Report

2013

The University of Manchester

MAPPING MANCUNIAN MULTILINGUALISM IN TWITTER

Maëla Mainguy

Yumi Nakai

Megumi Takayama

1 Introduction

The study we conducted aimed at mapping the multilingual tweets collected in Manchester on a two-months-period. Its goal was also to explore whether multilingualism on the Internet accurately reflects Mancunian multilingualism in the real world. Additionally, in the course of the research, we will discuss the accuracy of the language detection software used to analyze our data.

1.1 Method

We decided first to go roughly through the data and see if we could draw relevant information at first sight. The original spreadsheet contained approximately 28000 entries already sorted by language thanks to a language identification software, the Chromium Compact Language Detector. We initially wanted to verify the accuracy of the software, which proved harder than expected in regards to the number of tweets and our limited knowledge of most of the foreign languages listed in the spreadsheet. After noticing the unbelievable amount of English tweets classified as another language, we decided to delete them manually. Ed Manley (2012) had already noticed that the data classified as Tagalog were in fact all English, and our spreadsheet exhibited the same flawed identification, so we deleted all the tweets labelled as Tagalog. The other English tweets were deleted one by one. The second part of our study consisted in determining the predominant languages among the tweets and comparing the modified data to the original data. This allowed us to estimate a degree of accuracy concerning the language detection software used by Ed Manley, but also to determine the Mancunian multilingual landscape on Twitter. The software used to classify the data (Excel) allowed us to sort the data by language or user ID, as explained in the fieldwork plan. This made the calculation of the number of tweets per language very easy, but also allowed us to identify errors of classification for specific languages. This also enabled us to observe the number of contributions per user.

As intended, we also created an online map of Manchester plotting the tweets of the predominant languages. The process was extremely complicated as most mapping programmes are extremely expensive or require specific computing knowledge, and in the time lapse between our fieldwork plan and the actual creation of the map, the website initially chosen (batchgeo.com) had become a paying website. In the end, we managed to find a website plotting GPS data with latitude and longitude (http://www.gpsvisualizer.com). We isolated the GPS coordinates associated with each tweet and the language of the tweet as an indication for the map. We converted the file to a csv. file and just had to upload it. The map obtained is accessible online, but the likelihood of it being deleted led us to capture images of specific areas to illustrate our findings and discussion later on in the report. We compared the map obtained to the census-based map of the Guardian (2011).

In our fieldwork plan, we stated that we hoped to contact some of the users, another step which would have allowed us to obtain more qualitative data. We tried to send the users a questionnaire to determine whether the detected language of their tweets was their first language and what attitudes they had towards their language. However, this turned out much more difficult than we initially expected. We followed more than 200 twitter users whose tweets have been detected as Arabic, French and Malay, which were the three predominant detected languages according to the final spreadsheet, and we did not get any answer. This method has also revealed the limit of conducting a survey by Twitter as we were not allowed to follow more than 200 users when no one was following our account in return. Although we didn’t obtain any answers from them directly, we found out few things by observing their accounts.

2 Findings

2.1 Overall findings

As stated earlier in the report, we went through the data manually and discovered a number of English tweets classified as other languages. Generally, these tweets, also coined as Microblogs (Carter et al., 2013), were written in an informal English (no particular attention is given to grammar or spelling). The common use of hashtags (#) followed by unsegmented words was a recurrent feature as well. One of the other reasons for incorrect detections by the software were Internet slang such as '*xxx', 'soz*' for sorry and '*lol'*. It seems that the detection tool identifies such slangs as non-English languages. We noticed that most of the English tweets present were identified as Czech, Danish, Dutch, and German. In total, we deleted approximately16000 tweets, more than 57% of the initial spreadsheet's number of entries. Of the eight languages in which we had to delete more than 100 tweets, in six of them English tweets amounted to more than 80% (see table 1).

Languages	**Percentage of English tweets**
Czech	97,48%
German	96,63%
Danish	93,50%
Galician	90,60%
Finnish	86,55%
Dutch	86,02%
Indonesian	32,65%
Italian	30,72%

Table 1*, Percentage of English tweets in languages containing more that 100 of English tweets*

More than 4500 English tweets were classified as German and nearly 900 as Danish (see table 1 and Comparative Chart appendix). The few tweets classified as Basque

and Vietnamese were actually all English, and there was just one Estonian tweet left after the deletion of English tweets.

The probability of certain languages to display an important amount of English tweets brought our attention to the reasons behind this and we discovered that most of the tweets classified as Dutch in fact contained either names (such as the football player Van Persie who was the most recurrent, or even Justin Bieber!) or unsegmented words (see Table 2)

USERID	USERNAME	TEXT	DETECT_LANG
pau1luvsutd	Paul Hart	Ooooh robin van Persie	DUTCH
charli_louise96	CHARLI	#MagalufWeekender	DUTCH
LucKusumah	Luciana	#TweetLikeAFacebookStatus like for a rate	DUTCH

Table 2*, Examples of English tweets classified as Dutch, extract from the original spreadsheet*

After revision, we were able to determine which languages were highly spoken on Twitter in Manchester. Figure 1 illustrates the distribution of tweets in the predominant languages as it appeared in the original data, which means that it contains English tweets as well as non-English tweets. According to Figure 1, the top three languages tweeted in Manchester were Tagalog, German, and Arabic in the original data, and it also indicates that about 30% of the non-English tweets in Manchester, which amounts to approximately 8000 texts, were classified as Tagalog, although we all Tagalog tweets were deleted in the revised data as Figure 2 shows. According to Figure 2, which represents the number of tweets in the predominant language based on the revised data, the top three languages tweeted in Manchester are Arabic, French, and Malay according to the revised data. The language distribution in the revised data, however, is quite different from Ed Manley's original data.

