Why Do People Learn Foreign Languages?*

Victor Ginsburgh
ECARES, Université Libre de Bruxelles, Brussels, Belgium and
CORE, Université catholique de Louvain, Louvain-la-Neuve, Belgium

Ignacio Ortuno-Ortin
Department of Economics and IVIE, University of Alicante, Alicante, Spain

and

Shlomo Weber
Southern Methodist University, Dallas, Texas, U.S.A. and
CORE, Université catholique de Louvain, Louvain-la-Neuve, Belgium

May 2004

Abstract

We suggest a demand model for foreign languages and estimate demand functions for English, French, German and Spanish in 13 European countries. We show that three variables explain reasonably well the share of people who learn a foreign language: the larger the native population in the country, the less its citizens are prone to learn another language; the more the foreign language is spoken, the more it attracts others to learn it; the larger the distance between two languages, the smaller the proportion of people who will learn it.

1 Introduction

We are interested in the determinants that induce inhabitants from a country i (whose native language is i) to learn some other language j. Intuitively, the attractiveness of a foreign language j for a population that uses language i depends on the sizes of the two populations. The larger the population that speaks i, the lower the incentive (and the smaller the opportunity) of a citizen to learn another language, since he can trade and communicate with enough citizens who speak the same language as he does, either in his own country, or in any language friendly foreign country. However, the larger the population that speaks the foreign language j, the larger the attraction of that language on those who do not know it. Finally, learning a foreign language is easier if it is close to the native language, and one

* This paper was motivated by some of the views held in Van Parijs (2003). We are grateful to Abdul Noury and Philippe Van Parijs for comments on a previous version.

expects that the cost incured to learn it will definitely have an impact on the number of its students. This intuition, which leads to demand functions for foreign languages is confirmed by the theoretical model that is the subject of Section 2. Section 3 describes the data that will be used to estimate such demand functions, while estimation results are reported in Section 4. Section 5 is devoted to some concluding remarks.

2 Modeling the Learning of a Foreign Language

We consider two languages i and j, spoken in two regions or countries i and j, respectively by N_i and N_j citizens. For simplicity, we assume that all citizens are unilingual, but may consider learning the other language. We denote by N_{ij} (resp. N_{ji}) the number of citizens of country i (country j) who study language j (i). The language utility of each individual t depends on the number of those who speak the same language. It is represented by the utility function $U_t(x, y)$, where x is the (log of the) number of individuals who speak the same native language as t, while y is the (log of the) number of individuals who share with t a language that is not their native language.[1] We assume that the utility function is common to all individuals, so that $U_t(x, y) = U(x, y)$. Let n represent the logarithm of N.

More specifically, the utility of an i-speaker who learns j is $U(n_i, n_j)$, since she will be able to communicate with all j-speakers. The utility of an i-speaker who does not learn language j is $U(n_i, n_{ji})$: she will communicate with those who know her language in country j. For j-speakers these utilities are respectively $U(n_j, n_i)$ and $U(n_j, n_{ij})$. An individual who learns another language incurs a cost $C(d_{ij})$, where $d_{ij} = d_{ji}$ is the (log of the) linguistic distance D_{ij} between languages i and j.

We make the following standard assumptions:

Assumption A1. U(., .) is twice continuously differentiable and increasing in both arguments.
Assumption A2. The cross-derivative of U(., .) is positive.
Assumption A3. C(.) is increasing.

In a linguistic equilibrium, an individual will be indifferent between learning the foreign language and incurring the cost of learning it, and not learning the language. An

[1] Logarithms are used to link the model to the empirical results. This entails of course no loss of generality. Note that the formulation of the utility function allows for the special case $U_t(x + y)$ in which individual t derives the same utility from x and y.

(interior) *linguistic equilibrium* is therefore a solution of the following system of two equations:

$$U(n_i, n_j) - C(d_{ij}) - U(n_i, n_{ji}) = 0$$

$$U(n_j, n_i) - C(d_{ji}) - U(n_j, n_{ij}) = 0.$$

We assume that such an equilibrium exists, and leads to demand functions $n_{ij}(n_i, n_j, d_{ij})$ for language j of individuals whose native language is $i \neq j$; $n_{ji}(n_j, n_i, d_{ji})$ is the demand function of individuals whose native language is j.

