

## **Socio-economic and network dependencies of linguistic patterns in Twitter**

Marton Karsai

Ecole Normale Supérieure de Lyon, Laboratoire de l'Informatique du Parallélisme, Computer Science Department; IXXI Rhône Alpes Complex Systems Institute, INRIA, France.

Our usage of language is not solely reliant on cognition but is arguably determined by myriad external factors leading to a global variability of linguistic patterns. This issue, which lies at the core of sociolinguistics and is backed by many small-scale studies on face-to-face communication, is addressed here by constructing a dataset combining the largest French Twitter corpus to date with detailed socioeconomic maps obtained from national census in France. We show how key linguistic variables measured in individual Twitter streams depend on factors like socioeconomic status, location, time, and the social network of individuals. We found that (i) people of higher socioeconomic status, active to a greater degree during the daytime, use a more standard language; (ii) the southern part of the country is more prone to use more standard language than the northern one, while locally the used variety or dialect is determined by the spatial distribution of socioeconomic status; and (iii) individuals connected in the social network are closer linguistically than disconnected ones, even after the effects of status homophily have been removed. Our results inform sociolinguistic theory and may inspire novel learning methods for the inference of socioeconomic status of people from the way they tweet.