

Investigating a Common Lexical Core Across International Varieties of English

Leah Gilner* • Franc Morales**

Abstract

The objective of this investigation was to obtain lists of the most frequent words found across seven varieties of English as represented by corpora from the International Corpus of English (ICE; International Corpus of English, 2010). To this end, we developed a number of computer programs that have allowed us: First, to obtain a complete frequency list out of each corpus; second, to lemmatize each frequency list according to two criteria, namely, inflections and word families; and, third, to identify the most frequent words and word families across varieties. Results indicate that the seven varieties investigated share a relatively small common lexical core that accounts for 7 to 8 out of every 10 words used by fluent language users.

Background

Yano (2001: 129) proposes that English for international communication should be, among other things, “as socioculturally neutral as possible in order to attain high learnability and usability”. Indeed, pioneering research findings indicate that, when speakers use English as a lingua franca (ELF), they tend to adjust their speech so as to better match their interlocutor’s expectations (Jenkins, 2000; Seidlhofer, 2004), just as they do in the L1 (Dell and Brown, 1991). The result is that ELF communication often involves rephrasing and drawing on expressions and forms known by all participants (Cogo and Dewey, 2006). That is to say, speakers assess and make use of what they share in order to facilitate communication.

Yamaguchi (2002) argues that it is “extremely important to devise a common core of English to enable mutual intelligibility among the many varieties” since students need training that prepares them to be both global and local users of English. Jenkins’ work (2000) is a classic in this respect as it has intended “...to set up a common core for everyone to access and acquire” in the realm of the phonology of spoken interactions. Her identification of such a common core

* 准教授／応用言語学

** 語学教育IT研究開発者／応用言語学

led to a set of suggestions for English language instruction. Her suggestions were based on data collected over a long period of time, recordings and observations of colleagues and students, analyzed for communication breakdowns/trouble spots.

We propose that identifying linguistic common ground shared across speakers of world Englishes is one legitimate way to approach English language teaching (ELT) of English as an International Language (EIL). Seidlhofer (2003) speaks of a need for “comprehensive and reliable description of salient features of EIL” and “...what is needed is a description of EIL features as a basis for eventual codification” (p. 14). It is with this in mind that the study described here focuses on varieties of English and patterns of word use found in colingual (Schell, 2008) communication. Specifically, we investigated lexical frequency distributions in discourse from seven varieties of English. Our work parallels Jenkins’s efforts (2000) at the phonological level with this study at the lexical level. Our aim is to identify a common lexical core that facilitates communication.

The study

For this investigation, we have been given access by the Director of the International Corpus of English (ICE), Professor Gerald Nelson, to corpora corresponding to the following varieties: Canada, East Africa, Hong Kong, India, Jamaica, the Philippines, and Singapore. To ensure coherence among individual corpora, ICE enforces certain guidelines. Specifically, each corpus contains 500 texts of approximately 2,000 words each collected from 1990 on; 70% of a given corpus reflects spoken discourse represented by 100 private and 80 public dialogs as well as 120 unscripted monologues while 30% captures written discourse in the form of 30 letters and 20 student writings along with 150 printed texts originating in instructional, academic, literary, newspaper, and other domains. Speakers are both male and female, 18 years old or older, and educated in the respective country. In this manner, ICE provides a means of analyzing spoken and written discourse of a particular variety (Nelson, 2010).

The Canada corpus was compiled by the University of Alberta. It contains approximately 1.1 million tokens representing Canadian English. The East Africa corpus was compiled by the University of Bayreuth and Chemnitz University of Technology. It contains approximately 1.4 million tokens representing Tanzanian and Kenyan English. The Hong Kong corpus was compiled by Hong Kong Polytechnic University, the University of Hong Kong, and the Chinese University of Hong Kong. It contains approximately 1.5 million tokens. The Jamaica corpus was compiled by the University of Freiburg (Germany). It contains approximately 1.1 million tokens representing Jamaican English. The India corpus was compiled by Freie Universität Berlin. It contains approximately 1.1 million tokens representing Indian English. The Philippine corpus was compiled by De La Salle University-Manila. It contains approximately 1.1 million tokens. The

Singapore corpus was compiled by The National University of Singapore. It contains approximately 1.1 million tokens representing Singaporean English. Please refer to the manuals of each corpus regarding specifics. Together, these corpora add up to approximately 8.5 million running words.

