An Investigation of the Vocabulary Used in Iranian High School English Textbooks: A Corpus Linguistic Approach

Maryam Sodagar (M.A. in TEFL - University of Tehran)
email: m.sodagar@gmail.com
High school English teacher - Urumieh: Educational District No.**2**

چکیده

این تحقیق به بررسی لغات کتب درسی زبان انگلیسی که در حال حاضر در سطوح دبیرستان و پیش‌دانشگاهی در آموزش‌وپرورش ایران تدریس می‌شود می‌پردازد تا مناسب بودن این کتب را (از لحاظ لغات آن‌ها) برای دانش‌آموزان این مقطع تحصیلی مشخص نماید. معیار این بررسی «فراوانی لغت» در نظر گرفته شده که خود یکی از دستاوردهای «زبان‌شناسی پیکره‌ای» است. این تحقیق بر پایه‌ی این منطق بنا نهاده شده است که پیوستار «فراوانی لغت» دارای رابطه‌ی عکس با پیوستار «مهارت زبانی» یاد گیرنده می‌باشد؛ بدین معنی که هرچه مهارت و دانش زبانی دانش‌آموزان بالاتر می‌رود باید به آن‌ها لغات با فراوانی پایین‌تری تدریس شود و (برعکس) چرا که اصولاً کثرت لغات با فراوانی بالا در یک متن باعث کم بودن بار معنایی آن متن و متعاقباً باعث آسان شدن آن متن برای یادگیری می‌شود (و برعکس). بنابراین «نمودارهای فراوانی لغت» متون «خواندن و درک مطلب» به‌صورت کلی (مربوط به هر یک از این چهار کتاب) و به‌صورت اختصاصی (مربوط به هر یک از دروس) توسط برنامه‌ی کامپیوتری خاصی استخراج شد تا کتاب‌ها و هم‌چنین دروس موجود در هر کتاب، از لحاظ نسبت کلمات دارای فراوانی بالا

و پایین، باهم مقایسه شوند. نتایج به‌دست آمده نشانگر مطابقت کلی این کتاب‌ها با اصولی است که مبحث «فراوانی لغت» برای تألیف و تدوین کتب (هم‌چون اصول مربوط به «انتخاب» و «ترتیب ارائه»ی مطالب آموزشی) پیشنهاد می‌کند. بنابراین می‌توان گفت کتب درسی دبیرستان (از لحاظ واژگان) مناسب دانش‌آموزان این دوره‌ی تحصیلی است، ضمن این‌که بهتر است نقطه‌ی ضعفی نیز که در این کتاب‌ها شناسایی شده است مورد اصلاح قرار گیرد.

کلیدواژه‌ها: بررسی کتاب‌های درسی ـ فراوانی لغت ـ زبان‌شناسی پیکره‌ای ـ تألیف و تدوین کتب

Abstract:

This study is a frequency-based lexical analysis of four English textbooks which are currently being taught at high school and pre-university levels in the public education system of Iran. The purpose of the study is determining whether the lexical content of these textbooks is appropriate for the students of these levels of study or not and the criterion of the analysis is *word frequency* information which is one of the new outcomes of corpus-based analyses of language. In this study, it has been assumed that a large number of *low frequency* words are indicative of lexically rich environments whereas that of high frequency words are representative of *lexically poor* environments. Also, the continuum of *word frequency* has been assumed to be of a reverse relationship with the continuum of *learner proficiency.* the Lexical Frequency Profiles (LFPs) of all the reading texts in the textbooks (as indicators of the proportion of high / low frequency in those texts) were obtained by means of a computer program called VocabProfile (VP), the procedure which enabled us to compare and contrast the lexis in those textbooks are, in general, compatible with word frequency information and what it suggests for pedagogy, though a weak point has been detected which should be examined more closely by the materials developers.

