DISCUSSION PAPER SERIES

DP13002

THE ECONOMICS OF LANGUAGE

Shlomo Weber and Victor Ginsburgh

PUBLIC ECONOMICS

ISSN 0265-8003

THE ECONOMICS OF LANGUAGE

Shlomo Weber and Victor Ginsburgh

Discussion Paper DP13002
Published 19 June 2018
Submitted 19 June 2018

Centre for Economic Policy Research
33 Great Sutton Street, London EC1V 0DX, UK
Tel: +44 (0)20 7183 8801
www.cepr.org

THE ECONOMICS OF LANGUAGE

Abstract

The paper brings together methodological, theoretical, and empirical analysis into the single framework of linguistic diversity. It reflects both historical and contemporary research by economists and other social scientists on the impact of language on economic outcomes and public policies. We examine whether and how language influences human thinking (including emotions) and behavior, analyze the effects of linguistic distances on trade, migrations, financial markets, language learning and its returns. The quantitative foundations of linguistic diversity, which rely on group identification, linguistic distances as well as fractionalization, polarization and disenfranchisement indices are discussed in terms of their empirical challenges and uses. We conclude with an analysis of linguistic policies and shifts of languages and examine their welfare effects and the trade-offs between the development of labor markets and the social costs that they generate in various countries.

JEL Classification: F13, F22, G11, H11, J15, J3, O10, Z13, Z18

Keywords: Languages, Economic Behavior, Educational Linguistic Policies, Linguistic Distances, Diversity Indices. Welfare.

Shlomo Weber - sweber@smu.edu
New Economic School and CEPR

Victor Ginsburgh - vginsbur@ulb.ac.be
Free University of Brussels

The Economics of Language

Victor Ginsburgh
ECARES, Université Libre de Bruxelles,
and CORE, Université catholique de Louvain

Shlomo Weber
New Economic School, Moscow

June 15, 2018

Abstract

The paper brings together methodological, theoretical, and empirical analysis into the single framework of linguistic diversity. It reflects both historical and contemporary research by economists and other social scientists on the impact of language on economic outcomes and public policies. We examine whether and how language influences human thinking (including emotions) and behavior, analyze the effects of linguistic distances on trade, migrations, financial markets, language learning and its returns. The quantitative foundations of linguistic diversity, which rely on group identification, linguistic distances as well as fractionalization, polarization and disenfranchisement indices are discussed in terms of their empirical challenges and uses. We conclude with an analysis of linguistic policies and shifts of languages and examine their welfare effects and the trade-offs between the development of labor markets and the social costs that they generate in various countries.

Keywords: Languages, economic behavior, educational linguistic policies, linguistic distances, diversity indices. welfare

JEL classification: F13, F22, G11, H11, J15, J3, O10, Z13, Z18

1 Introduction

It is difficult to deny the role of languages in human development. Language makes information operational. One can argue that, as a social technology, it defines the human species. Language affects cultural identity, effective communication in business and international trade, employment opportunities, as well as tourism. The origin of the economics of language as a separate discipline is usually credited to Marschak (1965), who explicitly introduced economic concepts such as costs and benefits into linguistic analysis. It is possible that Marschak's own background, that included the command of a dozen of languages, and the service as Secretary of Labor of the multilingual Terek Soviet Republic in North Caucasus in 1919, sustained his interest in the economics of language.

A large number of articles soon followed that connected language, economics and business. These touched upon an even larger number of topics, including trade, migration, consumer choice, earnings, language acquisition, challenges of multinational corporations in the face of linguistic diversity, and most importantly, its effects on growth, institutional quality, redistribution, regional and national development.

The relatively early contributions by Fishman (1968), Pool (1972), Hočevar (1975), Bretton (1976), McManus et al. (1978), and Grenier (1984) had already stressed the importance of language on social, political, and economic outcomes to political scientists, sociologists, anthropologists and psychologists and, naturally, sociolinguists, though linguist Greenberg (1956) was already aware of the possible connections. But studying diversity to understand human experience has a long tradition that goes back to the Prussian geographer and explorer Alexander von Humboldt (1988, [1836]). Somewhat later, John Mills and Karl Marx indicated that culture and cultural diversity are central to economic interactions.[1] Fortunately or unfortunately, a growing number of economists recognizing their ignorance, eventually showed their interest in the economics of language and produced a fast-growing number of contributions that take on board some economic analysis in their study.

This paper brings together methodological, theoretical, and empirical aspects into the single framework of linguistic diversity, that reflects the history and contemporary interest of the topic. Language represents a singularly important facet of cultural diversity given its ubiquity and centrality to human experience. Its impact on economic outcomes and public policies in the contemporary era has been noticed and examined by economists and other social scientists.[2]

Humboldt's ideas were further developed during the twentieth century by Boas (1940), Sapir (1949) and Whorf (1956) whose contributions were synthesized and called the Sapir-Whorf hypothesis: Language and culture are interdependent, and it may be that language influences our behavior and decisions, though this got strongly rejected in 1955 by Chomsky, who was and still is followed by many linguists.

[1] This was picked up in a theoretical paper by Lazear (1999).

[2] See among others Guiso et al. (2006) for an early survey.

Our paper identifies common traits of languages as a part of culture and, using economic tools, addresses specific characteristics of linguistic diversity and its consequences.

In Section 2, we turn to whether and how language influences human behavior. We start with a brief description of the long march that transformed the (possibly) unilingual "people" into today's multilingual universe and its 7,000 spoken languages. The scope of the existing language menu immediately brings up the question of incentives for learning foreign or local languages. Historically, the main incentive was trade. The transportation, either by walking and later by boats or donkeys, required crossing various cities and ports, where communication and negotiations without a command of the local language must have been a problematic, inefficient and costly task.

The second reason for learning foreign languages is driven by migrations and wars. Some people were forced to leave their country and had to learn the language of their new homeland on the spot, while voluntary migrants were and are, to some degree, able to chose the country of their destination and possibly learn its language before leaving. The emerging multilingual world unleashed two different forces: (a) the need for translation and interpretation,[3] as well as the shifting *linguae francae* (Greek and Latin in the Ancient world, Arabic, Latin again, as scholarly language during the Middle Ages and the Renaissance in Europe, French, for diplomacy, German, for science and, finally, English), and (b) the interplay between two tendencies: Preserving and strengthening local languages versus uniformity and standardization.

We proceed with the Sapir-Whorf hypothesis, rejected by a large number of linguists and other scientists moved by the fear that it would open the door to "relativism" in science. Still economists found interesting enough to verify the assumption by running experiments (Roth et al., 1991), as do psycholinguists, or by analyzing grammars (Chen, 2013; Galor et al., 2016).

We complete the section by discussing linguistic emotions which arise when individuals find themselves in a linguistic environment different from the one they are used to. This could happen in two ways. One is migration, voluntary or not, when people have to live, work and write in a language different from their native tongue. We describe the pain, anguish and insecurity in some cases and the elation and satisfaction in others. The second case follows changes in the linguistic environment resulting from government policies. These may include a shift of the official languages and/or the languages of instruction in schools, and may create linguistic disenfranchisement for several population groups if their linguistic rights are restricted or denied. As we show the outcome of such policies could be very painful.

Section 3 on the multilingual world, lays down the quantitative foundations of linguistic diversity, which rely on group identification, linguistic proximities and diversity related to countries and regions. It is argued that, while some elements are similar to examining other aspects of cultural (religion, ethnicity) or genetic diversity, the investigation of linguistic traits is

[3]See for example Lewis (2004) in the context of the Middle East.

important and relevant in evaluating agents' decision-making and behavior. Introducing linguistic or ethnic characteristics rules out the notion of "representative citizen" (except, perhaps, in North Korea). This forces researchers to cope with the challenges of linguistic diversity, which itself can take many different forms: Patterns of linguistic proficiency (who speaks what); varying aptitudes in learning other languages (some are more skilled or motivated); and costs and returns to learning foreign languages.

We first introduce the difficulties of group identification and the division of a society into distinct (ethno)-linguistic groups. As Lewis (2009) points out: "not all scholars share the same criteria for distinguishing a language from a dialect." In other words, is Venetian an Italian dialect or a language, and should we consider Venetians as being distinct from Romans or Neapolitans? To mitigate the effects of possible ambiguities or errors, it is useful to introduce the notion of proximity, or distances, between languages and linguistic groups. We describe a wide variety of distances. They include lexicostatistical distances (based on similarities and supposed common roots of words in the vocabularies of various languages), Levenstein distances (the distance between the word in one language and the same word in the other one is determined by the number of insertions, deletions or substitutions of single characters (or their phonetic spelling) required to convert one word into the other), cladistic distances (based on linguistic trees which take into account many descriptors), phonetic distances, and others.

The next part of this section offers a formalization of fractionalization, polarization and disenfranchisement indices. Fractionalization indices are based on the division of a society into distinct ethno-linguistic groups. An additional dimension is taken into account in the construction of polarization indices: The identification of individuals with members of their own group and their alienation towards other groups. Disenfranchisement indices generalize fractionalization in a different direction which accounts for the possibility that agents may be multilingual. These indices basically count the number of agents who speak each language in their repertoire (accounting or not for linguistic distances) and allow judging, for instance, which languages to retain as official languages.

Both tools, linguistic distances and indices, combined or not with distances have been used by economists in formulating and estimating equations to explain international trade, migrations, financial transactions, or translations, as well as economic growth, regional development, institutional quality, redistribution, and conflicts. Exemples of these uses will be discussed.

The section concludes with the macroeconomic determinants of foreign language acquisitions, and the financial returns of knowing foreign languages, both for migrants who learned the local language and natives who know foreign languages and use them at work.

Section 4 analyzes some typical linguistic policies in multilingual settings and the possible inefficiency that may arise from using multiple languages in a given society. One of the most often used remedies to inefficiency is standardization, the foundations of which were laid down in Max Weber's (1968, [1910]) theory of rationalization. In our context, standardization im-

plies using a small number of languages (often, one) for official purposes in government, the press, education and trade. While advantages of standardization are often evident, an inevitable outcome is disenfranchisement since the linguistic rights of various groups of people are restricted or even denied. To address this issue, a certain number of drastic linguistic education policy shifts in various countries (Spain, Morocco, South Africa, Kazakhstan, Latvia, West Bengal and the United States) are briefly reported and analyzed. These changes were implemented as a result of a political change (Latvia, Kazakhstan, Spain), public demands (Morocco, India and the US) and changes in government policies for other reasons (South Africa before the abolition of apartheid). Switching the language of instruction to a native language is an important symbol of independence and national sovereignty, but the outcomes on labor markets may generate substantial social costs. We examine the theoretical underpinnings of linguistic policies and language learning in a bilingual setting. The results show that subsidization is sometimes needed to obtain a welfare optimum. Mandatory learning of the languages, however, comes at a considerable cost that could be viewed as the shadow price of national unity.

To examine specific effects of balancing efficiency and the social cost of linguistic disenfranchisement, three-language policies experimented in India, Nigeria and Kazakhstan are also reported. They amount to requiring all individuals to learn not only their own language, but also the language of another region, as well as a world language, most often English. The reform failed in India and Nigeria, but has a good chance to work in Kazakhstan since it benefits from the commitment of the government and this commitment is recognized by Kazakh citizens.

Section 5 concludes the paper and suggests further research.

2 The influence of language on the behavior of economic agents

Whether the first speaking humans came from Africa, or as has been recently hypothesized, from the Middle East (Hershkovitz et al., 2010, 2015), it is pretty clear that they started speaking a unique (or at most a couple of) language(s). This number increased over time, as some subgroups of the first population(s) started their long march to other regions, and languages evolved, changed, and died over the last 150,000-400,000 years. Or perhaps, this happened because God dispersed the builders of the Tower of Babel, which

> resulted in mass lay-offs of workers, but none of them complained. Those, who like me, worked on a foreign construction site, surrounded by workers with different languages and colors, know how difficult it is to start objecting, even if everyone roughly speaks a common language, using improbable accents. I try to imagine what would have happened as we were being fired, if we,

> Turks, Yugoslavs, Algerians, Portuguese and one Italian could not understand each other; but we had a first draft of a common language in a foreign country which made it possible for us to protest, while this was impossible in the valley of Shinar. Therefore, the people left, and dispersed. And of course, they forgot their words, but also the high degree of specialization reached in their work and never again tried to build heaven (de Luca, 1997).[4]

The quotation beautifully summarizes the long march from one to the 7,097 (Ethnologue, 2016) languages that still exist today. We all get born in (usually) one language that we treasure, but we also have economic and cultural incentives to learn other languages to survive in a globalized world as we show in Section 2.1. This is the reason for which we may have to go back to a smaller number of languages to be able to communicate.

Meanwhile, our native language probably also conditions part of our later behavior (Section 2.2), and we certainly get emotional if not infuriated when it is mistreated (Section 2.3).

2.1 Economics and language learning

Learning (or not learning) foreign languages results from several economic incentives. The main is obviously trade. Even in the Ancient World that is more familiar to us (Near East, including Mesopotamia, Egypt, Iran, the Levant, Greece and Rome), people were forced to communicate. Moore and Lewis (1999, p. 271, quoted by Holden, 2016, p. 295) suggest that "their international business achievements were probably more impressive than those of modern business people." And indeed, travelling from the Iranian border to Turkey through Mesopotamia,[5] probably took several weeks, needed transfers of goods between ships and donkeys, long walks in the desert, and speaking in a trustful way with a host of various and often hostile tribes. This needs either interpreters or a short list of common languages. During the Middle Ages, traders would travel between China and Italy along the Silk Road, and somewhat later, ships sailed between Europe and the recently discovered Americas, as well as between Europe, India, China and Japan, facing a long voyage since they had to make their way around Africa.

The second reason for learning foreign languages is driven by migrations and wars. Voluntary migrants or people forced into exile for political, racial or economic reasons have to learn the language of the country of destination. Some do so before leaving, and choose the country of destination accordingly.

[4] Our translation into English from the French translation.

[5] "Transportation was mainly ensured by boats, built in shipyards to be found in almost every city located along a river. Donkeys were used on land, when rivers were impossible to navigate, and, obviously, to cross deserts. Two main roads connected the southern delta and the west. The first led from Sippar to Mari and then, crossing the desert, it reached Qatna and the Phenician ports on the Mediterranean sea. Though longer, the second one was safer as it did not go through a desert; it also started in Sippar, followed the Tigris to Nineveh and then turned west towards the sea, through Nagar and Harran. Because of the mountains, communications with the eastern part of the country were much more problematic." See Bossuyt et al. (2001).

Others migrate and learn the language on the spot. This is not fully new either. Phoenicians travelled and settled in several regions surrounding the Mediterranean Sea. Greeks moved to Asia Minor during the Trojan War, and it took Ulysses a couple of years to travel back home, visiting island after island, where he even had to speak the language of the Mermaids. The Macedonian King Alexander the Great waged war to the East, as far as Persia and India, and to the south-western border of today's Egypt at the oasis of Siwa. Greeks moved to the Roman Empire (Southern Italy and Sicily). Romans occupied Palestine, and Constantinople. Christian crusaders travelled to save Jerusalem from the Arabs. Jews moved from Palestine to the whole Middle East, as well as to Spain and Portugal, from which they were expelled at the end of the 15th century and had to settle in Mediterranean Countries, the Middle East (Iraq, Iran, Syria), and Western Europe. Arabs invaded North Africa in the 7th century, and from there, came to Spain in the early 8th century. All this forced or fostered invaders, as well as locals and migrants, to learn languages.

Interpretation between languages was always needed, but with the birth of writing in Mesopotamia, people started translating. In the third century B.C., the Jewish Torah was translated from Aramaic into Greek by 72 Jewish scholars working in the Ptolemaic Library of Alexandria, and this translation became the most important text from which the Old Testament travelled into other languages.[6]

But the need for *linguae francae* was also felt. After the death of Muhammad in 632 A.D., the Arab invasions in parts of the Middle East and North Africa turned Arabic into a *lingua franca* and it still is so today. Latin was the *lingua franca* of scholars in the dark Middle Ages and the Renaissance. Newton's *Principia*, Spinoza's *Ethics*, Pascal's *Generatio conisectionum*, Gauss' *Disquisitiones Arithmeticae*, Galileo's *De Motu Antiquiora*, Descartes' *Principia Philosophiae*, Fermat's *Methodus ad disquirendam maximam et minimam* and *De tangentibus linearum curvarum*, Paracelsus' *Opus Chirugicum* and many other scientific works were written in Latin.

Somewhat later, French became the language used by diplomats, and German the one used by scientists. Crystal (1999, p. 105) estimates that English is nowadays used by 1.5 billion people, with only 400 million who speak it as their mother tongue.

2.2 Language, culture and the contemporary interpretation of the Sapir-Whorf hypothesis

In the early 19th century, German linguist Wilhelm von Humboldt (1988, [1836]) emphasized that language is more than a means of communication: It also represents the inner life of its speakers. He is said to be at the origin of Edward Sapir's, Benjamin Whorf's, as well as Noam Chomsky's work, though Chomsky and most of his followers strongly oppose Sapir's and

[6]Today the Bible has been translated into over 3,000 languages. Another 4,000 languages are still in the waiting line.

Whorf's ideas. Sapir (1949) and Whorf (1956) were influenced by anthropologist Franz Boas's (1940) cultural relativism and hypothesized that language and culture are interdependent, and more importantly, that the structure of the language that we use influences our way of thinking, observing, and behaving. Whorf (1956, p. 221) goes as far as writing that "users of markedly different grammars are pointed by their grammars toward different types of observations and different evaluations of externally similar acts of observation, and hence are not equivalent as observers but must arrive at somewhat different views of the world."

This bold hypothesis (also called linguistic relativity) that language determines or is at least isomorphic to the structure of our thinking and therefore our culture was met with resistance by the scientific community fearing that it may introduce "relativism" in scientific discoveries.

Actually what Sapir and Whorf wrote is not as strong as one may think. Here is what Sapir mentioned in one of his talks in 1929:[7]

> The fact of the matter is that the 'real world' is to a large extent unconsciously built up on the language habits of the group. No two languages are ever sufficiently similar to be considered as representing the same social reality. The words in which different societies live are different worlds, not merely the same world with different labels attached ... Even comparatively simple acts of perception are very much more at the mercy of the social patterns called words than we might suppose. If one draws some dozen lines, for instance, of different shapes, one perceives them as divisible into such categories as 'straight,' 'crooked,' 'curved,' 'zigzag' because of the classificatory suggestiveness of the linguistic terms themselves. We see and hear and otherwise experience very largely as we do because the language habits of our community predispose certain choices of interpretation.

Whorf's main work on the Hopi Indian language on which he based his assertions, has also been criticized by claims that he did not know the language well enough, and that he gave no citations of Hopi sentences and examples in his book.

It is therefore not surprising that Chomsky rejected the so-called Sapir-Whorf (SW) hypothesis, in its strong form (language determines thought) and even in its weak form (language influences thought) and was quickly followed by many linguists.

His deterministic *universal grammar* has as consequence that all languages share the same underlying structure and its potential to explaining "the striking similarity of language learning in children all over the world, [which] captured the imagination of a generation of scholars" (Li and Gleitman, 2002, p. 266). One of his well-known followers, Steven Pinker (1994, p.232) claims that "according to Chomsky, a visiting Martian scientist would surely conclude that aside from their mutually unintelligible vocabularies,

[7]Sapir (1949), cited by Leavitt (2011, pp. 138-139).

Earthlings speak a single language." If this were true "there is less scope for the large differences among languages that the more extreme linguistic relativists had imagined" (Swoyer et al., 2014, p. 7).

Chomsky's universal grammar is a set of principles that are the same in all languages, as well as some parametric variations. The definite article, for instance, comes in front of a noun in French and English; it is attached to the end of a noun in Swedish and there is no such article in Russian. In some sense, this is comparable to the principles of every spider producing its web. Spiders do not need to learn how to weave, they use their instinct, though there are some parametric variations in building between say, Brazilian and European spiders. But children need to hear their parents, or other people speaking, otherwise they will not speak, but may nevertheless have a grammar wired in their brain. The young spider needs no other spider to learn, but this is not necessarily the case for some mammals which can acquire some non-genetically transmitted actions by looking at their kinds.

There are also linguists who contradict Chomsky, at least partially. Dediu and Ladd (2007) accept that language is related to genes, but that the genetic determinant is not unique. Evans and Levinson (2009) suggest that existing languages result from a combination of historical accidents and other influences, and not from linguistic universals. Leavitt (2011, p. 143) argues that one of the main contributions of Whorf is that "he distinguished clearly between what is *possible* to think for speakers of any language, and what people *habitually* think, which may be strongly influenced by their language."

Nowadays, a weaker hypothesis is thought to hold, which posits that while reflecting cultural preoccupations and constraining our way of thinking, language is also a carrier of culture: "there are cultural differences in the semantic associations evoked by seemingly common concepts" (Kramsch, 1998, p. 13). The consequences, summarized by Kramsch (1998, p. 12), imply that "despite the possibility of translating from one language to another, there will *always* be an incommensurable residue of untranslatable culture associated with the linguistic structures of any given language," or, in somewhat different terms, by Gumperz and Levinson (1991, p. 614): "meaning is not fully encapsulated in lexicon and grammar."

During the very last years, a new group of cognitivist linguists started running experiments the results of which are very close to the SW hypothesis: "the languages we speak profoundly shape the way we think, the way we see the world, the way we live our lives" (Boroditsky, 2009).

The truth or the falsity of the SW hypothesis may be settled by empirical investigation. But this also sets a strong limitation to accepting or rejecting the SW or the Chomsky hypothesis, since it seems impossible to prove or disprove their universality. It may well be that in some cases, one hypothesis is shown to be satisfied, while in others it is not, and in both circumstances, the finding (right or wrong) is often later falsified or rediscussed. But doing this cannot lead to the claim that either of both hypotheses can be generalized. Many experiments have, for instance, been run with color terminologies,[8] or

[8]The fact that the number of terms for basic colors varies across languages (English has 11 words that everyone knows – black, white, red, green, yellow, blue, brown, orange,

spatial expressions in different languages, but they do not seem to lead to generalizations.[9] Moreover, it is not always obvious that some characteristics of language "cause" or are merely "correlated with" some economic facts, and Fabb (2016, p. 45) notes that causation may even go in the other direction: "cultural values of a group of people (ancestors) caused linguistic forms to come into existence perhaps in a situation where the linguistic system offers alternatives, and selection of the chosen alternative is biased by cultural values," which implies that culture and language may at best be correlated. This may even be stronger than what Whorf (1956) wrote: "I should be the last to pretend that there is anything so definite as 'a correlation' between culture and language."[10]

In what follows, we concentrate on a couple of cases discussed by economists, and refer to linguists, psychologists and cognitive scientists when needed. Economists almost seem to have read Casasanto's (2008, p. 67) paper "Who's afraid of the big bad Whorf," in which he recommends using "some sort of extralinguistic data to test" whether some form of the Sapir-Whorf hypothesis holds, since "otherwise, the only evidence that people who *talk* differently also *think* differently is that they *talk* differently." The discussion revolves around three questions that are very often evoked in the linguistics literature: time and how time is related to space,[11] gender of nouns and uses of politeness distinctions.[12]

The representation of time

Whorf started the controversy among linguists with the notion of "time." He studied the language spoken by Hopi Indians and found that it "contains no words, grammatical forms, construction or expressions that refer directly to what we call time, or to past, present or future, or to enduring or lasting" (Whorf, 1956, p. 57). As discussed above, this concept fell into disrepute, and the controversy was almost put to an end by Malotki (1983) who refuted Whorf's views. In a survey article of Malotki's book, linguist Comrie (1983), an authority on the typology of tenses concludes that "Malotki has performed the inestimable service of debunking one of the persistent myths of twentieth-century linguistics." Not fully, though, as cognitive scientists such as Boroditsky (2001), Boroditsky al. (2011), Casasanto and Boroditsky (2008) accept the idea that conceptions of time are different across languages. In particular, Mandarin speakers talk about time vertically (time has and upward or downward trajectory) and use vertical terms to de-

pink, purple, grey – while the Bolivian Amazonian language Tsimane has only three – black, white and red) and that warm colors (such as reds and yellows) are communicated more efficiently than cool colours (blues and greens) has been often studied. A recent paper by Gibson et al. (2017) shows that this has much less to do with perception than with how useful a color is.

