Using Twitter To Collect Global Multilingualism Statistics

Jelle Jan Bankert
Twente University
Drienerlolaan 5
Enschede, The Netherlands
j.j.m.bankert@student.utwente.nl

Erwin Starke
Twente University
Drienerlolaan 5
Enschede, The Netherlands
e.starke@student.utwente.nl

ABSTRACT

People communicate mostly in their own languages, both at home and online. Surveys try to research these language distribution, but this is mostly labour intensive and can only provide data for a sample of the population. With frequent use of Twitter, online language behaviour can be observed and measured properly. Using big data to analyse large quantities of tweets, a language distribution can be created for a country. This data can correlate with survey results, but results still vary.

Categories and Subject Descriptors

H.1.2 [**Models and Principles**]: User/Machine Systems—*Human information processing*;
H.2.8 [**Database Management**]: Database Applications—*Data mining*;
H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Distributed systems*

Keywords

multilingualism statistics, Twitter, Apache Pig

1. INTRODUCTION

When the telephone was first introduced, long-distance communication with people speaking a different language became possible without the need to be close to another. Since the late 19th century, these communications have gotten more commonplace and the need to speak foreign languages became more and more apparent. With the mass-adoption of the Internet and social networks, a person can now for example reach over 250 million people at Twitter with a single message. Languages have however not converged significantly in that period of time, and so people still need to use multiple languages to address a large, global group of persons.

The multilingualism research that has been performed up to now has several drawbacks. Firstly multilingualism is measured in specific contexts as a proxy for more general multilingualism. This in itself is not a problem, but different researchers and institutions use different proxies. This means that there is no unified, global metric to compare multilingualism. Secondly the current research is performed by asking people to self report, rather than through observing. This used to be the cheaper and faster solution. However, due to big data tools and the public availability of datasets such as Twitter, this has now reversed. A single person can gather data on millions of people in a matter of hours or days and without using any of their time. This has the additional benefit that it is insensitive to cultural biases that may occur with a self reporting methodology. Lastly, the current research is performed in intervals that are on the order of magnitude of years. Using big data to perform this analysis allows us to decrease this op to the point where it is performed in real-time.

In short the goal of our research is to extract statistics on multilingualism from a Twitter dataset and compare those to the official data. The Twitter dataset will be gathered through their Streaming API. For our research not only research papers but also government reports are essential to understanding the context and need for our proposed research.

In section 2, we will look at related work and create a context for multilingualism research with Twitter data. In section 3, we will describe our methods and materials used in this research. Section 4 will cover our results and discussion of those results. Section 5 will contain our conclusions drawn, with section 6 available for future relevant work.

2. RELATED WORK

2.1 Related Papers

Multilingualism research is based on the combination of language and location. Both of these can be gathered from the Twitter metadata in multiple ways. Location is asked for in a Twitter account, but this data is very unreliable [2]. The location field can be empty, contain false information, or be written in different styles for the same location (e.g.: NY, NYC, New York, New York City). Geodata is more reliable for determining a user's location, but only a small percentage of tweets contain this metadata. Timezones can also point to the general location, but this data can be unreliable as well.

The preconditions for language identification are more reliably satisfied, as false information or blank tweets don't occur. Different software packages exist to analyse a piece of text, but the small amount of characters doesn't make this an easy task. Recently, Twitter included its own language detection algorithm to their API, providing the language classification in the metadata[1].

Combining language and location can create interesting results [3]. In Europe, language borders can become apparent as several bordering countries have different languages. Even multilingual areas, such as Belgium or Montreal (Canada), can show hotspots for language activity.

2.2 Reports

Government reports provide baseline data against which we can compare our results. Different countries measure multilingualism in different ways. In all cases the researching institution ask people to self-report their data rather than obtaining the data through observations.

The United States Census Bureau provides their multilingualism data available online [1]. Their proxy for multilingualism is whether or not another language than English is spoken at home. The foreign languages are grouped into Spanish or Spanish Creole, Other Indo-European languages, Asian and Pacific Island languages and Other languages. The population is grouped by age into the 5-17 year, 18-64 year and 65 year partitions. An estimated 20.8% of the population speaks a language other than English, of which the Spanish or Spanish Creole accounts for 13.0 percent points.

