Determinants of TOEFL Score:

A Comparison of Linguistic and Economic Factors

By

Myung-Hee Kim

Division of English and Culture, Hanyang University, Korea

Hyun-Hoon Lee

Division of Economics and International Trade, Kangwon National University, Korea

Correspondence: Professor Hyun-Hoon Lee, Division of Economics and International Trade, Kangwon National University, Chuncheon, 200-701, Korea.
Email: hhlee@kangwon.ac.kr

Abstract

This paper aims to investigate, at the cross-country level, the linguistic and economic factors influencing the variation of proficiency in the use of English as a second language. As a proxy for English proficiency, we use the average TOEFL (Test of English as a Foreign Language) scores achieved by residents of a total of 88 countries. The regression results suggest that linguistic factors such as historical affinity and similarity in word order between English and a given language have an influence on the proficiency in English of its users. Among the economic factors, GDP per capita, average years of schooling, and the trade share of GDP have a positive relation with proficiency in English. The results seem robust, even when the computer-based TOEFL score is used in place of the paper-based TOEFL score, and even when the overall score is replaced by the respective score in listening comprehension, structure and written expression, and reading comprehension.

Table of Contents

1. Introduction

This paper aims to investigate, at the cross-country level, the linguistic and economic factors influencing the variation of proficiency in the use of English as a second language. As a proxy for English proficiency, we use the average TOEFL (Test of English as a Foreign Language) score achieved by residents of a total of 88 countries.

To the best of our knowledge, Snow (1998) has been the only study, which attempts to investigate the factors influencing English skills at the cross-country level. The present study extends Snow's (1998) study in many directions. First, while Snow (1998) uses the average TOEFL scores of 26 countries that include only native speakers of 26 different languages, we use the average TOEFL score of 88 countries whose official language is other than English and where data for the explanatory variables are available. Second, while Snow (1998) uses only the overall average TOEFL score, we also use the TOEFL score in three different categories – listening comprehension, structure and written expression, and reading comprehension. Third, instead of a single scalar measure of linguistic distance between English and other languages, we use two distinctive linguistic measures which should be contributing factors to linguistic distance.[1] One is a scalar measure of historical affinity between English and a given language, based on well-established studies on the Indo-European language family. The other is a binary variable to capture the difference in word order between English and a given language.

We find that both linguistic and economic factors have statistically significant impacts on the variation of TOEFL score. Specifically, it is found that residents of a country whose official language is linguistically "closer" to English, in terms of historical affinity and word order, attain a higher level of proficiency in English when other factors remain the same. It is also found that the TOEFL score is influenced by GDP per capita (proxy for the quality of education), average years of schooling (proxy for the quantity of education) and trade share of GDP (proxy for incentives for learning English language skills and level of everyday exposure to the English language).

The rest of the present paper is composed as follows. Section 2 briefly introduces the

---

[1] At the micro level, there have been a number of studies that investigated the effect that linguistic distance has on English language fluency for immigrants in the United States. See, for example, Chiswick and Miller (1998; 2001).

characteristics of the TOEFL and how it is performed in different countries. Section 3 discusses the linguistic and economic variables to be included in the subsequent regression analyses. Results are discussed in Section 4, and concluding remarks are offered in Section 5.

2. The TOEFL: Characteristics and National Performance

The Test of English as a Foreign Language (TOEFL) was initially developed in the early 1960s to measure the language proficiency of nonnative speakers of English who wished to study at colleges and universities in the United States and Canada. It is widely used as a gauge of English proficiency for employment and other purposes, as well. The original TOEFL, a paper-based test (PBT), consisted of three main sections: (a) Listening Comprehension, (b) Structure and Written Expression, and (c) Reading Comprehension. The computer-based test (CBT) was first introduced in 1998 with a writing section added. By using electronic technology, it became possible to enhance assessments by testing more complex skills. Two sections of the test—Listening and Structure—are computer-adaptive, meaning that the test is tailored to each examinee's performance level. A new internet-based version of TOEFL (TOEFL iBT) has been introduced progressively from September, 2005.

Since it was first introduced in 1998, the number of examinees taking the computer-based TOEFL test has been steadily increasing, although there are still people taking the traditional paper-based test. For instance, the number of examinees who took the computer-based TOEFL test from July 2004 through June 2005 was 554,942 whereas the number of examinees taking the paper-based test was 108,808. Of the people who took the CBT test, Korea topped the list, numbering 102,340, followed by Japan with 82,438. On the other hand, of the countries which ran the paper-based test, China had the most examinees, amounting to 67,085.

It is an intriguing fact to researchers as well as ordinary people that there has been a consistent variation among countries in the average TOEFL scores. For example, the average scores of Asian countries are consistently lower than those of European countries. During the period between July 2004 and June 2005, the Asian examinees obtained a mean score of 215.3, whereas the European examinees scored 238.4. On the other hand, in the period from July 1997 to June 1998, the mean score of the paper-based TOEFL test for Asian examinees was 529.5, lower than that for European

4

examinees, which was 563.1. Even within a region, there is a consistent variation among countries. For example, during the period from July 1997 to June 1998, the national mean of China was 560, whereas that of Japan was 498. Six years later, the Chinese got 215 whereas the Japanese got 191 on the TOEFL CBT.

We posit two different kinds of factors to account for these differences: linguistic and economic factors. The next section explains them in more detail.

3. Determinants of National Performance in the TOEFL

A. Linguistic Factors

We consider linguistic relatedness or 'linguistic distance' an important factor in learning a foreign language. The degree of difficulty a French native speaker has when she or he tries to learn English would be much lower than that of a Korean trying to learn English. Thus, the linguistic distance between English and the examinee's language is expected to have an effect on the score of the TOEFL test he or she gets.

Any two languages can be compared in any number of ways: phonetic similarity, shared vocabulary, syntactic affinity, etc. However, it is hard to explicate how to measure the linguistic distance between them quantitatively. For one thing, languages are not really amenable to quantitative treatments, and therefore linguists have been rather reluctant to compare languages by using numbers. For another, there are still a number of languages out there that we do not understand very well, so it is difficult, sometimes even dangerous, to compare a large number of languages by a certain criterion.

With the foregoing discussion in mind, we chose two linguistic factors, (a) historical affinity and (b) word order, for the practical reason that they are probably the most reliable quantitative linguistic measurements at present that can apply to over 4,000 languages around the world (Comrie 1990). If historical affinity is related to the lexical and phonological components of a language, word order has to do with its syntactic component.