[9] See Gumperz and Levinson (1991, p. 622).

[10] Whorf, cited by Leavitt (2011, p. 142).

[11] See Gijssels and Casasanto (2017).

[12] See also Fabb (2016, pp. 42-55) for other examples. Fabb thinks that these are surface properties that "may not represent what people actually know when they know a language."

scribe it, while English speakers see it horizontally (behind and ahead). In Malagasy (Madagascar's native language), the "future comes from behind" (Dahl, 1995). Kuuk Thaayorre, an Australian Aboriginal language, seems to represent time as moving from east-to-west (Boroditsky and Gaby, 2010; Gaby, 2012).

Chen (2013) takes a different way in analyzing the grammar of languages, and in particular into how languages mark the future (FTR for future time reference). He distinguishes the group of languages which *have to* mark the future in speaking, so called strong-FTR languages, and those that *do not have to* mark it, called weak-FTR languages. Paraphrasing somewhat Chen, assume that you want to say "Tomorrow it will rain" in French. Since the future tense of to rain ("pleuvra") exists, you have to say "Demain il pleuvra," and it would be strange if, instead you said "Demain il pleut", where "pleut" is the present tense. In English, you will have to add "will": "Tomorrow it will rain", and you can hardly say: "Tomorrow it rains." French and English are strong-FTR languages. In German, you can say "Morgen regnet es," using the same form of the verb "regnen" (to rain) for tomorrow (morgen), then you would for today: "Heute regnet es.". But you can also use other forms such as "Morgen wird es regnen" which can be translated as "Tomorrow it is going to rain." But German is considered a weak-FTR language.

Using regression analysis, Chen shows that strong- and weak-FTR native speakers differ in the way they account for the future in their savings (and other types of) behaviors. He finds that native speakers of weak-FTR languages "appear more future oriented in numerous monetary and non-monetary behaviors. [They] are 31 percent more likely to have saved in any given year, have accumulated 39 percent more wealth by retirement, are 24 percent less likely to smoke, are 29 percent more likely to be physically active, and are 31 percent less likely to be medically obese." Chen also finds that countries which have weak-FTR native languages save some six percent more of their GDP than strong-FTR countries. The reasoning that explains this counterintuitive result is that strong-FTR languages make the future more distant, and thus less important (Dahl, 2003).

Chen (2013, p. 721) is very careful in discussing whether the form of the language causes the behavior of agents: It may be that "language is not *causing* but rather *reflecting* deeper differences that drive savings beahavior," but he still claims that "much of the measured effects I find are causal," though he does not fully exclude that they also reflect cultural values.

His paper came under strong attacks by linguists, including Dahl, who used to be a convinced anti-Whorfian, but became more lenient and admits some degree of Whorfianism (Dahl, 2003). He also happens to be at the source of the description of tenses (Dahl and Velupillai, 2013) used by Chen. His main worries are threefold:

(i) there is not only one non-past form (such as the future tense) that refers to the future, such as "Tomorrow, it is going to rain" instead of "it will rain", which makes strong- and weak-FTR less dichotomous;

(ii) some strong-FTR languages such as French and Spanish also use the present tense for something they *intend* to do in the immediate future, and we probably more often make intention-based than prediction-based statements, which again blurs the distinction between strong- and weak-FTR;
(iii) there are many other ways to refer to the future than the existence or non-existence of the future tense.

Sutter et al. (2015) ran a controlled experiment to study the relationship between intertemporal choices among 860 six to eleven years old children grown up in German, a weak-FTR or Italian, a strong-FTR language in a bilingual town in Northern Italy. They confirm Chen's finding: German-speaking children are 46 percent more likely to delay gratification than their Italian peers.

Galor and Özak (2016) offer an alternative set of right-hand side variables to strong- and weak-FTRs used by Chen and Sutter that are correlated or in coevolution with time preference and therefore savings.[13] They show that returns to agricultural investment in the pre-1500 CE period represented by pre-1500 C.E. crop yields, crop growth cycles, caloric yields, and confounding geographical controls could also be used as "predictors" of saving habits.

The gender of nouns

The influence of grammatical gender for non-biologically gendered nouns has also generated some research. Though gender of nouns hardly exist in English,[14] they do in many other languages, such as German where "Tisch" for table is masculine, while in French "table" is feminine. The question is whether this has consequences on cognition.

Boroditsky et al. (2003) summarise the findings by psychologists, that indeed gender of nouns affects thinking, but not very deeply. For instance, people who speak two languages in which genders are different for a common noun, quickly realize that the assignment of genders is arbitrary, and has no deep meaning.[15] Still, Borodistky (2009) shows that, though her testing was administred in English, the descriptions given by German and Spanish speakers about a word such as "key" which is masculine in German and feminine in Spanish lead to "hard," "heavy," "metal," or "useful" by Germans, and to "golden," "lovely,", "shiny," and "tiny," by Spanish speakers. Boroditsky (2009) also notes that "abstract entities such as death, sin, victory, or time" are painted as men or women according to the grammatical gender the word takes in their language.

[13] As well as with some other economic behavioral attitudes such as education or smoking.

[14] With some exceptions such as *ship* which is feminine, though nowadays, the *Chicago Manual of Style* and even maritime authorities dislike the idea. See however, the very funny entry "Why is a ship she?" at http://www.glossophilia.org/?p=1411, last consulted June 1, 2017.)

[15] Though this has little to do with Sapir, Whorf and Chomsky, it is interesting to report that "female hurricanes are deadlier than male hurricanes" (Kiju et al., 2014). This is due to the subjective impression that a female-named hurricane is expected to be less severe, and people get less prepared to protect themselves.

Economists also started to play with this idea. Research by Mavisakalyan (2015) shows that in countries where the majority language is gender-intensive, women participate less in the labor force, and speakers of gender-intensive languages are more likely to think that women should not have equal access to jobs.[16]

Politeness distinctions

Tabellini (2008) distinguishes languages according to two rules that govern the use of pronouns :

(i) the use of first and second person pronouns in conversations: In some languages, Italian, for instance, the use of first and second pronouns in conversations can be dropped, in English they cannot;
(ii) in French you can address someone with "tu" if you know her or him well; otherwise, you would use "vous;" this is also the case in Spanish with the familiar "tu" and the more severe "usted." Such differences do not exist in English unless you meet the Queen.

He considers these grammatical rules to be "deep" enough to be used to instrument (in the econometric sense) cultural traits such as trust and respect, which are themselves used in the second stage estimation to explain the quality of institutions.[17]

Additional issues

Heblich et al. (2015) find that, in Germany, regional accents affect individuals' interactions. They tend to cooperate if they come from the same region, and have the same accent, but are more inclined to compete with those who speak with another accent.

Costa et al. (2014) point out that moral decisions can be different when one uses one's mother tongue or an acquired language. Their argument is that native languages elicit more emotional responses than foreign ones, and that lack of emotions lead to more utilitarian reactions.

Ramírez-Esparza et al. (2006) ran experiments on the *cultural frame switching* effect, "where bicultural individuals shift values and attributions in the presence of culture-relevant stimuli." They compared the reactions of US and Mexican monolinguals to those of Spanish and English bilinguals in the US and Mexico and found that bilinguals were more extraverted, agreeable and conscientious when they speak English rather than Spanish, which is consistent with the personality displayed in each culture. So again, the language used seems to have an impact on personality differences.

We thought it could be interesting to also cite the oldest paper that we could find on the influence of language on economic decisions. It describes an experiment run by Roth et al. (1991), actually without intention to

[16] See also Markovsky (2017).
[17] See Galor et al. (2016).

verify which linguistic theory was right or wrong. The authors ran two- and multiperson bargaining experiments (an ultimatum game and one-period market environment) in four different cultural and linguistic environments (Israel, Japan, the United States and Yugoslavia). This led them to conclude that the differences in behavior among countries in which native languages are obviously quite different cannot be attributed to linguistic differences, but probably to cultural differences.

The origins of language distinctions

Galor et al. (2016) go one step further by suggesting that some characteristics of languages (future tense, grammatical gender, and politeness distinctions discussed above) may have been defined by geographical, climatic and agricultural characteristics.

They run probit models where they regress a dummy variable defined by the presence or absence of the future tense on pre-1500 C.E. crop returns, geographic (latitude, elevation, ruggedness, coast length), weather conditions (rain, temperature and pre-1500 C.E. unproductive period), as well as regional fixed effects. They find that a one standard deviation increase in crop returns leads to a 12 to 23 percentage points decrease of the probability of the presence of the future tense, and, given their specification, they can obviously exclude reverse causality. More abundant crops are thus a reason for 'caring less' about the future.

Galor et al. (2016) also relate the existence of grammatical gendered languages and politeness distinctions to plow usage, plow negative crops[18] and ecological diversity in the area in which the languages are spoken. These results would imply that in Chen's (2013) regressions which make future tense variations a right-hand side variable in explaining savings behavior would be not be the main cause.

Humpty Dumpty has the last word

It may be worth quoting the following lines from Lewis Carroll's *Through the looking glass*,[19] which describe the encounter of Alice and Humpty Dumpty:

"My name is Alice, but – "
"It's a stupid name enough!" Humpty Dumpty interrupted impatiently, "What does it mean?"
"Must a name mean something?" Alice asked doubtfully.
"Of course it must," Humpty Dumpty said with a short laugh: "my name means the shape I am – and a good handsome shape it is, too. With a name like yours, you might be any shape, almost."
...

[18]Plow negative crops include root crops, tree crops, millet, sorghum, dry rice, and maize.

[19]Without forgetting that Carroll was not only a writer but also a logician and a mathematician.

"When I use a word," Humpty Dumpty said, in rather a scornful tone, "it means just what I choose it to mean – neither more or less."
"The question is, said Alice, "whether you can make words mean so many different things."
"The question is," said Humpty Dumpty, "which is to be master – that's all."

2.3 Linguistic emotions

Language is obviously a part of one's self, and many philosophers, sociologists, essayists and novelists discuss the subject, and express their emotions. Wismann (2012), a German-born philosopher who lives in France, claims that the French language allows for more complicity between speakers, which is precluded in German because the syntax imposes to finish a sentence with a verb.[20] The German philosopher and sociologist Tönnies (1887) preferred "English for irreverence, German for metaphysical speculation or French for expressions of sensual pleasures" (Bond and Ginsburgh, 2016, p. 234), while Charles V, the Holy Roman Emperor is quoted for having said that he speaks Spanish to God, Italian to women, French to men and German to his horse.

Some famous writers who had to leave their native country for racial problems, such as Elias Canetti (Bulgaria), Paul Celan (Romania) Imre Kertesz (Hungary) or Norman Manea (Romania) kept writing in their native language.[21] Others who went into exile for similar reasons switched to the language spoken in their country of adoption, because they could no longer stand their native language (Aharon Appelfeld), some did so because they felt it more comfortable (Samuel Beckett, Joseph Conrad, Eugen Ionesco). Franz Kafka was unhappy to write in his native German, and longing for Yiddish, but could not speak it. And so was philosopher Jacques Derrida, a French Jew, born and educated in Algeria, because he spoke and wrote in French, and was longing for Berber, Ladino and Hebrew.

Some people have no difficulty to forsake their native language feeling that communication with others who speak foreign languages is important. Most literary writings from the Middle Ages were in Latin, "the language of nobody" (Zink, 2014, pp. 8-19), and certainly not the language of the poor. Nowadays, novels are written in English in many countries all over the world, and writers from Nigeria, India, South Africa, Sri Lanka, Zanzibar, Pakistan, Rhodesia, Hong Kong, the West Indies or Egypt are awarded the British Man-Booker prize.

Some, who used to write in the language of their former colonisers, felt compelled to go back to their native language to "decolonize their mind" (See Ngugi wa Thiong'o, 1986.). But they were rare.

Many wanted "to identify with that which is the furthest removed from themselves; for instance, with other peoples' languages rather than their own" (Ngugi wa Thiong' o, 1986, p. 3). One such case is Léopold Sédar Senghor, a Senegalese poet, who was the first African member elected by

[20] See Bond and Ginsburgh (2016, p. 236).
[21] See Canetti (1977), Manea (2012), for example.

the Académie Française, and became the first President of the Republic of Senegal in 1960. He even wrote the text of the Senegalese national anthem in French, though it contains the following verse "Stand up, brothers, here is Africa assembled,"[22] and though it may be played by typical Senegalese music instruments. Here is why he preferred French to his mother tongue and made French to become Senegal's official language:

> What does using French represent for me as a black author? [...] I think in French; I express myself better in French than in my mother tongue [...] French is one of those great organs in which you can pull all the stops, creating every possible effect, from the most suave sweetness to the ferocity of the storm. It can alternate or simultaneously be the flute, the oboe, the trumpet, the tam tam and even the canon. And French has given us its words for the abstract – which are so rare in our mother tongues in which tears become gems. Our words bear the aura of sap and blood; words in French shed the rays of a thousand lights, like diamonds. Rockets which set alight our night (Senghor, 1956, and Senghor, 1962, p. 842).

In the wake of independence gained by British, French and Portuguese colonies in the 1960s and 1970s, many African politicians who had been educated in western Europe reacted in the same emotional way, and imposed the language of the coloniser as official language in their country. This had as obvious consequence that the "majority of Africans are governed in a language that they do not understand" (Phillipson and Skutnabb-Kangas, 1995, cited by Spolsky, 2004, p. 182).

A long time ago, Thomas Mowbray, the Duke of Norfolk, faced a similar situation, described in William Shakespeare's play *King Richard the Second*. The King tries to convince his relative, Henry Bolingbroke, and Thomas Mowbray, to stop quarrelling. Failing to do so, he punishes both, and Mowbray is banished forever from England. His reaction is interesting, since he does not lament over the loss of land or status. In his complaint of despair and hopelessness, he rather talks about the inability to speak his native language in exile:

> The language I have learn'd these forty years,
> My native English, now I must forego;
> And now my tongue's use is to me no more.

To complete this sad story, Mowbray passed away in Venice, about a year after his banishment, without having been able to understand a word of Venetian.

This is obviously not an isolated episode. Various decisions and policies tend to disenfranchise individuals or groups whose linguistic, cultural, and historical values and sensibilities are perceived to be under threat. Linguistic disenfranchisement (Ginsburgh and Weber, 2005, 2011) occurs when

[22] Debout, frères, voici l'Afrique rassemblée.

linguistic rights are restricted or even denied. It is unavoidable that linguistic policies may entail social sacrifices and one has to balance the common interest and those of the various groups. But the results could be painful and we briefly discuss two examples (Sri Lanka and Rwanda) which show how emotional and sometimes explosive linguistic policies can become.

Sri-Lanka is populated by two major ethnic and linguistic communities: Sinhalese, predominantly Buddhist, and Tamil, predominantly Hindu. They had peacefully coexisted during about 2,000 years. After 150 years under the British rule, the island attained self-governance in 1948, which triggered attempts by both groups to promote their language to replace English as the official language. Fearing Tamil dominance, the Sinhalese majority forced the signature of Sinhala-only Act, that prompted an emotional and violent reaction from the Tamil minority. A long civil war ensued which claimed thousands lives and devastated the country.

The other interesting case is Rwanda, where the official language was French after the Belgian colonisers had left in 1962, while both the Tutsi and the Hutu used to speak the same native language, Kinyarwanda. In 1994, during the dramatic events between the Tutsi and Hutu, many Tutsi who had, for some time, found refuge in English-speaking Uganda came back to Rwanda and ruled the country from that time on. French was the language of education since 1962. English was added in 1994, and became the unique language of education in 2008, in a country where the majority of inhabitants live in the countryside and speak Kinyarwanda. Two main reasons come to mind in explaining this change to English. First, since the new Rwandan leaders are Tutsi who lived in exile in Uganda, they may have thought that English would be less felt as a colonial language than French that was spoken in the country before its independence. But the decision may also be more political, since the Hutu had been supported, and some say, protected, by France during the 1994 slaughter.

Emotions are obviously activated by the languages that one speaks, and emotions have effects on economic, social and political decisions. But to analyze their effects, some formalization is needed. We surmise that they can be partly formalized using notions such as ethnolinguistic fractionalization, polarization and disenfranchisement that will be studied in Section 3.2.2.

3 A multilingual world

3.1 Identification

In analyzing the linguistic landscape of multilingual societies and its potential impact on economic outcomes, one immediately faces the need of two types of group identification. One concerns the determination of group boundaries – how does one define a partition of the country or countries into separate groups? Another is group association – how do individuals identify themselves with a community to which they presumably belong?

A major challenge in drawing a linguistic or ethnographic map is the prevalence of multiple identities. People may speak several languages using

them in communication across different cultural zones and under different circumstances. One needs several assumptions to proceed. An important one is the determination of the dominant linguistic identity. The first attempt of creating a comprehensive world *atlas* was undertaken by Soviet ethnographers in the Miklukho-Maklay Research Institute in Moscow. The result, called ELF (Ethno-Linguistic Fractionalization), was published in *Atlas Narodov Mira* (1964). This remarkable dataset was chosen by Western scholars, starting with Rustow (1967), Taylor and Hudson (1972) and for almost fifty years played the crucial role in analyzing the impact of linguistic diversity on growth, investment in public goods, quality of government services, corruption and so on. Fearon (2003), Alesina et al. (2003), Alesina and Zhuravskaya (2011), Desmet et al. (2009), among others, developed more advanced fractionalization datasets.

Another issue is that the identification of distinct languages or dialects may not be straightforward. Are Serbian and Croatian different languages? Should various dialects of Italian, German and Mandarin be treated as separate languages? Even though the answers to such questions should be provided by linguists, they matter for economic analysis. While economists are not qualified to determine whether Serbian and Croatian are the same language or not, they can hedge against the statement that these two languages are different, using the notion of linguistic proximity and take advantage of the fact that Serbian and Croatian are very close to each other. Thus, the notion of linguistic distance, which is extensively discussed in this section, allows mitigating the effect of the simple dichotomy of being either identical or different.

Still, if, for instance, one takes,the different dialects of Italian as constituting different groups, Italy would appear to be very diverse. If one considers these dialects to be only minor variations of Italian, then Italy turns out to be quite homogeneous. Desmet et al. (2012, 2017) propose another approach to identify groups. They use a linguistic tree based on the information on more than six thousand existing languages, and study different group structures that depend on the level of aggregation on the tree. Coarse linguistic divisions, obtained at high levels of aggregation, describe cleavages that emerged thousands of years ago, while lower levels generate finer partitions that emerged more recently. They also argue that cleavages may differ according to the research question (economic growth, patterns of redistribution, provision of public goods, conflicts, ...). The proper level of aggregation and the corresponding linguistic partition of the society should be determined by the econometric results from regressions of the variable one studies (say, redistribution, or economic growth), on indices based on ethnolinguistic partitions. They find that high levels of aggregation give better results to explain redistribution and conflicts, since they are tied to solidarity and empathy which go a long way back in time, while low levels are needed to explain economic growth, which is linked to more contemporaneous barriers such as coordination and communication between economic agents.

While the issue of objective identification is extensively discussed in the economic literature, the question of self-identification requires more atten-

tion. In their study of polarization, Esteban and Ray (1994) examine the abstract notions of identification with one's own group and alienation towards other groups. Castaneda-Dower et al. (2017) study the rise of alienation levels of various groups based on the historical patterns of English acquisition during the pre-colonial period in the protracted Sri Lankan civil war, where the linguistic divide played a crucial role and was claimed to be responsible for the outbreak of the war (Tambiah, 1986, DeVotta, 2003). But still more empirical research is needed to study individual identity choices to associate them with one linguistic group or another. In the U.S. context, for instance, how strong is the association of individuals with African American Vernacular English (AAVE) and New York Latino English (NYLE)?

Obviously, identification is driven by the fear of being rejected from one's own community if one chooses to speak Standard English instead of the vernacular that the majority of the community speaks. On the other hand, the societal "stigma" (Besley and Coate, 1992) may lessen the ties with one's own community. It could be that we should split the entire population into three groups: Those who learn Standard English as their first language, those who learn a nonstandard dialect of English natively, and those who do not learn English as their mother tongue (Baugh, 2009).

3.2 Linguistic distances and indices

3.2.1 Measuring distances

Linguists are interested in two types of linguistics that will be of use in this paper: *synchronic* linguistics and *historical* or *diachronic* linguistics. Synchronic linguistics deals with languages at one point in time, and leads, in particular, to computing *distances* between languages say, today: how far is English from Swedish, can an American communicate with a Swede if each speaks or writes in his native language, or will they have to learn a common language (or use an interpreter) if they want to communicate or trade.

But like in genealogical research which aims at building family trees and finding birth and death dates for the members of a family, historical linguistics thinks in terms of reconstructing the earlier stages of one or several languages. This also leads to linguistic trees (*cladistics*). However this still does not tell us at what time a language branched of from a common stem and created a new language. French, Spanish and Italian have a common root. When did they separate? This is the object of *glottochronology*.

Economists use the results of both synchronic and diachronic linguistics, and may even be led to combine the two approaches. They are, so far, not interested in glottochronology.

Several elements contribute to languages being related or not. *Vocabulary* is the most obvious – the French word *lune* for *moon* has the same origin as the Italian and Spanish word *luna*, and so are *moon*, *Mond* and *maan* the English, German and Dutch words for the same object. But French and English words are not close, and indeed, the subsets of languages they belong to consist of two different linguistic families (Romance and Germanic), that branched off from Indo-European, which may itself have a common ancestor

with other families, such as Afro-Asiatic or Niger-Congo which all separated from another Ur-language at an earlier point in time.

Phonetics comes next. The sound correspondence between *moon*, *Mond* and *maan* is close, and so is the one between *lune* and *luna*, which shows there are also phonetic differences which generate more or less relatedness. The final *e* in *lune* is mute, while the *a* in the Italian or Spanish *luna* is not, but the words are obviously related. Diphthongs, that is, sounds composed of two vowels joined to form one sound, as in *sound* also add to the difficulty of a language, and the diphthong *ou* in *sound* is not pronounced the same way in the English words *wound* or *tour*.

Syntax, the "way in which linguistic elements (as words) are put together to form constituents as phrases or clauses,"[23] illustrates another difference or relation between languages. *I would like to observe the moon* translates into *Ich möchte den Mond beobachten.* While the word *observe* is located in the middle of the English sentence, its translation, *beobachten*, ends the German sentence.

