Gil-Clavel, Sofia; Grow, André; Bijlsma, Maarten J.

Article — Published Version

Migration Policies and Immigrants' Language Acquisition in EU-15: Evidence from Twitter

Population and Development Review

Provided in Cooperation with:
John Wiley & Sons

Suggested Citation: Gil-Clavel, Sofia; Grow, André; Bijlsma, Maarten J. (2023) : Migration Policies and Immigrants' Language Acquisition in EU-15: Evidence from Twitter, Population and Development Review, ISSN 1728-4457, Wiley, Hoboken, NJ, Vol. 49, Iss. 3, pp. 469-497, https://doi.org/10.1111/padr.12574

Migration Policies and Immigrants' Language Acquisition in EU-15: Evidence from Twitter

SOFIA GIL-CLAVEL, ANDRÉ GROW AND MAARTEN J. BIJLSMA

In response to the increasingly complex and heterogeneous immigrant communities settling in Europe, European countries have adopted various civic integration measures. Measures aiming to facilitate language acquisition are considered crucial for integration and cooperation between immigrants and natives. Simultaneously, the rapid expansion of social media usage is believed to change the factors affecting immigrants' language acquisition. However, only a few previous studies have analyzed whether this is the case. This article uses a novel longitudinal data source derived from Twitter to (1) analyze differences in the pace of immigrants' language acquisition depending on the migration policies of destination countries and (2) study how the relative sizes of the migrant groups in destination countries, and the linguistic and geographical distances between origin and destination countries, are associated with language acquisition. Results show that immigrants who live in countries with strict language acquisition requirements for immigrants and conservative citizenship policies have the highest median times until language acquisition. Based on Twitter data, we also find that language acquisition is associated with classic explanatory variables, such as the size of the immigrant group in the destination country and the linguistic and geographical distance between origin and destination country similar to the previous studies.

Introduction

Since the beginning of the 21st century, policymakers across Europe have attempted to enforce the requirement that immigrants learn the

Sofia Gil-Clavel, Delft University of Technology, Max Planck Institute for Demographic Research, University of Groningen. E-mail: B.S.GilClavel@tudelft.nl.

André Grow, Max Planck Institute for Demographic Research. E-mail: grow@demogr.mpg.de.

Maarten J. Bijlsma, University of Groningen, Max Planck Institute for Demographic Research. E-mail: maarten.bijlsma@rug.nl.

national language through civic integration policies (Wright and Viggiano 2020). This was a reaction to the settlement of increasingly complex and heterogeneous immigrant communities in Europe, a phenomenon that has been called "superdiversity" (Vertovec 2007). Civic integration policies rest on the assumption that the successful incorporation of immigrants into the host society must go beyond their economic and political incorporation and should rely "also on individual commitments to characteristics typifying national citizenship, specifically country knowledge, language proficiency, and liberal and social values" (Goodman 2010, 754). As language acquisition is often regarded as critical for the integration of immigrants, and for cooperation between immigrants and natives (Eckert 2018; Forrest, Benson, and Siciliano 2018), many integration measures aim to facilitate language acquisition (Duncan 2020). Moreover, it is assumed that migrants who know the country's language are familiar with its culture and are therefore sufficiently integrated into the country (Goodman 2010). However, little is currently known about how such civic integration measures affect language acquisition. This is in part because of a lack of multinational data that can be used to compare the effects of different civic integration measures on different migrant groups (Frank van Tubergen and Kalmijn 2005). To address this knowledge gap, we use data on language use obtained from Twitter for the period from January 2012 to December 2016. We study the pace of migrants' acquisition of the destination country language 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 nonintrusive way among immigrants to the EU-15 countries with a large number of countries of origin. Unlike traditional data used in migration research, Twitter data can provide researchers with continuous access to transnational and comparable migration data. Because of these properties, Twitter data have been used to study different aspects of migration. For example, Mazzoli et al. (2020) showed that geo-located Twitter data can be used to monitor the 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) used data from 500,000 geo-located tweets to estimate migration flows from Twitter users in OECD countries, and Hawelka et al. (2014) used geo-located tweets to uncover global patterns of human mobility. While Twitter data have been used less frequently in research on integration, Lamanna et al. (2018) showed 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.

Following Lamanna et al. (2018), we study immigrant integration patterns by analyzing the language they use in their tweets. Studies have shown that there is a positive correlation between language acquisition and language usage (F. van Tubergen and Kalmijn 2008). This is because

immigrants who use the host language in day-to-day contexts are better able to learn it and because those who learn the language better use it more in their everyday life. Therefore, 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 about 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 classifying the EU-15 countries according to their requirements for civic integration and citizenship. Conceptually, we rely on the governmentality framework that theorizes the effects of governmental interventions 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). In our analysis, one complicating factor 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, factors that affected language acquisition in the past, such as the geographical distance between the origin and the destination country, may lose their importance (Komito 2011; Wright and Viggiano 2020). To assess this possibility, we also study the effects 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, as acquisition of the host country language facilitates cooperation between immigrants and natives (Eckert 2018; Forrest, Benson, and Siciliano 2018). Indeed, for immigrants, mastering the language of the destination improves their access to education and important institutions and is associated with higher income, more societal recognition, and more social contacts (Duncan 2020). Thus, learning the language of the host country facilitates the acquisition of human capital in the country of destination (Esser 2006). Because language acquisition plays a central role in the integration process, it has always been considered an important variable in the study of immigrant integration (Algan, B et al. 2012; Esser 2006), and it has been the focus of civic and integration policies (De Haas, Castles, and Miller 2020; Wright 2020).

The role of civic integration policies

Foucault (1991) was among the first to theorize that governmental programs have the capacity to change the behavior of the population. This

notion is captured in the term *governmentality*, which refers to the different effects governmental interventions have on individuals depending on their positions in relation to governmental programs (Li 2007). These interventions may be related to poverty, health, and demographic events, such as migration and fertility (Castro-Gómez 2010; Li 2007). Civic integration requirements represent a special category of governmental interventions in which immigrants are the target population. Civic integration requirements usually have a twofold nature. First, they are designed to assist newcomers in acquiring the local language, accessing basic services, and entering the labor market; that is, 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, 604). Menjívar and Lakhani (2016) found that immigrants gradually adopt new behaviors in response to governmental interventions and that the existence of host country citizenship requirements motivates immigrants to adopt new behaviors and lifestyles in both the short and the long term. According to Menjívar and Lakhani (2016), immigrants may adopt these behaviors in part out of a fear of being deported, and in part because they are seeking to fit into the legal categories through which they can gain admission to the United States.

Across countries, many types of civic integration policies have been implemented (Helbling 2013). In the European context, Goodman (2010) systematically examined three relevant policy field outputs (immigration entry, integration, and citizenship) with a special emphasis on language requirements (Helbling 2013). Language requirements are often included in civic integration policies, as it is assumed that knowing the country's language means that an applicant is familiar with the country's culture and is therefore sufficiently integrated into the country (Goodman 2010).

Goodman's (2010) classification was done by clustering the EU-15 countries based on their citizenship access and membership content policies. The notion of citizenship access comes from the Citizenship Policy Index (CPI), which evaluates the 2008 citizenship policies of the EU-15 countries (Goodman 2010; Howard 2010). Howard (2010) derived this index by developing theoretical arguments and analyzing cross-national empirical findings. The content of the index has been validated in Helbling (2013). According to Helbling (2013), CPI measures what it intends to measure: namely, the variance of outputs of citizenship policies. Moreover, the CPI is highly correlated with other indexes that cover similar policy components.

The notion of membership content is based on the Civic Integration Index (CIVIX), which analyzes the requirements for country knowledge, language, and values (Goodman 2010). Citizenship requirements are the rules that determine the extension of legal status and rights depending on state membership (entrance, settlement, or citizenship), while integration requirements are related to the degree to which newcomers have become

integrated into the host society (based on factors such as language acquisition and commitment to values) (Goodman 2010; Howard 2010).

Based on the CPI and CIVIX indexes, Goodman (2010) clustered the EU-15 countries into four groups: (1) prohibitive, (2) conditional, (3) enabling, and (4) insular. In the following paragraphs, we use Goodman's (2010) typology to further elaborate on the civic integration and citizenship measures adopted by the EU-15 countries.

The prohibitive group. The *prohibitive* group is made up of Austria, Denmark, and Germany. The countries in this group have relatively strict citizenship requirements (e.g., no dual nationality and a long period of residence in the country before citizenship can be acquired) (Howard 2010) and 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 in Austria and Denmark, anti-immigrant attitudes are relatively strong, and there is a lack of economic pressure to liberalize the citizenship requirements (Howard 2010).