Historical Affinity (AFFINITY)
The interest in historical relatedness between languages started during the eighteenth century, when Sir William Jones observed that Sanskrit, an ancient language of India,

had to be related to European languages such as Greek and Latin. The method that historical/comparative linguists use consists in observing systematic correspondences of sounds between significant numbers of words with related meaning in two or more different languages. When such correspondences occur in sufficient numbers, they cannot be due to chance and must mean that the differences in phonological shape must be the result of different phonetic developments from the same or similar word forms, i.e. the two languages must have developed from one and the same language. With this method, it has been possible to establish the 'genetic' relationships between hundreds of languages belonging to the same language family, for instance, the Indo-European Language Family. The hypothetical, ancestral language of all Indo-European languages is called 'Proto-Indo-European' (Cardona et. al (eds.) 1970; Lehman 1974). Although IE languages, which make up a total of 140 languages, are fewer than 3 percent of the number of languages in the world, nearly half of the world's population speaks them (Pyles & Algeo 1993).

We posit the relative distance between a given official language[2] of a country and English from a historical point of view as its Historical Affinity Value. However, it is not straightforward how to quantify the relative distance between two languages, even in the tree. Snow's (1998) IHRE (Index of Historical Relatedness to English) ranges from 1 to 7, with 1 for non-Indo-European languages and from 2 to 7 for various languages within the Indo-European Language Family Tree. The closer a language is to English in the IE tree, the higher value is assigned. One critical problem with this index is that while linguistic distance between two extreme IE languages is 6 (i.e., from 2 to 7), linguistic distance between an IE language with the lowest value and a non-IE language is only 1 (i.e., from 1 to 2), and thus the qualitative difference between IE and non-IE languages is not properly accounted for.

In our system, we assign values from 0 to 10, with 0 for non-IE languages and 10 for English. Contrary to Snow (1998), the minimum assigned value for an IE language is 5, so the smallest linguistic distance between an IE language and a non-IE language is 5. The value discrepancy between Indo-European languages and non-Indo-European

---

[2] Where there are more than two official languages, we chose the most widely-spoken language or an IE language, if any. Typically, the latter case occurred when a given country used to be a colony of an Indo-European country, and the relevant IE language is still used as a lingua franca among the speakers of different languages. It is expected that the language from the colonial period is taught in higher education as a prestigious language.

languages indicates the qualitative difference between them. The so-called Satem languages such as Slavic, Indo-Iranian, and Baltic languages are given value 5. The division of Proto-Indo-European into Centum and Satem sub-families is based on the development, in very ancient times, of the Proto-Indo-European [k] sound.[3] The Italic languages (so-called Romance languages), which had a great effect on English historically, are assigned value 7, whereas other Centum languages such as Hellenic languages are given 6. The Germanic language family, where English belongs, is divided into North Germanic, East Germanic, and West Germanic. Gothic, which is now a dead language, is the only East Germanic language. The North Germanic languages, which include Danish, Swedish, Norwegian, and Icelandic, are assigned value 8. The West Germanic languages, including Dutch, German, Frisian, and Yiddish, are most closely related to English are given value 9.

This method is intuitively appealing because English is closer to other West Germanic languages than the North Germanic languages and is closer to North Germanic languages than the Italic languages. Nonetheless, this method is arbitrary in the sense that the distance from one sub family to the next is assumed to be the same. For instance, the distance from English to Dutch, a West Germanic language, is assumed to be the same as the distance between Dutch and Swedish, a North Germanic language. However, we are not concerned with the "absolute" distance, but with the "relative" distance of different languages from English, and we believe that our scalar measure gives the most plausible "relative" distance between English and other languages.

Figure 1 shows a rough outline of the Indo-European Language Family Tree, including only the relevant parts. Table 1 presents the HAV for each sub-family. A list of all countries with their official languages and the value for historical affinity with English is presented in the Appendix Table.

---

[3] Satem and Centum are, respectively, the Avestan (an ancient Iranian language) and Latin words corresponding to 'hundred.'

Figure 1. Indo-European Language Family Tree
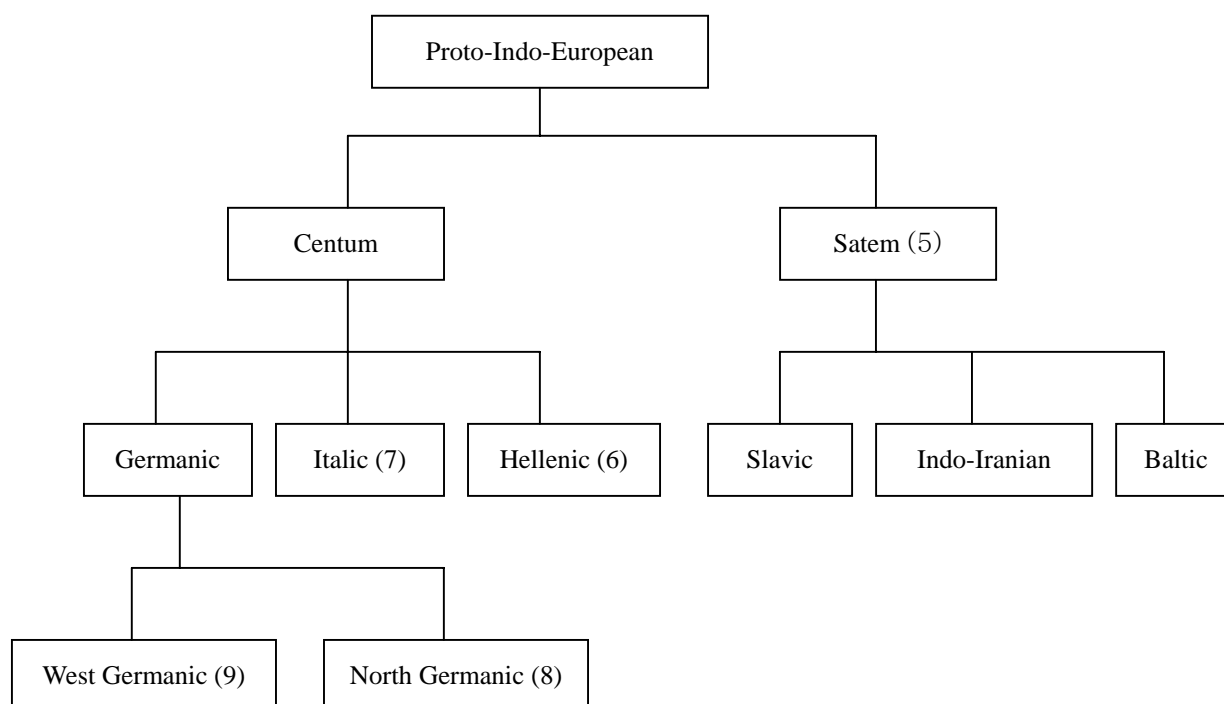
Table 1. Historical Affinity Value (modified from Katzner 2002, p.2)

