**ESTIMATING DEMOGRAPHICS AND MIGRATION PATTERNS USING TWITTER DATA**

Piotr Drzewiecki, Faraz Salahuddin, and Juan Carlos Salamanca

**Abstract:**

This study attempts to measure demographic breakdowns and changes in selected European cities by applying language processing software on large amounts of geolocated Twitter data. Multiple methods to scrape and analyze publicly available twitter data are discussed. Ultimately, the data collected and utilized in this study is not robust enough to allow for meaningful conclusions to be drawn. However, in full recognition of this limitation, our report can be better read as an outline of a research program. We conclude that it is possible to estimate demographic changes using social media data. Furthermore, with additional methodological refinements (discussed in Section IV), the research program set out in this report has the potential to surface rich and reliable demographic insights with regards to migration patterns in real time—significantly augmenting the information available to policy makers who previously might have relied only on time-lagged census or other survey-instrument data.

**SECTIONS:**

Introduction

1. Primary Data Collection and Dataset Integration
2. Language Detection and Analysis
3. Results and Discussion
   1. Trends in Language Composition Over Time
   2. Demographic Statistic Accuracy Assessment
   3. Language Tracking Through Space and Time
4. Limitations, Lessons, and Possible Improvements

**Introduction**

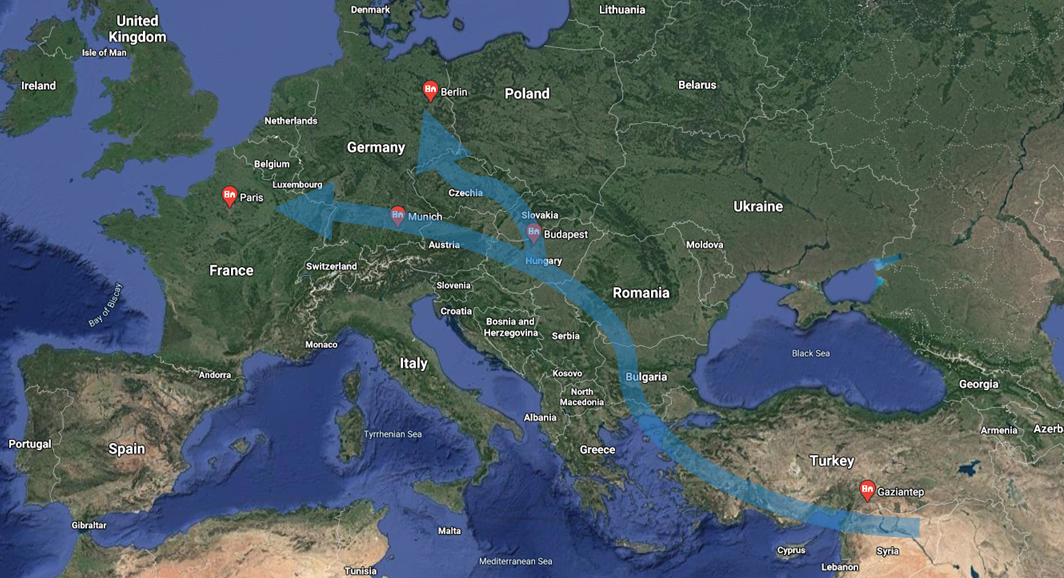
We started our project trying to address an important problem in public policy: the lack of updated demographic information, particularly in face of a migrant crisis. Considering the tools learned in our *Python for Global Affairs Course* and the information we have access to, we considered that the analysis of geo-located social media data could provide us the solution to our problem.

The case of Gaziantep, Turkey—one of the cities we study—helps illustrate exactly why this research project is necessary and the value that its results might hold for policy makers. Gaziantep, a city near the Turkey-Syria border has been a major intake point for refugees and migrants exiting conflicted affected zones in Syria ever since the start of the civil war in the region. The influx of individuals into Gaziantep has been large, and according to some estimates there are now ca. 500,000 Syrians in the city (over 25% of population)[[1]](#footnote-1). However, according to official demographic data published by the Turkish authorities the number of foreigners in the city is 16,000[[2]](#footnote-2). This is because refugees are given temporary protection status, and thus are not counted as foreign citizens with residence permit. Furthermore, the constantly evolving nature of conflicts and the forced migration that they produce mean that demographic changes in cities must be understood in as close to real time as possible. Censuses and other survey-based demographic counting methods inherently require extended measurement periods and produce an information gap. To address this need, new ways of counting demographics are required. That is precisely the policy space in which this study seeks to make an intervention.

One potential method to fill the information gap identified above is to use widely available social media data, and in particular, the language of publications made by users to estimate changes in population and migration. In this case, we have used historical data scraped from Twitter.com due to the popularity of the platform and the relative accessibility that Twitter grants to data scraping software.