The European Commission published their latest report on multilingualism in 2012 [4]. This is an extremely thorough and extensive report and we will present only the most relevant metrics with respect to our proposed research. Their proxy for multilingualism is that a person can hold a conversation in either 0, at least 1, at least 2 or at least 3 foreign languages. These numbers are given per country, per age group and per amount of use of the internet. For each of the countries the top-3 foreign languages and the percentage of people who can hold a conversation in it is given. Of the population 54% can hold a conversation in more than one foreign language, 25% in more than two and 10% in more than three. Luxembourg, Latvia and The Netherlands are the top-3 multilingual countries and Portugal, Italy and Hungary are the bottom-3.

The EU report also looks at other proxies for multilingualism. The most important secondary proxy is the percentage of people who can communicate online in English, French, German, Spanish and Russian. This data is given per country. It can possibly be used to relate our Twitter statistics to general ability to hold conversations in a language. Of the population 39% can communicate online in a foreign language. Latvia, Luxembourg and Estonia are the top-3 and Ireland, Hungary and Portugal are the bottom-3.

3. MATERIALS AND METHODS

Our data consists of 73.5M tweets collected between January 12, 18:38:29 CET and January 22, 10:21:03 CET, with some downtime in between due to program crashes. We gathered tweets from the Streaming API with a filter that selects tweets that fall within the [[-180,-80],[180,80]] bounding box[2]. Since this is (almost) the entire world we get all tweets that have location data. This yields around 10M tweets per day, which is around 2% of all tweets[3]. This is in line with the expectation that only a low percentage has geolocation data, but it's enough due to the large total volume. We were also provided with a sample set of 66.7M tweets regarding the FIFA World Cup 2014. In comparison to the dataset that we collected 2% of 66.7M tweets suddenly seems very small (1.3M) and so we chose not to use that data. Twitter attempts to automatically detect the originating country for tweets that have geolocation data. In our dataset the country code is given for 99.75% of the tweets. The dataset is 198GB when not compressed.

For this research, the CTIT cluster was made available to us. It has 48 hadoop workers[4] with the Hadoop filesystem and Apache Pig platform installed.

We analysed the collected data through Pig Latin scripts. The first round of parsing was performed using the Elephant Bird package on GitHub. This package was able to correctly convert the JSON structure used by Twitter into a Pig Latin-readable structure. Subsequently, we performed several operations on the data to obtain the tweeted languages per country. Most importantly, for each tweet we looked for the `lang` (language) attribute in the tweet's root and the `country_code` attribute in the tweet's `place` array. This data is combined to gain the number of tweets per country and the number of tweets per language per country.

In order to make the raw data easier to interpret, we used a modified Google GeoChart[5]. This allows us to show an interactive map of the world using tweet intensity and the distribution of languages per country. The tweet intensity is the 10-log of the amount of tweets in that country for the sample period, and the distribution of languages is represented using a pie chart created with the JFreeChart Java package.

All code and the visualisation are available through our github project at
`https://github.com/jjbankert/ManagingBigData`.

4. RESULTS AND DISCUSSION

With our data represented in a GeoChart, our results can be interpreted more efficiently and they can provide more insight into specific countries. Most of these results were as expected, but some countries gave us relatively unexpected results. A sample of our map can be seen in figure 1.

As can be seen on the GeoChart, there is a huge difference in Twitter usage between countries. We see that the four countries which tweet the most (United States, Brazil, Argentina and Indonesia) have made a bigger combined contribution (37.2M) than the other 241 countries combined.

There are 14 countries of which we gathered more than 1M tweets and 108 of which we gathered more than 10k tweets. In the other 137 countries we gathered less than 10k tweets. The Wald method for binomial distribution tells us that we can estimate proportions within 1% of the actual value with 95% certainty (2 standard deviations) when we have 10k tweets. This improves to within 0.1% with 95% certainty

