DISCUSSION PAPER SERIES

No. 10640

LINGUISTIC DISTANCES AND THEIR USE IN ECONOMICS

Victor Ginsburgh and Shlomo Weber

PUBLIC ECONOMICS

Centre for Economic Policy Research

ISSN 0265-8003

LINGUISTIC DISTANCES AND THEIR USE IN ECONOMICS

Victor Ginsburgh and Shlomo Weber

Discussion Paper No. 10640
May 2015
Submitted 25 May 2015

Centre for Economic Policy Research
77 Bastwick Street, London EC1V 3PZ, UK
Tel: (44 20) 7183 8801
www.cepr.org

LINGUISTIC DISTANCES AND THEIR USE IN ECONOMICS

Abstract

The paper offers an overview of the various approaches to compute linguistic distances (the lexicostatistic method, Levenshtein distances, distances based on language trees, phonetic distances, the ASJP project and distances based on learning scores) as well as distances between groups. It also briefly describes how distances directly affect economic outcomes such as international trade, migrations, language acquisition and earnings, translations. Finally, one can construct indices that take account (or not) of distances and how these indices are used by economists to measure their impact outcomes such as redistribution, the provision of public goods, growth, or corruption.

JEL Classification: F6, O21 and Z18
Keywords: development, economic outcomes, growth, linguistic disenfranchisement and linguistic distances

Victor Ginsburgh vginsbur@ulb.ac.be
CORE, Université catholique de Louvain

Shlomo Weber sweber@mail.smu.edu
Southern Methodist University, New Economic School and CEPR

Linguistic Distances and their Use in Economics[1]

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

Shlomo Weber
Southern Methodist University, Dallas
and New Economic School, Moscow

Abstract

The paper offers an overview of the various approaches to compute linguistic distances (the lexicostatistic method, Levenshtein distances, distances based on language trees, phonetic distances, the ASJP project and distances based on learning scores) as well as distances between groups. It also briefly describes how distances directly affect economic outcomes such as international trade, migrations, language acquisition and earnings, translations. Finally, one can construct indices that take account (or not) of distances and how these indices are used by economists to measure their impact outcomes such as redistribution, the provision of public goods, growth, or corruption.

Keywords: Linguistic distances, economic outcomes, growth, development, linguistic disenfranchisement.

JEL numbers: F6, O21, Z18.

1 Introduction

Ruhlen's (1994) quest allowed him to reconstruct twenty-seven words of the very first language. Some linguists raised eyebrows about the words themselves, but not so much about the idea that all our languages descend from one, or a very small number of, language(s).[2] Today, most linguist think

[1]The authors are grateful to Nigel Fabb for many useful comments. They also wish to acknowledge the support of the Ministry of Education and Science of the Russian Federation, grant No. 14.U04.31.0002, administered through the NES CSDSI.

[2]See Nichols (2012).

that the diversity of languages is the result of the migration 'out of Africa' of *homo sapiens sapiens* over the 50,000 to 100,000 last years (Michalopoulos, 2012, Ashraf and Galor, 2013). If this is so, languages can be represented in the form of a tree similar to genealogical trees, starting with a root representing the *first* language, or ancestor, and followed by branches and twigs for descendants. This implies of course that languages are related by their vocabulary, syntax, phonology, etc. in the same way as children are related to their parents and more distant ancestors by some of their genes. Genetic differences are relatively easy to trace and DNA analyzes have become common to check, for instance, in the case of disputed parenthood. It is, however, more difficult to 'count' the (dis)similarities between languages, since many characteristics, and not only vocabularies, are involved.

Linguistic, genetic, and cultural aspects of a society or a nation are thus often correlated, since all three are closely linked to nature, but also to learning and history, that is, nurture. One can therefore wonder when it is appropriate to choose one or the other in representing proximity of individuals or groups to which they belong. In some cases, one of the measures is obvious. This is so when describing the difficulty in acquiring a foreign language, though even here, there may be other types of proximities at work between, say, a Swedish and a Danish speaking individual than the mere proximity between the two languages. Otherwise, there is no obvious answer, and certainly no theory to invoke.

The paper is organized as follows. Section 2 focuses on the reasons for which distances between languages are important, since they allow by-passing the difficult separation between, and definitions of what is called *language* and what is called *dialect, creole, pidgin* and *trade language*. Section 3 surveys the various types of linguistic distances that are available and how they are computed. In Section 4, we introduce important applications in which linguistic distances as such play a role in explaining various economic outcomes: bilateral trade flows, migrations, language acquisition, and the problem of translations. In Sections 5 and 6, we extend our discussion to linguistic distances between groups of people (countries or regions) and analyze how linguistic distances can be introduced into fractionalization (or diversity) indices in order to take into account the proximity between groups of individuals or regions and countries in which they live.

2 Languages, dialects and trade languages

Before turning to the measurement issue discussed later in this paper, it is useful to know how many languages exist, how many will be left to our heirs, and what distinguishes a language from a dialect. The 2009 edition of *Ethnologue*[3] lists 6,909 languages that are currently spoken in the world. Whether this number is large or small is open for discussion, and so is the number itself, since it results from a rather subjective count. As *Ethnologue* (2009, p. 9) notes, 'every language is characterized by variation within the speech community that uses it. Those varieties are more or less divergent and are often referred to as dialects, which may be distinct enough to be considered separate languages or sufficiently similar' to be called dialects. Moreover, 'not all scholars share the same set of criteria for distinguishing a language from a dialect.' The criteria used by *Ethnologue* (2009, p. 9) to arrive at their count 'make it clear that the identification of a "language" is not solely within the realm of linguistics.'

A couple of examples are useful. *Ethnologue* lists 57 Zapotec languages in Mexico. Some of those, such as Zapotec (San Augustin Mixtepec), count less than 100 first-language speakers. The largest, Zapotec (Isthmus), has 85,000 speakers. In contrast, *Ethnologue* also lists five dialects in the Flemish part of Belgium (Antwerps, Brabants, Limburgs, Oostvlaams and Westvlaams) that are certainly all spoken by more than 100 people, and though they are considered variants of Dutch, can probably not be understood by a Netherlander from Amsterdam. Is Québéquois close to French? This can be questioned since on French TV, it happens that French Canadian series are subtitled in ... French.

There is not much known about what really happened between the birth of the first language and the large number of languages that exist today, nor on the pace at which some languages were born, while others died. Most of what we know was reconstructed by linguists on the basis of languages that exist today or extinct languages about which we have written documents, such as ancient Greek or Latin. Obviously, there are languages that have

[3] *Ethnologue. Languages of the World* is a comprehensive catalogue of the world living languages. The project started in 1951. The last 1,250 pages thick 2009 edition is a mine of information on languages, where they are spoken and the number of speakers in each country. If not otherwise mentioned, this is the information that is used in our book. See also *Ethnologue's* website http://www.ethnologue.com/ which contains updated information.

disappeared, or are no longer spoken as first languages. Latin is one such case, but there are certainly many others that we ignore. Hebrew that was no longer spoken as a first language, is now Israel's official language.

Slavery and explorations, followed by colonization had the effect of killing languages at a fast rate. So do, nowadays, mass tourism and globalization. To mitigate the frightening prospect of linguistic and cultural loss, there is "good news," as new languages and cultures are also being born. This is certainly what Claude Levi-Strauss would have thought, when he said in one of his last interviews in 2002:

> On peut se dire, en tout cas c'est un voeu pieux, qu'au moment où nous craignons une uniformisation universelle, en sous-main, sourdement, de nouvelles différences commencent à apparaître que nous ne pouvons même pas percevoir, mais que nos héritiers comprendront et qui, je l'espère, seront exploitées.[4]

Kibbee (2003, p. 51) adds: '[a] language is a behavior, not a physical characteristic. If two languages are in contact, then they influence each other. If a dog lives in the same house as a bird it does not grow wings, nor does the bird sprout paws.'

Pidgins and *creoles* have indeed grown and made communication easier. *Ethnologue* (2009, p. 29) lists 77 creole languages, that were created by Dutch, English, French, Portuguese, Russian or Spanish colonization. They result from a blending (and simplified grammar, vocabulary and style) of the local and the colonizer's language and have become the mother tongues of many communities. Some examples are Papiamentu (in the Netherlands Antilles), Virgin Islands Creole English (Virgin Islands), Saint Lucian Creole French (Sainte Lucie), Korlai Creole Portuguese (India) or Aleut Mednyj (Russian Federation).[5] In an interesting paper published by *Newsweek* on March 7, 2005, 'Not the Queen's English: Non-native speakers are transforming the global language,' the author distinguishes more than fifty varieties, including British English (BBC English, English English, Scottish English, Scots, Norn, Welsh English, Ulster Scots, Hiberno-English, Irish English), American English (Network Standard, Northern, Midland, Southern,

[4] We may hope that at the very same time during which globalization is taking place, somehow, in a secret way, new differences are appearing, that we cannot perceive but that will be understood by our heirs.

[5] Some historical linguists have even suggested that English is a creole. See Singh (2005).

Black English Vernacular, Gullah, Appalachian, Indian English) and Canadian English (Quebec English, Frenglish, Newfoundland English, Athabascan English, Inuit English). Spanglish is not mentioned by *Ethnologue*, but is becoming important in the United States where it is used by a growing community and won acclaim since MIT's Creative Literature Professor Junot Díaz received the 2007 National Book Critics Circle Award for Best Novel and the 2008 Pulitzer Prize for *The Brief Wondrous Life of Oscar Wao*, some parts of which use Spanglish.[6]

"Old" languages enrich each other and create exciting new blends, that obviously share some or many characteristics with their predecessors.[7] Nobody can claim that newcomers to the linguistic scene, such as the many Creoles or Spanglish, are inferior to Rabelais' French, Shakespeare's English or Cervantes' Spanish – none of which can be easily understood today by a non-specialist. The same is true for cultural, ethnic and religious traits, as is shown in the papers by Bisin and Verdier (2000, 2014), who describe the theoretical processes that can give birth to what we observe in terms of cultural mixtures. Sushis were "exported" to the United States, changed there to better embrace local tastes, and are now "reimported" as such by Japan.

It is also worth pointing out that many linguistically diverse countries or regions consisting of several (often recently) defined countries, have endogenously generated their "trade language." This is so for a couple of important languages in Nigeria (which otherwise has over 520 living languages) as well as for Swahili, a Bantou language that was contaminated by Arabic (slave trade) and Portuguese (with the trading posts established after 1488, when Bartolomeu Dias reached the Cape of Good Hope and entered the Indian Ocean), as well as English (in the former British colonies where it is used) which spread on the East Coast of Africa.

How can we decide that an idiom (language, dialect, variety, etc.) is different from another, and does this matter? Theoretically, one could take into account distances between every pair of languages, including dialects[8] and the unimaginably large number of varieties and decide on a threshold which would distinguish a "language" from small variations consisting of a few words, expressions, or ways to pronounce them. This would of course

[6] And for which there even exists an *Annotated Oscar Wao: Notes and translations for the Brief and Wondrous Life of Oscar Wao*. See http://www.annotated-oscar-wao.com.