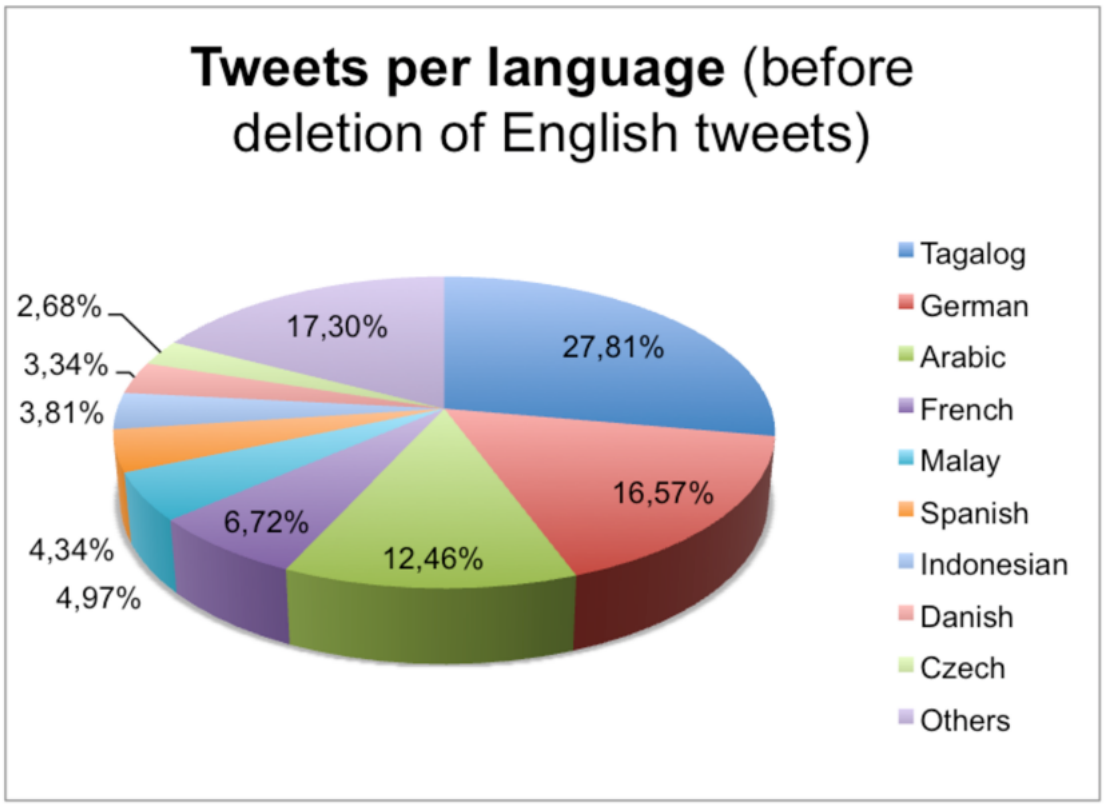


Figure 1, *Percentage of tweets per language before deletion of the English tweets.*

