2016/24

Ranking Languages in the European Union

Victor Ginsburgh, Juan D. Moreno-Ternero
and Shlomo Weber

50 YEARS OF
CORE
DISCUSSION PAPERS

CORE
Voie du Roman Pays 34, L1.03.01
Tel (32 10) 47 43 04
Fax (32 10) 47 43 01
Email: immaq-library@uclouvain.be
http://www.uclouvain.be/en-44508.html

Ranking languages in the European Union[1]

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

Juan D. Moreno-Ternero
Department of Economics, Universidad Pablo de Olavide, Seville, Spain,
and CORE, Université catholique de Louvain, Belgium

Shlomo Weber
Department of Economics, Southern Methodist University, USA,
and the New Economic School, Moscow, Russia

June 17, 2016

Abstract

This article presents a stylized framework to rank languages in multilingual societies. We consider several ranking methods, reflecting principles such as *minimal disenfranchisement*, *communicative benefits*, or *utilitarianism*, as well as game-theory-based rankings referring to the *Shapley Value*. We use data from the Special Eurobarometer survey in order to apply these methods to rank languages within the European Union. Although the methods largely differ on their normative grounds, they lead to very close results.

[1]We are very grateful to Israel Zang for comments.

1 Introduction

Multilingualism is a pervasive phenomenon that can be traced back to ancient times, as reflected by the story of the Babel tower. If people were members of small language communities, it was necessary to know two or more languages for trade or any other dealings outside one's own town or village (Holden, 2016). In more recent times, globalization and cultural openness is fostering multilingualism. A point in case is Europe, where there is no predominant language. English is often used as a communication language, but in multilingual countries such as Belgium (French, Dutch and German), Switzerland (German, French, Italian and Romanche) or Spain (Spanish, Catalan, Basque and Galician) it is common to see employees mastering two or even three of those languages. Some languages such as Danish, Swedish and Norwegian are close enough that it is generally more common to use their mother tongue rather than English in meetings.

In many situations, and mostly for practical reasons, it is necessary to select a language, or a group of languages. This could be the case for the choice of official languages, or for selecting the language under which a conversation among individuals with different mother tongues takes place. In this paper, we present a stylized framework to rank languages in multilingual societies. We are interested in ranking all (feasible and infeasible) languages rather than selecting a subset (including only one). The reason is simply that, depending on the context, we might be interested in different subsets and numbers of languages. This is particularly the case of the European Union.

We consider several ranking methods, reflecting focal principles. One is the principle of *Minimal Disenfranchisement*, which focuses on the number of individuals who are excluded by imposing one or several languages, because they hardly or do not know them. This is the principle studied by Ginsburgh and Weber (2005) and Ginsburgh et al. (2005), with a special emphasis on the European Union. It is equivalent to Van Parijs' *Minimex* option, which minimizes exclusion. As Van Parijs (2011) puts it, suppose you have to address simultaneously a set of people who each know to various extents a couple of languages and by all of whom you wish to be understood. Then you should use the language that is best known by the member of your audience who knows it least.

Another principle is that of *Communicative Benefits*, originally introduced by Selten and Pool (1991).[2] The principle states that a language is "superior" to another if it allows more pairs of individuals to communicate perfectly.[3] It is not difficult to show that this

[2]See also Ginsburgh et al., (2007) and Athanasiou et al., (2016).

[3]Without the adjective *perfectly*, the principle is equivalent to that of *Minimal Disenfranchisement*.

is equivalent to state that a language offers a higher number of speakers with perfect knowledge than another.

The two principles focus on the extreme cases of minimal and maximal language knowledge. The *Aggregate Knowledge* principle would allow to extend this to intermediate cases of language knowledge and would state that a language is preferred to another if its aggregate knowledge (weighing by individual degrees) is higher than that of another language. This is essentially the principle underlying *Utilitarianism*, a deeply rooted notion in economics and philosophy, which can be traced back to Jeremy Bentham and John Stuart Mill.

The three principles stated above would trivially lead to the same ranking under the premise of dichotomous knowledge of languages, in which agents either speak or do not speak a certain language. In the more general setting in which intermediate knowledge levels are also considered the principles may yield different rankings, as we show in Section 2.

We also consider two game-theory-based rankings using the *Shapley Value*, the best-known solution concept for cooperative games. Suppose the issue is to allocate the proceeds from a joint venture among the members of a group of agents. The *Shapley Value* implements the idea of focusing on the marginal productivity of each agent to the venture, and aggregating across agents. This concept relates to language rankings as follows. Suppose each agent speaks a set of languages that endorses with equal probability, or with probabilities that are proportional to the levels of his knowledge. Then, adding these probabilities across agents produces a scoring for each language and, thus, a ranking.

In what follows, we explore the different normative grounds of the five methods described above. Some of them echo voting procedures, such as *Approval Voting* and game-theory concepts, such as the *Shapley Value*. They are described in more detail in Section 2. We then use data from a Special Eurobarometer survey taken in 2006 to apply these methods to rank languages spoken in the European Union. Our empirical illustration surprisingly shows that, in spite of their different normative grounds, all rankings lead to very close results. This indicates that the ranking of languages within the European Union is more robust than what might be thought and incorporates several interesting axiomatic characteristics.

2 Defining and analyzing rankings

We describe the linguistic reality of a given society by a *language matrix*, A, whose rows refer to citizens (*agents*), and its columns to *languages*. Each entry a_{ij} of the matrix denotes the knowledge level of language j by agent i. We make the normalizing assumption that $a_{ij} = 0$ reflects no knowledge whatsoever, $a_{ij} = 1$ reflects perfect knowledge, and also assume that there exists a finite set of intermediate levels of knowledge ranging from 0 to 1; here is an example of such a matrix.