[7] See the very entertaining book on this issue by McWhorter (2001).

[8] This was undertaken for German dialects in a recent paper by Falck et al. (2009).

be a daunting project, but the theoretical and empirical basics exist and are discussed in what follows.

3 Distances between languages

Languages can be distant from each other in many different ways. *Vocabulary* is often thought of being the main reason, and indeed, though English and German are part of the Indo-European family and belong to the same branch of so-called Germanic languages, the English word *moon* is related to but differs from *Mond* in German. And they both differ strongly from *lune* in French, though French is also an Indo-European language, but belongs to the branch referred to as Romance languages. Words differ, but even *moon* and *Mond* have different genders: the first is neutral, the second is masculine and *lune* is feminine.

The pronunciation and spelling of *moon* is close to *Mond*. The final *e* in *lune* is mute, while the *a* in the Italian or Spanish *luna* is not. Diphthongs, that is, sounds composed of two vowels joined to form one sound, as in *sound* also add to the difficulty of a language. But the French student of English will wonder why the *ou* in *south* is not pronounced like the one in *tour.* Therefore, pronunciation also contributes to the distance between languages. *Phonetics* and *phonology* study how we use lips, tongue, teeth, and vocal chords to produce the various sounds (phonemes) in each language. Both the production and perception of phonemes differ across languages. Each language has phonemes that may prove difficult to acquire by non-native speakers. It seems easier for a native speaker of a language that contains a large number of phonemes to learn a language with less phonemes, since for her, both the production and perception of sounds are more developed.[9]

Syntax, the 'way in which linguistic elements (as words) are put together to form constituents as phrases or clauses'[10] illustrates another difference between languages. *I would like to observe the moon* translates into *Ich möchte (gerne) den Mond beobachten.* While the word *observe* is located in the middle of the English sentence, its translation *beobachten* ends the German sentence. It is quite difficult to grasp quickly the meaning of *fünfundzwanzig* Euros when the German taxi driver tells you how much you ought to pay for

[9] See Ladefoged and Maddieson (1996) for an extensive discussion.

[10] *Merriam-Webster Dictionary*, http://www.merriam-webster.com/dictionary/syntax.

the ride: the meaning is *twenty-five* but it is spelled out *five and twenty* in German (as well as in other Germanic languages, such as Dutch or Danish).

Grammar is of course another worry. For instance, while German has declinations, only very few of them remain in English such as the so-called possessive genitive in *the moon's last quarter.* None of the declinations that exist in Latin were inherited in contemporary (spoken) French, and a French student who did not have to learn Latin or Greek (or German) at school will hardly know the meaning of the word "declination."

These are just a few examples to illustrate that computing the distance between two languages is a stiff challenge, and this is even without going into the fundamental issue of whether languages have a common structure.[11]

The comparative method used in *historical linguistics* aims at reconstructing 'common ancestry, and descent through time with gradual divergence from [a] common source' (McMahon and McMahon, 2005, p. 3), and is not interested in the inter-comprehension of people who speak different languages today.[12] It consists of two parts: 'the demonstration of linguistic relatedness, and the reconstruction of a hypothetical common ancestral system' (McMahon and McMahon, 2005, p. 5). It uses morphological resemblances, lexical items, syntax, and sound correspondences, and identifies groups and subgroups of languages according to their similarity. The final result consists of a linguistic tree, with a root (the Ur-language or common ancestor) and branches which in turn grow into sub-branches, until one reaches final twigs, each of which corresponds to a unique language. Examples of such trees in different forms, including the old image representing a real tree, can be found by searching on the internet for the terms "language tree images."[13]

The result is comparable to what biologists (with the aid of scientists from other disciplines such as palaeontologists) adopt to construct biological and evolutionary trees. This is often called *cladistics*, a term that comes from the Greek root $\kappa\lambda\alpha\delta\o\varsigma$ (klados) for branch. To generate a tree, the method starts with a table which lists several animal or vegetal species as well as characteristics describing the various species supposed to have a common ancestor.Trees (also called cladograms) are then constructed using the information given by the descriptive characteristics, and the one considered the "best" is identified. Implicitly, cladistics also utilizes weights to generate

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

[12] We will come back to this point later.

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

the resulting tree, and therefore shares some of the issues outlined above.

The comparative method can be made looser (mass comparison) or tighter by using a unique characteristic of a language, its lexicon (which leads to lexicostatistics), but can also accommodate median situations.

Mass comparison was used by Greenberg (1987) to classify native American 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 can neither be repeated nor tested.

Lexicostatistical methods are based on one dimension only: the similarities and supposed common roots of words in the vocabularies (the lexicon) of various languages. Languages can be related or similar, and these 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 belonging to another branch: English, for example, contains some 30 percent of its lexicon borrowed from French after the Norman conquest in 1066 (and nowadays, French as well contains many English words); (c) two languages may descend from a common, older language.[14] This is the case for French, Italian, Spanish and Portuguese, which belong to the same branch, and have Latin as ancestor.

Lexical distances are built on so-called *cognate* words, occurring in languages with a historical chain linking them via an earlier language, thus

[14]Dyen, Kruskal and Black (1992) give the very poetic example of the word *flower* which is borrowed from the French *fleur*, while quite surprisingly, *blossom* and *fleur* descend from the same ancestral word.

ignoring not only borrowings[15] and accidental similarities, but also syntax and grammar. The reason is that linguists became (and still are) interested in constructing language trees, as well as estimating dates at which one language separated from another (glottochronology[16]).

Since it would be a formidable task to compare long lists of words for each couple of languages, linguists are forced to rely on a small selection of carefully chosen words, a so-called "list of meanings." Swadesh (1952) eventually introduced some rigor (Kessler, 2001, p. 31) in the choice of meanings that one can assume to be basic enough to exist in all languages and cultures (such as animal, bad, bite, black, child, die, eat, eye, hunt, numbers from *one* to *five*),[17] on which deductions can be based.[18] The list we are interested in consists of 200 basic meanings, which Swadesh later trimmed to 100. Both lists are still in use nowadays.

As we shall see, distances between couples of languages can be computed in several ways. We discuss distances simply based on the percentage of cognate words (usually called lexicostatistical distances), and Levenshtein distances based on analyzing and comparing words character by character, or their phonetical representations.

We examine in some detail the types of techniques used to compute linguistic distances which eventually are used to construct trees: (i) lexicostatistical distances; (ii) Levenshtein distances; (iii) distances based on linguistic trees; (iv) other methods.

[15]Ignoring borrowed words may rule out some factors that influence the closeness of two languages. The case of English and French is a good example. 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 any close 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.' The French linguist Hagège (2009, pp. 647-670) adds that English is particularly difficult to learn. He probably had in mind French speakers, but his remarks seem to be more general.

[16]The premises on which glottochronoly is based – in particular that the probability of a word losing its original meaning stays constant over time – are being questioned, and this area of research fell into some disrepute. See, however, Gray and Atkinson (2003), Searls (2003) and McMahon and McMahon, 2005, pp. 177-204) for a recent revival of the concept.

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

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

3.1 Lexicostatistical distances

Dyen, Kruskal and Black (1992) used Swadesh's basic list of meanings to classify 84 Indo-European speech varieties. 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. See Gamkrelidze and Ivanov (1990), and also Ruhlen (1991) for a general overview. They describe the lexicostatistical method as consisting of four phases:

(a) Collecting for each of the meanings in Swadesh's list the words used in each speech variety under consideration.

(b) Making cognate decisions on each word in the lists for each pair of speech varieties, that is, deciding whether they have a common ancestral word, or not, or whether no clear-cut decision can be made; this phase is performed by linguists who know the language family.[19]

(c) Calculating the lexicostatistical percentages, i.e., the percentages of cognates shared by each pair of lists; these percentages lie between one (if all words are cognate; actually, the largest number of cognates found by Dyen, Kruskal and Black was 154, leading to a percentage of 0.770) and zero (if there is no cognate).

(d) Partitioning the word lists into family trees; this is performed using one of the many existing clustering algorithms.

Table 1 which gives an example of a list of words for the basic meanings of numbers one to five (which are part of Swadesh's 200 meanings) in 15 languages. This is essentially what is done in Step (a), usually for a much larger list of meanings and languages. In this table, the first four groups are Indo-European languages, which are themselves subdivided in Germanic, Romance, Celtic and Slavic subgroups. 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 1 approximately here]

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

Steps (a) and (d) need no comments, though step (a) is not only time consuming, but difficult since the right choice has to be made among possible synonyms. We describe what is done in steps (b) and (c).

Step (b) is devoted to comparing words for every pair of languages.[20] This looks of course simple in Table 1, where languages are already grouped into families. The words in the first five languages (Danish, Dutch, English, German, Swedish) are all cognates, as the pairwise comparisons show. They all 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 has all five numbers that are cognate. But more importantly, the first three numbers in all three families can be seen to be cognates as well (and indeed, Germanic, Romance and Celtic language families are part of the larger family of Indo-European languages). Cognation decisions are less obvious and need more linguistic knowledge for the numbers four and five. Next come Slavic languages, which are very clearly related to each other and for which one also sees a relation of digits two and three with the previous groups, but this is less so for the other numbers. The words in the last two languages, Hungarian and Swahili have little in common with those of the four families of Indo-European languages, but may nevertheless have a faraway ancestor (or Ur-language).[21] In general, cognate decisions need trained linguists,[22] 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.*

Step (c) is easy once cognate decisions are made. It consists in counting the number of cognates for each pair of languages, and divide it by the total number of meanings. If the only meanings to be compared were our five num-

[20]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."

[21]See Ruhlen (1994) for a deep but also entertaining exposition. He makes the reader construct the cognates 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.

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

bers, 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 Brythonic *pedwar*) are certainly further apart. Assume that as linguists, we decided to classify them as non-cognate. Then, the distance between, say English and French, would be 3/5. Finally, in our example 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, the numbers given in Table 2 are equal to one minus the percentage of cognates. They concern the distances between 25 European languages[23] and the six European languages that are most spoken in the European Union: two are Germanic (English and German), three are Romance (French, Italian and Spanish) and the last is Slavic (Polish). It is easy to check that Danish, Dutch, English, German, Icelandic, Norwegian and Swedish are related. So are the Romance languages Catalan, French, Italian, Portuguese, Romanian and Spanish, and the Slavic ones, Bulgarian, Czech, Russian, Serbo-Croatian, Slovak, Slovene, Ukrainian, Latvian and, to some extent, Lithuanian. Albanian and Greek are distant from any language belonging to the three previous families.

[Insert Table 2 approximately here]

3.2 Levenshtein distances

In step (b) described in Section 3.1, knowledgeable linguists judge words (meanings) between two languages as being cognate (distance equal to zero), not cognate (distance equal to one) or ambiguous (and thus eliminated from the count). In step (c), the number of cognate words is computed and divided by the total number of words (minus the doubtful ones). The resulting number is considered to represent the distance between the two languages subject to the comparison. Levenshtein (1966) suggests an algorithm that

[23] Basque, Estonian, Finnish, Hungarian and Turkish (spoken in Cyprus) are excluded, since they do not belong to the Indo-European family. The distances with Indo-European languages are set to 1 as an approximation.

