URBANSCOPE: A LENS TO OBSERVE LANGUAGE MIX IN CITIES

Michela Arnaboldi[1], Marco Brambilla[1], Beatrice Cassottana[2], Paolo Ciuccarelli[1], and Simone Vantini[1]

1. Politecnico di Milano
2. University of Singapore

ABSTRACT

Cities of the XXI century are places where various actors interact, where physical systems, that are sometime geographically distant, are strictly dependent, where relational mechanisms become crucial, and where the boundaries between individual and collective, local and global, real and digital become more and more blurred. In this context, social media can be used as a digital lens to analyse the space and the territory of cities. In fact, they offer a great opportunity to individualise and understand the connections that might exist between different spheres. In this paper, we use Twitter to analyse the language mix of the city and to detect language communities within the city neighbourhoods. We then compare these "digital" communities, discovered through Twitter, with the "real" communities identified by the traditional census data. Milan, a city which is increasingly becoming an international melting pot, is chosen as a case study for this work.

URBANSCOPE: A LENS TO OBSERVE LANGUAGE MIX IN CITIES

1. INTRODUCTION

We need a tool, as precious as were the microscope and the telescope for the scientific knowledge of the universe, intended for all those who seek to understand the meaning and the place of their activities. We will call this tool the macroscope (macro, big; skopein, observe) (De Rosnay, 1977).

Exploring cities of the XXI century offers a great opportunity to understand the ever evolving modern society. In fact, cities are increasingly attracting new people, with half of the world's inhabitants living in urban areas (McKinsey & Company, 2013). Furthermore, cities are not only physical centres but also virtual hubs, where individuals and communities interact and exchange messages through social media. As a consequence of this dense network of interactions, a great amount of data, the so called Big Data (Batty, 2013), can be tracked. On the one hand, data created by the single identities of the city, i.e. inhabitants, are available, on the other hand, we need to capture the "big picture" by aggregating them. This situation allows us to observe both the activities at a micro-societal level and to draw the main features that characterise the city at a macro-societal level. Therefore, Big Data can be regarded as a lens to understand cities, or, using the words by De Rosnay (1977), as an urban "macroscope".

The exploration fields of the urban macroscope are infinite. Among all, one of the feature of interest for policy makers and cities managers (UN-HABITAT, 2016) is the extremely diversified composition of the language mix, or *multilingualism*. This interest is motivated by the increasing immigration flow towards cities (Sanderson, 2015), which results in rapidly changing population densities (Deville, et al., 2014). Multilingualism has also a broad scope in academia. In particular, different papers approach the issue of multilingualism from a historical perspective. Leimgruber (2013), for example, analyses the city of Singapore, García and Fishman (2001) the city of New York, Extra and Yagmur (2004) develop a cross-linguistic perspective on Gothenburg, Hamburg, The Hague, Brussels, Lyon and Madrid.

To this end, traditional data alone do not suffice in providing comprehensive information (Deville, et al., 2014). For instance, census data are aggregated at a macro-level and do not consider the inhabitants without official residency status (Extra & Yagmur, 2004), such as tourists and temporary residents. Therefore, they fail to capture a realistic picture of the society at a micro-level. Moreover, they are not updated in real-time, thus failing to capture the dynamic nature of the city. In order to fill this gap, we use social media as a powerful alternative data source. Big Data deriving from social media overcome

the trade-off between level of aggregation (Jiang, 2013) and time-horizon of the analysis (Arnaboldi & Coget, 2016), thus enabling us to continuously monitor a certain feature of the city (Ceron, 2014).

In this paper we build an urban macroscope to understand multilingualism in the city, more specifically we use Twitter to analyse the language mix of the city and to capture language communities within the city neighbourhoods. Throughout the paper, we refer to language communities as those groups of individuals sharing either the language used on Twitter or the language of their country of origin. We then compare these "digital" communities, discovered through Twitter, with the "real" communities identified by the traditional census data. Milan, a city which is increasingly becoming an international melting pot, is chosen as a case study for this work. We use quantitative tools to analyse the micro-level data collected from individuals, but we also develop visual solutions to show and navigate the aggregated results. Our dashboard is called Urbanscope, as the fusion the words "urban" and "macroscope". More specifically, the paper aims at: (1) describing the construction methodology of Urbanscope, and (2) applying this lens to city of Milan at a neighbourhood level, comparing the multilingual composition resulting from the analysis of the social media with the traditional picture provided by traditional data.

Our research could benefit both scholars and practitioners. It benefits the first group by offering an example of integration of Big data and traditional data. In fact, the Urbanscope application is the result of a complex disciplinary integration. On the other hand, it benefits city managers and policy makers by making the results of the analysis accessible through the online interface of Urbanscope, and therefore supporting decision-making processes.

To present our argument, the rest of the paper is organised as follows. The next section [2] contains the analysis of the relevant literature. Section [3] shorty introduces the reader to the research project and context, that is, the city of Milan. The methodology is described in section [4] and in section [5] the results are presented. Finally, the research questions are re-iterated in the conclusion [6], highlighting the limitations and advancing suggestions for further research.

2. MULTILINGUALISM IN CITIES

For the purpose of this paper, two streams of literature are important. First, the study of multilingualism on Twitter is relevant in terms of the methodology used to aggregate the micro-level data. This literature also helps us in the interpretation of the results. Secondly, the benchmark with different dashboards is useful in designing the aggregated information to be visualized, and represents the final step from the micro-level data to the macro-level information.