Example 1

$$A = \begin{pmatrix} 0 & 1/2 & 1/2 \\ 0 & 1/2 & 1/2 \\ 0 & 1/2 & 1/2 \\ 1 & 0 & 1/2 \\ 1 & 0 & 1/2 \\ 0 & 1 & 0 \\ 0 & 1 & 1/2 \\ 1 & 1/2 & 1/2 \end{pmatrix}.$$

This matrix refers to the case in which there are eight agents and three languages. Agents 1, 2 and 3 only have intermediate knowledge of languages 2 and 3, which we model by the number 1/2, and no knowledge of language 1. Agents 4 and 5 have maximal (perfect) knowledge of 1, intermediate knowledge of 3 and no knowledge of 2. Agent 6 only has maximal knowledge of 2, and does not know any other. Agent 7 has maximal knowledge of 2, intermediate knowledge of 3 and no knowledge of 1. Finally, agent 8 has maximal knowledge of 1, and intermediate knowledge of the other two languages.

2.1 Definitions

Let N denote the set of agents, and L the set of languages. For each pair of languages $j, k \in L$, we say that j dominates k if and only if

$$\sum_{i \in N} a_{ij} > \sum_{i \in N} a_{ik}.$$

We are interested in two situations.

In the first, we *count* speakers, whether they speak the language perfectly or not, as long as they speak it. This means that we are not interested in the speaking quality of the language: all positive entries in A are replaced by ones. Thus, in Example 1, language 1 has 3 speakers, language 2 has 6 and language 3 has 7. More generally, we define the *Minimal Disenfranchisement* (MD) ranking as the one that ranks languages according to the number of agents who *do not* know them, or know them but only *very superficially* (they are disenfranchised). The higher the number of disenfranchised agents, the lower the ranking of the language. Therefore, MD ranks language 3 first (7 speakers), then language 2 (6 speakers) and finally, language 1 (3 speakers).

In the other situation, the elements a_{ij} that describe the degree of knowledge of a language can be bounded from below by imposing $a_{ij} \geq \underline{a}_i$ and $a_{ik} \geq \underline{a}_k$, where both $\underline{a}_i$ and $\underline{a}_k$ take values between 0 and 1. If, say, both bounds are set to 1, we are only interested in perfect speakers, count only those who have a perfect knowledge of both languages. In the above example, language 1 would have 3 speakers, language 2, 2 speakers, and language 3 no speaker. More generally, we define the *Communicative Benefits* (CB) ranking as the one that ranks languages according to the number of agents who have perfect knowledge of each language. This is equivalent to rank languages according to the number of pairs that could communicate perfectly in each language.

If both $\underline{a}_i$ and $\underline{a}_k$ are set to $\varepsilon > 0$ sufficiently small, we also include agents who know the language, even if their knowledge is not perfect. In this case, and for the above example, language 1 would score 3, language 2 would score 4 and language 3 would score 3.5. Thus, language 2 would be first, language 3 would be second and language 1 would be last. More generally, we define the *Aggregate Knowledge* ranking (AK), as the one that ranks languages according to their aggregate knowledge level across society.

The first two rankings described above could be interpreted as translations to this context of *Approval Voting* (see Brams and Fishburn, 1978). This voting method allows each voter to cast his or her vote for as many candidates he or she wishes; each positive vote is counted in favour of the candidate. These votes are then added as usual, and the winner is the candidate who gets the largest number of votes. All other candidates can also be ranked, according to the number of votes they obtain. The *Minimal Disenfranchisement* ranking arises when an agent *approves* all languages with partial knowledge ($a_{ij} > 0$), whereas the *Communicative Benefits* ranking arises when an agent only *approves* all languages with perfect knowledge ($a_{ij} = 1$).

The third ranking can also be interpreted as a generalization of *Approval Voting*, known as *Range Voting* (see Smith, 2004) or *Evaluative Voting* (see Hillinger, 2005).[4] In

[4]The difference between the voting methods actually lies on whether alternatives (knowledge levels in our setting) can take any value in $[0, 1]$ or only a finite number of values.

that voting method, agents can express further evaluations of alternatives (exemplified in numbers between 0 and 1) beyond just approving (or disapproving) them.

An alternative to Approval Voting is *Cumulative Voting* (Glasser, 1959; Sawyer and MacRae, 1962). It allows individuals to distribute points among candidates in any arbitrary way. An interesting case is the one in which every individual is endowed with a fixed number of points that are evenly divided among all candidates for whom the individual votes. This would translate into our context as the ranking introduced next.

For each $i \in N$, let $L_i(A)$ denote the set of languages over which i has at least some knowledge, i.e., $L_i(A) = \{j \in L : a_{ij} \neq 0\}$ and let $l_i(A)$ denote its cardinality. Then, for each pair of languages $j, k \in L$, we say that j S-dominates k if and only if

$$\sum_{i \in N, j \in L_i(A)} \frac{1}{l_i(A)} > \sum_{i \in N, k \in L_i(A)} \frac{1}{l_i(A)}.$$

We refer to the resulting ranking of languages as the **Shapley** (S) ranking.[5] Now, consider the following matrix:

Example 2

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{pmatrix}.$$

In Example 2, there are only three agents and four languages. Agent 1 only knows language 1, while the two other agents know languages 2, 3 and 4. Then, as stated above, the first three ranking methods discussed earlier would endorse the ranking in which languages 2, 3 and 4 come first, and language 1 comes last. The Shapley ranking inverts the ranking. Indeed, summing over individuals shows that language 1 would be ranked first, while each of the three others would obtain 2/3 and be ranked after language 1.