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 (and phonetic, see below) 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 100 words list. The word is spelled *Nacht* in German and *notte* in Italian. A linguist would probably classify the words as cognate between all three languages. The Levenshtein distance between the English and the German words is two, since one needs to substitute *a* to *i* and *c* to *g*. The distance between English and Italian is four (substitute *o* for *i* and *t* for *g*; 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.[24] This is in accordance with our intuition: English and German are Germanic languages, while Italian is a Romance language, but all three are Indo-European.

The distance between two languages is now simply the average of the distances over all the meanings that are compared, and linguistic trees can be computed using clustering algorithms.

Computing Levenshtein distances is easy to program on a computer. This allows computing distances for a large number of languages, a tedious in the case of 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. 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. It is however less likely with Swadesh's 100 list, which compares words that have the same meaning in *different* languages.

3.3 Levenshtein phonetic distances

It should be quite obvious that the similarities of characters in words is insufficient if these are spelled differently in the various languages that are ana-

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

lyzed. This is especially true for vowels and diphthongs. Take for example the meaning *fire*, which is written *Feuer* in German. If the Levenshtein method discussed above is used, the distance is 3/5. But Levenshtein's method also allows accounting for phonetic similarities by transcribing the words into their phonetic equivalents, using existing or especially tailored phonetic alphabets.[25] Phonetically the words *fire* and *Feuer* differ only through the English *i* and the German diphthong *eu* (which in English sounds like the */oi/* in the word *boiling*). The phonetic distance would be smaller, since the remainder of the two words is (roughly) pronounced in the same way. Another totally different phonetic approach based on speech sound will be discussed in Section 3.6.

3.4 Cladistic distances

An alternative way of computing linguistic distances is to utilize linguistic tree diagrams, based on world classifications of languages such as *Ethnologue* and compute cladistic distances. These have two advantages over lexicostatistical distances: they account for various aspects that characterize languages, such as lexicon, syntax, phonology, grammar, and are available for almost all languages in the world. As will be seen, however, they are less precise than lexical distances. As Laitin (2000, p. 148) notes:

> First, written forms could be quite different from spoken, and there were no general criteria to judging whether or how far two pronunciations needed to be from one another to be coded as a different word. Second, many forms for each of the items on Swadesh's list, and for any word, [distance] could change based on which word in a set of synonyms is chosen. Third, [...] linguistics, especially in the United States, focused almost entirely on structure by the 1960s, and much less on meaning. Therefore languages were coded based on mostly syntactic structures rather than on world relationship.

Fearon and Laitin (1999), Laitin (2000), and Fearon (2003) who suggested this approach[26] use the distances between linguistic tree branches as

[25] See Sections 3.4 and 3.6

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

a proxy 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. The approach was later followed by many researchers.

We use the (simplified) Indo-European language tree of Table 3 to calculate such distances for some languages.[27] 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 be 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. Thus, 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 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 distance. To sustain this feature, in his derivation Fearon utilizes

[27] Though Indo-European languages were among the first to be discussed and represented under the form of a tree, this is now so for all the world's languages. For other families, see http://linguistic.org/multitree and click on 'Go to the MultiTree Browser.'

the square root of linguistic distances, rather than distances themselves.[28]

3.5 Phonetic distances

There also exist methods that use *phonetic* similarities. One group of such methods, which carries the name of *dialectometry* computes distances between the elements of a word in two different languages, often using the Levenshtein distance. 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).

According to McMahon and McMahon (2005, pp. 212-214), the technique may work for 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 [within words] would have taken place.' 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, McMahon and McMahon (2005). See also McMahon and McMahon (2005, pp. 214-239).

A different approach of phonetics is considered by the Functional Phylogenies Group, in which linguists, phoneticians, statisticians, mathematicians, palaeontologists, and an engineer in aeronautics work together and analyze phonetic sound properties '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, Aston and Evans, 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, 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. Then the second branch consisting of Spanish and French split into Spanish (cien)

[28] A variant of Fearon's formalization is used by Desmet, Ortuño-Ortín and Weber (2009).

and French (cent).

3.6 Adding typology to lexicostatistics: The ASJP project

A group of linguists associated to the Automated Similarity Judgment Program (ASJP)[29] combines lexicostatistics (using a subset of 40 words from Swadesh's 100 words list) with 85 phonological, grammatical, and lexical structures described in Dryer and Haspelmath's (2013) *World Atlas of Language Structures* (WALS). ASJP transcribes the meanings using 41 different symbols (7 vowels and 34 consonants). It relies on Levenshtein distances.

3.7 Distances based on learning scores

The approaches described so far all belong to historical linguistics. Their aim is to construct trees, and they are not much interested in throwing light on current inter-comprehension between populations. On the contrary, and as we shall see in Section 4, economists are interested in today's world and to what makes trades, migrations, or translations easier.

Given the difficulty or the relative arbitrariness of representing the distance between two languages by a unique encompassing number or giving weights to different characteristics of languages and aggregate them (as is done in the ASJP project), the "best" method (which takes into account all characteristics, as well as borrowed words) would be to follow the speed of the progress made by people who learn languages, and measure their proficiency in some objective way and at different moments during and at the end of their learning period. Such a measure was established by Hart-Gonzalez and Lindemann (1993) using a sample of native Americans who were taught a variety of languages. Chiswick and Miller (2007a) suggested that this measure could be positively correlated with the difficulty of inter-comprehension, and used the scores as distances between American English and some other languages spoken by immigrants. The scores vary between 1 for difficult – Japanese and Korean – to 3 for easy – Afrikaans, Norwegian, Romanian,

[29] See http://email.eva.mpg.de/ wichmann/ASJPHomePage.htm for the aims of the program, the database, and a list of publications.

Swedish).[30] Table 4 contains the full set of scores, some of which look quite surprising.[31]

Insert Table 4

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 in Section 3.1 to 3.6 for three reasons at least. First, they encompass most of the difficulties encountered in acquiring a language. Secondly, scores would not necessarily be symmetric (as is the case for all other methods), since learning language A for a native speaker of B may be more difficult than learning B for a native speaker of A. Finally, 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.

3.8 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 gets even worse with languages such as 'Spanglish' in the United States, 'Konglish' (spoken by an older generation in South Korea), or 'Globish' everywhere.

The phenomenon is obviously not limited to English. There are quite substantial differences between the German spoken in Germany and in Austria and there even exist Austrian-German dictionaries.[32]

[30] This distance is used in two papers by Chiswick and Miller (2007b, Chapter 1), as well as by Hutchinson (2003) and Ku and Zusmann (2010).

[31] Swahili, a Bantu language essentially used in Eastern Africa, is closer or at least easier to learn for an American than German or French. One can, therefore, wonder whether scores are calculated "all other things being equal," and on a sufficient number of observations.

[32] See, e.g., http://www.dictionaryquotes.com/quotations-subject/454/Language.php for a partial, but by no means exhaustive list of more than 250 meaning represented by different words in Austria and Germany.

The main problem is of course due to words borrowed in one language from other languages, which are often not accounted for as reported above, but nevertheless facilitate somewhat reading and learning, even if other dimensions such as grammar or pronunciation are different. Still, as we shall see now, "historical" distances appear to significantly affect economic outcomes, and even if they have defects, seem to be a good approximation of the differences between contemporary tongues.

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.

4 The effect of linguistic distances on economic outcomes

The linguistic distances discussed in the previous section have important applications and economists have shown that they significantly matter in many fields: international (and even national) trade and migrations, language acquisition, and translations. Intra-country differences and fractionalization aspects of linguistic diversity are examined in Section 5.

The issue raised by the linguistic distance between two countries, regions or groups of people is different from that of two languages, independently of where they are spoken. We will turn to this issue in Sections 5 and 6. Most applications using inter-country linguistic differences are based on what is known as the "gravity model," whose name comes from its analogy with Newton's 1687 Law of Universal Gravitation which postulates that any two objects in the universe exert gravitational attraction on each other with a force, denoted f_{AB}, that is proportional to the product of their masses, m_A and m_B, and inversely proportional to (the square of) the distance d_{AB} that separates the two objects. Distance thus has a negative effect on the attraction force.

Hägerstrand was probably the first (in 1957) to transpose this law and

apply it to migrational flows in Sweden.[33] Here f_{AB} represents a flow of migrants between two cities or regions A and B; m_A and m_B are measures of wealth or income (population, gross domestic product), and d_{AB} is the geographic distance between the main cities of the two regions.

Tinbergen (1962) suggested that it could be applied to study international trade flows between various countries. Now, the force f_{AB} represents the volume of exports from country A to country B, m_A and m_B are as before and d_{AB} is the geographic distance usually between the capitals of the two countries A and B.

In both cases distances can be more generally thought of costs and impediments to trade and migration.

The gravitational equation became very popular, and fitted the data very well. In addition to trade and migration, we also discuss its application to literary translations (Ginsburgh, Weber and Weyers, 2007).[34]

4.1 International Trade

The basic gravitational trade equation can be written:

$$x_{AB} = \gamma + \alpha \text{GDP}_A + \beta \text{GDP}_B + \delta \text{dist}_{AB}$$

where γ, α, β and δ are parameters the value of which have to be assessed, x_{AB} represents (the log of) exports from A to B, GDP_A and GDP_B represent the (the log of) economic power of both countries (often measured by their gross domestic products) and, finally dist_{AB} is the (log of the) "distance" between A and B. The expected signs of α and β are positive: the richer a country, the more it has a propensity to export or import; the expected sign of δ is negative, since an increase in distance should reduce its trade flows $\exp_{AB}$.[35]

[33] According to Kerswill (2006, p. 4), who himself quotes it from a book by Gareth Lewis, *Human Migration: A Geographical Perspective*, London: Croom Helm, 1984.

[34] There are applications of the gravitational equation in other fields, e.g., explanations of flows of money laundering (Walker, 2000). Linguistic distances are also used without applying the gravity model.

[35] Without going into any detailed discussion, we simply state that the the introduction of trade-theoretical arguments leads to a slightly different gravity-like equation that reads: $x = \gamma + \alpha(\text{GDP}_A + \text{GDP}_B) + \delta \text{dist}_{AB} + \text{other variables}$. Here x represents (the log of) total trade between A and B, that is exports from A to B plus exports from B to A.

Among the many types of distances used, linguistic distances are common, and it has often been shown that a common (or a close) language between countries A and B enhances their trades.

4.2 Migrations

The standard approach in analyzing the trade-offs of a decision to migrate is again based on evaluating costs and benefits. The prospects of higher wages or other benefits[36] are contrasted with the monetary and psychological costs, adjustment to a new culture and possible uprooting of the family. The cultural and linguistic frictions in a new country, based on earlier immigrants' experience, can profoundly influence individual decisions. The form of the typical immigration equation is very close to the trade equation, but its theoretical underpinnings are different. Researchers relate migrations to existing networks that 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 on the decision to migrate.