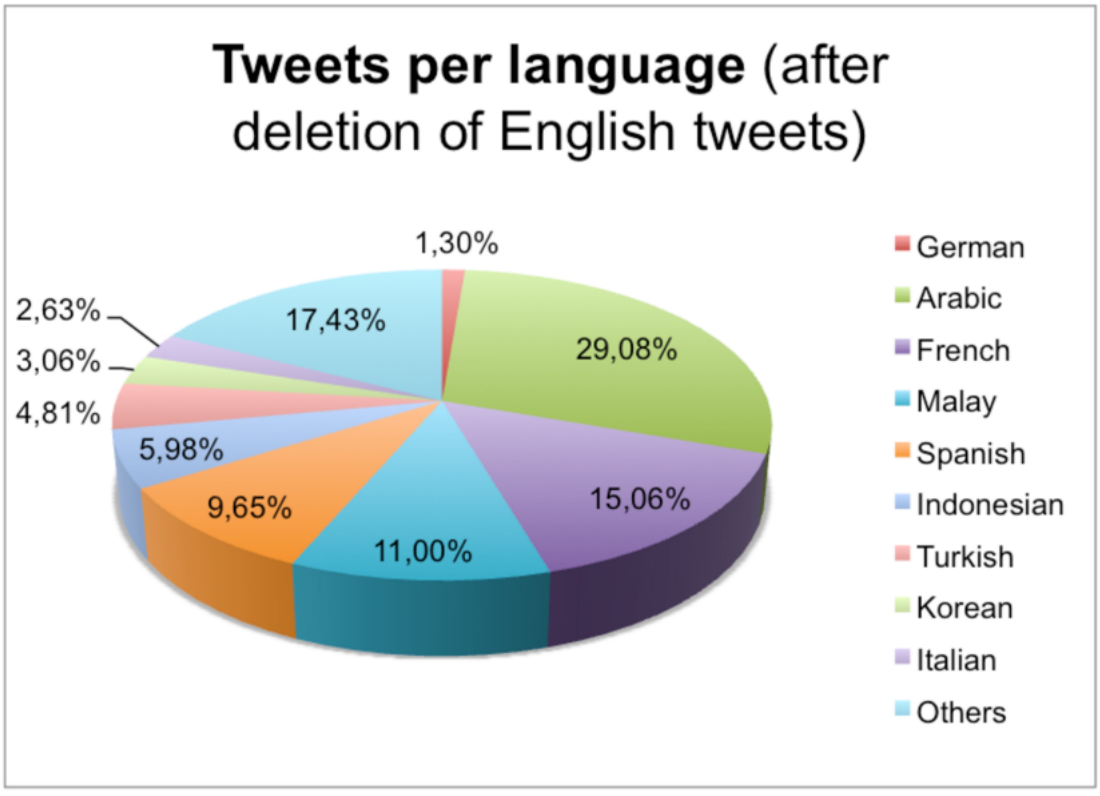


Figure 2, *Percentage of tweets per language after deletion of the English tweets.*

While analysing the data, we also observed several interesting facts. First of all, we found two Latin sentences that were not recognized as such by the language detection software. The second important information concerned the amount of data

per user; indeed, it appeared that while there seemed to be a great deal of tweets in some languages, many of those tweets were written by the same person. This was the case in particular for French. Although there were around 1800 French tweets recorded in the spreadsheet, we found out that only 126 different people produced them. As a matter of fact, the most active Twitter user had written around a thousand tweets over the period investigated. Numerous Arabic tweets were written in Latin script, and thus were not recognized as Arabic by the language detection software.
Finally, we found out that some of the users used code-switching on Twitter; most of the tweets displaying code-switching were labelled as Malay, hence our deduction that the language was in fact Manglish: (=> in discussion, LGD does not recognize creoles!).

E.g. "*Penatnyaaaa. macam travel Malaysia to London haaa T.T"*

E.g.2 "*@mSyuhadaa yup...absolutely..time2 bosan pon boleh tgok, I am CBO no1 fan.haha"*

2.2 Map

The realization of a map (http://www.gpsvisualizer.com/display/data/1369221214-05957-130.88.162.97.html) gave us a completely new approach to the data collected. Because of the limited capacity of the website supporting it, we chose to map only the most used languages (fourteen languages altogether; fifteen are displayed on the map for the sole reason that Chinese tweets were separated in two categories whether they used Latin script or not). Although we could not get a map as precise as Ed Manley's (2012), the result enabled us to see a repartition of Twitter languages in Manchester and to extract relevant information. The first overall look of the map allows us to see the density of multilingual tweets per area. The denser area is the city centre, and the areas of Bolton, Oldham and Rochdale also feature an important concentration of tweets (see map 1 in the appendix). As we zoom in the city centre we can clearly see the tweets following the main axes and roads, especially Oxford Street and Portland Street (see map 2). Other high-density zones are the Northern Quarter and Piccadilly, and along Oxford Road on the Curry Mile. The Eastern part of

the City Centre appears to be less dense. Some tweets tend be grouped, such as Arabic, Malay and Korean, while others seem to be spread across the map without any specific pattern. There are several very important concentrations of Arabic tweets, the biggest being around the Palace Theatre on Oxford Street and in Piccadilly (see map 4). Along the Curry Mile, there seems to be two major clusters of languages, which are Malay and Arabic (see map 3). Finally, we noticed that there were small concentrations of tweets around universities and student halls.

We compared our findings and map to the Guardian map based on the 2011 census. The density in the city centre appears quite similar; the Eastern part seems less multilingual that the West and the Northern Quarter. The areas along Oxford Road are very dense as well. When we compare the most important languages per area, the use of Arabic also seems to be high in comparison to most of the other languages, especially along the Curry Mile. However, Urdu and Kurdish are also part of the dominant languages in the Guardian map, but none of them appears on ours. Similarly, according to the Guardian map 17.2% of the population has Urdu as their main language and 12.43% Punjabi. According to our tweets, both Punjabi and Urdu came to a very low percentage among all the tweets.

2.3 Case study

When observing some of the accounts of users whose detected language was Arabic, Malay or French, we noticed some tendencies in the behaviour of users of each language. Some of the users were not actually living in Manchester, so we ruled them out for the most part. Many of the users who do live in Manchester use both English and the detected language in their tweets. Most of their profiles were written in two languages, especially for the Arabic users (see Appendix 2). They tend to mention their origin or their ethnicity and where they currently live. However, when it comes to their tweets, it seemed that they mostly use Arabic not only in the Arabic script but also in the Latin script. Almost half of the French detected users who live in Manchester use English as a main language, while the half uses mostly French. French speakers generally tend to use English in social networking services. Malay

speakers living in Manchester generally use English and when they respond to a certain person, they use Malay or code switching (Manglish).

