Multi-Lingual Societies: Challenges and Policies*

Victor Ginsburgh, ECARES, Free Univ. of Brussels

Ignacio Ortuño-Ortín, Dep. of Economics, Univ. Carlos III, Madrid

Shlomo Weber, Dep. of Economics, SMU, and CORE, Catholic Univ. of Louvain

March 25, 2008

Abstract. We consider a linguistically diversified society and examine the notion of *language disenfranchisement* when some individuals are denied the full access to documents and political process in their native tongues. To calculate the disenfranchisement indices we use the Dyen *percentage cognate* matrix of linguistic distances between Indo-European languages and apply survey and population data on language proficiency in the European Union. We then determine optimal sets of official languages that depend on society's sensitivity against disenfranchisement and comprehensiveness of the chosen language regime.

Key words: Languages, Disenfranchisement, Dyen matrix, European Union.

JEL Classification Numbers: D70, O52, Z13.

*Acknowledgements: Ginsburgh, ECARES and CORE, Ortuño-Ortín, Dep. of Economics, Univ. Carlos III, Madrid, Weber, SMU, CORE and CEPR. We should like to thank Philippe Van Parijs for useful insights. We are grateful to the Alexander von Humboldt Foundation and Fundación BBVA for their financial support.

1 Introduction

The challenges of multilingual societies are well documented over the course of human history. The most famous example is the consequence of the attempt of the"people" to build a tower in Shinar (Babylonia) to be closer to the sky. God disliked the idea, descended and "confuse[d] their speach, so that one person will not understand another's speach. God scattered them all over the face of the earth, and they stopped building the city."[1] The difficulties in modern societies are by no means smaller, the main reason being that "like religion, language does not lend itself easily to compromise."[2]

In this paper we consider a model of a society where individuals are distinguished on the basis of their language characteristics. There is a set of existing languages and every member of the society is characterized by her language skills represented by all the languages she is proficient in. The problem faced by the society is to select a subset of languages to be used for translation of official documents, communication between institutions and citizens, debates in official bodies, etc.

The choice of *official* languages may have a major impact on the well-being of some individuals since it will limit their access to laws, rules and regulations. In some cases, these limitations could even violate the basic principles of the society. For example, in the European Union "citizens must be able to take part in building and maintaining the Union. They have a right to participate on equal terms and must have access to information and legal texts in their national languages that affect their lives."[3] Article 2.11 of the Amsterdam Treaty allows every citizen of the Union to use his native language in dealing with the official institutions of the EU. Non-inclusion of some

[1] *Genesis*, 11, 1-9.

[2] See Laponce (1992, p. 599-600).

[3] R. Schaerer (2002).

languages in the set of the official ones goes beyond restrictions on the access to information. It may also alienate groups of individuals whose cultural, societal and historical values and sensibilities are not represented by the official languages and consequently create "language disenfranchisement." In the context of the European Parliament, "the right of an elected Member to speak, read and write in his or her own language lies in the heart of Parliament's democratic legitimacy. The case for multi-lingualism is based not only on fairness to Members, from whichever country they are elected. It is necessary to ensure the support of citizens in all Member states; if Parliament does not recognize their language, it is less likely that citizens will recognize it as being *their* Parliament."[4]

However, the cost of services required to maintain a larger number of official languages[5] could be quite substantial. Even before the 2004 enlargement the institutions of the European Union were the largest recruiter of interpreters and translators in the world.[6] In 1999 the total translation and interpretation costs for the Commission alone amounted to 30% of its internal budget.[7] The basic principles of political accountability and equality among citizens require that all, or at least a substantial part of the full-fledged translation services, will have to be maintained in some of

[4]Report of Secretary General, document PE 305.269/ BUR/fin, 2001.

[5]In this paper we use the term "official language" rather than "working language", but the distinction between these two concepts in the EU is not very clear. For example, in their first ever passed Regulation in 1958, the Council of Ministers of the European Community establishes the distinction between the two but no definition is provided. See Mamadouh (1995, p. 4). Currently, the European Union has eleven official and working languages, Danish, Dutch, English, Finnish, French, German, Greek, Italian, Portuguese, Spanish, Swedish, and one treaty language, Irish.

[6]Cole and Cole (1997, p.59).

[7]De Swaan (2001, p. 172).

the EU institutions (Council of the European Union, European Council, European Parliament).[8] Moreover, a failure to provide translation services by the EU would simply shift the provision of the service to individual countries, leading to duplications that may raise the total cost of services,[9] as well as to divergent translations and interpretations.[10] The burden of maintaining official languages is not limited to direct costs of translation and interpretation. Communication[11] constitutes an even more serious challenge in societies with a large number of official languages. Translation and interpretation errors as well as the delays caused by translations, may end up paralyzing multilateral discussions and negotiations.[12] But more importantly, language is so much associated to local culture that large subsets of the population may become at best insensitive, at worst opposed to the political process. As Bretton (1976, p. 447) points out: "Language may be the most explosive issue universally and over time. This mainly because language alone, unlike all other concerns associated with nationalism and ethnocentrism ... is so closely tied to the individual self. Fear of being deprived of communicating skills seem to rise political passion to a fever pitch."

