A Demolinguistic Approach to Study Immigrants' Integration Using Facebook Data

Short Abstract

(200 words)

Researchers acknowledge the key role played by language for a successful integration of immigrants in the destination societies. Demolinguistics, introduced by French demographers at the University of Montréal in the 1970s, provides an ideal framework for linking migration to the demographic composition of linguistic groups and distribution of populations, based on Jacques Henripin's conceptualization of demography, as a science serving society using social media to improve scientific knowledge (Henripin, 1974). This paper presents preliminary results in the field, as part of a Centre of Advanced Studies' project proposal aimed to build-up capacity in accessing and analysing social media, in coordination with European Commission Services. Three axes of research activities are defined. First, the validation of Facebook as source of data for demolinguistic studies through a comparative analysis between Canadian census and Facebook data. The method designs a reweighting method adjusting sample of Facebook users to account for coverage, selectivity and self-assessment biases. Second, the mapping of social interactions of immigrants living in the European Union, considering the language they use in Facebook network as a proxy for their propensity to be incorporated into the European culture. Finally, we investigate the role of English as *lingua franca* across the European Member-States.

LONG ABSTRACT

INTRODUCTION

Language is more than a means of communication. Researchers from different disciplines have recognised language as an expression of culture: ideally, language preserves the belonging to a cultural identity, transmitting its cognitive heritage to new generations. The correlation between language and social outcomes has been explored investigating the labour market employment and occupation status (Ghio and Bratti, 2019), marriage outcomes (Meng and Gregory, 2005 and Ducan and Trejo, 2007), and gender gaps in socio-economic position (Alesina et al, 2013). In the field of migration studies, researchers have established the key role played by language for a successful integration of migrant populations in the hosting societies (Dustmann and Fabbri, 2003).

Demolinguistics (or linguistics demography), first introduced by French demographers at the University of Montréal in the 1970s, provides an ideal framework for linking the linguistic structure of a population to migration and its outcomes. The key role of migration in determining the demographic distribution of linguistic groups in Québec was stated already in the first studies that Jacques Henripin and his colleagues carried out on the subject (Henripin, 1974).

Henripin's conceptualization of demolinguistics reflected also his idea about demography as a science serving society using social media to improve scientific knowledge (Henripin, 1974). This idea could not be more actual today, as we have entered an era of unprecedented data, digital footprints created everyday by cameras, sensors, smartphones, internet, social media, and administrative systems taking records of our collective actions. In addition, this increasing volume of individual traces can be processed using computational methods to model collective behaviours and describe aspects of our society that are not sufficiently well represented through social statistics, or to anticipate trends that are already present but not yet captured or quantified (Billari and Zagheni, 2017).

Giving a renewed meaning to Henripin's definition of demolinguistics, our paper prospects for a different role of social media: from a channel of dissemination, social media data can currently be used for studying social behaviours across countries. This paper is a part of the Centre of Advanced Studies (*CAS*) research project proposal, to build capacity in accessing and analysing private data through key partnerships and in coordination with European Commission Services. The general goal of the project is to streamline the work of scientists in accessing and processing social media data to support the policy making decision process. The package devoted to demolinguistics includes three lines of research. First, the validation of Facebook as source of data for demolinguistic studies through a comparative analysis between Canadian census data and Facebook data. The second research axis is the mapping of social interactions of immigrants living in European Union considering the language they use in their Facebook network (the ‘de facto’ language of their social network) as a proxy for their propensity to be incorporated into the European culture. The third line of research

investigates the role of English as *lingua franca* across the European Member States. In this paper, we present the preliminary findings of these three research activities.

THEORETICAL FRAMEWORK

Immigrants' integration in host societies is a multifaceted concept. In this research, we focus on linguistic integration, which presents its own issues of definition and measurement. Legal statements generally define the “official” language spoken in a country, but the language spoken by an individual cannot be forced by law. This is the reason why in the Canadian census, for instance, a distinction is made between the language spoken at home and at work. But even this distinction might not capture the full linguistics behaviour of an individual at one point in time. In addition, over time individuals do not necessarily have a mono-linguistic trajectory, with no distinction between their private and public language throughout their life. By contrast, individuals can have multi-linguistic trajectories, when the language spoken at home or with friends differs from the country’s official language.