Denote by $\log N_{ij}/N_i = D_i(n_i, n_j, d_{ij})$, the equilibrium share of individuals whose native language is i and who learn language j. The properties of $D_i()$ are decribed in the following proposition.

Proposition. If Assumptions A1-A3 are satisfied, then $D_i(\ )$ is (a) decreasing in n_i, (b) increasing in n_j and (c) decreasing in d_{ij}.

Proof. See Appendix.

3 Data

Our purpose is to the estimate the demand functions derived in Section 2 for English, French, German and Spanish by citizens from the European Union (E.U.) whose native languages are neither of these. The data at hand consist of knowledge of native and foreign languages in various E.U. countries, and distances between languages.

Language proficiency was the topic of a survey on languages ordered by the Directorate of Education and Culture of the E.U. in 2000.[2] In each of the15 E.U. countries, 1,000 interviews[3] were conducted on the use of languages. The information in which we are interested here is concerned with 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);

[2] INRA, Eurobaromètre 54 Special, Les Européens et les Langues, February 2001.
[3] With some minor variations: 1,300 interviews in the UK, 2,000 in Germany, 600 in Luxembourg.

(b) What other languages do you know? (show card [containing a list of languages];[4] read out; multiple answers possible).

There were four possible choices for (b), and 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 each of the tongues mentioned was "very good," "good" or "basic," but we did not take these answers into account, since such assessments are often subjective and, therefore, not very informative.

The results of such surveys can be questioned, since what individuals claim to know is hard to verify without deeper but very costly and time-consuming probing. We can however assume that there is some consistency across countries.

We restrict our attention to the "knowledge" in 13 E.U. countries[5] of four non-native languages: English, French, German, and Spanish. The first three are the most widely spoken in the E.U. Italian is more spoken than Spanish, but can hardly be considered to be as international as Spanish. Table 1, which also includes Italian and Dutch gives a general overview of language use in the E.U. and worlwide. Column (1) shows the number of native speakers, in fact the population in each country.[6] The second column shows the numbers of those who claim they know the language. For more details on how these numbers are calculated from the survey alluded to earlier, see Ginsburgh and Weber (2003). The last two columns show two estimates of the worldwide use of these languages. The first is the number of first language speakers, as given by www.ethnologue.com. Its advantage is consistency over countries. The other one gives estimates of worldwide knowledge. As can be seen, these numbers are much larger, since they include people who "can handle" the language.

Table 2 gives some details on the knowledge of languages in the 13 E.U. countries that are dealt with in our study. Column (2) contains the world population that speaks as first language, the language of the country listed in column (1). The other four columns give the percentage of people who (claim to) know English, French, German and Spanish.

Data on distances between languages among 95 Indo-European languages have been computed by Dyen et al. (1992). They are constructed on the basis of a set of cognition data. For each meaning in a list of 200 basic meanings, Dyen collected the words used in 95 Indo-European speech varieties, and classified these into cognate classes, that contain all the

[4] Danish, German, French, Italian, Dutch, English, Spanish, Portuguese, Greek, Irish, Swedish, Finnish, Luxembourgish, Arabic, Turkish, Chinese, Sign language, Other (specify first and second), None.
[5] Austria, Denmark, Finland, France, Germany, Greece, Italy, Ireland, The Netherlands, Portugal, Spain, Sweden and The United Kingdom. Belgium and Luxemburg are omitted because they are both biligual and would be more difficult to treat (and Luxemburg's population--0.4 million--is extremely small.)
[6] To simplify, we assume that immigrants speak the language of the country to which they migrated.

words for a given meaning that have an unbroken history of descendent from a common ancestral word. The distance between languages i and j is then computed as 1 minus the ratio of "cognate" and "cognate" plus "non-cognate" meanings,[7] and lies between 0 and 1. The distances used in this paper are given in Table 3.