Thus, in its most generalizable form, this research seeks to answer the question: *can changes in metropolitan area demographics be estimated using language of social media data as a proxy?* To figure this out, we posed a more limited question: *can we use the language of tweets to observe the changes in population in 2014-2020 in the cities of Gaziantep, Budapest, Munich, Berlin and Paris?*

What follows is an attempt to record and track changes in demographics across these five cities in the Eastern Meditterranean and Europe after the outbreak of the Syrian Civil War. We selected this event as we expected it would allow us to observe significant changes of the Arabic-speaking population in the selected cities.



We hope that the research process displayed in this study can serve as the basis for further investigations into the application of social media data to understand real-time demographic changes. Section IV discusses how our work may be used to further refine this type of analysis in significant detail.

Note that throughout our report, we will reference the location of the code we used within a series of Jupyter notebooks. All of them can be found in our GitHub Repository: **Python2020\_FinalProject\_ MigrationSocialMedia**.

**SECTION 1. Primary Data Collection and Dataset Integration**

**1. 1 Data collection**

To build the dataset of tweets we scraped Twitter using the Python package GetOldTweets3. For that purpose, we created a scrapping function that would allow us to easily select the dates and location of tweets to scrape. The function would save the scraped tweets in a CSV file.



Using this function, we scraped in the five cities, geo-located tweets of specific dates in quarterly intervals, starting in April 20, 2014 and concluding in April 20 of 2020. In total, we downloaded for each city 25 days tweets.It is important to note that we could only use Twitter data from users that had opted-in activating the geo-location data for their tweeting, and this limited significantly the number of tweets we could use. The next table displays the number of tweets scraped per city:

|  |  |
| --- | --- |
| **City** | **Total Number of Tweets Scraped** |
| Gaziantep | 17,445 |
| Budapest | 17,857 |
| Munich | 23,039 |
| Berlin | 45,314 |
| Paris | 123,222 |
| **TOTAL** | **226,877** |

The code used to scrape the tweets can be found in **Section 1.1 Collection - Get Old Tweets** of our Jupyter Notebook **1. Data collection, integration and langdetect.**

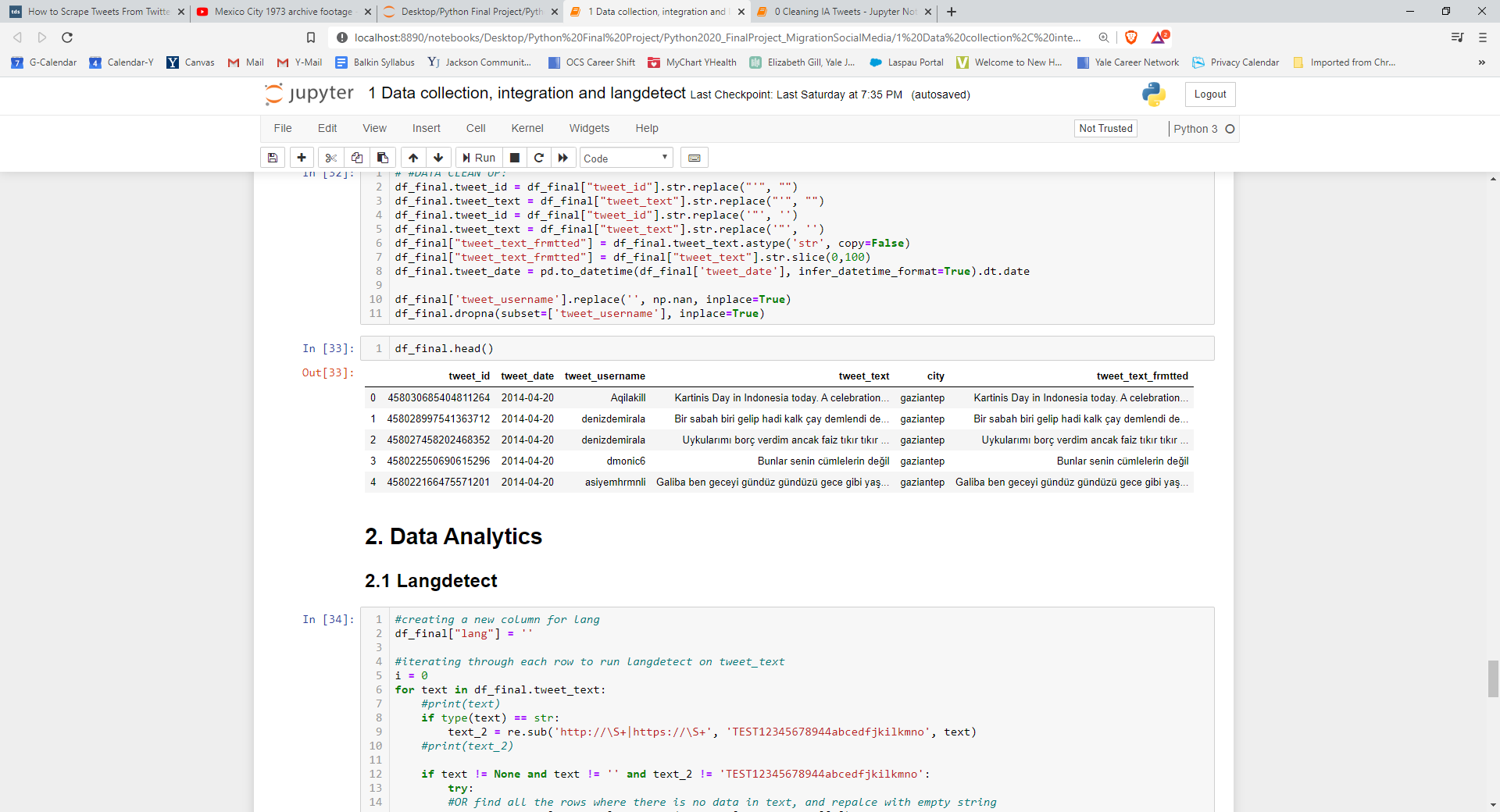
We decided to use GetOldTweets3 over Tweepy and Twitter’s API because the API does not allow the user to get tweets older than a week. Getting tweets older than a week was essential for us to identify patterns of language share change in different locations. Those patterns would not be observable within one week.