4.3 Language acquisition and earnings

Selten and Pool (1991) introduced a game-theoretical model of communicative benefits, that covers a wide range of economic, cultural and social advantages gained by learning languages. Somewhat later, 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 in each country, only one of the two populations learns the other language (corner solutions). Gabszewicz, Ginsburgh and Weber (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. The model was taken to data by Ginsburgh, Ortuño-Ortin and Weber (2007), and Ginsburgh, Melitz and Toubal (2014) who estimate an equation that relates learning decisions of 13 of the most important world languages by

[36] In migrations between developing countries prevention against risks rather that income maximization seems to be of relevance in the decision to migrate. See Guilmoto and Sandron (2001).

citizens who live in some 190 countries. They find that learning is influenced positively by trade (which is instrumented to avoid the endogeneity issues, since linguistic distances also have an effect on trades), as well as by the literacy rate in the learning country, and negatively by the linguistic distance between the home and the acquired language, and the number of speakers of the home language. The results also show that, contrary to what is often thought, Indo-European languages, and more specifically, English have no "special" status with respect to other important world languages. It may thus happen that Chinese or Arabic, and not English, will become the next *lingua franca*.

Economic returns of learning languages (in particular earnings) have been the subject of intensive research. However, one has to draw a distinction between migrants and native workers. Migrants usually have to acquire the language of the country to which they migrate, unless they knew it before migrating, while native residents in a country born in their home language may decide to acquire foreign languages that they use at the workplace. In some papers (Chiswick and Miller, 2007b), the distance between native and acquired language is taken into account and plays a significant role.

4.4 Cultural flows and translations

The necessity for translation is still an important issue in various national and international organizations such as the United Nations, the Work Bank, the IMF or the OECD, where there is translation and interpretation for selected languages only. This is not the case of the European Union which has to deal with 24 languages at a cost of some $1.5-2 billion; these could probably be put to better use. Past (as well as more recent) history shows that translations (treaties between countries, not to speak about most legal and of course literary texts), even by professionals, are not always accurate and may lead to different interpretations and economic losses (Lewis, 2004) or wars.[37] 'A less dramatic example of the problem is described by Wright (2007) who spent

[37]The Japanese word *mokusatsu* used as a response to the Potsdam declaration that required Japan to surrender on July 26, 1945, 'has no exact equivalent in the English language. It is a word which is ambiguous even in the Japanese. It might be translated roughly as "to be silent" or "to withhold comment" or "to ignore." "To withhold comment" probably comes the closest to its true meaning, implying that something is being held back, that there is something significant impending. Certainly that is what the Japanese meant The Japanese government soon discovered to its dismay, however, that the meaning of its policy of *mokusatsu* had been completely misinterpreted by the outside world... The

some time in the European Parliament Babylonian environment, and notes that since not enough native Portuguese could be found, 'the high number of Brazilian interpreters was an issue for the Portuguese. Respondents stated that they were irritated by interpretation into a language that they saw as allied to their own, but not their own.'

The literature on cultural transmission deals essentially with the media industries, and especially with movies and television programs.[38] Much less is written on music, that does not need translation, dubbing or subtitles. People do not only watch television, movies, or listen to music, they also read. Though television and broadcasting have changed considerably the way culture is transmitted, books (and more generally written material, including the web) remain essential. As Susan Sontag pointed out while receiving the Peace Prize at the Frankfurt Book Trade Fair in 2003:

> What saved me as a schoolchild in Arizona, waiting to grow up, waiting to escape into larger reality, was reading books, books in translation as well as those written in English. To have access to literature, world literature, was to escape the prison of national vanity, of philistinism, of compulsory provincialism, of inane schooling, of imperfect destinies and bad luck. Literature was the passport to enter a larger life; that is, the zone of freedom.

However, translations are sometimes accused of leading to a form of cultural domination by some languages. According to Melitz (2007), 'if one language is sufficiently larger than others in the sales of original-language works, it will tend to crowd out the rest in translations [and] those writing in the dominant language are privileged.' Ganne and Minon (1992) show that France, Italy, Spain and Germany translate much more (18, 25, 26 and 15 percent) than the United Kingdom (3.3 percent). They attribute this fact to the 'abundance of books that originate in the United States,' and that need no translation in the UK. They also show that English is the language that generates the largest number of translations in these countries. Heilbron (1999) describes the system as accounting for uneven flows between

Japanese government's intention of holding the matter open for an eventual favorably inclined response came to naught with the ensuing stiffening of the Allied attitude' (Kawai, 1950, 412-413) and the atomic bomb on Hiroshima and Nagasaki was the response.

[38]See Hoskins, McFadyen and Finn (1997) and the list of references therein. See also Hanson and Xiang (2011).

languages groups: On the European continent, 50 to 70 percent of the published translations are made from English. This is quite obvious since the English speaking population is large and spread over many countries with very different cultures (the US, Canada, India, East and West Africa, etc.), which leads to a wide variety of literatures.

These considerations, however, ignore another important factor in comparing the number of translations: the role of cultural proximities. Except for the sake of exoticism, a thriller that features New York is more likely to be translated from English into French than a Chinese or an Estonian thriller that unfolds in Shanghai or Tallinn. Just think of how hard it is to read Dostoiewski or Tolstoi with their many characters (which often requires a list of names describing who is who), before trying to get accustomed to Chinese or Estonian names of characters and streets.

Ginsburgh, Weber and Weyers (2007) construct and estimate a model that offers some insight into determinants of literary translations. Though the resulting demand equation for translations is very close to the gravity model, its theoretical roots are derived from microeconomic assumptions, that are not discussed here. Let us simply mention that the resulting equation specifies that he number of titles translated from a (source) language A to a (destination) language B is determined by the following variables: (a) the sizes of the populations that speak A and B as first language, (b) the distance between the two languages, (c) the literacy rate and the average income of the population speaking the language into which the title is translated. The model fits well the data and the conclusion of English (American) language hegemony in literature is based on incomplete reasoning: the number of books that are translated from language A to other languages is not necessarily an accurate indicator of the power of A. Account has to be taken of the number of books written in the source language, as well as the cultural distances between languages. It is obvious that the more titles are written in a language, the more will be translated into other languages, as long as cultural traits are similar: the smaller the distance, the larger the number of translations. Once the number of titles translated between languages takes the two factors into account, the English language hegemony hypothesis ceases to hold.

Moreover, Ginsburgh, Ortuño-Ortin and Weber (2007) show that though it does not fully represent cultural distances between languages, linguistic distance is highly significant and its estimate shows that an increase in the linguistic distance of 10 percent decreases the number of translations by 10

percent.

5 Linguistic distances between groups

The linguistic distance between two distinct population groups, such as countries, is different from the distance between languages, independently of where they are spoken.

A simple approach is to consider the distance as a dichotomous variable which takes the value 0 if the same language is used "extensively" in both countries (as is the case between Austria and Germany), and 1 otherwise.[39]

An alternative is to estimate the likelihood that citizens in two countries can speak a common language as reflected by Greenberg's (1956) H index of communication in a multilingual society. This estimate is determined by the probability that two members of the population chosen at random will have at least one language in common.

In order to illustrate the derivation of the index, consider a two-country case, say Germany and France. The communication distance between Germany and France is defined as the probability that a randomly chosen pair of French and German citizens speak no common language and are unable to communicate. Assume that all French and German citizens speak the language of their country, but in addition, 20 percent of Frenchmen speak German and 25 percent of Germans speak French; no other language is spoken. A French and German citizen will be unable to communicate only if they are both unilingual. Since 80 percent of Frenchmen and 75 percent of Germans are unilingual, the probability that they cannot understand each other is equal to $0.80 \times 0.75 = 0.60$.

The situation is more complicated when some Germans and Frenchmen can also communicate in, say English. Now the French population consists of four groups: 70 percent are unilingual, 15 percent speak French and German, 10 percent speak French and English, and 5 percent speak all three languages; in Germany 60 percent are unilingual, 15 percent speak German and French, 15 percent speak German and English, and 10 percent speak all three languages. For communication between a Frenchman and a German to fail, we need at least one of them to be unilingual. Indeed, if both speak at

[39]This "rough" approach can be refined. See, e.g., Melitz (2008), who replaces the notion of "extensive" by "widely spoken" if at least 20 percent of the population know the language.

least two languages, they will share a common language. Thus, 70 percent of unilingual Frenchmen cannot communicate with 75 percent of Germans (those who do not speak French). However, we need also to account for a possible interaction between 60 percent of unilingual Germans and 10 percent of Frenchmen who speak French and English, which is not covered by the previous case. The total percentage of those pairs of French and German citizens unable to communicate is therefore $0.70 \times 0.75 + 0.60 \times 0.10 = 0.585$.

A variant of such index in a multi-country setting, the so-called Direct Communication Index, is obtained by adding the products of the respective percentages of speakers over all relevant languages. See Melitz (2008).

Lieberson (1964, 1969) used Greenberg's approach to evaluate the degree of communication between distinct linguistic communities. He calculated distances between three large East Coast Canadian cities, Toronto, Montreal and Ottawa. Such distances are also used in international trade and migration models.

6 Fractionalization and disenfranchisement indices

While the H index discussed above is based on the communication structure of a diverse society, most studies on societal diversity in economics and other disciplines focus on identification and measurement of diversity. We will distinguish two main types of indices: *fractionalization* and *disenfranchisement* indices.

Fractionalization indices, based on a partition of the society into distinct (ethno)-linguistic groups, allow conducting cross-country or cross-regional comparisons and examining differences in various economic and political systems, institutions and outcomes influenced by linguistic diversity.

Disenfranchisement indices are related to the notion of linguistic disenfranchisement caused by government policies. A society (country) may have to or wish to "standardize," that is, reduce the multilingualism that prevails and choose a set of official languages that will be used for administrative, legal, and educational purposes, and "discard" other languages that are also spoken in the country. To analyze in a rigorous way the potential impact of such policies that unavoidably create "disenfranchised communities" requires a quantitative evaluation of disenfranchisement.

6.1 Distance-weighted fractionalization indices

6.1.1 Formulation

In evaluating linguistic diversity, one has to recognize that there are distinctive languages that identify the members of a given society. The presence of different attributes generates a partition of this society into *groups* distinguished by their linguistic characteristics. For simplicity, we assume that each group speaks its own native language only. This assumption is obviously somewhat restrictive. As Laitin (2003, 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.' Meanwhile, some groups speak languages that are close (say, Venetian and Italian), other groups do not (Turkish and Greek). Therefore, distances also matter and should be taken into account when measuring diversity.

The most widely used dataset of worldwide ethnolinguistic fractionalization was constructed by a group of about 70 Soviet ethnographers from the Miklukho-Maklai Research Institute in Moscow, which was part of the Department of Geodesy and Cartography at the USSR State Geological Committee (Atlas Narodov Mira, 1964). Their country-by-country construction, widely known as ELF (Ethnolinguistic Fractionalization), is based mainly on linguistic and historic origins of various groups. The findings of this remarkable and impressive project were introduced in the Western literature by Rustow (1967) and Taylor and Hudson (1972). The ELF dataset was expanded by Alesina et al. (2003), who disentangle the linguistic and ethnic aspects of fractionalization and construct separate datasets determined by linguistic, ethnic and religious affiliation. While the linguistic variable is calculated entirely on the basis of data from the Encyclopedia Britannica (2001), the construction of the ethnic data set necessitated using additional data from the CIA World Fact Book (2000), as well as Levinsohn (1998) and the Minority Rights Group International (1998). In summary, the impressive Alesina et al. (2003) datasets cover some 200 countries, 1,055 major linguistic and 650 ethnic groups. Alesina and Zhuravskaya (2008) went a step further and, using census data, extended the previous datasets to cover about 100 countries on a sub-national (regional) level. Desmet, Ortuño-Ortin and Wacziarg (2012) construct an alternative dataset using cladistic distances

based on Ethnologue's trees.

