*Estudios de
lingüística inglesa aplicada*

elia

LEXICAL SIMILARITY LEVEL BETWEEN ENGLISH AND PORTUGUESE

Maria Isabel Maldonado García
University of the Punjab, Pakistan
spanishprofessor1@gmail.com

Ana Maria Borges de Souza
University of the Punjab, Pakistan
ana_borges@live.com

Although the studies which calculate the similarity index between languages are abundant, to date the similarity index between English and Portuguese has not been calculated. When calculating the similarity index between English and Portuguese it is important to have in consideration the amount of Latin loan words English has received from Latin which may influence the similarity index.

However, English is a Germanic language whereas Portuguese is a Romance language. The purpose of this research is to calculate this level of lexical similarity through the computational lexical distance comparison of 500 high frequency words in English and Portuguese and the identification of cognates. The main hypothesis of this study is based on the fact that English and Portuguese degree of genetic difference may be less than 30% considering that the comparison of English with other

Romance languages has resulted in similar indexes

The study is aimed at assisting students of English or Portuguese as an L2 with their language acquisition through the identification of items which present similarities in both languages and in this manner aid vocabulary acquisition.

***Key words**: Cross-language similarities, lexical distance, genetic difference between languages, cognates, loan words, language acquisition.*

Se han realizado múltiples estudios para revelar el nivel de similitud entre diferentes lenguas. Sin embargo, el nivel de similitud entre el inglés y el portugués no se ha calculado todavía. El inglés ha recibido una considerable cantidad de préstamos del latín, lo cual puede aumentar el nivel de similitud entre las dos lenguas debido a que el portugués deriva directamente del latín. Sin embargo, el inglés es una lengua germánica, mientras que el portugués es una lengua romance. El propósito de esta investigación es calcular este nivel de similitud léxica a través de la comparación de la distancia computacional léxica de 500 palabras de alta frecuencia del inglés y portugués y la identificación de cognados. La principal hipótesis de este estudio se basa en el hecho de que el inglés y el portugués presentan cierto grado de diferencia genética, y como tal, su coeficiente de similitud probablemente será inferior al 30%. El estudio tiene como objetivo ayudar a los estudiantes de inglés o portugués como L2 en la adquisición de la lengua a través de la identificación de elementos que presentan similitud en ambas lenguas.

***Palabras clave**: Similitudes entre lenguas, distancia léxica, diferencia genética entre lenguas, cognados, préstamos léxicos, adquisición de lenguas.*

1. Introduction

This article is one of a series of investigations on cross-language similarities. The methodology utilized in this article replicates that utilized

previously by one of the researchers in other studies[1].Portuguese and English are Indo-European languages. Which means that both languages are genetic relatives. The comparison is relevant due to the fact that since in the last few years the use of English has become widespread all over the world it has come in close contact with the English language and Portuguese as well as Brazilian students are finding themselves forced to learn it. Furthermore, the data for the comparison was easily available and obtainable.

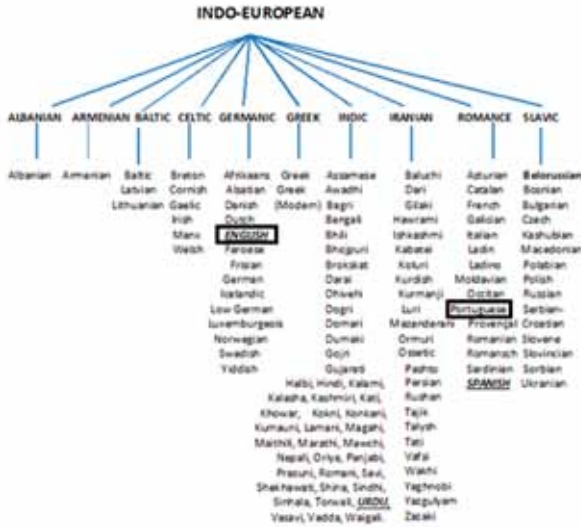The following figure illustrates the (relatively distant) relationship between English and Portuguese[2].