We also considered using Twitter data from the Internet Archive. The Internet Archive uploads a collection of JSON tweets from the general Twitter stream that amounts to 1% of all Tweets. The Internet Archive’s datasets had several advantages over GetOldTweets3 that, in future experiments, could help us improve our results. For instance, data is better spread throughout time and its 1% of Tweets is representative of the whole Twitter stream, something that is unguaranteed from the GetOldTweets grab. However, to work with these datasets we had to download all tweets from around the world, including tweets with and without geo-location data, which required disk space and analytical capacity we did not have at the time. For example, when running initial data explorations, we downloaded the data for April of 2015, and the dataset weighed 38.1 GB. The exploratory analysis of this dataset did reveal an important obstacle: this dataset only registers location declared by the user in his/hers profile, and not the geolocation of the user at the time of tweet. The exploratory analysis of this data set can be found in the Jupyter Notebook **0. Cleaning IA Tweets**.

**1. 2 Dataset integration and cleanup**

Once we had downloaded all of the tweets on CSV files, we loaded them as DataFrames using Pandas. We then appended them into one Data Frame. This process can be found in **Section 1.2 Data Combine** of our Jupyter Notebook **1. Data collection, integration and langdetect.** Afterwards, we proceeded to clean the strings in our data set to remove all of the characters that would be useless or redundant in our future analysis (e.g. removing “ ‘ “ from the text of tweets). The code for this is found in **Section 1.3 Data Cleanup** of the aforementioned Notebook.

The following table displays the first five rows of our dataset after combination and cleanup:

****

**SECTION 2. Language detection and analysis**

The second fundamental step required to complete a migratory pattern analysis using Twitter.com data is to assign each Twitter user a nationality of origin. To complete this assignment, a decision criteria is needed that can take in available information and extrapolate the user’s country of origin. Many such criteria are possible ranging from the crude to the overly sophisticated (see Section IV for additional discussion on what a more sophisticated nationality detection logic might look like). The method and criteria used in this study to assign Twitter users an origin-nationality is language of tweet. We have confidence in assuming language as a reasonable proxy for nationality because of the specifics of the case being observed.

