MAX PLANCK INSTITUTE
FOR DEMOGRAPHIC RESEARCH

Konrad-Zuse-Strasse 1 · D-18057 Rostock · Germany · Tel +49 (0) 3 81 20 81 - 0 · Fax +49 (0) 3 81 20 81 - 202 · www.demogr.mpg.de

MPIDR Working Paper WP 2022-012 I March 2022
https://doi.org/10.4054/MPIDR-WP-2022-012

Analyzing EU-15 immigrants' language acquisition using Twitter data

Sofia Gil-Clavel
André Grow I gil@demogr.mpg.de
Maarten J. Bijlsma

This working paper has been approved for release by: Emilio Zagheni (sekzagheni@demogr.mpg.de), Head of the Laboratory of Digital and Computational Demography.

Analyzing EU-15 Immigrants' Language Acquisition using Twitter Data

Sofia Gil-Clavel[1,2], André Grow[1], Maarten J. Bijlsma[2,1]

The increasingly complex and heterogeneous immigrant communities settling in Europe have led European countries to adopt civic-integration measures. Among these, measures that aim to facilitate language acquisition are often considered crucial for integration and cooperation between immigrants and natives. Simultaneously, the rapid expansion of the use of online social networks is believed to change the factors that affect immigrants' language acquisition. However, so far, few studies have analyzed whether this is the case. This article uses a novel longitudinal data source derived from Twitter to: (1) analyze differences between destination-countries in the pace of immigrants' language acquisition depending on the citizenship and civic-integration policies of those countries; and (2) study how the relative size of migrant groups in the destination-country, and the linguistic and geographical distance between origin- and destination countries, are associated with language acquisition. We focus on immigrants whose destination countries were in the EU-15 between 2012 and 2016. We study time until a user mostly tweets in the language of the destination-country for one month as a proxy of language acquisition using survival analysis. Results show that immigrants who live in countries with strict requirements for immigrants' language acquisition and low levels of liberal citizenship policies have the highest median times of language acquisition. Furthermore, on social media such as Twitter, language acquisition is associated with classic explanatory variables, such as size of the immigrant group in the destination country, linguistic distance between origin- and destination-language, and geographical distance between origin- and destination-country.

Note: Last version of the paper can be found here: https://doi.org/10.31235/osf.io/bs4hk.

---

[1] Max Planck Institute for Demographic Research.
Mail to: gil@demogr.mpg.de
[2] University of Groningen.

Introduction

Since the beginning of the 21st century, policy makers across Europe have aimed to enforce the national language among immigrants through civic-integration policies (Wright & Viggiano, 2020). This was a reaction to the increasingly complex and heterogeneous immigrant communities settling in Europe, a phenomenon coined as 'superdiversity' (Vertovec, 2007). Civic-integration policies rest on the assumption that immigrants' successful incorporation into the host-society goes beyond economic and political incorporation, by relying "also on individual commitments to characteristics typifying national citizenship, specifically country knowledge, language proficiency and liberal and social values" (Goodman, 2010, pp. 754). Language acquisition is often regarded as critical for integration and cooperation between immigrants and natives (Eckert, 2018; Forrest et al., 2018); therefore, many integration measures aim to facilitate language acquisition (Duncan, 2020). However, so far little is known regarding how such civic-integration measures affect language acquisition. One reason for this is a lack of multinational data that allows a comparison of different civic-integration measures across different migrant groups (van Tubergen & Kalmijn, 2005). To address this knowledge gap, we use data on language use obtained from Twitter, from January 2012 to December 2016. We study the pace of migrants' destination-language acquisition, and assess whether and how this pace is associated with different civic-integration policies in the EU-15, as categorized by Goodman (2010).

Our use of Twitter data enables us to study changes in language use in a longitudinal and non-intrusive way, among immigrants from a large number of countries of origin in the EU-15. Compared to traditional data used in migration research, Twitter data offers access to transnational and comparable migration data in a continuous manner. Because of these properties, Twitter data has been used to study different aspects of migration. For example, Mazzoli et al. (2019) show that geo-located Twitter data can be used to monitor migration routes, settlement areas and mobility of migrants, and that the data is correlated with official migration data from international agencies. Similarly, Zagheni et al. (2014) use data from 500,000 geo-located tweets to estimate migration flows from Twitter users in OECD countries, and Hawelka et al. (2014) use geo-located tweets to uncover global patterns of human mobility. When it comes to integration, Twitter data has been used less frequently so far, but Lamanna et al. (2018) show that language use patterns on Twitter can be used to study the interplay between migrant integration, social polarization, and spatial

segregation in different migrant communities in more than 50 cities. In our work, we follow Lamanna et al. (2018) and study immigrants' integration through the study of the language they use in their tweets. Our central assumptions are (1) that a switch from tweeting in the language of the country of origin to tweeting in the language of the country of destination is an indicator of language acquisition among migrants, and (2) that the time frame over which this switch happens provides insight into the pace of language acquisition.

To develop hypotheses as to how different civic-integration and citizenship policies affect language acquisition, we draw on the work of Goodman (2010) and Howard (2010). Goodman (2010) and Howard (2010) proposed to classify the EU-15 countries according to their requirements for civic-integration and citizenship, respectively. Conceptually, we rely on the governmentality framework that theorizes on the effects governmental interventions have on individuals (Foucault, 1991). In a nutshell, the governmentality framework holds that the government has the power to modify people's behavior through policy interventions (Foucault, 1991). One complicating factor here is that the use of social media itself may affect the process of language acquisition. Some scholars have argued that social media makes it easier for migrants to stay in touch with communities in their countries of origin. Therefore, traditional factors that typically affect language acquisition, such as the geographical distance between origin- and destination-country, may lose their importance (Komito, 2011; Wright & Viggiano, 2020). To assess this possibility, we also study the effect of factors that have traditionally been considered in studies of language acquisition, conditional on civic-integration and citizenship policies.

Background

Language is considered an important factor in the integration process and it facilitates cooperation between immigrants and natives (Eckert, 2018; Forrest et al., 2018). Indeed, mastering the language of the country of destination improves access to education and important institutions, and is associated with higher income, societal recognition, and social contacts (Duncan, 2020). As such, it facilitates the acquisition of human capital in the country of destination (Esser, 2006). Because of this central role, language acquisition has always been considered an important variable in the study of immigrants' integration (Algan et al., 2012a; Esser, 2006) and it has been the focus of civic and integration policies (De Haas et al., 2020; Wright, 2020).

The role of civic-integration policies

Foucault (1991) was among the first to theorize on how governmental programs have the capacity to change the behavior of the population. This notion is captured in the term *governmentality*, which refers to the effects governmental interventions have on individuals, depending on individuals' positions in relation to governmental programs (Li, 2007). This includes interventions related to poverty, health, and demographic events, such as migration and fertility (Castro-Gómez, 2010; Li, 2007). Civic-integration requirements are a special case of governmental interventions, where the target population are immigrants. Civic-integration requirements usually have a two-fold nature. First, they assist newcomers with acquiring the local language, accessing basic services, and entering the labor market, i.e., they promote migrants' individual autonomy (Duncan, 2020; Goodman, 2010). Second, civic-integration requirements are "intended to protect the host society from the presence of others becoming socially disruptive" (Duncan, 2020, pp.604). Immigrants' gradual adoption of new behaviors because of governmental interventions is documented by Menjívar and Lakhani (2016). They show that the existence of host-country citizenship requirements motivates immigrants to adopt new behaviors and life styles in the short- and long-term. According to Menjívar and Lakhani (2016), one reason is a fear to be deported. A second reason is the attempt to fit legal categories of admission to the United States.

Across countries, many types of civic-integration policies have been implemented. In the European context, Goodman (2010) provides a comparative analysis that classifies EU-15 countries according to the broad 'citizenship strategies' that they pursue. This is done by clustering the countries based on their citizenship access and membership content policies. The notion of citizenship access comes from the Citizenship Policy Index (CPI), which considers the 2008 citizenship policies of the EU-15 countries (Goodman, 2010; Howard, 2010). The notion of membership content is based on the Civic-Integration Index (CIVIX) that considers requirements for country knowledge, language, and values (Goodman, 2010). Citizenship requirements are the rules extending legal status and rights depending on state membership (entrance, settlement, or citizenship), while integration requirements are related with the performance and degree of incorporation of newcomers to the host society (such as language acquisition and values commitment) (Goodman, 2010; Howard, 2010). Based on these indexes, Goodman (2010) clusters the EU-15 countries in four groups (see details below): (1) prohibitive, (2) conditional, (3) enabling, and (4) insular.