Key Words: lexical analysis - word frequency - corpus linguistics - Lexical Frequency Profiles (LFPs) - VocabProfile (VP) program - materials development

Introduction:

Corpus linguistics is normally conceived of as the study of linguistic phenomena through corpora (singular: corpus)which have been defined as "large principled collections of natural texts" stored on a computer in a machine-readable form (Reppen & Simpson, 2002, p. 93). Corpus-based analyses, from the perspective of *formal/functional* linguistics, are much better suited to functional analyses of language, that is, "analyses that are focused... on describing the use of language as a communicative tool" (Meyer, 2002, p. 5) because corpora *contextualize* the language under study. The impact of corpus linguistics studies on classroom language teaching practices has already taken shape: No longer are pedagogical decisions based on intuitions and/or sequences that have appeared in textbooks over the years but they are rather grounded on the actually-recurring patterns in a language.

One of the major strong points of corpus-based analyses of a language is the 'objectivity' of the linguistic analyses that it yields. The upsurge of interest in applying empirical data as such in language pedagogy started in the early 1990s (Xiao & McEnery, 2005, section 1, 2). Among the scholars who believe in the incorporation of corpora-derived information in language pedagogy, one may refer to Widdowson (2000a) who has reiterated that this branch of linguistics (corpus linguistics) offers invaluable information regarding one of the features of language called *attestedness* according to Hyme's Scheme. Hyme (1972, as cited in Widdowson, 2000a, p. 22) had categorized the componets of communicative competence (as the reality of language) into four types of knowledge: possibility (conformity to grammatical rules), appropriacy (conformity to social conventions), feasibility (uttering what is easily processed and readily understood by the other interlocutor), and finally, attestedness (uttering what occurs in language frequently). It is based on this scheme that Widdowson (2000a) argues that corpus-based data, (not directly, of course) should inform pedagogic techniques.

There are many levels of information that can be gathered from analyses of corpora and one of the major ones (related to the present study) is the information regarding 'frequency of occurrence' of words in English. 'Word frequency' simply means "how often a given word occurs in normal use of the language" (Nation & Waring, n.d., fourth section 2006, Based on this information, several 'word frequency lists' have been developed till now which include: **a**) the list of the most common words in General English (GE) settings developed by West (1953). It has been called General Service List (GSL) and consists of the list of the first 1,000 plus the second 1,000 most frequent words in GE settings, and also **b**) the list of the most important words in academic settings which is called the Academic Word List (AWL). The AWL consists of the words of high frequency in academic settings which are, logically, of low frequency in GE settings.

A specific research tool used in this study is a computer program called *VocabProfile (VP)* which in its latest version is also known as *Range* program. This program (available on: www.vuw.ac.nz/lals/staff/Paul Nation) has been introduced and validated in a study made by Laufer and Nation (1995) and has been widely used in the domain of vocabulary studies. Since VP program is accompanied by special frequency-based word lists, it "deconstructs any text or corpus into its lexical components" by their frequency zones (Cobb, 2003) through the following procedures: It takes a given text as the raw input (the text may be typed, pasted or scanned into the program); checks the lexis of that text against its accompanying frequency-based word lists; and finally, as output, generates a lexical frequency profile (LFP) of that text in just a few seconds. The LFP generated as such describes the lexical content of a text in terms of four frequency zones which are actually representative of the four word lists ordinarily available in the program:

- The first 1,000 most frequent words in General English (GE),
- The second 1,000 most frequent words in GE (i.e. from 1,001 to 2,000),
- The Academic Word List (AWL),
- The words not included in any of the above lists (NIL or 'not in the Lists') so they are normally addressed as the 'difficult' words.

Word frequency information can provide pedagogical suggestions for the process of 'selection' and 'gradation' of teaching materials. According to Meara & Nation (2002, p. 39), "high frequency words need to be the first and main vocabulary goal of learners" simply because the most frequent words in English language are mostly function words which are empty of lexical content and at the same time crucial for grasping the idea of a text; therefore, preliminary knowledge of them facilitates consolidation of a basic GE knowledge among non-native students. On the other hand, it has been assumed that a large number of 'low frequency' words would mirror 'rich' lexical environments while a large number of 'high frequency, words would reflect 'poor' lexical environments; therefore, language textbooks are expected to contain a logical proportion of both high frequency words (conceived of as the first 2,000 most frequent words in English) and low frequency words (conceived of as the AWL and the NIL) in a way that the principle of 'systematic presentation' of materials to learners is catered for.