In order to interpret Shapley rankings, one could imagine that each agent is endowed with one vote which is shared among the candidates for whom he or she votes. Vote shares for each candidate are then added as above, and the candidate who gets the largest number of shares wins. All other candidates are ranked accordingly. As shown by Ginsburgh and Zang (2003), these numbers represent the *Shapley Values*, which could also be interpreted as the *power* of each candidate in the voting situation. The difference with respect to

[5]The name of the ranking comes from the rule introduced by Ginsburgh and Zang (2003) for the so-called museum pass game. Bergantiños and Moreno-Ternero (2015) extend the rule to more general museum pass problems. See also Ginsburgh and Zang (2013), who apply the rule for the ranking of wines.

Approval Voting is that each voter casts the same number of votes and has therefore the same rights. This is not so in Approval Voting, since voters can cast as many votes as they wish.

To conclude with the inventory of rankings, one could consider a reasonable generalization of the Shapley ranking, in which agents distribute points among languages proportionally to their knowledge levels. Then, for each pair of languages $j, k \in L$, j WS-dominates k if and only if

$$\sum_{i \in N, j \in L} \frac{a_{ij}}{\sum_{h \in L} a_{ih}} > \sum_{i \in N, k \in L} \frac{a_{ik}}{\sum_{h \in L} a_{ih}}.$$

We refer to the resulting ranking of languages as the *Weighted Shapley* (WS) ranking. Obviously, the last two rankings would coincide for examples in which knowledge is only maximal or minimal (such as in Example 2). Consider now the following matrix with four agents and three languages:

Example 3

$$A = \begin{pmatrix} 1 & 1/4 & 3/4 \\ 0 & 1/4 & 1 \\ 0 & 1 & 0 \\ 1/4 & 0 & 0 \end{pmatrix}.$$

Agent 1 speaks language 1 perfectly, and languages 2 and 3 partially (the last, better than the second). Agent 2 speaks language 2 partially and language 3 perfectly. Agent 3 only speaks language 2 (and she does so perfectly). Finally, agent 4 only speaks partially language 1.

The Shapley (S) scorings are as follows: Language 1 gets 1/3 from agent 1 and 1 from agent 4 which totals 4/3. Language 2 gets 1/3 from agent 1, 1/2 from agent 2, and 1 from agent 3 which totals 11/6. Language 3 gets 1/3 from agent 1 and 1/2 from agent 2 which totals 5/6. Therefore, language 2 would come first, followed by language 1 and then language 3.

As for the Weighted Shapley (WS) scorings, they would come as follows: Language 1 gets 1/2 from agent 1 and 1 from agent 4 which totals 3/2. Language 2 gets 1/8 from agent 1, 1/5 from agent 2, and 1 from agent 3 which totals 53/40. Language 3 gets 3/8 from agent 1 and 4/5 from agent 2 which totals 47/40. This produces the following ranking: language 1 would come first, followed by 2 and then 3.

2.2 A normative analysis

We now turn to an important issue, known as normative analysis, whose purpose is to characterize each ranking by a small number of reasonable assumptions (axioms) and which, possibly, imply uniqueness of the ranking under scrutiny. We first list and comment a certain number of such axioms defining *Minimal Disenfranchisement* (MD), *Communicative Benefits* (CB) and *Aggregate Knowledge* (AK) rankings, and follow with three propositions which characterize the rankings.

Axiom 1 Low Invariance. The ranking is not altered if non-perfect knowledge levels are replaced by no-knowledge levels, that is if every $a_{ij} < 1$ in matrix A is replaced by 0.

Axiom 2 High Invariance. The ranking is not altered if positive non-perfect knowledge levels are replaced by perfect knowledge level, that is if every $0 < a_{ij} < 1$ in matrix A is replaced by 1.

Axiom 3 Compensation. If all *but two* agents in a society have the same knowledge of two languages j and j', with the exception of agents i and i', and if i knows language j perfectly and has no knowledge of j', while the reverse holds for i', then both languages j and j' are ranked at the same level.[6]

Axiom 3' Generalized Compensation. Let the set of knowledge levels of every language and every agent be finite. If agent i increases his knowledge of language j to an immediate higher level, while agent i' decreases his knowledge of j to the immediate lower level, the ranking of languages does not change.[7]

Axiom 4 (Strong) Pareto Optimality. If all agents speak language j at least as well as language j', with at least one agent speaking j strictly better than j', then j is strictly ranked above j'.

Here are the propositions concerning rankings MD, CB and AK.[8]

Proposition 1 *MD is the unique ranking that satisfies Axioms 1, 3 and 4.*

Proposition 2 *CB is the unique ranking that satisfies Axioms 2, 3 and 4.*

Proposition 3 *AK is the unique ranking that satisfies Axioms 3' and 4.*

[6]This is a weaker version of an axiom recently introduced by Macé (2015) who uses the same terminology, and will be discussed later.

[7]This is the axiom recently introduced by Macé (2015) under the term *Compensation*.

[8]Appendix 1 contains more formal statements of the axioms, as well as the proofs of the propositions.