The Prohibitive Group

The *prohibitive* group is formed by Austria, Denmark, and Germany. This group has high citizenship requirements (e.g., no dual nationality and long time in the country before acquiring citizenship) (Howard, 2010) and high integration requirements (e.g., mandatory language requirements and country knowledge) (Goodman, 2010). According to Howard (2010), Germany has more liberal citizenship policies than Austria and Denmark. This is because of the relatively strong anti-immigrant attitudes held by the population of Austria and Denmark, together with a lack of economic pressures to liberalize the citizenship requirements (Howard, 2010).

The language-integration policies of countries in the prohibitive group have been characterized by a lack of tolerance towards different cultures, which are seen as a threat to the language and culture of the host society (Beauzamy & Féron, 2012; Brochmann & Hagelund, 2011; Schierup et al., 2006). It was not until 2002 that Austria adopted some measures to offer language training for immigrants. Before 2002, immigrants were expected to learn the language on their own (Höhne, 2013). In the case of Germany, the government began to finance language courses in the mid-1970s, which was before even establishing integration policies (Höhne, 2013). For Denmark, there are several studies that highlight lack of flexibility of Danish policies directed towards migrants. Migrants must show a perfect command of Danish to not suffer social and labor discrimination (Beauzamy & Féron, 2012; Lønsmann, 2020). Austria, Germany, and Denmark are among the European countries that before 2012 required a high level of language acquisition (B1) for permanent residence and citizenship (Höhne, 2013).

The Conditional Group

The *conditional* group consists of France, the United Kingdom, and the Netherlands. This group combines liberal criteria of citizenship with arduous integration requirements. Here citizenship is seen as a reward for integration. Therefore, migrants must acquire the language and country knowledge before obtaining citizenship, or even before moving to the country (Goodman, 2010). These countries are 'traditional' immigration countries with a colonial past (Brett, 2002) that since very early, by European standards, have tried to incorporate the immigrant population into the host-society by promoting an atmosphere of tolerance and cultural diversity (Algan et al., 2012b; Manning & Georgiadis, 2012). Furthermore, France and the United Kingdom are considered

historically liberal countries, while the Netherlands liberalized its citizenship policies between 1980 and 2008 (Howard, 2010).

France, the United Kingdom, and the Netherlands have been characterized by their legislations of cultural diversity tolerance, where citizenship for newcomers has been essential for their national identity (Castles et al., 2013). It was believed that this openness to diversity would led immigrants to feel part of the wider community. In the long term, governments felt they failed to create common core values, i.e. to integrate immigrants into the wider society (Beauzamy & Féron, 2012; Manning & Georgiadis, 2012). Therefore, before 2012 migrants were already required to pass a basic test on language (A1/A2), culture, and history, if they wanted to become citizens, or even if they wanted to be admitted to the country (Höhne, 2013; Manning & Georgiadis, 2012).

The Enabling Group

The *enabling* group is formed by Portugal, Finland, Ireland, Belgium, and Sweden. In this group, citizenship serves as a mechanism for establishing equal status and rights. Hence, citizenship enables integration instead of rewarding it, which is the opposite of the conditional group (Goodman, 2010). While Belgium and Ireland are considered historically liberal countries, in Portugal, Finland, and Sweden citizenship requirements became more liberal between 1980 and 2008 (Howard, 2010). The liberalization of citizenship requirements is a consequence of the low levels of far-right support from the population, and, among others, demographic change and the rise of international norms (Howards, 2012).

In terms of language integration strategies adopted before 2012, Portugal and Finland only required language certification for citizenship (Goodman, 2012). Ireland, Belgium, and Sweden neither required national language nor country knowledge for citizenship or permanent residence (Goodman, 2012; Höhne, 2013). Sweden is a particular case, as it was among the first countries implementing language courses for immigrants, where the government already financed the courses since 1965 (Höhne, 2013). However, it was not until 2009 that Swedish became the national language of the country (Bolton & Meierkord, 2013).

The Insular Group

The *insular* group is formed by Greece, Spain, Luxembourg, and Italy. In general, these countries have a restrictive approach to citizenship for immigrants (Castles et al., 2013), where citizenship

is mostly granted to descendants born abroad (Goodman, 2010). This is a consequence of the electoral support to far-right parties and the anti-immigrant attitudes held by the population (Howard, 2010).

The countries in this group have complex language landscapes, where linguistically-independent languages are spoken in different regions of the countries or are used for different official purposes[3] (Bruzos et al., 2018; Love, 2015; Sharma, 2018; Skourmalla & Sounoglou, 2021). In Italy and Luxembourg, policy mechanisms that aim to standardize and regulate official language usage were introduced at the beginning of the 21st century. However, these policies create conflict with the communities speaking different languages in the countries (Love, 2015; Sharma, 2018), and serve as a barrier to migrants' linguistic integration (Angela et al., 2016). In the case of Greece, language policies were introduced in the 70s to homogenize the language and cultural landscape of the country (Skourmalla & Sounoglou, 2021). Before 2012, Greece already required migrants to be proficient in Greek to acquire long-term residency (Tsoukalas et al., 2010). In Spain, there was not any type of regulation for official language usage before 2015, nor for migrants' language acquisition (Bruzos et al., 2018).

Citizenship-policy and Civic-integration Indexes

As showed in the previous sections, the countries that formed each of the groups are quite heterogenous. Therefore, we use the raw indexes CPI and CIVIX, as this gives us more variance in the analyses and allows us to capture differences that are blurred by a merely categorical variable. We employ both the CPI and CIVIX indexes to characterize the civic-integration policies of the countries, as they capture two important macro-level factors associated with immigrants' language acquisition (van Tubergen & Kalmijn, 2005): the political climate towards migrants and language integration policies. Political climate and migrants' language acquisition are related through anti-immigrant attitudes and left-wing majority governments. When the members of the receiving society hold strong anti-immigrant sentiments, immigrants have less exposure to the host-language (van Tubergen & Kalmijn, 2005). When left-wing parties form the majority of a

[3] This is the case of Luxembourg, where "Luxembourgish is the national language, French the legislative language, and German is the language of instruction in public schools." (Angela et al., 2016, 4067).

government, the political climate tends to be more tolerant toward immigrants and policies tend to favor linguistic pluralism, i.e. there is more tolerance towards other languages (van Tubergen & Kalmijn, 2005).

Challenges in the study of language acquisition, Twitter as an alternative

Analyzing the effects that different civic-integration policies across Europe have on language acquisition among immigrants comes with several difficulties that are related to data availability and data quality requirements. First, as highlighted by Beauchemin (2014), there is a lack of comparable databases that cover different countries and that contain information on multiple immigrant groups. Second, the study of language acquisition processes requires longitudinal information that captures the changes experienced by the person once they start living in a new country (Font & Méndez, 2013). Finally, language adoption and proficiency are normally measured by self-assessment, but research shows that these self-estimates only to some extent reflect actual skills as measured by standardized tests (Edele et al., 2015). To address these difficulties, we draw on a sample of Twitter data that was retrieved between January 2012 and December 2016.

Twitter data represents a novel and suitable source of information to study EU-15 immigrants' language acquisition in a non-intrusive way (i.e. researchers have access to users' digital traces, which are generated from users digital lives) (Lazer & Radford, 2017). Twitter is a microblogging social network on which users can release 140-character[4] messages called "tweets". Users can also follow other users to see their tweets displayed in their feeds, without the requirement of being followed back. Twitter does not provide visible limits on the number of either followees or followers that users can have (see McFedries, 2007; Krishnamurthy et al., 2008). It allows conversations in different languages to take place simultaneously on the platform worldwide. Tweets can also be geolocated. As we explain in the data section, geolocation makes possible to infer the users' place of residence and, if the geo-location changes, possible migration events (Armstrong et al., 2021). These features make it possible to study language usage at the country

[4] Twitter announced that the limit of characters was going to be increased to 280 in 2017. https://blog.twitter.com/official/en_us/topics/product/2017/Giving-you-more-characters-to-express-yourself.html. Accessed in October 21st, 2020.

level and to analyze travel and mobility patterns (Mocanu et al., 2013). One further advantage is that Twitter data is created in a passive manner, i.e., users create the data by interacting with others via posting or re-tweeting. This allows researchers to study the dynamics and behaviors of Twitter users in a non-intrusive manner (Lazer & Radford, 2017; Mejova et al., 2015).

According to Twitter financial releases, the number of Twitter monthly active users in 2014 was 271 million (Twitter Inc., 2014), while for 2021 the number of Twitter daily active users was 206 million worldwide (Twitter Inc., 2021a). The largest number of users is located in the US, accounting for about 20% of the total user base (Twitter Inc, 2021b). Between 2010 and 2012, the European countries with the highest Twitter penetration, in decreasing order, were the Netherlands, the UK, Ireland, Sweden, Spain, Belgium, Italy, France, and Germany (Mocanu et al., 2013, Fig.2, pp. 3). In terms of demographic characteristics, based on a US national survey, Hargittai (2020) reports that around 37%, 23% and 11% of the participants in the age groups 18-50, 51-62, and 63+ use Twitter, respectively. Women and men are equally likely to be on Twitter; and highly-educated and highly-internet-skilled individuals are more likely to use the platform compared to less educated and less-internet-skilled individuals (Hargittai, 2020). Comparing Twitter users' data against UK representative samples, Leak et al. (2018) show that Twitter users in the age-group 10-39 are over-represented, while those over 40 are under-represented. Female users are more prevalent for the age-group 10-19, while male users become dominant for the 20+ age-group (Leak et al., 2018). Finally, Asian, Black, and mixed-group groups are underrepresented, while white users are the majority of the population (around 90%). The final percentage is similar to the usual resident population of the UK (Leak et al., 2018).