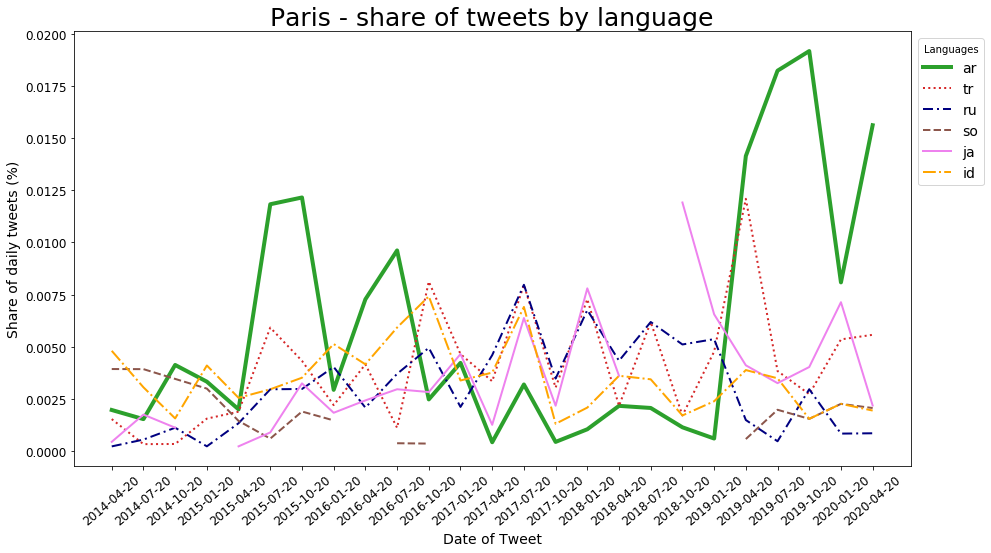
To enrich our Twitter.com data with an origin-nationality tag, we analyzed Tweet text for the language used and assigned the highest probability language to each record. The publicly available language detection Python library *langdetect* was used to complete this assignment. Langdetect was able to identify the language used in each tweet, however our dataset remained at tweet-level dataset. In order to properly count migration (a person-level phenomenon) the working dataset needed to be condensed to a user- or person-level dataset. A single user could have multiple tweets in multiple languages, and our dataset showed significant variation in tweet volume across users. Therefore tweet-language distributions cannot be used to understand user-nationality distributions.

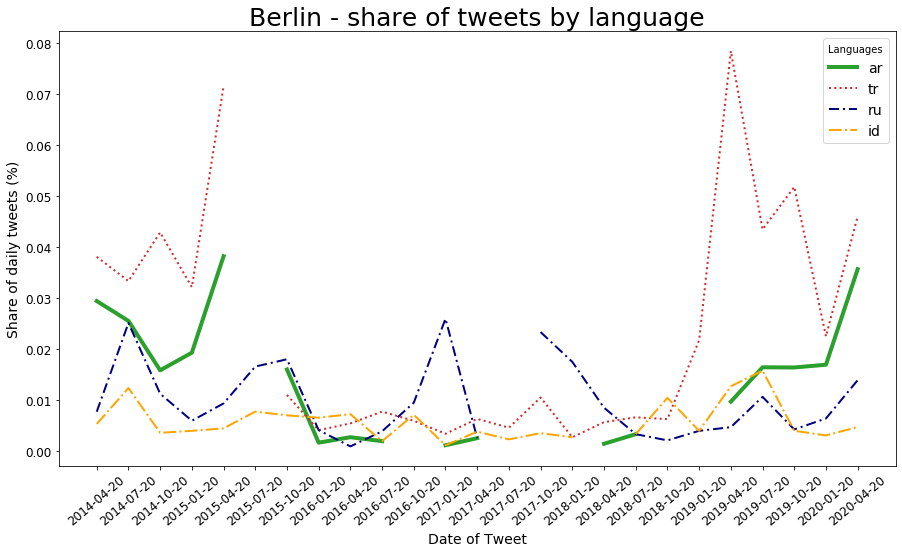
To collapse rows into a user-level dataset, we counted for each user the various languages used over the entirety of the study period (i.e. all days in the dataset) and ranked them in descending order of frequency. The language with rank = 1 was selected as each user’s primary language, and a new dataset was constructed that contained one row per Twitter.com user with only one primary language assigned. This primary language tag was now the variable to be used as a stand-in for nationality and the variable we have used to estimate demographic changes in all the cities studied.

**SECTION 3. Results and Discussion**

**3. 1 Trends in language composition over time**

Having constructed a built-to-purpose dataset, the first step in our analysis was to look at changes in language use over time for each city. The goal here is to compare rates of language usage against itself to see if the flow of foreign language speakers (i.e. foreign nationals) entering this geography was net positive over time. The ideal metric would be to use absolute counts of users tweeting in a language, however, due to limitations in how our data was originally sourced, we were unable to use raw counts (see Section 4, Limitation 1 for a detailed explanation of why). Share of tweets by language has been shown below to demonstrate the type of analysis we are after. Looking at Paris as a case example, we can see spikes in Arabic language use relative to other languages in 2015Q1 - 2015Q2, then again in 2016Q1-2016Q2 and 2019Q1 onwards. This is important insight already and provides a level of detail in demographic trends that one would be unable to get with official demographic data released by the City of Paris alone.





Berlin is an interesting example of how a relatively small number of Twitter users that opted for geolocation is an obstacle in using this social media platform as a proxy for demographic composition. The graph above presents sudden increases in the number of users tweeting in Arabic and Turkish, while at the same time on some of the dates there are no Arabic or Turkish users tweeting at all. Before trying to pin these changes to migratory, political or religious events, it is necessary to investigate the absolute values that are represented above as a percentage change. Upon closer examination the data reveals that Arabic, Turkish and any other minority languages at any given date are represented by only a handful (10 - 20) users. This small number of users who may not use the platform every day leads to a higher volatility in the results. Ergo, these daily samples are not representative of the underlying trends.