We chose to focus on our specific case to provide new characterization results for the first three rankings in our analysis. Note that some of the many existing characterizations of approval voting (see, for instance, Xu, 2010) can be used for the *Minimal Disenfranchisement* and *Communicative Benefits* rankings. Likewise, some of the many existing characterizations of *utilitarian* criteria (see, for instance, Blackorby et al., 2002) can be used for the *Aggregate Knowledge* ranking. However, such a route could not be taken to derive normative foundations for the *Shapley* (S) or *Weighted Shapley* (WS) rankings since we are not aware of any axiomatic characterization of (Equal and Even) Cumulative Voting.[9]

In order to provide normative foundations for the last two rankings, we consider a slightly different context. More precisely, suppose now that each agent has a *vote* to be allocated among the languages he speaks. The ranking would then be generated from the overall number of votes allocated to each language. Following Bergantiños and Moreno-Ternero (2015) we consider the next three axioms for such a context.

Axiom 5 Equal Treatment of Equals. If two languages have the same number of speakers, then they should receive the same number of votes.
Axiom 6 Dummy. If no agent speaks a language, then it should get no votes.
Axiom 7 Additivity. Given two groups of speakers, it is equivalent to consider them separately, or as the same group.

This leads to Proposition 4, which can be stated as follows:[10]

Proposition 4 *S is the unique ranking satisfying Axioms 5, 6 and 7.*

Such a result is thus equivalent to the seminal characterization of the *Shapley value* for TU-games (e.g., Shapley, 1953).[11] The Shapley value is the best-known solution concept for those games, which model the attempt to predict the allocation of resources in multi-

[9]Both Approval Voting and Cumulative Voting can be seen as members of a family of voting procedures dubbed as *Size Approval Voting*, which are characterized by Alcalde-Unzu and Vorsatz (2009).

[10]See Appendix 1 for the formal proof.

[11]A coalitional game is cooperative if the players can make binding agreements about the distribution of payoffs or the choice of strategies, even if these agreements are not specified or implied by the rules of the game. Transferable-utility games (in short, TU-games) are one category of cooperative games in which one specifies a function that associates with each nonempty coalition a real number indicating the worth of the coalition. If a coalition forms, then it can divide its worth in any possible way among its members. A comprehensive analysis of TU-games can be found, for instance, in Peleg and Sudholter (2007).

person interactions (see Winter, 2002). It is remarkable not only for its attractive and intuitive definition but also for its unique characterization by a set of reasonable axioms. In addition, the value is also viewed as an index for measuring the power of players (here, languages) in a game (see Shapley and Shubik, 1954). The value uses averages (or weighted averages in some of its generalizations) to aggregate the power of players in their various cooperation opportunities.

One of the axioms characterizing the Shapley value is *Equal Treatment of Equals*. This is a compelling axiom when information is limited. For some specific applications, however, we might possess more information about the environment, which motivates the asymmetry of the solution concept.[12] This interpretation led to the concept of *Weighted Shapley Value*, which has also been characterized in the literature, replacing the axiom *Equal Treatment of Equals* by *Partnership Consistency*, which refers to the treatment of players who can only generate value together (see Kalai and Samet, 1987). We consider another related alternative to *Equal Treatment of Equals*, which leads to the characterization of WS.

Axiom 5' Weighted Treatment of Equals. If a speaker speaks two languages, then he should allocate vote shares among them that are proportional to their knowledge levels.

Proposition 5. *WS is the unique ranking satisfying Axioms 5', 6 and 7.*[13]

3 Languages in the European Union

We now use the stylized framework described in Section 2 to rank the main languages used in the European Union (EU). We shall distinguish official languages (see below) from other languages that are spoken, but are not official.

Official languages. In 1958, the Treaty of Rome and Regulation 1/1958 recognized four languages Dutch (NL), French (F), German (G) and Italian (I) as official languages.[14] Danish (DK), English (GB), Finnish (FIN), Greek (GR), Portuguese (P), Spanish (SP) and Swedish (SW) were added later. The 2004 enlargement to Eastern Europe resulted in adding Czech (CZ), Estonian (EST), Hungarian (H), Latvian (LV), Lithuanian (LT), Maltese (M), Polish (PL), Slovak (SL), and Slovenian (SLO). Irish (IRL) was given the

[12] In our context, this is so if we allow for different knowledge levels of languages.

[13] See Appendix 1 for the formal proof.

[14] Though the Treaty included six countries, the three languages that are spoken in Belgium (French, Dutch and German) are also spoken in other partner countries (France, the Netherlands and Germany); Luxembourg agreed not to include Luxembourgish (LX).

same status in 2005 but it was agreed that the decision would be implemented only as of January 2007. We also included Bulgarian (BG) and Romanian (RO) which became official in 2007 only, but were already included in the survey to be discussed below.[15] All these languages, listed as *Official* in Table 1 enjoy the same privileges as the original four.

Other languages. Table 1 also includes seven other languages that are used in EU countries, but are not official. Russian is spoken in many former Eastern Bloc countries; Basque, Catalan and Galician are spoken in Spain, and Luxembourgish in Luxembourg; Arabic and Turkish are languages mainly spoken by immigrants.[16]

To determine who speaks what, we use the Special Eurobarometer 243 (2006) survey carried out in November 2005 in all member countries of the European Union, including Bulgaria and Romania (that were not yet members in 2005). In most countries, 1,000 citizens were interviewed, with the exception of Germany (1,500), the United Kingdom (1,300), Cyprus (500), Luxembourg (500) and Malta (500). The total number of usable interviews amounts to 26,700. Among the technical specifications, it is worth mentioning the following that were implemented:

> "[...] the survey covers the national population of citizens [...] that are residents in those countries and have a sufficient command of one of the respective national language(s) to answer the questionnaire. The basic sample design applied in all states is a multi-stage, random (probability) one. In each country, a number of sampling points was drawn with probability proportional to population size (for a total coverage of the country) and to population density.
>
> For each country a comparison between the sample and the universe was carried out. The universe description was derived from Eurostat population data or from national statistics offices. For all countries surveyed, a national weighting procedure, using marginal and intercellular weighting, was carried out based on this universe description. In all countries, gender, age, region and size of locality were introduced in the iteration procedure."