4 Estimation Results

The general idea is to estimate a demand function for any language j (j = English, French, German, Spanish) by those whose native language is i (i = Austria, Denmark, Finland, France, Germany, Greece, Italy, Ireland, The Netherlands, Portugal, Spain, Sweden and The United Kingdom) which takes the following logarithmic functional form:

(1) $$\log (N_{ij}/N_i)_{EU} = \alpha_0 + \alpha_1 n_i + \alpha_2 n_j + \alpha_3 d_{ij} + u_{ij},$$

where $(N_{ij}/N_i)_{EU}$ represents the proportion of inhabitants of E.U. country i who are proficient in language j (columns (2) to (5) in Table 2), while n_i and n_j represent respectively the (log of the) world populations whose native languages are i and j (column (2) in Table 2).

We first estimate a demand function for each foreign language j separately:

(2) $$\log (N_{ij}/N_i)_{EU} = \alpha_0 n_j + \alpha_1 n_i + \alpha_3 d_{ij} + u_{ij}.$$

Note that in each case, the intercept α_0 is multiplied by the world population that practices language j, and can be interpreted as α_2, This normalization (which does of course not change the other coefficients) will make it possible to give a first insight into the attraction power of individual foreign languages. Estimation results are reproduced in Table 4. They show that the fit is excellent and consistent with theory for English, and German. This is not the case for French and Spanish, though, distance picks a negative sign as it should. The coefficient for the country of origin, which should also be negative, is so for English and German only. Finally, observe that the (population-weighted) intercept terms are all positive, but their magnitudes differ widely.[8]

---

[7] See Dyen et al. (1992) for further details. See also Ginsburgh, Ortuno and Weber (2003).
[8] We also computed such equations using native populations in each country instead of the number of speakers of each language given in column (2) of Table 2. The results are qualitatively similar.

In Table 5, we list the estimation results for equations in which we pool observations. This pooling is useful for two reasons. The number of observations for single languages is rather small (11 or 12, see Table 4), and pooling makes it possible to estimate the basic equation using more observations. Pooling also checks whether the relations which govern the learning of a foreign language hold more generally than for each language individually. The first two columns give the results for the two Germanic (English and German) and the two Latin languages (French and Spanish) according to the following form:

(3) $$\log (N_{ij}/N_i)_{EU} = \alpha_{02}\delta\, n_2 + \alpha_1 n_i + \alpha_2 n_j + \alpha_3 d_{ij} + u_{ij},$$

where n_2 is the world German (resp. Spanish) speaking population; the dummy δ takes the value 1 for the observations relative to German (resp. Spanish), and 0 otherwise. This normalization makes it easy to test the null hypothesis that the two languages are equally attractive.[9] Estimation results are consistent with the model for the Germanic group, and show that the null of equal attraction is strongly rejected: German is significantly less attractive than English. For the Latin group, results are less clear cut. They show, however, that (a) distance between languages carries the negative sign and the effect is significantly different from zero, (b) Spanish is less attractive than French, but (c) the size of the population which learns the foreign language has a positive, though insignificant, effect.

Estimation of Eq. (1) on the full set of 46 observations leads to poor results that are not reported. The results given in column (3) of Table 5 are for an equation that has the same form as Eq. (3), except that it contains an $\alpha_{0j}\delta_j\, n_j$ term for each language j = French, German and Spanish. As can be checked, the α_{0j} parameters take significantly different values, and explain why the fit in an Eq. (1) type is poor. The results are consistent with the theoretical model (though α_1 is not significantly different from 0), but the four languages have different attraction powers. Note that the "distance" elasticity is not significantly different from -1, implying that a one percent increase in the distance between two languages decreases the number of its students by one percent. This is far from being negligible.

[9] Note that since there are only two possible values for n_j, the regression $\log (N_{ij}/N_i)_{EU} = \alpha_0 + \alpha_1 n_i + \alpha_2 n_j + \alpha_3 d_{ij} + u_{ij}$ leads to the same results for α_1 and α_3.