For immigrants, the first step towards linguistic integration is the achievement of proficiency in the official language of their destination country. If compared with spatial movements, linguistic mobility takes a longer period of 'risk exposure' to assimilation for becoming effective: the official language may lead a linguistic transfer (substitution) in the individual private sphere, but the dynamic is uncertain. As a consequence, policy interventions (i.e. investments in training and linguistic programs for migrant populations) may have a marginal impact in a short-term, yet, in a long term, intergenerational transfers could differently shape populations, and their linguistic composition. Thirdly, linguistic profiles depend on personal characteristics that can frequently change during the individual life history.

The language spoken by immigrants is the vehicle of interactions between persons and contexts, conventionally classified making a distinction between 'public sphere' and 'private sphere'. The public sphere is characterized by the use of the official language or *de iure,* i.e. the language spoken by institutions and national authorities. The use of language in the context of private relationships determines the language *de facto* spoken by individuals, including: - mother tongue; - language spoken at home or with relatives; - languages used in social activities.

Immigrants' linguistic adaptation is thus a multi-dimensional process which convers multiple linguistic patterns and the language spoken at home can be an indicator of the individual cultural identity. As a result, a successful incorporation into hosting societies can be assessed by the linguistic transfer of the language spoken at home, from the mother tongue to the official language of the hosting society. The estimated linguistic transfers offer a complementary and/or alternative measure of the naturalisation process, which conventionally identifies the incorporation of immigrants into the hosting societies by the legal change of their citizenship status.

DATA AND METHODS

In North America, official statistics include language variables. On the contrary, the European directive on migration statistics (Reg. n.862/2007) does not include language in the number of variables to be collected at the national level adopting harmonised definitions. Linguistics characteristics of the population are available for a limited number of European member states (i.e. Belgium) but they are not frequently and systematically collected. For this reason, in order to validate the utilization of Facebook data for demolonguistics the first goal of CAS was to exploit Canadian census data.

Through its advertising platform, Facebook allows customers to design targeted advertisements and publish them on the Facebook family of apps and services. Characteristics of the Facebook users can be selected, such as age, gender, location, country of previous residence, and language, to provide information on the number of daily active users (DAU) and monthly active users (MAU). In Facebook's advertising platform documentation, users who have "lived in country X" are defined as "people who used to live in country X and now live abroad"[1]. Facebook's advertising platform also classifies its users based on the languages that they use[2]; users' language classification is provided for approximately 78 different languages. According to Facebook, this estimation is a unique calculation based on users' self-reported demographic characteristics, and is not intended to be aligned with other official statistics. Facebook does not disclose details about the method used for classifying users as expats. A study by Facebook staff Herdagdelen et al. (2016) categorized Facebook users as expats based on their hometown, as reported in their profiles. However, empirical analyses have shown that Facebook also attributes for classifying its users as expats, among other attributes like geo-referenced information (Spyratos et al., 2018). The language spoken by immigrants in the European member states (the language de facto of their social network) versus the European lingua franca. It is possible to get Facebook audience estimates for users speaking two languages (for example: English and Italian) as follows:| (English AND Italian) | = | (English) | + | (Italian) | - | (English OR Italian)|. Basically, the API doesn't support the "AND" but it supports the "OR". As long as the user groups are big enough not to have rounding errors, the correlation can be established.

The method designs a reweighting approach adjusting sample of Facebook users to account for: a. coverage: population observed using Facebook vs population recorded using official data; b. selectivity and self-assessed indicators resulting from user profiles. Facebook Network users' representativeness varies by selected country and individual demographic profile, namely by sex and age. The reweighting method allows us the control of selection bias using Facebook data sources to study immigrants' linguistic adaptation behaviours.

[1] Facebook provides the "lived in country X" classification for 89 countries of previous residence.
[2] Note that the precise definition of "speaks" is unclear but, practically, it probably means "would click on an advertisement in the corresponding language".