[15] Croatia and Turkey were also candidates to accessing the EU, but this happened much later for Croatia, and did not happen for Turkey. Croatian is thus not included in our list. Turkish is, but only as a non-official language, spoken by migrants, and as a co-official language in Cyprus.

[16] Turkish is also, as mentioned above, a co-official language in Cyprus, which is a country included in our survey.

The data that we use are taken from the answers to the following questions:

(a) D48a. What is your mother tongue? (do not probe – do not read out – multiple answers possible). Follows a list of 34 languages that include the 23 member states' official languages, as well as Arabic, Catalan, Chinese, Croatian, Luxembourgish, Russian, Turkish, Basque, Galician, Other regional languages, Other.

(b) D48b to D48d. Which languages do you speak well enough in order to be able to have a conversation, excluding your mother tongue? (do note probe – do not read out – multiple answers possible). Follows the same list of 34 languages. This question was asked for first, second and third foreign languages.

(c) D48f. Is your (language cited in 48b, 48c and 48d) very good, good, basic? (show card with scale).

To illustrate the models described in Section 2, we code as 1 the responses *mother tongue* (question D48a), and as 1/2 the responses to other languages known (questions D48b to D48d), combined with the self-evaluation of knowledge *very good* and *good* (question D48f). All other responses, that is *basic*, or *not knowing the language* are coded as 0.

Table 1 summarizes our findings. Note that we separate Russian, Arabic, Turkish, Catalan, Basque, Galician as well as Luxembourgish, which appear at the bottom of the table, as they are not official.[17] These results are obtained by considering the whole EU as a unique country. Since the number of units surveyed are not proportional to the populations of the 27 countries included, we had to weigh the numbers of each country by its population.[18]

In the table, column (1) contains the name of the language and column (2) the countries where it is official. The remaining columns (3) to (7) contain the following rankings:

[17] With the caveat made before about Turkish in Cyprus.

[18] More precisely, the population (in millions) we used for each country was the following: Austria 8,2; Belgium 10,4; Bulgaria 7,8; Cyprus 0,7; Czech Rep. 10,2; Denmark 5,4; Estonia 1,3; Finland 5,2; France 60,6; Germany 82,5; Great Britain 60; Greece 11,1; Hungaria 10,1; Ireland 4,1; Italy 58,5; Latvia 2,3; Lithuania 3,4; Luxembourg 0,5; Malta 0,4; Netherlands 16,3; Poland 38,2; Portugal 10,5; Romania 21,7; Slovakia 5,4; Slovenia 2; Spain, 43; Sweden 9.

Minimal Disenfranchisement (MD), *Communicative Benefits* (CB), *Aggregate Knowledge* (AK), as well as the *Shapley* ranking (S) and the *Weighted Shapley* ranking (WS).[19]

(a) in column (3), the numbers represent the shares of the total EU population that do not know the language or whose knowledge is basic; for instance, the value 63.0 that appears in the first row of column (3), means that 63 percent of the EU population do not know or has only a basic knowledge of English.[20]

(b) in column (4), the numbers represent the shares of the total EU population which do not speak the language as mother tongue; the value 87.3, which appears in the first row of column (4), means that 87.3 percent of the EU population do not have English as mother tongue.

(c) in column (5), the numbers are less obvious to be interpreted in a simple way, as they result from adding knowledge levels (a_{ij}) of which some are equal to 1, others are equal to 1/2.

(d) the numbers in columns (6) and (7) that contain the rankings using *Shapley Values* have no meaning, other than giving a measure of the power of a language; what really counts here is the order, and to some extent, the differences between languages. One could surmise that English and Swedish are far away from each other, but there is no difference between Swedish and Bulgarian.

The main lesson to be drawn from our empirical analysis is that the rankings of official languages produced are extremely similar, in spite of the theoretical differences and distinct normative grounds of the ranking methods discussed above. Minor differences still occur that deserve some comments.

(a) The *Communicative Benefits* ranking, in column (4), points to a couple of differences with respect to the *Minimal Disenfranchisement* ranking in column (3). German is the most obvious one as it is the language with the highest number of native speakers, but also a language that less non-native individuals speak (even at intermediate levels), compared to English or French. Another difference refers to Slovak, which does worse than a few other languages (e.g., Danish, Finnish, Lituanian or Slovenian) under *Minimal Disenfranchisement*. The reason is that Slovak is also spoken

[19]The numbers that appear in column (3) guided the ordering in which the languages appear in the table. This is of course arbitrary, as any other column could have been chosen as well.

[20]The 100.0 number appearing in this column, as well as the next ones, are a consequence of rounding.

in the Czech Republic, but those who speak it there do not consider it a native language.[21] Also, Hungarian and Greek switch their positions in both rankings, but this might just be a consequence of rounding.

(b) The only difference between the *Aggregate Knowledge* ranking in column (5) and the *Minimal Disenfranchisement* ranking in column (3), is the position of Finnish, which does slightly worse than Lithuanian, Slovenian, Latvian, and Estonian in the AK ranking.