Lg Family	Subgroup	Branch	Major Languages	HAV
Indo-European	Germanic	West	English, Dutch, German, Flemish	9
		North	Danish, Swedish, Norwegian	8
	Italic		French, Italian, Portuguese, Spanish	7
	Hellenic		Greek	6
	Slavic		Russian, Bulgarian, Polish	5
	Indo-Iranian		Persian, Hindi, Bengali	
	Baltic		Lithuanian, Latvian	
Non-Indo-European				0

Word Order (WordOrder)

The second linguistic factor considered in this paper is word order. Since the pioneering work by Greenberg (1963), there have been a number of studies which explored linguistic universals based on word order characteristics of languages in the world (Dryer 1996, 1997, 2006; Hawkins 1979, 1983; Lehman 1973; Vennemann 1972). Based on a sample of 30 languages, Greenberg proposed 45 linguistic universals, 28 of which refer to the order or position of grammatical relations such as subject, object, and verb. His typology assumed that languages belong to one of three types, SVO (e.g., English), SOV (e.g., Korean), and VSO (e.g., Welsh). Although he described the other three types (VOS, OVS, OSV) as non-existent or extremely rare, all of them are now attested to exist (Keenan 1978; Tomlin 1986; Derbyshire and Pullum 1981). However, our database includes only the three common word order types: SVO, SOV, and VSO.

Determining the basic order of a language can be problematic because many languages exhibit a so-called free or flexible word order. When this happens, a number of criteria have been appealed to in determining basic order: relative frequency, distributional characteristics, pragmatic effect, etc. (Dryer 2006). That is, the more frequent, less restricted and pragmatically neutral word order will be considered basic. In this paper, if an examinee's official language has the same word order as English, namely SVO, the value 1 is assigned. If it is different, the value 0 is assigned. Again, if there are more than two official languages, the most-widely spoken language or the "closest" language to English, if any IE language is on the list, is chosen and assigned the WordOrder value.

A list of all countries with their official languages and the WordOrder value is shown in the Appendix Table.

B. Economic Factors

In addition to the linguistic factors, some economic factors are hypothesized to be related with TOEFL scores.

GDP per capita (PCGDP)

GDP per capita is chosen as a proxy for the overall quality of English education in a given country. That is, it is expected that higher-income countries in general tend to have better facilities for foreign language learning and provide higher-quality teachers for the students, in some cases by recruiting native English teachers. Thus, GDP per

capita is expected to have a positive coefficient when regressed on TOEFL scores. The data are drawn from the World Bank's WDI Online data[4] and is expressed in natural-logarithm form.

Average Years of Schooling (ReSchooling)

Average years of schooling is chosen to proxy the overall quantity of education. It should be noted, however, that the average length of schooling is highly correlated with GDP per capita. Therefore, we first regressed the log of the average years of schooling on the log of GDP per capita and obtained the residuals. We then used the residuals, which we call ReSchooling, as an income-free proxy for the overall quantity of education. That is, the higher the ReSchooling a country has, the longer the years of schooling in that country, in comparison with other countries at the same income level. Hence, it is expected that ReSchooling will have a positive relation with the TOEFL scores, when other factors are constant. The data for average years of schooling are drawn from the World Bank's WDI Online data.

Trade Share of GDP (TRADE)

This is to measure everyday exposure to English and the importance of English in a country, as international trade is conducted primarily in English. That is, residents of a country with a relatively high level of international trade are expected to have more opportunities to be exposed to English and to have more incentives to learn English as a second language, as English skills would give them more chances of getting higher-paying jobs. These data are also drawn from the World Bank's WDI Online data.

C. Other Determinants

Percentage of Population Taking the TOEFL (PerPop)

As in Snow (1998), this variable is to ensure the statistical comparability of TOEFL candidates across countries. It is expected that the smaller the proportion of the population taking the TOEFL, the higher will be the average score of the country, *ceteris paribus*. This variable is, in fact, closely related to GDP per capita (0.34 for 1997-98 and 0.37 for 2004-05, as seen in Table 2), and hence by including the two variables in the same equation, we control for the statistical under-estimation of PCGDP.

---

[4] http://publications.worldbank.org/WDI

<u>Colonial Experience by English-Speaking Countries (ExColony)</u>

If a country becomes a colony, it is in general forced to change its system of politics, economy, society, culture, education, language, etc., and to adopt that of its colonizer. Once that happens, it may persist even after post-colonial freedom. For example, the Philippines and India, which were colonies of the U.S. and U.K. respectively, have been using English as an official language even after they became independent. These English- speaking ex-colonies are deleted from the sample. Even among those countries which were colonies of English-speaking countries and do not use English as an official language may still use English more frequently in many aspects and hence their residents may be exposed to English more frequently than others. Therefore, in order to control for such a possibility, we include a binary variable which takes the value of one if a country was ever a colony of English-speaking countries, and zero otherwise. In our sample, seven countries fall into this category: Bangladesh, Cyprus, Egypt, Jordan, Malaysia, Sudan and Yemen, which were all once colonies of Great Britain. The data are drawn from Centre d'Etudes Prospectives et d'Informations Internationales (CEPII)'s website.[5]

To summarize, our regression equation for TOEFL score is as follows:

$$SCORE = \beta_1 AFFINITY + \beta_2 WordOrder + \beta_3 PCGDP + \beta_4 ReSchooling + \beta_5 TRADE + \beta_6 PerPop + \beta_7 ExColony + \alpha + \varepsilon$$

where
• SCORE = average TOEFL score,[6]
• AFFINITY = index of historical affinity of English to the official language of a country,
• WordOrder = a binary variable which takes one if the official language of a country has the same word order as English, and zero otherwise,
• PCGDP = log of GDP per capita,
• ReSchooling = residuals from the regression of log of average years of schooling on log of GDP per capita,
• TRADE = trade share of GDP,

---

[5] http://www.cepii.fr/distance/geo_cepii.xls
[6] We also tried to take the log of the TOEFL score, but found that the regression results remain unchanged qualitatively.

• PerPop = percentage of population taking the TOEFL,

• ExColony = binary variable which takes one if a country was ever a colony of English speaking countries, and zero otherwise,

• α = constant,

• ε = random disturbance term.

It is hypothesized that signs of all coefficients are positive, except for β_6. It should be noted that explanatory variables are constructed separately for two different years, July 1997 to June 98 and July 2004 to June 2005, based on the paper-based TOEFL scores and the computer-based TOEFL scores, respectively.

The basic descriptive statistics are reported in Table 1.

4. Estimation Results

A. Simple Correlation

Simple correlation among variables is reported in Tables 2 and 3, for 1997-98 and 2004-05. In both years, not only the overall average score but also the average score in the three different categories have a high degree of correlation with the linguistic variables (AFFINITY and WordOrder) and economic variables (PCGDP, ReSchooling, and TRADE). On the other hand, PerPop does not seem to have a correlation with the TOEFL score, while ExColony has a negative correlation with the TOEFL score (-0.23 for 1997-98; -0.15 for 2004-05).

