DISCUSSION PAPER SERIES

No. 4942

LEARNING FOREIGN LANGUAGES.THEORETICAL AND EMPIRICAL IMPLICATIONS OF THE SELTEN AND POOL MODEL

Victor Ginsburgh, Ignacio Ortuño-Ortín
and Shlomo Weber

PUBLIC POLICY

Centre for Economic Policy Research

www.cepr.org

Available online at: **www.cepr.org/pubs/dps/DP4942.asp**

ISSN 0265-8003

LEARNING FOREIGN LANGUAGES.THEORETICAL AND EMPIRICAL IMPLICATIONS OF THE SELTEN AND POOL MODEL

Victor Ginsburgh, ECARES, Université Libre de Bruxelles
Ignacio Ortuño-Ortín, Universidad de Alicante
Shlomo Weber, Southern Methodist University, CORE, Université Catholique de Louvain and CEPR

Discussion Paper No. 4942
March 2005

Centre for Economic Policy Research
90–98 Goswell Rd, London EC1V 7RR, UK
Tel: (44 20) 7878 2900, Fax: (44 20) 7878 2999
Email: cepr@cepr.org, Website: www.cepr.org

CEPR Discussion Paper No. 4942

March 2005

ABSTRACT

Learning Foreign Languages.Theoretical and Empirical Implications of the Selten and Pool Model*

In this paper we adopt the Selten-Pool model (1993) framework of language acquisition that is based on the notion of communicative benefits and learning costs. We consider a model with languages that serve as imperfect substitutes and show that under supermodularity of the communicative benefits function and some other mild conditions, there exists a unique interior linguistic equilibrium. We then derive a demand function for foreign languages that we estimate for English, French, German and Spanish in 13 European countries.

JEL Classification: C72, O52 and Z13
Keywords: communicative benefits, estimation of demand functions for languages, European Union, languages, learning costs and linguistic distances

Victor Ginsburgh
ECARES
Universite Libre de Bruxelles
Avenue F Roosevelt 50
1050 Bruxelles
Belgium
Tel: (32 2) 650 3839/4
Fax: (32 2) 650 3595
Email: vginsbur@ulb.ac.be

For further Discussion Papers by this author see:
www.cepr.org/pubs/new-dps/dplist.asp?authorid=101991

Ignacio Ortuño-Ortín
Department of Economics
University de Alicante
Campus de Sant Vicent
03071 Alicante
SPAIN
Tel: (34 96) 590 3400 ext 3223
Fax: (34 96) 590 3898
Email: ortin@merlin.fae.ua.es

For further Discussion Papers by this author see:
www.cepr.org/pubs/new-dps/dplist.asp?authorid=154176

Shlomo Weber
Department of Economics
Southern Methodist University
Dallas, TX 75275
USA
Tel: (1 214) 768 3577
Fax: (1 214) 768 1821
Email: sweber@mail.smu.edu

*The 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.

Submitted 16 February 2005

1 Introduction

In this paper we analyze the reasons that induce inhabitants of a country to learn other languages. Each individual's decision can be analyzed by examining the benefits and the costs that it generates. Economists who examine the benefits of acquiring an additional language link them with the increased earning potential,[3] but these studies often focus on immigrants who acquire the native language of the country in which they live. We adopt the framework of Selten and Pool (1993) who consider a general model of learning languages and do not limit their analysis to "earnings as a mechanism and to firms as a milieu of the incentive to learn languages." (Selten and Pool, 1993, p.66).

In their model every individual derives a gross benefit from the knowledge of a foreign language that depends on the number of other individuals with whom she can communicate. This "gross communicative benefit" is therefore positively correlated with the number of other individuals with whom she shares at least one common language. Naturally we assume that all languages are "communicative substitutes," and the communication between two individuals can take place in any common language they share. The substitution however, is, in general, imperfect, and the communicative benefits are different (larger) if communication is conducted in the languages that are native for both sides. Therefore, we distinguish advantages of communication in native and non-native languages. For any individual t we represent the gross communication benefit by means of an increasing function with two arguments: the number of individuals who share a common native language with t, and the number of individuals who speak a language known by t but do not share her native language. To reflect the fact that languages are substitutes, we assume that the benefit function is *supermod-*

[3] See e. g., MacManus, Gould and Welsch (1978), Granier (1985), Lang (1986), and Chiswick (1993)

ular.