The language acquisition policies of the countries in the prohibitive group have been characterized by a lack of tolerance of different cultures, which are seen as a threat to the language and the culture of the host society (Beauzamy and Féron 2012; Brochmann and Hagelund 2011; Schierup et al. 2006). Austria did not start offering language training to immigrants until 2002. Prior to that time, immigrants were expected to learn the language on their own (Höhne 2013). In Germany, the government began to finance language courses starting in the mid-1970s, which was before the country had even established integration policies (Höhne 2013). Several studies have highlighted the lack of flexibility of Danish immigration policies. In Denmark, migrants who lack a perfect command of Danish suffer from social and labor market discrimination (Beauzamy and Féron 2012; Lønsmann 2020). Austria, Germany, and Denmark are among the European countries that required a high level of language acquisition (B1 level based on the Common European Framework of Reference for Languages (CEFR)) as a condition for permanent residence and citizenship before 2012 (Höhne 2013).

The conditional group. The *conditional* group consists of France, the United Kingdom, and the Netherlands. The countries in this group combine liberal citizenship criteria with arduous integration requirements. In these countries, 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). As these countries are "traditional" immigration countries with a colonial past (Brett 2002), they have—relatively early by European standards—tried to

incorporate the immigrant population into the host society by promoting an atmosphere of tolerance and cultural diversity (Algan, Landais et al. 2012; Manning and Georgiadis 2012). While France and the United Kingdom are considered historically liberal countries, the Netherlands liberalized its citizenship policies between 1980 and 2008 (Howard 2010).

In France, the United Kingdom, and the Netherlands, tolerance of cultural diversity is embedded in law, and citizenship for newcomers is essential to the national identity (Castles, De Haas, and Miller 2013). In the past, the governments of these countries believed that this openness to diversity would lead immigrants to feel that they were part of the wider community. Over time, however, these governments became concerned that they were failing to create common core values; that is, to integrate immigrants into the wider society (Beauzamy and Féron 2012; Manning and Georgiadis 2012). Therefore, before 2012, immigrants to these countries were required to pass a basic language (A1/A2 level based on the CEFR), culture, and history test to become a citizen, or even to gain admission to the country (Höhne 2013; Manning and Georgiadis 2012).

The enabling group. The *enabling* group is made up of Portugal, Finland, Ireland, Belgium, and Sweden. For the countries in this group, citizenship serves as a mechanism for establishing equal status and rights. Hence, it is assumed that citizenship enables integration instead of rewarding it, which is the opposite approach to that 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). These countries were able to liberalize their citizenship requirements in part because the levels of support for far-right parties in the population were low, and in part because of other factors, including demographic change and the rise of international norms (Howard 2010).

Of these countries, only Portugal and Finland required language certification as a condition for citizenship before 2012 (Goodman 2012). Ireland, Belgium, and Sweden required neither national language nor country knowledge as a condition for citizenship or permanent residence (Goodman 2012; Höhne 2013). Sweden is a particular case, as it was among the first countries to implement language courses for immigrants. The Swedish government started financing these courses as early as 1965 (Höhne 2013). However, it was not until 2009 that Swedish became the national language of the country (Bolton and Meierkord 2013).

The insular group. The *insular* group consists of Greece, Spain, Luxembourg, and Italy. In general, these countries have a restrictive approach to granting citizenship to immigrants (Castles, De Haas, and Miller 2013), but often grant citizenship to descendants born abroad (Goodman 2010). This approach may be attributable to the electoral support for far-right parties

and the anti-immigrant attitudes held by the populations of these countries (Howard 2010).

The countries in this group have complex language landscapes, with linguistically independent languages being spoken in different regions of the countries or being used for different official purposes[1] (Bruzos, Erdocia, and Khan 2018; Love 2015; Sharma 2018; Skourmalla and Sounoglou 2021). In Italy and Luxembourg, policy mechanisms intended to standardize and regulate official language usage were introduced at the beginning of the 21st century. However, these policies created conflict with the communities that spoke different languages (Love 2015; Sharma 2018) and made linguistic integration more difficult for immigrants (Odero, Karathanasi, and Baumann 2016). In the case of Greece, language policies were introduced in the 1970s to homogenize the linguistic and cultural landscape of the country (Skourmalla and Sounoglou 2021). Before 2012, Greece required migrants to be proficient in Greek to acquire long-term residency (Tsoukalas et al. 2010). In Spain, there were no regulations regarding official language usage or language acquisition by immigrants before 2015 (Bruzos, Erdocia, and Khan 2018).

Citizenship policy and civic integration indexes

As shown in the previous section, the countries that make up each of the groups are quite heterogeneous. Therefore, we use the raw CPI and CIVIX indexes, as they allow us to account for more variance in the analyses, and to capture differences that are otherwise 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 macrolevel factors associated with immigrants' language acquisition (van Tubergen and Kalmijn 2005): the political climate surrounding immigration and language integration policies.

The political climate and language acquisition by migrants are related through anti-immigrant attitudes and left-wing majority governments. This is because, in a country where right-wing parties are in the majority, it is more likely that the people in that country hold strong anti-immigrant sentiments. This implies that immigrants to that country will have less exposure to the host language (van Tubergen and Kalmijn 2005). By contrast, in a country where left-wing parties are in the majority, the society and the political climate tend to be more tolerant toward immigrants, and the policies tend to favor linguistic pluralism; that is, there is likely to be more tolerance of other languages (van Tubergen and Kalmijn 2005). In view of both arguments, the election of left-wing parties could (unintentionally) reduce immigrants' exposure to the second language and the incentives of acquiring that language. Hence, immigrants in countries with a stronger presence

of left-wing parties in the government might have a lesser command of the destination language (van Tubergen and Kalmijn 2005).

Conservative citizenship policies are associated with strong anti-immigrant attitudes because countries whose citizens have relatively low levels of anti-immigrant sentiments are more likely to liberalize their citizenship policies. By contrast, countries with high levels of xenophobia tend to continue their restrictive citizenship policies (Howard 2010). A latent variable underlying these associations may be education (Dennison and Dražanová 2018). This is because the more educated a population is, the more proimmigration and the more democratic it tends to be (Dennison and Dražanová, 2018).

The first set of hypotheses concerns the effects that civic integration requirements and citizenship policies have on language acquisition. The first hypothesis consists of two competing alternatives. On the one hand, it is possible that the more conservative a country's citizenship policies are, the more time immigrants to the country will need to acquire the language (H1.1a). This is because, in countries with conservative citizenship policies, anti-immigrant attitudes tend to be strong, which implies that immigrants will have fewer chances to use the host language. On the other hand, it is also possible that the more liberal a country's citizenship policies are, the more time immigrants to the country will need to acquire the language (H1.1b). This is because, in countries with liberal policies that favor linguistic pluralism, immigrants may have fewer incentives to learn or to use the host country's language.

The second hypothesis holds that the more integration requirements a country has, the quicker immigrants to the country learn the language of the host country, as there are more incentives for them to learn it (H1.2). Finally, our third hypothesis of this set holds that there is an interaction between civic integration requirements and citizenship policies: that is, the more liberal a country's citizenship policies are and the more civic integration requirements the country has, the faster immigrants to the country learn the language (H1.3). This is because the imposition of more civic integration requirements provides immigrants with more incentives to learn the language, while more liberalization implies more tolerance of migrants, which should mean that members of the host population are more open to interacting with migrants.

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

When analyzing the effects that different civic integration policies across Europe have on language acquisition among immigrants, several difficulties can arise, mainly due to issues of data availability and data quality requirements. First, as highlighted by Beauchemin (2014), comparable databases

that cover different countries and that contain information on multiple immigrant groups are lacking. Second, to study language acquisition processes, researchers need longitudinal information that captures the changes experienced by immigrants either after they have started living in the new country (Font and Méndez 2013) or after they have started learning the new language (Meunier 2015). Finally, while language adoption and proficiency are normally measured through self-assessments, research has shown that these self-estimates only partially reflect actual language 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 and that is stored in the Internet Archive (Scott 2012; Internet Archive 1996).

Twitter data represent a novel and suitable source of information for studying the language acquisition of immigrants to EU-15 countries in a nonintrusive way (i.e., researchers have access to users' digital traces, which are generated from users' digital lives) (Lazer and Radford 2017). Twitter is a microblogging social network on which users can post 140-character[2] messages called "tweets." Users can also follow other users to see their tweets displayed in their feeds, even if the other users do not follow them in return. Twitter does not set known limits on the number of followers users can have (McFedries 2007; Krishnamurthy, Gill, and Arlitt 2008). Access to the Twitter data was stable (i.e., researchers had access to the same interface and its outputs) from 2012 to 2020 via its Application Programing Interface[3] (Zimmer and Proferes 2014). Before July 2020, this access extended only to prospective tweets, and not to tweets that were sent more than seven days in the past.[4,5] However, several 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 and Lin 2017). As we discuss in more detail below, we use data collected by the Internet Archive (Internet Archive 1996).