We investigated seven varieties because that was what was freely available from the International Corpus of English at the time of this study. When it was only five varieties, we worked with five varieties and we obtained similar results. Once the International Corpus of English includes additional varieties, we will again repeat our analyses. In other words, the task is an on-going one and will evolve to encompass analyses of other varieties as the data becomes available. These are the first of many steps that will need to be taken in order to arrive at an understanding of how English is used worldwide and, most importantly, how to provide our students with the best possible instruction.

The study, then, is based on corpora that amount to approximately 8.5 million running words or tokens. The key terminology to understand the data in this paper will be given an overview at this point so as to facilitate the interpretation of the analyses presented hereafter. In particular, we will briefly introduce the concepts of *type*, *token*, *token coverage*, and *lemmatization* for those unfamiliar with or unsure of the terminology.

A *type* is simply a word. For example, the phrase “*the woman on the stage*” has 4 types because the word “the” is repeated twice. Unique or different words are types. A token is the occurrence of a word. The example “the woman on the stage” has 5 tokens or running words. *Token coverage* is the percentage of running words or tokens that a given type provides. In the given example, “the woman on the stage”, 4 types account for all tokens, that is, they provide 100% token coverage of the phrase. However, not all types provide equal token coverage. The type or unique word “the” provides 40% token coverage of the sentence while the remaining 3 types provide 20% token coverage each. Therefore, putting aside other factors for the moment, the measure referred to as token coverage signifies the comparative need to know a particular word or type by its sheer quantitative presence in a given text. When considering texts of larger size than the example, such as passages, textbooks, or corpora, token coverage provides one criterion with which to assess the characteristics of the vocabulary at work.

Lemmatization can be interpreted and implemented in several ways. For this study, we have used two kinds of lemmatization each of which has provided its own statistical results. The first kind of lemmatization involved counting as a single type words corresponding to inflections of a given word, specifically, verbs (3rd person and tense forms), nouns (plural form), and adjectives (comparative and superlative forms). Irregular cases have also been taken into consideration. In sum, and for example, the type *jump* stands for the types *jump*, *jumps*, *jumping*, and *jumped*; the type *letter* stands for the types *letter* and *letters*; and the type *wide* stands for the types *wide*, *wider*

and *widest*. When we refer to a lemmatized type, inflected type, or lemma in this study, we are referring to the “base” form that represents all inflections.

The second kind of lemmatization involved counting as a single type words corresponding to inflections and derivations of a given word, that is, to the headword of a word family (Palmer, 1931; Nation, 2001). Specific to this study, a headword represents its own form in addition to inflections and derivations according to Bauer and Nation’s level 6 (1993). Level 6 “...includes the inflections, and the high-frequency, regular, productive, and transparent derivational affixes. These affixes can only be added to free forms” (Nation, 2006). In sum, and for example, the type access stands for the types, *access*, *accessed*, *accesses*, *accessing*, *accessible*, *inaccessible*, *accessibility*, *inaccessibility*. For this study, we will use the terms headword type, headword, and word family interchangeably.

Table 1 The amount of tokens and lemmatized types in each of the 7 corpora.

Varieties	Tokens	Lemmatized types
Canada	1,101,720	27,927
East Africa	1,416,183	26,596
Hong Kong	1,458,979	22,838
India	1,132,174	31,044
Jamaica	1,084,228	28,091
Philippines	1,132,360	28,029
Singapore	1,116,397	23,849

Table 1 shows the amount of lemmatized types that are used by speakers of each variety. There is a range of “wealth”, so to speak, regarding the amount of vocabulary used in each corpus. The extremes are India and Hong Kong and the interpretation is that the India corpus (and/or variety) captures a richer vocabulary than the Hong Kong corpus (and/or variety). If we assumed that the corpora faithfully reflected the varieties under investigation, we could aver that the vocabulary choices made by English speakers in India are more varied than those made by English speakers in Hong Kong. However, the likely source of the variance is the relatively small size of the corpora.

Table 2 The token coverage of each variety by its corresponding most frequent lemmatized types.