One curious occurrence, which we consider a possible false attribution in language detection by *langdetect,* is the steady and relatively significant number of users tweeting in Indonesian in all the locations. Once we established that the daily changes are not a good method to verify demographics, let us look into more aggregated data.

**3. 2 Demographic statistic accuracy assessment**

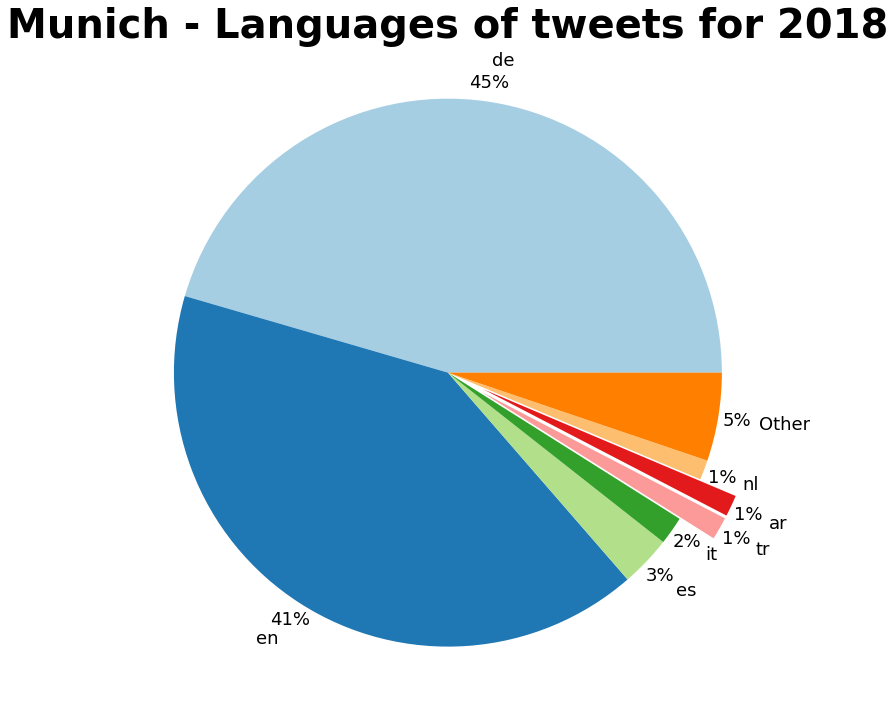
As a way of getting a better grasp of the shares of language spoken per year, per city, we visualized this information in a series of pie charts. We created a pie chart per city per year, displaying the share of languages of tweets (considering unique users). To calculate the data per year, we averaged the shares of the dates we had scraped for each year. This means that these pie charts are merely an estimate of the share per year based on the days we scraped. They should not be understood as the share of all of the tweets for that year. Additionally, we performed the same exercise for the sum of all of the tweets of all of the cities, per year.