There were some interesting individual cases amongst the users. There was one user who usually lives in Switzerland. By seeing her account, she seemed to be a Manchester City supporter and she had been in Manchester for a football game since she was tweeting about the home stadium of Manchester City. Another male user, who currently lives in Manchester, uses mainly English but also Spanish and French (see Appendix 3). This user seemed to be from Spain since he was communicating with who seemed to be his sibling. We also found one user who speaks mainly Portuguese and occasionally uses French.

3 Discussion

3.1 Accuracy and Software

First and foremost, it is very important to precise that this study is very experimental; therefore its percentage of accuracy is not always in satisfactory. We are aware of the limitations of the study, especially given the manual deletion of thousands of entries that was required. As a result, there might still be a few English tweets classified as another language in the spreadsheet we used, that is why we determined a possible 10% margin on error. However this does not falsify the data in any way. On the same note, we observed that the recognition of languages by the Google Chromium Compact Language Detector was highly biased, especially when dealing with Latin script. The problem was probably due to the informal nature of the tweet, which – as already proven by Carter et al. (2013) – confuses the language detection software and is less recognizable. It is interesting to note that the languages confounded most of the time, that is to say English with German and Dutch, are all Germanic languages, thus their probable mingling. Another important aspect of the deletion of the English tweets is that because they amounted to more than half of the tweets, the proportion of estimated multilingualism on Twitter diminishes. From 5% multilingual tweets in Manchester over the period examined,

the number dropped to 2.65%. However, we were impressed by the accuracy of the software concerning non-Latin script. We estimated the accuracy of the classification of Arabic, Chinese and Japanese tweets written in non-Latin script at 100%. The reason for that is of course that all three languages have their own script. All in all, the language detection software did not stand on its own, and a way to obtain a very solid study on Twitter would be to have speakers of each language manually check and identify the tweets one by one.

The significant difference between that appeared before and after the deletion of English tweets could imply the inaccuracy of Ed Manley's own data on multilingual tweets, knowing that he was the one who provide our data and did not mention any of the problems we encountered in his own paper.

3.2 Predominant languages and spatiality

As mentioned in the findings, Arabic was the most used language on Twitter, followed by French, Malay and Spanish. We can compare those results to the 2011 census and it is obvious that the results are not the same. According to the census, the main language in England, and in Manchester, was Polish, but the spreadsheet did not show evidence of more than a hundred Polish tweets. Although this census was very disputed, it still gives an indication of the main languages in the United Kingdom, and our results did not match those of the census. The only hypothesis that we can think of is that Twitter might be more popular among certain communities.

In an attempt to compare our findings with other studies, we can point out Veselinova & Booza's (2006) research in Detroit; they found out that some languages tend to form clusters while others are non-clustering languages. Similarly, we observed important clusters of Arabic and Malay tweets, and we could hypothesize that they indicate an important community. The comparison with the Guardian map tends to support this idea as the areas with an important density of Arabic tweets are also areas where a lot of people consider Arabic as their first language. Nonetheless, there is a limit to this hypothesis. Indeed one of the biggest clusters on our map is situated around the Palace Theatre, on the road; as this is not a residential area we

can deduce that people did not post those tweets from their home. This is where any study trying to locate people's residence with tweets will fail: most tweet are in fact sent from phones, hence the moving patterns and the high density of tweets along the main roads. Although a general overview of a map can give us an idea of the places more frequented by people according to their language, we cannot be wholly sure of the meaning of the tweets' provenance (unless we did a precise and complete follow up of each user, which would be highly challenging if not impossible). All in all, the mapping of tweets does not truly provide a spatial repartition of people according to their language, but it gives us an idea of how multilingual a city is, and allows us to determine clusters of languages that illustrate the movements of a given community. It also provides a hierarchy when it comes to the languages the most used on Twitter in Manchester.

3.3 Importance of users

Our findings also brought our attention to the importance of the user. First of all, the contribution of each user is variable and while some tweeted once or twice over the two-months-period observed, others were overly active and produced more than a thousand tweets. This is one of the reasons why the Twittersphere is in no instance comparable to the reality when it comes to multilingualism. We also observed that a lot of tweets were located around Universities and student halls. This is probably a hint concerning the mean age of Twitter users, which are apparently students for the most part. In fact, we can link this to the Huffington Post's article about Twitter (2012), in which it was indicated that around 73% of its users were between 16 and 25 years old.

The main issue of asking people to answer a questionnaire was that we did not really take the time to implement a solid plan for them to feel like they had to answer it. If we had conducted a survey on a long-term period, this would have been solved and we probably would have collected more qualitative data. Nevertheless, we have been able to observe around 200 accounts and found some relevant information about users' behaviours in a social network service. On the whole, the

Twitter users use English as well as other languages on the Internet and this partially reflects their real life because they need to socialize in English as they live in Manchester, which is an English-speaking region. The users' community has the capacity to extend beyond ethnicity since they live in a multilingual society and this involves the use of English to facilitate communication with others.

From the case studies we detected numerous examples of bilingualism and trilingualism, and therefore deduced that multilingualism in Manchester was also present on an individual level. The attitudes towards bilingualism and code-switching on the Internet led us to the conclusion that, as previously stated by Herring & Danet (2007), people can belong to two or more speech communities and display their own multiculturalism on SNS and particularly, in our case, on Twitter.