Grammar contributes to heterogeneity as well. German has declensions, but only very few of them remain in English such as the so-called possessive genitive in *the moon's last quarter.* None of the declensions that existed in Latin were inherited by today's French, Italian or Spanish, though all three descend from Latin. German has three genders, masculine (*der Mond*), feminine (*die Sonne*) and neutral (*das Wasser*), in English, all three nouns *the moon*, *the sun* and *the water* are neutral.

These are just a few examples to illustrate that computing the relatedness of languages has many facets, and this is even without going into the fundamental issue of whether languages have a common structure.[24]

Both the synchronic and the diachronic methods date back to the 19th century and are based on comparing languages. Diachronic methods go deeper than synchronic ones, since they try to determine "common ancestry, and descent through time with gradual divergence from [a] common source" (McMahon and McMahon, 2005, p. 3), but are not interested in the intercommunication of people who speak different languages today. They use morphological resemblances, such as lexical items, syntax, grammar and sound correspondences, and identify groups and subgroups of languages according to their similarity. The final result is a linguistic tree, with a root (the Ur-language) and branches which in turn grow into sub-branches, until one reaches final twigs, each of which corresponds to a unique language that may still be spoken, or is extinct, and breaks the extension of the tree starting from that twig.[25]

The result is comparable to what biologists (with the aid of scientists from other disciplines such as palaeontology) adopt to construct biological and evolutionary trees. To generate a tree, they start with a table which lists several animal or vegetal species as well as characteristics describing

[23] *Merriam–Webster Dictionary.*

[24] On the fights that this controversy still generates, see Harris (1993), and Baker (2001).

[25] See Nakhleh et al. (2005) for a description and an example of the construction of the Indo-European language tree.

the various species supposed to have a common ancestor. Trees are then constructed using the information given by the descriptive characteristics and computed by algorithms to identify the one that is considered "best."

Table 1 gives an example of such a comparison based on the lexicon only. It contains a (very short) list of words for the basic meanings of numbers one to five in 15 languages, the first of which is English. Linguists are supposed to find languages which are related and classify them into groups. The reader is invited to exercise his skills on guessing the other languages. He will quickly realize how much effort this would take if he were faced with a list of 1,000 words and 200 languages.

[Insert Table 1 approximately here]

In this case, it is not very difficult to group the languages (Table 2), with their names and the numbers that they were carrying in Table 1. The first four groups are Indo-European languages, which are themselves subdivided into Germanic, Romance, Celtic and Slavic families. The two last consist of Hungarian, a language spoken in Europe, but that belongs to the Uralic and not to the Indo-European family, and Swahili, one of the many Bantu languages spoken in Africa.

[Insert Table 2 approximately here]

The digits in the first five languages (Danish, Dutch, English, German, Swedish) are all related, as the pairwise comparisons show. They belong to the family of Germanic languages. The same can be verified for the next four languages (French, Italian, Portuguese, Spanish), which belong to the Romance family. The third group (Celtic) also shows that the words are related. But more importantly, the first three digits in all three families can be seen to be related as well (and indeed, Germanic, Romance and Celtic languages are part of the larger family of Indo-European languages). Such decisions are less obvious and need more linguistic knowledge for digits four and five. Next come Slavic languages, which are very clearly related to each other and for which there is also a relation for digits two and three with the previous groups, but this is less so for the other digits. The last two languages, Hungarian and Swahili have nothing in common with the four previous Indo-European families, but may nevertheless have a faraway ancestor.[26] In general, decisions need trained linguists,[27] as the following example, borrowed from Warnow (1997, p. 6586), shows. The Spanish word *mucho* has the same meaning as the English *much* and is obviously phonetically very similar. Sound change rules do, however, indicate that they *do not* come from a common ancestral word: *mucho* is derived from the Latin *multum* meaning *much,* while *much* is derived from the Old English *micel* meaning *big.*

[26]See Ruhlen (1994) for an entertaining, but nevertheless deep and instructive, exposition. He makes the reader construct similarities and guess which languages belong to the same family. The book reads like a very good crime story, in which the detective is not looking for a criminal, but for the very first language.

[27]Though there are now efforts and experiments to computerize this step. See McMahon and McMahon, 2005, pp. 68-88.

The comparative method can be made technically loose by using as much information as possible (*mass comparison*), or very tight by using a unique characteristic of a language, its *lexicon* (which leads to *lexicostatistics*), but can also accommodate median situations, which are reproducible (or falsifiable) by other researchers. We now turn to some of the most common methods used in synchronic and diachronic analyses.

Mass comparison

Mass comparison was used by Stanford linguist Joseph Greenberg (1955, 1963, 1987, 2000, 2002) to classify native American, Indo-Pacific, African, as well as Eurasiatic languages. It is, claim McMahon and McMahon (2005, p. 19, 22),

> so straightforward and non-technical that in the eyes of many historical linguists it scarcely qualifies as a method at all. As Wright (1991, p. 55, 58) puts it "First, forget all this stuff about rules of phonological correspondences. Second, forget all this stuff about reconstructing proto-languages. Third, write down words from a lot of different languages, look at them, and wait for similarities to leap out ... Greenberg doesn't spell out criteria for deciding when two words correspond closely enough to qualify as a match. Greenberg himself may not need such pedantry; his intuitive sense for linguistic affinity is the subject of some renown. But other linguists may. And science is supposed to be a game anyone can play,

which implies that what Greenberg did is difficult to repeat or to test statistically. Ruhlen (1994), a student of Greenberg, used this method to reconstruct twenty-seven words of the very first language.

Lexicostatistics

Lexicostatistical methods are based on one dimension only: the similarities and supposed common roots of words in the vocabularies of various languages, as is done in Tables 1 and 2. Languages can be related or similar, and their similarities can be explained by three mechanisms only: (a) There may be words that look common for accidental reasons; this is so for onomatopoeic words; (b) languages may also borrow words from other languages: Some 30 percent of English words were borrowed from French after the Norman conquest in 1066;[28] (c) the words in two languages may descend from a common, older language. These are the ones that matter and are used to compute lexical distances.[29]

[28] According to Janson (2002, pp. 157-158), "around 90 per cent of the words in an English dictionary are of French, Latin or Greek origin." This however does not make English closer to French, since "if one counts words in a text or in a recording of speech [in English], the proportion of Germanic words is much higher, for they are the most frequent ones, while most of the loans that figure in a dictionary are learned, rare items."

[29] Ignoring borrowed words may rule out some factors that influence the closeness of two languages, and the ease of learning the other if one knows one of them.

Since it would be a daunting task to compare long lists of words for each couple of languages, linguists rely on a small selection of carefully chosen words, a so-called *list of meanings*. Such a list was compiled by Swadesh (1952), who chose meanings that are basic enough to exist in all languages and cultures (such as animal, bad, bite, black, child, die, eat, eye, hunt, digits from *one* to *five*),[30] on which deductions can be based.[31] The list we are interested in consists of 200 meanings, which Swadesh later trimmed to 100. Both lists are still in use nowadays.

Distances between couples of languages can be computed in several ways. We shortly examine (i) *lexicostatistical distances*, based on the percentage of *cognate* words (that is, words based on a historical chain linking them via an earlier language) in each couple of languages; (ii) *Levenshtein distances*, based on analyzing and comparing words (or their phonetical representations) character by character; (iii) distances based on l*inguistic trees*; (iv) other methods.

Lexicostatistical distances

Dyen et al.(1992) used Swadesh's basic list of meanings to classify 84 Indo-European speech varieties.[32] The lexicostatistical method consists of three steps:

(a) Collecting for each meaning in Swadesh's list the words used in each speech variety under consideration;

(b) Making cognate decisions on each word in the list for each pair of speech varieties, that is, deciding whether they have a common ancestor, or not;[33]

(c) Calculating the lexicostatistical percentages, i.e., the percentages of cognates shared by each pair of languages; these percentages lie between one if all words are cognate, and zero if there is no pair of cognates.[34]

If the only meanings to be compared were the five digits in Tables 1 and 2, then the distances between each pair of the five Germanic languages would all be equal to 5/5. The same would be true for the pairwise distances within the two other families. Things get a little more difficult across the three families, since the words for digits four and five (compare the English *four* with the French *quatre* or the Welsh *pedwar*) are certainly further apart. Assume that

[30]Note that Swadesh's list has been slightly changed to accommodate Southeast Asian and Australian languages.

[31]See Kessler (2001, pp. 199-257) for the lists of meanings chosen by Swadesh.

[32]The conjecture that Aryan languages spoken in parts of India and European languages may have a common ancestor was already made by William Jones, in 1786, and is still holding to day. See Gray and Atkinson (2003), Gamkrelidze and Ivanov (1990), and Ruhlen (1994) for a general overview.

[33]See Warnow (1997) for further technical details.

[34]It is not always possible to make a clear distinction between cognate and non-cognate; therefore, linguists usually add a third group of "ambiguous decisions."

as linguists, we decide to classify them as non-cognate. Then, the distance between, say English and French, would be 3/5. Finally, the distance between any Indo-European language and Hungarian or Swahili would be equal to 0/5. To present these percentages in the form of distances that economists are used to, they are recalculated as one minus the percentage of cognates.

Levenshtein distances

Instead of expert advice used to compute lexicostatistical distances, Levenshtein (1966) suggested using an algorithm that enables measuring the distance between strings, for example those formed by words. The idea is to convert the word of one language into the word of the other one by inserting, deleting or substituting alphabetic (or phonetic) characters; the minimal number of such transformations, divided by the maximum number of characters between the two words is the Levenshtein – also called *edit* – distance between the two words.

As an example, let us consider the word *night*, one of those in Swadesh's list of 100 words. The word is spelled *Nacht* in German and *notte* in Italian. A linguist would probably classify the three words as cognate. The Levenshtein distance between the English and the German words is two, since one needs to substitute *i* by *a* and *g* by *c*. The distance between English and Italian is four (substitute *i* for *o* and *g* by *t*; delete *h*; insert the final *e*). It so happens that the three words have the same number of characters. So the distance between the English and the German words is 2/5 while the one between the English and the Italian words is 4/5.[35] This is in accordance with lexical distances: English and German are Germanic languages, while Italian is a Romance language, but all three are Indo-European.[36]

The distance between two languages is now simply the average of the distances over all the meanings that are compared. Here also, one can compute linguistic trees using clustering algorithms.

Calculating Levenshtein distances is easy to program on a computer, while it is tedious for lexicostatistical distances, which need decisions made by linguists. It may however lead to problems since obvious non-cognate words may be considered cognate by the computer.[37] Levenshtein's method is often performed by taking account of phonetic similarities after transcribing words into their phonetic equivalents, using existing or especially tailored phonetic alphabets.

According to McMahon and McMahon (2005, pp. 212-214), the technique is applicable to dialects, but it would "compromise the method if it were extended to comparisons between languages or across considerable spans of time, since it would then be more likely that changes in the order of segments

[35]If the number of characters is not the same, one divides by the largest number of characters.

[36]Phonetic spelling can be used to replace alphabetical spelling.

[37]A nice example is the small Levenshtein distance of 1/6 between *kitten* and *mitten* which only needs the substitution of the first character in *kitten* by an *m*, though the words have little to do with each other. This is less likely with Swadesh's 100 list, which compares words that have the same meaning in *different* languages.

[within words] would have taken place."[38] This approach was pioneered by Goebl (1982), and Kessler (1995) for Irish dialects, and more recently by Nerbonne, Heeringa and associates for Dutch dialects. See Nerbonne and Heeringa (1997).

Cladistic distances

An alternative way to compute distances is to use linguistic trees, based on world classifications of languages such as the ones in *Ethnologue*, and compute cladistic distances. These lie somewhere between mass comparison and lexicostatistical or Levenshtein distances, as they account for various aspects that characterize languages, including lexicon of course but also, syntax, phonology, grammar. Such trees are available for almost all languages in the world. As will be seen, however, they are less precise than lexical distances.

Fearon and Laitin (1999), Laitin (2000), and Fearon (2003) who suggested this approach[39] use the distances between the branches of a linguistic tree as a proxies for distances between linguistic groups. In the original Fearon and Laitin (1999) index, the score takes the level of the first branch at which the languages break off from each other for every pair of languages. The higher the number, the higher the similarity of languages. This approach was later used by many researchers.

The (very simplified) Indo-European language tree of Table 3 is used to calculate such distances for some languages.[40] Czech and Hungarian come from structurally unrelated linguistic families: Czech is an Indo-European language, while Hungarian belongs to the Uralic family. Therefore, the two languages share no common branches and break off on the first branch: their score is 1. Czech and Italian share one common level since they are both Indo-European, but separate immediately after that, making their score equal to 2. Czech and Russian share two classifications. They are both Indo-European and Slavic, and break off on the third branch, as Russian belongs to the Eastern branch of the Slavic group, while Czech is part of the Western branch: their score is 3. Czech and Polish share three common levels. In addition to being Indo-European and Slavic, both belong to the Western branch of the Slavic group, and their score is 4. Finally, Czech and Slovak belong to the Czech-Slovak sub-branch of the Western branch of the Slavic group, which sets their score at 5. In order to produce linguistic distances the similarity measure r_{ij} between languages i and j is first normalized to fit the interval $[0, 1]$. For a break on the first branch, $r = 0$, for a break on the second branch, $r = 0.2$, for a break on the third branch, $r = 0.4$, for a break

[38] They give the example of the words *bridde* and *friste* in Middle English, which become *bird* and *first* in Modern English. There are other similar issues that would make the use of Levenshtein distances inappropriate. They suggest adaptations that can be found in Heggarty et al. (2005). See also McMahon and McMahon (2005, pp. 214-239).

[39] According to McMahon and McMahon (2005, p. 125), a similar method had been suggested some years earlier by Poloni et al. (1997).

[40] Though Indo-European languages were among the first to be discussed and represented under the form of a tree, this is now the case for all the world's languages.

on the fourth branch, $r = 0.6$, for a break on the fifth branch, $r = 0.8$, and for identical languages, $r = 1$ (Laitin, 2000, p. 148). The linguistic distance d is then simply equal to $d = 1 - r$.

[Insert Table 3 approximately here]

Fearon (2003) produces a dataset for 822 ethnic groups in 160 countries. However, he points out that an early break-off between two languages in such a tree generates a higher degree of dissimilarity than later break-offs. Therefore, the resemblance function r_{ij} should increase at a lower rate for larger values of distances. Fearon suggests using the square root of linguistic distances, rather than distances themselves.[41]

Analyzing sound

The Functional Phylogenies Group analyzes the properties of phonetic sound "that include pulse, intensity, sound wave components, spectrum, and/or duration of the examined sound segment, as well as fundamental frequency, [which is what the] listener identifies as pitch, and relates to how fast the vocal folds of the speaker vibrate during speech" (Hadjipantelis et al., 2012, p. 4652) as well as speech sound evolution. Speech sounds are treated as (continuous) functions (instead of discrete points) that are studied using statistical methods (such as principal components (Aston et al., 2010), and regression models). The group hopes to construct cladistic trees. Aston et al. (2012) give an example of how the meaning of the number 100 in Latin (centum) later separated between Italian (cento), on the one hand, and Spanish and French, on the other with *cien* and *cent*, respectively.

Distances based on learning scores

The approaches described so far do dont take into account intercommunication between populations today. But they have served many other purposes, and their connection with genetics, migrations patterns and archeology are of particular importance.[42] But economists are also interested in today's world and to what makes trades, migrations, or translations easier.

The relative arbitrariness of representing the distance between two languages by a unique and encompassing number or giving weights to different characteristics of languages and aggregate them is reductive. An alternative method, which takes into account all characteristics, including borrowed words, would be to follow the speed at which people with different native languages learn foreign languages. Such a measure was established by Hart-Gonzalez and Lindemann (1993) using a sample of native Americans who were taught a variety of languages. If such distances were available for a large number of language pairs (and measured according to the same criteria), they would certainly be a very good alternative to the distances discussed earlier

[41] A variant of Fearon's formalization is used by Desmet et al. (2009).

[42] See Cavalli-Sforza (1997, 2000), Cavalli-Sforza and Cavalli-Sforza (1995), Cavalli-Sforza et al. (1994), Renfrew (1987) and Michalopoulos (2012) among many others.

since they encompass most of the difficulties encountered in acquiring a language, and borrowed words would also find their place in possibly easing the learning of the other language. To our knowledge, this is the only set of consistent data on learning, and one can hardly imagine the amount of money, time, and effort it would take to set up a coordinated project that would use this method, even if it were implemented "only" for the 2,450 combinations of the world's fifty most important languages.

Problems in using distances

The distance between British and American English is close to zero, if not zero, whatever the method used to measure it. This of course can raise eyebrows. There is a large (and ever increasing) number of meanings that are represented by different words in the United States and Great Britain, and this gets even worse with languages such as 'Spanglish' in the United States, 'Konglish' (spoken by an older generation in South Korea), 'Singlish' in Singapore or 'Globish' everywhere. This phenomenon is obviously not limited to English.

A further restriction is symmetry, which implies that the degree of difficulty experienced by a Spaniard to learn Portuguese is the same as the one experienced by a Portuguese to learn Spanish. This is probably true as far as vocabulary is concerned. However, given that Portuguese phonetics are richer than Spanish phonetics, it may be easier for a Portuguese to learn Spanish than the other way round. Learning scores would not necessarily be symmetric, as is the case for all other methods.

3.2.2. Ethno-linguistic fractionalization, polarization and disenfranchisement indices

The building blocs described earlier, namely ethnolinguistic group identification and linguistic distances, are combined to define fractionalization, polarization and disenfranchisement indices in a given society. In each case, we distinguish two types: dichotomous indices which take into account the dimension of groups only, and distance-weighted indices.

Fractionalization indices

Dichotomous fractionalization indices are defined for a multilingual society divided into distinct groups, each member of which speaks the same native language.[43] Let the society consist of K groups, $k = 1, \ldots, K$, where the population of group k is given by N_k and $\sum_{k=1}^{K} N_k = N$. Let also $n_k = N_k/N$ be the fraction of group k in the society.

The A-fractionalization index is defined as the probability that two individuals, randomly picked from the entire society, belong to two different

[43]Here one disregards proficiency in other languages of each group.

groups. It is formalized as follows

$$A = 1 - \sum_{k=1}^{K} n_k^2.$$

Introduced by Gini (1912) who called it the *mutuality index*, it was later rediscovered by Simpson (1949) and Greenberg (1956), who called it the *monolingual non-weighted index*.[44]

Obviously if a society is monolingual, all its members belong to the same group and the value of the A-index is zero. But if the society consists of many small groups, the probability that two randomly chosen individuals belong to two different groups may become quite large and the value of the index increases.

Another important dichotomous index is Shannon's (1948) and Wiener's (1948) entropy measure:

$$E = -\sum_{k=1}^{K} n_k \log n_k.$$

This measure is actually more often used in biology, statistics and information science, than in the social sciences. Both the A- and the E-indices have similar mathematical properties. Their formulation was unified through common axioms by Davydov and Weber (2016),[45] who suggested the following general form:

$$A^\alpha = 1 - \sum_{k=1}^{K} n_k^\alpha,$$

where α is a positive parameter different from one. Obviously, the value of A-index coincides with A^α for $\alpha = 2$. It is also quite easy to verify that A^α approaches E when α tends to one.

The value of α in the above formulation can be interpreted as the degree of societal sensitiveness towards diversity. Countries, regions, or cities may differ with respect to their own valuation of fractionalization. Some could feel threatened by diversity, others may welcome it. These attitudes may be correlated with profound economic and political outcomes. Ottaviano and Peri (2006), for example, show that Los Angeles, New York and San Francisco live with a substantially higher degree of linguistic diversity than, say, midwestern cities such as Cincinnati and Indianopolis. The differences in "perceived diversity" which signal how people feel about the impact of globalization – channeled through employment prospects and the presence of immigrants in their own communities – are much less obvious. In other words, different societies may choose different α's, and estimating this parameter could be an interesting topic of research.

The A-index may, however, produce unexpected results, as shown by Desmet et al. (2009) who compare its value in two European countries, Andorra and Belgium. In the small south-european principality of Andorra, half

[44]Note that it is also a reversed Hirschmann-Herfindahl index often applied to estimate the degree of industrial competitiveness.

[45]See also Hill (1973) and Simovici and Jaroszewicz (2002)

of the population of some 100,000 citizens, share Catalan as native tongue, whereas the other half speak Spanish. In Belgium, the split between the Dutch-speaking and the French-speaking populations is about 60 and 40 percent. Simple algebra shows that A is equal to $1 - 0.5^2 - 0.5^2 = 0.5$ in Andorra and $1 - 0.6^2 - 0.4^2 = 0.48$ in Belgium. Andorra seems thus linguistically more diverse than Belgium. The reason for this bizarre conclusion is that the A-index does not take into account the proximity between languages: Catalan, Spanish and French are Romance languages while Dutch is a Germanic language.The incorporation of linguistic proximities, Catalan and Spanish in one case, French and Dutch in the other would (and does) make Belgium more diverse than Andorra.

To address this issue, Greenberg (1956) introduced a monolingual weighted index, which, in simple words, accounts for an average linguistic distance between randomly chosen individuals within a society. In addition to the previously defined notation, let d_{ki} denote the linguistic distance between two groups $i, k = 1. \ldots, K$. This distance can be derived via any of the methods described in Section 3.2.1. Greenberg defined his index as:

$$B = \sum_{k=1}^{K} \sum_{i=1}^{K} n_k n_i d_{ki}.$$

It is easy to see that the B-index generalizes the A-index, in which dichotomous distances are replaced by an arbitrary distance metric. Indeed, if we assume that $d_{ki} = 0$ if $i = k$ and $d_{ki} = 1$ if $i \neq k$, B turns into A:

$$B = \sum_{k=1}^{K} n_k(1 - n_k) = 1 - \sum_{k=1}^{K} n_k^2 = A.$$

Bossert et al. (2011) offer an axiomatic foundation for a variant of the B-index. They, however, rely on primitives of individuals that are not preassigned and do not exogenously determine groups within a society in the way ethnic groups do.

The B-index obviously requires more data than the A-index, but the effort is rewarding. This is a clear outcome of the results by Desmet et al. (2009) in their cross-country analysis of redistribution patterns, as well as in the definition of societal indices by Castaneda-Dower et al. (2017) in the context of the Sri Lankan conflict.

In many applications such as the provision of public goods, redistribution, or the intensity of conflicts within a country, one may be led to account for the special status of some regions, in particular a centre c and all other $K-1$ "peripheral" regions.[46] This leads to a special case of the B-index in which only the distances between centre and periphery have to be accounted for:

$$CP = n_c \sum_{k=2}^{K} n_k d_{kc}.$$

[46] See for example Desmet et al. (2009, 2017).

Polarization indices

Both the A- and the B-fractionalization indices rely on a pre-existing partition into distinct ethnolinguistic groups. Polarization and the indices that it generates add self-identification, which comes through in two ways: Strong identification with those in one's own group, and alienation toward others. Esteban and Ray (1994) found a way (called *social effective antagonism*) to combine both identification and alienation, using an axiomatic approach. The functional form of their polarization index P is very close to the one of index B:

$$P = \sum_{k=1}^{K} \sum_{i=1}^{K} n_k^{1+\alpha} n_i d_{ki},$$

where the d_{ki} represent income differences between groups k and i and α is a positive parameter with values between 1 and 1.6.