Types	Canada	East Africa	Hong Kong	India	Jamaica	Philippines	Singapore	Average
100	57.61%	54.45%	58.44%	54.50%	57.16%	54.75%	57.00%	56.27%
200	65.01%	62.04%	66.46%	62.00%	64.54%	62.20%	64.44%	63.81%
500	74.65%	73.00%	76.46%	72.15%	74.52%	72.48%	74.37%	73.95%
1,000	81.78%	81.46%	83.88%	80.08%	82.07%	80.09%	81.87%	81.61%
1,500	85.76%	85.98%	87.86%	84.32%	86.18%	84.25%	85.96%	85.76%
2,000	88.32%	88.80%	90.38%	87.09%	88.77%	87.01%	88.63%	88.43%
2,500	90.11%	90.73%	92.12%	89.02%	90.57%	88.93%	90.51%	90.29%
3,000	91.44%	92.14%	93.40%	90.46%	91.90%	90.39%	91.93%	91.67%
3,500	92.48%	93.23%	94.39%	91.58%	92.92%	91.55%	93.03%	92.74%
4,000	93.31%	94.08%	95.17%	92.49%	93.74%	92.49%	93.92%	93.60%
4,500	94.00%	94.77%	95.79%	93.23%	94.40%	93.27%	94.63%	94.30%
5,000	94.58%	95.35%	96.31%	93.85%	94.96%	93.93%	95.23%	94.89%

Table 2 presents the token coverage of each variety by its corresponding most frequent inflected types, that is, each corpus has been profiled with its own list of the most frequent lemmatized words. Those familiar with these kinds of statistics will quickly see correlations with other studies. The 100 most frequent lemmas account for approximately half of the running words in each corpus. Adding the following 100 most frequent lemmas (2nd row, labeled 200) accounts on average for about 10% more tokens. That is, the most frequent 100 inflected types occur 5 times as often as the following 100 most frequent. As we continue to look down the rows we see this decreasing trend become gradually more evident.

Indeed, if we look at the rows labeled 2,500 and 3,000 respectively, we see that 500 more types only account for an average of 1% of the running words in each corpus. It is important to remark, once again, that each corpus has been profiled with a list of its own most frequent lemmas. That is, we have yet to investigate the token coverage provided by the shared frequent lemmas in all varieties. In order to arrive at such a list, we looked at the intersection of all the variety-specific lists, that is, at those lemmas that were most frequent as well as shared by all varieties. We have identified a list of approximately 1,100 inflected types or lemmas that provides, what we consider, adequate token coverage and concrete potential for instruction and learning. The work is preliminary and remains subject to revision and update.

Table 3 shows the token coverage the proposed list offers of each corpus together with the token coverage a similar amount of the most frequent lemmas of each variety provides for its own variety. In other words, and for example, the proposed list offers 81.45% token coverage of the Hong Kong corpus while the list of most frequent lemmas in the Hong Kong corpus provides 84.85% token coverage of the Hong Kong corpus. The difference in coverage is shown below to be 3.41%, that is, one third of a single word in a phrase of 10 words and about 3 words in a passage of 100 words.

Table 3 The token coverage of the proposed list as well as variety-specific lists.

Varieties	Canada	East Africa	Hong Kong	India	Jamaica	Philippines	Singapore
Canada	82.74%	n/a	n/a	n/a	n/a	n/a	n/a
East Africa	n/a	82.56%	n/a	n/a	n/a	n/a	n/a
Hong Kong	n/a	n/a	84.85%	n/a	n/a	n/a	n/a
India	n/a	n/a	n/a	81.12%	n/a	n/a	n/a
Jamaica	n/a	n/a	n/a	n/a	83.07%	n/a	n/a
Philippines	n/a	n/a	n/a	n/a	n/a	81.09%	n/a
Singapore	n/a	n/a	n/a	n/a	n/a	n/a	82.85%
Proposed List	80.77%	81.52%	81.45%	79.59%	81.43%	79.54%	81.49%
Differential	1.97%	1.04%	3.41%	1.53%	1.64%	1.55%	1.36%

The row labeled Differential represents the amount of token coverage lost by using a single word list (the proposed list) rather than word lists specific to each variety. While the differential percentage is relevant, it does not detract from the observation that all word lists account for 8 out of 10 words in every phrase, 80 out of 100 words in every passage, and so on. In all cases but Hong Kong values fluctuate around 1.5%.

It is, therefore, safe to say that far more is common than it is different. Also, do keep in mind that the proposed list only contains those words found to be most frequent in all varieties. If, for instance, we would have included lemmas found in all varieties but not necessarily equally frequent across, the proposed list would have obtained a higher token coverage.

Up to this point, we have shown the results of the investigation following the first kind of lemmatization, namely, when a type accounts for itself as well as its inflections. We will now look at how the results vary when using the second kind of lemmatization, namely, when a type is a *headword*, representing itself as well as its family members.