5 Concluding Comments

Our results show that three variables explain reasonably well the share of people who learn a foreign language. The larger the native population in the country, the less its citizens are prone to learn another language;[10] the more the foreign language is spoken, the more it attracts others to learn it; the larger the distance between two languages, the smaller the proportion of people who will learn it.

However, our results also show that the attraction powers of the four foreign languages are significantly different, and that other determinants, mostly historical, but also economic, must be at play.

Spanish, for instance, should attract Europeans much more than it does. With the exception of France, there is no country in which more that five percent of the population knows the language. The isolation of Spain until 1975, the year in which Franco died, explains partly this result, but the large population of native Spanish speakers (essentially in Mexico and South America) does only partially compensate for the lower level of economic development. Dynamics, past as well as current commercial and cultural relations that are absent from our model should obviously be part of the story: Attractiveness of a foreign language depends on more than the number of people who speak it worldwide. Therefore the questions of why English is becoming the *lingua franca* in Europe (and probably in the world), and why Spanish is relatively less spoken in Europe remain only partly captured by our model.

References

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

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

Ginsburgh, V., I. Ortuno-Ortin and S. Weber (2003), Language disenfranchisement in heterogeneous societies, manuscript.

Ginsburgh, V. and S. Weber (2003), Language disenfranchisement in the European Union, manuscript.

Van Parijs, Philippe (2003), Europe's three language problems, in R. Bellamy, D. Castiglione and C. Longman, eds., *Multilingualism in Law and Politics*, Oxford: Hart, forthcoming.

[10] Van Parijs (2003) holds the interesting view that this effect may be due to the dubbing of English speaking movies in countries that are large enough to afford the cost of dubbing (France, Spain, Germany).

Table 1
Main Languages Used in the European Union
(millions)

	Native speakers in the E.U.[1]	Use or know the language in the E.U.[2]	Use or know the language worldwide	
			First language[3]	All forms[4]
	(1)	(2)	(3)	(4)
English	62.3	208.6	341	1,800
French	64.5	127.8	77	169
German	90.1	118.3	100	126
Spanish	39.4	56.3	340	450
Italian	57.6	65.2	62	na
Dutch	21.9	24.3	20	na

[1] English is the native language in Great Britain and Ireland. French is the native language in France and is spoken by 40% of Belgians. German is the native language in Germany and Austria. Spanish and Italian are the native languages in Spain and Italy, respectively. Finally Dutch is the native language in The Netherlands and is spoken by 60% of Belgians.
[2] Computed from the survey on languages described in Ginsburgh and Weber (2003).
[3] Number of first language speakers as given by www.ethnologue.com. For Spanish, the number is the average between the two estimates given by www.ethnologue.com.
[4] Estimates for English ("number of people who can handle the language competently") and Spanish are from 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 German see http://germa.about.com/library/weekly/aa020298.htm.

Table 2
Knowledge of Languages in the European Union
(millions and %)

Country	Native language known by (millions)	Know English (%)	Know French (%)	Know German (%)	Know Spanish (%)
(1)	(2)	(3)	(4)	(5)	(6)
Austria (G)	100.0	46	11	100	1
Denmark (Dk)	5.3	75	5	37	1
Finland (Fi)	6.0	61	1	7	1
France (F)	77.0	42	100	8	15
Germany (G)	100.0	54	16	100	2
Greece (Gr)	12.0	47	12	12	5
Italy (I)	62.0	39	29	4	3
Ireland (E)	341.0	100	23	6	2
Netherlands (D)	20.0	70	19	59	1
Portugal (P)	176.0	35	28	2	4
Spain (S)	340.0	36	19	2	100
Sweden (Sw)	9.0	79	7	31	4
Un. Kingdom (E)	341.0	100	22	9	5

Notes. The native language in each country is given between brackets (G: German, Dk: Danish, Fi: Finnish, F: French, G: German, Gr: Greek, I: Italian, E: English, D: Dutch, P: Portuguese, S: Spanish, Sw: Swedish).
Sources. The numbers in the first column are from www. ethnologue.com. The percentages of people who know English, French, German and Spanish in each country are from Ginsburgh and Weber (2003).