(c) For the *Shapley* ranking, the change in the order between Dutch and Romanian is due to (i) the fact that the number of Netherlanders and Flemish (in Belgium) who speak foreign languages (English, German and French essentially) is large, which decreases the Shapley Value of Dutch and (ii) most Romanians only speak Romanian, which increases the Shapley Value of Romanian. The change in order between Danish and Finnish is due to identical reasons. Many Danes know German and/or English, while Finns are less proficient in foreign languages (with the exception of Swedish in the Western part of the country).

Russian is an interesting case as it belongs to the ten most important languages in the European Union. This is, of course, due to the Russian influence in Eastern Europe after Word War II. The forces in action nowadays may well decrease the knowledge of Russian among the younger generations, but there are also forces in the other direction, resulting from increased trade relations with Russia, that may foster the learning of Russian.

4 Further insights

In this paper, we provide a stylized framework to analyze the ranking of languages in multilingual societies. We introduce five ranking methods. Three of these reflect appealing principles such as minimizing exclusion (MD), maximizing communication benefits (CB), or aggregating knowledge levels (AK). The two last (S) and (WS) are inspired by game-theory tools used in cooperative situations. We explore the normative foundations for each method and apply them to rank languages in the European Union. The results are remarkably similar in spite of the different nature (and normative grounds) of the methods. Some differences occur (in particular, the relative positions of Dutch and Romanian and of

[21] This may be a consequence of the breakup in 1993 of Czechoslovakia into two sovereign states, the Czech Republic and Slovakia.

Danish and Finnish, as well as the role of Russian, which is not a EU language) for which we provided explanations. Nevertheless, the main lesson to be drawn from our analysis is that the ranking of official languages within the European Union is quite robust.

To conclude, it is worth mentioning that our work could shed some light on a potential additional effect of *Brexit*, which has largely been ignored.[22]

In 2005, the EU had some 490 million citizens. 182.5 of them were reasonably fluent in English. German follows with 128.5 million, French (100.7 million), Italian (68.4 million), Spanish (57.2 million) and Polish (43.5 million).[23]

Brexit will entail a loss of 60 million British natives speakers and here and there a few native English speakers who also speak other official EU languages, since these will no longer be considered as EU citizens if *Brexit* happens. If one only considers the number of speakers who will still be members of the EU after *Brexit*, the number of English speakers remains large (122.5 million). Nevertheless, German will turn out to be the most widely spoken language with 127.4 million EU citizens.

English is also heavily used in reports and rules set by EU bureaucrats. In 2008, 72.5 percent of the first versions of administrative and legislative documents were written in English, while 14 and 3 percent only were written in French and German, respectively. The situation has even become more skewed since English now makes for some 82.5 percent, while German dropped to 2 percent. To top things off, in 2014, 261.000 pages have been translated from other languages into English, 155.000 into French and 136.000 into German. This means that the EU essentially writes in English and translates into English. Will this still be the case, if English becomes the official language of 4.1 million Irish citizens only, who are also the "owners" of Gaelic, another official language in the EU?

5 References

Athanasiou, E., Moreno-Ternero, J.D., and Weber, S. (2016) Language learning and communicative benefits. In V. Ginsburgh and S. Weber (eds.), The Palgrave Handbook of Economics and Language, London: Palgrave Macmillan, 212-229.

Alcalde-Unzu, J., and Vorsatz, M. (2009) Size approval voting. Journal of Economic

[22] The following data come from our analysis above, as well as from Fidrmuc, Ginsburgh and Weber (2007).

[23] These numbers include native speakers, as well as non natives who acquired the language.

Theory 144, 1187-1210.

Bergantiños, G., and Moreno-Ternero, J.D. (2015) The axiomatic approach to the problem of sharing the revenue from museum passes. Games and Economic Behavior 89, 78-92.

Blackorby, C., Bossert, W., and Donaldson, D. (2002) Utilitarianism and the theory of justice,. In K. J. Arrow, A. Sen, and K. Suzumura (eds.), The Handbook of Social Choice and Welfare, Amsterdam: Elsevier, 543-596.

Brams, S., and Fishburn, P. (1978) Approval voting. American Political Science Review 72, 831-847.

Fidrmuc, J., Ginsburgh, V. and Weber, S. (2007), Ever Closer Union or Babylonian Discord? The Official-language Problem in the European Union, CEPR Discussion Paper 6397.

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

Ginsburgh, V., Ortuño-Ortin, I., and Weber S. (2005) Language disenfranchisement in linguistically diverse societies. The case of European Union. Journal of the European Economic Association 3, 946-965.

Ginsburgh, V., and Zang, I. (2003) The museum pass game and its value. Games and Economic Behavior 43, 322-325.

Ginsburgh, V., and Zang, I. (2012) Shapley ranking of wines. Journal of Wine Economics 7, 169-180.

Glasser, G. (1959) Game theory and cumulative voting for corporate directors. Management Science 5, 151-156.

Hillinger, C. (2005) The case for utilitarian voting. Homo Oeconomicus 23, 295-321.

Holden, N. (2016) Economic exchange and business languages in the ancient world: an exploratory review. In V. Ginsburgh and S. Weber (eds.), The Palgrave Handbook of Economics and Language, London: Palgrave Macmillan, 290-311.

Kalai, E., and Samet, D. (1987) On Weighted Shapley Values. International Journal of Game Theory 16, 205-222.

Macé, A. (2015) Voting with evaluations: When should we sum? What should we sum? AMSE Working Paper

Van Parijs, P. (2011) Linguistic Justice for Europe and for the World. Oxford: Oxford University Press.