Among the explanatory variables, it is noteworthy that GDP per capita is highly correlated with PerPop (0.34 for 1997-98; 0.37 for 2004-05), AFFINITY (0.37 for 1997-98; 0.36 for 2004-05), and WordOrder (0.21 for 1997-97; 0.22 for 2004-5). That is, high-income countries tend to have a greater proportion of population taking TOEFL, and their official languages are closely related with English in terms of historical affinity and word order. It is also noted that ReSchooling has a close relation with TRADE (0.35 for 1997-98; 0.32 for 2004-05), implying that at the same level of income, countries with longer years of schooling tend to trade more. ExColony, a binary variable for countries which were ever colonies of English speaking countries, has a negative correlation with AFFINITY (-0.27) and WordOrder (-0.34). This may be due to the fact that these countries use mostly non-Indo European languages.

As will be seen in the next sub-section, multicollinearity among explanatory variables is tolerable as significant results are obtained for most explanatory variables. One exception is the collinearity between AFFINITY and WordOrder. In both years, the simple correlation is 0.63, suggesting a close correspondence of historical affinity and word order. Therefore, in the regression analysis reported in the next sub-section, we will include each of the linguistic variables alternatively. In a separate specification, we will first regress AFFINITY on WordOrder and obtain the residual, calling it ReAFFINITY. We will then include both ReAFFINITY and WordOrder in the same equation.

B. Regression

Table 4 summarizes the results of Ordinary Least Squares (OLS) estimation for the overall average score of TOEFL for 88 different countries.[7] The first three columns report the results based on the paper-based TOEFL scores of examinees from July 1997 to June 1998. The last three columns report the results for the average score of the candidates who took the computer-based TOEFL test between July 2004 and June 2005.

Let us first look at the results for 1997-98. As noted above, Column 1 and 2 include AFFINITY and WordOrder alternatively because they are highly correlated. Column 3 includes ReAFFINITY and WordOrder together in the regression.

The overall goodness of fit is about 0.65 in terms of adjusted R^2, suggesting that our model explains about 65 percent of the variation in TOEFL scores among the 88 countries in the sample. As for the individual factors, AFFINITY, WordOrder and WordOrder-adjusted AFFINITY all have estimated coefficients which are positive and significant at the one percent level. Thus, where other determinants of English language proficiency are the same, the closer the linguistic distance in terms of historical affinity and word order, the greater is the candidate's English language proficiency illustrated by the TOEFL score.

Estimated coefficients for PCGDP and ReSchooling also have expected signs and are highly significant, while TRADE has no significant coefficient. The estimated

---

[7] We used the robust errors-correction procedure for heteroskedasticity in the STATA software package.

coefficient for PerPop has a negative sign, as expected, and this is statistically significant. ExColony has a positive coefficient as expected, but is not significant.

The results seem to remain robust, even when the computer-based TOEFL score is used in place of the paper-based TOEFL score. The only difference is that estimates for TRADE become statistically significant (Columns 5 and 6), and PerPop is no longer statistically significant at a reasonable level, even though its coefficient still has the expected negative sign.

Even when the overall score is replaced with the respective scores of Listening Comprehension (Table 5), Structure and Written Expression (Table 6), and Reading Comprehension (Table 7), the results remain qualitatively the same. Thus, we have evidence that the proficiency in listening comprehension, structure and written expression, and reading comprehension is in fact influenced by similar factors.

Table 8 contains the actual and predicted values of the TOEFL scores, as well as the residuals. Countries with positive residuals are those which enjoy over-performance of their residents in the TOEFL test, while countries with negative residuals are those which may need to improve the method of their English teaching. It is worth noting that countries which had positive residuals in 1997-98 tend to have positive residuals also in 2004-05, and *vice versa*.[8] This is consistent with the finding that the determinants of the TOEFL score are more or less the same in both years. This implies that even though the TOEFL test has changed from the paper-based test to the computer-based test, the general difficulty remains roughly the same for the candidates with different linguistic and economic backgrounds.

In the TOEFL test for the period from July 2004 to June 2005, the largest positive residuals are exhibited by Nepal (21.3), Romania (20.0), Morocco (17.8), Bulgaria (16.9), and Bangladesh (16.3), while the largest negative residuals are exhibited by Italy (-31.7), Japan (-24.9), Saudi Arabia (-24.8), Gabon (-23.9), and Cape Verde (-23.7). This may suggest that these under-performers have the greatest potential for improvement of TOEFL scores.

---

[8] Nonetheless, there are some countries whose average TOEFL score has improved, turning minus residuals to positive ones: Algeria, Argentina, , Bangladesh, Benin, Cambodia, Chile, Greece, Indonesia, Laos, Latvia, Malaysia, Mexico, Portugal, Senegal, Tunisia, and Venezuela.

5. Summary and Conclusions

This paper has been concerned with the determinants of English language proficiency at the country level. By using ordinary least squares (OLS) regression analysis, it has been shown that a main reason why countries differ in their residents' proficiency is the linguistic distance of their languages from the English language. Specifically, residents of a country whose official language is linguistically "closer" to English, in terms of historical affinity and word order, have been found to attain a higher level of proficiency in English when other factors remain the same.

The results suggest that knowledge about historical affinity and word order differences between languages is crucial for understanding differences across groups in the acquisition of a foreign language. This is in line with Crystal (1997) which argues "The structural closeness of languages to each other has often been thought to be an important factor in FLL (foreign language learning) (p.371)."

It has also been shown that to some extent English language proficiency illustrated by the TOEFL score differs due to the different quality of education (proxied by GDP per capita), different quantity of education (proxied by average years of schooling), and different incentives for investing in English language skills and different levels of everyday exposure to the English language (proxied by trade share of GDP).

The results hold not only for the overall average TOEFL score but also for different subcategories such as Listening Comprehension, Structure and Written Expression, and Reading Comprehension of the TOEFL.

The results suggest several caveats: First, the fact that the extent to which a language is away from English affects the level of English proficiency of the language users suggests that foreign language learning could be more effective if it began before the mother tongue became the "only" language for children. This is especially true when the mother tongue is very "distant" from the target language. This finding lends support to the argument for early education in a foreign language.

Second, the results of the present study suggest that it is important to enhance the learners' awareness of the differences between their first language and the target

language. Therefore, the learners should be given ample opportunities to understand how the target language differs from their mother tongue in specific areas such as phonological, syntactic, and lexical components of the language.

Finally, as linguistic factors are considered important in understanding a variety of social and economic phenomena in different countries of the world, it has become important to create objective quantitative measurements to compare different languages of the world. Language is the product of the complicated human cognition and the diverse cultures of the people. However difficult it is, the quantification of language phenomena is an important and inevitable job of linguists.