Assume that the society consists of K distinct groups, where $s_1, \ldots, s_K$ represent the shares of the groups in the total population. Obviously, $\sum_{k=1}^{K} s_k$ =1. Denote by $d(k,l)$ the linguistic distance between groups k and l ($0 \leq d(k,l) \leq 1$), where $d(k,l)$ can be any of the types of distances discussed earlier in this paper.

In his seminal paper Greenberg (1956) proposes a diversity index B that measures the average linguistic distance between two randomly chosen members of the society:[40]

$$B = \sum_{k=1}^{K} \sum_{l=1}^{K} s_k s_l d(k,l).$$

Desmet, Ortunño-Ortín and Weber (2009) suggest a variant in societies with a dominant group called "center." In evaluating the total degree of diversity in such a center-dominated society, their PI index takes into account the distances between the center and peripheral groups only, but not those between peripheral groups themselves. The functional form of PI is similar to B, except that the distance between every pair of peripheral groups is zero. Thus, in a society with a central group whose population share is s_c, the PI index is written:

$$PI = s_c \sum_{k=1}^{K} s_k d(k,c).$$

This index is further refined by Akchurina et al. (2014) for the purpose of cross-country analysis. The term s_c is replaced by $DOM(c)$ in order to account for the influence and dominance of the center with respect to peripheral groups that depend on their relative size and distribution of the population across different groups. This asymmetric treatment of groups is in line with Posner (2004) who argues that an appropriate index of diversity should distinguish groups on the basis of their involvement in political decisions and impact of economic policies they experience.

The majority of empirical and theoretical studies of diversity use a dichotomous distance between groups, where $d(k,l) = 1$ for any two distinct groups, and zero otherwise. It is easy to see that the B index turns into a simpler expression called A-index by Greenberg:[41]

[40] See also Nei and Li (1979), Rao (1982), Ricotta and Szeidl (2006), Desmet, Ortuño-Ortin and Wacziarg (2012), Bossert, d'Ambrosio and La Ferrara (2011).

[41] This is index is often called ELF, which is somewhat misleading. What is usually

$$A = \sum_{k=1}^{K} \sum_{l=1}^{K} s_k s_l,$$

for all $k \neq l$. Given that $(\sum_{k=1}^{K} s_k)^2 = 1$, this index can (and is often) also written:

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

Note that this index had already been introduced 100 years ago by Gini (1912) and that the expression $\sum_{k=1}^{K} s_k^2$ is the celebrated Hirschmann-Herfindahl Index (HHI) defined for an industry with K firms, where s_k stands for the market share of firm k.[42]

More than thirty years later, scholars across various disciplines almost simultaneously (and independently) addressed the issue of measuring diversity in their own field of research, reestablishing either the A-index itself or some closely related forms. In his one-page article Simpson (1949) produced what is now known as the Gini-Simpson index of biodiversity. A seminal contribution by Shannon (1948) also describes a diversity index, entropy, which influenced a large volume of research in information theory and statistics. It is given by the following expression:

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

Both A and E represent special cases of the more general Hill (1973) diversity index:

$$HI = \left(\sum_{k=1}^{K} s_k^a \right)^{\frac{1}{1-a}}.$$

It can indeed be verified that for $a = 2$, the Hill index is equivalent to (has the same properties as) the A index, though it has a slightly different form, while in the limit for $a = 1$, HI boils down to E.

Even though the A-index is the most often used "size-based" diversity index in empirical studies, it is by no means exclusive. Fishman (1968) and

called an ELF index in the literature, is, in fact, the A-index applied to the so-called Ethno-Linguisitic Atlas Narodov Mira (1964) dataset.

[42]HHI can be viewed as an indicator of the industry "degree of monopolization" and is widely applied in competition and anti-trust law.

Pool (1972) estimate linguistic homogeneity (the inverse of diversity) as the percentage of native speakers of the most widespread language in the country. Nettle (2000) uses the ratio between the number of languages spoken and total population. Gunnemark (1991) suggests computing the share (or the number) of members of each linguistic group for whom the language spoken at home is not the official or the country's most widely used language.

6.1.2 Applications

The indices described in the previous section are used in equations that explain the effect of fractionalization on economic or sociological outcomes y such as growth, redistribution, public goods' provision, or corruption. The equation reads:

$$y = \gamma IND + \sum_k \xi_k z_k + \epsilon,$$

where γ and the ξ_k are parameters to be estimated, $z_k, k = 1, 2, \ldots, m$ are exogenous control variables, ϵ is an error term, and IND represents one of the indices described above. The parameter of interest is of course γ, since its sign indicates whether fractionalization has a *causal* positive or negative effect on y. Causality imposes, among others, that IND is exogenous, which has often been disputed.

This follows Greenberg's (1956, p.109) suggestion that such measures should be used to compare dissimilar geographical areas, and correlate (which implies no direction of causality) the degree of linguistic diversity with political, economic, geographical, historical, and other non-linguistic factors.

The program was first picked up by political scientists. Hibbs (1973) appeals to ethnolinguistic diversity in his study of mass political violence. The so-called Fishman-Pool hypothesis based on the works by Fishman (1968), Pool (1972) and Nettle (2000), asserts that linguistic diversity has an impact on economic activities. Nettle (2000) points out that the index used by Fishman and Pool (the share of speakers of the most widespread language) does not fully account for the extent of multilingualism. Alternatively, using Nettle's index (number of languages divided by the entire population) may lead to puzzling conclusions. Consider, for example, the impact of the break-up of the Soviet Union on linguistic diversity in Russia. The much smaller population of the Russian Federation speaks roughly the same number of languages as the one spoken in the former USSR. This results in a dramatic

increase of the 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. The A and B indices discussed above avoid such problems.

Most papers (essentially written by economists) support the conclusion that linguistic fragmentation has a negative impact on economic development. Mauro (1995) argues that ethnic and linguistic fractionalization reduces institutional efficiency and increases the level of corruption generated by the lobbying activities of multiple groups. Easterly and Levine (1997), who coined the "Africa's growth tragedy" expression, highlight the negative impact of diversity on economic growth. Alesina et al. (2003) and La Porta et al. (1999) claim that ethnic and linguistic fragmentation reduce the quality of governments. According to Alesina, Baqir and Easterly (1999), ethnically fragmented communities run larger deficits and exhibit lower spending shares on basic public goods, including education. Annett (2001) points out that ethnic fractionalization leads to political instability and excessive government consumption that may, in turn, have a negative impact on growth. Alesina, Michalopoulos and Papaioannou (2012) show that the combination of linguistic diversity and economic inequality could be a cause of regional underdevelopment.

While most papers described summarized above use the A index (with dichotomous linguistic distances), there are also several authors who have introduced one or the other linguistic distances described in Section 3. Introducing non-dichotomous linguistic distances, that is using B-type indices, seems to have a much stronger explanatory power, as shown by Desmet, Ortuño-Ortin and Weber (2009) in their study of the influence of ethnolinguistic diversity on redistribution within a country.

But there are also examples where diversity could be a driving force for progress. 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. Lian and O'Neal (1997) and Collier (2001) argue that more fractionalized societies could, under some conditions, perform in a better way than more homogeneous ones. A more diversified environment attracts creative individuals, ventures, businesses and capital. The success of Silicon Valley in the late 1990s is often attributed to the background of scientists, engineers and entrepreneurs who flocked to California from all corners of the world, including India, China, Taiwan, and Israel: Saxenian (1999) points out that more than thirty percent of businesses

in Silicon Valley had an Asian-born co-founder. Diversity did not prevent and, in fact, even reinforced the commonality of worker's purpose and goals. Ottaviano and Peri (2005) investigate whether and how linguistic diversity affects wages rates in US cities and find that richer diversity is systematically associated with higher hourly wages.[43]

6.2 Distance-weighted disenfranchisement indices

6.2.1 Formulation

In Section 6.1, linguistic diversity is thus treated as an exogenous variable determined by the static linguistic fabric of society. Leaving aside the endogeneity issue created by previous changes of diversity over time and the effect this has on the estimation of parameter γ, we turn to measuring the degree of disenfranchisement (Ginsburgh and Weber, 2005) that voluntary restrictions of the number of languages may generate in a society.

Consider a multi-lingual society in which every member is characterized by her linguistic repertoire, represented by the languages she speaks. Assume that the society faces the problem of selecting a subset of languages (that will be called *core* languages in what follows) to be used in official documents, for communication between institutions and citizens, debates in official bodies, etc. Such restrictions may have a major negative impact on the wellbeing of some members by limiting their access to laws, rules and regulations and make them incur emotional pain and distress.

This makes us turn to the construction of disenfranchisement indices taking into account distances between languages. We calculate disenfranchisement using two approaches. The first considers that individuals are attached to their native language only, which indeed eases their ability to read, write and communicate with others. They suffer if their native language is not included in the core set as it may also affect their sense of national pride and negatively impact their involvement in the social, political and economic life of the society in which they live. The other alternative assumes that multilingual citizens also care for languages that are part of their linguistic repertoire, and not only for their native language.

In order to compute disenfranchisement indices, and similarly to what is done for fractionalization indices, we need to identify:

[43]The innovative and creative aspect of diversity is also studied in theoretical papers by Lazear (1999).

(a) the functional form of the index;
(b) data on language proficiency (both native and acquired) of individuals in a given country, or group of countries;
(c) distances between languages.

Point (a) poses the challenging question to which theory has not yet provided a satisfactory answer: How should one aggregate individual feelings of disenfranchisement into a collective index. For simplicity, we use a linear aggregation mechanism by simply adding individual evaluations of disenfranchisement. If the sizes of linguistic communities are very unequal, one probably needs to make an adjustment for the differences by using a formulation close to the Penrose (1946) Law which introduces a discount factor for the weight of larger communities. As for (b) the data are usually provided by surveys or census information. Point (c) is discussed in Section 3.

Suppose that the set of core languages C, is a subset of the entire linguistic menu of the country consisting of K languages. For every language k denote by $d(k, C)$ the minimal distance between language $k \in K$ and the languages in C. That is,

$$d(k, C) = \min_{l \in C} d(k, l),$$

where $d(k, l)$ is the linguistic distance between languages l and l. This leads us to the first disenfranchisement index:

$$DIS_B^n(C) = \sum_{k=1}^{K} s_k d(k, C),$$

where s_k is as before the share of the population that speaks language k. The index represents the average distance between native languages of all individuals and the set of core languages. This interpretation is similar to the one of Greenberg's index B, which explains the use subscript B in the notation. Superscript n stands for native languages.

