$9^{th}$ International Conference on Computational Social Science $IC^2S^2$
July 17-20, 2023, Copenhagen, Denmark

# Not sticking by the rules: *How the use of standard language varies with socio-economic status and social mixing*

*Keywords: socio-linguistics, social media, standard language, socio-economic status, assortativity*

## Extended Abstract

Language bounds individuals together just as much as it divides them. Upon reading the last sentence, one may first think of inter-language boundaries: if two individuals cannot understand each other's language, communication is effectively very limited. In this work, however, we delve into intra-language differences: when people speak the same language, but differently, which is also a strong source of social divides. Language variation in a population is driven by many factors, among which the socio-economic background of individuals is essential. As the sociologist Pierre Bourdieu put it, individuals possess different quantities of "linguistic capital" [1], which gives them more or less power both in the political and economic spheres. Socio-economic status (SES) and this linguistic variation are thus mutually sustaining: one inherits a SES along which comes a linguistic capital that contributes — among other things — to constrain them to their status of origin. Understanding the mechanisms that lead to this linguistic segregation is therefore of great importance.

A first step toward understanding is to quantify the phenomenon. It has already been measured by the PISA reports of the OECD [2], which survey, among others, the reading proficiency of 15-year-olds in 79 countries. They consistently show that students with lower socio-economic background tend to perform worse at this test. While these confirm there is an issue to tackle, they are not extensive enough. They do not test language production, and not the whole population, but only a sample of students of a specific age. Alternative empirical works are thus needed. In that vein, some past work has investigated the dependence on SES of the frequency of a few markers of non-standard language in France, as seen on Twitter [3], showing the potential of this data source for such an analysis.

In our work, we investigate this inter-dependence in the UK, by way of measuring how Twitter users abide by the rules of the standard variety of their language, which is the one taught at schools. To do so, we analyse the Tweets of identified UK residents with LanguageTool, an open-source grammar, style and spell checker. As the tool is implemented in 15 languages, our study can actually be replicated in other countries as well. It enables us to compute the frequency at which these users make different categories of mistakes, according to standard rules. To assign a SES to these users, we determine their area of residence from the geotags attached to their Tweets. The areas considered are administrative areas with a typical population of ten thousand. For those, we have the average net income of their residents from the census, which gives us a proxy for the SES of our Twitter users.

Focusing on grammar mistakes, we then find a significant, but weak correlation (Pearson r equal to -0.25) between the net income and the average frequency of mistakes in all areas of the country. We then focus on metropolitan areas because we expect them to host more linguistic and socio-economic heterogeneity than their rural counterparts. We also happen to have more

data in these on both mobility and language production, due to the fact that Twitter users tend to be more urban. We therefore consider the 8 largest metropolitan areas in England, and find large differences between these areas, with correlations ranging from -0.07 in Sheffield to -0.5 in Bristol. To find out what could make these cities so different in that regard, we measure the assortativity in the mobility patterns of their residents [4]. We thus determine how likely people from different socio-economic classes are to interact with each other. What we find is a very clear correlation between this assortativity measured at the city level, and the correlation between SES and the frequency of grammar mistakes, as shown in Figure 1(a). This indicates that the more mixing in the population, the less the frequency of mistakes is determined by the SES of origin.

Having identified the importance of social mixing, we wish to understand the mechanisms behind this effect with a simple model. Figure 1(b) provides a summary of the model. It considers agent who can have one of two SES classes (triangles and circles in the figure), living in two separate cells, and who can either speak standard, or not. The standard form has a higher prestige $l_v$ because it is for instance preferred in schools and in the media. Then, each SES class has a preference for one form, the lower class 1 is attached to the non-standard form with a factor $q_1$, and inversely the higher class 2 is attached to the standard one with $q_2$. When either of these three parameters has a value above 0.5, it means there is a preference for the respective form. Then for instance, when an agent of low SES speaking non-standard interacts with another agent speaking standard, they have a probability $l_v(1-q_1)$ to start using the standard form as well. The agents can move from their residence cell with a probability $m_{i,j}$ at each step, which thus controls the mixing of the two populations. We ran simulations with 200 agents spread evenly between two SES classes, with each class in its own residence cell. We set the standard form to be the more prestigious one, agents with higher SES to be indifferent and ones with lower SES to be more attached to the non-standard form. In Figure 1(c), we show the result of this simulation when we set a very low inter-cell mobility. For Figure 1(d), the mobility was greatly increased. We can see that the more the two populations mix, the closer they are to using the two forms in the same proportions, reflecting what we observed in the data. This promising result calls for further investigation of the model, and particularly how it would fare when initialised with the actual data that we have in the metropolitan areas of England.