Selten and Pool also assume that an individual who learns a new language incurs a cost. We make the assumption of "cost heterogeneity," that is the cost and difficulty for t of learning a new language depends on its linguistic proximity with her native language.[4] Indeed, it is natural to assume that a native speaker of Portuguese would find it easier to learn Spanish than Swedish. The fact that learning a foreign language is easier if it is close to the native language will have an impact on the number of those who learn it. The net communicative benefit that determines the individuals' behavior is the difference between the gross communicative benefit and the cost of acquiring a new language.

Consider the case of two native languages i and j with two populations. The attractiveness of a foreign language for a population that may learn this non-native language depends on the sizes of both populations. If the population that speaks i is large relatively to the other one, the incentive of an i-citizen to learn the other language is likely to be quite low, since she can trade and communicate with enough citizens in her own country. But a large population that speaks j may also attract citizens who speak i. The intuition on the expected properties of 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 results are reported in Section 4. Section 5 is devoted to some concluding remarks.

2 Modelling 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

[4] See Dyen, Kruskal and Black (1992), and Ginsburgh, Ortuño-Ortín and Weber (2003).

denote by N_{ij} (resp. N_{ji}) the number of citizens of country i (country j) who study language j (i). The communicative benefit 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.[5] 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 communicative benefit 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 benefit 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 the levels of communicative benefit are $U(n_j, n_i)$ and $U(n_j, n_{ij})$, respectively. An individual who learns another language incurs a cost $C(l_{ij})$, where $l_{ij} = l_{ji}$ is the (log of the) linguistic distance L_{ij} between languages i and j.

We impose the following assumptions:

Assumption A1: $U(\cdot, \cdot)$ is continuous and increasing on $\Re^2_+$. Moreover, U is *supermodular*, i.e., for every two pairs of positive numbers $\overline{n}_i, n_i$, $\overline{n}_j, n_j$ with $\overline{n}_i > n_i$, $\overline{n}_j > n_j$ the following inequality holds:

$$U(\overline{n}_i, \overline{n}_j) - U(\overline{n}_i, n_j) \geq U(n_i, \overline{n}_j) - U(n_i, n_j).$$

Assumption A2: The cost function $C(\cdot)$ is continuous and increasing on $\Re_+$.

The first part of Assumption A1 and Assumption A2 are standard, whereas the second part of A1 simply reflects the fact that the two languages are sub-

[5] Logarithms are used to link the model to the empirical results. This entails no loss of generality.

stitutes. If the function U is twice continuously differentiable, the condition amounts to the positivity of the cross derivative U_{ij} (Topkis, 1979).

To state our first result, we need an additional condition that guarantees that learning costs are not prohibitively high with respect to communicative benefits. It requires that if no j-speaker studies i, an i-speaker would get a positive net benefit from studying j so that his access to all j-speakers outweighs the language learning cost $C(l_{ij})$. If this assumption is violated, than no citizen of country i learns a foreign language. Similarly, we assume that it is worthwhile for a j-speaker to study i, if no i-speaker learns j. This very mild condition is formally stated as follows:

Assumption A3:

$$U(n_i, n_j) - U(n_i, 0) > C(l_{ij}) \text{ and } U(n_i, n_j) - U(0, n_j) > C(l_{ij}).$$

These assumptions make it possible to state the following two propositions.