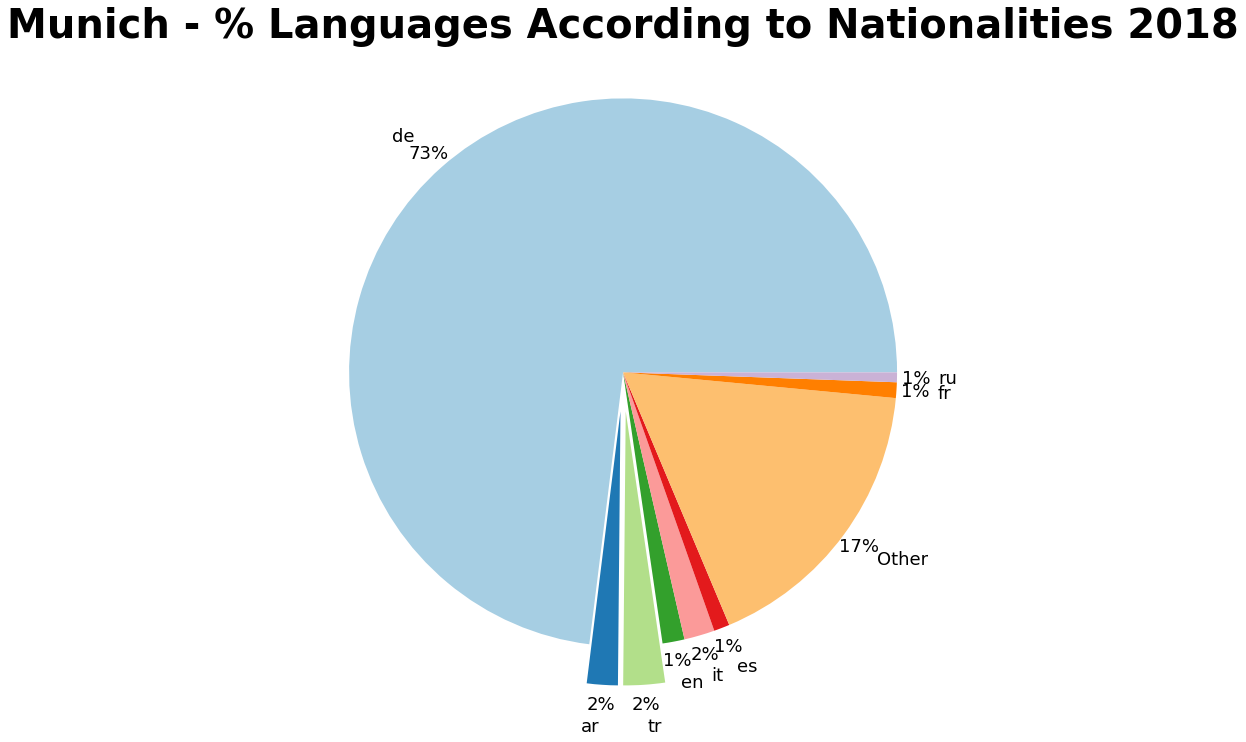
The code for the pie charts can be found in the Jupyter Notebook **3.2 Pie Chart,** and the 42 pie charts can be found in **Appendix 3.2** in our Repository.

The objective of calculating these pie charts is to better understand the proportion of languages observed in each city for each year. This serves two purposes:

1. As a preliminary sanity check, to see if, intuitively, the information makes sense considering what we know of the cities and the languages we would expect to encounter there. For instance, we would expect for Paris to have French as the most tweeted language and to have a significant share of tweets in other languages commonly spoken in Paris, like English, Spanish and Arabic.
2. To understand how precise our shares are in comparison with official census information, in order to better understand how representative our data is. We performed this exercise in the case of Munich in 2018, as we considered the data publicly available was of good quality.

The following pie chart shows the average share of languages in tweets, estimated per unique user, for the city of Munich in 2018. Then, we show a chart that represents an estimation of the languages spoken in Munich for 2018 according to official data. This estimation was done according to census data, assigning the official language to each national living in Munich.[[3]](#footnote-3)



****

This comparison allows us to see that our Twitter data is more useful in aggregate. However, an important adjustment has to be made in the share of English tweets, as it does not represent nationals coming from English speaking countries. We are convinced this is explained by the large number of German English speakers (as well as English speakers from other nationalities) that participate in Twitter. However, Arabic, Turkish, Spanish and Italian appear in both graphs as languages with a significant share, which means that the Twitter data is telling us something about the languages that have a significant share of speakers in Munich. Additional statistical tests that escape the scope of this project have to be made in order to estimate the languages of tweets are a good proxy to predict the population of native speakers of a particular language.