Reynal-Querol (2002) consider a dichotomous version (with $d_{ki} = 0$ if $k = i$, and $d_{ki} = 1$ if $k \neq i$) of the index, which reads:

$$RQ = \sum_{k=1}^{K} \sum_{i=1}^{K} n_k^2 n_i = \sum_{k=1}^{K} n_k^2 (1 - n_k),$$

where $\alpha = 2$. Using additional axioms, Geng (2012) later showed that the range of α's could be shrunk to a single point $\alpha = 1$ to obtain the Reynal-Querol functional form. It is worth pointing out the intuitive difference between A and RQ. Recall that A is determined by the probability that two randomly chosen individuals belong to two different groups. Thus, the value of A is the sum of the terms $n_k(1 - n_k)$, each identifying the probability that one individual belongs to group k, while the other does not. The value of RQ is determined by the probability that among three randomly chosen individuals, two belong to the same group, while the third belongs to another one. RQ is thus the sum of the terms $n_k^2(1 - n_k)$, in which two individuals belong to group k, while the third does not.

Esteban and Ray (1994) were interested in income polarization, where the distances between groups are naturally represented by income differences. Though one loses the elegance of their index which is derived from axioms, one can think of incorporating linguistic distances instead,[47] assuming that some additional factors, such as linguistic proximity or some historical path could play an important role in determining the degree of identification and alienation. In their study of the Sri Lankan civil war, Castaneda-Dower et al. (2017) introduce an ethnolinguistic polarization measure that allows for the impact on inter-group relations of historical factors driven by different patterns of English language acquisition during the colonial era. They use an RQ distance-weighted index for each Sri-Lankan district j:

$$D^j = \sum_{k=1}^{K} \sum_{i=1}^{K} (n_k^j)^2 n_i^j d_{ki},$$

[47] See also Montalvo and Reynal-Querol (2005) and Desmet et al. (2017).

where n_k^j and n_i^j denote the fraction of linguistic groups k and i in district j, and d_{ki} is the linguistic distance between k and i. Distances take into consideration the changes in English proficiency of the various groups during the precolonial period. They find that larger shares of English speakers resulting from colonial times trigger a larger number of war victims.

Disenfranchisement indices

The notion of linguistic disenfranchisement, formally introduced by Ginsburgh and Weber (2005),[48] provides a rigorous framework to examine linguistic discrimination. In terms of Skutnabb-Kangasand and Phillipson's (1989) formulation, linguistic discrimination and disenfranchisement amount to introducing "ideologies and structures which are used to legitimate, effectuate, and reproduce unequal division of power and resources (both material and non-material) between groups defined on the basis of language." Denying linguistic rights used to be and still is a common problem on all continents.[49] The most severe example is probably Africa since in most countries, the language of the former colonial power replaced the native language as "official language."

To provide a formal response, consider the set L of all language spoken in a country. L may coincide with all K native languages, but may also include some other world languages that are often not spoken as native tongues, like English in Russia, China, Romania or Nigeria. The additional information needed are proficiency data of all or of a sample of individuals in each of the L languages. These can be obtained from censuses or special surveys.

Consider languages (or a unique language) contained in subset $T \in L$ as candidate official language(s), to be used for official documentation, communication, media and educational purposes. Surveys make it possible to define the extent of disenfranchisement of individual m based on which languages (s)he speaks as well as her/his (often self-declared) proficiency in speaking them. Individual levels are then aggregated to evaluate societal disenfranchisement levels.

Let $k(m)$ denote the native tongue and $L(m)$ the set of all other languages spoken by m. There exist four ways to measure m's disenfranchisement:

(a) Index DN: Dichotomous distances and native languages only. m's disenfranchisement is equal to 0 if his native language $k(m)$ is included in T. Otherwise, it is equal to 1. Index DN is given by:

$$DN = \sum_{k:k\notin T} n_k.$$

(b) Index NN: Linguistic distances and native languages only. m's disenfranchisement is again equal to 0 if $k(m) \in T$. Otherwise, it is equal to the minimal distance between $k(m)$ and the languages in set T. Denote this

[48] See also Ginsburgh et al. (2005).

[49] See Ginsburgh and Weber (2011) for a number of cases.

distance by $d^{nat}(m,T)$. Index DN is then:

$$NN = \sum_{1}^{N} d^{nat}(m,T).$$

(c) Index DA: Dichotomous distances and all languages. m's disenfranchisement is equal to 0 if (s)he speaks at least one language in T. It is equal to 1 otherwise. Index DA is then:

$$DA = \sum_{m:L(m)\cap T\neq\emptyset} 1.$$

(d) Index NA: Linguistic distances and all languages. m's disenfranchisement is again equal to 0 if (s)he speaks at least one language in T. Otherwise it is equal to the minimal distance between $L(m)$ and the languages in T. Denote this distance by $d^{all}(m,T)$. Index NA is then equal to:

$$NA = \sum_{1}^{N} d^{all}(m,T).$$

3.3 Using linguistic distances in gravitational models

Distances were first and still are often used to represent frictions in international trade and migration models. More recently, they have been adopted in a larger group of applications, such as translations of literary works or of industrial patents, financial transactions, and money laundering. In one way or another, they make use of Newton's gravitational equation,[50] as suggested more than sixty years ago by Hägerstrand (1957) and Tinbergen (1962).

Hägerstrand (1957) was the first to use the metaphor of the gravity model in his study of migrations in Sweden. Tinbergen (1962) popularized the model by suggesting that it could be applied to study international trade flows between countries. Tinbergen's idea was followed by hundreds of trade papers, both applied and theoretical in order to explicit the assumptions and structural tenets that may lead to various forms of the original gravity equation, bringing it from a partial to a general equilibrium framework in different economic environments (perfect competition, monopolistic competition, heterogeneity of goods, etc).[51]

The simplest form of the equation (where left and right hand-side variables are expressed in logarithms) is:

$$x_{ij} = \alpha + \beta z_i + \gamma z_j + \delta d_{ij}$$

where i and j are countries, regions, languages, ..., x_{ij} is a flow from i to j, z_i and z_j are vectors of variables describing i and j and d_{ij} is a vector

[50] Any two objects i and j exert gravitational attraction on each other with a force, f_{ij}, that is proportional to the product of their masses, m_i and m_j, and inversely proportional to (the square of) the distance d_{ij} that separates the two objects.

[51] See Anderson and Wincoop (2004), Baier et al. (2017), Costinot and Rodriguez-Clare (2014), and Head and Mayer (2014) for recent surveys. See also Hanson and Xiang (2011) for heterogeneous goods.

of frictions, including languages, that may restrain flows between i to j (δ should thus be negative). Finally, α is a scalar, β, γ and δ are vectors of parameters. The gravitational equation became very popular, and usually fits the data very well.

International trade

In the basic trade equation x_{ij} represents the value of exports from country i to country j, the vectors z_i and z_j contain exporter- and importer-specific variables such as their GDPs, and d_{ij} are distances. We will be interested in linguistic distances, but many models also contain dummy variables representing adjacency, trade agreements, colonial ties, etc. between couples of countries, as well as geographic or other types of distances. The first papers introduced a dummy variable for language with value 0 if the language was a common official, common native, or common spoken language in the two countries, and 1 otherwise. Later on, and as better detailed statistics on languages spoken in countries became available, specifications became more subtle.

Egger and Lassman (2012) analyze the effect of 701 coefficients picked up by linguistic distances in 81 articles published between 1970 and 2011 in 24 refereed journals.[52] The average coefficient is equal to -0.49, implying, a lower distance of 10 percent between two countries increases trade by almost five percent. They also run a meta-analysis using the δ's as left hand-side variable which they regress on 18 variables, and find two interesting and robust (across regressions) conclusions: existing colonial ties decrease the (absolute) value of the distance coefficient, while the language effect (again in absolute value) increases over time.[53]

Language is represented by a unique variable in most papers, since they focus on trade rather than on language, and language is a control variable among many others. Melitz and Toubal (2014), on the contrary, use the bilateral trade model with the objective of disentangling the many effects that languages can exercise on trade.[54] They construct four types of bilateral distances between countries: common official language (COL), common native language (CNL), common spoken language (CSL) and linguistic distances (LD). As usual, COL is a binary variable; CNL and CSL both measure the probability that two randomly chosen individuals, one in each country, will be able to communicate; and LD is the (lexicostatistical or cladistic) distance between two languages. The authors suggest that each of them has a specific role in helping citizens from both countries to communicate: CNL and CSL differ since CNL is rather associated with common ethnicity, and thus trust; therefore, if CSL is significantly different from 0 in the presence of CNL, communication acts beyond ethnicity; if, in the presence of CNL and CSL, COL is also significant, then the contribution of official translations also helps

[52]The list of 81 articles can be found in their working paper available at http://hdl.handle.net/10419/54919, accessed on April 20, 2017.

[53]Egger (2008, p. 660) attributes this to the "changes in both prices and varieties".

[54]See also Melitz (2008).

communicating; finally if LD is significant in the presence of the three other measures, translations and interpreters can be used when native languages differ. Their panel consists of over 200,000 observations on trades over ten years (1998-2007); the following right-hand side variables are included: the four linguistic variables discussed above, GDPs, prices, geographic distances, contiguity, (former) colonial ties, common religion, common legal system, history of wars as well as exporter, importer and year fixed effects. Results on the effects of the various measures of languages are summarized in Table 4.[55] As can be checked, in columns (1) to (3) estimated parameters are highly significantly different from 0 when included one by one. They remain significant, though to a lower degree when they are taken together (except in column (4)), which shows that as suggested earlier, each of them has a specific role. It is worth noting that the aggregate effect, equal to -1.11, is much larger (in absolute value) than what is found by Egger and Lassman (2012) as average effect in other studies.

Melitz and Toubal (2014) also study whether particular languages such as English or other international languages (French, Spanish, German and Portuguese) could make a difference. They find that, finally, "all that really matters is a common language, whatever the language may be."

[Insert Table 4 approximately here]

Ku and Zussman (2010) point out that the proficiency of English (which is the only language they examine as possible common language) has changed over time. To measure English proficiency, they construct a data set based on the Test of English as a Foreign Language (TOEFL) results taken by all foreigners who wish to study in an English speaking country. They estimate trade equations that contain a variable measuring linguistic distance between two countries (equal to 0 if they share the same language, and to 1 otherwise) as well as the TOEFL distance of a certain number of languages to English. They show that the effect of English largely overshadows the effect of a common language. Moreover, this effect has the same order of magnitude whether countries share a common language or do not, which leads them to conclude that their results "demonstrate that acquired proficiency in English can assist countries in overcoming historically determined language barriers."

Migration decisions

The standard approach in analyzing the trade-offs of a decision to migrate is based on the difference between benefits and costs. The prospects of higher wages (times the probability of finding a job) or other benefits are contrasted with the monetary and psychological costs, adjustment to a new culture, a new language and a possible uprootedness or even collapse of the family.

Though the specification of the typical immigration equation is very close to the trade equation (where x_{ij} represents migratory flows between source and destination countries i and j divided by the population of the source

[55] Other RHS variables are ignored.

country), one of the most important right-hand side variables consists in existing networks of source country's previous immigrants who create positive externalities on those who migrate later, since they decrease their costs and facilitate their assimilation. Linguistic distances between the language(s) spoken by possible migrants and the country to which they intend to migrate also have an effect.

Many papers have recently studied migrations decisions. The paper by Adserà and Pytliková (2015) is probably the one which is the most recent in analyzing the role of language on migration decisions from 223 source countries to 30 OECD countries during the period 1980 to 2010, but it is by no means the only one.[56] The authors use several measures of linguistic distances (lexicographic, Levenshtein, cladistic, a self-constructed one using a combination of the previous ones) as well as GDPs per capita and unemployment rates in the source and destination country, public expenditure/GDP (as measure of welfare) in the destination country, existing linguistic networks in the destination country, political freedom indices in the source country. They find that migration rates are larger between countries whose first official languages are closer. Identical results are obtained with other measures, but the importance of language is smaller than that of ethnic networks, a result already obtained by Beine et al. (2011) and others who work in the field.[57] The effect of linguistic proximity increases with education.

Falk et al.'s (2012) paper is particularly interesting since it shows that such results also hold within a nation (Germany), where several dialects are spoken in some 440 districts. Their dependent variable consists of current migrations between districts observed during the period 2000-2006 (divided by the population living in the source district). The distances between dialects are based on phonological and grammatical differences that were collected between 1879 and 1888 (thus more that 100 years before the migrations took place) in 45,000 schools where students had to translate 40 German sentences into their local dialect. They find that indeed, "there are intangible cultural borders within a country that impede economic exchange," even if today, all German citizens share a common language, which was not the case in the late 19th century.

Chiswick and Miller (2007) suggested that distances based on learning scores measure could be positively correlated with the difficulty of intercomprehension, and used them as distances between American English and some other languages spoken by immigrants. Isphording (2013) estimates an equation in which the literacy skills of immigrants are influenced by linguistic distances between source and destination countries' languages. Immigrants with a distant language face higher costs to reach the same level of command of the destination language than close languages, especially in the case of immigrants who arrive when they are more than 12 years old.

[56] See Adserà and Pytliková (2016) for many other references.

[57] Bredtman et al. (2017) study the interaction between linguistic distance and networks. They show that the interaction effect is positive, and that the negative effect of distance decreases if networks are larger.

Financial transactions

Grinblatt and Keloharju (2001) deal with data on investment behavior by Finnish investors and companies and show that the language spoken has an effect on individual investment decisions. They had access to the daily trades over two years, as well as to the language spoken by each investor. Finnish is spoken by some 93 percent of the population, Swedish by 6 percent only, but both languages are official, and are quite distant from each other.[58] Swedish speakers are more active in trading and hold 23 percent of household share owner wealth. Among the 97 firms whose stocks are traded, 12 report in Finnish only, two report in Swedish only, and 83 reports are multilingual. Grinblatt and Keloharju also define the culture of the firm according to the name of its CEO: 83 are of Finnish and 14 of Swedish culture. Their results indicate that both language and culture have an effect on the investment behavior of households: Swedish speakers invest more in Swedish firms that publish their annual report in Swedish, and speakers of Finnish rather opt for Finnish firms. Moreover, Swedish speaking households prefer to hold and trade the shares of companies whose CEO is of Swedish origin: A report in Swedish increases the fraction of Swedish shareholders by eight percent, with respect to Finnish shareholders. The effect of a Swedish CEO exerts also a significantly positive though rather small influence. Languages exert therefore a stronger influence than financial returns.

Bellofatto (2017) concludes in a similar way for Belgium, where both French and Dutch are native languages: French (Dutch)-speaking investors invest more in French (Dutch) stocks, overweighting their portfolios in countries that share a common language with Belgium, though this may partly be due to geographical proximities as Dutch-speaking investors are closer to the Netherlands, while those who speak French are closer to France.

Walker and Unger (2009) make use of common languages between countries in estimating a gravity-type model of international money laundering flows. They find that money-laundering is larger between countries which share a common language.

Translations

Translations are sometimes accused of leading to a form of cultural domination by some other languages, in particular by English. These considerations, however, ignore the role of cultural and linguistic distances: In a nutshell, a thriller that features New York is more ikely to be translated from English into French than a Chinese or an Estonian thriller that unfolds in Shanghai or Tallinn.

Ginsburgh et al. (2011) develop a microeconomic model that leads to an equation in which the number of novels translated from a (source) language to a (destination) language is determined by the the sizes of the populations that speak both languages as first language (as a proxy for the number of

[58]The two languages belong to very different branches of the linguistic tree: Swedish is an Indo-European language, Finish is a Uralo-Altaic language.

books written and read, which are unfortunately both unavailable), the distance between the two languages, the literacy rate and the average income of the population speaking the language into which the title is translated. The resulting model (and the direction of the effects) is again similar to the Newton gravity equation, the estimation of which shows that the elasticity of the number of translations with respect to linguistic distance is -1.

Harhoff et al. (2016) study the effect of translations on patenting strategies in Europe. The system is such that once a patent is granted by the European Patent Organization (EPO), the applicant has to validate his patent in each country where he wants protection. This needs translations which happen to be very expensive in some countries, and linguistic distances may, therefore, affect patenting decisions. The authors test this on a dataset which includes all patents granted by the EPO in 2003. The results show that, once approved by EPO, a one percent increase in the linguistic distance between the source language in which the patent was written, and the one in which it has to be translated reduces the probability of validation by almost 16 percent.

3.4 Ethnolinguistic indices and economic outcomes

Fractionalization and polarization

Joseph Greenberg (1956, p. 109) was probably the first linguist to suggest that "developing quantitative measures of diversity in order to render impressions more objective, allow the comparing of disparate geographical areas, and eventually to correlate varying degrees of linguistic diversity with political, economic, geographic, historic, and other non-linguistic factors."

Greenberg's program was picked up a couple of years later. Fishman (1968) and Pool (1972) assert that linguistic diversity has an impact on economic activities, and propose as index the share of speakers of the most widespread language in each country or region. Some 30 years later, Nettle (2000) points out that the Fishman and Pool index does not fully account for the extent of multilingualism and suggest as index the number of languages divided by the entire population which may, however, lead to puzzling conclusions.[59]

The indices discussed in Section 3.2.2 avoid such problems, and are used in equations that link the possible effect of fractionalization or polarization and economic or sociological outcomes y such as growth, redistribution, provision of public goods, corruption and conflicts. The standard equation reads:

$$y = \alpha IND + \sum_k \beta_k z_k + \epsilon,$$

[59]Consider, for example, the impact of the break-up of the Soviet Union on linguistic diversity in Russia. Since the much smaller population of the Russian Federation speaks roughly the same number of languages as those spoken in the former USSR, this results in an increase of Nettle's index, despite the fact that the relative share of the dominant group of Russian speakers in the Federation is much larger than in the USSR.

where IND represents any of the indices described above, $z_k, k = 1, 2, \ldots, m$ are exogenous control variables, α and the β_ks are parameters to be estimated, and ϵ is an error term. The parameter of interest is α, since its sign signals whether fractionalization or polarization are positively or negatively correlated with outcome y. Whether IND "causes" y is debated, though the content of IND (groups and linguistic distances) is historical and usually predates the measurement of y, so that, if there is causality, it can hardly go from y to IND. In almost all cases, the equation is estimated using cross-sectional data, often countries, and the dichotomous A-index. The experiments carried out by Desmet et al. (2009) show that Greenberg's B-index performs better, especially when it is used with diverse forms of aggregating ethnolinguistic groups, which let the data speak.

Growth as outcome. Most papers (essentially written by economists) support the conclusion that linguistic fragmentation has a negative impact on economic development and growth. Easterly and Levine (1997), who coined the "Africa's growth tragedy" expression, highlight the negative impact of diversity on economic growth. They compare African and East Asian 1969 to 1990 growth rates and argue that about 40 percent of the annual growth differential of 3.5 percent between the two regions can be attributed to the effects of fractionalization. These conclusions were reinforced by the cross-country analyses ran by Annett (2001) and Alesina et al. (2003) who point out that fractionalization leads to political instability and excessive government consumption (Annett) or reduces the quality of government (Alesina et al.) that may, in turn, have a negative impact on growth. Collier (2001) and Alesina and La Ferrara (2005) find that the adverse effect of diversity on growth is mitigated in democratic societies, while, according to Lian and O'Neal (1997) it remains strong under dictatorships. Campos et al. (2011) detect no significant correlation of diversity and economic growth in 26 former communist countries of Eastern and Central Europe, Central Asia and Mongolia between 1989 and 2007, that is after the collapse of the Soviet Union.

Other outcomes. Mauro (1995) shows that ethnic and linguistic fractionalization reduces institutional efficiency and increases the level of corruption generated by the lobbying activities of multiple groups. La Porta et al. (1999) find that fractionalization has a negative impact on various public services and goods, including literacy rates, infant mortality, education and infrastructure. According to Alesina, et al. (1999), ethnically fragmented communities run larger deficits and exhibit lower spending shares on basic public goods, including provision of education, roads and sewage systems.[60] Alesina et al. (2016) show that the combination of linguistic diversity and economic inequality are associated to regional underdevelopment.

"Good" cases. Though cross-country results that associate fractionalization and economic outcomes are usually negatively correlated, the conclusions are

[60]See also Kuijs (2000) who confirms these results.

different at city or firm level, where diversity could, on the contrary, be a driving force for progress. The success of Silicon Valley in the late 1990s is often attributed to the background of foreign scientists, engineers and entrepreneurs who flocked to California from India, China, Taiwan, and Israel. Saxenian (1999) points out that over 30 percent of businesses in the Valley have an Asian-born co-founder. Lazear (1999) and Prat (2002) argue that to be successful, teams should show some cultural and linguistic diversity. Florida (2002), and Florida and Gates (2001) show that metropolitan regions with a higher degree of diversity in terms of education, cultural background, sexual orientation and country of origin, correlate positively with a higher level of economic development. Ottaviano and Peri (2005) investigate whether and how linguistic diversity affects wage rates in 160 US cities and find that more diversity is often associated with higher productivity and hourly wages.

Polarization. Hibbs (1973) is the first (political) scientist who made use of an A-index in his study of mass political violence. But later on, Fearon and Laitin (2003) as well as Collier and Hoeffler (2004) give some evidence that the A-index does not "predict" well enough the likelihood of civil conflicts. Montalvo and Reynald-Querol (2002, 2005) argue that their polarization index RQ performs in a better way and that larger values of RQ point to longer civil conflicts. See also Desmet et al. (2009) and Castaneda-Dower et al. (2017) for examples that use distance-weighted polarization indices.

Disenfranchisement

The European Union (EU) is faced with 24 official languages spoken in 28 countries,[61] including English whose status will be difficult to define after Brexit. Though today, it still is the most widely spoken language with 37 percent of Europeans knowing it well or very well, changes will happen after Brexit. Some 60 million British speakers will have left the EU, and German and French will become relatively more important.[62]

Maintaining these 24 languages costs some $1.5 billion per year in translation and interpretation, but it is impossible to slim their number since Regulation no. 1 (adopted by the European Council in 1958) states that the official language of a country joining the EU becomes an official language at the EU level. Since decisions on languages have to be unanimous, no country will accept to see its language disappear from the list, though Luxembourg nodded to this in 1958, at the time the Common Market was created.

Ginsburgh and Moreno-Ternero (2017) suggest that the EU could subsidize[63] each member country to encourage the learning of a common language,[64]

[61] Some languages are official in more than one country. German, for instance, is spoken in Germany and in Austria.

[62] See Ginsburgh et al. (2017).

[63] As will be shown in Section 4, subsidies are often needed to "persuade" citizens to acquire a new language.

[64] They also compute how much each country should receive, using two adjudicating rules (Aristotle's proportional rule and the so-called Talmud rule) to satisfy the claim of

though each country will keep its official language in its own territory, and in the EU. The proposal is meant to facilitate communication among all Europeans, and if possible, strengthen the ailing Union. It is based on the responses to a survey commissioned by the European Commission[65] which shows that the reactions of Europeans are quite positive. Here are the statements and questions that EU citizens had to answer, as well as their reactions:

(a) The European institutions should adopt one single language to communicate with European citizens. *Answer:* In 16 out of 27 countries, 50 percent or more voted in favor of a unique language, with an aggregate EU result of 54 percent.
(b) Everyone in the EU should be able to speak a common language. *Answer:* In all countries, with the exception of Bulgaria, 50 percent or more voted yes, with an aggregate EU result of 69 percent.
(c) Everyone in the EU should be able to speak one language in addition to their mother tongue. *Answer:* In all cases, the idea was adopted by more than 70 percent. Aggregate result: 83 percent.
(d) Which two languages, apart from your mother tongue do you think are the most useful to know for your personal development and career? (any 2 of the 24 official languages could be selected) *Answer:* English (67 percent), French (25 percent), German (22 percent), Spanish (15 percent). These are followed by Russian (3.4 percent), Italian (3.2 percent) and Chinese (1.5 percent). Beyond that, usefulness drops to less than one percent.