Peleg, B., and Suholter, P. (2007) Introduction to the Theory of Cooperative Games. Springer

Sawyer, J., and MacRae, D. (1962) Game theory and cumulative voting in Illinois: 1902-1954. American Political Science Review 56, 936-946.

Selten, R., and Pool, J. (1991) The distribution of foreign language skills as a game equilibrium. In Selten, R. (ed.). Game Equilibrium Models, Berlin: Springer-Verlag, 64-84.

Shapley, L. (1953) A value for n-person games. In H.W. Kuhn and A.W. Tucker (eds.), Contributions to the Theory of Games II (Annals of Mathematics Studies 28), Princeton: Princeton University Press, 307-317.

Shapley, L., and Shubik, M. (1954) A method for evaluating the distribution of power in a committee system. American Political Science Review 48, 787-792.

Special Eurobarometer 243, 2006.

Winter, E. (2002) The Shapley value. In R. Aumann, and S. Hart (eds.) Handbook of Game Theory, Volume 3. Elsevier, pp 2025-2054.

Xu, Y. (2010) Axiomatizations of approval voting. In J.-F. Laslier, and R. Sanver (eds.), Handbook on Approval Voting. Studies in Choice and Welfare, Berlin Heidelberg: Springer-Verlag, 91-102

6 Appendix 1

We provide in this appendix a formal treatment to the contents of Section 2.2.

For each language matrix A, we construct two matrices associated to it: A^L and A^H. Formally, for each $(i,j) \in N \times L$,

$$a_{ij}^L = \begin{cases} 0 & \text{if } a_{ij} < 1 \\ 1 & \text{otherwise} \end{cases}$$

and

$$a_{ij}^H = \begin{cases} 1 & \text{if } a_{ij} > 0 \\ 0 & \text{otherwise} \end{cases}$$

Let $\succsim_A$ denote the ranking of languages associated to A. Our axioms are formally stated as follows:

- *Low Invariance*: $\succsim_A \equiv \succsim_{A^L}$.
- *High Invariance*: $\succsim_A \equiv \succsim_{A^H}$.
- *Compensation*: Let $l, k \in L$ be such that $a_{il} = 1 = a_{jk}$, $a_{jl} = 0 = a_{ik}$, and $a_{hl} = a_{hk}$, for all $h \in N \setminus \{i,j\}$. Then, $l \sim_A k$.
- *Strong Pareto*: Let $l, k \in L$ be such that $a_{il} \geq a_{ik}$, for each $i \in N$, with at least one strict inequality. Then, $l \succ_A k$.

Proposition 1 *MD is the unique ranking satisfying Low Invariance, Compensation and Strong Pareto.*

Proof. Let $A = (a_{ij})_{(i,j)\in N\times L}$ be a language matrix, and $l, k \in L$ be a pair of languages. By *Low Invariance*, $l \succsim_A k \iff l \succsim_{A^L} k$. By iterated application of *Compensation*, if necessary, we obtain that $l \sim_{A^L} k \iff \sum a_{il}^L = \sum a_{ik}^L$. By *Strong Pareto*, $l \succ_{A^L} k \iff \sum a_{il}^L > \sum a_{ik}^L$. Altogether, we have that $l \succsim_A k \iff MD(l) \geq MD(k)$. ■

Proposition 2 *CB is the unique ranking satisfying High Invariance, Compensation and Strong Pareto.*

Proof. Let $A = (a_{ij})_{(i,j)\in N\times L}$ be a language matrix, and $l, k \in L$ be a pair of languages. By *High Invariance*, $l \succsim_A k \iff l \succsim_{A^H} k$. By iterated application of *Compensation*, if necessary, we obtain that $l \sim_{A^H} k \iff \sum a_{il}^H = \sum a_{ik}^H$. By *Strong Pareto*, $l \succ_{A^H} k \iff \sum a_{il}^H > \sum a_{ik}^H$. Altogether, we have that $l \succsim_A k \iff CB(l) \geq CB(k)$ ■

In order to define the next axiom, suppose now that $a_{ij} \in \{a_1, a_2, \ldots, a_n\}$, where $a_1 < a_2 < \cdots < a_n$.

- *Generalized Compensation*: Let $l, k \in L$ be such that $a_{il} = a_m = a_{jk}$, $a_{jl} = a_{m-1} = a_{ik}$, for some $m \in \{2, \ldots, n\}$, and $a_{hl} = a_{hk}$, for all $h \in N \setminus \{i, j\}$. Then, $l \sim_A k$.

Proposition 3 *AK is the unique ranking satisfying Generalized Compensation and Strong Pareto.*

Proof. Let $A = (a_{ij})_{(i,j)\in N\times L}$ be a language matrix, and $l, k \in L$ be a pair of languages. By iterated application of *Generalized Compensation*, if necessary, we obtain that $l \sim_A k \iff \sum a_{il} = \sum a_{ik}$. By *Strong Pareto* and *Generalized Compensation*, $l \succ_A k \iff \sum a_{il} > \sum a_{ik}$. Altogether, $l \succsim_A k \iff AK(l) \geq AK(k)$.[24] ■

Suppose now that each agent has a *vote* to be allocated among the languages he speaks. In this context, a **rule** is a mapping R that associates with each language matrix A an allocation $(R_l(A))_{l\in L}$ indicating the amount of votes each language gets. The Shapley rule allocates for each language $l \in L$ the following votes:

$$R_l^S(A) = \sum_{i\in N, j\in L_i(A)} \frac{1}{l_i(A)},$$

where recall that $L_i(A) = \{j \in L : a_{ij} \neq 0\}$ and $l_i(A)$ denotes its cardinality.

