sustained linguistic input and sufficient exposure. Furthermore, in condition 2, in which English speakers depart relatively late (at the 1-millionth step), it can be observed that although a majority of the words still stem from their language, some mid-frequency words are adopted from

the other European language, Dutch. For both conditions 2 and 3, no significant amount of creole innovation occurs (i.e., no convergence to creole words). In condition 1, the native speakers of the main language exit early, and while their language still occupies a large proportion of the final lexicon, Dutch is also used to denote some words that are of higher frequency.

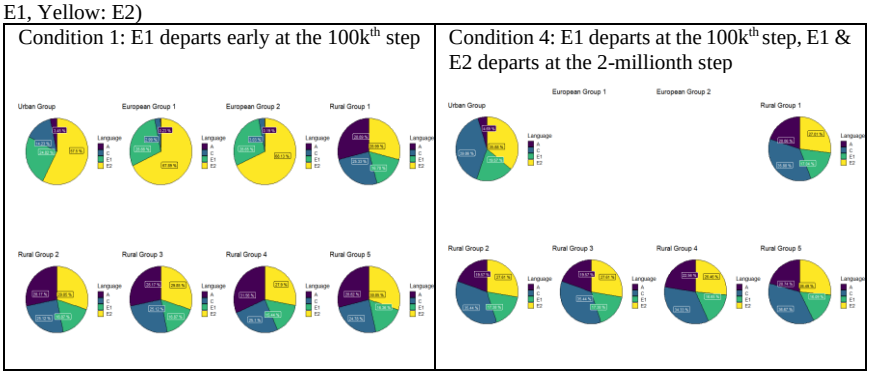
In figure 3, a comparison of Conditions 1 and 4 shows that creole words are more prominent in the latter condition. Condition 4, in particular, appears to be the closest approximation of the etymological distribution of Sranan, in which it is estimated that there are roughly 18% English words, 21.5% Dutch words, and 36% innovations (Romaine, 2001).

Figure 2: Etymological composition of the lexicon for different conditions after 5 million steps



The x-axis corresponds to the lexical items ranked by decreasing frequency from left to right, and the y-axis reports the average proportion of each language across all agents (Blue: E1; Purple: E2; Green: C; Red: A). Each plot represents the situation of a social group.

Figure 3: cumulative composition of the lexicon across different groups (Purple: A, Blue: C, Green: E1, Yellow: E2)



In addition, due to the high segregation of the society, there is a delay in the rural area in terms of linguistic exposure. This suggests a possible variation continuum between the urban and the rural areas, which is also observed in Suriname (Arends, 2017). The study of the interaction between language variety and social stratification could be an interesting direction for further exploration.

4. Conclusion

The present model offers a plausible way to computationally investigate socio-historically complicated and empirically specific scenarios of language contact. It offers reasonable predictions for a given social dynamics and segregated relationships between different social groups, and the group selection of lexical items appears to be highly sensitive to the interactions and language use between agents.

The current model could be extended to other (socio)linguistic phenomena, such as the social dynamics during second language learning. The observation of the different intermediate stages of language development could also explain how a variation continuum could appear between various social groups. Meanwhile, the model mostly focuses on horizontal transmission of language, and the role of vertical transmission is underplayed. It would be worthwhile to model the birth of children with initially blank-slate agents (in terms of linguistic repertoire and exposure), and observe how this affects the linguistic dynamic.

References

Abrams, D. M., & Strogatz, S. H. (2003). Modelling the dynamics of language death. *Nature*, *424*(6951), 900–900. https://doi.org/10.1038/424900a

Arends, J. (2017). *Language and Slavery: A social and linguistic history of the Suriname creoles* (Vol. 52). John Benjamins Publishing Company. https://doi.org/10.1075/cll.52

Baronchelli, A., Felici, M., Caglioti, E., Loreto, V., & Steels, L. (2006). Sharp transition towards shared vocabularies in multi-agent systems. *Journal of Statistical Mechanics: Theory and Experiment*, *2006*(06), P06014–P06014. https://doi.org/10.1088/1742-5468/2006/06/P06014

Cheung, C. Y. (2022). *Investigating contact languages with agent-based modelling*. The University of Hong Kong.

Furman, G., & Nitschke, G. (2020). Evolving an Artificial Creole. *Proceedings of The Genetic and Evolutionary Computation Conference 2020 (GECCO '20).*, 2.

Jansson, F., Parkvall, M., & Strimling, P. (2015). Modeling the Evolution of Creoles. *Language Dynamics and Change*, *5*(1), 1–51. https://doi.org/10.1163/22105832-00501005

Ke, J., Gong, T., & Wang, W. S.-Y. (2008). Language Change and Social Networks. *Commun. Comput. Phys.*, 15.

Loureiro-Porto, L., & Miguel, M. S. (2017). Language Choice in a Multilingual Society: A View from Complexity Science. In S. S. Mufwene, C. Coupe, & F. Pellegrino (Eds.), *Complexity in Language* (pp. 187–217). Cambridge University Press. https://doi.org/10.1017/9781107294264.008

Migge, B. (1998). Substrate influence in creole formation: The origin of give-type serial verb constructions in the Surinamese Plantation Creole. *Journal of Pidgin and Creole Languages*, *13*(2), 215–266. https://doi.org/10.1075/jpcl.13.2.02mig

Minett, J. W., & Wang, W. S.-Y. (2008). Modelling endangered languages: The effects of bilingualism and social structure. *Lingua*, *118*(1), 19–45. https://doi.org/10.1016/j.lingua.2007.04.001

Mufwene, S. S. (1996). The Founder Principle in Creole Genesis. *Diachronica*, *13*(1), 83–134. https://doi.org/10.1075/dia.13.1.05muf

Mufwene, S. S. (2007). Population Movements and Contacts in Language Evolution. *Journal of Language Contact*, *1*(1), 63–92. https://doi.org/10.1163/000000007792548332

Mufwene, S. S. (2008). *Language evolution: Contact, competition and change*. Continuum.

Postma, J. (1990). *The Dutch in the Atlantic Slave Trade, 1600–1815* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511528958

Romaine, S. (2001). Lexical structure in Pidgins and Creoles. In D. A. Cruse, F. Hundsnurscher, M. Job, & P. R. Lutzeier (Eds.), *Lexikologie* (pp. 1092–1095). De Gruyter. https://doi.org/10.1515/9783110171471.2.26.1092

Tria, F., Servedio, V. D. P., Mufwene, S. S., & Loreto, V. (2015). Modeling the Emergence of Contact Languages. *PLOS ONE*, *10*(4), e0120771. https://doi.org/10.1371/journal.pone.0120771

Voorhoeve, J., & Lichtveld, U. M. (Eds.). (1975). *Creole drum: An anthology of Creole literature in Surinam*. Yale University Press.