► Multilingual Societies from Twitter Data: Empirical Analysis and Theoretical Modelling

David Sanchez,
IFISC (UIB-CSIC), Spain, david.sanchez@uib.es
Thomas Louf,
IFISC (UIB-CSIC), Spain, thomaslouf@ifisc.uib-csic.es
Jose J. Ramasco,
IFISC (UIB-CSIC), Spain, jramasco@ifisc.uib-csic.es

Purpose: Cultural diversity encoded within the languages of the world is at risk, as many languages have become endangered in the last decades in the context of growing globalization. To preserve this diversity, it is first necessary to understand what drives language extinction, and which mechanisms might enable coexistence.

Design/methodology/approach: Here, we discuss the processes underlying language shift through a conjunction of theoretical and empirical perspectives. We quantify the linguistic diversity with the Earth-mover's distance (EMD), either at full county or region scale. This metric allows us to measure the discrepancy between two distributions embedded in a two-dimensional space and is shown to be a proper distance. To understand how different linguistic states can emerge and, especially, become stable, we propose a theoretical model in which coexistence of languages may be reached when learning the other language is facilitated, and when bilinguals favor the use of the endangered language. Then, we carry out simulations in a metapopulation framework.

Findings: A large-scale study of spatial patterns of languages in multilingual societies using Twitter and census data yields wide diversity. This ranges from an almost complete mixing of language speakers, including multilinguals, to segregation, with a neat separation of the linguistic domains and with multilinguals mainly at their boundaries. We calculate the EMD for both monolingual and multilingual groups. In the former, we find a large segregation for Switzerland and Belgium. In the latter, the segregation is maximal for Java and Estonia. The theoretical model is first analyzed in a single population in mean-field, which uncovers interesting stable states of extinction and coexistence, including with bilinguals alone sustaining a minority language. This stability is discussed with flow diagrams that take into account the language prestige and the bilingual preference at a fixed mortality rate. The metapopulation model highlights the importance of spatial interactions arising from population mobility to explain the stability of a mixed state or the presence of a boundary between two linguistic regions.

Research limitations/implications: Interestingly, the achieved state depends on the ratio between mortality rate and learning rate. When the ratio is low (high), the preferred state is spatial mixing (extinction/dominance or spatial separation). We also

investigate the dynamics of our model. We find that the evolution of the system once it undergoes a transition is highly history-dependent. It is easy to change the status quo, but going back to a previous state may not be simple or even possible.

Practical implications: We have shown that, quite counter-intuitively, increasing the ease to learn the other language may break the existing boundary and lead to extinction, and not to the desired coexistence with mixing of the languages. This calls for caution when designing policies since the final state is strongly history-dependent.

Originality/Value: Overall, our findings shed light on the role of heterogeneous speech communities in multilingual societies, and they may help shape the objectives and nature of language planning in many countries where accelerated changes are threatening cultural diversity.

Keywords: computational sociolinguistics, language dynamics, bilingualism, social media.

Research type: Research paper; see Louf, T., Sanchez D., & Ramasco, J.J. (2021). Capturing the diversity of multilingual societies. *Physical Review Research*, *3*, 043146.