Unless the set of official languages includes all languages, a linguistically diversified society is bound to face some degree of language disenfranchisement. An important feature of our analysis is that an individual derives her degree of disenfranchisement over the set of official languages as a whole, rather than dissecting it into preferences over single languages and we define the preferences of every member of the society over all subsets of languages. This has important implications on

[8] Mamadouh (1995, p. 55-56), and Council of the European Union (2002).

[9] Mamadouh and Hofman (2001).

[10] Van Parijs (2003b).

[11] De Swaan (2001, p. 173).

[12] Mamadouh (1998, p. 8).

the selection of optimal sets of official languages. For example, there are more citizens in the EU who speak German than French. However, this fact alone does not necessarily support the choice of German over French as one of the official languages. Indeed, the number of EU citizens who speak both English and French is larger than the number of those who speak English and German. Thus, preferences over larger sets of languages, especially those including English, could be more relevant and informative than preferences over single languages.

We calculate disenfranchisement using two alternative methods. One is *dichotomous*: An individual is disenfranchised if she speaks no official language; she is not if she speaks at least one official language. This assumption can be challenged: If an individual does not speak any official language, some of them may have common roots with her native tongue that would reduce the degree of her disenfranchisement. Indeed, consider a citizen who speaks only Portuguese and compare her attitude towards two potential sets of official languages, containing respectively Spanish or German. Even though our Portuguese citizen speaks none of these, given the cultural and linguistic proximity of Portuguese and Spanish, the degree of her linguistic disenfranchisement will be lower if Spanish rather than German is chosen as one of official languages. This leads to what we call the *Dyen*[13] disenfranchisement index.

Both indices can be computed using two basic datasets on the number of people who do (or do not) speak languages. First, we compute disenfranchisement indices and optimal language sets using country populations and their native language only. To account for multilingual citizens, we also use surveys on language proficiency in each country.

We also examine optimal sets of official languages, which are determined by two parameters. One is the sensitivity of the society towards language disenfranchisement of its members. The other

[13]The term refers to Isidore Dyen who led the research for collecting the data and for computing such distances.

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. Under *minimal interpretation*, nothing is translated. In practice, the language regime is chosen somewhere between these two extremes. This is already the case in the European Union nowadays, where only some documents are translated into all official languages and the discussions (and translations) in several elected bodies are limited to a small number of languages.[14]

The paper is organized as follows. In Section 2, we discuss our model and introduce language disenfranchisement indices. Section 3 is devoted to the two important special cases of disenfranchisement indices based on distances between languages: *dichotomous* disenfranchisement and *Dyen* disenfranchisement. In Section 4, we compute our indices and show that they lead to similar results. The two main groups of European languages (Latin and Germanic) have to be represented in order to reduce disenfranchisement in the EU15. We extend our analysis to the Union after the enlargement using the population based disenfranchisement indices. In Section 5, we derive the optimal sets of official languages for different values of sensitivity towards language disenfranchisement and comprehensiveness of the language regime. We show that the introduction of the Dyen matrix of linguistic distances has a major impact on our results. In particular, it highlights the importance of Latin languages, such as French, Spanish or Italian. It may come as a surprise that the pair of two

[14]There are other examples. (a) In the European Court of Justice the language of the hearings is chosen by the defendant (among eleven official languages plus Irish) at beginning of the procedure, and the proceedings are translated into French, the permanent working language of the Court. (b) Even though the Commission publishes its official documents in all official languages, its internal working languages are French and English, and to lesser extent German. See Mamadouh (1998, p. 5). (c) It is often suggested to use some pivotal languages in the Parliament, to which and from which all other languages that are used in the Parliament will be translated.

major European languages, English and German, generate more disenfranchisement than Spanish and Dutch. The reason is that proximity dominates the effect of number of native speakers, as in the case of the linguistic closeness of French, Italian, and Portuguese to Spanish, and of Dutch to German. Section 6 contains some concluding remarks. The theoretical derivations utilized in Section 4 are presented in the Appendix.

2 The Model

We consider a *society* N that consists of n members,[15] who speak different languages from a given set $\mathcal{L} = \{1, \ldots, L\}$. For every individual $i \in N$ we denote by $P(i)$ the subset of languages in $\mathcal{L}$ spoken by i.[16]

Given a set of official languages T, those members of the society who speak no language from T, will be *disenfranchised*. However, an empty intersection of the sets $P(i)$ and T may be insufficient to determine the degree of disenfranchisement of individual i. As alluded to in the introduction, a unilingual Portuguese speaker who speaks neither German nor Spanish may prefer the set which contains Spanish. To account for this possibility, we introduce the distance function Γ, defined over pairs of subsets of languages, where $\Gamma(S, S')$ indicates how "linguistically close" the sets S and S' are. Thus, for every set of languages T, the value $\Gamma(P(i), T)$, the distance between the set of languages $P(i)$ spoken by i and the set T, will be considered as a degree of (individual) language disenfranchisement of individual i. Thus, if the set T is chosen as the set of official languages, the

[15] We use the word society here in order to encompass communities, regions, countries, continents, or any other political and geographical structure.