Figure1. Indoeuropean languages- the genetic relationship between English and Portuguese.

The analysis of the similarity is done through cross-language cognate recognition. The variety utilized in this case is the Modern Standard European Portuguese variety as supposed to the Brazilian variety of Portuguese language. The particular reason for using this variety is

that its use is more extended than the Brazilian variety. This is due to the fact that it is the preferred variety of the African countries not just of the European Union. In terms of English language, the same parameters have been utilized. The variety utilized for the corpus is that of American English which seems to be the variety with the largest number of speakers[3].

The word cognate derives from the Latin *cognatus* and according to the Oxford Dictionary it has two definitions; one

> 'having the same linguistic derivation as another' and 'formal related; connected: related to or descended from a common ancestor[4].'

Different authors have attempted to define the term over time. Johnston defined cognate as 'born together, kindred, same stock' (1939: 4). This definition is, of course, linguistically inaccurate, misleading and leaving too many open possibilities. Holmes & Ramos define it as "items of vocabulary in two languages which have the same roots and can be recognized as such" (1993: 88).

For them, a common etymology is crucial for recognition of cognates. If we adjust strictly to this definition, the question arises of what would be the situation of cross-language synonyms, which are also phonologically and orthographically similar, although etymologically they are not. For Whitley, on the other hand, "a given word W from Language X and a word W from Language Y are termed 'cognates' if and only if they have been inherited from the same ancestor language of X and Y" (2002: 305).

And if their similarity is a coincidence they are not considered true cognates (2002: 305). Cognates are those terms which have traveled down the family tree as languages evolve and develop. They have proven of considerable importance for determining genetic relationships among languages not only in terms of contemporary languages but also in determining what the parent language of sibling languages might have looked like.

On the other hand, different authors have different viewpoints on the same topic. Lotto & de Groot define cognates as similar words, from

a phonological and /or morphological aspect, to their equivalent in the second language (1998: 31-69). In this sense, the authors do not consider etymology necessary for a pair of terms to be cognates. Similarly, Diane August et al. refer to the identification of cognates through their linguistic overlap according to phonetics, semantics and orthography (2005: 52).

Both groups of definitions are equally acceptable; those which require a common origin as well as those whose level of similarity are based on semantics, morphology, phonology and spelling.

Lexical solidarities are formed by semantic relationships between words that are apparent in combined restrictions. Coseriu in his book, *Las Solidaridades Léxicas*, in an effort to clarify the study of Porzig on lexical solidarities, defined them as:

> determinación semántica de una palabra por medio de una clase, un archilexema o un lexema, precisamente, en el sentido de que una clase determinada, un determinado archilexema o un determinado lexema funciona como un rasgo distintivo de la palabra considerada. Dicho de otro modo, se trata del hecho de que una clase, un archilexema o un lexema pertenece a la definición semántica de esa palabra, en el plano de las diferencias semánticas mínimas. (1977: 148)

For our lexical similarity study, one of the important aspects of lexical semantic relations and solidarity, (apart from the etymology) is undoubtedly synonymy. Baldinger considered synonymy as an important issue in semantics (1970: 205). López García states that synonymy is based on onomasiology, one meaning can be evoked by several signs (1990, 42). The semantic relation of synonymy between two terms is one of the principal basis of cognate identification. According to Francisco Marcos-Marín (agreeing with López García) (1990) the definition of synonymy stands as: "different signs have the same meaning" (1980: 452).

In this research we will identify 500 high frequency words in English (American variant) and find the corresponding terms in Portuguese language (Modern European variant). Until now the lexical similarity between Portuguese and English has not been established. On the other

hand, we do know the similarity level between Spanish and Portuguese which is of 89%. Both languages are Romance languages, hence the elevated similarity level since they are genetic siblings. However, English is a Germanic language and since Portuguese is a Romance language as well as French, and the similarity level between English and French is of 27%, we expect the similarity level to be less than 30%.