These features make it possible to use Twitter data to analyze the effects that different civic integration policies across Europe have on language acquisition among immigrants. First, Twitter data allow for the creation of comparable databases that cover different countries and that contain information on multiple immigrant groups (Lamanna et al. 2018). This is because Twitter is used internationally. Thus, conversations in different languages can take place simultaneously on the platform, which has a worldwide reach (Mocanu et al. 2013). Second, Twitter provides longitudinal information that captures the changes experienced by immigrants after they have started living in a new country. This is because tweets can be geo-located. As we explain in the data section, geo-location makes it possible to infer each user's place of residence, and, if the user's geo-location changes, potential migration events (Armstrong et al. 2021). Furthermore, as different independent organizations offer access to historical archives of

Twitter data (Scott 2012; Sequiera and Lin 2017), longitudinal databases can be built. Finally, Twitter data are created in a passive manner; that is, 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 nonintrusive manner (Lazer and Radford 2017; Mejova, Weber, and Macy 2015). Therefore, the study of language acquisition using Twitter data does not rely on users' self-assessments.

Language and social media usage

While data from digital sources such as Twitter offers 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 use of social network sites became widespread, migrants' language adoption at the macrolevel was inverse to (see Chiswick and Miller 2001; Esser 2006) (1) the number of immigrants from the same origin country living in the host country, (2) the linguistic distance between the mother tongue and the official languages of the destination country, and (3) geographic distance between the origin country and the destination country. Some scholars have argued that today, these variables may no longer be associated with language adoption. 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 and Viggiano 2020). In this paper, we consider the possibility that the use of social media may affect the language acquisition process. We do so 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 the origin and the destination country, the linguistic distance between the origin and the destination language, and the geographical distance between the origin and the destination country are associated with the length of time until an immigrant starts tweeting in the language of the destination country. At the macrolevel, it is also important to consider that English is the most used second language in Europe (Bolton and Meierkord 2013; Cromdal 2013); therefore, we also control for the percentage of the host population who speak at least one foreign language.

Our second set of hypotheses is a direct consequence of the aforementioned associations. First, we expect that the more Twitter users from the origin country there are on the platform, the slower the pace of language acquisition is (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 advent of social network sites. On the other hand, as Twitter does not have borders, migrants can keep

communicating with people from their origin country despite living in another country. However, Twitter users may have more incentives to tweet in a given language when there is a bigger audience with whom they can interact using that language, which is in line with the findings of traditional research (Chiswick and Miller 2001; Esser 2006). Second, the greater the linguistic distance between the origin country and the destination country is, the slower the pace of language acquisition is expected to be (H2.2). This is because when the host country language is more difficult for an immigrant to learn because it differs greatly from their mother tongue, it will usually take longer for the immigrant to use it. Finally, the greater the geographic distance between the origin country and destination country is, the faster the pace of language acquisition is expected to be (H2.3).

Data

To process the data, we used the programming language Python version 3.7 (Python Software Foundation 2020).

Twitter and the Internet Archive

Twitter provides access to the free current 1% sample of all public tweets through their public streaming Application Programing Interface (API) (Kumar, Morstatter, and Liu 2015; Pfeffer, Mayer, and Morstatter 2018). The data are accessed by either filtering or sampling the tweets (Pfeffer, Mayer, and Morstatter 2018). If it is accessed by filtering, then the user can specify the parameters: for example, tweet keywords, user IDs, and geographical boundaries. If it is accessed by sampling, then the API does not provide parameters to filter the tweets. Therefore, every user who retrieves data by sampling would receive the same set of tweets (Joseph, Landwehr, and Carley 2014). The sample is created by assigning tweets a millisecond stamp the moment the tweet arrives at Twitter's servers. Any tweet that arrives between milliseconds 657–666 will be available to retrieve (Kergl, Roedler, and Seeber 2014).

Based on those samples, researchers have been able to study the Twitter population. The results indicate that between 2010 and 2012, the European countries with the highest Twitter penetration (average number of Twitter users over population size) were, in decreasing order, the Netherlands, the United Kingdom, Ireland, Sweden, Spain, Belgium, Italy, France, and Germany (Mocanu et al. 2013, fig. 2). In a comparison of Twitter users' data with representative samples of the UK population, Leak et al. (2018) showed that Twitter users are overrepresented in the 10–39 age group and are underrepresented in the 40+ age group, and that female Twitter users are more prevalent in the 10–19 age group, while male Twitter users are more prevalent in the 20+ age group (Leak et al. 2018). Finally, the

authors found that Asian, Black, and mixed-ethnicity groups are underrepresented on the platform, while whites make up the majority of Twitter users (around 90%). The final percentages are similar to those for the usual resident population of the United Kingdom (Leak et al. 2018).

The streaming API has two main limitations. First, it does not return the demographic characteristics of the users, such as age, gender, and level of education. To access this information, researchers have used pattern recognition software to extract users' demographic characteristics depending on their profile picture, username, and tweets (Leak et al. 2018; Mejova, Weber, and Macy 2015; Yin, Chi, and Van Hook 2018). Second, before July 2020,[6] the streaming API did not allow users to retrieve tweets older than seven days. To retrieve older tweets, researchers relied on historical samples gathered by specific organizations, such as the Internet Archive (Internet Archive 1996).

The Internet Archive is the biggest and the oldest archive of the web, and it has been operating continuously since 2000 (Thelwall and Vaughan 2004). It aims to perpetually store collections of digital information, such as Media Collections (books, audio, and images) and the Wayback Machine (more than 500 TB of web pages) (Jaffe and Kirkpatrick 2009; Thelwall and Vaughan 2004). The Internet Archive also hosts a repository that contains a 1% real-time sample of tweets (Scott 2012) collected every hour from 2011 to 2018 through the Twitter sample streaming API.

In this work, we use the Twitter data stored in the Internet Archive that correspond to the period between January 2012 and December 2016. According to Sequiera and Lin (2017), the Twitter databases stored in the Internet Archive are a good replacement for 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 compared with the tweets retrieved using the Twitter streaming API. A study that examined whether Twitter data from the Internet Archive is suitable for making migration estimates concluded that the Twitter data stored in the Internet Archive 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. 2021).

Processing the data

In this work, we use the Twitter data stored in the Internet Archive that corresponds to the period between January 2012 and December 2016. The Twitter data stored in the Internet Archive is a 1% real-time sample of tweets (Scott 2012) collected every hour through the Twitter sample streaming API. The Twitter streaming API returns three different variables from which each user's geo-location can be inferred[7]: geo, place, and location. The variable "geo" consists of the coordinates from which the tweet was

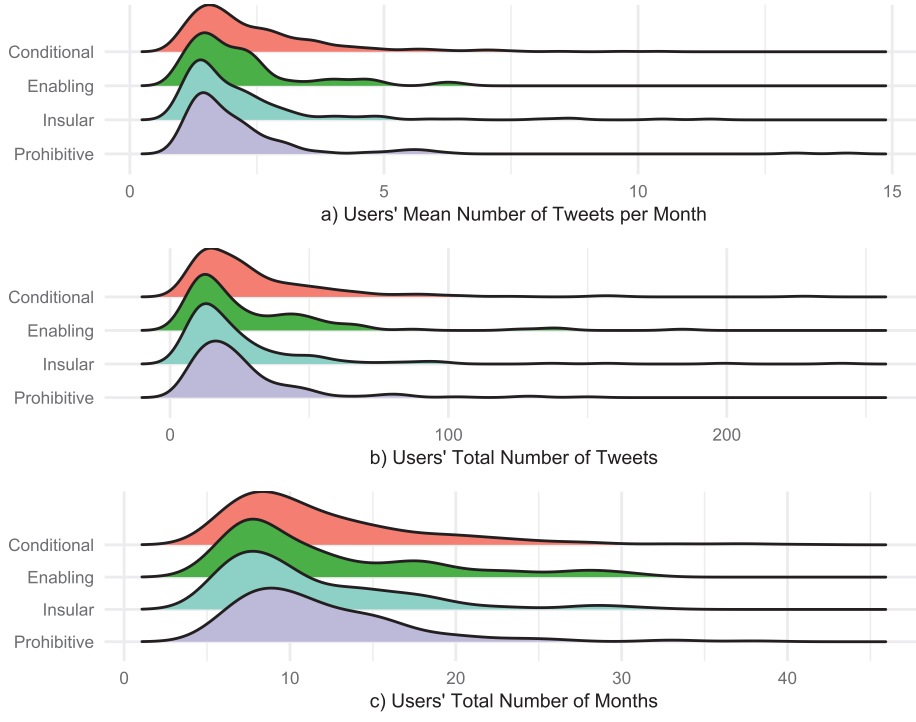
sent. The variable "place" contains the country, the country code, and the bounding box of coordinates—that is, four coordinates—from which the tweet was sent. The variable "location" contains either a user-self-written description or the geo-location of the place where the user is currently living. In our work, we only use the first two: "geo" and "place." If the tweet contains either the "geo" or the "place" information, 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 the country code. Of the 2.64 terabytes of Tape Archive File (TAR) tweets processed, 4% contained geo-location information, which is similar to the percentage Morstatter et al. (2013) reported.