These answers show that 83 percent of EU citizens accept the idea of a common language. Their most cited choices are English (67 percent; note that the survey predates the Brexit decision), French (25 percent) and German (22 percent). This, however, would by no means imply that countries lose their identity and home language, and the current translation and interpretation system would remain as it is.

In what follows, we use disenfranchisement indices to check which language should be chosen as *lingua franca* within a reasonable number of years. In this exercise, we make the assumption that the whole population would be as fluent as today's 15 to 29 years old generation, which is much more fluent than their elders.[66] Table 5 contains the results of the calculations for two indices, DN and NN.

The first three columns show DN-indices by country and for the whole EU if English, German or French were chosen as *lingua franca*. As is obvious, English would be the first language to choose. It is unknown by 44.5 percent of the population that would have to be taught the language (but in many countries, especially in Eastern Europe and France, this percentage is much

each country. The claim is assumed to be proportional to the number of young individuals (12-26 age bracket) who do not speak the chosen subsidized *lingua franca*, adjusted for the lexical distance between the official language of the country and the *lingua franca*.

[65]Special Eurobarometer (2006). See also Fidrmuc et al. (2007).

[66]Many more young people of that generation are more fluent than the rest of the population, and we assume that within 20 to 30 years the whole population in each country would be as fluent as the 15-29 years old are today.

larger), and ends up being less expensive to implement than German (73.9 percent) and French (77.8 percent). And this will still be so after Brexit, though some 60 million English speakers disappear: The 44.5 percent of pre-Brexit disenfranchised citizens would increase to 50.6 post-Brexit, but this is still much lower than the disenfranchisement rates generated by German (70.1 percent) and French (74.5 percent).

The last three columns give the values for NN-indices, which are in all cases lower than their DN counterparts, since they are corrected for the (lexicostatistical) distance between the chosen *lingua franca* and the one spoken in each country. But English will no longer be the best language to teach. It will be overtaken by French, which is much closer to Italian and Spanish, the official languages in Italy (58.5 million inhabitants) and Spain (43 million) and will obviously not need extra teaching in France (60.6 million inhabitants).

[Insert Table 5 approximately here]

3.5 Foreign language acquisition

We now turn upside down the questions that were examined earlier, by looking at the determinants that prompt agents to learn foreign languages. We estimate an equation that is based on the reduced form of a game-theoretical model of communicative benefits introduced by Selten and Pool (1991), who show that there exists a Nash equilibrium. Church and King (1993) construct a simplified two-languages two-populations model in which the communicative benefit of an individual increases with the number of those with whom he can communicate using a common language. Due to their assumption that aptitudes to learn are homogeneous across individuals in each country, only one of the two populations learns the other language, which results in corner solutions only. Gabszewicz et al. (2011) introduce heterogeneous aptitudes in the Church and King framework, which lead to the existence of interior Nash equilibria: Both populations learn to some extent the non-native language.

Their equilibrium learning formulation is taken to data by Ginsburgh et al. (2017) who estimate an equation that relates learning decisions of 13 of the most important world languages[67] by citizens who live in some 190 countries. This equation can be written:

$$x_{cj} = \alpha + \beta N_i + \gamma N_j + \delta d_{ij} + \tau T_{cj} + \phi I_c$$

where c is a country whose native language is i and j is the target language of learners. The dependent variable x_{cj} is the share of the population in country c that learns language j.

Five factors are assumed to influence learning of language j by citizens in country c whose native language is i: N_i, the world population of speakers of language i (expected negative effect, since a native language spoken by many should lead to less learning); N_j, the population that speaks j (expected

[67] Chinese, English, Spanish, Arabic, Russian, French, Portuguese, German, Malay, Japanese, Turkish, Italian and Dutch, in descending order of number of speakers.

positive effect, since a language spoken by many attracts more people to learn it); d_{ij} the linguistic distance between languages i and j (with a negative effect). This is very close to the gravity equation discussed earlier. There are two important additional variables: T_{cj}, the ratio of the total trade of country c with the j-speaking world (expected positive effect, since more trade should increase learning) and I_c, the literacy rate in learning country c (expected positive effect). Note that the trade between country c and acquired language j is instrumented to avoid the endogeneity issues, since knowing languages also has an effect on trades.[68]

Table 6 illustrates the main results for both the full sample of over 2,300 observations that include many zeroes when no learning is observed (columns (1) and (2)), and the much smaller sample (240 observations) with positive learning (columns (3) and (4)). Probit (and instrumental variables Probit) is used in the first case and OLS (and TSLS) in the second one. As the first column shows, all five parameters are highly significant and carry the expected signs. Column (2) gives the results after instrumentation of trade. Coefficients drop but remain significant. Based on the estimates, the largest effect by far on learning appears to be trade. Specifically, there is a 13 percent probability that a doubling of trade will result in more learning of the destination language. Columns (3) and (4) deal with the results conditional on positive learning. The world population of speakers of the target language and literacy in the learning country cease to be significant. After correction for endogeneity in column (4), the level of significance of all coefficients remains roughly the same. The coefficient for trade is substantially higher than before in column (4). A ten percentage-point increase in the ratio of trade with native speakers of the destination language would increase learning of the language by 14 percentage points. The negative significant effect of native language on learning is also of some consequence. As expected, linguistic distance always has a significantly negative effect on learning. The surprising effect that can be derived from the model and its estimated coefficients is that learning English is subject to the same principles as learning other languages. Since trade has a quite large effect on learning: growth in, say Chinese/English trade should promote the learning of Chinese in native English countries, as well as the learning of English in China. Which language will gain the battle is, however, difficult to predict. But the demographics of larger birth rates in Arabic and Spanish-speaking populations will prompt more people to learn these two languages, while the Arab and Spanish-speaking populations will learn less foreign languages. This goes in the direction of Graddol's (2006) prediction that Arabic and Spanish will become relatively more important.

[Insert Table 6 approximately here]

[68]The idea for the instrument is borrowed from Frankel and Romer (1999), who faced a similar problem. They needed ratios of trade to GDP that were independent of economic growth; here, T_{cj} values must be independent of language learning. Their solution was to base trade values strictly on variables such as national land area, status as landlocked, common border, geographic distance and population size. An identical procedure was used in this case.

3.6 Returns to language learning

Most papers on returns to language learning are concerned with immigrants who have an incentive to learn the language of the host country if they want to assimilate with locals and find a job,[69] though there are also a couple of studies that look at the issue of nationals, whose objective may be to acquire foreign languages that they use at their workplace.[70]

The main issue here is to estimate the effect of language proficiency on wages, but this leads to the well-known problem that unobserved heterogeneity affects the attempt to estimate returns to languages as it does in the returns to education literature, where assessing the effect of education on earnings, both education and earnings may depend on unobservable individual skills and talent, and the right-hand side variables may be correlated with the error terms in the equation. Instrumental variable techniques are often used to correct for this situation. The 'predicted' linguistic proficiency, obtained in the first stage is then used in the second stage Mincer-type equation which contains educational achievement, tenure and potential experience as right-hand side variables.

Results of the many studies on returns to languages for migrants are summarized by Chiswick and Miller (2014). Returns vary between 5 and 35 percent, depending on datasets, source and destination countries and languages, and on gender. Two papers based on the European Community Household Panel Survey 1994-2000 avoid this problem and collect comparable information. Adserà and Chiswick (2007) study immigrants and Ginsburgh and Prieto-Rodriguez (2007) look at native workers who use foreign languages at their workplace. Still, earnings differ greatly across countries, but in both cases, knowing the language of the immigration country, or speaking and using a foreign language at the workplace have positive returns.

Within the same country, returns for the same language may differ across groups. Indeed, Levinsohn (2006) finds that, in South Africa, returns to speaking English increased between 1993 and 2000, for whites but not for blacks. In Kazakhstan, Kazakh and Russian are both official languages, but returns for Russian are larger for people who *do not* speak Kazakh than for those who speak *only* Russian. The reason seems to be that those who speak Russian and Kazakh signal they attended a Kazakh and not a Russian school with better schooling quality (Aldashef and Danzer, 2014).

[69] Australia (Chiswick and Miller, 1995); Canada (Abbott and Beach, 1992; Aydemir and Skuterud, 2005; Chiswick and Miller, 1995); Germany (Dustmann and Van Soest, 2002); Israel (Beenstock et al., 2001; Berman et al., 2003; Chiswick and Miller, 1995; Chiswick, 1998); the United Kingdom (Leslie and Lindley, 2001); and the United States (Bleakley and Chin, 2004; Bratsberg et al., 2002; Chiswick and Miller, 1995, 2002; Hellerstein and Neumark, 2003).

[70] Canada (Shapiro and Stelcner, 1997), countries of the European Union (Williams, 2006; Ginsburgh and Prieto-Rodriguez, 2007), Hungary (Galasi, 2003), Switzerland (Cattaneo and Winkelman, 2005), and the United States (Fry and Lowell, 2003).

4 The multilingual world: Solutions and policies

The importance of linguistic policies in multilingual societies, countries, unions and international organizations moved to the forefront of public debate in many countries and regions. Recognizing that the multiplicity of languages inevitably requires, explicitly or implicitly, some degree of standardization, is an element of Max Weber's (1968, [1910]) rationalization theory. The requirements of calculability, efficiency, predictability and control over uncertainties leads to bureaucratization and to the creation of a common legal system and, hence, a common language used for administrative purposes. As Jain (2017, p. 475) indicates "using uniform languages can lower the cost of communication, facilitate education and expand economic growth. However, attempts to impose official languages can meet considerable resistance, both because language is an important component of identity, and because learning a new language is difficult."

Linguistic rationalization or standardization can be achieved by different means. One possibility is simply choosing the language of the majority group: French in France, Han Chinese in China, or Kuotsugo Japanese in Japan. Another way is to recognize a *lingua franca*, also a language spoken by the majority of the population, but that is not the mother tongue of any large group in the society. This was the case of Swahili in Tanzania, and more generally in East Africa, Bahasa in Indonesia, or even English in the United States.[71]

Challenges remain important even in the case of two languages, one that is native and the "other." In terms of education, some former colonies opted for native-language instruction in public schools (Morocco, Malaysia, Pakistan, and India) while others kept going with the colonial language (much of sub-Saharan Africa and the Philippines). Native-language instruction may indeed reinforce national identity and make schooling more accessible. But, since top jobs in government and business often continue to use the colonial language, native-language instruction may reduce economic opportunities for the poor.

The economic advantages of standardization, important as they are, represent only part of the equation. The threat of survival and the feeling of disenfranchisement by those who face restrictions of their linguistic rights, have to be taken into account. History, including of the contemporary era, shows how oppression or suppression of languages or cultures may lead to conflicts and even wars. It is shocking and horrifying to think of the cost of linguistic policies in terms of human lives, but the failure to do so ignores the passion and violence generated by the defence of, or the attack on one's own culture and language.[72]

Respecting the "will of the people" is a necessary condition for any sustainable success of long-term policies in a democratic setting. Excluding parts of the population from the process of creation, especially in our in-

[71] See Laitin (1989, 1994) and Ginsburgh and Weber (2011).

[72] See for example Kadochnikov (2016) for the Russian empire or De Votta (2004) on the war in Sri Lanka.

creasingly globalized and competitive environment, simply does not make much economic sense. In addition, it should come as no surprise that it also generates strong emotions. This is made clear by a unique sentence in Fernando Pessoa's *Book of Disquiet*: "Minha pàtria é a lingua portuguesa", my homeland is the Portuguese language. As Section 2.3 shows, being deprived of or persecuted because of one's native language is difficult to accept. There exist many attempts to bring about the benefits of standardization but the feeling of being disenfranchised must be respected. Though some policies, such as the so-called *three language formula* experimented in India, Kazakhstan and Nigeria, or the 24 official language policy in the European Union, respect native languages, their success was and still is, at best, mixed (Laitin, 1994 and Ginsburgh and Weber, 2011).

Many papers and books on linguistic policies have been written. It is probably the largest subject treated by sociolinguists. Spolsky's (2004, 2012) *Language Policy* and *Handbook of Language Policy* give excellent overviews, and many detailed descriptions of most issues, past and present. The examples in Section 4.1 limit the discussion to a few cases for which the consequences of a reform on labor markets were observed and studied by economists. Section 4.2 is specifically devoted to the various *three-language formula* essays in India, Nigeria and Kazakhstan. Section 4.3 aims at showing that massive learning of another language needs subsidies to be successful.

4.1 Examples of simple linguistic policies and their outcomes

Linguistic policies and the changes they bring about have indeed effects on educational systems and on labor markets. Angrist and Lavy (1997) point out that the "promotion of native languages suggest that language is an important symbol of independence and national sovereignty. Governments appear to be willing to tolerate the many social costs of a language transition in return for the perceived social benefits." The effects of such a transition are of interest given the economic value of the human capital embodied in language skills, but, even neglecting emotions, outcomes are unclear. School quality and the economic returns to schooling could either increase or decrease after a change in the language of instruction. They may also increase for some citizens but decrease for others.[73] In most examples that follow, the short-run consequences indicate that language switches have limited influence on the performance of students.

Aspachs-Bracons et al. (2008) study the impact of the change in linguistic education in Catalonia and the Basque Country in 1983, a time at which Catalan and Basque[74] were given the status of official languages in addition to Spanish. Teaching Catalan or Basque to all children became the norm in the two provinces. It turns out that, in Catalonia, the new system had a much

[73]The possibility that individuals with better language skills may earn more for reasons other than those skills is a particular concern in the labor economics literature, and are often difficult to ascertain.

[74]As well as Galician.

deeper effect on the formation of regional identity since the reform was compulsory. In the Basque Country parents could choose the language in which they wanted their children to be educated. To strengthen the point about the impact of identity in Catalonia, Clots-Figueras and Masella (2013) show that those who were exposed for a longer period to Catalan, have stronger Catalan feelings, including those whose parents do not have a Catalan origin.

In 1983, Morocco switched its language of instruction in grade six and above from French to Arabic. Angrist and Lavy (1997) estimate the effect of French language skills on test scores and earnings and suggest that the elimination of compulsory French instruction led to a substantial reduction in the returns to schooling for those affected by the change. This reduction appears to be largely attributable to the loss of French writing skills. Students must fit into the existing economic and social system dominated by the French language and culture, but at the same time, the insistence of instruction in a foreign language constitutes a barrier for poor or rural students who seem to learn more effectively when schooling is given in their native language. Strong native language skills may prove to be of more enduring value in local labor markets than French language skills. In the short-run, however, the language reform contributed to the popular impression that Moroccan schools "are now turning out bilingual illiterates."

A similar situation is studied by Erikkson (2014) who examines the impact of the 1955 Bantu Education Act which required that the local Bantu mother tongue should be taught for six instead of four years, replacing English or Afrikaans in subsidized South African schools. Using the 1980 South African census, she estimates a difference-in-difference model and finds that the two years difference generated positive effects, contrary to what happened in Morocco. The effects could even have been larger in the absence of labor market discrimination against blacks under apartheid.

Latvia, another former Soviet Republic, implemented a similar reform in 2004. Ivlev and King (2014) examine the other side of the coin, namely the educational level in Russian-language schools. They estimate the effect of the switch from Russian (now a minority language) instruction to a composite of 60 percent of Latvian and 40 percent of Russian. Using data from centralised exam results for all Latvian secondary schools between 2002 and 2011, they observe that a significant deterioration of performance took place in minority schools during the early years following the reform.

In Spain and South Africa, bilingual or multi-lingual programmes were instituted to protect ethnolinguistic minorities, while Morocco, Kazakhstan and Latvia share the common feature of introducing the language of the ethnolinguistic majority which aimed to protect itself. South Africa and Kazakhstan subsidized the switch. The comparison seems to imply that introducing minority languages to protect them is almost a necessary condition to make a reform successful, but subsidization may also be needed. Note also that the Bantu Education Act was not imposed in response to popular demand. This distinguishes it from the policy changes in Morocco and Puerto Rico which were the result of national debates.

An interesting example of the shift and its reversal is given by the ex-

perience of the Indian state of West Bengal, where, in 1983, the communist government abolished English teaching at the primary level in public schools. Roy (2005) argues that this lowered academic standards, and despite its avowed objective of making education more accessible, there is no evidence of a positive effect, even on the poorest income quartiles. On the contrary, Kapur and Chakraborty (2008) point out that this change generated a significantly high English skill premium in the labor market: A one percent decrease in the probability of learning English decreased weekly wages by 1.6 percent, which, on average, implied a 68 percent reduction in wages due to the policy change. Ironically, the program increased the wage gap that it was designed to close.

Jain (2017) investigates the impact of official language policies on education using state formation in India. Colonial provinces consisted of some districts where the official language matched the district's language and some where it did not. Linguistically mismatched districts experienced 18.8 percent lower literacy rates and 27.6 percent lower college graduation rates. Educational achievement caught up in mismatched districts after the 1956 reorganization of Indian states on linguistic lines, suggesting that political reorganization can mitigate the impact of mismatched language policies.

Finally, Lleras-Muney and Shertzer (2015)[75] estimate the effect of US statutes requiring English as the language of instruction and compulsory schooling laws on school enrolment, literacy, and English fluency of immigrant children during the 1910-1930 Americanization period. It turns out that the English-only statutes moderately increased literacy rates of certain foreign-born children, particularly of those living in cities or whose parents were not fluent in English. The laws had no impact on immigrants' labor market outcomes or measures of social integration.

And time does not help either, as the last example shows. In 1898, Spain lost the war against the United States. Puerto Rico became American, though not yet an American state, and the language of instruction became English in most post-primary grades in public schools. Education switched back to Spanish fifty years later (in 1948-49). Though American policymakers claimed that English instruction increased English-speaking ability among Puerto Rican natives, Angrist et al. (2008) show that the conclusion was supported by "naïve estimates", and that "more serious estimates" found no effect.

4.2 The three-language formula

One of the remarkable examples of a linguistic policy aimed at balancing efficiency, national pride, fairness and equality of economic opportunity was introduced in India in 1965-66 and adopted by the Parliament in 1968. It followed another recommendation made just after the country's independence in 1948-49 by the University Education Commission, called the *three-language formula* making Hindi the language of the government, while hoping that it would eventually replace English.

[75] See also Lleras-Muney (2002), Goldin and Katz (2011), Clay et al. (2012).

The 1965-66 recommendation, also called the *three-language formula*,[76] was initiated as a national response to bitter complaints against the previous one from Tamil Nadu and other Southern states which claimed that the use of Hindi in government services imposed formidable barriers since they were required to become proficient in two non-native languages, English and Hindi, whereas speakers of Hindi had to learn English only. Therefore, the new formula, that varied across states, required children in Hindi-speaking states to study Hindi, English and one of the Southern languages, whereas children in non-Hindi speaking states were supposed to learn their own regional language, Hindi and English. This masterful and well-crafted formula that seemed to foster group identity, preservation of mother tongues and traditions through the study of regional languages, national pride and unity by acquiring Hindi, and administrative efficiency and standardization by using English, was not successful either. In addition to the inadequate and lukewarm backing by the regional administrations, the formula failed to generate wide public support both in the North and the South. In Hindi regions relatively little effort or resources were spent on studying English and even less so in learning other languages. There was more enthusiasm for English and Tamil in Tamil Nadu, but almost no interest in acquiring Hindi. The lack of public commitment and of resources needed to implement the recommendation caused its failure.

It is worth pointing out that several variants of the *three-language formula* were experimented in Nigeria. "Official regional" languages Hausa, Igbo, and Yoruba, were suggested here, and were considered, like in India, as being a unifying device (Laitin, 1986). The idea was killed by the same challenges that slowed down and often stopped the implementation of the formula in India: No government resources and lack of commitment of students, their families and regional authorities.

Language policy in the Soviet Union that evolved over the years is also worth considering. The Soviet society was viewed as a new type of a super-ethnic unity after the Great Patriotic War. The convergence and fusion of nationalities was officially declared as a step in building the communist society and was adopted in 1962 (Kadochnikov, 2016). The program guaranteed the development of all nationalities and languages and called for the study of Russian and other native languages. In practice, it led to the introduction of a variant of the *three-language formula*: Russian, a regional language, and English for international communication.[77]

An interesting example of the recent implementation of that formula belongs to Kazakhstan, the last Soviet republic to declare its independence from the Soviet Union in 1991, with its two large ethnic groups, Kazakhs and Russians. 74 percent of the 12 million-wide population speak, or at least understand, Kazakh, while 94 percent speak, or at least understand, Russian (Smailov, 2011). Under the Constitution adopted in 1993, Kazakh became the "state" language, whereas Russian, spoken by most Kazakhs,

[76] See Laitin (1989).

[77] Note that English already replaced German as international language after World War II.

was declared an "official" language, used routinely in business, government, and inter-ethnic communication. Using two newly assembled data sets, Aldashev and Danzer (2014) uncover negative returns to speaking Kazakh and a negative wage premium in the first years of Kazakh independence. In spite of the official support of the language at that time, scholastic achievements were substantially lower for pupils taught in Kazakh, a probable consequence of the comparatively poor quality of schools that taught in Kazakh. It is important to point out that changing the official state language without appropriate investments is unlikely to cure the economic disadvantage of a previously marginalized language. However, by a presidential decree in 2011, Kazakhstan introduced a *three-language formula* that, in addition to Kazakh and Russian, includes English as the "world" language. Unlike in India and Nigeria, Kazakhstan committed considerable resources in improving infrastructure, incentives to learn English, developing the culture of language use, raising demand for all three languages in government programs, and preserving linguistic diversity. This required upgrading the low level of education in Kazakh schools during the first years of independence (Aldashev and Danzer, 2014). An important factor that contributed to the success of the formula was the public willingness to buy into a program that strengthened the role of the national Kazakh language, preserved the current and future role of Russian as a vehicle of regional communication and support for social stability in the country. It also highlighted the role of English as an indispensable factor of international communication and business cooperation. The Kazakh level of commitment to the *three-language formula* makes its prospects more promising than in India and elsewhere.

4.3 Welfare issues

Shifts of the existing linguistic policies and the introduction of new ones. raise questions about their stability, efficiency and costs. To examine this question, we use a simplified variant of the two-language model developed in Gabszewicz et al. (2011), already briefly discussed in Section 3.5.[78]

There are two regions, E and F, where agents speak only their native (regional) language, also denoted E and F. The decision by agents to (or not to) acquire the other language is modelled as a non-cooperative normal-form game. All agents have two strategies. They can learn the non-native language by incurring a cost c or remain unilingual. The populations of the regions are N_E and N_F, with $N = N_E + N_F$. We assume that $N_E > N_F$.