Considering the preceding studies about the importance of frequency information in pedagogy, this study can be conceived of as a 'lexical text analysis' within a 'whilst-use' materials evaluation. On the significance of materials evaluation there is no doubt among materials developers and textbook writers because of the enlightening role it has in the process of revision and improvement of teaching materials. Needless to say, each of the three types of evaluation - "pre-use", post-use" and "whilst-use" materials evaluation (Tomlinson, 1998, p. xi) - is of its own particular advantages and contributes ultimately to this process.

The present study, then, is aimed at answering the following research questions:

1. Is there any significant difference between *the English high school textbooks* in terms of the extent to which they have made use of *the first 1,000* most frequent words of English?

2. Is there any significant difference between *the English high school textbooks* in terms of the extent to which they have made use of *the second 1,000* most frequent words of English?

3. Is there any significant difference between *the English high school textbooks* in terms of the extent to which they have made use of *the academic vocabulary (AWL)?*

4. Is there any significant difference between *the English high school textbooks* in terms of the extent to which they have made use of the words not included in the three previous lists (NIL)?

5. Is there any significant difference between *the lessons in each of the English high school textbooks* in term of the extent to which they have made use of the words of *high frequency* (considered as the first 2,000 most frequent words in GE) and those of *low frequency* (considered as *beyond* those 2,000 words which are all the words included in AWL+NIL)?

Method:

Data collection:

At first, all the 'reading passages' together with the 'new words' sections in the textbooks were scanned into the VP computer program lesson by lesson using a scanner device. Then, some modifications were made on the scanned texts, for example, all "proper nouns" and "numbers" found in the scanned texts were omitted because their inclusion in the analysis would result in a misleading increase in the number of the words which belong to the first and the fourth word lists. Specifically speaking, "proper nouns" do not belong to the lexis of any given language and, accordingly, they are not included in any of the first three word lists; therefore, they inevitably fall into the fourth category (NIL) which results in a misleading increase in the percentages of the words which belong to that category. On the other hand, "numbers" normally belong to the first word list; therefore, for the purpose of avoiding the illusion that the textbooks have made more use of the first word list, their omission from the texts was necessary, too.

After collecting the data related to each lesson (which is required for an intra-textbook analysis in relation to the fifth reseach question), the scanned lessons related to each textbook (nine lessons in textbook 1, seven lessons in textbook 2, six lessons in textbook 3 and eight lessons in textbook 4) were put together and saved as separate files so that the four textbooks themselves could be compared and contrasted against each other (in order to do an inter-textbook anaylysis and to find answers to the first four research questions). Then, the lexical frequency profile (LFP) of each file was obtained using VocabProfile (VP) computer program (See 'introduction section' for the procedures of producing an LFP via VP). In this way, 34 LFPs (lexical frequency profiles) constituted our collected data needed for the analysis (30 LFPs for the lessons and 4 LFPs for the textbooks).

Results:

The data were, then, submitted to statistical analysis using chi-square.

The results related to the first four research questions which all deal with the existence of any significant difference between *textbooks* in terms of any of those four word frequency lists (i.e. the first 1,000, the second 1,000, the AWL, and the NIL) are presented in table 1 in two sections: The first section of the table reports the results of Vocabulary Profile (VP) analysis which yields the LFPs of each of those four textbooks under study and the second section of the table demonstrates the results of chi-square tests for comparison of (the LFPs of) those four textbooks in terms of any of the word lists in the study.

