Automatic Isogloss Maps from Twitter Data: The Catalan Case

Alexandre Nobajas
Keele University

Thanks to the popularization of social media platforms an unprecedented wealth of written data is now available to researchers (Kumar *et al.* 2014). This ever-growing resource has some special characteristics which make it quite interesting to social researchers, as it is not part of a structured experiment or interview, and therefore much less prone to research biases such as response bias or procedural bias (Bertrand *et al.* 2001). Another important characteristic of this new data source is that, contrary to traditional methods such as surveys, interviews or questionnaires, it is almost immediate and inexpensive to gather, which makes it an excellent complement to the aforementioned methods. However, each social network has its own demographic characteristics which need to be taken into account when drawing conclusions, such as a big age bias towards the younger population (Schoonderwoerd 2013).

This paper's objective is to explore how one of the most popular social networks, Twitter, can be used to create dialect maps and how do they compare to official results created using traditional methods. In order to achieve this, a tweet gathering system has been designed and data mining methods have been used to obtain millions of conversations written in Catalan. In this context, Catalan language is an interesting case study as it is spoken by 9 million people (Strubell and Boix-Fuster, 2011) who have access to the technology and who are quite clustered on the eastern side of the Iberian Peninsula in Europe. It also presents a dynamic speaker community spread across international boundaries with different levels of legal recognition. Finally, its dialectological characteristics have been very well studied and dialect maps have been already been produced, which provides the necessary ground to test the suitability of the method, which could potentially be applied to other languages.