Agents in each population differ on the basis of their learning cost described by a parameter $\theta \in [0, 1]$, which is the inverse of their ability to learn a foreign language. Those with small θ are more apt to learn than those with high θ, and, in particular, an agent with $\theta = 0$ can learn the language in her sleep. The cost of learning of the other language in either population is given by $c\theta$, where c is a positive parameter. The distribution of the aptitude parameter is assumed to be uniform in both populations.

[78] For a more complicated three-language model, see Davydov et al. (2017).

The payoff of agent i is given by the utility function u_i which represents her communicative benefit (Selten and Pool, 1991), or the total number of individuals with whom she can communicate with, net of the learning cost (if language learning takes place). If i endowed with aptitude parameter θ learns the other language, she will be able to communicate with the entire population and her utility is $u_i = N - c\theta$. If i in region E does not learn language F, her communication network will consist of N_E and $\alpha_F N_F$, where α_F is the proportion of agents in F who study E. Her utility will be given by $u_i = N_E + \alpha_F N_F$. Similarly, if agent i in F refrains from studying E, she can communate with the total number of $N_F + \alpha_E N_E$ agents where α_E is the fraction of agents in E who study F. Her utility will be given by $u_i = N_F + \alpha_E N_E$.

Gabszewicz et al. (2011) show that both corner and interior solutions can exist. In a corner or pooling equilibrium, either the entire population of the region learns the other language or nobody in that region does.[79] Here we focus on interior or separating equilibria, where there exist cut-off values of θ_E and θ_F in each population that identify agents who are indifferent between studying the other language and refraining from doing so. The agents in E with $\theta < \theta_E$ learn F, while those with $\theta > \theta_E$ do not. Similarly, for population F. It is easy to verify that if $N_E < c$ (and as assumed, $N_E > N_F$), then there exists a stable interior equilibrium, where the fraction of learners in the two populations is given by

$$\alpha_E^* = \frac{cN_F - N_E N_F}{c^2 - N_E N_F}$$

and

$$\alpha_F^* = \frac{cN_E - N_E N_F}{c^2 - N_E N_F}.$$

We now turn to efficient outcomes, where the aggregate welfare of the country is simply the sum of agents' utilities Thus, again denoting by α_E and α_F the fraction of learners in both populations, the total welfare of agents in E is:

$$W_E = (1 - \alpha_E)N_E(N_E + \alpha_F N_F) + \alpha_E N_E(N_E + N_F) - c\int_0^{\alpha_E N_E} \frac{\theta}{N_E} d\theta.$$

The first term is the welfare of the $(1 - \alpha_E)N_E$ agents who do not learn language F; each of them gets a benefit equal to $N_E + \alpha_F N_F$, since they can communicate with that number of E-speakers. The second term describes the benefit of individuals in E who learn F; their number is $\alpha_E N_E$, and the benefit that each of them gets is N; the third terms is the total cost of those who learn. The expression can be rewritten as:

$$W_E = N_E^2 + N_E N_F(\alpha_E + \alpha_F - \alpha_E \alpha_F) - \frac{c}{2}\alpha_E^2 N_E.$$

Similarly,

$$W_F = N_F^2 + N_E N_F(\alpha_E + \alpha_F - \alpha_E \alpha_F) - \frac{c}{2}\alpha_F^2 N_F,$$

[79] See also Church and King (1993).

and the total welfare $W = W_E + W_F$ is:

$$W = N_E^2 + N_F^2 + 2N_E N_F(\alpha_E + \alpha_F - \alpha_E\alpha_F) - \frac{c}{2}(\alpha_E^2 N_E + \alpha_F^2 N_F).$$

Simple algebra shows the first order conditions with respect to α_E and α_F hold for $N_E < \frac{c}{2}$. Efficient shares of learners in each population are given by:

$$\alpha_E^0 = \frac{2cN_F - 4N_E N_F}{c^2 - 4N_E N_F}$$

$$\alpha_F^0 = \frac{2cN_E - 4N_E N_F}{c^2 - 4N_E N_F}.$$

If, in addition, $N_E < \frac{c}{3}$, then (i) $\alpha_E^0 > \alpha_E^*$, $\alpha_F^0 > \alpha_F^*$ and (ii) $\alpha_E^0, \alpha_F^0 < \frac{2}{3}$. Inequalities in (i) imply that there is insufficient learning in both populations, and room for government intervention. Two cases have to be distinguished: (a) Mutual understanding and (b) National cohesiveness.

To achieve mutual understanding, it is sufficient to subsidize the agents in one of the two regions, say E so that all of them will learn language F spoken in the other region. This will cost less than subsidizing all agents in both regions, but may lead to controversies, since language F will be spoken by a majority of the population of the country, and may become dominant.

If the country wants to succeed in achieving national unity and cohesiveness, inequalities in (ii) imply that the government may have to subsidize the entire learning cost of at least one third of of the population in each region. Moreover, it has to do so for agents whose language cost are the highest and who declined the option to learn the other language in equilibrium. The total cost of such a policy is:

$$c(\int_{\alpha_E^0 N_E}^{N_E} \frac{\theta}{N_E} d\theta + \int_{\alpha_F^0 N_F}^{N_F} \frac{\theta}{N_F} d\theta) = \frac{c}{2}[N_E(1 - (\alpha_E^0)^2 + N_F(1 - (\alpha_F^0)^2].$$

Given (ii) and the fact that $c > 3N$, the cost exceeds $\frac{5}{6}N^2$ and shows that the building of national unity could be costly.

5 Conclusions, problems, and further research

It is remarkable that in the very last years of the 20th century, economists became interested in the effects of ethnolinguistic diversity and, knowingly or not, followed a suggestion made by Greenberg (1958) some fifty years earlier. The same happened with the examination of the impact of language on human behavior. Here also, Casasanto (2008), a cognitive scientist who works on languages, expressed the wish that others than linguists should test the Sapir-Whorf vs. Chomsky hypothesis since "otherwise, the only evidence that people who *talk* differently also *think* differently is that they *talk* differently." His wish was fulfilled somewhat later. The idea of using linguistic distances was introduced by Tinbergen (1962): a common language has a positive impact on inter-country trade, but language was, for quite some time, represented by a dummy variable equal to 0 if the couple of

countries had a common official language and to 1 otherwise. In 1992, Dyen et al. (1992) produced the first matrix of distances between all couples of Indo-European languages based on the lexical proximities of a set of 200 words. Later on, several other methods extended such distances to many more languages, and economists started using them in other applications than international trade, including migrations, foreign language learning, and fractionalization indices (again an idea that had already been suggested by Greenberg, 1956). Selten and Pool (1991) were also instrumental in defining language learning equilibria, though they did not derive closed form solutions to their equations that could have been tested by economists.

Language has now become a study field by economists and econometricians, and as every field in its beginnings, there exist problems that have to be overcome. We discuss a couple of those in our concluding comments. They are concerned with the difficulty in defining groups of individuals to construct fractionalization and other indices, the discussion of whether it is language, culture or other factors that *cause* our behavior, why we study other languages, and how we make those choices.

Defining groups

While recognizing the role of linguistic, religious or ethnic differences in studying diversity, one has to tackle the issue of identifying groups and the resulting fractionalization of the society.

It is indeed difficult to identify cultural groups within countries that often result from artificial constructions whose ethnic spreads do not coincide with official borders and may host large numbers of linguistic, ethnic, and religious varieties. As Laitin (2000, p.143) points out "people have multiple ethnic heritages, and they can call upon different elements of those heritages at different times. Similarly, many people throughout the world have complex linguistic repertoires, and can communicate quite effectively across a range of apparently diverse cultural zones."

This has a practical impact on the definition of the concept of diversity and its measurement. For the sake of simplicity, language may be, and is often, used as a proxy for culture and/or ethnicity, and it is therefore not surprising that the first and most influential country-by-country identification widely known as ELF conducted in the Soviet Union some fifty years ago, was based mainly on linguistic and historic origins of various groups.

The partition into linguistic groups is not, however, failsafe. Members of the same group may exhibit very distinct patterns of behavior and even fight each other. A striking example is Tutsis and Hutus, who share almost the same language (Kinyarwanda in Rwanda and Kirundi in Burundi), which did not prevent the horrible conflict between the two populations which still goes on in Burundi. Another issue is how fine the linguistic partition should be. Consider for example Nigeria with its 527 spoken languages, not including the vehicular ones. Hausa is spoken by 27 percent of the country population, Yoruba by 17 percent, Igbo by 15 percent, and Fulfulde by 9 percent.[80]

[80]The data are approximately for 2000. See *Ethnologue* website, consulted on November

Should the fractionalization index account for all 527 groups or should it be limited to five groups (Hausa, Yoruba, Igbo, Fulfulde and others)?

The second question is whether language is the right concept? Fearon (2003, p. 197) illustrates this problem using the United States:

> "What are its ethnic (or racial) groups? Let us make things much easier by restricting attention to groups with at least 1 percent of country population. If we consult official census categories, we get three 'races' – white, African American, and Asian – and an additional group, Hispanic, which the government emphatically declares is 'not a race.' Is this the right list for the United States? Why not disaggregate Hispanic into Puerto Rican Americans, Cuban Americans, Mexican Americans, and so on, or likewise for Asian? Why not distinguish between Arab Americans, Irish Americans, Italian Americans, German Americans, and so on? And why should we use the current census categories, when earlier censuses formulated the categories quite differently."

Alesina et al. (2003, Tables 13, 14 and 15) computed measures based on ethnic, religious as well as linguistic fractionalization for some 190 countries, and find that in some cases they reproduce the same consequences as those discussed in Easterly and Levine (1997) who had used ELF:

Effects of fractionalization on growth. The effect of ethnicity is negative though it fails to be significantly different from 0 at the five percent probability level in the presence of some additional controls such as financial depth, black market premium and number of telephones. More linguistic diversity produces the same negative result as usual. Religious diversity has no significant effect.

Effect of fractionalization on the quality of government policies and institutions. Ethnic diversity has varying (often not significant) effects on most indicators of government quality (business climate, corruption and bureaucratic quality, taxation, size of the public sector, size of government, provision of public goods, schooling and literacy, political rights).

While the three dimensions of fractionalization indices that address ethnic, religious and linguistic partitions are dealt with separately, the challenge would be to examine their combined effect on economic outcomes such as, say, growth and quality of the government. This is precisely discussed in some recent research by Desmet et al. (2012, 2017) and by Davidov and Weber (2016).

The recent and very promising paper by Desmet et al. (2017) explores a fourth type of diversity index based on *cultural* norms, values and attitudes[81] using surveys conducted by the World Values Surveys (2009) taken in 76

23, 2017.

[81] See also Hofstede (1980, 2001) and Schwartz (1992, 2014) for earlier approaches.

countries between 1981 and 2008. They construct *cultural fractionalization indices* $CF = 1 - \sum_{j=1}^{J}(1 - v_j)^2$, where v_j represents the probability that a randomly chosen individual disagrees with answer j. CF gives thus the probability that two randomly drawn individuals from a given population or society give different answers to the set of J questions from their survey. This index is the equivalent for cultural fractionalization of the ethnolinguistic diversity index $A = 1 - \sum_{k=1}^{K} n_k^2$ index, where k is an ethnic group (Section 3.2.2), though it is not based on more "visible" groups such as ethnic, linguistic or religious groups.

They also define a new index χ^2 which depends on the average distance between the observed share of answers to question j by ethnolinguistic group k and the expected share if the distributions of groups and answers were independent. If χ^2 is small, there is little overlap: the cultural answers in each ethnic group are close to the country's average answers. This index thus combines both cultural and ethnic diversities. It would be challenging to construct such indices which could represent more than two types of diversity, since they may all contribute together to economic outcomes.

Using linguistic trees, Desmet et al. (2012, p. 322) also suggest, as already mentioned earlier, to "compute measures of diversity at different levels of linguistic aggregation and let the data inform us which linguistic cleavages are most relevant for a range of political outcomes, rather than making ad hoc choices." And indeed, they show that deeper (that is older) cleavages matter to explain civil conflicts and redistribution, while more recent cleavages are necessary to explain the provision of public goods and economic growth. In some sense, the fractionalization or polarization index is "endogenously" determined by the data, though it is used in the right-hand side of the typical equations that are estimated.

Davydov and Weber (2016) bring to light the axioms needed to construct the following family of indices that includes both Greenberg's A-index and Shannon's entropy index:

$$A^{\alpha} = 1 - \sum_{k=1}^{K} n_k^{\alpha},$$

where α is a positive parameter different from one, and not necessarily equal to the standard value of 2.

Each society i (city, region, or country) is characterized by its own diversity index $A^{\alpha_i} = 1 - \sum_{k=1}^{K} n_k^{\alpha_i}$. Consider, say, Los Angeles and some midwestern town T. The former consists of a large number of various linguistic groups (Mexicans, Koreans, Chinese, Iranians, Armenians and others), while T is relatively homogenous with a small number of recent immigrants arrived from the Middle East or some other distant destinations. If the same diversity index is applied to both localities, Los Angeles' diversity will be much larger than T's. However, if both indices are adjusted to the societal attitude, or to cultural frictions (Fujita and Weber, 2010), comparing diversity levels in Los Angeles and T is much more difficult, and the negative impact of diversity on economic development, found in earlier studies on the topic may become less important and even disappear or become positive (see

Lazear, 1999 and Ottaviano and Peri, 2006). A similar point has been made in the context of the effect of linguistic diversity on civil wars and conflicts. Laitin and Feron (2003) had shown that standard fractionalization indices did not explain the likelihood of civil wars and conflicts. Montalvo and Reynal-Querol (2002, 2005), argued, however, that the incidence, severity and length of civil wars and conflicts could rather be explained by polarization indices, which have another form that incorporates the attitudes of distinct groups towards each other, rather than simply accounting for linguistic diversity of the population.

And since polarization indices proposed by Esteban and Ray (1991) and diversity indices introduced by Davidov and Weber (2016) are characterized by the range of a parameter α, they could both be estimated for each society on the basis of value surveys.

Here we suggest that the choice of an appropriate value of α_i could and, in some cases, should be based on relevant surveys which encompass *cultural* norms, values and attitudes, such as the World Values Survey considered by Desmet et al. (2017), and made "endogenous" as in Desmet et al. (2012).

The last observation underscores an important point. Instead of coming up with a predetermined index to address a specific research question, one should choose the opposite route: The research question that we investigate should determine the index that we use. This obviously requires developing families of indices.

Econometric problems

The empirical econometric literature surveyed in this paper is based on *association*, *coevolution*,' or *correlation* between variables. Scholars usually avoid the term *causality*, though sometimes it unfortunately and probably unwillingly slips into their writings in different forms (such as *explains* which is not far from *causes*) and the not very careful reader would take it for granted that in a regression $y = \alpha + \beta x + \epsilon$, x *causes* y (existence or absence of future tenses in languages *cause* saving behavior, high ethnolinguistic fractionalization in a country *causes* slow economic growth).

Causality. We take the term *causal* in the way it is defined today by econometricians: A causal effect is "the effect on an outcome y of a given action or treatment x as measured in an ideal randomized controlled experiment," since "in such an experiment, the only systematic reason for differences in outcomes between the treatment and the control groups is the treatment itself" (Stock and Watson, 2015, p. 52).[82] Psychologists indeed have run some experiments on comparing the grammars of languages which seem to confirm the analysis carried out by economists: psychologists Sutter et al., 2015 and economist Chen, 2013 find that strong- and weak-future tense have an influence on savings in the broad sense (see Section 2.2). However if strong- and weak-future tense are not exogenous and may, as suggested by Galor et al.

[82]Note that the issue of causality, and sometimes its very existence, is still discussed by contemporary philosophers. See De Pierris and Friedmann (2013) and Schaffer (2016).

(2016), be explained by other variables, the estimated parameters picked up by the right-hand side variable(s) (in this case, tenses) may have no meaning and in the worst case, carry a wrong sign.

But it may obviously be expensive, difficult, and undesirable if not unethical to run experiments on varying the degree of ethnolinguistic fractionalization in the less-developed world to check whether their growth rates would be affected. Fortunately, there obviously exist other econometric methods that, in some cases, and if the data permit, make it possible to 'mimic' a randomized experiment. The real question is, however, whether we need to make causal statements that, as Angrist and Pischke (2009, p. 3) point out are "useful for making predictions about the consequences of changing circumstances or policies; it tells us what would happen in alternative (or 'counterfactual') worlds." There is little that can be done to change languages or the ethnolinguistic groups to make the world any better (see however Section 4 which describes mixed results), and we may satisfy ourselves with observing correlation, and not necessarily causality, though the estimated coefficient β may be strongly biased if x is correlated with the error terms ϵ.

Still, some papers mention directly or indirectly *causality* but also make reservations about it in the same paper. Here are some examples (underlinings are ours):

(a) Easterely and Levine's (1997, p. 1207-1208) title for Section 2 of their landmark paper on Africa's growth tragedy reads "[u]sing cross-country regressions to explain growth," in which the wording 'explain' may resonate as 'causes,' but in the section itself, they write: "this section shows that many indicators [appearing in the right-hand side] have a close association with growth and account for a substantial amount of the cross-country variation in growth rates over the last 30 years," which plays down the term *explain* to *associates*.

(b) Alesina and La Ferrara (2005, p. 772) are also tempted to point to causality by writing "the estimates suggest that ceteris paribus, going from perfect homogeneity to maximum heterogeneity [that is changing the value of the fractionalization index they use from 0 to 1] would reduce a country's growth rate by 2 percentage points per year," and (p. 776) "in any case, neither of these studies argues that ethnic fragmentation is the only cause of 'poor quality of government': La Porta et al. (1999), for instance, argue that legal origins are at least as important."

(c) In their abstract, Galor et al. (2016) write that their "research explores the economic causes and consequences of language structures," but, somewhat later in the paper (p. 15), they mention that "still, the results might be biased due to omitted variables, precluding a causal interpretation of the estimated coefficients."

Reverse causality. Reverse causality can be excluded, since x, the right-hand side variable of interest and y, the left-hand side variable are often measured

at different points in time. This is so for, say, Chen's (2013) analysis of the influence of strong- and weak-future tense on savings behavior. It is also the case in equations in which an ethnolinguistic fractionalization appears in the right-hand side of the equation, and the left-hand side variable is growth or provision of public goods during the second half of the 20th century. It should be clear that the grammar of a language cannot be caused or even influenced by today's saving behavior, and contemporary growth measures cannot have an effect on much older ethnolinguistic divisions in a region or a country. However, there may be reverse causality in Galor et al. (2016) who try to explain the various types of future tense by pre-1500 C.E crop returns. It is very likely that the distinction between languages with or without strong-future tense already existed in the Ur-languages, that is long before 'pre-1500 C.E.' agricultural practices.

Other research issues

The seminal paper by Selten and Pool (1991) on the communicative benefits of language learning has generated some theoretical research, including Church and King (1993), Lazear (1999), Ginsburgh et al. (2011), Gabszewicz et al. (2011), Ginsburgh and Weber (2013) among others, but did not lead to much empirical research other than the one on returns to learning. This was initiated by Chiswick in the early 1970s and has produced a huge body of literature (see Adsera and Pytlikova, 2016 for a survey), without a clear theoretical model of the relation between language learning and labor markets. We are able to infer the wage premium of migrants who learn the language of the host country, as well as the one of those who were born in a country but acquire another language than the one spoken in that country. But we have no microeconomic foundations for making those decisions. There is common work to do by economists and neuroscientists.

We also believe that the issue of linguistic and social identity should attract the attention of economists. Both are critical determinants of our individual or collective behavior, especially at times of social and economic critical situations. There is little doubt that one can observe strong signs of linguistic identity, as is illustrated by the attachment to African American Vernacular English and New York Latino English in the US, or Catalan in Europe, as well as by the societal, economic and even judicial backlash effects of identity choices. One also has to address the yet unexplored issue of measurement of identity, its construction and transmission, including the possibility of multiple identities or the voluntary changes of identity in some cases. "Identity" languages get not only studied, but also revived: Hebrew is of course the case per excellence, but, in the face of our globalizing world, Catalan, Basque, Welsh, Maori in New Zealand, or Inuktitut in Eastern Canada are also. As mentioned earlier, Aspachs-Bracons et al. (2008) stressed the compulsory nature of introducing of Catalan in Catalonia; this generated the dramatic rise of identity that we observe nowadays between the governments of Spain and Catalonia. The voluntary option of studying Basque in the Basque country failed to produce the same result, and

so did the many attempts to reintroduce local languages such as Breton or Provençal in French schools during the late 1960s.

Another interesting research area is to incorporate linguistic diversity into production functions. This was first suggested by Alesina and La Ferrara (2005), Fujita and Weber (2010), Osang and Fujita (2017) and Desmet et al. (2017) who introduced as a variable in societal utility functions, the number of distinct linguistic groups, or the degree of cultural and linguistic distinctions between natives and immigrants. More specifically, in their models, individuals are endowed with a utility function $u(G, c)$ where G and c represent the public and the private good. The production of G is financed by a proportional tax, and its cost is increasing in the level of diversity (for example, the average ethnolinguistic distance between individuals), as more diversity, that is more players with diverse interests, makes more difficult to reaching a decision on taxation. This suggests that it would be useful and promising to introduce various layers of diversity in economic calculus in general.

Last but not least, Esteban et al. (2012) developed a new line of research on the impact of ethnic divisions on conflict. They link conflict intensity within a country to three indices of ethnic distribution: polarization, fractionalization, and linguistic diversity represented by Greenberg's B-index. Their empirical analysis supports the assertion that all three distributional measures are significantly correlated to conflict. They also introduce country-specific measures of group cohesion based on linguistic distances and the importance of public goods, combining them with the distributional measures, and point out that though linguistic distances are exogenous to conflict, they can be expected to drive or at least influence antagonisms across groups. One could take this argument even further and show that the combination of linguistic distances and income inequality could be a deciding factor of the emergence and length of conflicts.

Table 1. Words for Numbers 1 to 5 in 15 Languages

Language	one	two	three	four	five
1. English	one	two	three	four	five
2.	um	dois	três	quatro	cinco
3.	uno	due	tre	quattro	cinque
4.	odin	dva	tri	chetyre	piat'
5.	eins	zwei	drei	vier	fünf
6.	moja	mbili	tatu	nne	tano
7.	een	twee	drie	vier	vijf
8.	unan	daou (m)	tri (m)	pevar (m)	pemp
9.	en	to	tre	fire	fem
10.	jeden	dwa	trzy	czetry	pieć
11.	en	två	tre	fyra	fem
12.	un	deux	trois	quatre	cinq
13.	un	dau (m)	tri (m)	pedwar (m)	pump
14.	uno	dos	tres	cuatro	cinco
15.	egy	kettö	három	négy	ött

(m) is for masculine

Table 2. Grouping Words for Numbers 1 to 5 in Some Indo-European Languages, Hungarian, and Swahili

Language	one	two	three	four	five
Germanic languages					
1. English	one	two	three	four	five
5. German	eins	zwei	drei	vier	fünf
7. Dutch	een	twee	drie	vier	vijf
9. Danish	en	to	tre	fire	fem
11. Swedish	en	två	tre	fyra	fem
Romance languages					
2. Portuguese	um	dois	três	quatro	cinco
3. Italian	uno	due	tre	quattro	cinque
12. French	un	deux	trois	quatre	cinq
14. Spanish	uno	dos	tres	cuatro	cinco
Celtic languages					
8. Breton	unan	daou (m)	tri (m)	pevar (m)	pemp
13. Welsh	un	dau (m)	tri (m)	pedwar (m)	pump
Slavic languages					
4. Russian	odin	dva	tri	chetyre	piat'
10. Polish	jeden	dwa	trzy	czetry	pieć
Uralic languages					
15. Hungarian	egy	kettö	három	négy	ött
Bantu languages					
6. Swahili	moja	mbili	tatu	nne	tano

(m) is for masculine

Table 3. Simplified Indo-European Language Tree

1. Eurasiatic
   2. Uralic-Yukaghirc
      . . . Hungarian
   2. Indo-European
      3. Germanic
      3. Italic
         4. Romance
            5. Italo-Western
               6. Italo-Dalmatian
                  7. Italian
      3. Slavic
         4. East
            5. Belarusan
            5. Russian
            5. Ukrainian
         4. West
            5. Czech-Slovak
               6. Czech
               6. Slovak
            5. Lechitic
               6. Polish
      3. Albanian
      3. Armenian
      3. Baltic
      3. Celtic
      3. Greek
      3. Indo-Iranian

. . .