This rule obviously leads to the Shapley ranking we introduced above. As shown by the next result, which replicates Theorem 1 in Bergantiños and Moreno-Ternero (2015), the Shapley rule is characterized by the following three axioms:

- *Equal treatment of equals*: Let $l, k \in L$ be such that $N_l(A) = \{i \in N : a_{il} \neq 0\} = \{i \in N : a_{ik} \neq 0\} = N_k(A)$. Then, $R_l(A) = R_k(A)$.

- *Dummy*: Let $l \in L$ be such that $N_l(A) = \emptyset$. Then, $R_l(A) = 0$.

- *Additivity*: For each pair of subsets of agents (N^1, N^2), and their corresponding language matrices A^1 and A^2, $R(A^1 \cup A^2) = R(A^1) + R(A^2)$, where $A^1 \cup A^2$ denotes the resulting matrix from merging the rows of both matrices.

Proposition 4 *A rule satisfies equal treatment of equals, dummy and additivity if and only if it is the Shapley rule.*

[24] This proof essentially mimicks the proof of Theorem 3 in Mace (2015).

Proof. It is obvious that the Shapley rule satisfies the axioms. Let R be a rule satisfying the three axioms in the statement. For each $i \in N$, let A_i denote the corresponding i-th row of A. By *dummy*, $R_l(A_i) = 0$ for each $l \notin L_i(A)$. By *equal treatment of equals*, $R_l(A_i) = R_k(A_i)$ for each pair $l, k \in L_i(A)$. As $\sum\limits_{l \in L} R_l(A_i) = 1$, we deduce that $R_l(A_i) = \frac{1}{l_i(A)}$ for each $l \in L_i(A)$. Consequently, it follows, by *additivity*, that $R_l(A) = \sum_{i \in N, j \in L_i(A)} \frac{1}{l_i(A)}$, for each $l \in L$, as desired. ■

To conclude, we consider the Weighted Shapley rule, which allocates for each language $l \in L$ the following votes:

$$R_l^{WS}(A) = \sum_{i \in N, l \in L} \frac{a_{il}}{\sum_{h \in L} a_{ih}}.$$

As shown in the last proposition, the following alternative axiom to *equal treatment of equals* leads to characterize the Weighted Shapley rule.

- *Weighted treatment of equals*: Let $i \in N$ and $l, k \in L_i(A)$. Let A_i denote the corresponding i-th row of A. Then, $\frac{R_l(A_i)}{R_k(A_i)} = \frac{a_{il}}{a_{ik}}$.

Proposition 5 *A rule satisfies weighted treatment of equals, dummy and additivity if and only if it is the Weighted Shapley rule.*

Proof. It is obvious that the Weighted Shapley rule satisfies the axioms. Let R be a rule satisfying the three axioms in the statement. For each $i \in N$, let A_i denote the corresponding i-th row of A. By *dummy*, $R_l(A_i) = 0$ for each $l \notin L_i(A)$. By *weighted treatment of equals* $\frac{R_l(A_i)}{R_k(A_i)} = \frac{a_{il}}{a_{ik}}$ for each pair $l, k \in L_i(A)$. As $\sum\limits_{l \in L} R_l(A_i) = 1$, we deduce that $R_l(A_i) = \frac{a_{il}}{\sum_{h \in L} a_{ih}}$ for each $l \in L_i(A)$. Consequently, it follows, by *additivity*, that $R_l(A) = \sum_{i \in N, l \in L} \frac{a_{il}}{\sum_{h \in L} a_{ih}}$, for each $l \in L$, as desired. ■

Table 1. Rankings of Languages in the European Union

Languages		Rankings				
Name (1)	Native in (2)	MD (3)	CB (4)	AK (5)	S (6)	WS (7)
Official						
English	GB-IRL	63.0	87.3	75.4	106	96
German	G-A-B	75.1	82.4	78.7	78	54
French	F-B	80.2	87.3	84.0	64	43
Italian	I	86.7	88.3	88.5	49	35
Spanish	E	89.0	92.0	90.6	38	26
Polish	PL	91.6	92.9	92.0	32	22
Dutch	NL-B	95.1	95.0	95.3	12	10
Romanian	RO	95.5	95.8	95.5	19	28
Hungarian	H	97.3	97.7	97.5	11	8
Greek	GR-CY	97.3	97.6	97.5	10	7
Portuguese	P	97.5	97.8	97.5	10	7
Czech	CZ	97.5	98.0	97.7	8	6
Swedish	S	97.8	98.2	98.0	6	5
Bulgarian	BG	98.4	98.3	98.6	6	4
Slovak	SL	98.5	100.0	98.8	4	3
Danish	DK	98.7	99.0	98.8	3	3
Finnish	FIN	98.8	99.0	100.0	4	3
Lituanian	LT	99.3	99.3	99.4	2	2
Slovenian	SLO	99.4	99.6	99.6	1	1
Latvian	LV	99.6	99.7	99.6	1	1
Estonian	EST	99.7	99.9	99.8	<1	<1
Irish	IRL	99.8	100.0	99.8	<1	<1
Maltese	M	100.0	100.0	100.0	<1	<1
Others						
Russian	Non EU	95.3	99.3	97.3	10	6
Catalan	SP	98.8	99.3	99.0	2	2
Galician	SP	99.4	99.7	99.6	1	1
Arabic	Non EU	99.5	99.8	99.6	1	1
Turkish	CY	99.5	99.7	99.6	1	1
Basque	SP	99.7	99.9	99.8	<1	<1
Luxemb.	LX	99.9	100.0	99.9	<1	<1