[16] We do not distinguish here between native and non-native tongues, and, more generally, we do not introduce the degree of command of a certain language, which is anyway very difficult to assess.

aggregate disenfranchisement index, $D^{\Gamma}(T)$, is defined by:

$$D^{\Gamma}(T) = \sum_{i \in N} \Gamma(P(i), T).$$

Note that for every distance function Γ, the disenfranchisement index D^{Γ} decreases if the set of official languages expands:

$$T \subset S \ \rightarrow \ D^{\Gamma}(T) > D^{\Gamma}(S),$$

where $T \subset S$ means that the set T is contained in the set S and is different from S. That is, a more inclusive set of official languages reduces disenfranchisement. Thus, if the reduction of disenfranchisement is the *only* goal of the society, the entire set of languages $\mathcal{L}$ would be the unambiguous choice. In this case only individuals who speak no language in $\mathcal{L}$ would contribute to disenfranchisement. However, cost considerations for maintaining official languages make the choice of the optimal set more complicated. Denote by $C(T)$ the cost of maintaining the set T of official languages and assume that the cost function increases if the set of official languages expands:

$$T \subset S \ \rightarrow \ C(T) < C(S).$$

Thus, there is a trade-off between disenfranchising citizens and the translation, interpretation and communication costs generated by a large number of languages. Formally, the society's objective is to find a set of languages T that minimizes the weighted sum of the total disenfranchisement index $D^{\Gamma}(T)$ and the cost $C(T)$:

$$\min_{T \subset \mathcal{L}} \alpha D^{\Gamma}(T) + C(T),$$

where the positive parameter α represents the society's "sensitivity" parameter attached to members' disenfranchisement.[17]

Let us turn to a brief examination of the cost function. There are cases in which the proper functioning of official institutions becomes impractical if too many languages are used. Imagine a meeting where every participant speaks her own language without being understood by the majority of other participants. This generates a cost function whose values are prohibitively high if the number of official languages exceeds a certain threshold. But even if this is not the case, the total cost of sustaining several languages depends on the nature of the language regime imposed by the society. Assume that there is a fixed cost c generated by translation, interpretation, communication, and printing of all documents between any two official languages and that there is a uniform stream of demands from all languages. Under a "full interpreting regime" that requires every important document to exist in all official languages, the total cost of sustaining k languages would be given by ck^2. If the society adopts a "minimal standard interpreting regime," that requires no translation into any other official language, the total cost of sustaining k languages will be ck. The society can also adopt an "intermediate standard interpreting regime," in which case the cost would take values ck^β, where $1 < \beta < 2$. To accommodate various language regimes, we assume that $C(T) = c|T|^\beta$, where $|T|$ stands for the cardinality of the set T, and the parameter β $(1 \leq \beta \leq 2)$ represents the degree of comprehensiveness of the language regime, including two polar cases $\beta = 1$ and $\beta = 2$.

[17] See also Grin (2003) who argues that there must be an optimum, since "it is reasonable to assume that the benefits of diversity increase at a decreasing rate, while its costs increase at an increasing rate." As one of the referees has suggested, along these lines one can introduce a societal objective function that is strictly concave, rather than linear, in $D^\Gamma(T)$. In our attempt to keep the analysis in this paper as simple as possible, we leave this extension to the future research.

Without loss of generality, we set $c = 1$. Then the society's problem is to choose T that solves

$$\min_{T \subset \mathcal{L}} G^{\Gamma}(T, \alpha, \beta), \tag{1}$$

where

$$G^{\Gamma}(T, \alpha, \beta) \equiv \alpha D^{\Gamma}(T) + |T|^{\beta}. \tag{2}$$

The following proposition is straightforward.

Proposition 1: (i) The function $G^{\Gamma}(T, \alpha, \cdot)$ is increasing in β for every T and α. That is, raising the standard of the interpreting regime increases society's language costs.

(ii) The function $G^{\Gamma}(T, \cdot, \beta)$ is increasing in α for every T and β.

For every α and β, let the solutions of (1) be denoted by $T^{\Gamma}(\alpha, \beta)$ and assume that they are well-defined. We have the following observation:

Proposition 2: There exists α^* such that $T^{\Gamma}(\alpha, \beta) = \mathcal{L}$ for every $1 \leq \beta \leq 2$ whenever $\alpha > \alpha^*$. That is, if the society exhibits a sufficiently high degree of intolerance to disenfranchisement, no language should be excluded from the list of official languages.

Note that the second term in (2) depends only on the number of languages in T. Thus, if the examination is restricted to sets of languages that consist of $k \leq L$ elements, the task is reduced to identifying those k languages that minimize disenfranchisement.[18] Indeed, let k be given. Denote

$$T_k^{\Gamma} = \arg \min_{|T|=k} D^{\Gamma}(T).$$

Then the optimal set $T^{\Gamma}(\alpha, \beta)$ is determined by:

$$T^{\Gamma}(\alpha, \beta) = \arg \min_{k=1,\ldots,L} G^{\Gamma}(T_k^{\Gamma}, \alpha, \beta).$$

[18] This is what Van Parijs (2003a) calls "the principle of minimal exclusion" for single languages.

In the next section we investigate the solutions of problem (1).