2.1 Building the micro-level multilingualism of Twitter

Language distributions in Twitter have been covered in many papers. The most recurring result is the predominance of English among the languages used to communicate on Twitter (Eleta & Golbeck, 2014; Hong, Convertino, & Chi, 2011; Mocanu, et al., 2013; Takhteyev, Gruzd, & Wellman, 2012). Different are the mathematical techniques available to analyse the tweets, and their utilization depends upon the scope of the study. For example, some papers are concerned with analysing the different language communities from a network perspective. Eleta and Golbeck (2014) use logistic regression to demonstrate that multilingual users are located between language groups. In the same stream of literature, Kim, Weber, Wei, and Oh (2014) quantify the distribution of a language in a certain region using the Shannon entropy. Takhteyev, et al. (2012) explain the Twitter network in terms of number and type of ties existing between users.

Other authors are concerned with understanding the factors affecting the user influence on Twitter. For this purpose, different rank methods are used to build influence tree relative to various accounts. Bakshy, Hofman, Mason, and Watts (2011) measure influence as the number of times a tweet is reposted by friends, friends of friends, and so on, which is a similar method to the Google PageRank. Ravikumar, Balakrishnan, and Kambhampati (2012) measure the popularity of a tweet as a function of the number of reposts that show agreement with it. Finally, Cha, Haddadi, Benevenuto, and Gummadi (2010) assess the user influence as an aggregated index of multiple influence indicators.

Finally, a section of literature deals with behavioural differences related to language in Twitter. Hong, et al. (2011), for example, find that different language communities use Twitter for different purposes. A similar distinction is made by Kim, et al. (2014) with regards to the language chosen by multilingual speakers.

2.2 Building the macro-level vision

The rise in available urban Big Data has encouraged city governments, academia and private companies to start investing in projects to collect, analyse and visualise the vast amount of information being created through urban services (Lee, Felix, He, Offenhuber, & Ratti, 2015) and by the users themselves on social media. Many of these initiatives are shared with the public, enabling people to access the data which is displayed through online, dynamic and interactive systems, often defined as “city dashboards” (Kitchin, 2014). A seminal group of city dashboards are used for crime-related questions. One of the first dashboards ever conceived is CompStat, the crime map model for aggregating and mapping crime statistics in the city of New York (Mattern, 2015). Following the example of the city of New York, the city of Baltimore, Maryland, also implemented a dashboard called CitiStat used to reduce crime by tracking and mapping crimes through a GIS (Geographic Information System) application (Gullino, 2009). Similarly, ChicagoCrime.org allows users to search

for specific types of crime in a certain location and date (Kraft, 2007). These dashboards, despite focusing on a single issue, introduced the interesting feature of displaying geo-referenced data of the city.

In the same main stream, another group of dashboards has more recently been developed around environmental issues; examples are in Amsterdam (AMS Institute[1]), Barcelona with the City Eye (Lee, et al., 2015), London (City Dashboard[2]), Rio de Janeiro (Kitchin, 2014). These dashboards introduced the effective feature of data displayed and visualised within the city area, but they concentrate on the functions and services of the city, such as transport, energy and safety, ignoring aspects such as demographics and multilingual variations.

The demographic element sometimes appears in a second type of dashboard, which is based on the balanced scorecard frameworks proposed for corporations (Kaplan, 1992). The balanced scorecard was first introduced to provide an overall and precise way for companies to monitor their performance, according to four perspectives: financial, customer, internal processes, learning and growth. Inspired by this model, cities have developed dashboards with a focus on various areas, where demographic data started to enter (Edwards & Thomas, 2005, DublinDashboard[3], Open Michigan MiDashboard[4]).

Summarising, there are several examples of dashboards, but only limited attention has been spent on analysing demographic data and, in particular, studying multilingual evolution within cities. When demographic analyses are present, they are based on traditional methods, such as census data and surveys, which are unable to monitor cities on a timely basis. Furthermore, these visualisations are based on a balanced scorecard approach, with no spatial reference to city sub-areas. Our work enhances these previous researches combining the multilingualism analysis of Twitter with the development of a city dashboard.

3. RESEARCH SETTING: URBANSCOPE AND MILAN

The study presented in this paper is carried out within a larger interdisciplinary project called Urbanscope[5], which involves monitoring the digital traces in the urban context. Urbanscope acts as a tool to obtain added-value information from raw data and displays this data on an online dashboard.

Since the experimentation of this paper is centred around the city of Milan, we shortly introduce the reader to its historical background. The profile of the city of Milan has been changing at an exponential speed in the last few years. One of the main feature that characterises the city is its internationality, which had its maximum manifestation in EXPO, the universal exposition held from May until October

[1] http://www.ams-institute.org/
[2] http://www.citydashboard.org/london
[3] http://www.dublindashboard.ie
[4] https://midashboard.michigan.gov
[5] http://www.urbanscope.polimi.it

2015 in the neighbourhood called Cascina Triulza – Expo. The international profile of the city does not only result from its touristic attractiveness, but it is rooted in its history of immigration. In fact, the city has seen a mass immigration over the last 40 years especially from North Africa and South America, corresponding to the Arabic and the Spanish language communities, respectively. This immigration wave reached its peaked in the 1980s, 1990s and early 21st century. This phenomenon does not merely concern Milan, but the entire country. In the period 2001–06, for example, Italy saw the largest growth in the legally resident immigrant population from 1.3 million in October 2001 to 2.67 million in January 2006 (Geddes, 2008). Nonetheless, after 2006, the quota of foreign citizens without a regular permit started to increase, reaching 12.8% in the city of Milan in 2011 (Costa & Sabatinelli, 2012). The new comers populated mostly the peripheral areas of the city. At the same time, Milan's residential population began to decline and age. Milan has nowadays a negative birth rate, which is offset by the immigrant population.

