Population effects on languages:
Modelling population dynamics and language transmission from the perspective of language learning, contact and change

Interdisciplinary Workshop* — *November, 20 2017* — *Lyon, France
https://poplang.sciencesconf.org/

ANALYSING LANGUAGE VARIATION THROUGH BIG DATA: THE CASE OF TWITTER AND GOOGLE BOOKS

Lucía LOUREIRO-PORTO[1] and David SÁNCHEZ[2]

[1] University of the Balearic Islands
[2] Institute for Cross-Disciplinary Physics and Complex Systems (CSIC-UIB)

In the last years, the use of big data for linguistic purposes has opened up new paths in corpus linguistics, since it offers new opportunities for the investigation of large-scale language variation (see, among others, Russ 2012, Doyle 2014, Grieve 2015). For instance, microblogging platforms such as Twitter provide a deluge of textual data that can be employed for the analysis of informal communication between millions of individuals. Because most of these data are geolocalized, the resulting corpus can be used to conduct studies on regional variation including many more speakers and much larger texts than traditional dialectology. By analysing a database with nearly 4,000 million geolocalized tweets, we address the study of lexical and orthographical variation in Spanish and English. Different methods are applied: cluster analysis, information-theoretic measures and polarization metrics. While linguistic variation in Spanish is found to show a clear distinction between urban and rural speeches (Gonçalves and Sánchez 2014, Donoso and Sánchez 2017), the English language is characterized by two dominant varieties, namely, British and American, the latter becoming dominant outside the United Kingdom (Gonçalves et al. submitted). Finally, we address the diachronic evolution of the two inner-circle varieties of English using the Google Books database. Here, we observe a significant Americanization of English in the last two centuries, with tendencies shaped by cultural and historical milestones.

REFERENCES

Doyle, G. 2014. "Mapping dialectal variation by querying social media". *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics* (EACL).
Gonçalves, B, L. Loureiro-Porto, J. Ramasco & D. Sánchez. *Submittted*. "The fall of the Empire: The Americanization of English". https://arxiv.org/abs/1707.00781.
Gonçalves, B. & D. Sánchez. 2014. "Crowdsourcing dialect characterization through Twitter". *PLOS One* 9, e112074.
Donoso, G. & D. Sánchez. 2017. "Dialectometric analysis of language variation in Twitter". *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pp. 16-25.
Grieve, J. 2015. "Mapping Lexical Spread in American English". *American Dialect Society Annual Meeting*, Portland, Oregon, January 8, 2015.
Russ, B. 2012. "Examining large-scale regional variation through online geotagged corpora", *ADS Annual Meeting*. Retrieved from http://www.briceruss.com/ADStalk.pdf.