3 Dichotomous and Dyen Disenfranchisement Indices

Let us assume that for any two sets of languages S and S', the distance function $\Gamma(S, S')$ takes values between 0 and 1 and that $\Gamma(S, S') = 0$ only if S and S' contain a common language. If either S or S' is empty, we set $\Gamma(S, S') = 1$. We consider two special cases.

Dichotomous case. Here the value of the distance function, denoted $\Gamma^d(S, T)$, is equal to 1 for every two sets S and S' with an empty intersection. That is,

$$\Gamma^d(S, S') = \begin{cases} 0 \quad if \quad S \bigcap S' \neq \emptyset \\ 1 \quad if \quad S \bigcap S' = \emptyset. \end{cases}$$

Given the set of official languages T, the only factor in determining the degree of disenfranchisement of individual i is whether she speaks a language from T or not, and no consideration is given to languages which i does not speak. This formulation leads to a *dichotomous disenfranchisement* index, denoted $D^d(T)$, which represents the number of members who do not speak a language in T:

$$D^d(T) = \sum_{\{i \in N : P(i) \bigcap T = \emptyset\}} 1.$$

Dyen case. If an individual speaks at least one official language, she is not disenfranchised, that

is, the degree of her disenfranchisement is equal to zero. However, if she speaks none of the official languages, her degree of disenfranchisement may depend on the linguistic proximity between the set of languages that she speaks and the set of official languages. To account for this important feature, we consider the linguistic function Γ^y, derived from the matrix of "percentage cognate" Indo-European languages constructed by Dyen et. al (1992).[19] The matrix consists of the distances $y(l,m)$ between any two languages $(l,m) \in \mathcal{L}$. They take values between 0 and 1, with $y(l,m) = 0$ if and only if $l = m$. For two sets S and S', the value of the linguistic distance function $\Gamma^y(S,S')$ is then determined as the minimal distance between languages in S and T:

$$\Gamma^y(S,T) = \min_{l \in S, m \in T} y(l,m).$$

The corresponding Dyen disenfranchisement index $D^y(T)$ is the sum of Dyen distances between the language sets $P(i)$ of all members of the society and the set of official languages T:

$$D^y(T) = \sum_{\{i \in N : P(i) \bigcap T = \emptyset\}} y(P(i),T).$$

Since for every i who speaks a language that belongs to T, the linguistic distance $y(P(i),(T)$ is equal to zero, it follows that the Dyen index is, in fact, the sum of the Dyen linguistic distances between the set T and the language sets $P(i)$ for all those individuals who speak no language from T. This is in contrast to the dichotomous index that counts them as one.

4 Computing Disenfranchisement Indices

The disenfranchisement indices D^d and D^y are computed by using two sets of data. The first is a survey on language proficiency. Since some doubt is often cast on such surveys, we also calculate

[19] This matrix is actually the inverse to the *resemblance function* of Greenberg (1956).

two indices with respect to native populations of each country. In the latter case we assume, for simplicity, that the entire population of each country (or region, as in the case of Belgium) speaks its unique official language. Our derivations lead to four indices exhibited in Table 1.

Insert Table 1

4.1 Survey-Based Disenfranchisement

In 2000, the Directorate of Education and Culture of the European Union ordered a survey on languages, that was conducted by INRA (2000). In each of the 15 then-members of the EU, 1,000 interviews[20] were conducted on the use of languages. The information used in this paper is derived from answers to the following two questions:

(a) What is your mother tongue? (note to the interviewer: do not probe; do not read [the list of languages] out; if bilingual, state both languages);

(b) What other languages do you know? (show card [containing a list of languages];[21] read out; multiple answers possible).

There were four possible choices for (b). We assumed that the first two choices that came to the mind of the person interviewed were the languages that she knew best. There were also questions on whether the knowledge of the language was "very good," "good" or "basic," but we did not take these answers into account, since such qualifications are usually very subjective, vary across individuals and are, therefore, not very informative.[22]

[20] With some minor variations: 1,300 interviews in the UK, 2,000 in Germany, 600 in Luxembourg.

[21] Danish, German, French, Italian, Dutch, English, Spanish, Portuguese, Greek, Irish, Swedish, Finnish, Luxembourgish (one of the official languages of Luxembourg), Arabic, Turkish, Chinese, Sign language, Other (specify first and second), None.

[22] The examination of language knowledge in this type of surveys is open to a criticism. Non-native speakers of a

In order to derive disenfranchisement indices, we need some notation. For every subset T of the set of languages $\mathcal{L}$, we denote by $n^E(T)$ the number of individuals who speak all languages in T and no other language:

$$n^E(T) = |\{i \in N : P(i) = T\}|.$$

However, the survey results are given in terms of the number of individuals, denoted by $n^A(T)$, who speak all the languages in T and, possibly, some others:

$$n^A(T) = |\{i \in N : T \subseteq P(i)\}|.$$

Obviously, for every T, the inequality $n^E(T) \leq n^A(T)$ holds. The derivation of the values $n^E(T)$ from those of $n^A(T)$ is presented in the Appendix.