References

Cardona, George, Henry M. Hoenigswald and Alfred Senn, eds., (1970), *Indo-European and Indo-Europeans*, Philadelphia: University of Pennsylvania Press.

Chiswick, B. and P. Miller (1998), "English Language Proficiency among Immigrants in the United States," *Research in Labor Economics*, 17, 151-200.

Chiswick, B. and P. Miller (2001), "A Model of Destination Language Acquisition: Application to Male Immigrants to Canada," *Demography,* 38, 391-409.

Chiswick, B. and P. Miller (2004), "Linguistic Distance: A Quantitative Measure of the Distance Between English and Other Languages," Discussion Paper No. 1246, Institute for the Study of Labor.

Comrie, Bernard ed. (1990), *The World's Major Languages*. New York & Oxford: Oxford University Press.

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

Derbyshire, Desmond and Geoffrey Pullum (1981), "Object-initial Languages." *International Journal of American Linguistics*, 47, 192-214.

Dryer, Matthew S. (1996), "Word Order Typology." in J. Jacobs ed., *Handbook on Syntax, Vol. 2*, 1050 – 1065, New York: Walter de Gruyter.

Dryer, Matthew S. (1997), "On the 6-way Word Order Typology." *Studies in Language* 21: 69-103.

Dryer, Matthew S. (2006), "Word order," in Timothy Shopen ed., *Language Typology and Syntactic Description, Vol. 1. Clause Structure*, Cambridge: Cambridge University Press.

Greenberg, Joseph (1963), "Some Universals of Grammar with Particular Reference to the Order of Meaningful Elements," in Joseph Greenberg ed., *Universals of Language*,

73-113. Cambridge, MA: MIT Press.

Hawkins, John A. (1979), "Implicational Universals as Predictors of Word Order Change," *Language*, 55, 618-648.

Hawkins, John. A. (1983), *Word Order Universals*, New York: Academic Press.

Hutchinson, William (2001), "Linguistic Distance as a Determinant of Bilateral Trade," Working Paper No. 01-W30, Department of Economics, Vanderbilt University.

Katzner, Kenneth (2002), *The Languages of the World*, 3rd edition. London & New York: Routledge.

Keenan, Edward (1978), "The Syntax of Subject-final Languages." in Winfred P. Lehman ed., *Syntactic Typology: Studies in the Phenomenology of Language*, 267-327, Austin: University of Texas Press.

Lehman, Winfred P. (1973), "A Structural Principle of Language and its Implications," *Language*, 49, 42-66.

Lehman, Winfred P. (1974), *Proto-Indo-European Syntax*, Austin: University of Texas Press.

Pyles, Thomas & John Algeo (1993), *The Origins and Development of the English Language*, 4th Edition, Fort Worth: Harcourt College Publishers.

Snow, Marcellus S. (1998), "Economic, Statistical, and Linguistic Factors Affecting Success on the Test of English as a Foreign Language (TOEFL)," *Information Economics and Policy*, 10, 159-172.

Tomlin, Russell (1986), *Basic Word Orders: Functional Principles*, London: Croom Helm.

Vennemann, Theo (1972) "Explanation in Syntax." in John Kimball ed., *Syntax and Semantics*, Vol. 2, 1-50. New York: Academic Press.

Table 1. Descriptive Statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
Score (1997−98)	88	541.9	31.9	475.0	612.0
Listening (1997−98)	88	54.4	3.6	46.0	63.0
Structure (1997−98)	88	53.9	3.2	46.0	61.0
Reading (1997−98)	88	54.3	3.1	46.0	60.0
Score (2004−05)	88	223.2	20.5	172.0	262.0
Listening (2004−05)	88	22.5	2.4	16.0	27.0
Structure (2004−05)	88	22.1	2.0	17.0	26.0
Reading (2004−05)	88	22.3	2.1	17.0	26.0
AFFINITY	88	4.5	3.3	0.0	9.0
WordOrder	88	0.8	0.4	0.0	1.0
PCGDP (1997)	88	7.6	1.6	4.6	10.5
PCGDP (2004)	88	7.8	1.5	4.6	10.6
ReSchooling (1997)	88	0.3	2.0	−4.8	5.0
ReSchooling (2004)	88	0.1	1.9	−4.6	5.0
TRADE (1997)	88	59.2	29.9	14.6	162.7
TRADE (2004)	88	64.4	32.0	19.9	174.8
PerPop (1997)	88	0.0	0.1	0.0	0.4
PerPop (2004)	88	0.0	0.0	0.0	0.2
ExColony	88	0.1	0.3	0.0	1.0

Table 2. Correlation Matrix – 1997-98

	Score	Listening	Structure	Reading	PCGDP	ReSchooling	TRADE	PerPop	ExColony	AFFINITY	WordOrder
Score	1.00										
Listening	0.94	1.00									
Structure	0.97	0.84	1.00								
Reading	0.96	0.84	0.94	1.00							
PCGDP	0.65	0.68	0.57	0.61	1.00						
ReSchooling	0.39	0.38	0.34	0.38	0.02	1.00					
TRADE	0.17	0.20	0.13	0.16	0.02	0.35	1.00				
PerPop	0.00	0.03	0.01	−0.05	0.34	−0.07	−0.02	1.00			
ExColony	−0.23	−0.17	−0.23	−0.25	−0.09	−0.16	0.03	0.29	1.00		
AFFINITY	0.51	0.48	0.50	0.51	0.37	−0.05	−0.10	0.00	−0.27	1.00	
WordOrder	0.45	0.40	0.44	0.46	0.21	0.13	0.06	−0.03	−0.34	0.63	1.00

Table 3. Correlation Matrix – 2004-05

	Score	Listening	Structure	Reading	PCGDP	ReSchooling	TRADE	PerPop	ExColony	AFFINITY	WordOrder
Score	1.00										
Listening	0.95	1.00									
Structure	0.96	0.88	1.00								
Reading	0.94	0.84	0.89	1.00							
PCGDP	0.66	0.63	0.57	0.68	1.00						
ReSchooling	0.40	0.42	0.40	0.32	0.01	1.00					
TRADE	0.25	0.30	0.25	0.19	0.07	0.32	1.00				
PerPop	0.10	0.08	0.11	0.08	0.37	−0.05	−0.08	1.00			
ExColony	−0.15	−0.11	−0.13	−0.22	−0.10	−0.16	0.01	0.17	1.00		
AFFINITY	0.50	0.44	0.47	0.54	0.36	−0.06	−0.15	0.00	−0.27	1.00	
WordOrder	0.41	0.34	0.39	0.46	0.22	0.12	0.10	−0.06	−0.34	0.63	1.00