# References

[1] Pierre Bourdieu. *Language and Symbolic Power*. Ed. by John B. Thompson. Trans. by Gino Raymond and Matthew Adamson. Cambridge: Polity Press, 2009. 302 pp.

[2] OECD. *Where All Students Can Succeed*. Vol. II. PISA 2018 Results. Paris: Organisation for Economic Co-operation and Development, 2019.

[3] Jacob Levy Abitbol et al. "Socioeconomic Dependencies of Linguistic Patterns in Twitter: A Multivariate Analysis". In: *The Web Conference 2018 - Proceedings of the World Wide Web Conference, WWW 2018*. 2018, pp. 1125–1134. DOI: 10.1145/3178876.3186011.

[4] Rafiazka Millanida Hilman, Gerardo Iñiguez, and Márton Karsai. "Socioeconomic Biases in Urban Mixing Patterns of US Metropolitan Areas". In: *EPJ Data Science* 11.1 (1 Dec. 2022), pp. 1–18. DOI: 10.1140/epjds/s13688-022-00341-x.

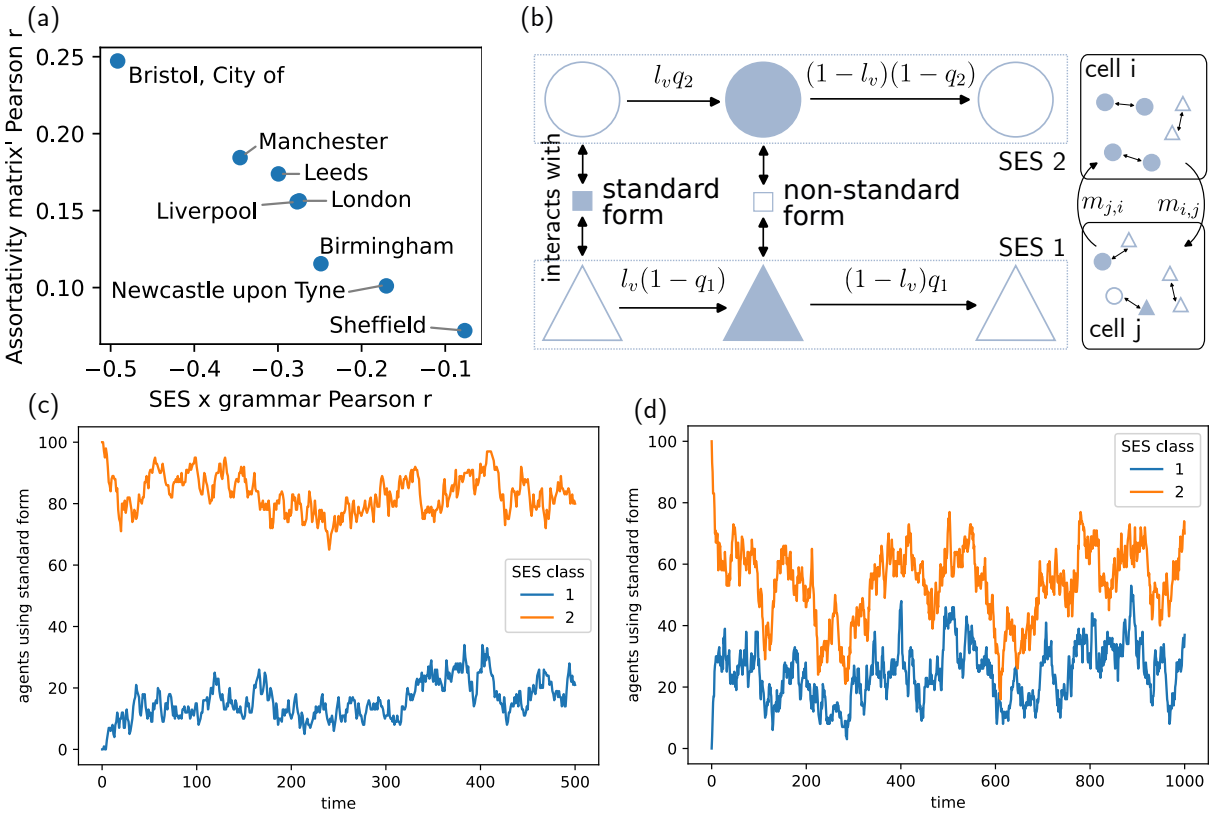


Figure 1: The influence of social mixing on standard use. (a) How the assortativity in mobility correlates with the correlation between SES of origin and the frequency of grammar mistakes, as observed using Twitter data in 8 metropolitan areas of England. (b) Diagram summarizing our agent-based model for the adoption of a linguistic variety. We ran simulations of the model for a prestigious standard form, but an attachment of the lower SES class to the non-standard, when (c) the two classes are mixing well and (d) when they are mixing much less.