Access to Twitters' data was stable (i.e. researchers have had access to the same interface and its outputs) from 2012 to 2020 via its Application Programing Interface[5] (Zimmer & Proferes, 2014). This access only concerned prospective Tweets, but not Tweets that were sent more than seven

[5] Twitter launched the Twitter API V.2 in August 2020 (Cairns & Shetty, 2020). With that they replaced the version V1.1 launched in September 2012 (Costa, 2012).

days in the past[6]. However, different organizations have stored Twitter samples in a systematic manner, which allows researchers to study the behaviors in a longitudinal way (Morstatter et al., 2013; Sequiera & Lin, 2017). As we discuss in more detail below, we use data collected by the Internet Archive[7].

Language and Social Media Usage

While data from digital sources, such as Twitter, offer new opportunities to study language acquisition, some scholars have argued that the advent of social media itself may have affected the process of language use and maintenance (Komito, 2011; Wright, 2020). Before the dawn of social network sites, at the macro level, migrants' language adoption was inverse to (see Chiswick & Miller, 2001; Esser, 2006): (1) number of immigrants from the same origin-country living in the host-country; (2) linguistic distance between the mother tongue and the official languages of the destination-country; and (3) geographic closeness between the origin- and destination-country. Some scholars have argued that nowadays these variables might not be associated with language adoption anymore. This is because information and communication technologies have enabled the emergence of transnational identities as a new factor in the traditional patterns of migration and integration, assimilation, or diversity in host societies (Wright & Viggiano, 2020). In this paper, we consider the possibility that the use of social media may affect the language acquisition process. We consider this by exploring whether factors that are traditionally associated with language acquisition are also associated with language use on Twitter. Specifically, we explore whether the number of Twitter users from origin- and destination-country, linguistic distance between origin- and destination-language, and geographical distance between origin- and destination-country are associated with the time until an immigrant starts tweeting in the language of the destination-country. At the macro level, it is also important to consider that English is the most used second language in Europe (Bolton & Meierkord, 2013; Cromdal, 2013); therefore, we also control for the percentage of the host population that speak at least one foreign language.

---

[6] https://developer.twitter.com/en/docs/twitter-api/v1/tweets/search/api-reference/get-search-tweets. Accessed on May 1, 2021.

[7] https://archive.org/. Accessed October 28th, 2020.

Hypotheses

Based on the foregoing, we formulate the following two sets of hypotheses. The first set of hypotheses concerns the effect that civic-integration requirements and citizenship policies have on language acquisition. Here the first hypothesis consists of two competing alternatives. On the one hand, the less liberalized the citizenship policies the greater the time to acquire the language (H1.1a). This is because low levels of liberalization are related with high levels of anti-immigrant attitudes, which implies low chances for immigrants to use the host-language. On the other hand, the more liberalized the citizenship policies the greater the time to acquire the language (H1.1b). This is because high levels of liberalization imply policies that tend to favor linguistic pluralism, which implies low incentives for immigrants to learn or use the host-language. The second hypothesis holds that the more integration requirements, the quicker the immigrant would learn the language, as there are more incentives to learn the language of the host-country (H1.2). Finally, our third hypothesis of this set holds that there is an interaction between civic-integration requirements and citizenship policies; the more liberalized the citizenship policies and the more civic-integration requirements, the faster immigrants learn the host language (H1.3). This is because more civic-integration requirements mean more incentives for migrants to learn the language, while more liberalization means more tolerance towards migrants, which should make the host population more open to interact with migrants.

Our second set of hypotheses are a direct consequence of the associations described in the subsection Language and Social Media Usage. First, we expect that the more Twitter users from the origin-country in the platform, the slower the pace of language acquisition (H2.1). This is because, on the one hand, migrants' language adoption was inverse to the number of immigrants from the same origin-country living in the host-country before the dawn of social network sites. On the other hand, Twitter does not have borders; therefore, a migrant can keep communicating with people from their origin-country despite living in another. However, Twitter users may have more incentives to tweet in a language when there is a bigger audience with whom they can interact using that language, which is in line with what traditional research has shown (Chiswick & Miller, 2001; Esser, 2006). Second, the greater the linguistic distance between the origin-country and the destination-country, the slower the pace of language acquisition (H2.2). This is because when a language is more difficult to learn, relative to the mother tongue, then it takes longer to use it.

Finally, the bigger the geographic distance between the origin-country and destination-country, the faster the pace of language acquisition (H2.3).

Data

In this section, we first describe the general sample of tweets that researchers have access to and that are stored in the Internet Archive, which we also describe. Second, we describe the sample of tweets use in this study and their characteristics. Third, we explain the steps we follow to prepare the data for analysis. To process the data, we used the programming language Python version 3.7 (Python, 2020).

Twitter and the Internet Archive

Twitter provides access to a free-current 1% sample of all the public tweets through the streaming Application Programming Interface (API) (Kumar et al., 2015), which has the parameters tweet-keywords, user-IDs, and geographical boundary. The Twitter streaming API returns at most 1% of all the tweets produced on Twitter and "[O]nce the number of tweets matching the given parameters surpasses 1 percent of all the tweets on Twitter, Twitter begins to sample the data returned to the user" (Kumar et al., 2015, pp. 40). Morstatter et al. (2013) compared the 1% sample from the Streaming API against the full tweets retrieved using the Twitter Firehose (which is the paid API version that returns the full tweets that matched the target characteristics). Morstatter et al. (2013) retrieved data for 28 days and used as parameters specific keywords, user-IDs, and as bounding-box the geo-location of Syria. They found that the representativeness of the sample at Twitter tweets level is negatively associated with the number of parameters to match, where the streaming API is not representative of the trending topics of Twitter at that moment. For geolocation, when a bounding-box is used as parameter, the streaming API returns almost the complete set of the geo-tagged[8] tweets despite sampling. If the bounding box is not used, then the sample coverage follows a similar distribution as the one from the full tweets.

The streaming API has two main limitations. First, it does not return demographic characteristics of the users, such as age, gender, and level of education. For this, researchers have relied on pattern

[8] Geo-tagged means that the user shares their geolocation.

recognition software to extract users' demographic characteristics depending on their profile picture, username, and tweets (Leak et al., 2018; Mejova et al., 2015; Yin et al., 2018). Second, the streaming API does not allow users to retrieve tweets older than seven days. To retrieve older tweets, researchers have relied on historical samples gathered by specific organizations, such as the Internet Archive[9].

The Internet Archive is a repository that contains a 1% real-time sample of Tweets[10] collected every hour from 2011 to 2018 through the Twitter streaming API. According to Sequiera and Lin, (2017), the Twitter databases stored in the Internet Archive are a good replacement from those retrieved using the Twitter streaming API. There are no significant differences among these databases and only 5% of the tweets from the Internet Archive were missing in comparison with those retrieved using the Twitter streaming API. Data from the Internet Archive has been used to evaluate the consistency of Twitter data for migration estimates. The research concluded that the data can be used to analyze long-term and seasonal migration, as long as a temporal window (buffer) greater than 12 weeks is used to classify users as migrants (Fiorio et al., 2020).

Processing the Data

The Twitter streaming API returns three different variables from which the users' geo-location can be inferred[11]: geo, place, and location. The variable 'geo' consists of the coordinates from which the tweet was sent. The variable 'place' contains the country, country-code, and bounding box of coordinates– i.e., four coordinates– from which the tweet was sent. The variable 'location' contains either a user-self-written description or the geo-location of where the user is currently living. In our work, we only use the first two, 'geo' and 'place'. If the tweet contains the information of either 'geo' or 'place', then our algorithm extracts the country code. If the country code is missing but the coordinates are given, then the algorithm uses the package reverse_geocoder (Thampi, 2016) to transform coordinates into country code. From the 2.64 terabytes of Tape Archive File

[9] https://archive.org/. Accessed October 28th, 2020.

[10] https://archive.org/details/twitterstream. Accessed October 28th, 2020.

[11] https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/overview/geo-objects. Accessed October 28th, 2020.

(TAR) tweets processed, 4% contained geo-location information, which is similar to what Morstatter et al. (2013) have reported.