To classify users as migrants, we follow four steps. First, we look for all of the tweets in the filtered data coming from the same user and keep only those users who have tweeted at least five times in a given year. We use this lower bound primarily because we need to capture the moment when a user moves and the moment when they start tweeting in another language. Users who tweet often produce more fine-grained data, which is easier to analyze. A secondary benefit of this lower bound is that it is less computationally demanding, as we need to construct a dictionary to store all of 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 is a new collection of datasets containing the paths to the tweets by the user. Second, from the sample produced in the first step, we select all the users who tweeted from more than one country.

Third, we categorize a user as a migrant if the user tweeted for at least three months from one country and for at least the last three months from a second country. The user's location by month is the location from which the user tweeted the most during that month. The month is considered if the user tweeted at least once. In our sample, many users did not tweet consecutively. This could be because of how the Twitter API samples the tweets or because the users did not tweet during those months. A user is identified as a migrant if, for example, the person starts tweeting in Mexico and then moves to Germany, and the geo-location of their tweets, therefore, changes 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. (2021) that the data can be used to analyze long-term migration if a temporal window (buffer) greater than 12 weeks is implemented to classify users as migrants.

Finally, from the sample, we keep the migrants who moved to one of the EU-15 countries and for whom the official languages of their origin country and their destination country are different. This gives us a final database of around 1,210 unique users and around 35,448 tweets. In order

FIGURE 1 Distributions of users' tweets by cluster.

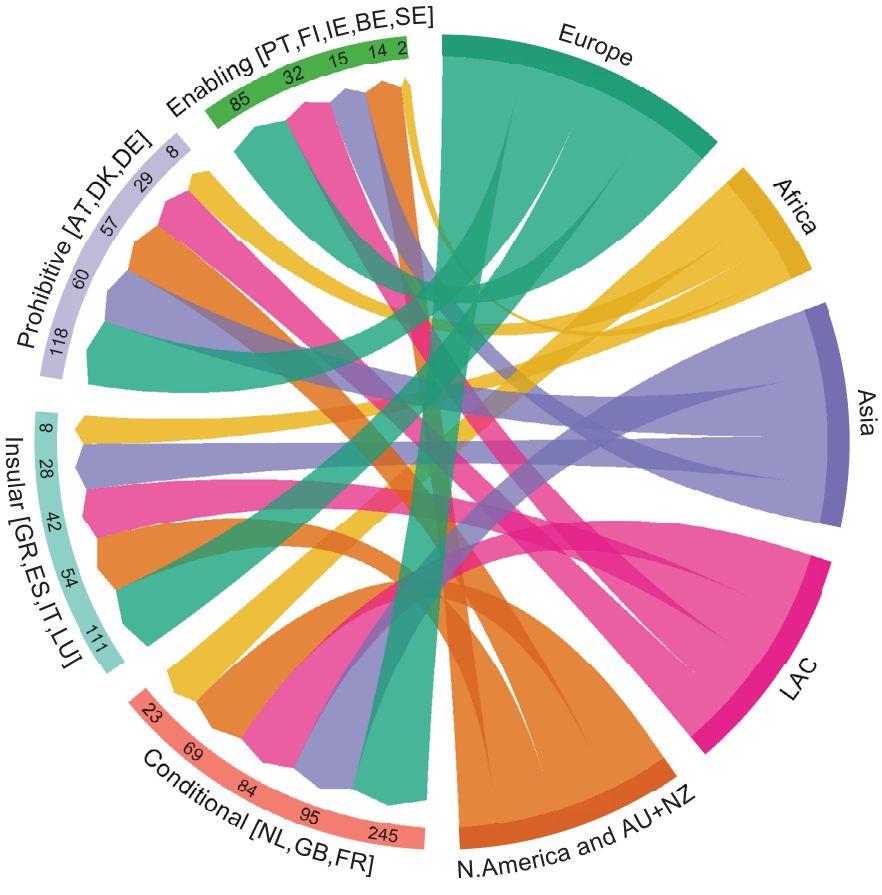


NOTE: The y-axis corresponds to the name of the cluster.

to classify the language used in their tweets, we use the package pycld2 (Al-Rfou 2019) together with our own algorithm (online Appendix A). From this processed data, we aggregate the information by month and frequency (Figure A1 of online Appendix A).

To the final database, we add for each user the time (in months) between their arrival in the country of destination and the time when the user started tweeting primarily in the official language of destination for one month; in other words, the point when the user's tweets started reflecting the geo-location of the destination country, and the point when more than 50% of the tweets the user posted each month were in the host country language. Figure 1 shows the distribution of the users' tweets. Figure 1a shows that the users' mean number of tweets per month was around 2.5 for each of the clusters. Figure 1b shows that the users' average total number of tweets by cluster was around 30. Finally, Figure 1c shows that the users' average total number of tweeted months by cluster was around 12.5. Overall, the median number of tweets per user was 20, and the median number of months per user was 10. From these results, we can conclude that the distributions of the tweets by cluster are similar and that the sample of users' tweets returned by the Twitter API is not biased to certain regions.

FIGURE 2 Migration flows from the regions of origin to EU-15 civic integration groups.



NOTE: 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.

The users' countries of origin are quite diverse; in total, our sample contains users with 81 countries of origin. Given this considerable diversity, we categorize the users' origins into five regions in order to visualize them more effectively: Europe; Africa; Asia; Latin America and the Caribbean (LAC); and North America, Australia, and New Zealand (N. America and AU+NZ). The total number of individuals who migrated from these regions is 564, 46, 205, 195, and 200, respectively. Figure 2 shows the migration flows from these regions of origin to the countries in the civic integration clusters described above: prohibitive, enabling, insular, and conditional. While we could not identify any users who could be classified as immigrants to Luxembourg, we kept the code in Figure 2 as part of the insular group. The total number of immigrants each of these clusters received is 276, 152,

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

NOTE: Account status corresponds to the information the Twitter API V2.2 returned on August 10, 2021.

243, and 539, respectively. Table T1 of online Appendix B shows the number of immigrants by country of destination and the percentage who started to tweet in an official language of the destination country.

We also checked the distribution of the users by gender and account status by civic integration group (Table 1). Users' gender was inferred from their user names using the databases Social Security Administration (2020) and Demografix ApS (2021). For this purpose, we have built a dictionary with the weighted probability of a name being male or female according to these databases. From this point onward, we expect that the distributions will be similar in each of the groups; if they are not, we can assume that the sample of users that the Twitter API returns is biased to certain regions. Table 1 shows that the percentages of female, male, and unknown users are equally distributed across the groups, as is the current users' account status. The percentages of female and male users are similar to those reported in previous findings (Zagheni et al. 2014). The percentages of deleted and suspended accounts are also in line with previous estimates (Armstrong et al. 2021).

Before performing the analysis, we validated that these users are (were) migrants by performing a qualitative analysis of a 10% sample of users. We analyzed their tweets and their tweets metadata, as suggested by Armstrong et al. (2021). The qualitative analysis shows that some of the users tweeted as a student in a foreign country, while others became a resident in the new country. For the students, their status is deduced from their tweets indicating that they were sharing their experiences as a newcomer in the country. For the residents, their status is deduced from their tweets, and, for some of them, from their current Twitter status profile in which they share that they are from country A and are currently living in country B. For a small proportion of the users, we could not infer their motivations for moving; nonetheless, we kept them for the analysis. No individuals were detected who could with certainty be classified as nonmigrants.

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: more than 50% of the tweets the user posted each month were in the host

country language. For this analysis, we use the programming language R (R Core Team 2020) and the *survival* package (Therneau and Grambsch 2000).

To choose the best parametric model to fit our data, we test the linearity of the Kaplan–Meier survival values by plotting $\ln(-\ln(\hat{S}(t)))$ vs. $\ln(t)$ (Kleinbaum and Klein 2012, 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 (Figure A2 of online Appendix C). The Weibull parametrization we follow is given by Kleinbaum and Klein (2012) (Equation 1).

$$S(t) = \exp(-\lambda t^{p}), \tag{1}$$

To study the factors that enhance language acquisition, we model the accelerated failure time (AFT) ratios of ***T***: *time until a user mostly tweets in the language of destination for one month*. We decided to use this model because the results are interpreted as the median survival time until the language is acquired, which we consider to be more directly interpretable than proportional hazards.