Ethnologue states about lexical similarities

The percentage of lexical similarity between two linguistic varieties is determined by comparing a set of standardized wordlists and counting those forms that show similarity in both form and meaning. Percentages higher than 85% usually indicate a speech variant that is likely a dialect of the language with which it is being compared[5].

It further states that lexical similarity is bidirectional and reciprocal. Furthermore, this method can be applied to any language pair combination, no matter how distant its genetic relationship. If the scripts of the languages are different, the method is applied through the phonetic characterization of each language. This can also be done if the languages are written in the same script, although in this case we have applied it using orthography.

1.1. Other similarity indexes between English and other languages

Ethnologue establishes some levels of similarities among certain languages[6] as can be seen in the following table:

	German S.I.	French S.I.	Russian S.I.	Portuguese S.I.
English	60%	27%	24%	-

Table 1.Lexical similarity indexes (S.I.)

According to this table, the similarity index between German and English is of 60% as compared to 27% between English and French

and 24% between English and Russian. The elevated similarity between English and German is mostly due to the fact that both languages belong to the Germanic family of languages. French, on the contrary, belongs to the Romance family of languages while Russian to the East Slavic family of languages. The elevated percentages indicate that those languages belong to the same family of languages and are closely related. In this context, the rationale of this research is to calculate the S.I. between English and Portuguese.

2. Method

Different methods can be used for cognate identification; such as those utilized by Holmes & Ramos (1993); Rama, List (2012); and others. The majority of linguistic survey studies require the collection of word lists. For lexical similarity assessment, semantics, as well as sound/orthographic correspondences need to be established. "Cognate identification is usually based on a similarity or distance score (e.g., edit-distance) calculated from the number of matches and mismatches in the alignment" (List, 2012).

Some of the initial investigations on calculation of cross-language lexical similarities occurred in 1948, when Morris Swadesh manually applied his lexicostatistical method in a series of articles (Swadesh, 1948; 1950; 1952; 1955). At the beginning Swadesh started his list including 215 items and later on, in 1955 reduced it to the length of 100. These lists were not made taking into consideration high frequency use of different terms within a language, but based on Swadesh's experience and capabilities as a linguist. There have been studies in this regard later on, such as Oswalt (1971) and the Automated Similarity Judgement Program (ASJP) by Brown, Holman, Wichmann & Velupillai (2008).

Our motivation for conducting this study is mainly based on two principal factors. One that there are limited studies in regards to cross-language similarities percentages (principally those of Ethnologue), and two, that these percentages and common vocabulary lists constitute in

fact, a great source of knowledge and useful tool, which second language instructors can utilize for the teaching of target language vocabulary to students such in Dressler et al. (2011); Hancin-Bhatt & Nagy (2004); as well as Ramirez, Chen & Pasquarella (2013).

For the purpose of this research we will follow the method, already used by one of the authors of this research in her doctoral thesis, which is similar to Ethnologue and ASJP, although it included a lengthier list of high frequency words. In this case we will use the Levenshtein algorithm, a highly efficient dynamic programming algorithm, utilized by Kessler (1995) for computational comparison of dialects.

The lexical-statistic method (Swadesh, 1952) is as follows:

1. 500 high frequency words of English language will be identified. The reason why only high frequency words will be utilized is their importance within a language. This means that high frequency words usually do not follow phonic rules and are more resistant to change or evolution than others. That is, they evolve slower than the rest.

2. The Portuguese equivalents of these terms will be identified and both lists of terms put on an excel spreadsheet with the purpose of string comparison and calculation of edit distance.

