# Sketching Linguistic Borders: Mobility Analysis on Multilingual Microbloggers

Muhammad Syafiq Mohd Pozi
Yukiko Kawai
Kyoto Sangyo University
{syafiq,kawai}@cc.kyoto-su.ac.jp

Adam Jatowt
Kyoto University
adam@dl.kuis.kyoto-u.ac.jp

Toyokazu Akiyama
Kyoto Sangyo University
akiyama@cc.kyoto-su.ac.jp

## ABSTRACT

Twitter has been used for various kinds of sociological studies including also multi-lingual analysis. In this paper we study the phenomenon of multilingualism in Twitter from a novel viewpoint. We advance the existing studies by correlating user mobility and multilingualism. The results we show can be used for explaining the usage of languages based on user location and mobility.

## Keywords

Twitter, microblogging, multilingual, tourism

## 1. INTRODUCTION

Numerous researches made use of Twitter as the main source of data in a broad range of fields and objectives, including ones in the context of multilingual studies [3]. Especially in the current era of mass travel and increasing Internet coverage, Twitter has been used to generate data across many languages and cultures on a large scale from different parts of world [4], as people from different backgrounds like to share their travel experiences online.

Although some prior researches focused on the study of languages used in Twitter [6, 1], few works approached user multilingualism in Twitter [3, 5], that is, the case when the same users share content in different languages. We contribute to that study by examining particular, yet, important factors behind multilingualism - the effect of user mobility and the effect of user's mother language on her or his propensity of using different languages. This work offers conceptual analysis of multilingual travelers and their language of choices conducted on a relatively long snapshot of Twitter data. Our study could benefit tourism research especially service-oriented tourism [2].

## 2. ANALYSIS

### 2.1 Data Collection and Preprocessing

Our analysis is based on geo-tagged Twitter data (using Twitter API) over western and central part of Europe gathered within approximately 6 months from 30th Apr 2016 to 21st Dec 2016 consists of 16.5 million of tweets accumulating to 5 gigabytes in memory size. The dataset consists of people who are multilingual and who visited at least 2 countries within the time frame of analysis. In the context of our dataset, this means the fact of posting tweets from two or more different countries by the same user. A user is deemed capable of speaking a given language if she issued more than $\alpha$ ($\alpha = 5$) tweets in that language. Fig. 1 shows the distribution of languages (we show only European languages) accumulated from multilingual users from each European country, while Table 1 lists user distribution per each multilingual category.

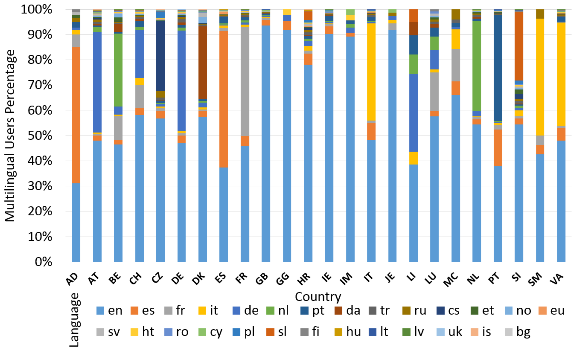

**Figure 1: European language distribution across different European countries.**

### 2.2 Mobility vs. Multilingualism

We first investigate how the mobility influences multilingualism. To answer this question, we perform two analyses. The first one determines how many countries are visited by travelers according to their ability of speaking given number of languages. Fig. 2 shows the average number of countries visited by travelers vs. average number of used languages. We experiment with different $\alpha$ values as a threshold that must be exceeded for considering a user as able to converse in a given language.

For all values of *alpha* the positive correlation can be observed, meaning that speakers of many languages tend to visit more countries compared to travelers who use few number of languages.

**Table 1: Counts of users in each n-lingual group.**

| #languages | Total Users | Percentage |
|---|---|---|
| 1 | 57,240 | 26.69 |
| 2 | 64,503 | 30.08 |
| 3 | 38,974 | 18.18 |
| 4 | 22,142 | 10.33 |
| 5 | 12,430 | 5.80 |
| > 5 | 19,147 | 5.25 |
| Total | 214,436 | 100 |

*WWW'17 Companion,* April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4914-7/17/04.
http://dx.doi.org/10.1145/3041021.3054261