4. RESEARCH METHODOLOGY

In order to study multilingualism in the city of Milan, the research has been divided into four phases, as shown in Table 1. As highlighted in the last column, these phases are not sequential, but overlap time-wise. The first phase is the preliminary analysis of the data sources and the roles. Two main data sources are identified as relevant. First, the dataset provided by the Municipality of Milan[6] regarding the geographical zones in which the city is divided, which are addressed by the city managers as NILs (Nuclei di Identità Locale). This dataset provides demographic data alongside data concerning various services for each NIL. The second dataset derives from Twitter. Twitter is considered the most suitable source for the purpose of our analysis, since it is largely based on written text and features the option of geo-locating the posts (although only a very limited share of users actually opt for using this feature). The second phase of the research consists of data downloading. Data are downloaded through the public API (Application Programming Interface) provided by Twitter. We downloaded the complete payload of each tweet, including the main text and a large set of metadata, spanning geographical data, tags, mentions, images, links, timing and language. For this study, we consider only the tweets geo-tagged within the boundaries of the municipality of Milan for the period August 2014-December 2015. It should be pointed out that not all the tweets are tagged with geolocation; therefore, the ones missing that piece of information are not considered in our study. Recent analyses show that the share of geolocated tweets in Italy has been consistently around 7% in the last years[7]. Notice that geolocation on Twitter is a controversial point. In particular, recently Twitter has applied a new policy in the user interface of its mobile application, which now asks the user to explicitly opt in for the precise

[6] http://dati.comune.milano.it/

[7] https://www.statista.com/statistics/592882/twitter-geolocation-use-italy/

geotagging for every new post. Alternatively, only the tagging with the city-level generic location is applied, even in case geotagging is selected. For our study only the tweets with precise geotagging are useful.

The language used in the tweets is part of the metadata returned by Twitter. Twitter's language detection algorithm elaborates the language automatically. Tweets containing words written in different languages or whose language could not be detected are classified as "undefined language" and are excluded from our dataset. The entire dataset contains 1,109,693 tweets, with 1,007,314 being associated to a defined language. There are 793,838 tweets whose metadata links them precisely to one of the 88 NILs within the Milan municipality area. No further data cleaning or filtering has been applied to the retrieved tweets before the analysis. The only cleaning operation we performed is related to how tweets are displayed on the user interface, where all tweets containing explicit material, vulgarity, pornography, or violence have been removed. The reason for this is to avoid public display of such contents. However, notice that this kind of filtering is applied only in the user interface when the content of the tweet is displayed. Such tweets are still considered and contributing to the statistical analysis and to the display of the aggregated values to the user. Therefore, this filtering doesn't affect the results of the quantitative analysis.

The third and four phase, mathematical analysis and data visualisation, are detailed further below as part of our results.

Table 1 Phases of the study

Phase	Output	Duration	Owner
Preliminary analysis of data sources and roles	List of data available, with their respective accessibility and utility within the research	January 2014 – November 2014	Performance Management Group
Data Download	Dataset of tweets to be analysed divided into NILs and identification of language used in each tweet	June 2014 - Ongoing	Computer Science Group
Mathematical Analysis	Identification of transparent NILs and classification of non-transparent NILs in terms of the most popular language used	October 2014 - July 2015	Mathematics Group
Data Visualisation	Creation of displays showing the analysis on the online dashboard Urbanscope	January 2015- July 2015	Design Group

5. RESULTS

Results are twofold. First, we illustrate how the Urbanscope tool combines mathematical analysis and visualization techniques to extract information from Big Data. Second, we describe and interpret the results in connection with the city of Milan.

5.1 The Urbanscope lens

The analysis of Big Data is a challenge that is addressed by researchers with both calculative solutions and visualisation techniques.

5.1.1 Calcultative solutions

The calcultative solution consists of three steps:. First, the definition of the proper level of data aggregation, with respect to time, space and language. Second, the elimination process, which consists of identifying and eliminating those NILs for which the sample size is too limited to infer any statistically significant conclusions (i.e. transparent NILs). Third, the classification of the NILs. in terms of the most popular spoken language on Twitter (i.e. coloured NILs).

With respect to data aggregation, we aggregate data on months, NILs and language. The proper level of aggregation should be high enough to provide a signal-to-noise ratio sufficiently large to estimate the number of tweets reliably, and should be low enough to monitor linguistic variations over time and space. In order to be consistent with the census data, we fix the space-related aggregation at the NIL level and we vary the dimensions of time and language. More technically, we carry out a set of Fisher's exact tests to assess the stochastic independence between NILs and language for each time unit candidate (i.e. months, two-month periods, quarters...). In this first analysis, each tweet is considered as an instance of two categorical random variables, namely the NIL and the language of the tweet. We iterate the same procedure to test the stochastic independence between NIL and time units for each language candidate and find that an optimal aggregation setting for the statistical analysis consists of quarter time units and English, Italian and other languages as aggregated macro-languages.