To classify users as migrants we follow the next four steps. First, we look for all the tweets in the filtered data coming from the same user and keep only those users that have tweeted at least five times in a year. We use this lower bound primarily because we need to capture the moments in which a user moves and when they start tweeting in another language. Users who tweet often produce more fine-grained data which in turn can be analyzed better. A secondary benefit of this lower bound is that it is less computationally demanding, as we need to build a dictionary to store all the paths to the tweets for each user. This lower bound has been used in similar studies (Lamanna et al., 2018), and the computational magnitude of this secondary benefit should not be underestimated. The outcome of this first step are new datasets containing the paths to the tweets by user. Second, from the sample produced in the first step, we select all the users that tweeted from more than one country.

Third, we categorize a user as a migrant if the user tweets for at least three months from one country and for at least the last three months from a second country. As an example, imagine a user who starts tweeting in Mexico and then moves to Germany and, therefore, changes the geo-location of their tweets from Mexico to Germany. This user would be classified as a migrant whose origin-country is Mexico and destination-country is Germany. We chose a window of at least three months following the argument by Fiorio et al. (2020) that the data can be used to analyze long-term migration, if a temporal window (buffer) greater than 12 weeks is used to classify users as migrants. Finally, from the sample, we keep the migrants that moved to one of the EU-15 countries and for whom the official language from their origin-country and destination-country are different. This gives us a final database of around 1,210 unique users and around 35,448 tweets. In order to classify the language used in their tweets, we use the package pycld2 (Al-Rfou, 2019) together with our own algorithm (Appendix A). From this processed data, we aggregate the information by month and frequency.

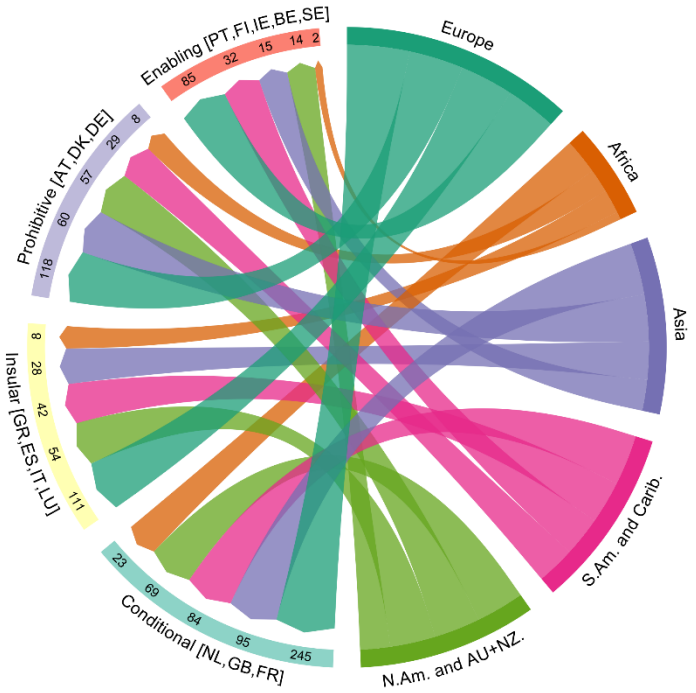

Figure 1: Migration flows from the regions of origin to EU-15 civic-integration groups. Numbers represent migration flows. The country codes of the receiving countries are in square brackets (AT: Austria, BE: Belgium, DE: Germany, DK: Denmark, ES: Spain, FI: Finland, FR: France, GB: Great Britain, GR: Greece, IE: Ireland, IT: Italy, LU: Luxembourg, NL: Netherlands, PT: Portugal, SE: Sweden). AU and NZ correspond to Australia and New Zealand, respectively.

To the final database, we add for each user the time (in months) between arriving in the country of destination and the time they started tweeting in the official language of destination for mostly one month. In other words, once their tweets reflect the geolocation of the destination-country; and once the number of times they tweeted in the host-country language was higher than 50% of their monthly tweets. The median tweets per user was 20, and the median number of months per user was 10. Users' countries of origin are quite diverse; in total our sample contains 81 countries of origin. Given this large diversity, we categorize their origins in five regions to visualize them: Europe; Africa; Asia; South America and the Caribbean (S.Am. and Carib.); and North America, Australia, and New Zealand (N.Am. and AU+NZ). The total number of individuals that migrate from these regions are 564, 46, 205, 195, and 200, respectively. Figure 1 shows the migration flows from these regions of origin to the civic-integration clusters described before: prohibitive, enabling, insular, and conditional. From here there was not any user that could be classified as an immigrant to Luxembourg, but we kept the code in the figure as is part of the Insular group. The total number of immigrants they receive are 276, 152, 243, and 539, respectively. Table B of

Appendix B shows the number of immigrants by country of destination and the percentage that started to tweet in a destination-country official language.

We also check the distribution of the users by gender and account status depending on the group (Table1). From here, we expect that the distributions are similar in each of the groups; otherwise, this would mean that the sample of users that the Twitter API returns is biased to certain regions. Table 1 shows that the percentage of Female, Male and Unknown users are equally distributed in the groups; and this is also the case for the current users' account status. The percentages of female and male users are similar to what others have reported (Zagheni et al., 2014); this is also the case for the percentages of deleted and suspended accounts (Armstrong et al., 2021).

Table 1: Percentage of users by gender and current account status.

		Gender (%)			**Account-Status (%)**		
Group	**Total**	Female	Male	Unknown	Active	Deleted	Suspended
Conditional	539	32.84	58.25	8.90	77.36	18.74	3.89
Insular	243	37.04	53.49	9.46	77.78	20.16	2.06
Enabling	152	29.60	60.52	9.86	75	22.37	2.63
Prohibitive	276	34.42	59.42	6.15	80.43	15.59	3.99

Notes: Users' gender was inferred from their user-names using the databases Social Security Administration (2020) and Demografix ApS (2021). Account-status corresponds to what the Twitter API V2.2 returned in August 10, 2021.

Before the analysis, we validate these users are (were) migrants by performing a qualitative analysis of a 10% sample of users. From which we analyze their tweets and their tweets metadata, as suggested by Armstrong et al. (2021). The qualitative analysis shows that some of the users tweeted as students in a foreign country, and others became residents in the new country. For the students, this is deduced from their tweets, as they share their experiences as newcomers in the country. For the residents, this is deduced from their tweets and, for some of them, from their current Twitter status profile, where they share they are from country A currently living in country B. For a small proportion of them, we could not infer users' motivations to move, but we kept them for the analysis.

Methodology

We model the variable ***T:*** *time until a user mostly tweets in the language of destination for one month* using survival models ($S(t)$). Where *mostly* means: the number of times a user tweeted in

the host-country language was higher than 50% of their monthly tweets. For this analysis, we use the programming language R (R Core Team, 2020) and the *survival* package (Therneau & Grambsch, 2000).

We first plot the Kaplan-Meier curve and then check which parametric model (such as Exponential, Weibull, or Gamma) fits the Kaplan-Meier curve best. We test the linearity of the Kaplan-Meier survival values by plotting $\ln(-\ln(\hat{S}(t)))$ vs $\ln(t)$ (Kleinbaum & Klein, 2012, pp. 305). This visual test shows that the best model is Weibull, as the values show a linear behavior and the slope of the line is different from one (Appendix C, Fig. C1).

The Weibull parametrization we follow is given by Kleinbaum and Klein (2012) (Eq. 1).

$$S(t) = exp(-\lambda t^{p}) \quad \ldots \quad \text{Equation 1}$$

To study the factors that enhance language acquisition, we model the Accelerated Failure Time (AFT) ratios of $\boldsymbol{T}$. We decided to use this model because the results are interpreted as the median survival time to acquire the language, which we consider to be more directly interpretable than proportional hazards.

We model time until a user mostly tweets in the language of destination for one month as a function of the following seven variables. First, CPI, where the higher the value the more liberal the citizenship requirements of the destination country. Second, CIVIX, where the higher the value the stronger the integration requirements of the country of destination. These two variables are continuous and range from 0 to 6. Third, an interaction term between both CPI and CIVIX. This variable is continuous and range from 0 to 36. We do not transform any of these variables in order to facilitate interpretation, given the interaction term.

Fourth, the logarithm of the ratio of Twitter users in the country of origin to Twitter users in the country of destination. Here a positive value means there are more Twitter users in the origin-country than in the destination-country, and vice-versa if the value is negative. The fifth and sixth variables are linguistic distance and geographic distance. These variables come from the databases "Language" and "Gravity", respectively, from the Centre d'Études Prospectives et d'Informations

Internationales[12]. Linguistic distance is the variable LP2, which according to Melitz and Toubal (2014), the smaller the value the closer the languages are in terms of vocabulary and grammar. Geographical distance is the distance in km between the capitals of origin- and destination-countries[13]. Finally, the percentage of the destination population that self-reported to know at least one foreign language (EUROSTAT, 2011). These variables are continuous and standardized (meaning we subtract the mean and divide by the standard deviation).