Instead of $d(k, C)$, one can also consider a dichotomous distance measure, in which case $d(k, C)$ is equal to zero if k is included in set C of core languages, and equal to one otherwise. This leads to a simplified version of the previous index:

$$DIS_A^n(C) = \sum_{k \notin C} s_k,$$

where summation is taken over all linguistic groups whose native language is not included in C. Again, subscript A is used since this is a formulation that is close to Greenberg's A index.

We now move to the other alternative that takes account of the entire linguistic repertoires instead of native languages only. Here we can no longer rely on the notion of linguistic groups identified by their unique native language since individuals with the same native language may speak different non-native languages. We thus need to identify the set of language that every individual i speaks and partition each of the K linguistic groups into clusters of individuals with identical linguistic repertoires. As in our earlier discussion in Section 5 on communication distances with two languages, French and German, the group of French speakers can be divided into two clusters, unilingual French speakers and bilingual speakers of French and German. German speakers are divided into two clusters as well. Thus, in total, a society with two languages would consist of three clusters: unilingual French and unilingual German speakers, and all bilingual fellows.[44]

Denote the collection of all clusters in the society by Q, where for each cluster $q \in Q$, q identifies the set of languages spoken by the members of that cluster. We now need to define the distance between all linguistic clusters q and the set of core languages C by finding the shortest linguistic distance among all possible pairs of languages, when one language is in q and another is in C:

$$d(q, C) = \min_{k \in q, l \in C} d(k, l).$$

This leads to two additional disenfranchisement indices that are related to all languages spoken by individuals and not only to native ones. The first is:

$$DIS_B^a(C) = \sum_{q \in Q} s_q d(q, C),$$

where s_q denotes the population share of cluster q and superscript a stands for *all* languages and not only native ones and replaces superscript n used

[44]If English is added, the population of French speakers can be divided into four clusters: unilingual French, bilingual French and German (who do not speak English), bilingual French and English (who do not speak German), and trilingual French, German and English speakers. A similar partition applies to German speakers, and in total there will exist six clusters: unilingual French and unilingual German speakers, bilingual French and German, bilingual French and English, bilingual German and English, and all trilingual speakers.

in the first index. $DIS_B^a(C)$ represents the average societal linguistic distance between the set of language spoken in the society and the set of core languages.

The second is a simplified version that obtains if we choose dichotomous distances. Here the distance $d(q, C)$ is equal to zero if there is at least one common language between the linguistic repertoire of the cluster q and the set C of core languages, and one if this intersection is empty, so that:

$$DIS_A^a(C) = \sum_q s_q$$

where the summation is all over those clusters that have no common language with the set of core languages C.

6.2.2 Applications

We show how the various indices can be used to judge the impact on disenfranchisement of various choices of core languages. The example used is the European Union with its 23 (without Croatian) official languages that are supposed to be given equal treatment. Reality, including in the Parliament, is however very different, as described by Wright (2007, 2009). It is therefore unavoidable that at some point, the EU will have to consider (or admit that it decided to implement) a certain degree of linguistic standardization.

The effects on disenfranchisement of reducing the number of languages are based on Fidrmuc, Ginsburgh and Weber (2007) who formulate a procedure for selecting subsets of languages among all eligible official languages in order to minimize the EU-wide disenfranchisement index, which measures the share of citizens (in the EU as a whole, but also in each member country) who would be unable to communicate under a particular restricted set of languages.[45] Data are based on a survey carried out in 2005 all EU countries (Special Eurobarometer 243, 2006), including Bulgaria and Romania, that were candidates but not yet member states.

Table 5 exhibits disenfranchisement indices (using index $DIS_A^n(C)$ where each set C consist of a unique language) in each EU country for the seven most spoken languages in the EU. Results show that though English is, as

[45] In our definition of disenfranchisement, a citizen is considered disenfranchised in a language (a) if he does not speak it (that is, if he does not cite it among the languages that he "knows") or (b) if, when asked how proficient he is in a language that he "knows" he responds that his knowledge is only basic.

expected, the most widely spoken language, it would nevertheless disenfranchise 62.6 percent of EU citizens if it were the only core language; moreover, it would disenfranchises more than 50 percent of the populations in 20 out of the 27 member states. Any other language (German, French, Italian, Spanish, Polish or Dutch) would do worse, both in the EU as a whole, and in most individual countries.

[Insert Table 5 approximately here]

We now address the question of whether a subset of official languages could do better. The procedure used selects the sequence of subsets that minimize disenfranchisement in the EU as a whole, for every given number of languages. Let m take the values 1, 2, 3, ..., 23. Then, for every m, denote by T_m the subset of the 23 languages that minimizes the disenfranchisement index over all sets with m languages, ending up with a set T_m for every m between 1 and 23.[46]

Results of these computations are reproduced in Table 6, using index $DIS_A^a(C)$ where the sequence of sets C increases by one additional language at each step. Each column indicates the language that should be added to the subset formed by the languages reported in the preceding columns in order to minimize EU's disenfranchisement index. The optimal one language set is English. For two languages, the optimal set contains English and German, and so on.[47]

[Insert Table 6 approximately here]

The marginal contribution of each additional language to reducing disenfranchisement falls under one percent of the EU population once the number of languages exceeds 13 and the differences between marginal contributions attributable to further candidate languages are often minute. To save space, only the first 11 languages are reported.

English is clearly the first language in any sequence as it is spoken well or very well by one third of the EU population. German and French are in close race for the second position; German, with a 49.3 percent disenfranchisement

[46]Details of the procedure are described in Fidrmuc, Ginsburgh and Weber (2007).

[47]Note that there are instances where two languages result in approximately the same reduction in disenfranchisement at a particular step in the sequence. For example, the tenth language could be Czech or Greek.

index, fares slightly better than French with 50.6 percent. The bundle of three languages leads to a disenfranchisement index of 37.8 percent. Italian, Spanish or Polish would each make almost the same contribution to reducing disenfranchisement further. Spanish, in turn, performs only marginally better than Polish. With the six largest languages included, 16 percent of the EU population would still remain disenfranchised. Adding Romanian brings the residual disenfranchisement index further down to 13 percent. Of course, important differences across countries remain. The most dramatic case is Hungary, where only 16 percent of the population can speak one of the first seven languages. Not surprisingly, Hungarian becomes the eighth language in the sequence. This also has a positive impact on Slovakia whose disenfranchisement index declines from 70 to 57 percent. Portuguese is the ninth language, followed by Czech and Greek tied in tenth position.

The disenfranchisement indices in Table 6 are a snapshot of the situation at the time of the survey (end of 2005). Fidrmuc, Ginsburgh and Weber (2007) also calculate a sequence of optimal sets based on the disenfranchisement indices of the youngest generation (15 to 29 years old) only. They show that German which was second to enter in Table 6, would be replaced by French. This is essentially due to the fact that among the younger generation in Germany and in Austria, 60 percent of the population knows English, so that German becomes less necessary. Beyond the first three languages, the sequence is essentially the same as before. With ten languages, the disenfranchisement index that would prevail is 3.9 percent. This percentage is even likely to decrease further as more and more children in upper secondary education study languages (essentially English, but to some extent, also French and German).

Table 7 reports results in which lexicostatistical distances (defined in Section 3.1) are accounted using index $DIS_B^a(C)$. In the single-language (English-only) scenario, accounting for linguistic proximity of other languages reduces the EU-wide disenfranchisement considerably, from 62.6 to 43.1 percent. French which now comes in as second reduces disenfranchisement in all Romance-language countries, bringing the EU-wide index to 24 percent, and Polish is third. These are deviations with respect to the two sequences discussed earlier: French and Polish beat German (which is at close distance of English) that becomes fourth. Italian is the fifth language followed by Hungarian and Spanish. Greek ties with Romanian for the eighth position. Nine core languages would be sufficient to decrease (the distance-adjusted) disenfranchisement to 2.9 percent.

[Insert Table 7 approximately here]

Ginsburgh, Ortuño-Ortín, and Weber (2005) take the examination of this issue a step further (though they used an older survey taken at a time Poland was not yet a member of the EU). They look at optimal sets of official languages, which are determined by two parameters. One is the society's sensitivity towards language disenfranchisement of its members. The other is the degree of comprehensiveness of its *language regime*, which can take any intermediate form between the following two polar cases. Under *full interpretation* all documents and discussions in meetings are translated into all languages, whereas under *minimal interpretation*, all documents are translated into one core language. In practice, the language regime is chosen somewhere between these two extremes. The society's objective is, as above, to find a set of languages that minimizes a weighted sum of total EU disenfranchisement and costs. The weight attached to total disenfranchisement represents society's "sensitivity" towards the linguistic concerns of its citizens. If the sensitivity parameter is high, then society cares about disenfranchisement of its citizens, and will implement a large number of (maybe even all) core languages. If sensitivity is low and cost considerations become more important, the society would shrink the set of core languages. They run simulations with different values of the weights given to the two parameters. These simulations show that it could be unwise to select English as unique working language, not only because it is not always optimal, but also because it is optimal only for very small values of the weight which represents sensitivity to disenfranchisement. The best choice of three languages is English, French and German, though Italian could be a very reasonable substitute to French.[48] Spanish is obviously not a good choice within the Union if no account is taken of Mexico and Latin America, and its growing importance in the South and the West of the United States. The authors argue that it might be reasonable for the EU to adopt four working languages, three of which (English, French and German) for general use, while Spanish would be added for its importance in the rest of the world.[49]

The sequences of sets which minimize EU's global index of disenfranchisement can be used to simulate the political feasibility of linguistic reforms and

[48]The 2004 enlargement to countries of the former Eastern Block have improved the situation of German with respect to French.

[49]See also Pool (1996) for a different approach of the problem.

have the European Council (or the Parliament) casting votes on their preferred set.[50]

[50]See Fidrmuc, Ginsburgh and Weber (2009).

7 References

D. Akchurina, D. Davydov, D. Krutikov, A. Khazanov and S. Weber (2015) 'Measurement of diversity: Theory and socio-economic applications', *Mathematical Methods in Economics* 62(2) 2-7.

A. Alesina, R. Baqir, and W. Easterly (1999) 'Public goods and ethnic divisions', *Quarterly Journal of Economics*, 114, 1243-1284.

A. Alesina, A. Devleeschouwer, W. Easterly, S. Kurlat and R. Wacziarg (2003) 'Fractionalization', *Journal of Economic Growth*, 8, 155-194.

A. Alesina, S. Michalopoulos and E. Papaioannou (2012) Ethnic inequality, NBER Working Paper 18512.

A. Alesina and E. Zhuravskaya (2008) Segregation and the quality of government in a cross-section of countries, NBER Working Paper 14316.

A. Annett (2001) 'Social fractionalization, political instability, and the size of the government', *IMF Staff Papers*, 46, 561-592.

Q. Ashraf and O. Galor (2013) 'The "Out of Africa" hypothesis, human genetic diversity, and comparative economic development', *American Economic Review* , 103, 1-46.