In the elimination process, data are considered as instances of a 3D multinomial random variable defining the number of tweets in English, Italian and other languages for a NIL in a quarter. The number of tweets in one of the three macro-languages for a given NIL and quarter was modelled as a binomial random variable $B(n, p)$, where, given a specific NIL in a specific quarter, p is the probability that the macro-language of a tweet is the language selected, and n is the total number of tweets. In this framework, p is estimated as the observed macro-language percentage. Its standard deviation is upper-bounded by $1/(2\sqrt{n})$. This bound allows us to determine the minimum number of tweets required for each NIL in each quarter. Specifically, we require that each NIL should have a maximum standard deviation of 5% (over all quarters), which corresponds to at least 100 tweets for each quarter.

The classification process identifies NILs with anomalously large numbers of tweets counted for at least one of the three macro-languages in a given quarter (compared to the other NILs for the same quarter). In particular, we group non-transparent NILs into three categories, according to their macro-language percentage of tweets counted compared to the other NILs for the same quarter. In the displays (see section 5.1.2), those NILs with a large macro-language percentage of tweets counted are darkly

coloured, those with a moderate percentage are lightly coloured, and those with a small percentage are grey coloured. We use a standard outlier detection method, namely the Box-and-Whisker Plot (Murrell, 2005), to formally define the three categories above. According to this method, for each macro-language in each quarter, 75% of non-transparent NILs are grey and 25% lightly or darkly coloured.

5.1.2 Visualisation solution

The Urbanscope online platform is composed of four main modules, which are used to investigate different aspects of the phenomena analysed. The Urbanscope entry-point is a homepage giving a brief description of each module (see Figure 1) .Looking at the online dashboard Urbanscope, the section concerning the multilingualism analysis of Twitter for the city of Milan is called “Cities within Cities” and is, in turn, divided into two visualisations: Explore Tweets and Analysed Tweets. The visualisation called “Explore Tweets” displays all the tweets posted within the city of Milan since Urbanscope was first set up (see Figure 2). In the map, supported by Openstreetmap, NILs are shaded according to the density of the geo-referenced tweets posted. The interactive map displays data and information very simply. By surfing from the initial visualisation page, the user can select time window, macro-language (English, Italian, other languages) and can easily identify the NIL by moving the cursor on the map. If the user selects “Other” from the macro-language option, details are given about the language distribution for each NIL (see Figure 3).

Figure 1 Home page

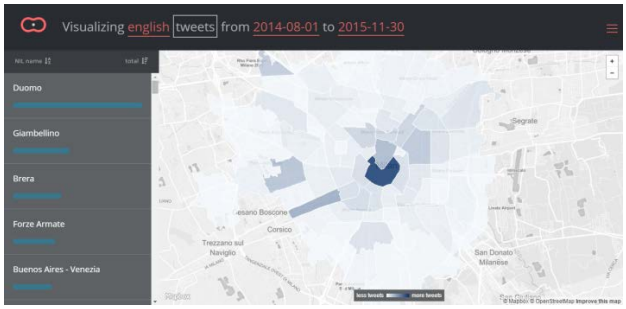

Figure 2 Explore Tweets. NILs are colored according to the density of tweets (dark color indicates high density)

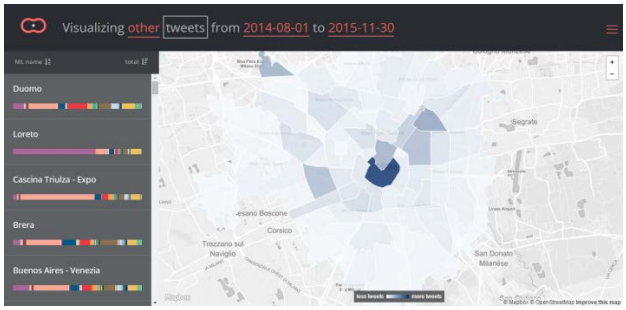

Figure 3 Explore Tweets, Other languages. NILs are colored according to the density of tweets (dark color indicates high density)

The visualisation called “Analysed Tweets” gives details about the density of tweets in each NIL for each quarter. Three maps are displayed, with the NILs are coloured in red, blue or yellow according to whether English, Italian or another language is predominant. The intensity of the colour is defined in section 5.1.1.

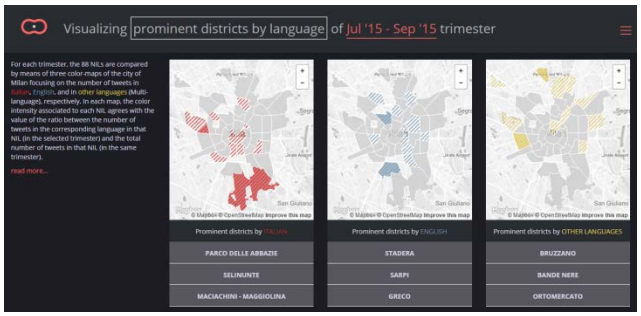

Figure 4 Analysed Tweets. The color of the NIL indicates the predominance of tweets written in English (red), Italian (blue), or another language (yellow). The intensity of the color is proportional to the density of the tweets written in the respective language calculated with the Box-and-Whisker Plot outlier detection method (for each language, 75% of non-transparent NILs are grey and 25% lightly or darkly coloured).