Bibliography

Bergsma, S., McNamee, P., Bagdouri, M., Fink, C., & Wilson, T. (2012). Language Identification for Creating Language-Specific Twitter Collections. In *Proceedings of the 2012 Workshop on Language in Social Media*. Montreal, Canada in June 7th 2012. pp. 65-74.

Carter, S. ,Weerkamp, W., & Tsagkias, M. (2013). Microblog Language Identification: Overcoming the Limitations of Short, Unedited and Idiomatic Text. *Language Resources and Evaluation Journal*, 47, pp. 195-215.

Herring, S. C. & Danet, B. (2007). *The Multilingual Internet: Language, Culture, and Communication Online.* New York: Oxford University Press.

Honigman, H. (2012). *100 Fascinating Social Media Statistics and Figures From 2012.* The Huffington Post. [online] Available at: http://www.huffingtonpost.com/brian-honigman/100-fascinating-social-me_b_2185281.html [Accessed 24/05/2013].

Manley, E. (2012). Detecting Languages in London's Twittersphere. [online] Available at: http://urbanmovements.co.uk/2012/10/23/detecting-languages-in-londons-twittersphere/ [Accessed 24/05/2013].

The Guardian. (2011). *Languages mapped: what do people speak where you live?* [online] Available at: http://www.guardian.co.uk/news/datablog/interactive/2013/jan/30/languages-mapped-england-wales-census?zoom=12&lat=53.47802508552413&lng=-2.231103185058605 [Accessed 24/05/2013].

Veselinova, L., & Booza, J. C. (2006). Using GIS to map the multilingual city. In *Proceedings of the 26th ESRI International User Conference*. San Diego, CA in August 7th-11st 2006.

Appendix

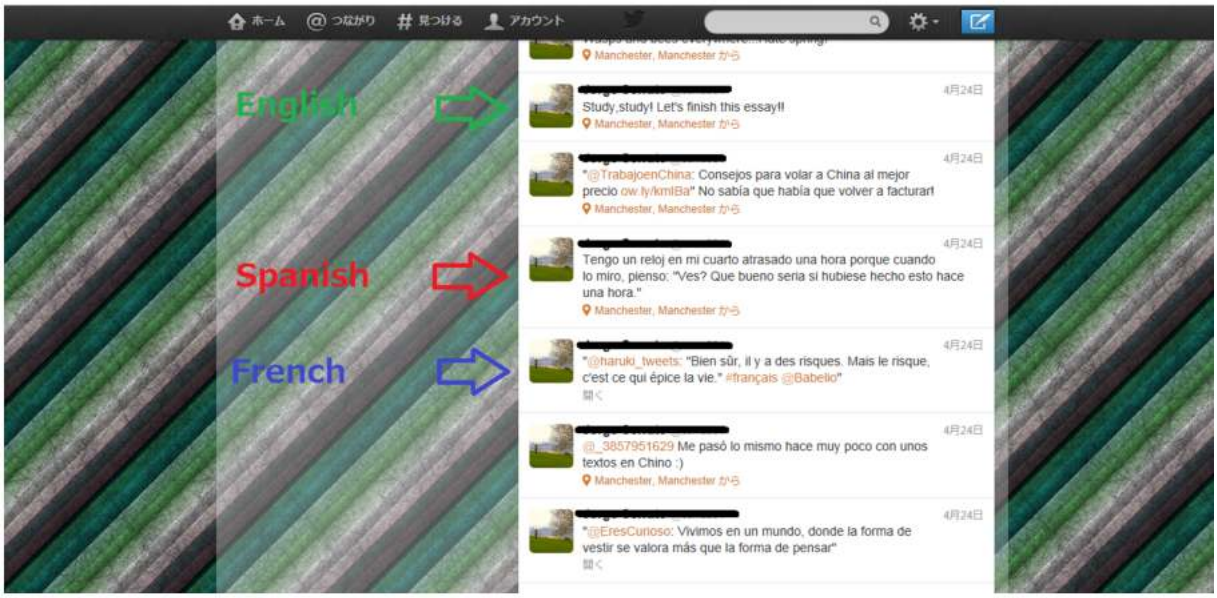
1. Comparative Chart

Languages	Original	Revised	
Afrikaans	24	2	−22
Albanian	63	6	−57
Arabic	3502	3512	10
Azerbaijani	10	10	±0
Basque	10	0	−10
Bengali	1	1	±0
Bulgarian	11	11	±0
Catalan	105	64	−41
Chinese	156	150	−6
Chinese T	15	15	±0
Croatian	56	19	−37
Czech	754	19	−735
Danish	938	61	−877
Dutch	658	92	−566
Estonian	24	1	−23
Finnish	119	16	−103
French	1890	1819	−71
Galician	117	11	−106
German	4659	157	−4502
Greek	112	107	−5
Haitian Creole	2	2	±0
Hebrew	2	2	±0
Hungarian	23	11	−12
Icelandic	5	4	−1
Indonesian	1072	722	−350
Inuktitut	2	2	±0
Irish	25	19	−6
Italian	459	318	−141
Japanese	160	155	−5
Korean	373	370	−3
Latin	0	2	2
Latvian	23	18	−5
Lithuanian	227	137	−90
Macedonian	2	2	±0
Malay	1398	1328	−70
Maltese	17	6	−11
Norwegian	214	161	−53
Persian	80	80	±0
Polish	115	101	−14
Portuguese	295	275	−20
Punjabi	1	1	±0
Romanian	51	37	−14
Russian	190	190	±0