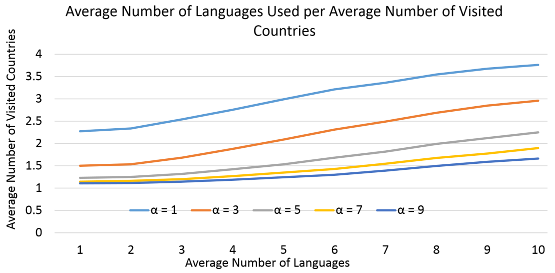

**Figure 2: Average number of visited countries vs. average number of languages used.**

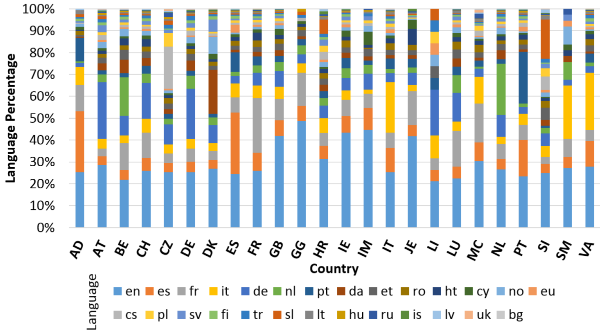

**Figure 3: Language distribution among multilingual travelers based on their home country.**

The second analysis explores the language usage by multilingual travelers according to their home (or base) country. Here, the home country for a given user is defined as the country from where he or she posted the largest number of tweets. Fig. 3 shows the language distribution of multilingual travelers when they are in their home countries. We can observe that English is less common than in Fig. 1, which considered all tweets in the dataset (not only ones issued from one's home country as in Fig. 3). Yet, still we often observe significant percentages of languages different from the official languages of these countries (e.g., Spanish, French, Italian, German in UK or German, Danish in the Netherlands). These are likely from foreigners who immigrated or travellers who stay longer time at different countries.

### 2.3 Seasonal Effect

We next use DBSCAN to determine the centroid for a given population of geo-tagged users on a map. Fig. 4 shows the distribution of multilingual users over two different seasons: spring and summer. The color ranges from darker blue (less density of multilingual travelers) to bright red (higher density).

The distribution of multilingual travelers differs across the two seasons in particular locations. For example, it is higher at the seaside during summer (June, July and August) compared to spring (April and May). Popular seaside places such as Barcelona and attractive, lively islands (e.g., Balearic Islands, especially, Ibiza being popular destination for European youth) and historically and culturally attractive cities (e.g., Venice, Prague, Berlin, Lisbon) are the most popular vacation spots attracting multilingual travellers more in summer than in spring.

### 2.4 Language Association

Lastly, we look into inter-language associations. Let language $a_i$ be defined by a vector consisting of the numbers $n$ of users that use this language and some other language, $a_i = [n_{a_1}, n_{a_2}, ..., n_{a_l}]$ (e.g. English = {Polish: 300, Spanish: 500} would mean there are 300 English speakers who also use Polish and 500 English speakers who also use Spanish). Based on such vectors, we calculate distance based on the generalized Jaccard distance, such that $J(x,y) = \sum_i min(x_i, y_i) \;/\; \sum_i max(x_i, y_i)$ and $d_j(x,y) = 1 - J(x,y)$. Fig. 5 shows the resulting agglomerative based dendogram defined by the Jaccard distance.

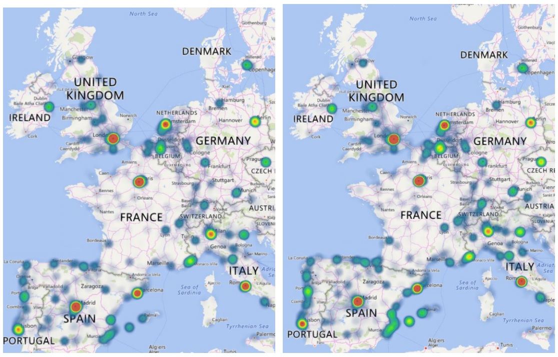

**Figure 4: The distribution of multilingual travelers during *Spring* (left) and *Summer* (right).**

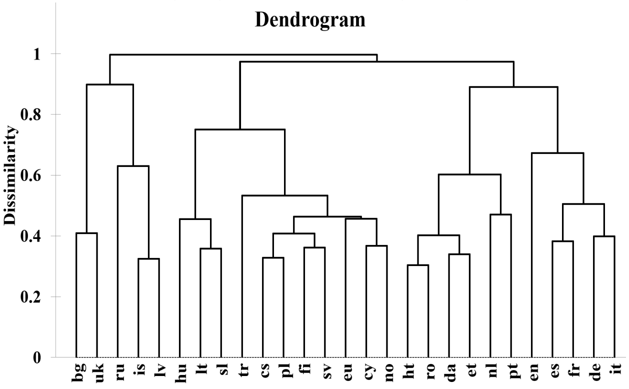

**Figure 5: Associations of languages.**

## 3. CONCLUSIONS

We have investigated the relation of mobility, places and home countries of multilingual travelers based on Twitter data. Future work will analyze the effect of particular events on the use of different languages and more closely the temporal shifts in language use in the same places.

## 4. ACKNOWLEDGMENTS

This work was partially supported by MIC SCOPE (150201013), and JSPS KAKENHI Grant Numbers 15K00162 and 16H01722.

## 5. REFERENCES

[1] C. Catal et al. A sentiment classification model based on multiple classifiers. *Applied Soft Computing*, 50:135–141, 2017.

[2] F. Constantin et al. Multilingual online communications in corporate websites: Cases of romanian dental practices and their application to health tourism. In *Tourism and Culture in the Age of Innovation*, pages 185–196. Springer, 2016.

[3] I. Eleta et al. Multilingual use of twitter: Social networks at the language frontier. *Comp. in Human Behavior*, 41:424–432, 2014.

[4] G. Hogan-Brun. *Linguanomics: What is the Market Potential of Multilingualism?* Bloomsbury Publishing, 2017.

[5] C. McCollister. *Predicting Author Traits Through Topic Modeling of Multilingual Social Media Text.* PhD thesis, University of Kansas, 2016.

[6] K. Rudra et al. Understanding language preference for expression of opinion and sentiment: What do hindi-english speakers do on twitter? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1131–1141, 2016.