5.2 The application of Urbanscope to the City of Milan

While it is easy to explain the prevalence of Italian tweets as Italian speakers are predominant in many NILs, we are mostly interested in understanding where and why English or other languages are

prevalent elsewhere. Unsurprisingly, the NILs that come next are those where English is the prevalent language. In accordance with literature, English is widely adopted because it is perceived as counting more within the Twitter community (Kim, et al., 2014), which means that it is difficult to recognise the nationality of those posting in English. More interesting are the NILs where a language other than English or Italian is prevalent, which suggests that the last group of NILs is frequented by mostly non-native Italians.

Figures 5.a-5.b show the representation fo language, where for each macro-language in each quarter, 75% of non-transparent NILs are grey and 25% lightly or darkly coloured. As specified before, for each NIL to appear in the analysis we require maximum standard deviation of 5% over all quarters, i.e., at least 100 tweets per quarter.

Figure 5.a October 2014-December 2014

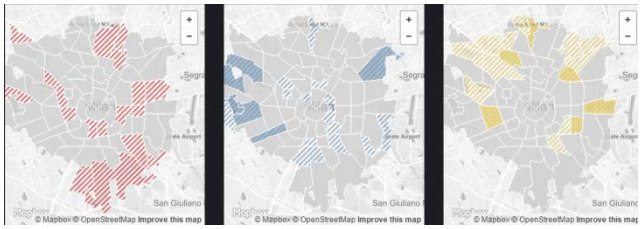

Figure 5.b January 2015-March 2015

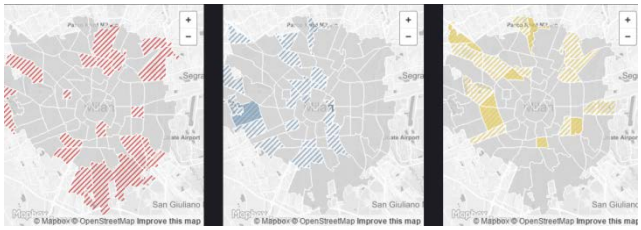

Figure 5.c April 2015-June 2015

Figure 5.d July 2015-September 2015

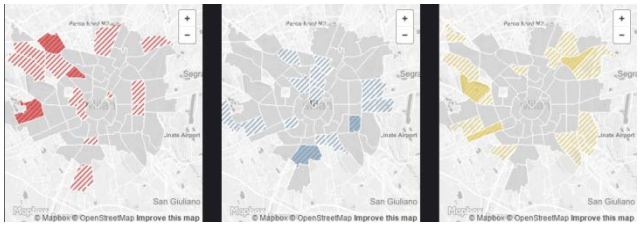

Figure 5.e October 2015-December 2015

Figure 5. Analysed Tweets visualizations in the five quarters. The color of the NIL indicates the predominance of tweets written in English (red), Italian (blue), or another language (yellow). The intensity of the color is proportional to the density of the tweets written in the respective language calculated with the Box-and-Whisker Plot outlier detection method (for each language, 75% of non-transparent NILs are grey and 25% lightly or darkly coloured).

It is interesting to note that the NILs where the main languages are neither English nor Italian are more likely to be located in the outskirts rather than being central. This is probably due to the historical settlement of immigrants into the suburb areas. Italian is used in the southern NILs (e.g. Parco delle Abbazie and Ticinello) for most of the months analysed and English in the central NILs (e.g. Parco Sempione, Sarpi and Guastalla), which correspond to the most attractive destinations for tourists.

Figure 6 shows the map of Milan, where the NILs are coloured on the basis of number of tweets posted in that NIL throughout the period of analysis, from August 2014 to December 2015. Duomo is the NIL with the highest number of tweets, followed by Loreto, Cascina Triulza – Expo, Brera and Buenos Aires – Venezia.

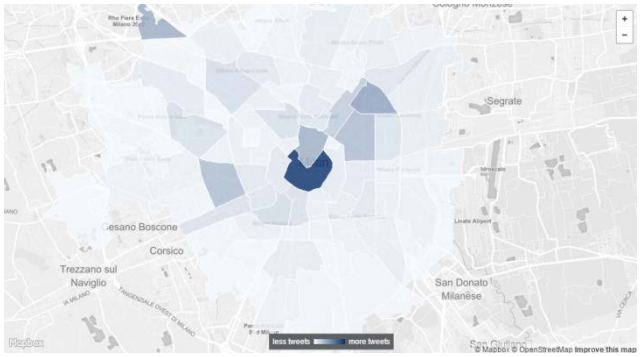

Figure 6 NILs shown according to number of geo-referenced tweets

In order to detect the international communities within the city of Milan, we can explore the NILs further by disaggregating them into their various languages (excluding English and Italian) and considering the entire period of analysis (August 2014-December 2015). Table 2 gives the relevant summarised results.

In Trenno, Loreto, Umbria – Molise, Ortomercato, Parco Forlanini – Ortica, and Quarto Oggiaro there is a prevalence of tweets written in Arabic. Spanish tweets make up a consistent percentage of non-

English and non-Italian tweets for many NILs (above 20% for most of the NILs), indicating that the Spanish Twitter community is more spread out within the city of Milan.