Proposition 1: Under A1 and A3, there exists a unique interior[6] linguistic equilibrium, where all individuals are indifferent between learning the foreign language and incurring the cost of learning it, and not learning the language. This equilibrium is a solution of the following system of two equations:

$$U(n_i, n_j) - C(l_{ij}) - U(n_i, n_{ji}) = 0, \tag{1}$$

$$U(n_j, n_i) - C(l_{ji}) - U(n_j, n_{ij}) = 0. \tag{2}$$

Proof: The proof follows immediately from Assumptions A1 and A2. Indeed, A2 together with the continuity of U in the second argument yields the unique n_{ji} that satisfies (1), whereas A2 together with the continuity of

[6] There also exist two "corner" equilibria, where one language is learnt by everybody in the foreign country whereas the other language is learnt by nobody (see Church and King, 1993). We do not examine these equilibria here.

U in the first argument guarantees the uniqueness of n_{ij} that satisfies (2). □

The interior linguistic equilibrium yields functions $n_{ij}(n_i,\ n_j,\ l_{ij})$ and $n_{ji}(n_j, n_i, l_{ji})$ for languages j and i of individuals whose native language is i and j, respectively. Denote by $\log(N_{ij}/N_i) = D_i(n_i, n_j, l_{ij})$, the equilibrium share (demand function) of individuals whose native language is i and who learn language j. The properties of D_i are described in the following proposition:[7]

Proposition 2: Suppose that Assumptions A1-A3 hold. Then:

(a) If, in addition, U is concave in the second variable, $D_i(\cdot, n_j, l_{ij})$ is decreasing in n_i;

(b) $D_i(n_i, \cdot, l_{ij})$ is increasing in n_j;

(c) $D_i(n_i, n_j, \cdot)$ is decreasing in l_{ij}.

Proof: (a) We can rewrite equation (2) as:

$$U(n_j, n_i) - C(l_{ji}) - U(n_j, n_i + D_i(n_i, n_j, l_{ij})) = 0. \qquad (3)$$

Assume that n_i increases to $\overline{n}_i$. Concavity of the U in the second variable implies that

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

Since

$$U(n_j, \overline{n}_i) - C(l_{ji}) - U(n_j, \overline{n}_i + D_i(\overline{n}_i, n_j, l_{ij})) = 0,$$

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

$$D_i(\overline{n}_i, n_j, l_{ij}) < D_i(n_i, n_j, l_{ij}).$$

(b) Consider (3) and assume that n_j increases to $\overline{n}_j$. The supermodularity

[7] Obviously, the same properties hold for D_j.

of U implies that

$$U(\overline{n}_j, n_i) - C(l_{ji}) - U(\overline{n}_j, n_i + D_i(n_i, n_j, l_{ij})) > 0.$$

But since

$$U(\overline{n}_j, n_i) - C(l_{ji}) - U(\overline{n}_j, n_i + D_i(n_i, \overline{n}_j, l_{ij})) = 0,$$

and, by A1 (U is increasing in the second argument), it follows that

$$D_i(n_i, n_j, l_{ij}) < D_i(n_i, \overline{n}_j, l_{ij}).$$

(c) Follows from $A1$ and $A2$. Indeed, let $\bar{l}_{ij} > l_{ij}$, which, by $A2$, yields $C(\bar{l}_{ij}) > C(l_{ij})$. Since the function U is increasing in the second argument, equation (2) immediately implies that the number of j-learners in country i would decline under a higher value of the linguistic distance. □

3 Data

We estimate the demand functions derived in Section 2 for English, French, German and Spanish by citizens from the European Union (EU) whose native languages are none of these. The data consist of knowledge of native and foreign languages in various EU 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 EU in 2000.[8] In each of the 15 EU countries, 1,000 interviews[9] were conducted on the use of languages. The information in which we are interested here is concerned with answers to the following two questions:

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

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

(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).

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. Note that 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.