Serbian	4	2	−2
Slovak	171	100	−71
Slovenian	18	14	−4
Spanish	1219	1165	−54
Swahili	35	30	−5
Swedish	117	91	−26
Tagalog	7819	0	−7819
Thai	45	45	±0
Turkish	644	581	−63
Urdu	21	21	±0
Vietnamese	41	0	−41
Welsh	21	12	−9
55 languages			
Total	28115	12077	−16038

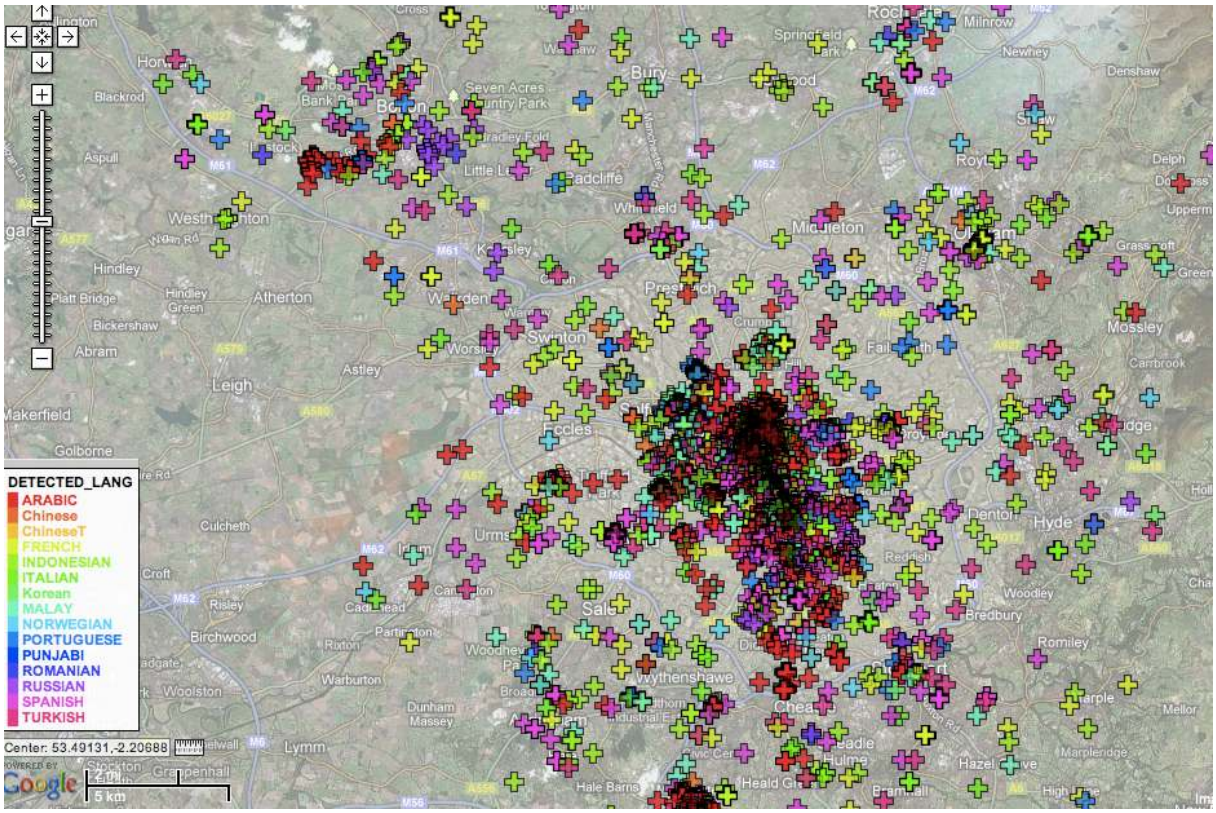
2.Arabic twitter user profile

3. The use of three languages in one account

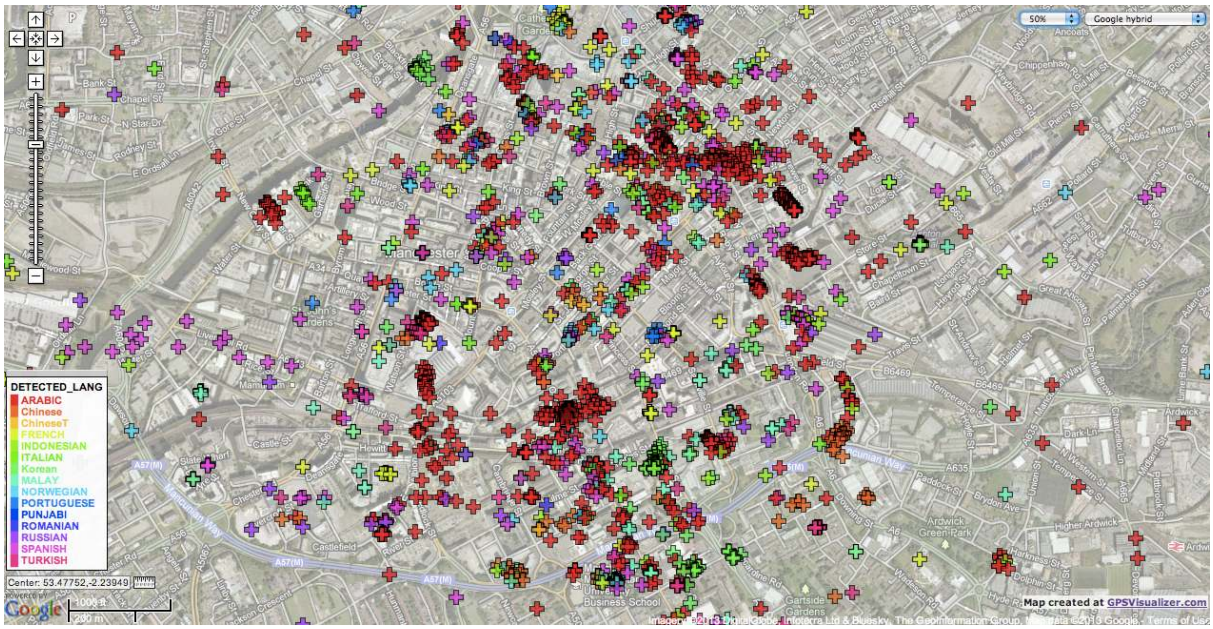


Maps :

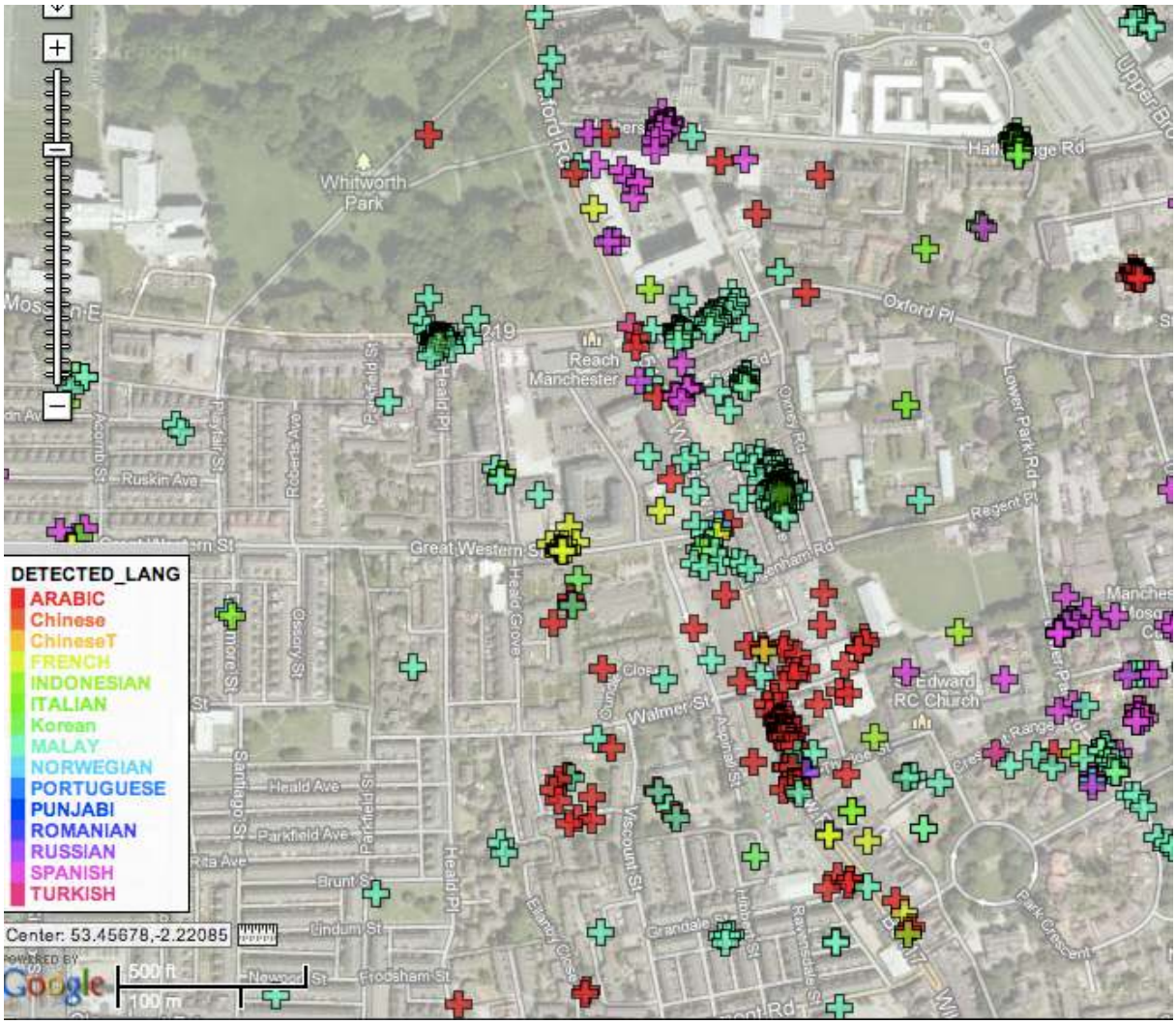
1.