To adjust the survey results to our framework, we consider the set N of the residents of the European Union, and restrict our attention to the set $\mathcal{L}$ of six languages most widely spoken in the EU before the enlargement: Dutch, English, French, German, Italian and Spanish. To simplify, we disregard the small group of individuals who know four or more languages and assume that $n^E(T) = n^A(T) = 0$ if the set T contains more than three languages. By using the derivations relegated to the Appendix, we obtain the values of the functions $n^E(\cdot)$ given in Table 2.

Insert Table 2

language do not use the right idiomatic expressions, mistranslate, misinterpret the real meaning of words or sentences (Piron (1994, p. 67)). To be known, a language needs 12,000 hours of study and practice (Piron (1994, p. 79)) and a survey like the one we use certainly exaggerates the number of people who speak the language in some depth. Our argument for using the survey is twofold. First, it contains numbers, which are better than the usual guesswork on which discussions on knowledge of languages and the decisions that may follow, are based (Fettes (1991), Piron (1994, p. 69), and Crystal (1997, pp.55-61)). Second, this is the most complete and recent dataset that exists, and unless one has 15,000 people taking linguistic exams in several languages, it will be difficult to do any better.

These values allow for the direct derivation of the dichotomous disenfranchisement indices $D_s^d(T)$.[23] Moreover, combining them with the Dyen distance matrix, given in Table 3, Dyen disenfranchisement indices $D_s^y(T)$ can be easily computed. (Both indices D_s^d and D_s^y are given in Table 4.)

Insert Tables 3 and 4

4.2 Population-Based Disenfranchisement

Here we take the extreme assumption that only those citizens who live in a country speak its native language. It is quite obvious that this assumption will negatively affect native languages in less populated countries, and favor native languages in larger countries.[24] Both sets of indices D_p^d and D_p^y are presented in Table 4.

It is worthwhile to extend the examination of population-based indices. for the ten countries that have joined the Union on May 1, 2004. Detailed results are provided in Table 5.

Insert Table 5

German comes out as optimal choice if only one language is retained, but English and Italian are very close competitors. For three languages, the choice English-French-German is again optimal (or second-best), though the triples English-German-Italian or French-German-Italian are close substitutes.

[23] See Ginsburgh and Weber (2003), and Stroobants (2002).

[24] English, for example, is the native language of 62.3 million inhabitants (58.6 in the United Kingdom and 3.7 in Ireland), while German is spoken by 90.1 native speakers (82 million Germans and 8.1 million Austrians). Even French is the native language of more citizens than English (60.4 million Frenchman and 4 million French-speaking Belgians).

5 Optimal Choices of Official Languages: Empirical Analysis and Discussion

Since for given number of official languages k, given value of society's sensitivity to disenfranchisement α, and its degree of the language interpreting regime β, the solutions of the minimization problem (1) depend on disenfranchisement indices only, we can derive optimal sets T_k^Γ in Table 6 by using the data from Table 4.

Insert Table 6

It turns out that survey-based dichotomous and Dyen first-best choices coincide. English is obvious if society restricts its choice to a single official language. If two languages are chosen, then the second language should be reasonably distant from the first and known by a reasonably large number of non-natives. Therefore English-French is also an obvious choice, though Italian and Spanish come close to French. The successive optimal choices (if society opts to go to three, four, five and six languages) oscillate between a Germanic and a Latin language. For three, German is added, then Italian (or Spanish, which ties with Italian), then Spanish (or Italian), not because of their linguistic proximity, but because they are spoken by more citizens than Dutch, and finally, Dutch. It is also interesting to examine second-best choice sets, i.e., those with the second-lowest values of the indices. Under dichotomous disenfranchisement, the pairs English-French and English-German are very close. The Dyen index makes the choices English-French, English-Italian and English-Spanish almost identical; and so are the triples English-French-German, English-Italian-German and English-Spanish-German.

As expected, population-based optimal sets are different. Indeed, English loses its lead, since German and French are spoken by more natives than English, and Italian and Spanish are lin-

guistically closer than English and German.[25] However, if the Union settles for three working languages, English, French and German are the first-best choices according to three criteria, and is a second-best according to the Dyen population-based criterion. Note, however, that French could be replaced by Italian or Spanish without substantially altering the level of disenfranchisement.[26]

English-French-German is the group of languages that the European Commission uses nowadays (though German is used to a lesser extent), and these will probably be the pivotal languages, to which and from which other languages will be translated. Our results show that this is indeed the optimal choice. Since Spanish is widely spoken in some regions outside of the EU, it could, for that reason, serve as a serious alternative to French, even though French is optimal within the European Union.[27] This shows that when distances between languages are accounted for, the balance shifts towards Latin languages, providing a strong argument *against* English as a unique *lingua franca*.

Figures 1-6 illustrate the sets of optimal languages $T^d(\alpha, \beta)$ and $T^y(\alpha, \beta)$, respectively, for all values of α and β. The darkest area in the left of each figure represents the pairs (α, β), for which only one language (English) is chosen as the official language. The next areas to the right represent the sets of (α, β)-values for which two, three, four or five languages are optimal

[25]The Dyen distance between Italian and Spanish is 0.212, while it is 0.422 between English and German. See Table 3.

[26]The results would remain almost the same if we consider the EU after the enlargement. The only difference is that instead of Italian and German being first and second best single choices according to the Dyen-population index before the enlargement, German and French lead the way.