3. The Levenshtein, or edit distance will be calculated orthographically in this case, as the script of both languages is the Latin script. If the languages had different scripts the phonetics of the languages would be compared since the method is applicable to all languages. For each pair of words in the third column the numerical figure of the corresponding edit distance will be placed. This method is a popular alignment method and has been successful in determining differences in pronunciation in phonetic strings (Kessler, 1995; Heeringa, 2004)

The Levenshtein distance between two strings is defined as the minimum number of operations utilized for conversion of one string into another, where the operations may be deletion, insertionor substitution of a character.

From a mathematical point of view the distance between two strings of symbols a and b is given by lev a,b (|a|, |b|) where:

$$lev_{a,b}(i,j) = \begin{cases} \max(i,j) & \min(i,j) = 0, \\ \min \begin{cases} lev_{a,b}(i-1,j)+1 \\ lev_{a,b}(i,j-1)+1 \\ lev_{a,b}(i-1,j-1)+[a_i \neq b_j] \end{cases} \end{cases}$$

Figure 2. Levenshtein algorithm[7]

Utilizing this formula, we have produced the following computed example:

In this example we illustrate how many changes it will take to convert one string into the other.

		G	A	T	O
	0	1	2	3	4
G	1	1	2	3	4
A	2	2	1	2	3
T	3	3	2	1	2
E	4	4	3	2	1

Figure 3. Computed example of Levenshtein distance between the strings 'gato' and 'gate'.

The Levenshtein distance between '*gato*' and '*gate*' is of 1 since the following modifications transform one into the other one and there is no physical method to transform it with less than 1 modification:

1. Gato → Gate (substitution of o for *e*)

Another example with distance 3:

> 1.science → ciencia (deletion of *s*)
> 2.cience → ciencia   (substitution of e for *i*)
> 3.cienci → ciencia   (insertion of *a*)

1. At this point cognate identification can be performed according to orthographic similarity measures. The measures will be broken down according to different categories of similarity groups. The reason for this is that two pairs of terms may present the same distance and in fact one pair can present a more elevated similarity than the other one. This is due to the length of every string. The longer the length of the word pairs is and the smaller the distance between the words, the higher the percentage of similarity between two lexical strings. When collecting and processing word lists, certain surveyors may determine intuitively the similarity of certain words, for which we have utilized the criteria described in point number 2.

2. In order to determine similarity, a number of categories have been set. This was previously implemented by Blair (1990). In our case, the following categories have been set:

Category 1:

a. Exact cognates. Those which present 0 distance (100% similarity index)

b. Terms which differ only on one character or have a distance of 1, and the rest of the characters occur in the same position in every term.

c. Terms which differ on two characters or present a distance of 2.

Category 2:

a. Terms which present a distance of 3 or more characters.

Category 3:

a. Terms which are not orthographically similar

3. Results

The identification of the high frequency terms was done through a search of the internet in google search engine, according to the following search criteria: 1. Recent date of the corpus, 2. List of high frequency words in English language, 3. The search was directed towards finding lists with contained an appropriately large number of words and so a high frequency English words list was identified which constitutes the corpus of this study. The search yielded a 500 high frequency word list which constitutes the corpus of our investigation[8].

Once the words were identified, their equivalents in Portuguese were identified and compared in a spreadsheet. In this manner, the words from English were placed in the first column and in the second column the words from Portuguese were placed. The third column was reserved for the Levenshtein distance.

The table presents the following format:

English	Portuguese	Distance
a	um	2
able	capaz	5
about	sobre	5

Table 2. Edit distance between English and Portuguese

The results are as follows:

Category 1:

a. Terms which are Exact Cognates (Distance 0 or 100% similarity): *Animal-animal, oh-oh, real-real* (3 pairs).

b. Terms which differ only on one character or distance 1: *Air-ar, American Americano, animals-animais, area-área, at-a, class-classe, complete-completo, fact-facto, fine-fino, for-por, idea-ideia, important-importante, just-justo, list-lista, long-longo, map-mapa, name-nome, no-não, or-ou, order-ordem, paper-papel, part-parte, plants-plantas, power-poder, problem-problema, simple-simples, special-especial, top-topo*, (28 pairs).

c. Terms which differ on two characters or present a distance of 2: *Ball-bola, car-carro, center-centro, course-curso, day-dia, different-diferente, example-exemplo, form-formar, front-frente, group-grupo, hour-hora, line-linha, live-viver, me-mim, my-meu, not-não, on-em, person-pessoa, point-ponto, poor-pobre, put-pôr, question-questão, rock-rocha, six-seis, sun-sol, system-sistema, use-usar,* (27 pairs).