We model the time until a user mostly tweets in the language of destination for one month as a function of the following seven variables. The first variable is the CPI developed by Howard (2010), in which the higher the value is the more liberal the citizenship requirements of the destination country are. Howard (2010) built this index by aggregating the following factors: whether or not a country grants *ius soli*; the minimum length of residency required for naturalization; and whether or not naturalized immigrants are allowed to hold dual citizenship. In addition, he penalized countries that have added civic integration requirements (such as language and civic tests). The second variable is the CIVIX developed by Goodman (2010), in which the higher the value is the stronger the integration requirements of the country of destination are. Goodman (2010) built this index by giving points to four different categories of requirements:

> [whether] third-country nationals [are] accountable, specifically family unification; whether civic conditions are required for entry, settlement or citizenship; the number of requirements across the civic targets of country knowledge, language and values, including integration courses, tests, contracts, oath ceremonies and interviews; and, finally, the severity of requirements along the path to citizenship (for example, a "high" level of language proficiency or cost). (Goodman 2010, 759)

These two variables are continuous and range from zero to six. The third variable is an interaction term between both the CPI and the CIVIX. This variable is continuous and ranges from zero to 36. We do not transform any of these variables in order to facilitate interpretation, given the interaction term.

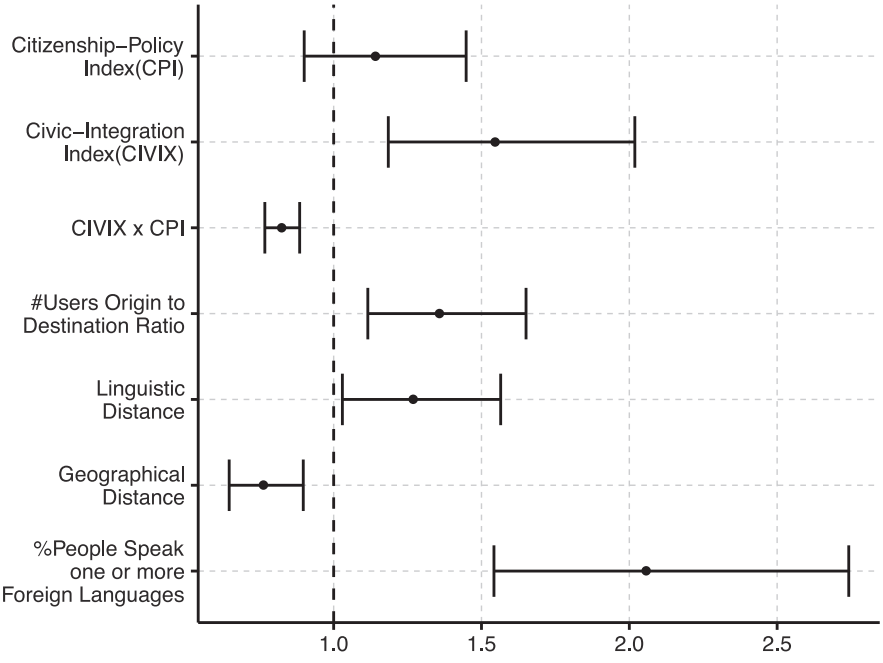
The fourth variable is the logarithm of the ratio of Twitter users in the country of origin to Twitter users in the country of destination. These

latter users are all users who tweeted from the same country during their entire Twitter history. Here, a positive value means there are more Twitter users in the origin country than in the destination country, and a negative value means the opposite. The fifth and sixth variables are linguistic distance and geographic distance. These variables come, respectively, from the "Language" and "Gravity" databases from the Centre d'Études Prospectives et d'Informations Internationales.[8] Linguistic distance is the variable LP2 (Melitz and Toubal 2014). According to Melitz and Toubal (2014), LP2 shows how close two different native languages are based on the similarity of words with identical meanings. It is interpreted as the smaller the value is the closer the languages are in terms of vocabulary and grammar. Geographical distance is the distance in kilometers between the capitals of the origin and the destination countries (Conte, Cotterlaz, and Mayer 2021). The seventh variable is the percentage of the destination population who self-reported knowing at least one foreign language in 2011, except for the United Kingdom, where we interpolated (Eurostat 2021). These variables are continuous and standardized (meaning we subtracted the mean and divided by the standard deviation).

In the model, we do not account for the variables gender and age because there is no evidence that leads us to assume that the age and gender distributions of Twitter users differed between the integration regimes. In the case of gender, this is shown in Table 1, which indicates that all of the regimes have around the same percentages of female, male, and unknown gender users. We tested this argument by running the model (Equation 2 beneath) adjusting and without adjusting for gender (Figure A3 of online Appendix D), and found that the results were similar. We do not know of a satisfactory way to impute age.[9] Furthermore, while migrants who are Twitter users may, on average, be younger and better educated than other migrants, we have no reason to suspect that the overrepresentation of young and highly educated people differs between the integration regimes. We, therefore, assume that the bias caused by this overrepresentation is the same between the integration regimes. If this assumption is correct, the relative comparison (ratio of accelerated failure times) between the regimes will not be affected, and the comparison is valid even when age is not controlled for.

The Weibull AFT function is $t = [-\ln S(t)]^{\frac{1}{p}}\lambda^{-\frac{1}{p}}$, where $\lambda^{-1/p}$is parameterized with regression coefficients (Equation 2) (Kleinbaum and Klein 2012, 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 in which the coefficient is positive leads to an increase in the median (or other quantile) survival time until the language is acquired. If the coefficient is negative, then an increase in the variable would lead to a decrease in the median survival time until the language is acquired.

FIGURE 3 Exponential of AFT coefficients with their corresponding 95% confidence intervals



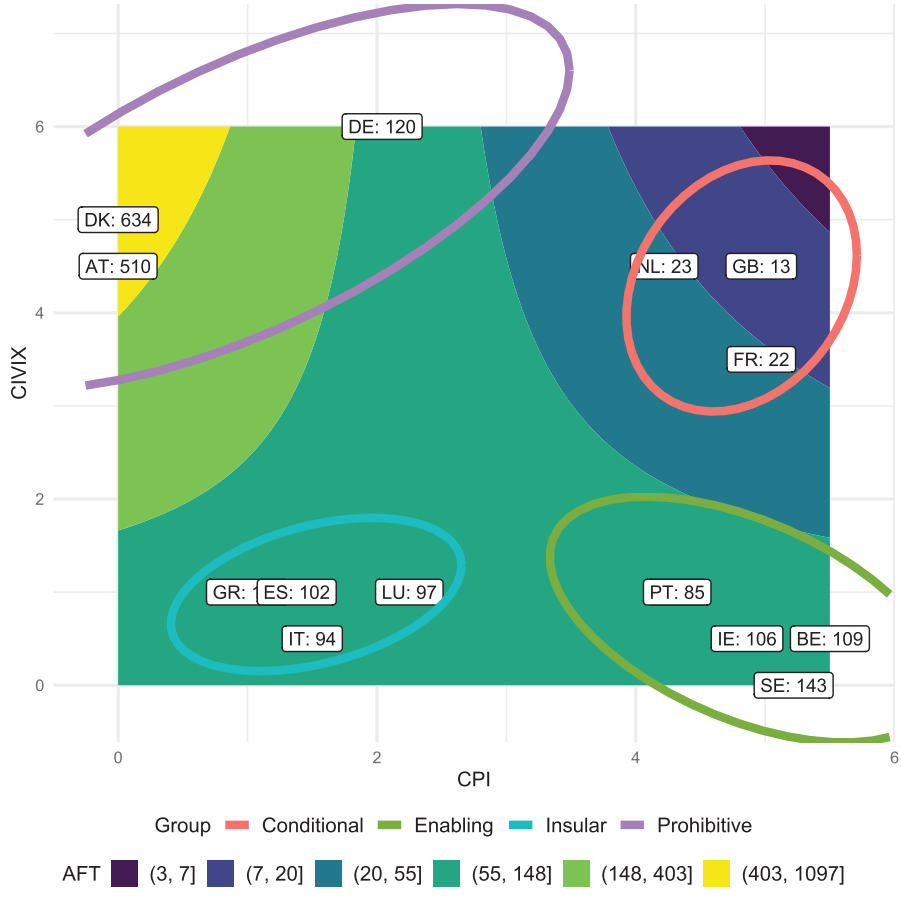
$$\lambda_i^{-\frac{1}{p}} = \exp\left(\alpha_0 + \alpha_1 CPI_i + \alpha_2 CIVIX_i + \alpha_3 CPI_i \times CIVIX_i + \alpha_4 \log\left(ratio\right)_i + \alpha_5 Ling.Dist._i + \alpha_6 Geo.Dist._i + \alpha_7 \%\geq 1 Foreign\ Lang._i\right) \quad (2)$$

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 the origin country by the number of Twitter users from the destination country; *Lin. Dist*. is linguistic distance; *Geo. Dist*. is geographical distance; and *%≥1 Foreign Lang*. is the percentage of the destination country population who speak more than one foreign language.

