Mapping urban multilingualism through Twitter

Enrique Manjavacas
Freie Universität Berlin
enrique.manjavacas@gmail.com

Ben Verhoeven
CLiPS, University of Antwerp
ben.verhoeven@uantwerpen.be

With the advent of globalisation, the interweaving of different languages in urban environments has acquired a high degree of complexity and has become an attractive topic for sociolinguistic research in recent years. Conveniently, in 2009 the microblogging service Twitter added an option that allowed users to attach geolocation information to their broadcasted tweets. Although only 1,6% (Leetaru et al. 2012) of the Twitter stream is actually shipped with geolocation information, these geo-referenced tweets can be utilised to address general issues of urban multilingualism that otherwise elude research due to the high cost of data collection.
In this project, we test, as a proof of concept, the viability of utilising Twitter data to analyse large-scale distributional patterns of language use in four European cities (Amsterdam, Antwerp, Berlin and Brussels, alphabetical order). We use our growing database of tweets, geolocated to the cities of interest, as collected and filtered from the Twitter stream. We experiment with a majority vote approach based on preexistent language identification system in order to guarantee precise language identification, given that this task is an essential preparatory step of the project. Moreover, different preprocessing steps are carried out and evaluated in order to account for potentially distorted data - filtering out bots, identifying tweets by tourists etc,...
Finally, we devise and test various techniques for aggregating and visualising geolocated data and evaluate the data against known demographic figures (official censuses and open source datasets on population and similars).

The resulting dataset is suitable for addressing questions such as the distribution of languages across urban areas, spotting touristic areas, analysing differences of languages across day/night population patterns, etc...
As a result, we aim to obtain insight into the urban multilingual picture and simultaneously into the bias that Twitter data may introduce therein.