Category 2:

d. Terms which present a distance of 3 or more characters:

Distance 3: *Certain-certo, common-comum, distance-distância, during-durante, English-Inglês, family-família, fire-fogo, found-fundar, language-língua, letter-letra, mind-mente, more-mais, new-novo, number-número, river-rio, second-segundo, space-espaço, state-estado, study-estudar, sure-seguro, three-três, time-tempo, voice-voz* (23 pairs).

Distance 4: *Boat-barco, city-cidade, cut-cortar, great-grande, mother-mãe, much-muitos, night-noite, now-agora, other-outro, page-página, past-passado, possible-possível, round-redondo, school-escola, stay-estadia, surface-superfície* (16 pairs).

Distance 5: *Finally-finalmente, really-realmente, story-história* (3 pairs).

Distance 6: *There are no terms which present a distance of 6.*

Distance 7: *Longer-mais longo* (1 pair).

Distance 8: *There are no terms which present a distance of 8.*

Distance 9: *There are no terms which present a distance of 9.*

Distance 10: *There are no terms which present a distance of 10.*

Distance 11: *United Estates-Estados Unidos* (1 pair)[9].

Category 3:

e. Terms which are not orthographically similar: *The rest.*

After the number of pairs was computed and added in order to calculate the similarity index (SI):

DISTANCE	NUMBER OF PAIRS
0	3
1	28
2	27
3	23
4	16
5	3
7	1
11	1
TOTAL	**102**

Table 3. Total number of pairs which present lexical similarity

The analysis of the pairs according to the mentioned categories revealed that 102 pairs do, in fact, present similarity and can be used for the purpose of second language acquisition.

TOTAL LEXICAL SIMILARITY INDEX ENGLISH-PORTUGUESE	
LANGUAGE PAIRS	**ENGLISH-PORTUGUESE**
N. Pairs	102 pairs
TOTAL S.I.	20.4%

Table 4. Total lexical similarity index

4. Conclusion

In this study we described and applied a statistically driven algorithm for the task of calculating lexical similarity among genetically related languages. Languages are compared on the basis of word strings. A string consists of the concatenation of linguistic signs (orthographical). Attempting to quantify distances in orthography between languages, we based the comparison and measurements on orthographic segment relations. The representations were made on the basis of segments being exact or equal, consecutively followed with one degree of variation. The degree of variation augments progressively as the edit-distance augments as well. The higher the number of variations, the distance or difference between strings is larger. The results show the relations between the elements.

The lexical comparison result yielded the computation of 102 pairs in which lexical similarity has been observed. The percentage of lexical similarity between English and Portuguese or S.I. is then of 20.4%.

	German S.I.	French S.I.	Russian S.I.	Portuguese S.I.
English	60%	27%	24%	**20.4%**

Table 5. Improved lexical similarity indexes

From the point of view of phylo-genetics, Portuguese and English are Indo-European languages from different families. As was our hypothesis, a S.I. of less than 30% has been revealed through the analysis and comparison of the word lists. With a S.I. of 20.4% we can conclude that the two languages are distantly related in spite of loan words from Latin in English.

From the point of view of second language acquisition, the importance of this research rests on the fact that the resulting similarity list of terms or shared vocabulary (Annex 3) will be a useful tool for students learning Portuguese or English as an L2. It will be an important aid for

students whose L1 is either English or Portuguese and who are learning the other language as an L2.