Results

Figure 3 shows the estimated AFT ratios of the Weibull model with 95% confidence intervals. In the case of CIVIX, we find that the stronger the civic integration requirements are the longer the median survival time until the language is acquired; therefore, H1.2 is not supported. In the case of CPI, it does not appear to play a role in the median survival time until the language is acquired conditional on the other variables in the model. However, the interaction variable shows that the greater the CPI and the CIVIX are the lower the median survival time until the language is acquired. This indicates

FIGURE 4 Contour map of the accelerated failure time (AFT) relative to the civic integration index (CIVIX) and the citizenship policy index (CPI)



that CPI does play a role, but only in conjunction with particular CIVIX levels.

To clarify the interaction effect, we show the predicted survival time until the language is acquired using a contour map relative to the CIVIX and CPI indexes in Figure 4. These predicted values are obtained by multiplying the model coefficients by the different combinations of CIVIX and CPI values while keeping the rest of the variables constant at their means (which are zero because of the standardization). Figure 4 shows that when the CIVIX index is below 1.75, the CPI index is not associated with the median survival time until immigrants have acquired the language. This can be seen in the median survival values of the insular and enabling groups (excluding Sweden), which range from 85 to 109 months regardless of the CPI values. Once the CIVIX values are higher than one, the CPI index becomes associated with the median survival time until the immigrants' language ac-

quisition, the more liberalized the country is the lower the median length of time it takes for immigrants to acquire the language (conditional group), and the less liberalized the country is the higher the median length of time it takes for immigrants to acquire the language (prohibitive group). In this analysis, Sweden seems to follow a different pattern from the rest of the countries. This could be because, compared to the rest of the countries in the insular and enabling groups, a higher percentage of Sweden's population speak at least one foreign language.

Returning to Figure 3, in line with our secondary hypotheses, the median survival time until the language is adopted increases if the number of Twitter users in the country of origin is larger than in the country of destination (H2.1). This might be because Twitter users may have more incentives to tweet in a language when there is a larger audience with whom they can interact using that language, as previous research has shown (Chiswick and Miller 2001; Esser 2006). Linguistic distance also has a positive association. This means that the larger the distance between the origin and the destination language is, the longer it takes to acquire the destination language (H2.2). A potential explanation for this finding is that when a language is more difficult for immigrants to learn because it differs considerably from their mother tongue, it takes longer for them to use it (Chiswick and Miller 2001; Esser 2006). For geographic distance, we find that the larger the distance between the origin and the destination country is the shorter the median survival time is until the language is acquired (H2.3). This was expected, as geographic distance is associated with greater incentives to invest in language skills (Chiswick and Miller 2001; Esser 2006). Therefore, the classic variables used to explain immigrants' language acquisition have the same associations with language acquisition as before the advent of social network sites.

Finally, for the control variable, an increase of a percentage point in the share of the destination country population who speak a foreign language doubles the mean number of months until the language is acquired. This may be because migrants may lack incentives to learn the destination country language when the destination population can communicate with them in a language migrants might know (such as English) (Bolton and Meierkord 2013; Cromdal 2013).

Discussion and conclusions

In this work, we studied immigrants' language acquisition through a longitudinal analysis of the languages they used in their tweets. To do so, we drew on Goodman's (2010) and Howard's (2010) work to formulate how citizenship and civic integration policies may have affected immigrants' language acquisition. Conceptually, we relied on the governmentality framework that theorizes on the effects that governmental interventions have

on individuals. We used survival models to analyze the pace of immigrants' language acquisition depending on (1) citizenship and civic integration policies and (2) the relative sizes of migrant groups in the destination country and the linguistic and geographical distances between the countries of origin and destination. Specifically, we analyzed the length of time it took until a user was mostly tweeting in the language of the destination country for one month. We used starting to tweet in the language of the country of destination as a proxy for language acquisition.

Our findings point to an interaction effect between civic integration requirements and citizenship policies, whereby immigrants in countries with loose or no civic integration requirements had similar median times until language acquisition regardless of how liberalized the citizenship policies were. This was the case for countries that had heterogeneous citizenship policies but few civic integration requirements. However, among these countries, Sweden appeared to be a particular case. In Sweden, the median time until language acquisition was similar to that of countries with strict civic integration and citizenship requirements. Among the potential explanations for this finding are that a high percentage of the Swedish population speaks at least one foreign language (Bolton and Meierkord 2013) and that Sweden has high levels of multiculturalism, which could discourage immigrants from learning Swedish (van Tubergen and Kalmijn 2005).

We found that in the countries with strict civic integration and citizenship requirements (Denmark, Austria, and Germany), the time it took for immigrants to acquire the language was longer than in the other countries. While this may be a consequence of the anti-immigrant attitudes of the majority population, it has also been suggested that strict requirements placed on immigrant groups may be the result of right-wing parties trying to constrain immigrants from having access to rights equal to those of natives (M. B. Jørgensen 2009; Beauzamy and Féron 2012; Bolton and Meierkord 2013; Lønsmann 2020). Research has shown that these types of negative interactions between authorities and migrants can lead to language balkanization and to immigrants rejecting learning the language of the destination country (J. N. Jørgensen 2003).

For those countries with onerous civic integration requirements, we found that the more liberalized immigrants' access to citizenship was the faster they acquired the host country language. This was shown to be the case for France, the Netherlands, and the United Kingdom, which have strict civic integration requirements but are also considered historically liberal countries (Howard 2010). However, this result might also be explained by the early integration requirements that immigrants had to fulfill, such as learning the language before moving to the country (Goodman 2010), which would then be more related to a selection effect than to migration policies.

Our results also showed that the evidence of immigrants' language acquisition on Twitter was associated with the same classic macrolevel explicative variables employed before the advent of social network sites. This result is relevant in two ways. On the one hand, it supports the notion that the Twitter data actually captures migration. On the other hand, it helps to shed light on the question of whether the transnational property of social network sites has affected the association between immigrants' language acquisition and classic macroexplicative variables. For this sample of Twitter users, the results showed that this has not been the case. However, this may change in the future, as the use of information and communication technologies is becoming more and more pervasive across the globe.

Limitations

This work has several limitations that we would like to acknowledge. Twitter data are not representative of the general population. Twitter users tend to be young adult men who are highly educated and highly Internet skilled (Hargittai 2020). For this study, this point is especially important, as the results are not representative of all the migrants who moved to the European countries analyzed. As such, the number of possible migrants found was quite small compared to all the terabytes of information analyzed. Furthermore, in the data, certain vulnerable migrant populations, such as illegal migrants, may not be represented, while other populations, such as international students, may be overrepresented. Therefore, caution is advised when interpreting the results.

While our sample was a random selection of tweets, the identification of the users' location depended on longitudinal Voluntarily Geographic Information (Haklay 2016), that is, the analysis was limited to highly active users that contribute to place-based systems and who are considered content producers. Hence, the final sample is subject to selection. However, as far as we know, there is insufficient evidence to ascertain whether content producers, conditional on the other variables adjusted for (the size of the immigrant group in the destination countries and the linguistic and geographical distance between origin and destination countries) differ from other Twitter users in their time to destination language acquisition. If content producers do differ in this way, then this would result in biased estimates of time-to-destination language acquisition. However, it would not necessarily result in biased ratios of time to language acquisition between the country blocs, the variable of interest in this study, unless different "types" of content producers are also more likely to migrate to different country blocs. This could be a subject for future research.

Despite these clear data limitations, we showed that Twitter data can be used to study immigrants' language acquisition and that the data shows

patterns similar to those found in analyses done with representative samples collected before the advent of social network sites (Chiswick and Miller 2001; Esser 2006). However, it is important to continue studying and developing statistical techniques, training databases, and machine learning algorithms 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. For the analyses, we relied on public data from the Internet Archive, and we studied only the language of the users' tweets. For research purposes only, we also read the public description of the profiles of a small sample of the users.

Acknowledgments

We would like to thank Lee Fiorio (University of Washington), Clara Mulder (University of Groningen), and Emanuele del Fava (Max Planck Institute for Demographic Research) for their feedback.

Conflict of interest

The authors declare no conflict of interest.

Reproducibility

Given Twitter's terms and conditions, we do not share the final database, as this can lead to the disclosure of user IDs. All of the code needed to replicate this work is available in Gil-Clavel's GitHub repository: https://github.com/SofiaG1l

Data availability statement

The data to reproduce this work are freely available in the Internet Archive: https://archive.org/details/twitterstream

Notes

1 This is the case for Luxembourg, where "Luxembourgish is the national language, French the legislative language, and German is the language of instruction in public schools" (Odero, Karathanasi, and Baumann 2016, 4067).

2 Twitter announced that the character limit would 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 on October 21, 2020.

3 Twitter launched the Twitter API V.2 in August 2020 (Cairns and Shetty 2020). It replaced the version V1.1 launched in September 2012 (Costa 2012).

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

5 In July 2020, Twitter launched the second version of the API. This included the Academic Research Application, which gives access to retrospective tweets based on tailored queries. https://developer.twitter.com/en/blog/product-news/2021/enabling-the-future-of-academic-research-with-the-twitter-api. Accessed on July 4, 2022.