[1] `https://blog.twitter.com/2013/introducing-new-metadata-for-tweets`

[2] `https://dev.twitter.com/streaming/overview/request-parameters#locations`

[3] `https://blog.twitter.com/2013/new-tweets-per-second-record-and-how`

[4] `https://fmt.ewi.utwente.nl/redmine/projects/ctit_user/wiki`

[5] `https://developers.google.com/chart/interactive/docs/gallery/geochart`

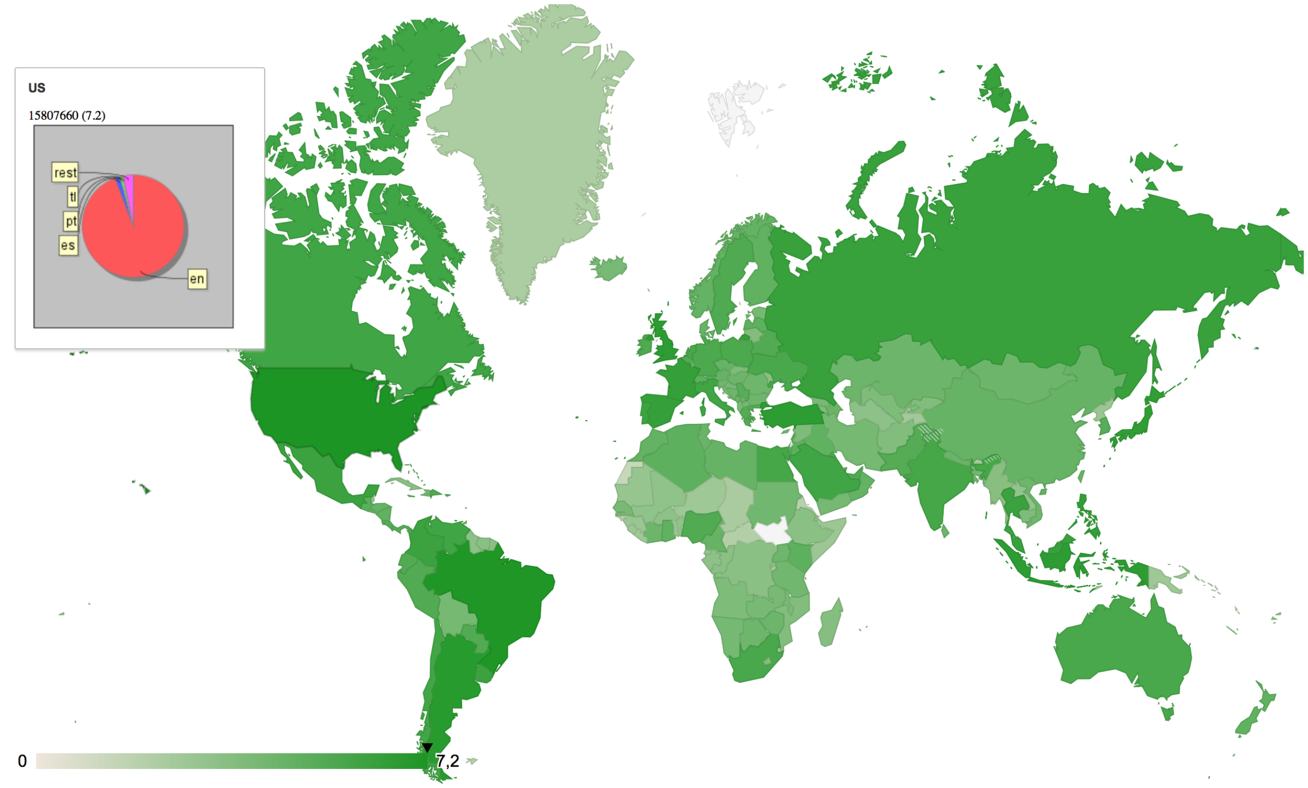


Figure 1: The map, targeted at the United States for a language analysis.

for countries with 1M tweets. The Wald method of course does not account for nonzero mean systematic errors.

4.1 Noteworthy results

Some countries gave a different distribution of languages than we anticipated, as we can see in figure 2:

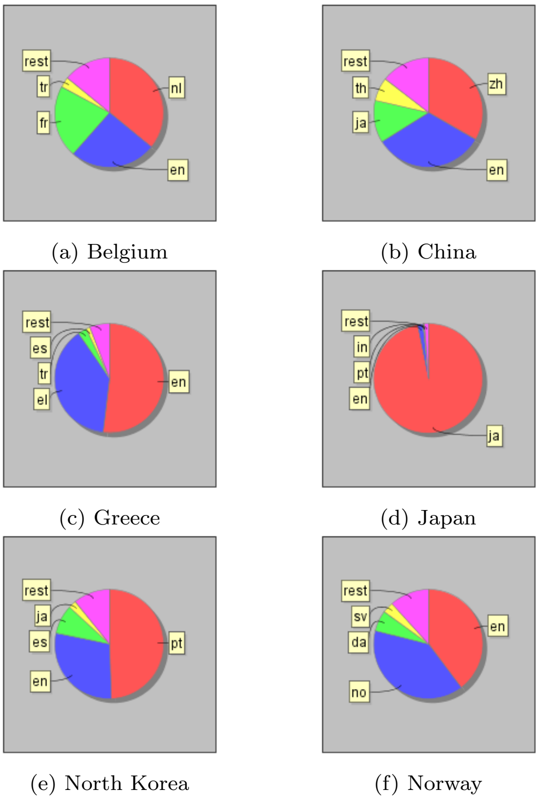
Belgium Although Belgium has three official languages (Dutch, French and German), English was the second most tweeted language (26%) with Dutch at 36% and French at 21%.