Table 3
Distances between languages
(x 1,000)

	English	French	German	Spanish
Danish	407	759	293	750
Dutch	392	756	162	742
English	0	764	422	760
Finnish	1000	1000	1000	1000
French	764	0	756	266
German	422	764	0	747
Greek	838	843	812	833
Italian	753	197	735	212
Portuguese	760	291	753	126
Spanish	760	266	747	0
Swedish	411	756	305	747

Sources. Dyen et al. (1992) for further details. See also Ginsburgh, Ortuno and Weber (2003).

Table 4
Estimation Results for Individual Languages

	English	French	German	Spanish
Population speaking language i (α_1)	-0.153* (0.021)	0.355* (0.138)	-0.361* (0.072)	0.032 (0.168)
Distance between languages i and j (α_3)	-0.408* (0.082)	-0.512 (0.416)	-1.362* (0.214)	-0.560 (0.385)
Intercept (α_0)	0.733* (0.016)	0.193 (0.121)	0.586* (0.077)	0.091 (0.109)
R-square	0.919	0.599	0.910	0.232
No. of observations	11	12	11	12

Standard errors are given between brackets, under the coefficients. A * indicates that the coefficient is significantly different from zero at the 5% (or even 1%) probability level. The number of observations is 12 for France and Spain, since French or Spanish is a foreign language in 12 of the 13 countries. This number is 11 for English (spoken in the UK and Ireland) and German (spoken in Austria and Germany).

Table 5
Estimation Results for Groups of Languages

	English and German	French and Spanish	English, German, French and Spanish
Population speaking language i (α_1)	-0.272* (0.057)	0.205 (0.109)	-0.058 (0.069)
Population speaking language j (α_2)	0.744* (0.046)	0.326* (0.103)	0.625* (0.057)
Distance between languages i and j (α_3)	-1.102* (0.188)	-0.531 (0.284)	-0.954* (0.200)
French speaking population (α_{0F})	-	-	-0.112 (0.062)
German speaking population (α_{0G})	-0.394* (0.037)	-	-0.233* (0.061)
Spanish speaking population (α_{0S})	-	-0.262* (0.052)	-0.514* (0.050)
R-square	0.906	0.647	0.758
No. of observations	22	24	46

Standard errors are given between brackets, under the coefficients. Starred coefficients are significantly different from 0 at the 5% (or even 1%) probability level. See also the notes in Table 2.

Appendix. Proof of the Proposition

(a) We rewrite the second equation in (1) as:

$$U(n_j, n_i) - C(d_{ji}) - U(n_j, n_i + D_i(n_i, n_j, d_{ij}) = 0.$$

Assume now that n_i increases to $\tilde{n}_i$. The concavity of U (Assumption (A1)), and the fact that $n_{ij} < n_i$ imply that

$$U(n_j, \tilde{n}_i) - C(d_{ji}) - U(n_j, \tilde{n}_i + D_i(n_i, n_j, d_{ij})) < 0.$$

Since by (A1), U is increasing, it follows that

$$D_i(\tilde{n}_i, n_j, d_{ij}) < D_i(n_i, n_j, d_{ij}),$$

which proves assertion (a) that D_i is decreasing in n_i.

Applying the Implicit Function Theorem to (1), we have

$$\frac{\partial N_{ij}}{\partial N_j} = - \frac{U_1(n_j, n_i) - U_1(n_j, n_{ij})}{- U_2(n_j, n_{ij})},$$

where U_1 and U_2 are the derivatives of U w.r.t. the first and the second arguments. But since by (A2), U_1 is increasing in the second argument and $U_2 > 0$, by (A1), the last expression is positive, which proves assertion (b) that D_i is increasing in n_j.

(c) Assertion (c) follows immediately from (A1) and (A3). ◊◊