Table 4 The amount of tokens and headword types in each of the 7 corpora.

Varieties	Tokens	Headword types
Canada	1,101,720	23,319
East Africa	1,416,183	21,584
Hong Kong	1,458,979	18,429
India	1,132,174	26,347
Jamaica	1,084,228	23,678
Philippines	1,132,360	23,229
Singapore	1,116,397	19,260

The amount of tokens cannot change from Table 1 since the total amount of running words remains the same. The amount of headword types naturally does change; a quick comparison between Tables 1 and 4 shows that there are several thousand fewer headword types than inflected types. This is unsurprising since a word family includes derivations as well as inflections, in other words, inflected types are a subset of word families. This observation does not change the fact that the India and Hong Kong varieties remain the extremes regarding vocabulary “wealth”. Interestingly, and despite an average drop of over 4,000 types, the standard deviation remains remarkably similar at 2,793 for inflected types and 2,735 for word families.

Table 5 presents the token coverage of each variety by its corresponding most frequent headwords types, that is, each corpus has been profiled with its own list of the most frequent word families. Table 5 is remarkably similar to Table 2 in terms of values and trends. The 100 most frequent headwords also account for approximately half of the running words in each corpus. Adding the following 100 most frequent headwords (2nd row, labeled 200) also accounts on average

for about 10% more tokens. That is, the most frequent 100 word families also occur 5 times as often as the following 100 most frequent word families. As we continue to look down the rows we also see this decreasing trend become gradually more evident.

Table 5 The token coverage of each variety by its corresponding most frequent headword types.

Types	Canada	East Africa	Hong Kong	India	Jamaica	Philippines	Singapore	Average
100	58.86%	55.51%	60.33%	55.59%	58.30%	56.47%	57.78%	57.55%
200	66.45%	63.50%	68.20%	63.21%	65.89%	63.92%	65.62%	65.25%
500	76.65%	75.52%	78.69%	74.17%	76.60%	74.65%	76.48%	76.11%
1,000	84.25%	84.58%	86.53%	82.65%	84.66%	82.74%	84.55%	84.28%
1,500	88.31%	89.17%	90.53%	87.11%	88.79%	87.05%	88.75%	88.53%
2,000	90.82%	91.84%	92.91%	89.83%	91.27%	89.80%	91.34%	91.11%
2,500	92.47%	93.54%	94.44%	91.64%	92.92%	91.60%	93.07%	92.81%
3,000	93.64%	94.73%	95.52%	92.91%	94.08%	92.91%	94.31%	94.01%
3,500	94.52%	95.58%	96.31%	93.86%	94.94%	93.91%	95.23%	94.91%
4,000	95.21%	96.24%	96.91%	94.60%	95.61%	94.70%	95.95%	95.60%
4,500	95.77%	96.75%	97.38%	95.21%	96.13%	95.33%	96.51%	96.15%
5,000	96.23%	97.16%	97.75%	95.71%	96.55%	95.84%	96.96%	96.60%

And if we look at the rows labeled 2,500 and 3,000 respectively, we also see that 500 more word families only account for an average of 1% of the running words in each corpus. Again, it is important to remark that each corpus has been profiled with a list of the most frequent word families corresponding to each corpus. That is, we have yet to investigate the token coverage provided by the shared frequent word families in all varieties. In order to arrive at such a list, we looked at the intersection of all lists, that is, at those word families that were most frequent as well as shared by all varieties. We have identified a list of approximately 1,200 types that provides, what we consider, adequate token coverage and concrete potential for instruction and learning. As before, the work is preliminary and remains subject to revision and update.

Table 6 shows the token coverage the proposed list of headwords offers of each corpus together with the token coverage a similar amount of the most frequent headwords of each variety provides for its own variety. In other words, and for example, the proposed list offers 85.25% token coverage of the Hong Kong corpus while the list of most frequent word families in the Hong Kong corpus provides 88.40% token coverage of the Hong Kong corpus. The difference in coverage is shown below to be 3.15%, that is, one third of a single word in a phrase of 10 words and about 3 words in a passage of 100 words. Despite the overall increase in token coverage provided by headwords versus lemmas (please compare Tables 2 and 6), the difference between averages of corpus-specific lists versus the proposed lists remain similar and less than 0.3%.

Table 6 shows the token coverage by the proposed list as well as variety-specific lists.