The Weibull AFT function is $t = [-\ln S(t)]^{1/p}\,\lambda^{-1/p}$ where $\lambda^{-1/p}$ is parametrized with regression coefficients (Eq. 2) (Kleinbaum & Klein, 2012, pp. 308). In general, the AFT is a ratio of survival times corresponding to any quantile (q) of survival time ($S(t) = q$). In this model, an increase in a variable which coefficient is positive leads to an increase in the median (or other quantile) survival time of acquiring the language. If the coefficient is negative, then an increase in the variable would lead to a decrease in the median survival time of acquiring the language.

$$\lambda_i^{-\frac{1}{p}} = \exp(\alpha_{0i} + \alpha_{1i}CPI + \alpha_{2i}CIVIX + \alpha_{3i}CPI \times CIVIX + \alpha_{4i}\log(ratio) +$$

$$\alpha_{5i}\,Ling.Dist. + \alpha_{6i}\,Geo.Dist. + \alpha_{7i}\,\%\geq 1\,Foreign\,Lang.)\ldots \text{ Equation 2}$$

Where *CPI* is the Citizenship Policy Index; *CIVIX* is the Civic Integration Index; *CPI*×*CIVIX* is the interaction term; *log(ratio)* is the logarithm of the ratio of the number of Twitter users from origin country by the number of Twitter users from destination country; *Lin. Dist.* is linguistic distance; *Geo. Dist.* is geographical distance; and *%≥1 Foreign Lang.* is the percentage of destination country population that speaks more than one foreign language.

Translating our hypotheses to the results of the AFT model leads to the following expected findings. For the first main hypotheses, on the one hand, a more liberalized citizenship policy resulting in greater median survival time to acquire the language would corroborate hypothesis H1.1a. On the other hand, a less liberalized citizenship policy resulting in a greater median survival

[12] http://www.cepii.fr/CEPII/en/bdd_modele/presentation.asp?id=19. Accessed April 12th, 2021.

[13] http://www.cepii.fr/DATA_DOWNLOAD/gravity/doc/Gravity_documentation.pdf. Accessed May 5th, 2021.

time would corroborate hypothesis H1.1b. The second main hypothesis (H1.2) is supported if the stronger the integration requirements the smaller the median time to acquire the language. The last main hypothesis (H1.3) holds, if the median survival time to acquire the language decreases when the more liberalized the citizenship policies and the stronger the integration requirements.

In the case of our secondary hypotheses, if the more Twitter users from country of origin using the platform results in a greater time to acquire the language, then hypothesis H2.1 is supported. If the greater the linguistic distance between the origin-country and the destination-country the greater the median survival time to acquire the language then hypothesis H2.2 is corroborated. Finally, if a bigger geographic distance between the country of origin and country of destination results in a smaller median survival time to acquire the language then hypothesis H2.3 is supported.

Results

Figure 2 shows the estimated AFT ratios of the Weibull model with 95% confidence intervals. In the case of CIVIX, the stronger the civic-integration requirements the larger the median survival time of acquiring the language; therefore, H1.2 is not supported. In the case of CPI, it does not appear to play a role on the median survival time of acquiring the language conditional on the other variables in the model. However, the interaction variable shows that the greater the CPI and CIVIX the smaller the median survival time of acquiring the language. This indicates that CPI does play a role, but only in conjunction with particular CIVIX levels.

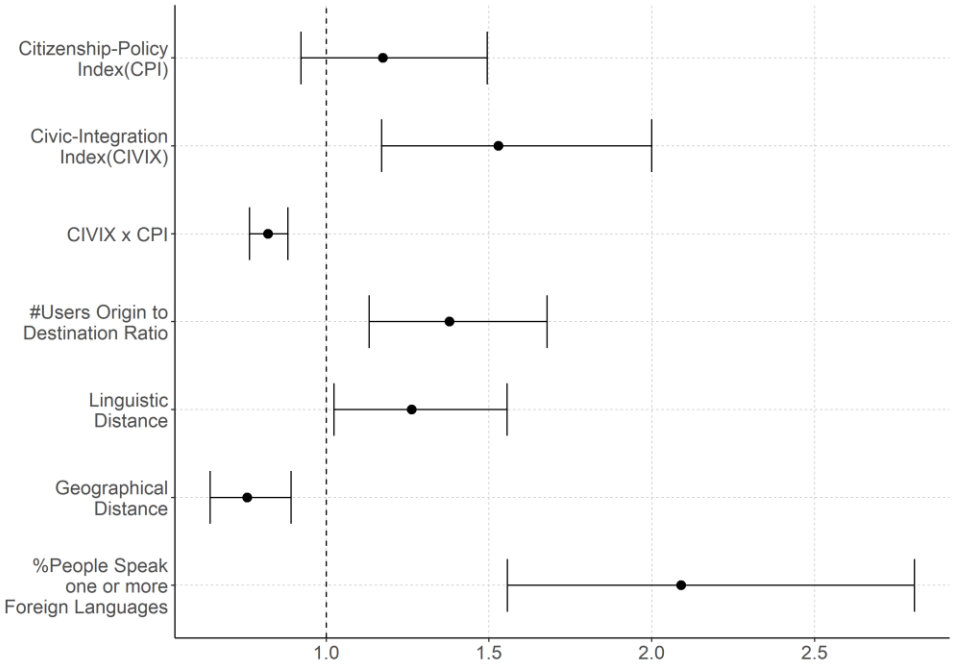

Figure 2 Exponential of AFT Ratio Coefficients with their corresponding 95% Confidence Intervals.

To clarify the interaction effect, we show the predicted survival time of acquiring the language using a contour map relative to the CIVIX and CPI indexes in Figure 3. These predicted values are obtained by multiplying the model coefficients with the different combinations of CIVIX and CPI values, while keeping the rest of the variables constant on their means (which are zero because of the standardization). Figure 3 shows that when the CIVIX index is below 1.75, the CPI index is not associated with the median survival time of immigrants' language acquisition. This can be seen on the median survival values of the insular and enabling groups (excluding Sweden), which range between 55 and 148 months regardless of the CPI values. Once the CIVIX values are over one, the CPI index becomes associated with the median survival time of immigrants' language acquisition; the more liberalized the country the smaller the median times of immigrants' language acquisition (conditional group); and the less liberalized the higher the median times of immigrants' language acquisition (prohibitive group). In this analysis, Sweden seems to follow a different pattern from the rest of the countries. This could be related with the high percentage of its population that speaks at least one foreign language, in comparison with the low levels in the rest of the countries in the insular and enabling groups.

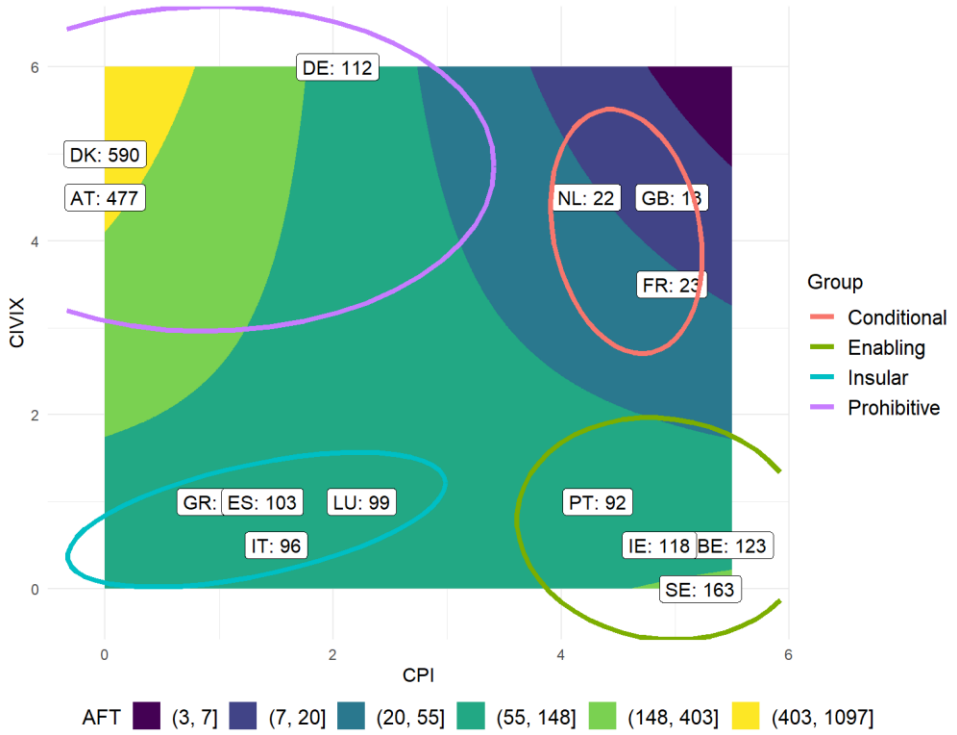

Figure 3 Contour map of the Accelerated Failure Time (AFT) Ratios relative to Civic Integration Index (CIVIX) and Citizenship Policy Index (CPI).