Table 4. The Determinants of TOEFL Score - Total

	Year 1997-98			Year 2004-05		
	(1)	(2)	(3)	(4)	(5)	(6)
WordOrder	20.60***		21.58***	11.43***		12.19***
	(3.87)		(4.21)	(3.31)		(3.84)
AFFINITY		3.15***			2.26***	
		(4.60)			(5.37)	
AFFINITY residual			2.36***			2.11***
			(2.79)			(3.97)
PCGDP	13.30***	11.77***	11.92**	8.68***	7.42***	7.44***
	(9.12)	(7.86)	(8.02)	(8.87)	(7.83)	(7.81)
ReSchooling	4.97***	5.62***	5.39***	3.82***	4.12***	4.09***
	(4.49)	(5.23)	(5.02)	(5.10)	(6.01)	(5.91)
TRADE	0.03	0.07	0.05	0.04	0.09**	0.08*
	(0.37)	(0.95)	(0.76)	(0.85)	(2.11)	(1.95)
PerPop	-112.21***	-98.32**	-102.21***	-60.11	-46.76	-46.72
	(2.75)	(2.48)	(2.60)	(1.55)	(1.31)	(1.30)
ExColony	3.45	1.78	4.59.	5.01	5.54	6.01
	(0.40)	(0.22)	(0.55)	(0.93)	(1.15)	(1.22)
CONSTANT	424.30***	434.73***	-431.96***	144.43***	149.36***	150.28***
	(36.06)	(38.72)	(37.14)	(18.74)	(21.38)	(20.76)
Number of Observations	88	88	88	88	88	88
Adjusted R^2	0.637	0.659	0.665	0.633	0.693	0.690

Notes: 1. All estimates are made with the Ordinary Least Squares method. 2. See the text for the definitions of the variables. 3. t-statistics are shown in parentheses. 4. ***, **, and * denote one, five, and ten percent level of significance, respectively, for a two-tailed test.

Table 5. The Determinants of TOEFL Score - Listening

	Year 1997-98			Year 2004-05		
	(1)	(2)	(3)	(4)	(5)	(6)
WordOrder	1.96***		2.07***	0.92**		1.01**
	(3.36)		(3.67)	(2.18)		(2.57)
AFFINITY		0.32***			0.23***	
		(4.25)			(4.39)	
AFFINITY residual			0.25***			0.24***
			(2.71)			(3.70)
PCGDP	1.63***	1.47***	1.49**	1.01***	0.87***	0.87***
	(10.19)	(9.00)	(9.08)	(8.43)	(7.41)	(7.35)
ReSchooling	0.56***	0.62***	0.61***	0.46***	0.48***	4.49***
	(4.61)	(5.32)	(5.13)	(4.98)	(5.71)	(5.69)
TRADE	0.01	0.01	0.01	0.01	0.01**	0.01**
	(0.84)	(1.38)	(1.24)	(1.48)	(2.51)	(2.52)
PerPop	-12.88***	-11.49***	-11.81***	-8.26*	-6.71	-6.71
	(2.88)	(2.66)	(2.73)	(1.73)	(1.51)	(1.51)
ExColony	1.04	0.93	1.17.	0.74	0.90	0.85.
	(1.11)	(1.05)	(1.28)	(1.12)	(1.51)	(1.39)
CONSTANT	40.10***	41.10***	-40.91***	13.38***	13.80***	-14.05***
	(31.03)	(33.47)	(31.95)	(14.15)	(15.93)	(15.65)
Number of Observations	88	88	88	88	88	88
Adjusted R^2	0.661	0.684	0.686	0.598	0.656	0.652

Notes: 1. All estimates are made with the Ordinary Least Squares method. 2. See the text for the definitions of the variables. 3. t-statistics are shown in parentheses. 4. ***, **, and * denote one, five, and ten percent level of significance, respectively, for a two-tailed test.

Table 6. The Determinants of TOEFL Score - Structure

	Year 1997-98			Year 2004-05		
	(1)	(2)	(3)	(4)	(5)	(6)
WordOrder	2.14***		2.24***	1.17***		1.25***
	(3.36)		(3.62)	(3.06)		(3.46)
AFFINITY		0.33***			0.22***	
		(4.03)			(4.69)	
AFFINITY residual			0.25**			0.20***
			(2.47)			(3.39)
PCGDP	1.15***	0.98***	1.00**	0.69***	0.57***	0.57***
	(6.57)	(5.47)	(5.56)	(6.35)	(5.27)	(5.26)
ReSchooling	0.44***	0.51***	0.49***	0.36***	0.39***	0.39***
	(3.32)	(3.93)	(3.74)	(4.39)	(5.07)	(4.97)
TRADE	-0.00	0.00	0.00	0.00	0.01*	0.01*
	(0.00)	(0.49)	(0.34)	(0.85)	(1.94)	(1.77)
PerPop	-8.51*	-7.05	-7.44***	-2.95	-1.66	-1.66
	(1.74)	(1.48)	(1.57)	(0.68)	(0.41)	(0.41)
ExColony	0.04	-0.12	0.16	0.54	0.58	0.64.
	(0.04)	(0.12)	(0.16)	(0.91)	(1.06)	(1.14)
CONSTANT	43.63***	41.72***	-44.45***	15.52***	16.03***	16.01***
	(30.96)	(33.09)	(31.60)	(18.13)	(20.16)	(19.55)
Number of Observations	88	88	88	88	88	88
Adjusted R^2	0.493	0.519	0.523	0.512	0.572	0.568

Notes: 1. All estimates are made with the Ordinary Least Squares method. 2. See the text for the definitions of the variables. 3. t-statistics are shown in parentheses. 4. ***, **, and * denote one, five, and ten percent level of significance, respectively, for a two-tailed test.

Table 7. The Determinants of TOEFL Score – Reading

	Year 1997-98			Year 2004-05		
	(1)	(2)	(3)	(4)	(5)	(6)
WordOrder	2.14***		2.23***	1.38***		1.45***
	(3.99)		(4.29)	(3.81)		(4.28)
AFFINITY		0.31***			0.24***	
		(4.47)			(5.28)	
AFFINITY residual			0.22**			0.20***
			(2.57)			(3.54)
PCGDP	1.25***	1.11***	1.12**	0.92***	0.80***	0.81***
	(8.53)	(7.27)	(7.46)	(9.03)	(7.91)	(7.96)
ReSchooling	0.47***	0.54***	0.51***	0.30***	0.33***	0.32***
	(4.23)	(4.91)	(4.70)	(3.80)	(4.51)	(4.39)
TRADE	-0.00	0.01	0.00	0.00	0.01	0.01
	(0.27)	(0.84)	(0.63)	(0.31)	(1.53)	(1.26)
PerPop	-13.07***	-11.69***	-12.14***	-7.32*	-6.06	-6.05
	(3.19)	(2.90)	(3.05)	(1.80)	(1.59)	(1.59)
ExColony	0.21	-0.01	0.32	0.02	-0.01	0.11
	(0.24)	(0.01)	(0.38)	(0.03)	(0.02)	(0.21)
CONSTANT	43.12***	44.19***	-43.84***	14.01***	14.58***	-14.56***
	(36.37)	(38.65)	(37.16)	(17.39)	(19.50)	(18.92)
Number of Observations	88	88	88	88	88	88
Adjusted R^2	0.617	0.632	0.642	0.628	0.673	0.674

Notes: 1. All estimates are made with the Ordinary Least Squares method. 2. See the text for the definitions of the variables. 3. t-statistics are shown in parentheses. 4. ***, **, and * denote one, five, and ten percent level of significance, respectively, for a two-tailed test.