The upper part of the tree in the first part of the table is based on Greenberg (2000, 279-281). The tree for Indo-European languages is constructed using Ethnologue's website, starting with the root at http://www.ethnologue.com/subgroups/indo-european, and then following the various branches. Details are given for the languages used in the text only (Hungarian, Italian, Russian, Czech, Slovak and Polish, in bold) to illustrate the calculation of distances.

Table 4. Effects of Languages on Bilateral Trade

Variable	(1)	(2)	(3)	(4)	(5)
Common official language (COL)	-0.51			-0.32	-0.35
	(13.5)			(6.8)	(7.6)
Common spoken language (CSL)		-0.77		-0.50	-0.40
		(14.6)		(6.6)	(4.9)
Common native language (CNL)			-0.86	-0.06	-0.28
			(11.2)	(0.57)	(2.3)
Linguistic distance (LD)					-0.08
					(4.2)

Source: Melitz and Toubal (2014, p. 357).
The dependent variable is the log of bilateral trade.
Student *t*s between brackets; 209,276 observations.
The paper looks at two measures of linguistic distances: lexicographic and cladistic.
Results are almost the same. The table reports on the result with lexicographic distances.

Table 5. *DN*- **and** *NN*-**Disenfranchisements by Country in Three Languages** (in % of the total population in each EU country)

	DN-index			*NN*-index		
	English	German	French	English	German	French
Austria	41	1	85	17	0	64
Belgium	39	90	25	19	36	11
Bulgaria	57	87	95	44	67	75
Croatia	34	76	99	26	58	77
Cyprus	18	98	91	15	80	77
Czech R.	64	80	98	48	60	75
Denmark	9	68	98	4	20	75
Estonia	33	85	100	33	85	99
Finland	29	95	97	29	95	97
France	67	95	0	51	72	0
Germany	38	1	88	16	0	67
Greece	40	93	94	33	75	79
Hungary	76	82	99	76	82	99
Ireland	2	94	85	0	40	65
Italy	54	94	85	41	69	17
Latvia	55	96	99	44	76	79
Lituania	49	93	99	39	72	77
Luxembourg	50	8	3	0	0	0
Malta	10	99	92	10	99	92
Netherlands	11	59	88	4	10	67
Poland	57	83	97	43	62	76
Portugal	62	99	87	47	74	25
Romania	69	97	82	53	73	34
Slovak R.	57	66	98	43	49	75
Slovenia	22	72	97	16	52	76
Spain	65	98	92	50	73	24
Sweden	5	89	96	2	27	73
Un. Kingdom	2	95	90	0	40	68
Pre-Brexit EU	44.5	73.9	77.8	31.4	49.2	33.7
Post-Brexit EU	50.5	70.1	74.5	35.5	50.3	30.9

Source: Own calculations based on Fidrmuc et al. (2007).

Table 6. Foreign Language Acquisition

Variable	Full sample		Positive sample	
	Probit	IV Probit	OLS	2SLS
	(1)	(2)	(3)	(4)
Speakers of acquired language	0.014	0.002	0.024	0.006
	(4.35)	(2.70)	(1.84)	(0.29)
Speakers of native language	-0.015	-0.003	-0.024	-0.025
	(3.99)	(4.02)	(4.41)	(4.04)
Trade with acquired language countries	0.46	0.13	0.79	1.40
	(9.24)	(3.19)	(4.69)	(3.04)
Linguistic distance	-0.32	-0.05	-0.35	-0.33
	(6.97)	(5.56)	(2.20)	(2.14)
Literacy in learning country	0.25	0.03	0.06	0.14
	(5.32)	(4.03)	(0.57)	(1.00)
First stage instrumentation of trade		0.59		0.44
		(7.66)		(3.00)
(Pseudo) R-squared	0.234	-	0.236	0.150
No. of countries	193	193	94	94
No. of observations	2,365	2,365	240	240

Source: Ginsburgh et al. (2017).
In columns (1)-(2), the dependent variable is binary (0 if no learning, 1 otherwise);
in columns (3)-(4) it is the share of the population in country i that learns language j.
Student ts between brackets (based on robust standard errors clustered at country level).

References

Abbott, Michael and Charles Beach (1992), Immigrant earnings differentials in Canada: A more general specification of age and experience effects, *Empirical Economics* 17, 221-238.

Adserà, Alicia and Barry Chiswick (2007), Are there gender and country of origin differences in immigrant labor market outcomes across European destinations?, *Journal of Population Economics* 20, 495-526.

Adserà, Alícia and Mariola Pytliková (2015), The role of language in shaping international migration, *The Economic Journal* 125, F49-F81.

Adserà, Alícia and Mariola Pytliková (2016), Language and migration, In Victor Ginsburgh and Shlomo Weber, Eds., *The Palgrave Handbook of Economics and Language*, Houndmills, Basingstoke: Palgrave Macmillan.

Aldashev, Alisher and Alexander M. Danzer (2014), Economic returns to speaking the right languages? Evidence from Kazakhstan's shift in state language and language of instruction, CReAM Discussion Paper Series 1440, Department of Economics, University College London.

Alesina, Alberto, Reza Baqir and William Easterly (1999), Public goods and ethnic divisions, *Quarterly Journal of Economics* 114, 1243-84.

Alesina, Alberto, Arnaud Devleeschauwer, William Easterly, Sergio Kurlat and Romain Wacziarg (2003), Fractionalization, *Journal of Economic Growth* 8(2), 155-94.

Alesina, Alberto and Eliana La Ferrara (2005), Ethnic diversity and economic performance, *Journal of Economic Literature* 43, 762-800.

Alesina, Alberto, Stelios Michalopoulos and Elias Papaioannou (2016), Ethnic inequality, *Journal of Political Economy* 124(2), 428-88.

Alesina, Alberto and Ekaterina Zhuravskaya (2011), Segregation and the Quality of Government in a Cross-section of Countries, *American Economic Review* (2011) 101, 1872-1911.

Anderson, James and Eric van Wincoop (2004), Trade costs, *Journal of Economic Literature* 42, 691-751.

Angrist, Joshua, Aimee Chin and Ricardo Godoy (2008), Is Spanish-only schooling responsible for the Puerto Rican language gap?, *Journal of Development Economics* 85, 105-28.

Angrist, Joshua and Victor Lavy (1997), The effect of a change in language of instruction on the returns to schooling in Morocco, *Journal of Labor Economics* 15, S48-76.

Angrist, Joshua and Jörn-Steffen Pischke (2009), *Mostly Harmless Econometrics. An Empiricist's Companion*, Princeton and Oxford: Princeton University Press.

Annett, Anthony (2001), Social fractionalization, political instability, and the size of the government, *IMF Staff Papers*, 46, 561-92.

Aspachs-Bracons, Oriol, Irma Clots-Figueras, Joan Costa-Font and Paolo Masella (2008), Compulsory language educational policies and identity formation, *Journal of the European Economic Association* 6 , 434-44.

Aston, John, Dorothy Buck, John Coleman, Colin Cotter, Nick Jones, Vincent Macaulay, Norman MacLeod, John Moriarty and Adrew Nevins (2012), Phylogenetic inference for function-valued traits: Speech sound evolution, *Trends in Ecology and Evolution*, 2, 160-6.

Aston, John Jen-Min Chiou and Jonathan Evans (2010), Linguistic pitch analysis using functional principal component mixed effect models, *Journal of the Royal Statistical Society, Series C*, 59, 297-317.

Atlas Narodov Mira (1964), The Miklucho-Maklai Ethnological Institute at the Department of Geodesy and Cartography of the State Geological Committee of the Soviet Union.

Aydemir, Abdurrahman and Mikael Skuterud (2005), Explaining the deterioration entry earnings of Canada's immigrant cohorts, 1996-2000, *Canadian Journal of Economics* 38, 641-72.

Baier, Scott, Amanda Kerr and Yoto Yotov (2017), Gravity, distance and international trade, CESifo Working Paper 6357.

Baker, Mark (2001), *The Atoms of Language*, Oshkosh, WI: Basic Books.

Barrett, Lisa Feldman, Kristen Lindquist, Maria Gendron (2007), Language as context for the perception of emotion, *Trends in Cognitive Sciences* 11, 327-332.

Baugh, John (2009), Econolinguistics in the USA, In Wayne Herbert, Ed., with help from Sally McConnell-Ginet, Amanda Miller and John Whitman, *Language and Poverty*, Bristol, UK: Multilingual Matters, 67-77.

Beenstock, Michael, Barry Chiswick and Gaston Repetto (2001), The effect of linguistic distance and country of origin on immigrant language skills: Application to Israel, *International Migration* 39, 33-60.

Beine, Michel, Frédéric Docquier and Çaglar Özden (2011), Diasporas, *Journal of Development Economics* 95, 30-41.

Bellafatto, Anthony (2017), How does Language Impact Foreign Investing in a Multilingual Country?, Working Paper, Louvain School of Management.

Berman, Eli, Kevin Lang, and Erez Siniver (2003), Language skill complementarity: Returns to immigrant language acquisition, *Labour Economics* 10, 265-90.

Besley, Timothy and Stephen Coate (1992), Understanding welfare stigma: Taxpayer resentment and statistical discrimination, *Journal of Public Economics* 48(2), 165-183.

Bleakley, Hoyt and Aimee Chin (2004), Language skills and earnings: Evidence from childhood immigrants, *The Review of Economics and Statistics* 86, 481-96.

Boas, Franz (1940), *Race, Language and Culture*, Chicago, IL.: University of Chicago Press.

Bond Nigel and Victor Ginsburgh (2016), Language and emotion, In Victor Ginsburgh and Shlomo Weber, Eds., *The Palgrave Handbook of Economics and Language*, Basingstoke, UK: Palgrave-Mac Millan.

Boroditsky, Lera (2001), Does language shape thought? Mandarin and English speakers conceptions of time, *Cognitive Psychology* 43, 1-22.

Boroditsky, Lera (2009), How does our language shape the way we think?, Edge Foundation (https://www.edge.org/conversation/lera_boroditsky-how-does-our-language-shape-the-way-we-think (last consulted June 12, 2018)

Boroditsky, Lera, Orly Fuhrman and Kelly McCormick (2011), Do English and Mandarin speakers think about time differently?, *Cognition* 118, 123-9.

Boroditsky Lera and Alice Gaby (2010) Remembrance of times east: aboriginal Australianrepresentations of time, *Psychological Science* 21, 1635-9.

Boroditsky, Lera, Lauren Schmidt and Webb Phillips (2003), Sex, syntax and semantics, In Dedre Gentner, Susan Goldin-Meadow, Eds., *Language in Mind: Advances in the Study of Language and Cognition*, Cambridge, MA: MIT Press.

Bossert, Walter, Conchita d'Ambrosio and Eliana La Ferrara (2011), A generalized index of fractionalization, *Economica*, 78, 723-50.

Bossuyt, Audrey, Laurence Broze and Victor Ginsburgh (2001), On invisible trade relations between Mesopotamian cities during the third millennium B.C., *The Professional Geographer* 53, 374-83.

Bratsberg, Bernt, James Ragan, and Zafir Nasir (2002), The effect of naturalization on wage growth: A panel study of young male immigrants, *Journal of Labor Economics* 20, 568-97.

Bredtmann, Julia, Klaus Nowotny and Sebastian Otten (2017), Linguistic distance, networks and the regional location decisions of migrants to the EU, Paper presented at the European Economic Association Meeting, 2017.

Bretton, Henry (1976), Political Science, Language, and Politics, In William O'Barr and Jean O'Barr, Eds., *Language and Politics*, The Hague: Mouton.

Campos, Nauro, Ahmed Saleh, and Vitaliy Kuzeyev (2011), Dynamic ethnic fractionalization and economic growth, *Journal of International Trade and Economic Development* 20(2), 129-52.

Canetti, Elias (1977), *Die Gerettete Zunge. Geschichte einer Jugend*, München: Carl Hanser Verlag.

Casasanto, Daniel (2008), Who's afraid of the big bad Whorf? Crosslinguistic differences in temporal language and thought, *Language Learning* 58 (Suppl. 1), 63-79.

Casasanto, Daniel and Lera Boroditsky (2008), Time in the mind: Using space to think about time, *Cognition* 106, 579-93.

Castaneda-Dower, Paul, Victor Ginsburgh and Shlomo Weber (2017), Colonial legacy, polarization and linguistic disenfranchisement: The case of the Sri Lankan war, *Journal of Development Economics* 127, 440-448.

Cattaneo, Alejandra and Rainer Winkelman (2005), Earning differentials between German and French speakers in Switzerland, *Swiss Journal of Economics and Statistics*, 141, 191-212.

Cavalli-Sforza, Luigi Luca (1997), Genes, peoples and languages, *Proceedings of the National Academy of Sciences of the USA*, 94, 7719-7724.

Cavalli-Sforza, Luigi Luca (2000), *Genes, Peoples, and Languages*, Berkeley, CA: University of California Press.

Cavalli-Sforza, Luigi Luca and Fracesco Cavalli-Sforza (1995), *The Great Human Diasporas. The History of Diversity and Evolution*, Cambridge, MA: Perseus Books.

Cavalli-Sforza, Luigi, Luca, Paolo Menozzi and Alberto Piazza (1994), *The History and Geography of Human Genes*, Princeton, NJ: Princeton University Press (abridged edition).

Chen, Keith (2013), The effect of language on economic behavior: Evidence from savings rates, health behaviors, and retirement assets, *American Economic Review* 103, 690-731.

Chiswick, Barry (1998), Hebrew language usage: Determinants and effects on earnings among immigrants in Israel, *Journal of Population Economics* 11, 253-71.

Chiswick, Barry and Paul Miller (1995), The endogeneity between language and earnings: International analyses, *Journal of Labor Economics* 13, 246-88.

Chiswick, Barry and Paul Miller (2002), Immigrant earnings: Language skills, linguistic concentrations and the business cycle, *Journal of Population Economics* 15, 31-57.

Chiswick, Barry and Paul Miller (2007), Linguistic distance. A quantitative measure of the distance between English and other languages, In Barry Chiswick and Paul Miller, Eds., *The Economics of Language, International Analyses*, London and New York: Routledge.

Chiswick, Barry and Paul Miller (2014), International migration and the economics of language, In Barry Chiswick and Paul Miller, Eds., *Handbook of the Economics of International Migration*, vol. 1, Amsterdam: North-Holland, 211-373.

Church, Jeffrey and Ian King (1993), Bilingualism and network externalities,*Canadian Journal of Economics* 26, 337-45.

Clay, Karen, Jeff Lingwall and Melvin Stephens Jr. (2012), Do schooling laws matter? Evidence from the introduction of compulsory attendance Laws in the United States, NBER Working Paper 18477.

Clots-Figueras, Irma and Paolo Masella (2013), Education, language and identity, *The Economic Journal* 123, 332-57.

Collier, Paul (2001), Implications of ethnic diversity, *Economic Policy*, 16, 129-55.

Collier, Paul and Anke Hoeffler (2004), Greed and grievance in civil war, *Oxford Economic Papers* 56, 563-95.

Comrie, Bernard (1983), Book review of Ekkehart Malotki, Hopi Time: A Linguistic Analysis of the Temporal Concepts in the Hopi Language, Australian Journal of Linguistics 4, 131-3.

Costa, Abert, Alice Foucart, Sayuri Hahakama, Melina Aparici, Jose Apesteguia, Joy Heafner and Boaz Keyzar (2014), Your morals depend on language, *PLOS One*, April 24 DOI: 10.1371/journal.pone.0094842.

Costinot, Arnaud and Andrés Rodriguez-Clare (2014) Trade theory with numbers: Quantifying the consequences of globalization, In Gita Gopinath, Elhanan Helpman and Kenneth Rogoff, Eds., *Handbook of International Trade*, vol. 4, Amsterdam: Elsevier.

Crystal, David (1999), *A Dictionary of Languages*, Chicago: University of Chicago Press.

Dahl, Östen (2003), Stuck in the futureless zone, Diversity linguistics comments blog, http://dlc.hypotheses.org/360 (recovered February, 22, 2016).

Dahl, Östen and Viveka Vellupillai (2013), Tense and aspect, In Matthew Dryer and M. Haspelmath, Eds., *The World Atlas of Language Structures Online*, Leipzig: Max Planck Institute for Evolutionary Anthropology.

Dahl, Ovind (1995), When the future comes from behind: Malagasy and other time concepts and some consequences for communication, *International Journal of Intercultural Relations* 19, 197-209.

Davydov, Denis, Aleksander Shapoval and Shlomo Weber (2018), Linguistic equilibrium with local and world languages: challenges of globalization, *The World Economy* 41, 1-22.

Davydov, Denis and Shlomo Weber (2016), A simple characterization of the family of diversity indices, *Economics Letters* 147, 121-123.

Dediu, Dan and D.Robert Ladd (2007), Linguistic tone is related to the population frequency of the adaptive haplogroups of two brain size genes, ASPM and microcephalin, *Proceedings of the National Academy of Sciences* 104, 10944-10949.

de Luca, Erri (1997), *Ora Prima*, Monastero di Bose: Qiqajon.

De Pierris, Graciela and Michael Friedman (2013), Kant and Hume on causality, *Stanford Encyclopedia of Philisophy* (last accessed November 7, 2017 at
https://plato.stanford.edu/entries/kant-hume-causality/

Desmet, Klaus, Ignacio Ortuño-Ortín and Romain Wacziarg (2012), The Political Economy of Ethnolinguistic Cleavages, *Journal of Development Economics* 97, 322-338.

Desmet, Klaus, Ignacio Ortuño-Ortín and Romain Wacziarg (2017), Culture, ethnicity and diversity, *American Economic Review*, 107, 2479-2513.

Desmet, Klaus, Ignacio Ortuño-Ortín and Shlomo Weber (2009), Linguistic diversity and redistribution, *Journal of the European Economic Association*, 7, 1291-1318.

Desmet, Klaus, Ignacio Ortuno-Ortin and Shlomo Weber (2017), Peripheral diversity: Transfers versus public goods, *Social Choice and Welfare*, forthcoming.

DeVotta, Neil (2004), *Blowback: Ethnolinguistic Nationalism, Institutional Decay and Ethnic Conflict in Sri Lanka*, Stanford: CA: Stanford University Press.

Dustmann, Christian and Arthur Van Soest (2002), Language and the earnings of immigrants, *Industrial and Labor Relations Review* 55, 473-492.

Dyen, Isidore, Joseph Kruskal, and Paul Black (1992), An Indo-European classification: A lexicostatistical experiment, *Transactions of the American Philosophical Society*, 82/5 (Philadelphia: American Philosophical Society).

Easterly, William and Ross Levine (1997), Africa' growth tragedy: Policies and ethnic divisions, *The Quarterly Journal of Economics* 112(4), 1203-1250.

Egger, Peter (2008), On the role of distance for bilateral trade, *The World Economy* 32, 653-662.

Egger, Peter and Andrea Lassmann (2012), The language effect in international trade: A meta-analysis, *Economic Letters* 116, 221-224.

Erikkson, Katherine (2014), Does the language of instruction in primary school affect later labour market outcomes? Evidence from South Africa, *Economic History of Developing Regions* 29, 331-335.

Esteban, Joan, Laura Mayoral, and Debraj Ray (2012), Ethnicity and conflict: An empirical study, *American Economic Review* 102, 1310-42.

Esteban, Joan and Debraj Ray (1994), On the measurement of polarization, *Econometrica*, 62, 819-851.

Ethnologue (2016), *Languages of the World*, 19th edition, Dallas, TX: SIL International Publications.

Evans, Nicholas and Stephen Levinson (2009), The myth of language universal: Language diversity and its importance for cognitive sciences, *Behavioral and Brain Sciences* 32, 429-492.

Fabb, Nigel (2016), Linguistic theory, linguistic diversity and Whorfian economics, In Victor Ginsburgh and Shlomo Weber, Eds., *The Palgrave Handbook of Economics and Language*, Houndmills, Basingstoke, UK: Palgrave MacMillan.

Falk, Olivier, Stephan Heblich, Alfred Lamelli and Jens Südekum (2012), Dialects, cultural identity, and common exchange, *Journal of Urban Economics* 72, 225-239.

Fearon, James (2003), Ethnic and cultural diversity by country, *Journal of Economic Growth* 8, 195-222.

Fearon, James and David Laitin (1999), Weak states, rough terrain, and large ethnic violence since 1945, Paper presented at the annual meetings of the American Political Science Association, Atlanta, GA.

Fearon, James and David Laitin (2003), Ethnicity, insurgency, and civil war, *American Political Science Review* 97(1), 75-90.

Fidrmuc, Jan, Victor Ginsburgh and Shlomo Weber (2007), Ever closer Union or Babylonian discord? The official-language problem in the European Union, CEPR Discussion Paper 6367.

Fischman, Joshua (1968), *Readings in the Sociology of Language*, The Hague, Paris: Mouton.

Florida, Richard (2002), *The Rise of the Creative Class: And How It's Transforming Work, Leisure, Community, and Everyday Life*, New York: Perseus Book Group.

Florida, Richard and Gary Gates (2001) Technology and tolerance: The importance of diversity to high-tech growth, Brookings Institute Discussion Paper.

Frankel, Jeffrey and David Romer (1999), Does trade cause growth?, *American Economic Review* 89, 379-99.

Fry, Richard and Lindsay Lowell (2003), The value of bilingualism in the U.S. labor market, *Industrial and Labor Relations Review* 57, 128-40.

Fujita, Masahisa and Shlomo Weber (2010), Immigration quotas in the globalized world, *Journal of the New Economic Association* 7, 10-23.

Gabszewicz, Jean, Victor Ginsburgh, and Shlomo Weber (2011), Bilingualism and communicative benefits, *Annals of Economics and Statistics* 101/102, 271-86.

Gaby, Alice (2012), The Thaayorre think of time like they talk of space, *Frontiers in Psychology* 3, article 300.

Galasi, Peter (2003), Estimating wage equations for Hungarian higher education graduates, Budapest Working Papers on the labor market.

Galor, Oded and Ömer Özak (2016), The agricultural origins of time preference, *American Economic Review*, 106, 3064-3103.

Galor, Oded, Ömer Özak and Assaf Sarid (2016), Geographical origins and economic consequences of language structures, CESIFO Working Paper 6149, October.

Gamkrelidze, Thomas and Vjacheslav Ivanov (1990), The early history of Indo-European languages, *Scientific American*, March, 82-89.