[27]French is used worldwide by 169 million people, Italian, by 70 million, and Spanish by 450 million. For Spanish see Dalby (2002, p. 31). For French which is also the lingua franca in most West-African countries, see http://www.france.diplomatie.fr/francophonie/francais/carte.html, the website of the French diplomatic service. Dalby's (2002, p. 31) estimate is somewhat lower (130 million people "use French"). For Italian, the number comes from http://www.ethnologue.com/show_language.asp?code=ITN (or DUT).

according to the criterion considered. Finally, in the white area, all six languages are needed. As the figures show, sensitivity to disenfranchisement (α) has to be very low in order to sustain a unique official language for all types of interpreting regimes (β). In general, the set of optimal languages expands under higher values of sensitivity to disenfranchisement and shrinks under a higher degree of comprehensiveness of the language regime.

Insert Figures 1-6

6 Conclusions

Our results show that it could be unwise to select English alone as a working language, not only because it is not always optimal, but also because it is optimal only for very small values of the coefficient which represents sensitivity to disenfranchisement. What is remarkable, however, is that whatever index is chosen, the best choice of three languages is English, French and German, though Italian could be a very reasonable substitute to French. This is so for the E. U. before and after the 2004 enlargement. 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. It may therefore be reasonable for the European Union to adopt four working languages, three of which (English, French and German) for general use, while Spanish is added for its importance in the rest of the world.

7 Appendix

The following useful result allows us to derive the values $n^E(T)$ from those of $n^A(T)$.

Proposition 3: For every $T \subset \mathcal{L}$ we have

$$n^E(T) = n^A(T) - \sum_{k=1}^{L-|T|} \sum_{S \in \mathcal{L}^T_{|T|+k}} (-1)^k n^A(S), \tag{3}$$

where for every integer $k, |T| \leq k \leq L$, $\mathcal{L}^T_k$ denotes the set of all subsets of $\mathcal{L}$ that consist of k elements and contain the set T.

Proof: Let individual i be such that there is a set $S \in \mathcal{L}^T_{|T|+k}$ such that $S \subset P(i)$. If $k = 0$ then i is included once on both sides of (3). If $k > 0$, then i does not appear on the left side of (3), but is included $(1-\binom{k}{1}+\binom{k}{2}-\cdots+(-1)^k\binom{k}{k})$ times[28] which completes the proof of the proposition. □

Since we ignore those individuals who speak at least four languages, (3) implies that for every $i, j, k \in \mathcal{L}$

$$n^E(\{i\}) = n^A(\{i\}) - \sum_{T \in \mathcal{L}^{\{i\}}_2} n^A(T) + \sum_{T \in \mathcal{L}^{\{i\}}_3} n^A(T),$$
$$n^E(\{i,j\}) = n^A(\{i,j\}) - \sum_{T \in \mathcal{L}^{\{i,j\}}_3} n^A(T),$$
$$n^E(\{i,j,k\}) = n^A(\{i,j,k\}).$$

The values n^E are presented in Table 2.

8 References

Bretton, Henry (1976), Political Science, Language, and Politics, in W.M. O'Barr and J.F. O'Barr, eds., *Language and Politics*, The Hague: Mouton.

Cole, John, and Francis Cole (1997), *A Geography of the European Union*, London: Routledge (second edition).

[28] Given a set of cardinality n, $\binom{n}{m}$ denotes the number of its m-element subsets, where $m \leq n$.

Council of the European Union (2002), Use of Languages in the Council in the Context of an Enlarged Union, Report of the Presidency, document 15334/1/02 December 6, 2002.

Crystal, David (1997), *English as a Global Language*, Cambridge: Cambridge University Press.

Dalby, Andrew (2002), *Languages in Danger*, London: Allen Lane, The Penguin Press.

De Swaan, Abram (1993), The Evolving European Language System: A Theory of Communication Potential and Language Competition, *International Political Science Review* 14(3), 241-255.

De Swaan, Abram (2001), *Words of the World*, Cambridge: Polity Press.

Dyen, Isidore, Joseph B. Kruskal, and Paul Black (1992), An Indo-European Classification: A Lexicostatistical Experiment, *Transactions of the American Philosophical Society* 82(5), Philadelphia: American Philosophical Society.

European Union (2001), Preparing for the Parliament of the Enlarged European Union, Report of the Secretary General, document PE 305.269/ BUR/fin, adopted by the Bureau on September 3, 2001.

Fettes, Mark (1991), Europe's Babylon: Towards a Single European Language, *History of European Ideas* 13, 201-202.

Ginsburgh, Victor and Shlomo Weber (2003), Language Disenfranchisement in the European Union, *Journal of Common Market Studies*, forthcoming.

Greenberg, Joseph (1956), The Measurement of Linguistic Diversity, *Language* 32, 109-115.

Grin, François (2003), On the Costs of Cultural Diversity, University of Geneva, working paper.

INRA, Eurobaromètre 54 Special, Les Européens et les Langues, February 2001.

Laponce, J.A. (1992), Language, and Politics, in M. Hawkesworth and M. Hogan, eds., *Encyclopedia of Government and Politics*, vol. 1, 587-602, London: Routledge.