Returning to Fig. 2, in line with our secondary hypotheses, the median survival time of adopting the language increases if the number of Twitter users in the country of origin is bigger than in the country of destination (H2.1). Linguistic distance also has a positive association, this means that the bigger the distance between the origin- and destination-language the higher the time to acquire the destination language (H2.2). For geographic distance, the smaller the distance the smaller the median-survival time of acquiring the language (H2.3). Therefore, the classic variables used to explain immigrants' language acquisition have the same associations with language acquisition as before the dawn of social network sites. Finally, for the control variable, an increase of a percentage point of people speaking a foreign language doubles the mean number of months of acquiring the language.

Discussion and Conclusions

In this work, we study immigrants' language acquisition through a longitudinal analysis of their tweets language. To do so, we draw on Goodman's (2010) and Howard's (2010) work to formulate

how citizenship and civic-integration policies may affect immigrants' language acquisition. Conceptually, we rely on the governmentality framework that theorizes on the effects governmental interventions have on individuals. We use survival models to analyze immigrants' language-acquisition pace depending on: (1) citizenship and civic-integration policies; (2) relative size of migrant groups in the destination country and the linguistic and geographical distance between countries of origin and destination. Specifically, we analyze the time until a user mostly tweets in the language of destination for one month. We use starting to tweet in the language of the country of destination as a proxy of language acquisition.

Our findings point to an interaction effect between immigrants' civic-integration requirements and citizenship access liberalization, where immigrants in countries with low or not civic-integration requirements share similar median times of language acquisition regardless of how liberalized citizenship policies are. This is the case for countries with heterogeneous citizenship policies, but that have few civic-integration requirements. However, among these countries, Sweden is a particular case. Sweden seems to share immigrants' median times of language acquisition as countries with strict civic-integration and citizenship requirements. This could be explained by: the high percentage of Swedish population that speaks at least one foreign language (Bolton & Meierkord, 2013); or the high levels of multiculturalism, which could disincentivize immigrants from learning Swedish (van Tubergen & Kalmijn, 2005).

In the case of countries with strict civic-integration and citizenship requirements (Denmark, Austria, and Germany), immigrants take the highest times to acquire the language. While this may be a consequence of the anti-immigrant attitudes from the majority population, it has also been proposed that strict requirements imposed to immigrant groups may be also a consequence of right-wing parties that try to constrain immigrants' access to equal rights as natives (M. B. Jørgensen, 2009; Beauzamy & Féron, 2012; Bolton & Meierkord, 2013; Lønsmann, 2020). Research shows that these type of negative interactions between authority and migrants can lead to immigrants' language balkanization and rejection towards learning the language of destination (Jørgensen, 2003).

For those countries with high civic-integration requirements, we found that the more liberalized their citizenship access the faster the immigrants acquire the host-country language. This is the case for France, the Netherlands and the United Kingdom, which have high civic-integration

requirements, but are also considered historically liberal countries (Howard, 2010). Though, this might also be explained by the early integration requirements that immigrants fulfill, such as learning the language before moving to the country (Goodman, 2010).

Our results also show that immigrants' language acquisition in Twitter is associated with the same classic macro-level explicative variables employed before the dawn of social network sites. This result is relevant in two ways. On the one hand, it supports the notion that the data from Twitter is actually capturing migration. On the other hand, it helps to shed light on whether the transnational property of social network sites has affected the association between immigrants' language acquisition and classic macro-explicative variables. For this sample of Twitter users the results show that this has not been the case. However, this might change in the future, as the use of information and communication technologies are becoming more and more pervasive in the world.

Limitations

This work has several limitations that we would like to acknowledge. First, Twitter data is not representative of the general population. Twitter users tend to be young adult men that are highly educated and that are highly internet skilled (Hargittai, 2020). Furthermore, because our work depends on longitudinal Voluntarily Geographic Information (Haklay, 2016), the analysis is constrained to highly active users that are considered as the content producers. Despite these clear data limitations, we show that Twitter data can be used to study immigrants' language acquisition and that the data shows patterns that had been found in work done with representative samples collected before the dawn of social network sites (Chiswick & Miller, 2001; Esser, 2006). While the findings have to be interpreted with caution, it is important to continue the study of the use of statistical techniques to model data from social network sites to study hard to reach populations, such as migrants.

Research Ethics

This work obtained ethical approval from the data protection department of the Max Planck Institute for Demographic Research and the Max Planck Society. For the analyses, we rely on public data from the Internet Archive and we study only the language from the users' tweets. For

a 10% sample of the users, for research purposes only, we also read the public description of their profiles.

Acknowledgments

We would like to thank Lee Fiorio (Univ. Washington), Clara Mulder (Univ. of Groningen), and Emanuele del Fava (MPIDR) for their feedback.

Reproducibility

Given Twitter terms and conditions, we do not share the final database, as this can lead to user-IDs disclosure. All the code to replicate this work are available in Gil-Clavel's Github repository.

Appendix A: Language Classification

First, let's establish some terminology, if a tweet's language is classified by reverse_geocoder as unknown then it is labeled as *un*, when it is not then it is labeled *code_i*, where *i* is an index that indicates one of the possible languages used by the user. Then the next rules apply:

- If the first user' tweet language is *un*, but the next one is *code_1*, then *un* becomes *code_1*.
- If the last user' tweet language is *un*, but the previous one is *code_1*, then un becomes *code_1*.
- If for a given user, there is a tweet language sequence like: *code_1-un-code_1*, then *un* becomes *code_1*.

The use of reverse_geocoder together with the aforementioned rules, result in the next language distribution.

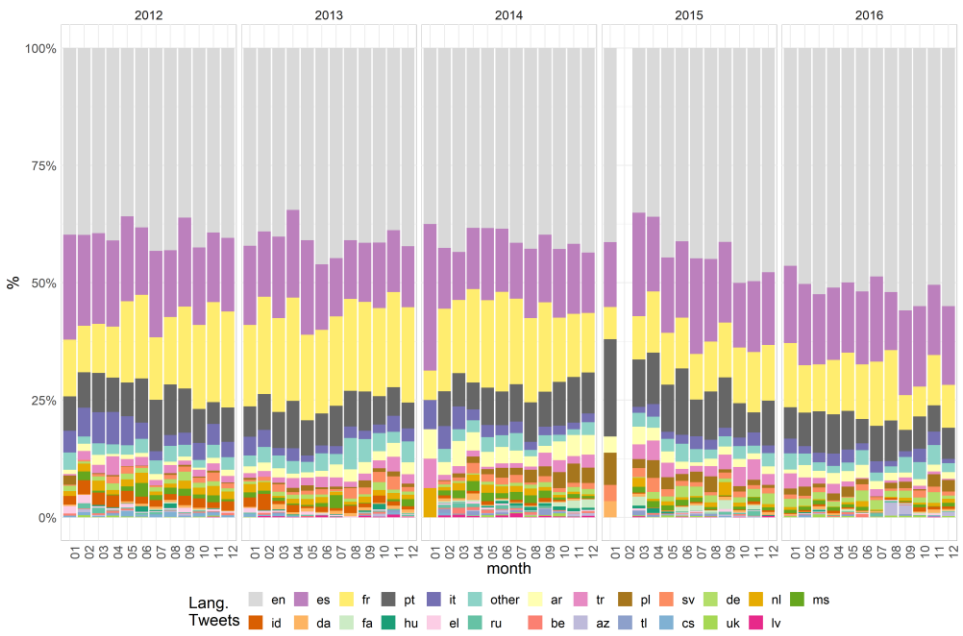

Figure A: Proportion of tweets language by month and year. Color corresponds to language. Langages are ordered by the total frequency of appearence.

Appendix B: Users classified as Immigrants by Destination-Country

Table B: Number of Twitter users classified as immigrants (n) and the percentage that started to tweet in a destination-country official language (%) by destination-country.