Varieties	Canada	East Africa	Hong Kong	India	Jamaica	Philippines	Singapore
Canada	86.12%	n/a	n/a	n/a	n/a	n/a	n/a
East Africa	n/a	86.73%	n/a	n/a	n/a	n/a	n/a
Hong Kong	n/a	n/a	88.40%	n/a	n/a	n/a	n/a
India	n/a	n/a	n/a	84.73%	n/a	n/a	n/a
Jamaica	n/a	n/a	n/a	n/a	86.61%	n/a	n/a
Philippines	n/a	n/a	n/a	n/a	n/a	84.70%	n/a
Singapore	n/a	n/a	n/a	n/a	n/a	n/a	86.49%
Proposed List	82.26%	84.99%	85.25%	82.43%	83.91%	83.12%	83.72%
Differential	3.86%	1.73%	3.15%	2.30%	2.70%	1.58%	2.76%

As in Table 3, the row labeled *Differential* represents the amount of token coverage lost by using a single headword list (the proposed list) rather than headword lists specific to each variety. As before, while the differential percentage is relevant, it does not detract from the observation that all word lists account for 8 out of 10 words in every phrase, almost 84 (on average) out of 100 words in every passage, and so on. In all cases differential values fluctuate around 2.58% with a standard deviation of 0.87%.

A more detailed comparison of Tables 3 and 6 shows that the average token coverage of each corpus by its corresponding list yields 82.85% for inflected types and 86.49% for headword types. This difference contrasts with the average token coverage provided by the proposed list, that is, 81.49% for inflected types and 83.72% for headword types. The margin of “loss” is greater in the case of headwords (2.77%) than in the case of lemmas (1.36%) even though the token coverage is greater in the case of the former. Therefore, a decision as to which list is more appropriate for instruction or research is left to the approach and preferences of the teacher or researcher. The proposed lists (named ICE-CORE) are freely available for download from the following URL: <http://www.sequencepublishing.com/academic.html>

Conclusion

The objective of this investigation was to obtain word lists of the most frequent words shared across seven varieties of English. Thanks to the work of many individuals and the framework provided by the International Corpus of English, we have been able to analyze a corpus amounting to some 8.5 million tokens and representative of individual corpora of over 1 million tokens each.

The lexical material that makes up these corpora provides invaluable information regarding the vocabulary chosen and preferred by speakers of each variety. Let us not forget that the total lexicon of English well exceeds the few thousand most frequent words in the proposed lists. However, as has been known for about a century, most words in the lexicon are hardly ever used while relatively few words do most of the work (Ayres, 1915; Faucett and Maki, 1932; Faucett, Palmer, Thorndike,

and West, 1936; Horn, 1926; Lorge, and Thorndike, 1938; Palmer, 1931; Thorndike, 1921 and 1931; West, 1926–7).

Results have shown this characteristic of language use with pristine clarity. Only 100 words, together with their repetitions, account for over half of all the words used by speakers of any variety. The 5,000 most frequent words, together with their repetitions, account for an average 95% of all words used by speakers regardless of variety.

Moreover, our analyses have shown that the most frequent words are one of the many features shared by varieties of English. Two word lists has been extracted using two different criteria of lemmatization and we have carried out analyses that identify the intersection of the most frequent inflected and headword types in all varieties. The results show unequivocally that the proposed lists obtained account for a similar token coverage as variety specific ones. as detailed in Tables 3 and 6.

It is of utmost relevance to understand that the proposed lists compiled in this study and following the methodology described are preliminary. This is so for various reasons. First, it is dependent on the corpora available and, in particular, bound by the sample size. Over 1 million running words per corpus is more than sufficient for some analyses but, perhaps in this case, results should be taken with caution. Second, the priceless work carried out by the ICE team as well as institutions and individuals involved in the actual compilation of the corpora still leaves us with only seven varieties to study. Proper core lists that could serve all varieties would have a more solid basis if more varieties and larger corpora were available.