Geng, Difei (2012) Identifying the unique polarization index: A mean-preserving axiomatic approach, *Journal of Public Economic Theory* 14, 791-812.

Gibson, Edward, Richard Futrella, Julian Jara-Ettingera, Kyle Mahowalda, Leon Bergena, Sivalogeswaran Ratnasingamb, Mitchell Gibsona, Steven T. Piantadosic, and Bevil R. Conwayb (2017), Color naming across languages reflects color use, *Proccedings of the National Academy of Sciences* 104, 10785-90.

Gijssels, Tom and Daniel Casasanto (2017), Conceptualizing time in terms of space: Experimental Evidence, In B. Dancygier, Ed., *Cambridge Handbook of Cognitive Linguistics*, Cambridge: Cambridge University Press.

Gini, Corrado (1912), Variabilità e mutabilità, in *Studi Economico-Giuridici della R. Universitadi Cagliari* 3, 3-1590. Reprinted in Enrico Pizetti and Tomasso Salvemini, Eds., *Memorie di metodologica statistica*, Roma: Libraria Eredi Virilio Vechi (1955).

Ginsburgh, Victor, Jacques Melitz and Farid Toubal (2017), Foreign language learning and trade, *Review of International Economics* 25, 320-61.

Ginsburgh, Victor and Juan Moreno-Ternero (2018), Compensation schemes for learning a *lingua franca* in the European Union, *The World Economy* DOI: 10.1111/twec.126 44.

Ginsburgh, Victor, Juan Moreno-Ternero and Shlomo Weber (2017), Ranking languages in the European Union: Before and after Brexit, *European Economic Review* 93, 139-51.

Ginsburgh, Victor, Ignacio Ortuno-Ortin and Shlomo Weber (2005), Disenfranchisement in linguistically diverse societies. The case of the European Union, *Journal of the European Economic Association* 3 (2005), 946-64.

Ginsburgh, Victor and Juan Prieto-Rodriguez (2007), Returns to foreign languages of native worker in the EU, *Industrial and Labor Relations Review* 64, 599-618.

Ginsburgh, Victor and Shlomo Weber (2005), Language disenfranchisement in the European Union, *Journal of Common Market Studies* 43, 273-86.

Ginsburgh, Victor and Shlomo Weber (2011), *How Many Languages Do We Need? The Economics of Linguistic Diversity*, Princeton, NJ: Princeton University Press.

Ginsburgh, Victor, Shlomo Weber and Sheila Weyers (2011), The economics of literary translation. A simple theory and evidence, *Poetics* 39, 228-46.

Goebl, Hans (1982), *Principien und Methoden des Einsatzes der Numerischen Taxonomie im Bereich der Dialektgeographie*, Wien: Austrian Academy of Sciences.

Goldin, Claudia and Lawrence Katz (2011), Putting the "co" in education: Timing, reasons, and consequences of college coeducation from 1835 to the present, *Journal of Human Capital* 5, 377-417.

Graddol, David (2006), *English Next*, London: British Council.

Gray, Russel and Quentin Atkinson (2003), Language-tree divergence times support the Anatolian theory of Indo-European origin, *Nature* 426, 35-439.

Greenberg, Joseph (1955), *Studies in African Linguistic Classification*, New Haven, CN: Compass Publishing Company.

Greenberg, Joseph (1956), The measurement of linguistic diversity, *Language* 32, 109-15.

Greenberg, Joseph (1963), *The Languages of Africa*, Bloomington: Indiana University Press.

Greenberg, Joseph (1987), *Language in the Americas*, Stanford, CA: Stanford University Press.

Greenberg, Joseph (2000), *Indo-European and Its Closest Relatives: The Eurasiatic Language Family. 1: Grammar*, Stanford, CA: Stanford University Press.

Greenberg, Joseph (2002), *Indo-European and Its Closest Relatives: The Eurasiatic Language Family. 2: Lexicon*, Stanford, CA: Stanford University Press.

Grenier (1984), The effects of language characteristics on the wages of Hispanic American males, *Journal of Human Resources* 19, 35-52.

Grinblatt, Mark, and Matti Keloharju (2001), How distance, language and culture influences stockholdings and trade, *The Journal of Finance* 56, 1053-73.

Guiso, Luigi, Paola Sapienza and Luigi Zingales (2006), Does culture affect economic outcomes, *Journal of Economic Perspectives* 20, 23-48.

Gumperz, John and Stephen Levinson (1991), Rethinking linguistic relativity, *Current Anthropology* 32, 613-623.

Hadjipantelis, Pantelis, John Aston and Jonathan Evans (2012) Characterizing fundamental frequency in Mandarin: A functional principal component approach utilizing mixed effect models, *Journal of the Acoustical Society of America* 13, 4651-4664.

Hägerstrand, Torsten (1957), Migration and area, In David Hannerberg, Torsten Häger-strand and Bruno Odeving, Eds., *Migration in Sweden*, *Lund Studies in Geography* 13, 27-158.

Hanson, Gordon and Xiang, Chong (2011), Trade barriers and trade flows with product heterogeneity: An application to US motion picture exports, *Journal of International Economics* 83, 14-26.

Harhoff, Dietmar, Karin Hoisl, Bruno van Pottelsberghe de la Potterie and Charlotte Vandeput (2016), Languages, fees and the international scope of patenting, In Victor Ginsburgh and Shlomo Weber, Eds., *The Palgrave Handbook of Economics and Language*, Houndmills, Basingstoke: Palgrave Macmillan.

Harris, Randy (1993), *The Linguistic Wars*, Oxford: Oxford University Press.

Hart-Gonzalez, Lucinda and Stephanie Lindemann (1993) Expected achievement in speaking proficiency, Foreign Service Institute, Department of State: School of Language Studies.

Head, Keith and Thierry Mayer (2014), Gravity equations: Workhorse, toolkit and cookbook, In Gita Gopinath, Elhanan Helpman and Kenneth Rogoff, Eds., *Handbook of International Trade*, vol. 4, Amsterdam: Elsevier.

Heblich, Stephan, Alfred Lameli and Gerhard Riener (2015), The effect of perceived regional accents on individual economic behaviour: A lab experiment on linguistic performance, cognitive ratings and economic decisions, *PloS ONE* February 11, do: 10.1371/journal.pone.0113475.

Heggarty, Paul, April McMahon and Robert McMahon (2005), From phonetic similarity to dialect classification: A principled approach, In Nicole Delbecque, Dirk Geeraerts and Johan van der Auwera Eds., *Perspectives in Variation: Sociolinguistic, Historical, Comparative*, Amsterdam: Mouton de Gruyter.

Hellerstein, Judith and David Neumark (2003), Ethnicity, language, and workplace segregation: Evidence from a new matched employer-employee data set, *Annales d'Economie et de Statistique* 71/72, 19-78.

Hershkovitz, Israel, Patricia Smith, Rachel Sarig, Rolf Quam, Laura Rodriguez, Rebeca Garcla, Juan Luis Arsuaga, Ran Barkai, and Avi Gopher (2010), Middle Pleistocene dental remains from Qesem cave (Israel), *American Journal of Physical Anthropology* 144, 575-92.

Hershkovitz, Israel, Gerhard Weber, Cinzia Fornai, Avi Gopher, Ran Barkai, Viviane Slon, Rolf Quamf, Yankel Gabet, Rachel Sarig (2015), New Middle Pleistocene dental remains from Qesem cave (Israel), *Quaternary International*,
http://dx.doi.org/10.1016/j.quaint.2015.08.059.

Hibbs, Douglas (1973), *Mass Political Violence: A Cross-National Causal Analysis*, New York, NY: Wiley.

Hill, Mark (1973) Diversity and evenness: A unifying notation and its consequences, *Ecology*, 54, 427-32.

Hofstede, Geert (1980), *Culture's Consequences: International Differences in Work-related Values*, Bevery-Hills, CA: Sage.

Hosfstede, Geert (2001), *Culture's Consequences: Comparing Values, Behaviors, Institutions, and Organizations Across Nations*, B every Hills, CA: Sage.

Hočevar, Toussaint (1975), Equilibria in linguistic minority markets, *Kyklos* 28, 337-357.

Holden Nigel (2016), Economic exchange and business language in the ancient world: An exploratory review, In Victor Ginsburgh and Shlomo Weber, Eds., *The Palgrave Handbook of Economics and Language*, Houndmills, Basingstoke, UK: Palgrave MacMillan.

Humboldt, Wilhelm von (1988, [1836]), *The Diversity of Human Language-Structure and its Influence on the Mental Development of Mankind*, Cambridge, UK: Cambridge University Press.

Isphording, Ingo (2013), Disadvantage of linguistic origin: Evidence fro immigrant literacy scores, IZA Discussion Paper 7360.

Ivlevs, Artjoms and Roswitha King (2014), Minority education reform and pupil performance in Latvia, *Economics of Education Review* 38, 151-66.

Jain, Tarun (2017), Common Tongue: The Impact of Language on Educational Outcomes, *Journal of Economic History* 77, 473-510.

Janson, Tore (2002), *Speak: A Short Story of Languages*, Oxford: Oxford University Press.

Kadochnikov, Denis (2016), Languages, regional conflicts and economic development: Russia, In Victor Ginsburgh and Shlomo Weber, Eds., *The Palgrave Handbook of Economics and Language*, Houndmills, Basingstoke: Palgrave Macmillan.

Kapur Shilpi and Tanika Chakraborty (2008), English Language Premium: Evidence from a policy experiment in India, Washington University in St. Louis Discussion Paper.

Kessler, Brett (1995), Computational dialectology in Irish Gaelic, In *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics*, Dublin: European Chapter of the Association for Computational Linguistics.

Kessler, Brett (2001), *The Significance of Word Lists*, Stanford, CA: Center for the Study of Language and Information.

Kiju, Jung, Sharon Shavit, Madhu Viswanathan, and Joseph Hilbe (2014), Female hurricanes are deadlier than male hurricanes, *Proceedings of the National Academy of Sciences of the United States of America* 111, 8782-7.

Kramsch, Claire (1998), *Language and Culture*, Oxford: Oxford University Press.

Ku, Hyejin and Asaf Zussman (2010), Lingua franca: The role of English in international trade, *Journal of Economic Behavior & Organization* 75, 250-60.

Kuijs, Louis (2000), The impact of ethnic heterogeneity on the quantity and quality of public spending, IMF Working Paper WP0049.

Laitin, David (1986), *Hegemony and Culture: The Politics of Religious Change Among the Yoruba*, Chicago, IL, University of Chicago Press.

Laitin, David (1989), Language policy and political strategy in India, *Policy Studies* 22, 415-36.

Laitin, David (1994), The Tower of Babel as a coordination game: Political linguistics in Ghana, *American Political Science Review* 88, 622-34.

Laitin, David (2000), What is a language community?, *American Journal of Political Science* 44, 142-55.

La Porta, Rafael, Florencio Lopez de Silanes, Andrei Shleifer and Robert Vishny (1999), The quality of government, *Journal of Law, Economics and Organization* 15, 222-279.

Lazear, Edward (1999), Culture and language, *Journal of Political Economy* 107, S95-S127.

Leavitt, John (2011), *Linguistic Relativities. Language Diversity and Modern Thought*, Cambridge, UK: Cambridge University Press.

Leslie, Derek and Joanne Lindley (2001), The impact of language ability on employment and earnings of Britain's ethnic communities, *Economica* 68, 587-606.

Levenshtein, Vladimir (1966), Binary codes capable of correcting deletions, insertions, and reversals, *Cybernetics and Control Theory* 10, 707-10.

Levinsohn, James (2006), Globalization and the returns to speaking English in South Africa, In Ann Harrison, Ed., *Globalization and Poverty*, Chicago, IL: Chicago University Press.

Lewis, Bernard (2004), *From Babel to Dragomans: Interpreting the Middle East*, Oxford: Oxford University Press.

Lewis, Bernard (2009), *Islam: The Religion and the People*, Upper Saddle River, NJ: Prentice HallWharton School Publishing.

Li, Peggy and Lila Gleitman (2002), Turning the tables: language and spatial reasoning, *Cognition* 83, 265-94.

Lian, Brad and John O'Neal (1997), Cultural diversity and economic development: A cross-national study of 98 countries, 1960-1985, *Economic Development and Cultural Change* 46, 61-77.

Lleras-Muney, Adriana (2002), Were compulsory education and child labor laws effective? An analysis from 1915 to 1939, *Journal of Law and Economics* 45, 401-35.

Lleras-Muney, Adriana and Alison Shertzer (2015), Did the Americanization movement succeed? An evaluation of the effect of English only and compulsory school laws on immigrants' education, *The American Economic Journal: Economic Policy* 7, 238-57.

Malotki, Ekkehart (1983), *Hopi Time: A Linguistic Analysis of the Temporal Concepts in the Hopi Language*, Berlin, New York, Amsterdam: Mouton.

Manea, Norman (2012), The exiled language, In Norman Manea, *The Fifth Impossibility. Essays on Exile and Language*, New Haven: Yale University Press.

Marschak, Jacob (1965), Economics of language, *Behavioral Science* 10(2), 135-40.

Markowsky, Eva (2017), Speaking and gender: Does language affect labor market outcomes, Paper presented at the Conference on Language Skills for Economic and Social Inclusion, Berlin, October 2017.

Mauro, Paolo (1995), Corruption and growth, *The Quarterly Journal of Economics*, 110, 681-712.

Mavisakalyan, Astghik (2015), Gender in language and gender in employment, *Oxford Development Studies* 43, 403-24.

McMahon, April and Robert McMahon (2005), *Language Classification by Numbers*, Oxford: Oxford University Press.

McManus, Walter, William Gould and Finis Welch (1983), Earnings of Hispanic Men: The Role of English Language Proficiency, *Journal of Labor Economics* 1, 101-30.

Melitz, Jacques (2008), Language and foreign trade, *European Economic Review* 52, 667-99.

Melitz, Jacques and Farid Toubal (2014), Native language, spoken language, translation and trade, *Journal of International Economics* 93, 351-363.

Michalopoulos, Stelios (2012), The origins of linguistic diversity, *American Economic Review*, 102, 1508-39.

Montalvo, Jose and Marta Reynal-Querol (2002), Why ethnic fractionalization? Polarization, ethnic conflict and growth, UPF Working Paper 660, University Pompeu Fabra, Barcelona.

Montalvo, Jose and Marta Reynal-Querol (2005), Ethnic polarization, potential conflict and civil war, *American Economic Review* 95, 796-816.

Moore, Karl and David Lewis (1999), *Birth of the Multinational: 2000 Years of Ancient Business History from Ashur to Augustus*, Copenhagen: Copenhagen Business School Press.

Nakhleh, Luay, Tandy Warnow, Don Ringe and Steven Evans (2005), A comparison of phylogenetic reconstruction methods of an Indo-European dataset, *Transactions of the Philological Society* 103, 171-92.

Nerbonne, John and Wibert Heeringa (1997), Measuring dialect difference phonetically, In John Coleman, Ed., *Workshop on Computational Phonology*, Madrid: Special Interest Group of the Association for Computational Linguistics.

Nettle, Daniel (2000), Linguistic fragmentation and the Wealth of Nations: The Fishman-Pool hypothesis reexamined, *Economic Development and Cultural Change* 48(2), 335-48.

Ngugi wa Thiong'o (1986), *Decolonizing the Mind: The Politics of Language in African Literature*, Oxford: James Currey.

Osang, Thomas and Shlomo Weber (2017), Immigration policies, labor complementarities, population size and cultural frictions: Theory and evidence,*International Journal of Economic Theory* 13, 95-111.

Ottaviano, Gianmarco and Giovanni Peri (2005), Cities and cultures, *Journal of Urban Economics*, 58, 304-37.

Ottaviano, Gianmarco and Giovanni Peri (2006), The economic value of cultural diversity: Evidence from US cities, *Journal of Economic Geography*, 6, 9-44.

Phillipson, Robert and Tove Skutnabb-Kangas (1995), Language rights in postcolonial Africa, In Robert Phillipson, Mart Rannut and Tove Skutnabb-Kangas, Eds., *Linguistic Human Rights: Overcoming Linguistic Discrimination*, Berlin and New York: Mouton De Gruyter.

Pinker, Steven (1994), *The Language Instinct*, New York, NY: Harper Perennial Modern Classics.

Poloni, Estelle Omella Semino, Giuseppe Passarino, A. Silvana Santachiara-Benerecetti, Isabelle Dupanloup, André Langaney and Laurent Excoffier (1997), Human genetic affinities for Y-chromosome P49a,f: TaqI haplotypes show strong correspondences with linguistics, *American Journal of Human Genetics* 61, 1015-35.

Pool, Jonathan (1972), National development and language diversity, In Joshua Fishman, Ed., *Advances in the Sociology of Language*, The Hague: Mouton.

Prat, Andrea (2002), Should a team be homogeneous?, *European Economic Review* 46(7), 1187-1207.

Ramírez-Esparza, Nairán, Samuel Gosling, Veónica Benet-Martinez, Jeffrey Potter and James Pennebaker (2006), Do bilinguals have two personalities? A special case of cultural frame switching, *Journal of Research in Personality* 40, 99-120.

Renfrew, Colin (1987) *Archeology and Language*, London: Jonathan Cape.

Reynal-Querol, Marta (2002), Ethnicity, political systems, and civil wars, *Journal of Conflict Resolution* 46, 29-54.

Roth, Alvin, Vesan Prasnikar, Masahiro Okuno-Fujiwara and Shmuel Zamir (1991), Bargaining and market behaviour in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An experimental study, *American Economic Review* 81, 1068-1095.

Roy, Joydeep (2005), Redistributing Educational Attainment: Evidence from an Unusual Policy Experiment in India, Georgetown University Discussion Paper.

Ruhlen, Merritt (1994), *The Origin of Language*, New York: John Wiley and Sons.

Rustow, Dankwart (1967), *A World of Nations: Problems of Political Modernization*, Washington, DC: Brookings Institution.

Sapir, Edward (1949), *Selected Writings of Edward Sapir in Language, Culture, and Personality*, Berkeley, CA: University of California Press.

Saxenian, AnnaLee (1999) *Silicon Valley's New Immigrant Entrepreneurs*, San Francisco, CA: Public Policy Institute of California.

Schaffer, Jonathan (2016), The metaphysics of causation, *Stanford Encyclopedia of Philosophy* https://plato.stanford.edu/entries/causation-metaphysics/ (last accessed on November 7, 2017).

Schwartz, Shalom (1992), Universals in the content and structure of values: Theory and empirical tests in 20 countries, *Advances in Experimental Social Psychology* 25, 1-65.

Schwartz, Shalom (2014), National culture as value orientations: Consequences of value differences and cultural distance, In Victor Ginsburgh and David Throsby Eds., *Handbook of the Economics of Arts and Culture*, Vol. 2, Amsterdam: Elsevier.

Selten, Reinhard and Jonathan Pool (1991), The distribution of foreign language skills as a game equilibrium, In Reinhard Selten, Ed., *Game Equilibrium Models*, vol. 4, Berlin, Springer-Verlag, 64-84.

Senghor, Léopold Sédar (1956), *Ethiopiques, Oeuvre Poétique*, Paris: Editions du Seuil.

Senghor, Léopold Sédar (1962), Le français, langue de culture, *Esprit*, Novembre, 837-844.

Shannon Claude (1948), A mathematical theory of communication, *Bell Systems Technical Journal* 27, 379-423, 623-56.

Shapiro, Daniel and Morton Stelcner (1997), Language earnings in Quebec: Trends over twenty years, 1970-1990, *Canadian Public Policy* 23, 115-140.

Simovici, Dan and Szymon Jaroszewicz (2002), An axiomatization of partition entropy, *IEEE Transactions on Information Theory* 48, 21382142.

Simpson, Edward (1949), Measurement and diversity, *Nature* 163, 688.

Skutnabb-Kangas, Tove and Phillipson, Robert (1989), Mother tongue: The theoretical and sociopolitical construction of a concept, In Ulrich Ammon, Ed., *Status and Function of Languages and Language Varieties*, Berlin and New York: de Gruyter, 450-477.

Smailov, Arman (2011), Results of the 2009 National Population Census of the Republic of Kazakhstan. Analytical report, Astana, Kazakhstan: The Agency on Statistics of the Republic of Kazakhstan.

Special Eurobarometer 243 (2006), Europeans and their Languages.

Spolsky, Bernard (2004), *Language Policy*, Cambridge: Cambridge University Press.

Spolsky, Bernard, Ed. (2012), *The Cambridge Handbook of Language Policy*, Cambrige, UK: Cambridge University Press.

Stock, James and Mark Watson (2015), *Introduction to Econometrics*, 3d edition, Harlow, England: Pearson.

Sutter, Matthias, Silvia Angerer, Daniela Glätzle-Rützler and Philipp Lergetporer (2015), The effect of language on economic behavior: Experimental evidence from children's intertemporal choices, CESIFO working Paper 5532.

Swadesh, Morris (1952), Lexico-statistic dating of prehistoric ethnic contacts, *Proceedings of the American Philosophical Society* 96, 121-137.

Swoyer, Chris, Baghramian, Maria and Adam Carter (2014), Relativism The linguistic relativity hypothesis, In Edward Zalta, Ed., *Stanford Encyclopaedia of Philosophy Archive*.

Tabellini, Guido (2008), Institutions and culture, *Journal of the European Economic Association* 6, 255-294.

Tambiah, Stanley (1986), *Sri Lanka. Ethnic Fratricide and the Dismantling of Democracy*, Chicago, WI: University of Chicago Press.

Taylor, Charles and Michael Hudson (1972), *World Handbook of Social and Political Indicators*, Ann Arbor, MI: ICSPR.

Tinbergen, Jan (1962), *Shaping the World Economy: Suggestions for an International Economic Policy*, New York: The Twentieth Century Fund.

Tönnies, Ferdinand (1887), *Gemeinschaft und Gesellschaft*, Leipzig: Fuess Verlag.

Walker, John and Brigitte Unger (2009), Measuring global money laundering. The Walker gravity model, *Review of Law and Economics* 5, 821-853.

Warnow, Tandy (1997), Mathematical approaches to comparative linguistics, *Proceedings of the National Academy of Sciences of the USA* 94, 6585-6590.

Weber, Max (1968, [1910]), *Economy and Society*, Berkeley, CA: University of California Press.

Whorf, Benjamin (1956), *Language, Thought and Reality: Selected Writings of Benjamin Lee Whorf*, Cambridge, MA: MIT Press.

Wiener, Norbert (1948), *Cybernetics: or Control of Communication in the Animal and the Machine*, Paris: Librairie Herman, Cambridge, MA: MIT Press.

Williams, Donald (2006), The economic returns to multiple language usage in Western Europe, IRISS Working Paper Series 2006-7, CEPS, Luxembourg.

Wismann, Heinz (2012), *Penser entre les langues*, Paris: Albin Michel.

World Values Survey (2009), *WVS 1981-2008 official 5-wave aggregate v.20090902*, Madrid: World Values Survey Association, available at www.worldvaluessurvey.org

Wright, Robert (1991), Quest for the mother tongue, *Atlantic Monthly*, April, 36-68.

Zink, Michel (2014) Ouverture. Quelle langue est mienne, In Michel Zink, Ed., *D'autres langues que la mienne*, Paris: Odile Jacob.