J. Aston, D. Buck, J. Coleman, C. Cotter, N. Jones, V. Macaulay, N. MacLeod, J. Moriarty and A. Nevins (2012) 'Phylogenetic inference for function-valued traits: speech sound evolution', *Trends in Ecology and Evolution*, 2, 160-166.

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.

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

A. Bisin and T. Verdier (2000) 'Beyond the melting pot: Cultural transmission, marriage and the evolution of ethnic and religious traits', *Quarterly Journal of Economics*, 115, 955-988.

A. Bisin and T. Verdier (2014) 'Trade and cultural diversity' In V. Ginsburgh and D. Throsby (eds.) *Handbooks in Economics. Art and Culture* (Amsterdam: Elsevier).

W. Bossert, C. d'Ambrosio and E. La Ferrara (2011) 'A generalized index of fractionalization', *Economica*, 78, 723-750.

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

B. Chiswick and P. Miller (2007b) *The Economics of Language, International Analyses* (London and New York: Routledge).

J. Church and I. King (1993) 'Bilingualism and network externalities', *Canadian Journal of Economics*, 26, 337-345.

CIA World Fact Book (2000) available at https://www.cia.gov/news-information/press-releases-statements/press-release-archive-2001/pr09182001.html

P. Collier (2001) 'Implications of ethnic diversity', *Economic Policy*, 16, 129-155.

D. Crystal (2003) *English as a Global Language* (Cambridge: Cambridge University Press).

K. Desmet, I. Ortunño-Ortín and S.Weber (2009) 'Linguistic diversity and redistribution', *Journal of the European Economic Association*, 7, 12911318.

K. Desmet, I. Ortuño-Ortín and R. Wacziarg (2012) The political economy of linguistic cleavages,*Journal of Development Economics*, 97, 322-338.

M. Dryer and M. Haspelmath (eds.) (2013) *The World Atlas of Language Structures Online* (Leipzig: Max Planck Institute for Evolutionary Anthropology) (Available online at http://wals.info, accessed on July 27, 2014.)

I. Dyen (1965) 'A lexicostatistical classification of the Austronesian languages', *International Journal of American Linguistics*, Memoir 19.

I. Dyen, J. Kruskal, and P. Black (1992) 'An Indo-European classification: A lexicostatistical experiment', *Transactions of the American Philosophical Society*, 82 (Philadelphia: American Philosophical Society).

W. Easterly and R. Levine (1997) 'Africa's growth tragedy: policies and ethnic divisions', *Quarterly Journal of Economics*, 112, 1203-1250.

Ethnologue (2009) *Languages of the World*, M. Paul Lewis (ed.) (Dallas, TX: SIL International).

O. Falck, S. Heblich, A. Lameli, and J. Södekum (2009) Dialects, cultural identity, and economic exchange, IZA Working Paper 4743.

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

J. Fearon and D. 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.

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

J. Fishman (1968) 'Some contrasts between linguistically homogeneous and linguistically heterogeneous polities' In J. Fishman, C. Ferguson, and J. Dasgupta (eds.) *Language Problems of Developing Nations* (New York: Wiley).

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

R. Florida and G. Gates (2001) Technology and tolerance: the importance of diversity to high-tech growth, Brookings Institute Discussion Paper.

J. Gabszewicz, V. Ginsburgh and S. Weber (2011) 'Bilingualism and communicative benefits', *Annals of Economics and Statistics*,101/102, 271-286.

T. Gamkrelidze and V. Ivanov (1990) 'The early history of Indo-European languages', *Scientific American*, March, 82-89.

V. Ganne and M. Minon (1992) 'Géographies de la traduction' In Françoise Barret-Ducrocq (ed.) *Traduire l'Europe* (Paris: Payot).

C. Gini (1912/1955) 'Variabilità e mutabilità', *Studi Economico-Giuridici della R. Universitadi Cagliari*, 3, 3-159. Reprinted in Enrico Pizetti and Tomasso Salvemini (eds.) *Memorie di metodologica statistica* (Roma: Libraria Eredi Virilio Vechi).

V. Ginsburgh, I. Ortuño-Ortin and S. Weber (2005) 'Disenfranchisement in linguistically diverse societies. The case of the European Union', *Journal of the European Economic Association*, 3, 946-964

V. Ginsburgh, I. Ortuño-Ortin and S. Weber (2007) 'Learning foreign languages. Theoretical and empirical implications of the Selten and Pool model', *Journal of Economic Behavior and Organization*, 64, 337-347.

V. Ginsburgh, J. Melitz and F. Toubal (2014) Foreign language learning: An econometric analysis, Manuscript.

V. Ginsburgh and S. Weber (2005) 'Language Disenfranchisement in the European Union', *Journal of Common Market Studies*, 43, 273-286.

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

V. Ginsburgh, S. Weber and S. Weyers (2011) 'The economics of literary translation. A simple theory and evidence' *Poetics*, 39, 228-246.

H. Goebl (1982) 'Dialektometrie, Prinzipen und Methoden des Einsatzes der Numerischen Taxonomie im Bereich der Dialektgeographie' *Denkschriften der Österreichischen Akademie der Wissenschaften, philosophisch-historische Klasse*, 157, 1-123 (Wien: Verlag der Österreich-ischen Akademie der Wissenschaften).

R. Gray and Q. Atkinson (2003) 'Language-tree divergence times support the Anatolian theory of Indo-European origin', *Nature*, 426, 435-439.

J. Greenberg (1956) 'The measurement of linguistic diversity', *Language*, 32, 109-115.

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

J. Greenberg (2000) *Indo-European and Its Closest Relatives* (Stanford, CA: Stanford University Press).

C. Guilmoto and F. Sandron (2001) 'The international dynamics of migrations networks in developing countries', *Population: An English Selection*, 13, 135-164.

E. Gunnemark (1991) *Countries, Peoples, and Their Languages: The Linguistic Handbook* (Gothenburg, Sweden: Lanstryckeriet).

P. Hadjipantelis, J. Aston and J. 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.

C. Hagège (2009) *Dictionnaire amoureux des langues* (Paris: Plon and Odile Jacob).

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

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

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

J. Heilbron (1999) 'Towards a sociology of translation. Book translations as a cultural world system', *European Journal of Social Theory*, 2, 429-444.

P. Heggarty, A. McMahon and R. McMahon (2005) 'From phonetic similarity to dialect classification: A principled approach' In N. Delbecque, D. Geeraerts and J. van der Auwera (eds.) *Perspectives in Variation: Sociolinguistic, Historical, Comparative* (Amsterdam: Mouton de Gruyter).

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

M. Hill (1973) 'Diversity and evenness: a unifying notation and its consequences', *Ecology*, 54, 427-432.

C. Hoskins, S. McFadyen and A. Finn (1997) *Global Television and Film* (Oxford: Clarendon Press).

W. Hutchinson (2003) Linguistic distance as determinant of bilateral trade, Working Paper 01-W30R, Department of Economics, Vanderbilt University.

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

K. Kawai (1950) '*Mokusatsu*, Japan's response to the Potsdam declaration', *Pacific Historical Review*, 19, 409-414.

P. Kerswill (2006) 'Migration and language' In K. Mattheier, U. Ammon and P. Trudgill (eds.) *Sociolinguistics. An International Handbook of the Science of Language and Society* (Berlin: De Gruyter, Second edition).

B. Kessler (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).

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

D. Kibbee (2003) 'Language policy and linguistic theory' In Jacques Maurais and Michal Morris (eds.) *Languages in a Globalising World* (Cambridge: Cambridge University Press).

H. Ku and A. Zussman (2010) 'The role of English in international trade', *Journal of Economic Behavior & Organization*, 75, 250-260.

P. Ladefoged, and I. Maddieson (1996) *The Sounds of the World's Languages* (Oxford: Blackwell).

D. Laitin (2000) 'What is a language community?', *American Journal of Political Science*, 44, 142-155.

R. La Porta, F. Lopez de Silanes, A. Shleifer and R. Vishny (1999) 'The quality of government', *Journal of Law, Economics and Organization*, 15, 222-279.

E. Lazear (1999) 'Globalization and the market for team-mates', *Economic Journal*, 109, 15-40.

V. Levenshtein (1966) 'Binary codes capable of correcting deletions, insertions, and reversals', *Cybernetics and Control Theory*, 10, 707-710.

D. Levinsohn (1998) *Ethnic Groups Worldwide. A Ready Reference Handbook* (Phoenix: Oryx Press).

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

B. Lian and J. 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.

S. Lieberson (1964) 'An extension of Greenberg's linguistic diversity measures' , *Language*, 40, 526-531.

S. Lieberson (1969) 'Measuring population diversity', *American Sociological Review*, 34, 850-862.

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

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

J. McWorther (2001) *The Power of Babel* (New York, NY: Perennial Harper).

J. Melitz (2007) 'The impact of English dominance on literature and welfare' *Journal of Economic Behavior and Organization*, 64, 193-215.

J. Melitz (2008) 'Language and foreign trade', *European Economic Review*, 52, 667-699.

S. Michalopoulos (2012) 'The origins of linguistic diversity', *American Economic Review*, 102, 1508-1539.

Minority Rights Group International (1998) *World Directory of Minorities* (London: Minority Rights Group International).

L. Nakhleh, T. Warnow, D. Ringe and S. Evans (2005) 'A comparison of phylogenetic reconstruction methods of an Indo-European dataset', *Transactions of the Philological Society*, 103, 171-192.

M. Nei and Wen-Hsiung Li (1979) 'Mathematical model for studying genetic variation in terms of restriction endonucleases' *Proceedings of the US National Academy of Sciences*, 76, 5269-5273.

J. Nerbonne and W. Heeringa (1997) 'Measuring dialect difference phonetically' In John Coleman (ed.) *Worshop on Computational Phonology* (Madrid: Special Interest Group of the Association for Computational Linguistics).

D. Nettle (2000) 'Linguistic fragmentation and the wealth of nations: The Fishman-Pool hypothesis reexamined', *Economic Development and Cultural Change*, 48, 335-348.

J. Nichols (2012) 'Monogenesis or polygenesis: a single ancestral language for all humanity' In M. Tallerman and K. Gibson (eds.) *The Oxford Handbook of Language Evolution* (Oxford: Oxford University Press).

G. Ottaviano, and G. Peri (2005) 'Cities and cultures', *Journal of Urban Economics*, 58, 304-337.

L. Penrose (1946) 'The elementary statistics of majority voting', *Journal of the Royal Statistical Society*, 109, 53-57.

E. Poloni, O. Semino, G. Passarino, S. Santachiara-Benerecetti, I. Dupanloup, A. Langaney and L. Excoffier (1997), 'Human genetic affinities for Y-chromosome P49a,f: TaqI haplotypes show strong correspondences with linguistics', *American Journal of Human Genetics*, 61, 1015-1035.

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

J. Pool (1996) 'Optimal language regimes for the European Union', *International Journal of the Sociology of Languages*,121, 159-179.

D. Posner (2004) 'Measuring ethnic fractionalization in Africa', *American Journal of Political Science*, 48, 849-863.

C. Rao (1982) 'Diversity and dissimilarity coefficients: a unified approach', *Theoretical Population Biology*, 21, 24-43.