China The Chinese language (all versions) was nearly surpassed by the English language in China: 33.3% for Chinese versus 32.6% for English. This could be related to the Great FireWall of China[6], where Twitter can be blocked for all users in China, as we only recovered 30.329 tweets form China.

Greece Where 39% of the tweets were in the native Greek language, 52% were written in English, a big difference between the official and the most used language.

Japan 97% of the tweets from Japan were in Japanese, the highest amount for a language in a country with a sizeable (more than 2000 tweets) data set.

North Korea With Korean as its official language, it was quite a surprise to see only 0.4% (3 tweets) in Korean here. The most used language in North Korea appeared to be Portuguese at 49%. English is also quite



(a) Belgium (b) China (c) Greece (d) Japan (e) North Korea (f) Norway

Figure 2: Language distribution in a couple of countries.

[6] http://en.wikipedia.org/wiki/Golden_Shield_Project

large at 29%. We only collected 628 tweets from North Korea in our time frame, therefore the margin of error is quite large in comparison to other countries. Lack of free access to Twitter or spoofing location data could also play a part in this.

Norway Another one of the few countries where one of its official or de facto languages is not the most tweeted language: English beat Norwegian with 39.7% versus 39.3%.

United States For the US specifically we can estimate the multilingualism proportions to within 0.03% with 95% certainty. As a comparison, the US Census data is accurate to 'only' ± 0.1% [1]. We found that 94.28% ± 0.03 of tweets are in English and 1.54% ± 0.03 in Spanish. This is quite different from the US census data with states that 79.2% ± 0.1% speaks English and 13.0% ± 0.1% speaks Spanish at home.

4.2 Discussion

We have only accounted for random errors. Systematic errors caused by vacation time and sudden inactivity could be filtered out better by collecting samples over a longer period.

With the terrorist attacks in Paris several days prior to the start of our data collection, the commonly tweeted phrase "Je suis Charlie"[7] could have been detected as a French tweet, but there was no hard evidence this polluted our data set, as French never came up as a frequent language in countries where it was not spoken regularly.

The EU Commission report focuses on which languages a person can use, in stead of which one he uses most. This meant that it was hard to make a suitable comparison with our data.

5. CONCLUSIONS

The use of big data to analyse languages on Twitter can be used to efficiently determine the language distribution in a country. Most countries also adhere to their official or de facto languages, but with others, English turns up as a bridging language, possibly to other countries. When excluding the English language from the data set, nearly all countries follow expected behaviour.

6. FUTURE WORK

A research for the unique languages a user speaks per country (say: 90% speaks English and 50% speaks German) is within possibilities, but would require different scripts to calculate. Building on that, a study for the most multilingual country can be performed.

A different research would be to expand this study to include more types of online media (e.g. Facebook posts), provided they can be accessed freely. With more channels of information, the surveys conducted periodically can eventually be eliminated by big data research.

7. REFERENCES

[1] U. S. C. Bureau. *American Community Survey.* 2013, `http://factfinder.census.gov/` [Online, accessed 8-Januari-2015].

[2] M. Graham, S. Hale, and D. Gaffney. Where in the world are you? geolocation and language identification in twitter. *Professional Geographer*, 66(4):568–578, 2014.

[3] D. Mocanu, A. Baronchelli, N. Perra, B. Gonçalves, Q. Zhang, and A. Vespignani. The twitter of babel: Mapping world languages through microblogging platforms. *PloS one*, 8(4):e61981, 2013.

[4] T. O. . Social. *Europeans And Their Languages.* European Commission, 2012, `http://ec.europa.eu/public_opinion/archives/ebs/ebs_386_en.pdf` [Online, accessed 8-Januari-2015].

[7] `http://en.wikipedia.org/wiki/Je_suis_Charlie`