DETECTED_LANG
ARABIC
Chinese
ChineseT
FRENCH
INDONESIAN
ITALIAN
Korean
MALAY
NORWEGIAN
PORTUGUESE
PUNJABI
ROMANIAN
RUSSIAN
SPANISH
TURKISH
Center: 53.49131,-2.20688
5 km
Google
Bolton
Bury
Rochdale
Salford
Eccles
Leigh
Atherton
Wigan
Middleton
Oldham
Hyde
Stockport
Altrincham
Sale
Manchester
Trafford

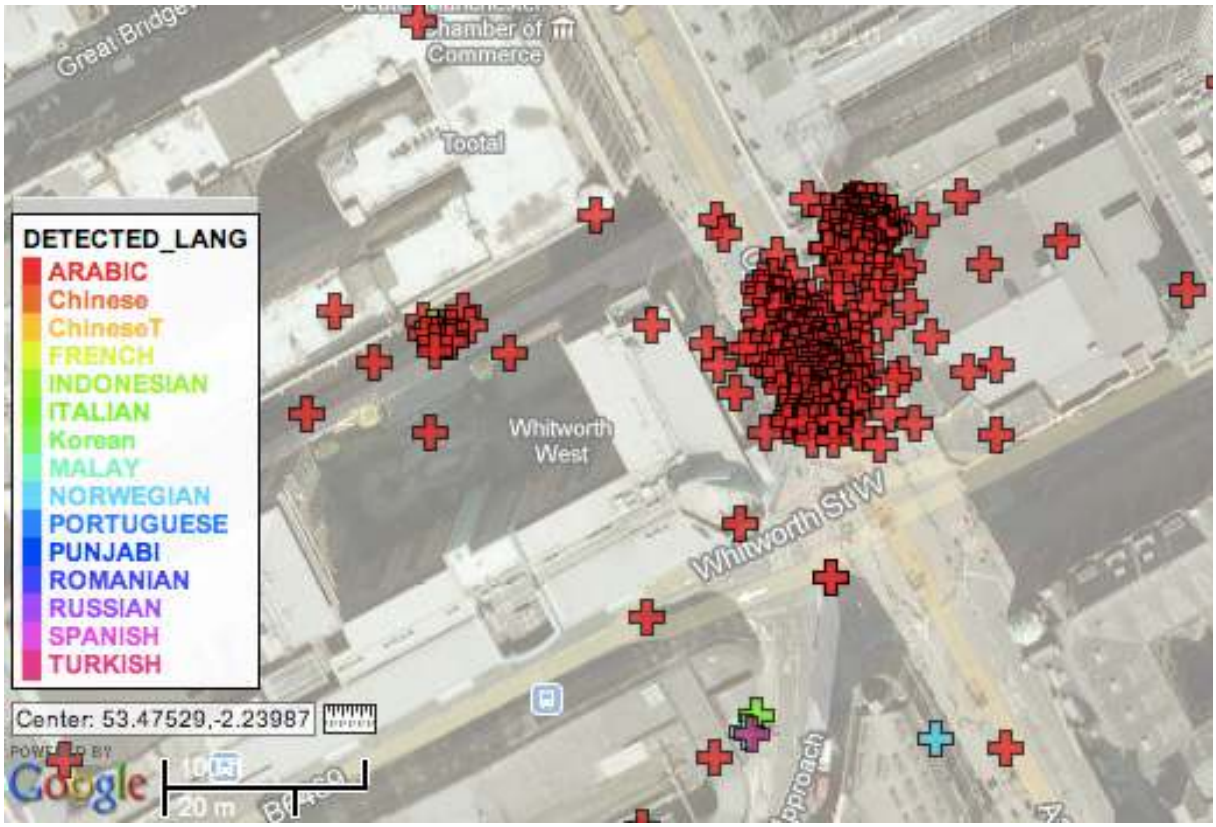
2.


DETECTED_LANG
ARABIC
Chinese
ChineseT
FRENCH
INDONESIAN
ITALIAN
Korean
MALAY
NORWEGIAN
PORTUGUESE
PUNJABI
ROMANIAN
RUSSIAN
SPANISH
TURKISH
Center: 53.47752,-2.23949
Google hybrid
Map created at GPSVisualizer.com

3.


DETECTED_LANG
ARABIC
Chinese
ChineseT
FRENCH
INDONESIAN
ITALIAN
Korean
MALAY
NORWEGIAN
PORTUGUESE
PUNJABI
ROMANIAN
RUSSIAN
SPANISH
TURKISH
Center: 53.45678,-2.22085
Whitworth Park
Reach Manchester
Great Western St
Walmer St
Oxford Rd
Oxford Pl
Heald Pl
Heald Ave
Parkfield Ave
Parkfield St
Brunt St
Lindum St
Frodsham St
Viscount St
Ruskin Ave
Regent Pl
Lower Park Rd
Park Crescent
Edward RC Church
500 ft
100 m
Google

4.


Great Bridge
Chamber of Commerce
Tootal
Whitworth West
Whitworth St W
DETECTED_LANG
ARABIC
Chinese
ChineseT
FRENCH
INDONESIAN
ITALIAN
Korean
MALAY
NORWEGIAN
PORTUGUESE
PUNJABI
ROMANIAN
RUSSIAN
SPANISH
TURKISH
Center: 53.47529,-2.23987
Google
20 m