Nevertheless, some NILs are shown to have a strong Spanish influence, like Quintosole, where 100% of all tweets are in Spanish. The detection of Arabic and Spanish as the most diffuse foreign languages can find its explanation in the immigration waves from North Africa and South America that affected Milan over the last 40 years. Other commonly used languages in the NILs are Tagalog (a language mostly spoken in the Philippines), Portuguese, Indonesian and Turkish. Some NILs in particular differ from the others in being the only place where a certain language is used, meaning that tweets written in that language make up over 50% of the tweets written in languages other than English and Italian.

Table 2 Relevant Results

NIL	***% of Tweets (***tweets not in English or Italian***)***
Trenno	72% Arabic
Loreto, Umbria – Molise, Ortomercato	64% Arabic
Parco Forlanini - Ortica	63% Arabic
Quintosole	100% Spanish
Parco dei Navigli	69% Spanish
Gallaratese	66% Spanish
Cascina Triulza- Expo	57% Spanish
Ex Om – Morivione	56% Spanish
Mecenate	51% Spanish
Padova	59% Tagalog
Giambellino	57% Tagalog
Villapizzone	56% Tagalog
Bruzzano	59% Portuguese
Parco Nord	55% Dutch
Chiaravalle	75% Norwegian

5.2.1 Comparison between Twitter Results and Census Data

According to the census data for the municipality of Milan, recorded as at December 31st 2014, there were 253,334 foreign people residing in Milan, about 18.9% of the total population. Among them, the greatest community is that of Filipinos (41,237 people), followed by Egyptians (35,597 people),

Chinese (25,928) and Peruvians (20,462)[8]. Figure 8 reports the top countries (as percentage of foreign residents).

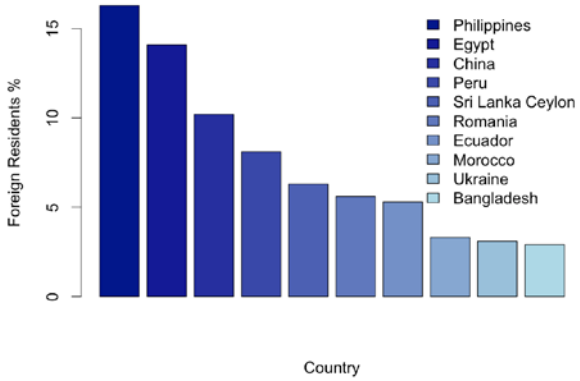

Figure 8 Foreign people residing in Milan by country

By assigning the appropriate language to each NIL according to the census data, we can compare Twitter results with traditional data. The columns in Table 3 give the percentages of non-Italian residents (grouped by spoken language) for languages spoken by at least 20% of the non-Italians in at least one NIL. The rows give only the NILs where at least one of these languages is spoken by at least 20% of the non-Italian residents. The first column gives the main language used on Twitter in that NIL. The cells highlighted in yellow indicate the largest non-Italian community in each NIL.

As shown in the correlation analysis displayed in Figure 9, only in some cases does the language community detected through Twitter correspond to the majority of non-Italian residents in any given NIL. This is the case for some of the Arabic and Spanish communities. With respect to the correlation line drawn in Figure 8, we note that Portuguese, Dutch, Norwegian and Albanian are underexposed. Tagalog, Ukrainian and Romanian are obviously overexposed, since these language groups are not apparent in the census data. The most noticeable overexposed languages are Arabic and Spanish. These language communities might consist of those generations descending from North African and South American immigrants in the 1980s and in the 1990s. In fact, while the new generations have acquired Italian citizenship, they might have maintained a double language identity and might use their original language to communicate within their community.

[8] http://allegati.comune.milano.it/Statistica/Popolazione/Stranieri%202014/10naz%20prev%202014.pdf

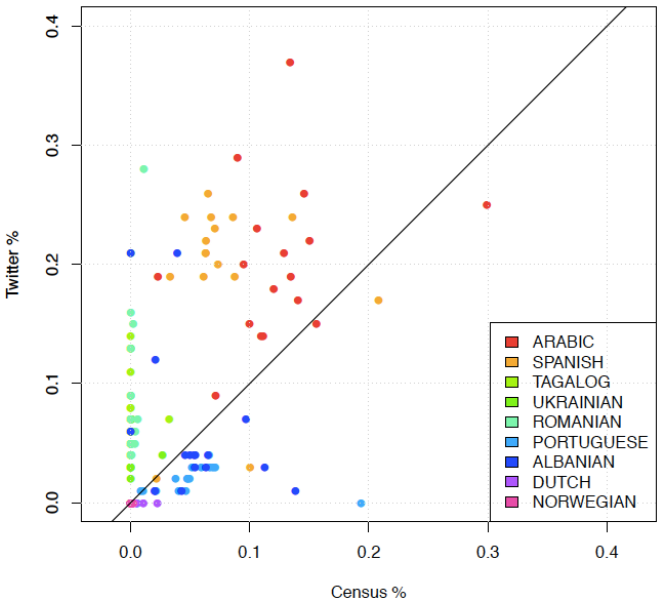

Figure 9 Correlation analysis: Twitter vs. official residents

Interesting fact, in Sarpi, which is typically seen as a Chinese neighbourhood, there is not a prevalence of tweets written in Chinese. This is probably because Twitter is not used in China (Mocanu, et al., 2013), or rather, it is blocked there and replaced by Sina Weibo (www.sina.com.cn) (Crampton, 2011).