According to this table which, in fact, reports all the results needed for an inter-textbook analysis, there is no significant difference between these four textbooks in terms of the first 1000 most frequent words, the second 1,000 most frequent words and the NIL word lists; whereas, there is a significant difference between them in terms of AWL (P=.02); therefore, the null hypotheses formulated for question number 1, 2, and 4 were confirmed whereas the null hypothesis for question number 3 was rejected.

Regarding the fifth research question dealing with the existence of any significant difference between the lessons included in each textbook (in an intra-textbook analysis), the results of chi-square tests

Table 1: The analise of Word Frequency Profiles of English Textbook

VP analysis					Chi-square test	
Word list	Textbook 1	Textbook 2	Textbook 3	Textbook 4	X2	P
1st 1,000	74.5	70.5	68.2	62.4	1.30	.72
2nd 1,000	16.3	18.2	15.4	13.7	.55	.90
AWL	0.8	1.2	6.4	8.2	9.50	.02
NIL	8.4	10.1	10.0	15.7	3.27	.35

Note. The values in 'VP analysis' section represent percentages (rather than absolute values). AWL = Academic Word List; NIL = Not In the Lists. 'Chi-square test' estimated at p < .05 with df = 3.

applied to their LFPs revealed that there is no significant difference between them neither in terms of high frequency words (the first 2,000 most frequent words) nor in terms of low frequency ones (beyond those 2,000); therefore, the fifth null hypothesis was confirmed, too. *[Note: the LFPs of the lessons as well as X2 and P values found for their comparison have not been demonstrated in this brief paper. Interested readers may contact the researcher for a full list of results]*.

Discussion:

In this study, the findings on the first three hypotheses (confirmation of the first two null hypotheses as well as rejection of the third null hypothesis) are all desirable from the perspective of frequency-based pedagogical considerations, revealing that high school and pre-university English textbooks have already catered for what 'word frequency' information suggests for language pedagogy regarding 'selection' and 'gradation' of teaching materials in the whole process of syllabus design and materials development. The more specific reasons behind this justification are as follows:

The lack of significant difference between the textbooks in terms of the first 1,000 most frequent words and the fact that the words of this word list constitute the larger part of all these textbooks (74.5% of textbook 1, 70.5% of textbook 2, 68.2% of textbook 3, and 62.4% of textbook 4) is not an unexpected phenomenon because this word list normally comprises 'function' words in English and, needless to say, 'function words' are abundantly used in almost any text and, at the same time, they are crucial for grasping the content of a text; therefore, the first finding is quite justified in being compatible with our expectations from the perspective of 'frequency' information.

Also, the lack of a significant difference between the textbooks in terms of the second word list (which consists of the most common lexis used in GE settings) can be interpreted as another sign of suitability of these textbooks for the students of these levels of study. That is because it indicates that these textbooks, regardless of students' proficiency levels, expose students to somehow an equal number of the most common lexical word of English in GE settings throughout their four years of (partial) studying English at high school and pre-university levels; in other words, this finding provides evidence supporting the idea that all these four textbooks are rich in terms of GE vocabulary which is the mostly-needed vocabulary for students at these levels.

Moreover, the existence of a significat difference between the textbooks in terms of AWL can also be interpreted as a strong point in the development of these textbooks because, on the one hand, we had aready assumed that ‘low frequency’ words are representative of ‘lexically rich’ environments (and vice versa) and, on the other hand, there *must* always exist a basis for division and difference among textbooks of various grades of study so that they can best represent the textbooks assigned to be taught in various proficiency levels (grades of study). Accordingly, the existence of a significant difference among the textbooks in terms of the most frequently-used vocabulary in academic settings (AWL) can be considered as a good sign of variation among these textbooks, the point which makes them lexically appropriate.

With regard to the confirmation of the fourth null hypothesis, however, a weak point was recognized in these textbooks - It was, surprisingly enough, revealed that all these textbooks contain almost an equal number of ‘difficult words’(included in NIL word list). Evidently, this finding is in clear contrast with one of the important tenets of materials development, that is, learners should start with easy materials and end up with difficult ones. In this way, it can be argued that the principle of ‘systematic presentation’ of materials has been violated in the development of these textbooks because of incorporation of materials of the same difficulty at all levels.