Table 8. Actual vs. Predicted Scores in 1997-97 and 2004-05

Country	Actual (1997–98)	Predicted (1997–98)	Residual (1997–98)	Actual (2004–05)	Predicted (2004–05)	Residual (2004–05)
Albania	540	535.6	4.4	219	217.1	1.9
Algeria	514	518.5	−4.5	208	207.5	0.5
Argentina	569	572.6	−3.6	245	241.6	3.4
Armenia	553	537.8	15.2	219	220.8	−1.8
Austria	597	575.6	21.4	257	244.4	12.6
Azerbaijan	532	510.9	21.1	216	205.0	11.0
Bangladesh	510	514.2	−4.2	227	210.7	16.3
Belarus	555	558.7	−3.7	235	237.5	−2.5
Belgium	599	601.9	−2.9	257	268.8	−11.8
Benin	515	520.5	−5.5	208	207.0	1.0
Bolivia	544	558.9	−14.9	224	235.1	−11.1
Brazil	553	565.8	−12.8	230	237.3	−7.3
Bulgaria	580	543.1	36.9	242	225.1	16.9
Burkina Faso	506	497.3	8.7	190	190.2	−0.2
Burundi	532	506.0	26.0	215	199.3	15.7
Cambodia	514	514.2	−0.2	206	204.2	1.8
Cape Verde	496	547.9	−51.9	203	226.7	−23.7
Chile	556	562.5	−6.5	236	235.4	0.6
China	560	520.4	39.6	215	205.6	9.4
Colombia	545	542.6	2.4	221	221.5	−0.5
Costa Rica	578	546.4	31.6	233	224.9	8.1
Cote d'Ivoire	507	517.1	−10.1	200	205.9	−5.9
Croatia	575	557.6	17.4	242	232.5	9.5
Cyprus	529	520.2	8.8	228	224.3	3.7
Czech Republic	567	565.5	1.5	238	239.8	−1.8
Denmark	607	586.7	20.3	261	250.9	10.1
Egypt	531	525.9	5.1	220	214.5	5.5
El Salvador	551	545.2	5.8	222	223.9	−1.9
Estonia	580	560.4	19.6	244	235.8	8.2
Finland	596	588.7	7.3	257	252.8	4.2
France	557	568.3	−11.3	237	237.1	−0.1
Gabon	501	553.6	−52.6	204	227.9	−23.9
Georgia	545	524.9	20.1	218	207.8	10.2

Germany	593	580.1	12.9	253	246.4	6.6
Greece	548	559.1	−11.1	234	233.1	0.9
Hungary	569	552.8	16.2	236	229.1	6.9
Iceland	589	575.9	13.1	252	243.9	8.1
Indonesia	517	526.5	−9.5	214	208.7	5.3
Iran	535	528.7	6.3	215	216.8	−1.8
Italy	554	569.1	−15.1	205	236.7	−31.7
Japan	498	532.8	−34.8	191	215.9	−24.9
Jordan	517	526.1	−9.1	206	221.8	−15.8
Kazakhstan	536	550.0	−14.0	221	229.1	−8.1
Korea (South)	522	526.5	−4.5	215	217.3	−2.3
Kyrgyzstan	537	548.7	−11.7	221	229.9	−8.9
Laos	506	515.1	−9.1	204	201.0	3.0
Latvia	554	562.5	−8.5	238	236.4	1.6
Lebanon	540	535.4	4.6	225	226.7	−1.7
Lithuania	556	564.5	−8.5	230	238.0	−8.0
Malaysia	530	541.9	−11.9	230	228.5	1.5
Mali	475	503.6	−28.6	183	196.1	−13.1
Mauritania	503	495.3	7.7	199	194.1	4.9
Mexico	551	552.1	−1.1	230	227.0	3.0
Moldova	554	539.0	15.0	237	225.0	12.0
Mongolia	502	511.6	−9.6	199	207.4	−8.4
Morocco	516	506.5	9.5	217	199.2	17.8
Mozambique	512	507.5	4.5	199	200.4	−1.4
Nepal	521	508.2	12.8	224	202.7	21.3
Netherlands	612	584.2	27.8	262	251.2	10.8
Niger	497	497.0	0.0	172	190.8	−18.8
Norway	586	581.5	4.5	258	247.3	10.7
Panama	547	548.8	−1.8	223	224.8	−1.8
Paraguay	537	549.9	−12.9	224	228.2	−4.2
Peru	544	559.5	−15.5	227	233.6	−6.6
Poland	558	562.7	−4.7	230	235.5	−5.5
Portugal	575	575.1	−0.1	252	242.8	9.2
Romania	581	550.0	31.0	249	229.0	20.0
Russia	555	550.3	4.7	231	227.1	3.9
Saudi Arabia	476	513.9	−37.9	177	201.8	−24.8

Senegal	490	515.2	−25.2	212	204.4	7.6
Slovakia	566	559.7	6.3	240	237.8	2.2
Slovenia	586	572.4	13.6	256	242.3	13.7
Spain	566	574.2	−8.2	235	241.7	−6.7
Sudan	515	486.2	28.8	195	187.5	7.5
Sweden	585	591.8	−6.8	249	257.2	−8.2
Switzerland	581	579.9	1.1	250	246.3	3.7
Syria	530	506.0	24.0	204	198.7	5.3
Tajikistan	510	531.8	−21.8	206	223.5	−17.5
Thailand	502	530.9	−28.9	202	217.0	−15.0
Togo	503	520.8	−17.8	193	206.2	−13.2
Tunisia	527	530.5	−3.5	219	217.0	2.0
Turkey	537	517.4	19.6	217	205.5	11.5
Ukraine	552	549.5	2.5	231	231.4	−0.4
Uruguay	572	567.3	4.7	244	238.1	5.9
Uzbekistan	529	516.5	12.5	219	208.3	10.7
Venezuela	539	545.5	−6.5	224	221.8	2.2
Vietnam	511	526.9	−15.9	204	214.7	−10.7
Yemen	488	505.6	−17.6	184	202.8	−18.8