Mamadouh, Virginie (1995), De Talen in het Europees Parlement (Languages in the European Parliament), ASGS Vol. 52, University of Amsterdam, Institute for Social Geography.

Mamadouh, Virginie (1998), Supranationalism in the European Union: What About Multilingualism, paper presented at the World Political Map Conference on Nationalisms and Identities in a Globalized World, May-nooth and Belfast, August 1998.

Mamadouh, Virginie and Kaj Hofman (2001), The Language Constellation in the European Parliament, 1989-2004, Report for the European Cultural Foundation, Amsterdam.

Piron, Claude (1994), *Le défi des langues*, Paris: L'Harmattan.

Schaerer, Rolf (2003), Multilingualism in the Enlarged European Union and its Institutions. Report presented at a brainstorming session on languages, European Commission, Brussels, January 27-28, 2003.

Stroobants, J.-P. (2002), Une Europe à 21 Langues, Nouveau Boulet Budgétaire pour la Commission, *Le Monde*, July 3, 2002, p. 6.

Swadesh, Morris (1952), Lexicostatistic Dating of Prehistoric Ethnic Contacts, *Proceedings of the American Philosophical Society* 96, 452-463.

Van Parijs, Philippe (2003a), Europe's Three Language Problems, in R. Bellamy, D. Castiglione and C. Longman, eds., *Multiligualism in Law and Politics*, Oxford: Hart, forthcoming.

Van Parijs, Philippe (2003b), private communication.

Table 1

Disenfranchisement Indices

	Dichotomous	Dyen
Survey-based data	D_s^d	D_s^y
Population-based data	D_p^d	D_p^y

Table 2

Number of EU Citizens Who Know Only 1, 2 or 3 Languages in $\mathcal{L}$

Languages	No. of speakers (in millions)	Languages	No. of speakers (in millions)
E	58.7	EGF	19.2
G	40.9	EGI	2.1
F	35.4	EGS	2.0
I	27.1	EGD	7.9
S	22.4	EFI	11.5
D	4.2	EFS	13.3
		EFD	4.4
EG	37.9	EIS	1.5
EF	24.7	EID	0.3
EI	11.4	ESD	0.1
ES	10.8	GFI	1.0
ED	2.8	GFS	0.6
GF	2.2	GFD	2.0
GI	0.9	GIS	0.1
GS	0.5	GID	ng
GD	1.1	GSD	ng
FI	7.8	FIS	0.8
FS	3.6	FID	0.1
FD	1.1	FSD	0.1
IS	0.4	ISD	ng
ID	ng		
SD	ng		

Notes. "ng" means less than 0.05 million. E = English, F = French, G = German, I = Italian, S = Spanish, D = Dutch.

Table 3

The Dyen Matrix of Linguistic Distances

	Dk	D	E	F	G	Gr	I	Po	S	Sw
Dk	0	0.337	0.407	0.759	0.293	0.817	0.737	0.750	0.750	0.126
D	0.337	0	0.392	0.756	0.162	0.812	0.740	0.747	0.742	0.308
E	0.407	0.392	0	0.764	0.422	0.838	0.753	0.760	0.760	0.411
F	0.759	0.756	0.764	0	0.756	0.843	0.197	0.291	0.291	0.756
G	0.293	0.162	0.422	0.756	0	0.812	0.735	0.753	0.747	0.305
Gr	0.817	0.812	0.838	0.843	0.812	0	0.822	0.833	0.833	0.816
I	0.737	0.740	0.753	0.197	0.735	0.822	0	0.227	0.212	0.741
Po	0.750	0.747	0.760	0.291	0.753	0.833	0.227	0	0.126	0.742
S	0.750	0.742	0.760	0.291	0.747	0.833	0.212	0.126	0	0.747
Sw	0.126	0.308	0.411	0.756	0.305	0.816	0.741	0.742	0.747	0

Notes. Since Finnish is not a Indo-European language, it is not included here. Given the linguistic remoteness of Finnish, its Dyen distance to every language in the table was set equal to 1. Dk = Danish, D = Dutch, E = English, F = French, G = German, Gr = Greek, It = Italian, Po = Portuguese, S = Spanish, Sw = Swedish.

This matrix is based on cognate data collected by Isidore Dyen in the 1960s (see IE-DATA1 at www.ntu.edu.au/education /langs/ielex/IE-DATA1). For each entry from the list of 200 basic meanings selected by Swadesh (1952), Dyen (see Dyen *et al* (1992)) collected the words used in 95 Indo-European speech varieties (languages and dialects) and classified these into *cognate classes*. For a given meaning, such a class contains all the words from different speech varieties, that have an unbroken history of descent from a common ancestral word. An entry of this matrix is equal to $n_{lm}/(n^0_{lm} + n_{lm})$, the " percentage cognate" between languages l and m, where n_{lm} is the number of meanings for which l and m are classified as "cognate" and n^0_{lm} is the number of meanings for which the speech varieties l and m are "not cognate." (The number of "doubtfully cognate" meanings does not enter into the calculation of such percentages). Note that the higher this number, the more "similar" the two languages. Since we use a "distance" matrix, it is more convenient to consider the "percentage of not cognate," $y(l,m) = n^0_{lm}/(n^0_{lm} + n_{lm})$. The diagonal elements $y(l,l)$ are set to zero.