Austria		Belgium		Germany		Denmark		Spain		Finland		France	
n	%	n	%	n	%	n	%	n	%	n	%	n	%
25	8	23	9	226	8	21	29	129	29	21	0	165	32
Great Britain		Greece		Ireland		Italy		Netherlands		Portugal		Sweden	
n	%	n	%	n	%	n	%	n	%	n	%	n	%
284	69	12	0	45	44	102	20	67	12	18	28	41	22

Appendix C: Weibull Model Goodness of Fit

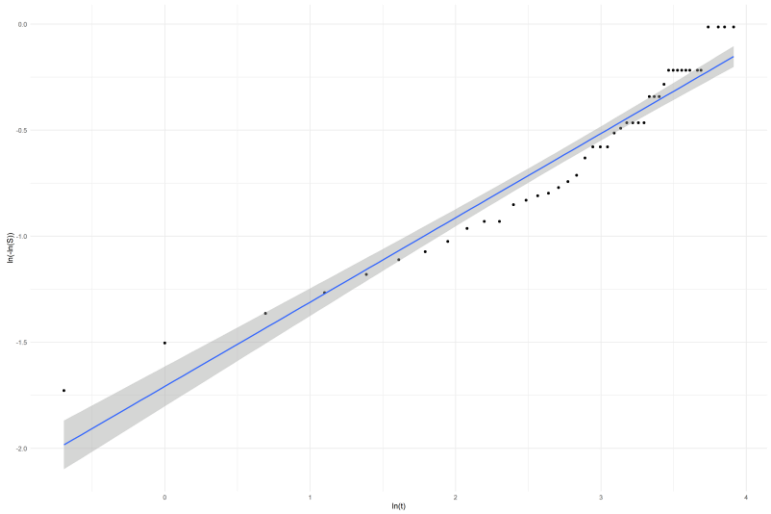

Figure C1: Visual test for the use of Weibull model for Kaplan-Meier survival values.

Table C1: Result from the linear regression model testing Weibull suitability

R^2	Term	$\hat{\beta}$	Std. Error	p-value
0.9377	Intercept	-1.70789	0.04644	<0.001
	ln(t)	0.3975	0.01581	<0.001

Note: $\lambda = exp(\hat{\beta}_{Intercept})$ and $p = \hat{\beta}_{ln(t)}$.

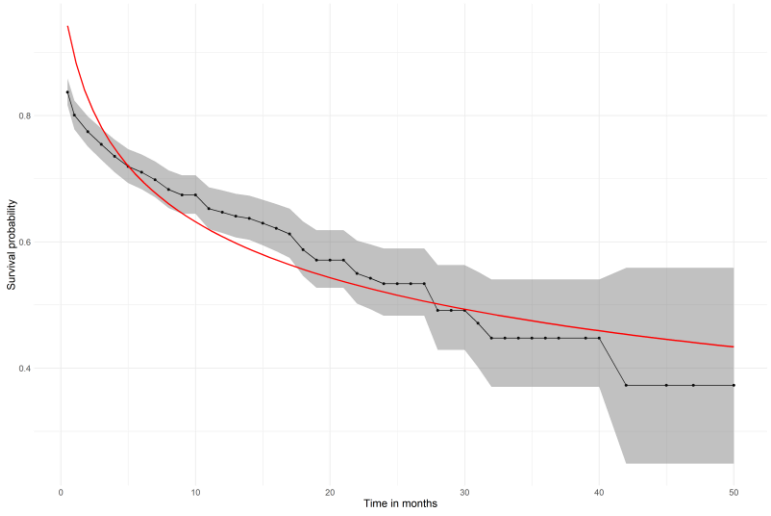

Figure C2: The red curve is the Weibull adjusted model for the Kaplan-Meier survival curve. The 95% Confidence Intervals correspond to the Kaplan-Meier curve.

Figure B2 shows the Weibull model fitted to the Kaplan-Meier survival curve. The results of the nonlinear least-squares estimate of the parameters of the Weibull survival model are: $\lambda = 0.1003$ and p=0.3502.

References

Algan, Y., Bisin, A., Manning, A., & Verdier, T. (Eds.) (2012a). Cultural integration of immigrants in Europe (p. 359). Oxford University Press.

Algan, Y., Landais, C., & Senik, C. (2012b). Cultural Integration in France. In Y. Algan, A. Bisin, A. Manning, & T. Verdier (Eds.), *Cultural integration of immigrants in Europe*. Oxford University Press.

Al-Rfou, Rami. 2019. "PYCLD2." https://pypi.org/project/pycld2/.

Odero, A., Karathanasi, C., & Baumann, M. (2016). The Integration Process of Non-EU Citizens in Luxembourg: From an Empirical Approach Toward a Theoretical Model. International Journal of Humanities and Social Sciences, 9(11), 4066-4073.

Armstrong, C., Poorthuis, A., Zook, M., Ruths, D., & Soehl, T. (2021). Challenges when identifying migration from geo-located Twitter data. EPJ Data Science, 10(1), 1.

Beauchemin, C. (2014). *A Manifesto for Quantitative Multi-sited Approaches to International Migration*. International Migration Review, 48(4), 921–938. https://doi.org/10.1111/imre.12157.

Beauzamy, B., & Féron, E. (2012). *Otherism in Discourses, Integration in Policies?: Comparing French and Danish educational policies for migrants*. Nordic Journal of Migration Research, 2(1), 66. https://doi.org/10.2478/v10202-011-0028-7

Bolton, K., & Meierkord, C. (2013). *English in contemporary Sweden: Perceptions, policies, and narrated practices*. Journal of Sociolinguistics, 17(1), 93–117. https://doi.org/10.1111/josl.12014

Brochmann, G., & Hagelund, A. (2011). *Migrants in the Scandinavian Welfare State: The emergence of a social policy problem*. Nordic Journal of Migration Research, 1(1), 13. https://doi.org/10.2478/v10202-011-0003-3.

Bruzos, A., Erdocia, I., & Khan, K. (2018). *The path to naturalization in Spain: Old ideologies, new language testing regimes and the problem of test use*. Language Policy, 17(4), 419–441. https://doi.org/10.1007/s10993-017-9452-4

Cairns, I., & Shetty, P. (2020, July 16). Introducing a new and improved Twitter API. https://blog.twitter.com/developer/en_us/topics/tools/2020/introducing_new_twitter_api.

Castles, S., De Haas, H., & Miller, M. J. (2013). *The Age of Migration International Population Movements in the Modern World*. 5th ed. Palgrave Macmillan UK.

Castro-Gómez, S. (2010). Siglo XVIII: el nacimiento de la biopolítica. Tabula Rasa, (12), 31-45.

Chiswick, B. R., & Miller, P. W. (2001). A model of destination-language acquisition: Application to male immigrants in Canada. Demography, 38(3), 391-409.

Costa, J. (2012, September). Current status: API v1.1. https://blog.twitter.com/developer/en_us/a/2012/current-status-api-v1-1.

Cromdal, J. (2013). Bilingual and second language interactions: Views from Scandinavia. International Journal of Bilingualism, 17(2), 121–131. https://doi.org/10.1177/1367006912441415

De Haas, H., Castles, S., & Miller, M. J. (2020). The age of migration: International population movements in the modern world (6th ed.). The Guildford Press.

Duncan, H. (2020). Trends in international, national and local policies on migrant entry and integration. In C. Inglis, W. Li, & B. Khadria (Eds.), *The Sage handbook of international migration* (pp. 592-607). SAGE Publications Ltd, https://dx.doi.org/10.4135/9781526470416.n40.

Eckert, E. (2018). Immigration, Language, and Conflicting Ideologies: The Czech in Texas. In S. D. Brunn & R. Kehrein (Eds.), *Handbook of the Changing World Language Map*. (pp. 1–17). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-73400-2_31-1.

Edele, A., Seuring, J., Kristen, C., & Stanat, P. (2015). Why bother with testing? The validity of immigrants' self-assessed language proficiency. Social science research, 52, 99-123. https://doi.org/10.1016/j.ssresearch.2014.12.017.

Esser, H. (2006). Migration, language and integration. WZB.

EUROSTAT. (2011). Statistics | Eurostat [Data Browser]. Number of Foreign Languages Known (Self-Reported) by Sex. https://ec.europa.eu/eurostat/databrowser/view/EDAT_AES_L21/default/table?lang=en&category=sks.sks_ssr.sks_ssaes.edat_aes_l2

Fiorio, L., Zagheni, E., Abel, G., Hill, J., Pestre, G., Letouzé, E., & Cai, J. (2021). *Analyzing the Effect of Time in Migration Measurement Using Georeferenced Digital Trace Data*. Demography, 58(1), 51–74. https://doi.org/10.1215/00703370-8917630.

Font, J., & Méndez, M. (2013). Introduction: The methodological challenges of surveying populations of immigrant origin. In J. Font & M. Méndez (Eds.), *Surveying Ethnic Minorities and Immigrant Populations: Methodological Challenges and Research Strategies*. Amsterdam University Press. https://doi.org/10.26530/OAPEN_450851.

Forrest, J., Benson, P., & Siciliano, F. (2018). Linguistic Shift and Heritage Language Retention in Australia. In S. D. Brunn & R. Kehrein (Eds.), *Handbook of the Changing World Language Map* (pp. 1–18). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-73400-2_37-1.

Foucault, M (1991). Governmentality. In G. Burchell, C. Gordon, & P. Miller (Eds.), *The Foucault Effect: Studies in Governmentality* (pp. 87–104). University of Chicago Press.

Goodman, S. W. (2013). Integration requirements for integration's sake? Identifying, categorizing and comparing civic integration policies. In Migration and Citizenship Attribution (pp. 49-68). Routledge. https://doi.org/10.1080/13691831003764300.