Appendix Table. Historic Affinity and Word Order of Official Languages

Country	Official language	Language family of official languages	Historic Affinity	Word order (a)	Word order (b)
Albania	Albanian	Satum	5	SVO	1
Algeria	Arabic	Non−Indo−European	0	VSO	0
Argentina	Spanish	Italic	7	SVO	1
Armenia	Armenian	Satum	5	SVO	1
Austria	German	West Germanic	9	SVO	1
Azerbaijan	Azerbaijani	Non−Indo−European	0	SOV	0
Bangladesh	Bengali	Satum	5	SOV	0
Belarus	Belarusian, Russian	Satum	5	SVO	1
Belgium	Dutch, French, German		9	SVO	1
Benin	French	Italic	7	SVO	1
Bolivia	Spanish, North Bolivian Quechua, South Bolivian Quechua, Central Aymara	Italic	7	SVO	1
Brazil	Portuguese	Italic	7	SVO	1
Bulgaria	Bulgarian, Turkish	Satum	5	SVO	1
Burkina Faso	French	Italic	7	SVO	1
Burundi	Rundi, French	Italic	7	SVO	1
Cambodia	Khmer	Non−Indo−European	0	SVO	1
Cape Verde	Portuguese, Kabuverdianu	Italic	7	SVO	1
Chile	Spanish	Italic	7	SVO	1
China	Chinese Mandarin	Non−Indo−European	0	SVO	1
Colombia	Spanish	Italic	7	SVO	1
Costa Rica	Spanish	Italic	7	SVO	1
CÔte d'Ivoire	French	Italic	7	SVO	1
Croatia	Croatian, Italian	Italic	7	SVO	1
Cyprus	Greek, Turkish	Hellenic	6	SVO	1

Czech Republic	Czech	Satum	5	SVO	1
Denmark	Danish, German	West Germanic	9	SVO	1
Egypt	Arabic	Non-Indo-European	0	VSO	0
El Salvador	Spanish	Italic	7	SVO	1
Estonia	Estonian	Non-Indo-European	0	SVO	1
Finland	Finnish, Swedish	North Germanic	8	SVO	1
France	French	Italic	7	SVO	1
Gabon	French	Italic	7	SVO	1
Georgia	Georgian	Non-Indo-European	0	SVO	1
Germany	German	West Germanic	9	SVO	1
Greece	Greek	Hellenic	6	SVO	1
Hungary	Hungarian	Non-Indo-European	0	SVO	1
Iceland	Icelandic	North Germanic	8	SVO	1
Indonesia	Indonesian	Non-Indo-European	0	SVO	1
Iran	Persian [Western Farsi]	Satum	5	SOV	0
Italy	Italian	Italic	7	SVO	1
Japan	Japanese	Non-Indo-European	0	SOV	0
Jordan	Arabic	Non-Indo-European	0	VSO	0
Kazakhstan	Kazakh, Russian	Satum	5	SVO	1
Korea (South)	Korean	Non-Indo-European	0	SOV	0
Kyrgyzstan	Kirghiz, Russian	Satum	5	SVO	1
Laos	Lao	Non-Indo-European	0	SVO	1
Latvia	Latvian	Satum	5	SVO	1
Lebanon	Arabic, French	Italic	7	SVO	1
Lithuania	Lithuanian	Satum	5	SVO	1
Malaysia	Malay	Non-Indo-	0	SVO	1

		European			
Mali	<u>French</u>, Bamanankan, Bomu, Tieyaxo Bozo, Toro So Dogon, Maasina Fulfulde, Hasanya Arabic, Mamara Senoufo, Kita Maninkakan, Soninke, Koyraboro, Senni Songhay, Syenara Senoufo, Tamasheq, Xaasongaxango	Italic	7	SVO	1
Mauritania	Hassaniya Arabic	Non-Indo-European	0	VSO	0
Mexico	Spanish	Italic	7	SVO	1
Moldova	Romanian	Italic	7	SVO	1
Mongolia	Mongolian	Non-Indo-European	0	SOV	0
Morocco	Arabic	Non-Indo-European	0	VSO	0
Mozambique	Portuguese	Italic	7	SVO	1
Nepal	Nepali	Non-Indo-European	0	SOV	0
Netherlands	<u>Dutch</u>, Achterhoeks, Drents, Western Frisian, Gronings, Limburgisch, Sinte Romani, Vlax Romani, Sallands, Stellinwerfs, Twents, Veluws, Western Yiddish	West Germanic	9	SVO	1
Niger	<u>French</u>, Arabic, Fulfulde, Gourmanchema, Hausa, Manga Kanuri, Tamajaq, Zarma	Italic	7	SVO	1
Norway	Norwegian	North Germanic	8	SVO	1

Panama	Spanish	Italic	7	SVO	1
Paraguay	Guarani, Spanish	Italic	7	SVO	1
Peru	Spanish	Italic	7	SVO	1
Poland	Polish	Satum	5	SVO	1
Portugal	Portuguese	Italic	7	SVO	1
Romania	Romanian	Italic	7	SVO	1
Russia	Russian	Satum	5	SVO	1
Saudi Arabia	Arabic	Non-Indo-European	0	VSO	0
Senegal	Balanta-Ganja, Hassaniyya, Jola-Fonyi, Mandinka, Mandjak, Mankanya, Noon, Puaar, Serer-Sine, Soninke, Wolof, French	Italic	7	SVO	1
Slovakia	Slovak	Satum	5	SVO	1
Slovenia	Slovenian, Hungarian, Italian	Satum	5	SVO	1
Spain	Spanish	Italic	7	SVO	1
Sudan	Arabic	Non-Indo-European	0	VSO	0
Sweden	Swedish	North Germanic	8	SVO	1
Switzerland	French, German, Italian, Romansch	West Germanic	9	SVO	1
Syria	Arabic	Non-Indo-European	0	VSO	0
Tajikistan	Tajiki	Satum	5	SOV	0
Thailand	Thai	Non-Indo-European	0	SVO	1
Togo	Éwé, Kabiye, French	Italic	0	SVO	1
Tunisia	Arabic	Non-Indo-European	0	VSO	0
Turkey	Turkish	Non-Indo-European	0	SOV	0
Ukraine	Ukrainian	Satum	5	SVO	1
Uruguay	Spanish	Italic	7	SVO	1

Uzbekistan	Uzbek	Non-Indo-European	0	SOV	0
Venezuela	Spanish	Italic	7	SVO	1
Vietnam	Vietnamese	Non-Indo-European	0	SVO	1
Yemen	Arabic	Non-Indo-European	0	VSO	0

Note: Where there are more than two official languages, we took the language which is "closest" to English. Underlined are those languages.