Table 4

Dichotomous and Dyen Disenfranchisement Indices

in EU 15 for 1, 2, 3, 4, 5 and 6 languages in $\mathcal{L}$

Languages	D_s^d	D_s^y	D_p^d	D_p^y	Languages	D_s^d	D_s^y	D_p^d	D_p^y	Languages	D_s^d	D_s^y	D_p^d	D_p^y
E	169	108	314	197	EGF	70	20	160	46	EGFI	43	13	102	32
G	259	142	286	177	EGI	83	21	166	45	EGFS	48	13	120	34
F	250	144	312	182	EGS	92	25	185	52	EGFD	66	20	138	43
I	312	151	319	177	EGD	114	84	202	146	EGIS	57	15	127	35
S	321	161	337	186	EFI	85	32	192	77	EGID	78	20	138	41
D	353	153	36	186	GFI	115	41	164	58	EGSD	87	24	163	49
					GFS	121	41	183	60	EFIS	62	27	153	67
EG	119	85	224	150	GFD	148	48	200	67	EFID	80	19	170	43
EF	114	40	250	91	GIS	147	46	189	61	EFSD	85	19	188	45
EI	130	44	257	89	GID	185	53	207	66	EISD	95	21	195	47
ES	140	46	275	97	EFS	90	32	210	79	GFIS	82	33	125	49
ED	160	91	292	162	EFD	108	27	228	58	GFID	108	38	142	53
GF	156	51	222	73	EIS	103	35	217	80	GFSD	113	38	161	55
GI	198	57	229	71	EID	122	27	235	56	GISD	135	42	167	56
GS	206	62	247	78	ESD	131	31	253	63	FISD	154	43	193	58
GD	246	135	264	171	GSD	193	58	225	73					
FI	206	130	254	164	FIS	170	122	215	61	EGFIS	20	8	63	22
FS	212	132	273	168	FID	189	51	232	68	EGFID	39	12	80	29
FD	233	62	290	82	FSD	195	51	251	70	EGFSD	43	12	98	30
IS	259	140	279	168	ISD	235	58	257	71	EGISD	52	14	105	32
ID	288	70	297	81						EFISD	57	14	131	34
SD	297	75	315	88						GFISD	75	30	103	44
										EGFISD	16.5	7.5	41	19

Notes. E = English, F = French, G = German, I = Italian, S = Spanish, D = Dutch.

Table 5

Dichotomous and Dyen Population-Based Disenfranchisement Indices in EU 25, for 1, 2, 3, 4, 5 and 6 languages in $\mathcal{L}$

Languages	D_p^d	D_p^y	Languages	D_p^d	D_p^y	Languages	D_p^d	D_p^y
E	377	245	EGF	223	94	EGFI	165	79
G	349	224	EGI	229	92	EGFS	183	81
F	375	224	EGS	248	100	EGFD	201	90
I	382	225	EGD	265	194	EGIS	190	82
S	400	235	EFI	255	125	EGID	201	89
D	424	235	EFS	273	127	EGSD	226	96
			EFD	291	106	EFIS	216	115
EG	287	198	EIS	280	128	EFID	233	91
EF	313	140	EID	297	104	EFSD	251	93
EI	319	138	ESD	316	112	EISD	258	95
ES	338	145	GFI	227	106	GFIS	188	96
ED	355	210	GFS	245	108	GFID	205	100
GF	285	120	GFD	263	115	GFSD	223	102
GI	292	118	GIS	252	109	GISD	230	104
GS	310	126	GID	270	113	FISD	256	106
GD	327	218	GSD	288	121			
FI	317	212	FIS	278	109	EGFIS	126	70
FS	336	217	FID	295	116	EGFID	143	76
FD	353	131	FSD	314	118	EGFSD	161	78
IS	342	216	ISD	320	119	EGISD	168	79
ID	360	129				EFISD	194	82
SD	378	136				GFISD	166	91
						EGFISD	104	66

Notes. E = English, F = French, G = German, I = Italian, S = Spanish, D = Dutch.

Table 6
Optimal Languages Sets in EU15

	Number of languages					
	One	Two	Three	Four	Five	Six
First best choices						
Dich. survey-based	E	EF	EFG	EFGI	EFGIS	EFGISD
	169	114	70	43	20	16
Dyen survey-based	E	EF	EFG	EFGI*	EFGIS	EFGISD
	108	40	20	13	8	7
Dich. pop.-based	G	GF	EFG	EFGI	EFGIS	EFGISD
	286	222	160	102	63	41
Dyen pop.-based	I	GI	EGI	EFGI	EFGIS	EFGISD
	177	71	45	32	22	19
Second best choices						
Dich. survey-based	F	EG	EGI	EGFS	EGFID	
	250	119	83	48	39	
Dyen survey-based	G	EI	EGI	EGIS	EGFID†	
	142	41	21	15	13	
Dich. pop.-based	F	EG	FGI	EGFS	EGFID	
	312	224	164	120	80	
Dyen pop.-based	G	FG	EFG	EFGS	EFGID	
	182	73	46	34	29	

*Ties with EFGS.

†Ties with EGSFD.