Table 1 gives a general overview of the six languages most used by native speakers in the EU. It also shows to what extent the language is spoken elsewhere than in its country of "origin." Column (1) shows the number of native speakers, in fact the population in each country.[11] Columns (2) and (3) show two estimates of the worldwide use of each language as mother tongue. Though orders of magnitude are similar, there is some variation especially for English and Spanish. The last column gives estimates of worldwide knowledge as mother tongue and otherwise.[12] English, French and German are the languages most widely spoken in the EU. Though Italy has a larger population than Spain, and the number of natives speakers of

[10] Danish, German, French, Italian, Dutch, English, Spanish, Portuguese, Greek, Irish, Swedish, Finnish, Luxembourgish, Arabic, Turkish, Chinese, Sign language, Other (specify first and second), None.

[11] To simplify, we assume that immigrants speak the language of the country to which they migrated.

[12] Dalby (2002, p. 31) gives much higher estimates for English (1.8 billion) and Spanish (450 million). The French diplomatic service (website http://www.france.diplomatie.fr/francophonie/francais/carte.html) provides the estimation of 169 million people who use French. Note that for Dutch, the numbers are not fully consistent.

Italian is larger, the language is hardly spoken outside of Italy,[13] which is of course not the case of Spanish. Dutch is spoken only in the Netherlands and Belgium. Therefore, we restrict our attention to the "knowledge" of four non-native languages (English, French, German, and Spanish) in 13 EU countries.[14]

Table 2 gives details about the knowledge of languages in the 13 EU countries that are dealt with here. Column (2) contains the world population that speaks the language of the country listed in column (1) as first language.[15] The other four columns give for each EU country the share of the total population that (claims to) know English, French, German and Spanish.

Data on distances between languages among 95 Indo-European languages have been computed by Dyen, Kruskal and Black (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 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 the ratio between "non-cognate" and "cognate" plus "non-cognate" meanings,[16] and lies between 0 and 1. The distances used in this paper are displayed in Table 3.

[13]Italian is known by 9 percent of the population in Austria, 6 percent in Belgium and Greece, 5 percent in France, and one percent in other countries of the EU.

[14]Austria, Denmark, Finland, France, Germany, Greece, Italy, Ireland, The Netherlands, Portugal, Spain, Sweden and The United Kingdom. Belgium and Luxembourg are omitted because they are both bilingual and would be more difficult to treat (and Luxembourg's population–0.4 million–is extremely small.)

[15]Following Ethnologue.com.

[16]See Dyen, Kruskal and Black (1992), and also Ginsburgh, Ortuño-Ortín and S. Weber (2003) for further details.

4 Estimation Results

We estimate the following demand function for languages 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):

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

where $(N_{ij}/N_i)_{EU}$ represents the proportion of inhabitants of EU country i who are proficient in language j (columns (3) to (6) in Table 2); 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) and l_{ij} is the (log of the) distance between languages i and j (Table 3).

The results of this equation appear in column (1) of Table 4. Though the signs of all three variables conform with theory, and both α_2, the coefficient picked by the acquired language, and α_3, the coefficient for distance are significantly different from zero, the adjustment is poor, as can be seen from the adjusted R^2. The reason for this appears in the equation reproduced in column (2), which separates the four acquired languages (world populations who speak English, French, German and Spanish are interacted with dummies for each of them). The results show that the four languages do not have the same attraction power, which is explained by factors that are not fully captured by the population that knows the language.

Spanish is obviously much less attractive than the three other languages. In column (3), we pool the three other languages, leaving Spanish separate. The fit deteriorates only very marginally with respect to the previous equation. The null hypothesis (using an F-test) that English, French and German are equally attractive, but that Spanish is less so, cannot be rejected at the 10% probability level, but is rejected at the 5% level.

The introduction of trade shares (which do not appear in the theoretical model) as possible determinants (or incentives) between the 13 EU countries and the regions in which English, French, German and Spanish are spoken, does not lead to much change. The parameter picked by the trade shares variable is almost significant at the 5% probability level, and the parameters for acquired languages are slightly reduced.