Finally, the confirmation of the fifth null hypotheses which deals with intra-textbook analysis can be considered as another indicator of the appropriate organization of the lexical content of each of these textbooks. In other words, the lack of significant difference among the lessons of each textbook whether in terms of ‘high frequency’ or ‘low frequency’ words is another desirable finding in line with pedagogic considerations because it is always recommended in syllabus design that the lessons of any given textbook should be congruent with one an other in order to best represent the lessons of *one* given textbook assigned to be taught to students of *one* specific (not various) proficiency level.

Conclusion and Pedagogical Implications:

Generally speaking, most of the results of this inter-textbook and intra-textbook lexical analysis indicated that the lexical content of these four textbooks is compatible with what ‘word frequency’ information implies for language pedagogy. Based on the obtained results, these four textbooks are considered as suitable and lexically appropriate textbooks for students at this level of study, though, a weak point was also identified (i.e. the lack of difference across textbooks, in terms of incorporation of ‘difficult words’) which is hoped to be improved by materials developers.

The present study may be of an enlightening role for those who are involved, in some way or another, in TEFL (Teaching English as a Foreign Language) as it focused on the crucial role that 'frequency information' (in particular) and corpora-derived information (in general) can play in the process of syllabus design. Materials developers, for example, may be encouraged to consult frequency-based word lists and also the outcomes of corpus-based analyses of language (information on collocation and phraseology, for instance) in developing new textbooks. At the same time, reading this study may encourage language teachers to take more advantage of the insightful dictionaries written on the basis of 'frequency of occurrence 'during teaching 'vocabulary' or 'syntactic patterns' to learners if they are interested to expose them to the vocabulary or syntactic patterns of various frequencies in accordance with their proficiency levels and / or their special needs.

References:

Academic Word List (AWL), Available online, http://www.vuw.ac.nz/lals/divl/awl/

Cobb, T. (2003). Analyzing Late Interlanguage with Learner Corpora: Quebec

replications of three European studies. *Canadian Modern Language Review, 59(3),* 393-423. Retrieved May 14, 2005, from

http://www.er.uqam.ca/nobel/r21270/cv/LC3.html

General Service List (GSL), Available online, http://jbauman.com/gsl.html

Laufer, B. & Nation, P. (1995). Vocabulary size and se: Lexical richness in L2 Written Production, *Applied Linguistics*, 16, 307-322.

Meara, P. & Nation, P. (2002). Vocabulary. In N. Schmitt (Ed.), *An Introduction to Applied Linguistics* (pp. 35-54). New York: Oxford University Press.

Meyer, C. F. (2002). *English Corpus Linguistics an Introduction.* Cambridge: Cambridge University Press.

Nation, P & Waring, R. (n.d)Vocabulary Size, Text Coverage and Word Lists.

Retrieved May 13, 2006, from http://www.1harenet.ne.jp/~waring/papers/cup.html

Reppen, R. & Simpson, R. (2002). Corpus Linguistics. In N. Schmitt (Ed.) *An Introduction to Applied Linguistics* (pp.92-111). New York: Oxford University Press.

Tomlinson, B. (1998). Glossary of Basic Terms for Materials Development in Language Teaching. In B. Tomlinson (Ed.), *Materials Development in Language Teaching* (pp.viii-xiv). Cmbridge: Cambridge University Press.

VocabProfile of Range program available from Paul Nation's Website http://www.vuw.ac.nz/lals/staff/Paul_Nation

West, M. (1953). *A general Service List of English Words.* London: Longman.

Widdowson, H. G. (2000 a). Object Language and the Language Subject: On the Mediating Role of Applied Linguistics. Annual Review of *Applied Linguistics, 20, 21-33.*

Xiao, R. & McEnery, T. (2005). Corpora and Language Education. Retrieved March 25, 2007, from: http://forum.corpus4u.org/showthread.php?t=75