6. CONCLUSIONS

In this paper, we use Twitter and official Census data to analyse multilingualism in cities and the to compare the "digital" language communities with the "real" language communities.

The finding of this study are related to two areas. The first is the development of the Urbanscope lens and methodology, through which we merge refined calculative practices with a communicative visual tool. Previous studies focused on one of the two aspects, without considering the potentiality to render more articulated information available to the large public. Lens, such as Urbanscope, offer to everyone the possibility to observe the micro and the macro aspects of our digital traces, where each individual becomes actor of the information used to build the lens. Urbanscope has the potential to be used for experimenting relational sociology in the social media age (Fourcade, 2007).

The second area of results is the related to the application of the lens to the City of Milan and the detection of "digital" language communities within the neighbourhoods of the city of Milan. The divergence we discovered between the prevalent language in any NIL detected through Twitter and through census data clearly points out the weaknesses and strengths of the two data sources, supporting the need for (and wishing for) a future fruitful integration between social network data and census data in language distribution studies. For example, Twitter is able to intercept (in part) those who are tourists, city-users or residents (although further and deeper analysis would be needed to distinguish between these categories). On the contrary, the census data is not able to intercept tourists and city-users, but does recognise the totality of residents. Twitter data vary with time while census data are

static (at least within each year). Twitter data are biased by the different rate in which communities adopt the social networking platform and by the general tendency to use languages of a wide spread appeal (ie English) rather than their native language. Census data are biased by the presence of non-recorded residents (e.g. tourists). In addition, as the data is based on the residents' country of origin it is not possible to infer from this data which languages are spoken by the residents, especially when coming from multi-language countries (i.e. Belgium, Canada...) or when residing in the country for a long time (in which case they could have switched to the local language). Data obtained from social media need to be treated with caution. The information must be interpreted carefully, especially when trying to look at aspects that are characteristic of specific areas of the city. The population of social media users differs from the actual population, and generalising the results to the whole city would have to be validated through several different sources of data. It is important to remember that, since our study uses Twitter as a lens through which to study the multilingual distribution of Milan, the results shown are affected by all the inherent limitations of this choice, such as Twitter being adopted at different rates by different populations. A future development of this paper is to study how the language identities of the different neighbourhoods of Milan change over time, recognising the seasonal patterns of language distribution.

Table 3 Twitter language communities vs. non-Italian residents

	Main language used on Twitter	Percentages of non-Italian residents (grouped by language)									
		Arabic	Spanish	Tagalog	Ukrainian	Romanian	Albanese	Norwegian	Portuguese	Dutch
Trenno	Arabic	29%	19%	7%	6%	5%	1%	0%	1%	0%
Villapizzone	Tagalog	14%	23%	6%	4%	5%	4%	0%	3%	0%
Umbria - Molise	Arabic	22%	19%	7%	3%	5%	3%	0%	2%	0%
Quintosole	Spanish	37%	3%	6%	0%	13%	21%	0%	0%	0%
Parco Nord	Dutch	21%	0%	11%	21%	4%	7%	0%	0%	0%
Parco Forlanini - Ortica	Arabic	23%	21%	8%	7%	15%	1%	0%	2%	0%
Parco dei Navigli	Spanish	9%	2%	0%	9%	28%	21%	0%	2%	0%
Padova	Tagalog	14%	26%	6%	4%	5%	4%	0%	3%	0%
Ortomercato	Arabic	25%	24%	14%	2%	7%	1%	0%	1%	0%
Mecenate	Spanish	18%	24%	8%	5%	6%	3%	0%	2%	0%
Loreto	Arabic	15%	22%	5%	3%	4%	4%	0%	3%	0%
Giambellino	Tagalog	19%	20%	6%	3%	5%	4%	0%	3%	0%
Gallaratese	Spanish	17%	21%	7%	5%	7%	4%	0%	3%	0%
Ex.OM - Morivione	Spanish	15%	24%	13%	5%	5%	3%	0%	4%	0%
Chiaravalle	Norwegian	20%	19%	13%	4%	9%	1%	0%	1%	0%
Cascina Triulza - Expo	Spanish	26%	17%	3%	0%	16%	12%	0%	1%	0%
Bruzzano	Portuguese	19%	24%	7%	4%	7%	6%	0%	1%	0%

REFERENCES

Agostino, D. (2013). Using social media to engage citizens: a study of Italian municipalities. *Public Relations Review*, 39 (3), pp.232-234.

Armstrong, A., & Hagel, J. (2000). The real value of online communities. *Knowledge and communities*, 74(3), 85-95.

Arnaboldi, M., & Coget, J. F. (2016). Social media and business. *Organizational Dynamics,* (45), 47-54.

Bakshy, E., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011). Everyone's an influencer: quantifying influence on twitter. *ACM International Conference on Web Search and Data Mining,* Hong Kong, 2011. ACM Press.

Batty, M. (2013). Big data, smart cities and city planning. *Dialogues in Human Geography*, *3*(3), 274-279.

Boyd, J. (2002). In community we trust: online security communication at eBay. *Journal of Computer Mediated Communication*, 7(3), 210-230.

Ceron, A. C. (2014). Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. *New Media & Society*, *16*(2), 340-358.

Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, P. K. (2010). Measuring User Influence in Twitter: The Million Follower Fallacy. *International AAAI conference on Weblogs and Social media,* Whashington, DC, 2010.

Costa, G., & Sabatinelli, S. (2012). City Report: Milan. Polythecnic of Milan, Italy. *WILCO Publication* no. 23.

Cova B, & Pace S. (2006). Brand community of convenience products: new forms of customer empowerment — the case "My Nutella" community. *European Journal of Marketing;* 40(9/10), 1087–105.

Crampton, T. (2011). Social Media in China: The Same, but Different Participating in China's unique and diverse social media environment is key to winning over online consumers. *China Business Review*, *38*(1), 28.

De Rosnay, J. (1977). *Il macroscopio: verso una visione globale*. Bari, Italy: Dedalo.

Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F. R., Gaughan, A. E., . . . Tatem, A. J. (2014). Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences*, *111*(45), 15888-15893.

Edwards, D., & Thomas, J. C. (2005). Developing a Municipal Performance-Measurement System: Reflections on the Atlanta Dashboard. *Public Administration Review*, *65*(3), 369-376.

Eleta, I., & Golbeck, J. (2014). Multilingual use of Twitter: Social networks at the language frontier. *Computers in Human Behavior*, *41*, 424-432.

Extra, G., & Yagmur, K. (2004). *Urban multilingualism in Europe: Immigrant minority languages at home and school*. Clevedon: Multilingual matters.

Fourcade, M. (2007). Theories of Markets and Theories of Society. *American Behavioral Scientist*, *50*(8), 1015-1034.

García, O. & Fishman, J. A. (2001). *The multilingual apple: languages in New York City*. Berlin: Mouton.

Geddes, A. (2008). Il rombo dei cannoni? Immigration and the centre-right in Italy 1. *Journal of European Public Policy*, *15*(3), 349-366.

Gruzd, A., Wellman, B., & Takhteyev, Y. (2011). Imagining Twitter as an imagined community. *American Behavioral Scientist*, *55*(10), 1294-1318.

Gullino, S. (2009). Urban regeneration and democratization of information access: CitiStat experience in Baltimore. *Journal of environmental management*, *90*(6), 2012-2019.

Hong, L., Convertino, G., & Chi, E. H. (2011). Language Matters In Twitter: A Large Scale Study. *International AAAI Conference on Weblogs and Social Media,* Barcelona, 2011.

Jiang, B. (2013). Head/tail breaks: A new classification scheme for data with a heavy-tailed distribution. *The Professional Geographer*, *65*(3), 482-494.

Kaplan, R. S. (1992). The Balanced Scorecard: Measures That Drive Performance. Harvard business review, *83*(7), 172.

Kim, S., Weber, I., Wei, L., & Oh, A. (2014). Sociolinguistic analysis of twitter in multilingual societies. *ACM Conference on Hypertext and Social Media*, Santiago, 2014. ACM Press.

Kitchin, R. (2014). The real-time city? Big data and smart urbanism. GeoJournal, *79*(1), 1-14.

Kraft, M. A. (2007). Mashing up the internet. In *Wood MS, editor. Medical librarian 2.0: Use of Web 2.0 technologies in reference services.* Binghamton, New York, 2007. Haworth Information Press.

Lee, D., Felix, J. R., He, S., Offenhuber, D., & Ratti, C. (2015). CityEye: Real-time Visual Dashboard for Managing Urban Services and Citizen Feedback Loops. *Computers in Urban Planning and Urban Management Conference,* Cambridge, 2015.

Leimgruber, J. (2013). The management of multilingualism in a city-state. *Multilingualism and Language Diversity in Urban Areas: Acquisition, identities, space, education*, 227-256.

Mathwick C, Wiertz C, & de Ruyter K. (2008). Social capital production in a virtual P3 community. *Journal of Consumer Research*, 34, 832–49. (April).

Mattern, S. (2015). History of the Urban Dashboard. *Places Journal*.

McKinsey & Company (2013). How to make a city great. Retrieved from http://www.mckinsey.com/insights/urbanization/how_to_make_a_city_great.

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

Murrell, P. (2005). R Graphics. *CRC computer science & data analysis*. Chapman & Hall, Boca Raton.

Ravikumar, S., Balakrishnan, R., & Kambhampati, S. (2012). Ranking tweets considering trust and relevance. *ACM Ninth International Workshop on Information Integration on the Web*, Scottsdale, Arizona, 2012. ACM Press.

Sanderson, M. R. (2015). Are world cities also world immigrant cities? An international, cross-city analysis of global centrality and immigration. *International Journal of Comparative Sociology*, 0020715215604350.

Takhteyev, Y., Gruzd, A., & Wellman, B. (2012). Geography of Twitter networks. *Social networks*, 34(1), 73-81.

UN Habitat (2016). Urbanization and development: emerging futures. *World cities report 2016, United Nations Human Settlements Programme*, Nairobi, 2016.

Wang, Y., & Fesenmaier, D.R. (2004). Towards understanding members' general participation in and active contribution to an online travel community. *Tourism Management*, 25, 709-22.

Wellman, B., & Leighton, B. (1979). Networks, neighborhoods and communities. *Urban Affairs Quarterly*, *14*, 363-390

Wu, J. J., & Chang, Y. S. (2005). Towards understanding members' interactivity, trust, and flow in online travel community. *Industrial Management & Data Systems*, 105(7), 937-954.