Goodman, S. W. (2012). Fortifying citizenship: Policy strategies for civic integration in Western Europe. World Politics, 64(4), 659-698. https://doi.org/10.1017/S0043887112000184.

Haklay, M. E. (2016). Why is participation inequality important?. Ubiquity Press. https://doi.org/10.5334/bax.c.

Hargittai, E. (2020). Potential Biases in Big Data: Omitted Voices on Social Media. *Social Science Computer Review*, *38*(1), 10–24. https://doi.org/10.1177/0894439318788322.

Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., & Ratti, C. (2014). Geo-located Twitter as proxy for global mobility patterns. Cartography and Geographic Information Science, 41(3), 260-271. https://doi.org/10.1080/15230406.2014.890072.

Höhne, J. (2013). Language integration of labour migrants in Austria, Belgium, France, Germany, the Netherlands and Sweden from a historical perspective (No. SP VI 2013-101). WZB Discussion Paper.

Howard, M. M. (2010). The Impact of the Far Right on Citizenship Policy in Europe: Explaining Continuity and Change. Journal of Ethnic and Migration Studies, 36(5), 735–751. https://doi.org/10.1080/13691831003763922.

Jørgensen, J. N. (2003). Bilingualism and minority languages. *International Journal of the Sociology of Language*, *2003*(159). https://doi.org/10.1515/ijsl.2003.010.

Jørgensen, M. B. (2009). National and Transnational Identities: Turkish Organising Processes and Identity Construction in Denmark, Sweden and Germany [Aalborg Universitet]. Spirit PhD Series No. 19.

Kleinbaum, D. G., & Klein, M. (2012). Survival analysis: a self-learning text (Vol. 3). New York: Springer. https://doi.org/10.1007/978-1-4419-6646-9.

Komito, L. (2011). Social media and migration: Virtual community 2.0. Journal of the American Society for Information Science and Technology, 62(6), 1075–1086. https://doi.org/10.1002/asi.21517

Krishnamurthy, B., Gill, P., & Arlitt, M. (2008, August). A few chirps about twitter. In Proceedings of the first workshop on Online social networks (pp. 19-24). https://doi.org/10.1145/1397735.1397741.

Kumar, S., Morstatter, F., & Liu, H. (2015). Analyzing Twitter Data. In Y. Mejova, I. Weber, & M. W. Macy (Eds.), Twitter: A Digital Socioscope. Cambridge University Press.

Lamanna, F., Lenormand, M., Salas-Olmedo, M. H., Romanillos, G., Gonçalves, B., & Ramasco, J. J. (2018). Immigrant community integration in world cities. PloS one, 13(3), e0191612. https://doi.org/10.1371/journal.pone.0191612.

Lazer, D., & Radford, J. (2017). Data ex Machina: Introduction to Big Data. Annual Review of Sociology, 43(1), 19–39. https://doi.org/10.1146/annurev-soc-060116-053457.

Leak, A., Lansley, G., Longley, P., Cheshire, J., & Singleton, A. (2018). Geotemporal Twitter Demographics. In Consumer Data Research (pp. 152–165). UCL Press. http://www.jstor.org/stable/j.ctvqhsn6.14.

Li, T. M. (2007). Governmentality. Anthropologica, 49(2), 275-281. https://www.jstor.org/stable/25605363.

Lønsmann, D. (2020). Language, employability and positioning in a Danish integration programme. International Journal of the Sociology of Language, 2020(264), 49–71. https://doi.org/10.1515/ijsl-2020-2093

Love, S. V. (2015). Language testing, 'integration' and subtractive multilingualism in Italy: Challenges for adult immigrant second language and literacy education. Current Issues in Language Planning, 16.1(2), 26–42.

Manning, A., & Georgiadis, A. (2012). Cultural Integration in the United Kingdom. In Y. Algan, A. Bisin, A. Manning, & T. Verdier (Eds.), Cultural integration of immigrants in Europe. Oxford University Press.

Mazzoli, M., Diechtiareff, B., Tugores, A., Wives, W., Adler, N., Colet, P., & Ramasco, J. J. (2020). Migrant mobility flows characterized with digital data. PloS one, 15(3), e0230264.

McFEDRIES, P. (2007). Technically speaking: All a-twitter. IEEE spectrum, 44(10), 84-84. https://doi.org/10.1109/MSPEC.2007.4337670.

Mejova, Y., Weber, I., & Macy, M. W. (Eds.). (2015). Twitter: a digital socioscope. Cambridge University Press.

Melitz, J., & Toubal, F. (2014). Native language, spoken language, translation and trade. Journal of International Economics, 93(2), 351-363. https://doi.org/10.1016/j.jinteco.2014.04.004.

Menjívar, C., & Lakhani, S. M. (2016). Transformative effects of immigration law: Immigrants' personal and social metamorphoses through regularization. American Journal of Sociology, 121(6), 1818-1855. https://doi.org/10.1086/685103.

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.https://doi.org/10.1371/journal.pone.0061981.

Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. (2013). Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 7, No. 1, pp. 400-408).

Python Software Foundation. 2020. Python Language Reference, version 3.7. Available at http://www.python.org.

R Core Team. 2020. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Schierup, C.-U., Schierup, D. of the I. for R. on M. E. and S. C.-U., Hansen, P., & Castles, S. (2006). Migration, Citizenship, and the European Welfare State: A European Dilemma. OUP Oxford.

Sequiera, R., & Lin, J. (2017, August). Finally, a downloadable test collection of tweets. In Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 1225-1228). https://doi.org/10.1145/3077136.3080667.

Sharma, A. (2018). Migration, Language Policies, and Language Rights in Luxembourg. Acta Universitatis Sapientiae, European and Regional Studies, 13(1), 87–104. https://doi.org/10.2478/auseur-2018-0006.

Skourmalla, A.-M., & Sounoglou, M. (2021). Human Rights and Minority Languages: Immigrants' Perspectives in Greece. Review of European Studies, 13(1), 55. https://doi.org/10.5539/res.v13n1p55.

Thampi, Ajay. 2016. "Reverse Geocoder (Reverse_geocoder)." https://pypi.org/project/reverse_geocoder/.

Therneau, T. M., and Grambsch, P. M. (2000). Modeling Survival Data: Extending the Cox Model. New York: Springer.

Tsoukalas, S., Ntalianis, F., Papageorgiou, P., & Retalis, S. (2010, June). The impact of training on first generation immigrants: Preliminary findings from Greece. In 2010 2nd International Conference on Education Technology and Computer (Vol. 3, pp. V3-235). IEEE. https://doi.org/10.1109/ICETC.2010.5529555.

Twitter Inc. (2014). Twitter Reports Second Quarter 2014 Results. 7.

Twitter Inc. (2021a). Twitter Financial Releases. Financial Releases. https://investor.twitterinc.com/financial-information/financial-releases/default.aspx.

Twitter Inc. (2021b). Q2 2021 Letter to Shareholders. Twitter Inc. https://s22.q4cdn.com/826641620/files/doc_financials/2021/q2/Q2'21-Shareholder-Letter.pdf.

Van Tubergen, F., & Kalmijn, M. (2005). Destination-language proficiency in cross-national perspective: A study of immigrant groups in nine western countries. American Journal of

Sociology, 110(5), 1412-1457. American Journal of Sociology, 110(5), 46. https://doi.org/0002-9602/2005/11005-0005.

Vertovec, S. (2007). New complexities of cohesion in Britain: Super-diversity, transnationalism and civil-integration.

Wright, S. (2020). Migration, linguistics and sociolinguistics. In C. InglisW. Li, & B. Khadria (Eds.), *The Sage handbook of international migration* (pp. 142-158). SAGE Publications Ltd, https://dx.doi.org/10.4135/9781526470416.n10.

Wright, S., & Viggiano, C. (2020). Language and incorporation. In C. InglisW. Li, & B. Khadria (Eds.), The Sage handbook of international migration (pp. 481-495). SAGE Publications Ltd, https://dx.doi.org/10.4135/9781526470416.n32.

Yin, J., Chi, G., & Van Hook, J. (2018, November). Evaluating the representativeness in the geographic distribution of Twitter user population. In Proceedings of the 12th Workshop on Geographic Information Retrieval (pp. 1-2). https://doi.org/10.1145/3281354.3281360.

Zagheni, E., Garimella, V. R. K., Weber, I., & State, B. (2014, April). Inferring international and internal migration patterns from twitter data. In Proceedings of the 23rd international conference on world wide web (pp. 439-444). https://doi.org/10.1145/2567948.2576930.

Zimmer, M., & Proferes, N. J. (2014). A topology of Twitter research: Disciplines, methods, and ethics. Aslib Journal of Information Management, 66(3), 250–261. https://doi.org/10.1108/AJIM-09-2013-0083.