Data and Model Analysis of Socioeconomic Influence on Linguistic Variation

David Sánchez (IFISC, UIB-CSIC)
In collaboration with: Thomas Louf, José J. Ramasco, and Márton Karsai

The socioeconomic background of people and how they use standard forms of language are not independent, as demonstrated in pioneering sociolinguistic studies [1,2]. However, the extent to which these correlations may be influenced by the mixing of people from different socioeconomic classes remains relatively unexplored from a quantitative perspective. Here, we leverage geotagged tweets and transferable computational methods to map deviations from standard English on a large scale. We focus on the standard variety because this is defined by norms set by the language ideology of a society's major institutions and as such is attributed with a market value [3].

We perform our analysis in seven thousand administrative areas of England and Wales [4]. We combine these data with high-resolution income maps to assign a proxy socioeconomic indicator to home-located users. Strikingly, across eight metropolitan areas we find a consistent pattern suggesting that the more different socioeconomic classes mix, the less interdependent the frequency of their departures from standard grammar and their income become. We base our observations on mobility mixing matrices stratified by socioeconomic classes (assortativity). Further, we propose an agent-based model of linguistic variety adoption that sheds light on the mechanisms that produce the observations seen in the data. We assess the fixed points of the model and their stability, showing when the different varieties dominate or coexist. Importantly, our model not only takes into account the status of a linguistic variant that determines the rates of change but also the speakers' preference to adopt standard or. nonstandard forms.

Overall, the work provides a solid foundation for future works of the same vein. It could first be extended to other countries where similar data could be obtained in sufficient amounts. Also, our observations were made in the social context of Twitter, but individuals may choose to use a different language in other environments. Observing the language production of individuals in different social contexts on a scale such as the one discussed here poses a great challenge, but it would definitely help further modelling endeavour and thus greatly contribute to our understanding of these linguistic phenomena.

[1] W. Labov, *The Social Stratification of English in New York City* (Cambridge University Press, Cambridge, UK, 1966).
[2] P. Trudgill, *The Social Differentiation of English in Norwich*, no. 13 in Cambridge Studies in Linguistics (Cambridge University Press, Cambridge, UK, 1974).
[3] P. Bourdieu, Language and Symbolic Power (Polity Press, Cambridge, UK, 2009).
[4] T. Louf, J. J. Ramasco, D. Sánchez, and M. Karsai, arXiv:2307.10016 (preprint).