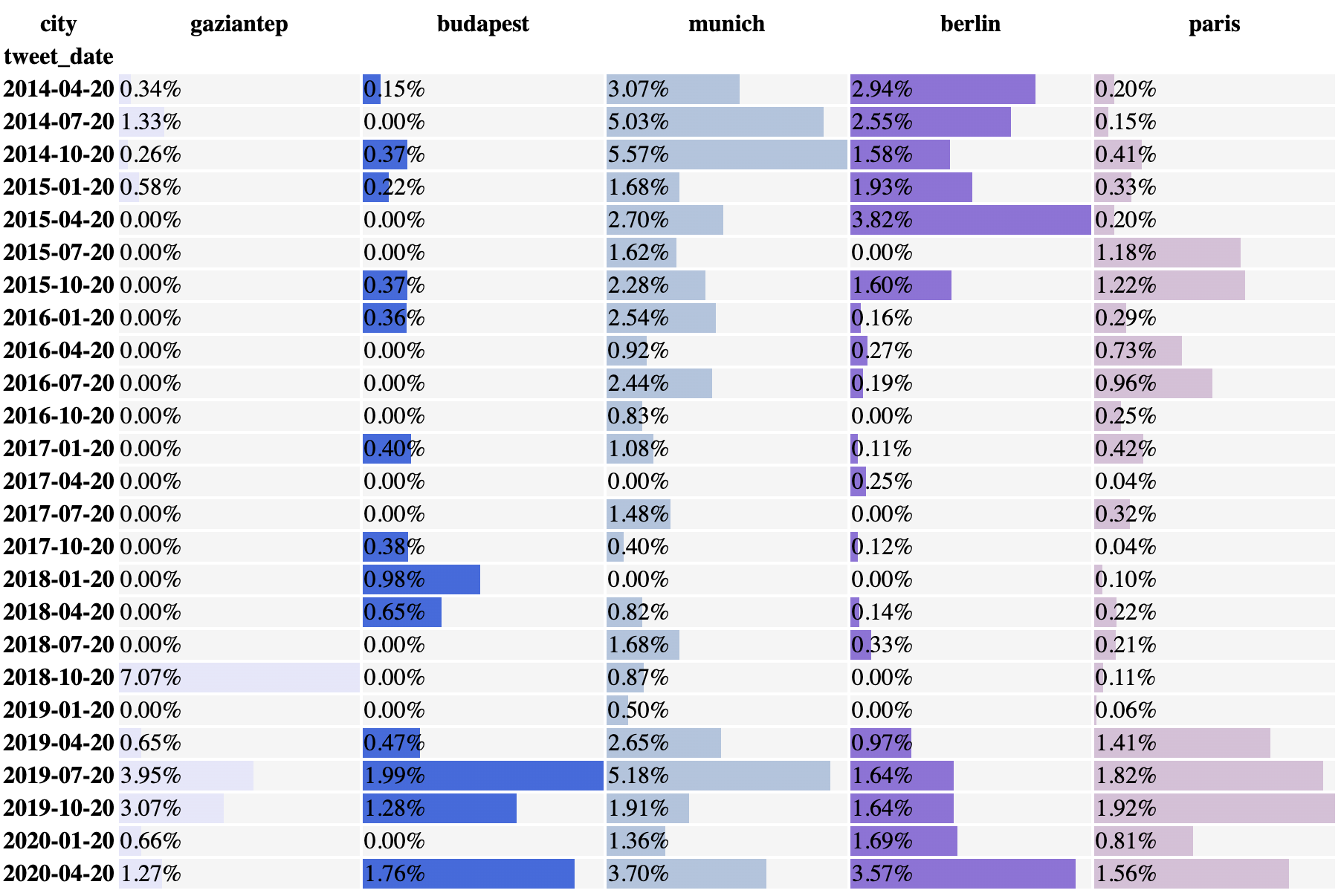
**3. 3 Language tracking through space and time**

After having assessed demographics in particular cities over time, and via comparative snapshots, the next step is to view changes in demographics over time for all cities together. Doing so will allow us to make inferences about particular migratory pathways and where along a known migratory pathway bulges of bottlenecks exist. Figure 3.3 puts all the data components visualized in Section 3.1 and 3.2 together to do exactly this.

The inset bar charts within each column, have been included to help the reader quickly and easily identify relative change in Arabic language share relative to each city’s own maximum value. Taking the example of Berlin, one can tell at quick glance, that the Arabic language usage in tweets peaked in late 2014 and then again in middle 2020. In comparison, Paris does not show a bimodal distribution and Arabic language share seems to grow more steadily over the duration of the study-period. However, owing to the inherently compromised nature of the dataset, we will refrain from extensively analyzing the inter- and intra-city trends we have calculated for the Arabic share. Rather, and in keeping with the mission of this report to outline a viable research program, we will discuss the relative merits and rationale for visualizing Twitter-demographic data in the specific formations as we have reproduced here.

When conducting the migration tracking analysis that we have proposed, a multi-indexed table with cities (columns) ordered to follow the migratory path of refugees and otherwise displaced peoples should be used. Table has been sorted on the date column, in ascending order. If data is reported in this fashion then to read the table from top-left to bottom-right would be to follow migrants along their journeys across the Eurasian landmass. In the specific case of migration out of conflict affected zones in Syria, one should expect to see an increasing share of Arabic language in this table as one moves diagonally from the top left to the bottom right starting in about 2015. Furthermore, this same table template may be used to test other hypothesized migration pathways as well. New geographic points of measurement (cities) could be selected and the column order adjusted. If the table (and underlying bar chart lengths) show gradual or staggered increase as one moves diagonally through the table, then there would be credible evidence to suggest the existence of a migratory pathway for individuals exiting conflict zones after a particular migratory trigger.

***Figure 3.3. Arabic language tweets as a percentage of all tweets, 2014-2020***

****

*(Note: the table reproduced below was created using the Dataframe.style property within the pandas data analysis library in Python3)*

**SECTION 4. Limitations, lessons, and possible improvements**

**4. 1 Limitations**

Inconsistent GetOldTweet3 Performance.Final analyses and checks on the compiled dataset revealed significant inconsistencies in how GetOldTweets3 was able to source tweet-level information for our geographies of interest. We observed that for all five cities, tweet volume was high (in relative terms) in dates early and late in the study period (2014Q2 - 2020Q2). Dates falling in the middle section of the study period showed total tweet counts up to 4 and 5 times lower than tweet counts for dates on either end. Given that this drop in volumes is visible across all five cities for nearly the same date range, we conclude that this is not attributable to broader Twitter utilization but specific to the data-collection methods used in this study.