C. Ricotta and L. Szeidl (2006) 'Towards a unified approach to diversity measures: bridging the gap between the Shannon diversity and Rao's quadratic entropy', *Theoretical Population Biology*, 70, 237-243.

M. Ruhlen(1991) *A Guide to World's Languages. Classification* (London: Edward Arnold).

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

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

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

D. Searls (2003) 'Linguistics: Trees of life and language', *Nature*, 426, 391-392.

R. Selten and J. Pool (1991) 'The distribution of foreign language skills as a game equilibrium' In R. Selten (ed.) *Game Equilibrium Models*, vol. 4 (Berlin: Springer-Verlag).

C. Shannon (1948) 'A mathematical theory of communication', *Bell Systems Technical Journal*, 27, 379-423 and 623-656.

I. Singh (2005) *The History of English* (Cambridge: Hodder Arnold Publications).

E. Simpson (1949) 'Measurement of diversity', *Nature*, 163, 688.

Special Eurobarometer 243 (2006) *Europeans and their languages* (Brussels: European Commission).

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

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

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

J. Walker (2000) 'Money laundering: Quantifying international patters', *Australian Social Monitor*, 2, 139-147.

T. Warnow (1997) 'Mathematical approaches to comparative linguistics', *Proceedings of the National Academy of Sciences of the USA*, 94, 6585-6590.

S. Wright (2007) 'English in the European Parliament: MEPs and their language repertories', *Sociolinguistica*, 21, 151-165.

S. Wright (2009) 'The elephant in the room: Language issues in the European Union', *European Journal of Language Policy*, 1, 93-119.

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

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

(m) is for masculine

Table 2. Distances Between Selected Indo-European Languages
(value times 1,000)

	English	French	German	Italian	Spanish	Polish
Albanian	883	878	870	877	883	871
Bulgarian	772	791	769	769	782	369
Catalan	777	286	764	236	270	784
Czech	759	769	741	753	760	234
Danish	407	759	293	737	750	749
Dutch	392	756	162	740	742	769
English	0	764	422	753	760	761
French	764	0	756	197	291	781
German	422	756	0	735	747	781
Greek	838	843	812	822	833	837
Icelandic	454	772	409	755	763	758
Italian	753	197	735	0	212	764
Latvian	803	793	800	782	794	668
Lithuanian	784	779	776	758	770	639
Norwegian	452	770	367	754	761	762
Polish	761	781	754	764	772	0
Portuguese	760	291	753	227	126	776
Romanian	773	421	751	340	406	784
Russian	758	778	755	761	769	266
Serbo-Croatian	766	772	764	755	768	320
Slovak	750	765	742	749	756	222
Slovene	751	782	733	760	772	367
Spanish	760	291	747	212	0	772
Swedish	411	756	305	741	747	763
Ukrainian	777	781	759	774	782	198

Source: Dyen, Kruskal and Black (1992, pp. 102-117).

Table 3. Simplified Indo-European Language Tree

0. Ur-language
    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. Scores of Foreign Students Learning English

Score	Language of origin
3.00	Afrikaans, Norwegian, Rumanian, Swedish
2.75	Dutch, Malay, Swahili
2.50	French, Italian, Portuguese
2.25	Danish, German, Spanish, Russian
2.00	Amharic, Bulgarian, Cambodian, Czech, Dari, Farsi, Finnish, Hebrew, Hungarian, Indonesian, Mongolian, Polish, Serbo-Croatian, Tagalog, Thai, Turkish
1.75	Bengali, Burmese, Greek, Hindi, Nepali, Sinhala
1.50	Arabic, Lao, Mandarin, Vietnamese
1.25	Cantonese
1.00	Japanese, Korean

Source: Chiswick and Miller (2007a, p. 578).

Table 5. Disenfranchisement Indices in the EU
Results for Unique Core Languages

	English	German	French	Italian	Spanish	Polish	Dutch
Austria	55	1	94	95	98	100	100
Belgium	59	87	29	97	97	99	32
Bulgaria	84	94	96	99	99	100	100
Cyprus	49	98	95	99	99	100	100
Czech R.	84	81	98	100	100	98	100
Denmark	34	73	97	99	98	100	100
Estonia	75	92	100	100	100	100	100
Finland	69	95	99	100	100	100	100
France	80	95	1	95	93	100	100
Germany	62	1	92	99	98	98	100
Greece	68	94	95	98	100	100	100
Hungary	92	91	100	99	100	100	100
Ireland	1	98	91	100	99	99	100
Italy	75	96	90	3	97	100	100
Latvia	85	97	100	100	100	99	100
Lithuania	86	96	99	100	100	87	100
Luxembourg	61	12	11	95	99	100	99
Malta	32	99	95	65	99	100	100
Netherlands	23	43	81	100	97	100	1
Poland	82	90	99	99	100	2	100
Portugal	85	98	91	99	96	100	100
Romania	86	97	90	98	99	100	100
Slovak R.	83	82	99	100	100	98	100
Slovenia	59	79	98	91	99	100	100
Spain	84	98	94	99	2	100	100
Sweden	33	88	97	99	99	100	100
Un. Kingdom	1	98	91	99	98	100	100
EU	62.6	75.1	80.1	86.7	88.9	91.6	95.1

Source: Fidrmuc, Ginsburgh and Weber (2007).

Table 6. Disenfranchisement Indices in the EU
Optimal Sequence of Subsets of Core Languages

Number Languages	1 EN	2 GE	3 FR	4 IT	5 SP	6 PL	7 RO	8 HU	9 PT	10a CZ	10b GR	11 CZ&GR
Austria	55	0	0	0	0	0	0	0	0	0	0	0
Belgium	59	56	18	18	18	18	18	18	18	18	18	18
Bulgaria	84	81	79	79	78	78	78	78	78	77	77	77
Cyprus	49	49	49	48	48	48	48	48	48	48	0	0
Czech R.	84	69	69	69	69	67	67	66	66	0	66	0
Denmark	34	31	31	31	31	30	30	30	30	30	30	30
Estonia	75	70	70	70	70	69	69	69	69	69	69	69
Finland	69	67	67	67	67	67	67	67	67	67	67	67
France	80	77	1	0	0	0	0	0	0	0	0	0
Germany	62	1	1	1	1	1	1	1	1	1	1	1
Greece	68	64	63	63	63	63	63	63	63	63	0	0
Hungary	92	85	85	85	85	85	84	0	0	0	0	0
Ireland	1	1	1	1	1	1	1	1	1	1	1	1
Italy	75	74	69	1	1	1	1	1	1	1	1	1
Latvia	85	83	83	83	83	82	82	82	82	82	82	82
Lithuania	86	82	82	82	82	71	71	71	71	71	71	71
Luxembourg	61	8	1	1	1	1	1	1	1	1	1	1
Malta	32	31	31	31	31	31	31	31	31	31	31	31
Netherlands	23	18	18	18	18	18	18	18	18	18	18	18
Poland	82	77	76	76	76	1	1	1	1	1	1	1
Portugal	85	84	81	81	79	79	79	79	0	0	0	0
Romania	86	85	81	80	79	79	1	1	1	1	1	1
Slovak R.	83	72	72	72	72	70	70	57	57	44	57	44
Slovenia	59	50	50	45	45	45	45	45	45	45	45	45
Spain	84	84	81	80	1	1	1	1	1	1	1	1
Sweden	33	33	33	33	33	33	33	33	33	33	33	33
Un. Kingdom	1	1	1	1	1	1	1	1	1	1	1	1
EU	62.6	49.3	37.8	29.5	22.4	16.4	12.9	10.9	9.2	7.7	7.7	6.2

Source: Fidrmuc, Ginsburgh and Weber (2007). One language is added to the previous ones in each column. In columns 10a, and 10b, two languages result in the same percentage reduction in disenfranchisement. In column 11, they are both added to the set. Languages are abbreviated as follows: Czech (CZ), English (EN), French (FR), German (GE), Greek (GR), Hungarian (HU), Italian (IT), Spanish (SP), Polish (PL), Portuguese (PT), Romanian (RO).

Table 7. Disenfranchisement Distance-ajusted Indices in the EU
Optimal Sequence of Subsets of Core Languages

Number Languages	1 EN	2 FR	3 PL	4 GE	5 IT	6 HU	7 SP	8a GR	8b RO	9 GR&RO	10a CZ	10b FI	10c BG
Austria	23	23	23	0	0	0	0	0	0	0	0	0	0
Belgium	33	8	8	3	3	3	3	3	3	3	3	3	3
Bulgaria	64	62	29	28	28	28	28	28	28	28	20	33	2
Cyprus	41	40	40	39	39	39	39	0	39	0	0	0	0
Czech R.	59	58	19	16	16	15	15	15	15	15	0	14	14
Denmark	14	14	13	9	9	9	9	9	9	9	9	9	9
Estonia	60	60	35	34	34	28	28	28	28	28	15	11	15
Finland	65	65	65	64	64	45	45	45	45	45	45	0	45
France	60	0	0	0	0	0	0	0	0	0	0	0	0
Germany	26	26	24	0	0	0	0	0	0	0	0	0	0
Greece	55	54	53	50	50	50	50	0	50	0	0	0	0
Hungary	88	87	86	84	84	0	0	0	0	0	0	0	0
Ireland	1	1	1	1	1	1	1	1	1	1	1	1	1
Italy	57	15	15	14	1	1	1	1	1	1	1	1	1
Latvia	65	64	27	27	27	27	27	27	27	27	8	8	8
Lituania	64	64	27	26	26	26	26	26	26	26	13	13	13
Luxemburg	28	3	3	0	0	0	0	0	0	0	0	0	0
Malta	31	31	31	31	30	30	30	30	30	30	30	30	30
Netherlands	9	9	9	3	3	3	3	3	3	3	3	3	3
Poland	61	60	1	1	1	1	1	1	0	0	0	0	0
Portugal	64	24	24	24	18	18	10	10	10	10	10	10	10
Romania	66	35	35	34	28	26	25	25	1	1	1	1	1
Slovak R.	59	59	19	17	17	13	13	13	13	13	3	10	10
Slovenia	41	39	20	17	16	16	16	16	16	16	15	16	16
Spain	64	22	22	22	18	18	1	1	1	1	1	1	1
Sweden	14	14	14	10	10	10	10	10	10	10	10	10	10
UK	1	1	1	1	1	1	1	1	1	1	1	1	1
EU	43.1	24.0	16.6	11.4	9.0	6.9	5.2	4.0	4.1	2.9	2.1	2.1	2.1

Source: Fidrmuc, Ginsburgh and Weber (2007). One language is added to the previous ones in each column. In columns 8a, and 8b, two languages result in the same percentage reduction in disenfranchisement. In column 9, they are both added to the set. In columns 10a, 10b and 10c, three languages compete for the 10th place. Languages are abbreviated as follows: Bulgarian (BG), Czech (CZ), English (EN), French (FR), German (GE), Greek (GR), Finnish (FI), Hungarian (HU), Italian (IT), Spanish (SP), Polish(PL), Romanian (RO).