Last, our motivation behind the extraction of these word lists of the most frequent inflected and headword types shared across varieties of English is fundamentally pedagogical. We understand and want to serve those instructors and learners that have the need to teach and learn, respectively, English as an international language and without bias towards a particular variety. We consider this investigation a step in that direction. Seidlhofer (2003) explains that “the potential for pedagogy would...reside in knowing which features tend to be crucial for international intelligibility...” (p.18). Seidlhofer adds that Jenkins’ (2000) ability to identify a pedagogical core “after establishing which pronunciation features impeded mutual intelligibility... provided an empirical basis” (p. 15) for suggesting that pedagogy would serve students by focusing on items that appear to be essential for intelligible pronunciation and, consequently successful communication. The idea is that students need to be provided with basis that they can use to learn from, “fine-tuning subsequently (after leaving school) to any native or non-native varieties and registers that are relevant for their individual requirements” (Seidlhofer, 2003, p. 23). Our work parallels Jenkins (2000) on the lexical level; we have identified a common lexical core that facilitates communication as this core is shared by and of equally high frequency in several varieties of English.

“By better understanding how English is used at the present, we can more accurately predict how it will be used in the future, and given a sense of what our future needs will be, we can begin preparing to meet them” (Morrow, 1987, p. 61).

“EIL is practically non-existent in language teaching curricula and materials... [EIL] has not had any major impact on how the subject ‘English’ is actually conceptualized, linguistically described and pedagogically prescribed for learning” (Seidlhofer, 2003, p. 12).

References

- Ayres, L.P. (1915). *Measuring Scale for Ability in Spelling*. New York: Russell Sage Foundation.
- Bauer, L. and Nation I.S.P. (1993). Word families. *International Journal of Lexicography*, 6(4), 253-279.
- Cogo, A. and Dewey, M. (2006). Efficiency in ELF communication: From pragmatic motives to lexicogrammatical innovation. *The Nordic Journal of English Studies*. 5(2), 59–93.
- Dell, G.S. and Brown, P.M. (1991). Mechanisms for listener-adaptation in language production: Limiting the role of the “model of the listener”. In D. Napoli and J. Kegl (Eds.), *Bridges between Psychology and Linguistics* (pp. 105–129). New York: Academic Press.
- Faucett, L. and Maki, I. (1932). A Study of English-word Values Statistically Determined from the Latest *Extensive Word Counts*. Tokyo: Matsumura Sanshodo.
- Faucett, L., Palmer H., Thorndike, E.L. and West, M. (1936). *Interim Report on Vocabulary Selection*. London: P.S. King and Son, Ltd.
- Horn, E. (1926). *A Basic Writing Vocabulary, 10,000 Words Most Commonly Used in Writing*. Iowa: College of Education, University of Iowa.
- Jenkins, J. (2000). *The Phonology of English as an International Language: New Models, New Norms, New Goals*. Oxford: Oxford University Press.
- Lorge, I. and Thorndike, E.L. (1938). *A Semantic Count of English Words*. New York: Teachers College, Columbia University.
- Morrow, P.R. (1987). The users and uses of English in Japan. *World Englishes*, 6(1), 49-62.
- Nation, I.S.P. (2001). *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.
- Nation, I.S.P. (2006). Range32 (Version 1.32) [README.DOC]. Available at: <http://www.victoria.ac.nz/lals/staff/paul-nation/nation.aspx>
- Nelson, G. (2010). International Corpus of English. Available at: <http://ice-corpora.net/ice>
- Palmer, H. (1931). *Second Interim Report on Vocabulary Selection Submitted to the Eighth Annual Con-*

ference of English Teachers Under the Auspices of the Institute for Research in English Teaching. Tokyo: Institute for Research in English Teaching.

Seidlhofer, B. (2003). A concept of international English and related issues: From 'real English' to 'realistic English'? Strasbourg: Language Policy Division, DG-IV - Directorate of School, Out-of-School and Higher Education, Council of Europe.

Seidlhofer, B. (2004). Research perspectives on teaching English as a Lingua Franca. *Annual Review of Applied Linguistics*, 24(1), 209-239.

Schell, M. (2008). Colinguals among bilinguals. *World Englishes*, 27(1), 117-130.

Thorndike, E.L. (1921). *The Teacher's Word Book*. New York: Teachers College, Columbia University.

Thorndike, E.L. (1931). *A Teacher's Word Book of 20,000 Words*. New York: Teachers College, Columbia University.

West, M. (1926-7). *The New Method Readers (New Series)*. Bombay and Calcutta: Longmans, Green.

Yano, Y. (2001). World Englishes in 2000 and beyond. *World Englishes*, 20, 119-132.

Yamaguchi, C. (2002). Towards international English in EFL classrooms in Japan. *The Internet TESL Journal VIII*(1). Available at: <http://iteslj.org/Articles/Yamaguchi-Language.html>

(2011.9.22 受稿, 2011.11.10 受理)