5 Concluding Comments

Our results show that, in conformity with the theoretical model, three variables explain reasonably well the share of people who learn a foreign language, without taking into account the incentives every individual has to acquire a language. The larger the native population who speaks the language, the less speakers 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. However, our results also show that the attraction powers of the four foreign languages are significantly different, and that other determinants, mostly historical, 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 than 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, and increasingly so in the United States) does not seem to generate large incentives to learn the tongue. Dynamics, past as well as current cultural relations that are absent from our model should obviously be part of the story: Attractiveness of a foreign language depends on more than just 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.

6 References

Chiswick, B. (1993), Hebrew language usage: determinants and effects on earnings among immigrants in Israel, manuscript, University of Illinois.

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

Crystal, D. (2001), *A Dictionary of Language*, Chicago: Chicago University Press.

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 (5).

Ginsburgh, V., I. Ortuño-Ortín, and S. Weber (2003), Language disenfranchisement in linguistically diverse societies. The case of European Union, *Journal of the European Economic Association*, forthcoming.

Ginsburgh, V. and S. Weber (2003), Language disenfranchisement in the European Union, *Journal of Common Market Studies*, forthcoming.

Grenier, G. (1985), Bilinguisme, transferts linguistiques et revenus du travail au Québec, quelques éléments d'interaction, in Francois Vaillancourt, ed., *Economie et Langue*, Québec: Editeur officiel, 243-287.

Lang, K. (1986), A language theory of discrimination, *Quarterly Journal of Economics* 100, 363-381.

MacManus, W., W. Gould, and F. Welsch (1978), Earnings of Hispanic men: The role of English language proficiency, *Journal of Labor Economics* 1, 101-130.

Selten, R. and J. Pool (1991), The distribution of foreign language skills as a game equilibrium, in R. Selten, ed., *Game Equilibrium Models,* vol. 4, Berlin: Springer-Verlag, 64-84.

Topkis, D. (1979), Equilibrium points in non-zero sum n-person submodular games, *SIAM Journal of Control and Optimization* 17, 773-787.

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

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

	Native speakers in the EU	Mother tongue		Woldwide use
		Ethnologue	Crystal	
	(1)	(2)	(3)	(4)
English	62.3	341	400	1,000-1,500
French	64.5	77	72	122
German	90.1	100	100	120
Spanish	39.4	340	270	350
Italian	57.6	62	63	63
Dutch	21.9	20	20	20

Column (1): English is the native language in Great Britain and Ireland. French is the native language in France and is spoken by 40 percent 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 percent of Belgians.
Column (2): 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.
Column (3): Estimates by Crystal (2001).
Column (4): Estimates by Crystal (2001). Note that the 1 billion users of English is a conservative estimate. Crystal also gives a more "liberal" estimate of 1.5 billion users.

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

Country	Native language known by (millions)	Percentage who know English	French	German	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

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). 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 et. al (2003).

Table 4. Estimation Results

	(1)	(2)	(3)	(4)
Population speaking language i (α_1)	-0.059 (0.120)	-0.049 (0.069)	-0.046 (0.070)	-0.055 (0.070)
Population speaking language j (α_2)	0.543* (0.207)		0.760* (0.123)	0.600* (0.067)
Dummy English (α_E)		3.602* (0.333)		
Dummy French (α_F)		2.202* (0.344)		
Dummy German (α_G)		1.779* (0.351)		
Dummy Spanish (α_S)		0.617 (0.341)	-0.374* (0.041)	-0.340* (0.044)
Distance between languages i and j (α_3)	-0.876* (0.341)	-0.942* (0.199)	-0.915* (0.200)	-0.789* (0.205)
Trade share (α_4)				0.249 (0.134)
Adjusted R^2	0.185	0.726	0.720	0.712

Standard errors are given between brackets, under the coefficients. Starred coefficients are significantly different from 0 at the 5 % (or 1 %) probability level. The number of observations is equal to 46 in all equations. Intercepts are not reported.