Why do certain dates return lower volume across the board when GetOldTweets3 mimics Twitter’s own search function to fetch old tweets will need to be assessed separately and our study does not touch on this exact issue. However, our research process does reveal that simple fact that the Python library GetOldTweets3—as versatile and useful as it is! —does not fetch historical tweets in a uniform or consistent manner. This particular dynamic requires that we make a fundamental assumption before moving on to data analysis and discussion of analysis results. We assume that GetOldTweets3 is basically unbiased in its fetching of old tweets, even in cases where total volumes of tweets fetched for a day are significantly lower in absolute terms than on other days. This assumption allows us to carry forward the task of data analysis, however it would need to be validated to confirm the validity of conclusions drawn.

One further consequence of the inconsistent performance observed in GetOldTweets3 is that we cannot compare, using absolute volume figures, changes in language use across time for individual geographies. It is not possible to claim that language use decreased or increased over time if we do not have confidence in comparability of daily tweet samples that we have gathered—which we do not. As such, our analyses have to limit themselves to measuring changes in language share in a given city over time. This is permissible precisely because of the assumption stated in the paragraph above.

**4. 2 Project Refinements**

Through our research process and the challenges faced along the way, we have identified multiple avenues for methodological refinement that future iterations of this study should employ to arrive at more robust and accurate demographic estimates. Each refinement is addressed below.

Refinement 1: increase sample size.First and foremost, the sample size of tweets used for this analysis should be expanded to be as wide as possible. As discussed in Limitation 1 above, GetOldTweets3 has proven to be an unreliable method to scrape historical tweets in a uniform manner. It is unclear if the sample scraped by GetOldTweets3 is properly random or if there is bias informing which tweets are secured. We believe that compiling aggregate data using more data points would yield more promising results. Alternatively, for an ongoing projects, using Twitter’s developer access and scrapping the tweets daily may be a better approach.

Relatedly if researchers have the requisite access, Facebook.com utilization and posting data should be preferred to Twitter, since Facebook is likely to have greater penetration among migrating communities than Twitter[[4]](#footnote-4). Most importantly, geolocation is a feature that users have to opt in to on Twitter.com, and tweets are not geotagged by default. The characteristics of those who manually opt into geolocation on Twitter.com could be systematically different to the wider Twitter population. It is also unlikely to expect that those fleeing conflict zones will opt into geolocation at high rates. This represents a major weakness of using Twitter data for conducting the kinds of analysis described here.

Refinement 2. build a more sophisticated nationality identification method. Assigning a language of origin to each Twitter user in the working dataset is a fundamental component of our analysis. Consequently, to improve the overall accuracy of this research activity, ensuring that the initial tagging of language of origin based on information contained in tweets is robust to various deviations (including multiple nations speaking similar languages and bilingualism among twitter users) is crucial. In the absence of geolocated data that can consistently place users in particular geographies, we suggest using language processing scripts on variables contained in Twitter user bios, including name, to add another probabilistic measure of regionality.

Our nationality identification method treats the issue of bilingualism among twitter users very simply. Only the most frequently used language is incorporated into the analysis, and the user is effectively assumed to belong to that primary language alone. However, if we are to pay closer attention to individuals' online behavior and if we recognize that due to certain historically specific reasons, migrants from one region may already speak languages of another region (e.g. native Arabic speakers from Syria may also speak and tweet primarily in English). Our approach would miscount such individuals. Instead, primary language assignment could be changed to allow for users who display even a minority of tweets in a language.

1. As witnessed by one of us during visits to the city and in accordance to available media reports: https://www.theguardian.com/cities/2019/jun/19/gaziantep-turkish-city-successfully-absorbed-half-a-million-migrants-from-syria [↑](#footnote-ref-1)
2. Türkiye İstatistik Kurumu: İllere ve vatandaşlığa göre Türkiye'ye gelen ve Türkiye'den giden göç, 2016-2018 [↑](#footnote-ref-2)
3. Data obtained from the official site of the government of Munich: <https://www.muenchen.de/rathaus/dam/jcr:89a2dcdb-76bb-427d-8930-61a956092c08/jt190115.pdf> [↑](#footnote-ref-3)
4. Carleen Maitland, Ying Xu: “A Social Informatics Analysis of Refugee Mobile Phone Use: A Case Study of Za’atari Syrian Refugee Camp”, TPRC 43: The 43rd Research Conference on Communication, Information and Internet Policy Paper, 2015. [↑](#footnote-ref-4)