6 https://developer.twitter.com/en/blog/product-news/2021/enabling-the-future-of-academic-research-with-the-twitter-api. Accessed on July 4, 2022.

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

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

9 The methods to approximate age have mostly been developed using training datasets that are biased or not representative (D'Ignazio and Klein 2020). They work by relying on specific texts and images that cannot be used to classify users from all over the world. Furthermore, their accuracy is still low in terms of machine learning classification (Jung et al. 2018). Therefore, their use would lead to the inaccurate classification of users' ages, which would introduce more noise and lead to erroneous conclusions (Höfler 2005).

References

Algan, Yann, Alberto Bisin, Alan Manning, and Thierry Verdier, eds. 2012. *Cultural Integration of Immigrants in Europe*. Studies of Policy Reform. Oxford: Oxford University Press.

Algan, Yann, Camille Landais, and Claudia Senik. 2012. "Cultural Integration in France." In *Cultural Integration of Immigrants in Europe*, edited by Yann Algan, Alberto Bisin, Alan Manning, and Thierry Verdier, 49–69. Studies of Policy Reform. Oxford: Oxford University Press.

Al-Rfou, Rami. 2019. "PYCLD2." Windows. Python. PYCLD2 - Python Bindings to CLD2. https://pypi.org/project/pycld2/.

Armstrong, Caitrin, Ate Poorthuis, Matthew Zook, Derek Ruths, and Thomas Soehl. 2021. "Challenges When Identifying Migration from Geo-Located Twitter Data." *EPJ Data Science* 10: 1. https://doi.org/10.1140/epjds/s13688-020-00254-7

Beauchemin, Cris. 2014. "A Manifesto for Quantitative Multi-Sited Approaches to International Migration." *International Migration Review* 48(4): 921–38. https://doi.org/10.1111/imre.12157

Beauzamy, Brigitte, and Elise Féron. 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, Kingsley, and Christiane Meierkord. 2013. "English in Contemporary Sweden: Perceptions, Policies, and Narrated Practices." *Journal of Sociolinguistics* 17(1): 93–117. https://doi.org/10.1111/josl.12014

Brett, Klopp. 2002. *German Multiculturalism: Immigrant Integration and the Transformation of Citizenship*. Westport, CT: Greenwood Publishing Group.

Brochmann, Grete, and Anniken Hagelund. 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, Alberto, Iker Erdocia, and Kamran Khan. 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, Ian, and Priyanka Shetty. 2020. "Introducing a New and Improved Twitter API." Accessed July 16, 2020. https://developer.twitter.com/en/blog/product-news/2020/introducing-new-twitter-api

Castles, Stephen, Hein De Haas, and Mark J. Miller. 2013. *The Age of Migration: International Population Movements in the Modern World*. 5th ed. London: Palgrave Macmillan UK.

Castro-Gómez, Santiago. 2010. "Siglo xviii: El nacimiento de la biopolítica1 18th Century: The emergence of biopolitics." *Tabula Rasa* 15: 31–45.

Chiswick, Barry R, and Paul W Miller. 2001. "A Model of Destination-Language Acquisition: Application to Male Immigrants in Canada." *Language Acquisition* 38(3): 19.

Conte, Maddalena, Pierre Cotterlaz, and Thierry Mayer. 2021. "The CEPII Gravity Database," 40.

Costa, Jason. 2012. "Current Status: API v1.1." September 2012. https://blog.twitter.com/developer/en_us/a/2012/current-status-api-v1-1

Cromdal, Jakob. 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, Hein, Stephen Castles, and Mark J. Miller. 2020. *The Age of Migration: International Population Movements in the Modern World*. 6th ed. New York City: The Guildford Press.

Demografix ApS. 2021. "Genderize.Io." https://genderize.io/

Dennison, James, and Lenka Dražanová. 2018. "*Public Attitudes on Migration : Rethinking How People Perceive Migration : An Analysis of Existing Opinion Polls in the Euro-Mediterranean Region*." San Domenico di Fiesole: European University Institute. https://cadmus.eui.eu/handle/1814/62348

D'Ignazio, Catherine, and Lauren Klein. 2020. *Data Feminism*. <Strong>Ideas Series. Cambridge, MA: MIT Press. https://data-feminism.mitpress.mit.edu/

Duncan, Howard. 2020. "Trends in International, National and Local Policies on Migrant Entry and Integration." In *The Sage Handbook of International Migration*, 592–607. London: SAGE Publications Ltd. https://doi.org/10.4135/9781526470416

Eckert, Eva. 2018. "Immigration, Language, and Conflicting Ideologies: The Czech in Texas." In *Handbook of the Changing World Language Map*, edited by Stanley D Brunn and Roland Kehrein, 1–17. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-73400-2_31-1

Edele, Aileen, Julian Seuring, Cornelia Kristen, and Petra Stanat. 2015. "Why Bother with Testing? The Validity of Immigrants' Self-Assessed Language Proficiency." *Social Science Research* 52(July): 99–123. https://doi.org/10.1016/j.ssresearch.2014.12.017

Esser, Hartmut. 2006. *Migration, Language and Integration*. AKI Research Review 4. Berlin: WZB.

Eurostat. 2021. "Number of Foreign Languages Known (Self-Reported) by Sex (EDAT_AES_L21)." [Dataset page]. Eurostat Data Browser. 2021. https://ec.europa.eu/eurostat/databrowser/view/EDAT_AES_L21/default/table?lang=enamp;category=sks.sks_ssr.sks_ssaes.edat_aes_l2

Fiorio, Lee, Emilio Zagheni, Guy Abel, Johnathan Hill, Gabriel Pestre, Emmanuel Letouzé, and Jixuan Cai. 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, Joan, and Mónica Méndez. 2013. "12 Surveying Immigrant Populations: Methodological Strategies, Good Practices and Open Questions." In *Surveying Ethnic Minorities and Immigrant Populations : Methodological Challenges and Research Strategies*, edited by Joan Font and Mónica Méndez. IMISCOE Research Series. Amsterdam: Amsterdam University Press. https://doi.org/10.26530/OAPEN_450851

Forrest, James, Phil Benson, and Frank Siciliano. 2018. "Linguistic Shift and Heritage Language Retention in Australia." In *Handbook of the Changing World Language Map*, edited by Stanley D Brunn and Roland Kehrein, 1–18. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-73400-2_37-1

Foucault, Michel. 1991. "Governmentality." In *The Foucault Effect: Studies in Governmentality*, edited by Graham Burchell, Colin Gordon, and Peter Miller, 87–104. Chicago: University of Chicago Press.

Goodman, Sara Wallace. 2010. "Integration Requirements for Integration's Sake? Identifying, Categorising and Comparing Civic Integration Policies." *Journal of Ethnic and Migration Studies* 36(5): 753–772. https://doi.org/10.1080/13691831003764300

Goodman, Sara Wallace. 2012. "Fortifying Citizenship: Policy Strategies for Civic Integration in Western Europe." *World Politics* 64(4): 659–698. https://doi.org/10.1017/S0043887112000184

Haklay, Mordechai Muki. 2016. "Why Is Participation Inequality Important?" In *European Handbook of Crowdsourced Geographic Information*, edited by Cristina Capineri, Muki Haklay, Haosheng Huang, Vyron Antoniou, Juhani Kettunen, Frank Ostermann, and Ross Purves, 35–44. London: Ubiquity Press. http://dx.doi.org/10.5334/bax

Hargittai, Eszter. 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, Bartosz, Izabela Sitko, Euro Beinat, Stanislav Sobolevsky, Pavlos Kazakopoulos, and Carlo Ratti. 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

Helbling, Marc. 2013. "Validating Integration and Citizenship Policy Indices." *Comparative European Politics* 11(5): 555–576. https://doi.org/10.1057/cep.2013.11

Höfler, Michael. 2005. "The Effect of Misclassification on the Estimation of Association: A Review." *International Journal of Methods in Psychiatric Research* 14(2): 92–101. https://doi.org/10.1002/mpr.20

Höhne, Jutta. 2013. "Language Integration of Labour Migrants in Austria, Belgium, France, Germany, the Netherlands and Sweden from a Historical Perspective." *WZB Berlin Social Science Center*, WZB Discussion Paper, No. SP VI 2013–101: 28.

Howard, Marc Morjé. 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

Internet Archive. 1996. "Internet Archive: About IA." 1996. https://archive.org/about/

Jaffe, Elliot, and Scott Kirkpatrick. 2009. "Architecture of the Internet Archive." In *Proceedings of SYSTOR 2009: The Israeli Experimental Systems Conference*, 1–10. Haifa Israel: ACM. https://doi.org/10.1145/1534530.1534545

Jørgensen, J. Normann. 2003. "Bilingualism in the Køge Project." *International Journal of Bilingualism* 7(4): 333–52. https://doi.org/10.1177/13670069030070040101

Jørgensen, Martin Bak. 2009. "*National and Transnational Identities: Turkish Organising Processes and Identity Construction in Denmark, Sweden and Germany*." Denmark: Aalborg Universitet. Spirit Ph.D. Series No. 19.