As has been proven, the findings of the present study have important implications on the fields of phylo-genetics (since in the process of linguistic evolution all languages receive influence from others, suffer a process of evolution in which contact, borrowing and imposition occurs, as well as mutations in the form of innovations, phonological shifts, etc. and in this research we have calculated the similarity index of both English and Portuguese) as well as second language acquisition. The encouragement of vocabulary acquisition through cognate knowledge, reciprocally will be facilitated in this manner.

The limitation of this study is based on the fact that while the English vocabulary was of high frequency, the corresponding Portuguese vocabulary may or may not be high frequency. This is yet to be determined in a future study where high frequency lists can be used for both languages[10].

Acknowledgements

Notes

[1] Maldonado García, María Isabel (2013). Comparación del Léxico Básico del Español, el Inglés y el Urdu. Doctoral Dissertation. Madrid: UNED.

[2] The Balto-Slavic branch is notrepresented in this figure due to lack of space andirrelevance.

[3] "Table 53. Languages Spoken At Home by Language: 2009", The 2012 Statistical Abstract (U.S. Census Bureau), retrieved 1-10-2014

[4] Oxford Dictionary. http://www.oxforddictionaries.com/definition/english/cognate?q=cognate. Retrieved on 17-1-2014.

[5] Ethnologue is a language database which contains information about more than seven thousands languages in a statistical form.

Ethnologue: http://www.ethnologue.com/about/language-info

[6] Information in this table has been extracted from Ethnologue: http://www.ethnologue.com

[7] Levenshtein, V.I. (1965). Binary codes capable of correcting spurious insertions and reversals. Cybernetics and Control Theory, 10:707–710.

[8] Date of identification of the list 6-6-2011. List published by Freedom Elementary High School, Clovis, California. http://freedom.cusd.com/documents/
Can be retrieved from the website.

[9] If we had compared this pair United with Unidos and Estates with Estados the similarity level would have been higher. However, the original order in the languages has been taken into consideration.

[10] As in Frequency Dictionary of Portuguese (2007). The dictionary is based on the 20 million words from the 1900s portion of the 45 million word Corpus do Português. It is the first frequency dictionary of Portuguese that is based on a large corpus from several different genres. (Co-authored with Prof. Ana Preto-Bay of the Department of Spanish and Portuguese at BYU, and published by Routledge.)

References

August, D., Carlo, M., & Calderon, M. (2005). Development of literacy in Spanish-speaking English-language learners: Findings from longitudinal study of elementary school children. *Perspectives, 31*(2) 52.

Baldinger, K. (1970). *Teoría semántica. Hacia una semántica moderna*. Madrid: Alcalá.

Blair, F. (1990). Survey on a shoestring: A manual for small-scale language survey. Arlington: Summer Institute of Linguistics and the University of Texas at Arlington.

Brown, C., Holman, E., Wichmann, S., & Velupillai, V. (2008). Automated classification of the world's languages: A description of the method and preliminary results. *Sprachtypologie und Universalienforschung, 61* (4), 285-308.

Coseriu, E. (1977). Las Solidaridades Léxicas. *Principios de semántica estructural*. Madrid: Gredos.148.

Dressler, C., Carlo, M., Snow, C., August, D., & White, C. (2011). Spanish-speaking students' use of cognate knowledge to infer the meaning of English words. *Bilingualism, Language, and Cognition, 14* (2), 243-255.

Gutiérrez Ordóñez, S. (1989). *Introducción a la semántica funcional*. Madrid: Síntesis.131.

Hancin-Bhatt, B., & Nagy, W.(1994).Lexical transfer and second language morphological development. *Applied Psycholinguistics, 15*, 289-310.

Heeringa, W. (2004). *Measuring dialect pronunciation differences using Levenshtein distance*. Unpublished doctoral dissertation, Rijksuniversiteit, Groningen.

