Disenfranchisement and Diversity in Multilingual Societies.

Victor Ginsburgh[*] Ignacio Ortuño-Ortín [†] Shlomo Weber[‡]

March 2006

Abstract

We consider a linguistically diversified society that has to select a set of languages to be used for official purposes. We examine the notions of *language disenfranchisement* and *linguistic standardization* that is created when one or more languages fail to be included in the list of the official ones. For the analysis of our quantitative disenfranchisement indices, we use the Dyen *percentage cognate* matrix of linguistic distances between languages. We then apply survey and population data on language proficiency in the European Union before and after its latest enlargement in May 2004.

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]

At present, there are 6,912 distinct languages spoken in 271 countries all over the world. About one third of the world's nation-states have official language provisions in their constitutions. Multilingualism therefore is undoubtedly an important part of the current political debate almost everywhere. One must recognize that attempts to maintain multilingual societies (and to avoid the fate of Babylonia) require willingness on behalf of the participating linguistic groups to make compromises and to accept some sort of language standardization.

[*] ECARES, Université Libre de Bruxelles, and CORE, Louvain-la-Neuve, Belgium.

[†] Department of Economics and IVIE, University of Alicante, Alicante, Spain.

[‡] Department of Economics, SMU, Dallas, USA, CORE, Louvain-la-Neuve, Belgium, and CEPR.

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

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

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 chosen languages are called *official* or *working* languages. In selecting its official languages, the society should recognize an often fierce intensity of feelings on these issues.

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. 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 languages in the set of the official ones 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 multilingualism 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."

However, the cost of services required to maintain a larger number of official languages 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.[3] In 1999 the total translation and interpretation costs for the Commission alone amounted to 30% of its internal budget.[4] 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,[5] as well as to divergent translations and interpretations. The burden of maintaining official languages is not limited to direct costs of translation and interpretation. Communication[6] 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. 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

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

[4] De Swaan (1993) and (2001, p. 172).

[5] Mamadouh and Hofman (2001).

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

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 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*[7] disenfranchisement index.

2 The Model

We consider a *society* N that consists of n members, 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.

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

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

the set T is chosen as the set of official languages, the 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.

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 $\beta = 2$. 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

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

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: 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. 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:

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

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,

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

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).[8] 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. (See Table 1 for the values of the Dyen matrix.)

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 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.

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[9] were conducted on the use of languages. The information used

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

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

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];[10] read out; multiple answers possible).

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.[11]

It is worthwhile to extend the examination of population-based indices. for the ten countries that have joined the Union on May 1, 2004. 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.

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.

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

[10] 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.

[11] 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).

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 linguistically closer than English and German.[12] 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.[13]

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.[14] 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.*

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,

[12] The Dyen distance between Italian and Spanish is 0.212, while it is 0.422 between English and German. See Table 1.

[13] 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.

[14] 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).

French and German) for general use, while Spanish is added for its importance in the rest of the world.

7 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).

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.

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.

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 (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.

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.

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 file IE-DATA1 at www.ntu.edu.au/education /langs/ielex/IE-DATA1. The matrix is described in Dyen *et al.* (1992). For each meaning from a list of 200 basic meanings selected by Swadesh (1952), Dyen collected the words used in 95 Indo-European speech varieties (i.e., 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 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.