Joseph, Kenneth, Peter M. Landwehr, and Kathleen M. Carley. 2014. "Two 1%s Don't Make a Whole: Comparing Simultaneous Samples from Twitter's Streaming API." In *Social Computing, Behavioral-Cultural Modeling and Prediction*, edited by William G. Kennedy, Nitin Agarwal, and Shanchieh Jay Yang, 75–83. Lecture Notes in Computer Science, Vol. 8393. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-05579-4_10

Jung, Soon-Gyo, Jisun An, Haewoon Kwak, Joni Salminen, and Bernard J Jansen. 2018. "Assessing the Accuracy of Four Popular Face Recognition Tools for Inferring Gender, Age, and Race." In *Proceedings of the Twelfth International AAAI Conference on Web and Social Media (ICWSM 2018)*, 4. Menlo Park, CA: AAAI Press.

Kergl, Dennis, Robert Roedler, and Sebastian Seeber. 2014. "On the Endogenesis of Twitter's Spritzer and Gardenhose Sample Streams." In *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*, 357–64. Beijing: IEEE. https://doi.org/10.1109/ASONAM.2014.6921610

Kleinbaum, David G., and Mitchel Klein. 2012. *Survival Analysis: A Self-Learning Text, Third Edition*. 3rd ed. Statistics for Biology and Health. New York: Springer-Verlag. https://doi.org/10.1007/978-1-4419-6646-9

Komito, Lee. 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, Balachander, Phillipa Gill, and Martin Arlitt. 2008. "A Few Chirps about Twitter." In *Proceedings of the First Workshop on Online Social Networks - WOSP '08*, 19. Seattle, WA: ACM Press. https://doi.org/10.1145/1397735.1397741

Kumar, Shamanth, Fred Morstatter, and Huan Liu. 2015. "Analyzing Twitter Data." In *Twitter: A Digital Socioscope*, edited by Yelena Mejova, Ingmar Weber, and Michael W. Macy. New York: Cambridge University Press.

Lamanna, Fabio, Maxime Lenormand, María Henar Salas-Olmedo, Gustavo Romanillos, Bruno Gonçalves, and José J. Ramasco. 2018. "Immigrant Community Integration in World Cities." *PLoS ONE* 13(3): e0191612. https://doi.org/10.1371/journal.pone.0191612

Lazer, David, and Jason Radford. 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, Alistair, Guy Lansley, Paul Longley, James Cheshire, and Alex Singleton. 2018. "Geotemporal Twitter Demographics." In *Consumer Data Research*, 152–165. Berkley, CA: UCL Press. http://www.jstor.org/stable/j.ctvqhsn6.14

Li, Tania Murray. 2007. "Governmentality." *Anthropologica* 49(2): 275–281.

Lønsmann, Dorte. 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, Stephanie 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, Alan, and Andreas Georgiadis. 2012. "Cultural Integration in the United Kingdom." In *Cultural Integration of Immigrants in Europe*, edited by Yann Algan, Alberto Bisin, Alan Manning, and Thierry Verdier. 260-284. Studies of Policy Reform. Oxford: Oxford University Press.

Mazzoli, Mattia, Boris Diechtiareff, Antònia Tugores, Willian Wives, Natalia Adler, Pere Colet, and José J. Ramasco. 2020. "Migrant Mobility Flows Characterized with Digital Data." Edited by Jordi Paniagua. *PLoS ONE* 15(3): e0230264. https://doi.org/10.1371/journal.pone.0230264

McFedries, Paul. 2007. "Technically Speaking: All A-Twitter." *IEEE Spectrum* 44(10): 84–84. https://doi.org/10.1109/MSPEC.2007.4337670

Mejova, Yelena, Ingmar Weber, and Michael W. Macy. 2015. *Twitter: A Digital Socioscope*. New York: Cambridge University Press.

Melitz, Jacques, and Farid Toubal. 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, Cecilia, and Sarah M. Lakhani. 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

Meunier, Fanny. 2015. "Developmental Patterns in Learner Corpora." In *The Cambridge Handbook of Learner Corpus Research*, edited by Sylviane Granger, Gaetanelle Gilquin, and Fanny Meunier, 379–400. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.017

Mocanu, Delia, Andrea Baronchelli, Nicola Perra, Bruno Gonçalves, Qian Zhang, and Alessandro Vespignani. 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, Fred, Juergen Pfeffer, Huan Liu, and Kathleen M Carley. 2013. "Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose." In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, 400–408. Menlo Park, CA: Association for the Advancement of Artificial Intelligence Press.

Odero, Angela, Chrysoula Karathanasi, and Michèle Baumann. 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.

Pfeffer, Jürgen, Katja Mayer, and Fred Morstatter. 2018. "Tampering with Twitter's Sample API." *EPJ Data Science* 7(1): 50. https://doi.org/10.1140/epjds/s13688-018-0178-0

Python Software Foundation. 2020. "Python Language Reference." 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, Carl-Ulrik, Peo Hansen, and Stephen Castles. 2006. *Migration, Citizenship, and the European Welfare State: A European Dilemma*. Oxford; OUP.

Scott, Jason. 2012. "Archive Team: The Twitter Stream Grab." *Internet Archive*. 2012. https://archive.org/details/twitterstream

Sequiera, Royal, and Jimmy Lin. 2017. "Finally, a Downloadable Test Collection of Tweets." In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1225–28. Shinjuku Tokyo Japan: ACM. https://doi.org/10.1145/3077136.3080667

Sharma, Abhimanyu. 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, Argyro-Maria, and Marina Sounoglou. 2021. "Human Rights and Minority Languages: Immigrants' Perspectives in Greece." *Review of European Studies* 13(1): 55. https://doi.org/10.5539/res.v13n1p55

Social Security Administration, USA. 2020. "Distribution of Given Names of USA Social Security Number Holders." *Baby Names*. 2020.

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

Thelwall, Mike, and Liwen Vaughan. 2004. "A Fair History of the Web? Examining Country Balance in the Internet Archive." *Library & Information Science Research* 26(2): 162–176. https://doi.org/10.1016/j.lisr.2003.12.009

Therneau, Terry M., and Patricia M. Grambsch. 2000. "*Modeling Survival Data: Extending the Cox Model*." New York: Springer.

Tsoukalas, Spyridon, Filotheos Ntalianis, Petros Papageorgiou, and Symeon Retalis. 2010. "The Impact of Training on First Generation Immigrants: Preliminary Findings from Greece." In *2010 2nd International Conference on Education Technology and Computer*, V3-235–V3-238. Shanghai, China: IEEE. https://doi.org/10.1109/ICETC.2010.5529555

Tubergenvan, F., and M. Kalmijn. 2008. "Language Proficiency and Usage Among Immigrants in the Netherlands: Incentives or Opportunities?" *European Sociological Review* 25(2): 169–182. https://doi.org/10.1093/esr/jcn043

Tubergenvan, Frank, and Matthijs Kalmijn. 2005. "Destination-Language Proficiency in Cross-National Perspective: A Study of Immigrant Groups in Nine Western Countries." *American Journal of Sociology* 110(5): 46. https://doi.org/10.1086/428931

Vertovec, Steven. 2007. "*New Complexities of Cohesion in Britain: Super-Diversity, Transnationalism and Civil-Integration*." Oxford: University of Oxford.

Wright, Sue. 2020. "Migration, Linguistics and Sociolinguistics." In *The Sage Handbook of International Migration*, 142–58. London: SAGE Publications Ltd. https://doi.org/10.4135/9781526470416

Wright, Sue, and Claudia Viggiano. 2020. "Language and Incorporation." In *The Sage Handbook of International Migration*, 481–95. London: SAGE Publications Ltd. https://doi.org/10.4135/9781526470416

Yin, Junjun, Guangqing Chi, and Jennifer Van Hook. 2018. "Evaluating the Representativeness in the Geographic Distribution of Twitter User Population." In *Proceedings of the 12th Workshop on Geographic Information Retrieval - GIR'18*, 1–2. Seattle, WA: ACM Press. https://doi.org/10.1145/3281354.3281360

Zagheni, Emilio, Venkata Rama Kiran Garimella, Ingmar Weber, and Bogdan State. 2014. "Inferring International and Internal Migration Patterns from Twitter Data." In *Proceedings of the 23rd International Conference on World Wide Web - WWW '14 Companion*, 439–444. Seoul, Korea: ACM Press. https://doi.org/10.1145/2567948.2576930

Zimmer, Michael, and Nicholas John Proferes. 2014. "A Topology of Twitter Research: Disciplines, Methods, and Ethics." *Aslib Journal of Information Management* 66(3): 250–61. https://doi.org/10.1108/AJIM-09-2013-0083