Holmes, J. & Ramos, R. (1993). False friends and reckless guessers: Observing cognate recognition strategies. In T. Huckin, M. Haynes, & J. Coady (Eds.), *Second Language Reading and Vocabulary Learning* (pp. 86-108). Norwood, NJ: Ablex.

Johnston, Marjorie Cecil (1939). *Cognate Relationships Between English and Spanish Vocabularies as a Basis for Instruction*. Austin, TX: Unpublished Doctoral Dissertation.P.4.

Kessler, Brett (1995). Computational dialectology in Irish Gaelic. *In Proceedings of the seventh conference on European chapter of the Association for Computational Linguistics*, Inc.60-66, San Francisco: Morgan Kaufmann.

Levenshtein, V. I. (1965). Binary codes capable of correcting spurious insertions and reversals. *Cybernetics and Control Theory,* 10, 707-710.

Lewis, M. P. (2009). *Ethnologue: Languages of the World*. 17th edition. Dallas: SIL International.

List, Johann-Mattis (2012). Lexstat: Automatic detection of cognates in multilingual wordlists. *In Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*. Avignon: Association for Computational Linguistics. 117-125.Online page: http://www.alweb.org/anthology/W12-0216 Accessed on May 13, 2014.

Lotto, L., & De Groot, A. (1998). Effects of learning method and word type on acquiring vocabulary in an unfamiliar language. *Language Learning, 48* (1), 31-69.

López García, Ángel. (1990). Sinonimia intralingüística y sinonimia interlingüística. *Actas del Primer Coloquio Internacional de Traductología,* 41-45.

Maldonado García, M.I. (2013). *Comparación del léxico básico del español, el inglés y el urdu*. Unpublished Doctoral Dissertation, UNED, Madrid.

Marcos Marín, F. (1980). *Curso de Gramática Española*. Madrid: Cincel.452.

Rama, T., & Borin, L. (2014). N-Gram Approaches to the Historical Dynamics of Basic Vocabulary. *Journal of Quantitative Linguistics, 21* (1), 50-64.

Rama, Taraka, Kolachina, Prasant & Kolachina Sudheer (2013) *Two methods for automatic identification of cognates*. In Proceeding of Quantitative Investigations in Theoretical Linguistics (QITL-5). Leuven, Belgium. 76.

Ramirez, G., Chen, X., & Pasquarella, A. (2013). *Cross-Linguistic Transfer of Morphological Awareness in Spanish-Speaking English Language Learners: The Facilitating Effect of Cognate knowledge*. Top Lang Disorders. 33 (1), 73-92.

Oswalt, R. (1971). Towards the construction of a standard lexicostatistic list. *Anthropological Linguistics, 13* (9), 421-434.

Oxford English Dictionary. Retrieved on 17-01-2014 from http://www.oxforddictionaries.com/definition/english/cognate?q=cognate

Sunderman, G., & Schwartz, A. (2008). Using cognates to investigate cross-language competition in second language processing. *TESOL Quarterly, 42,* 527-536.

Swadesh, M. (1948).*The time value of linguistic diversity*. Paper presented at the Viking Fund Supper Conference for Anthropologists, New York.

Swadesh, M. (1950). Salish internal relationships. *International Journal of American Linguistics, 16* (4), 157-167.

Swadesh, M. (1952). Lexico-statistic dating of prehistoric ethnic contacts: With special reference to North American Indians and Eskimos. *Proceedings of the American Philosophical Society, 96* (4), 452-463.

Swadesh, M. (1955). Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics, 21* (2), 121-137.

US Census Bureau. Language spoken at home: 2010 American Community Survey 1-year estimates. Available at: http://factfinder2.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_10_1YR_S1601&prodType=table . Accessed 2011-12-27.

Whitley, M. S. (2002). *Spanish/English Contrasts: A Course in Spanish Linguistics*. 2nd Edition. Washington, DC: Georgetown University Press.