

Farhad Maszlum Zavaragh, English teacher, Maraghe
& PhD Student of TEFL in Teacher Training University, Tehran,
email: mazlumzf@yahoo.com

A Corpus-based Analysis of Iranian High School English Textbooks

چکیده

مطالعه‌ی حاضر سه هدف عمده داشته است: الف) بررسی متون خواندن کتاب‌های زبان انگلیسی دوره‌ی دبیرستان برای تعیین میزان استفاده‌ی آن‌ها از اولین (K1) و دومین (K2) ۱۰۰۰ کلمه‌ی رایج در زبان انگلیسی؛ ب) مقایسه هم‌نشینی لغات جدید با نتایج یا برودادهای برنامه‌های تطبیقی؛ ج) بررسی میزان استفاده‌ی مؤلفین از رایج‌ترین ۱۲ فعل مکالمه‌ای در بخش‌های گفت‌وگو و کاربردهای زبانی (براساس یافته‌های بایبر و کنراد این ۱۲ فعل ۴۵ درصد کل افعال مکالمات را تشکیل می‌دهند). بررسی کلمه محور برودادها نشان داد که متون خواندن هماهنگ با درصدهای نمونه و معیار تدوین شده‌اند، ولی بررسی خانواده محور همین متون مشخص کرد که کتب سه سال دوره‌ی دبیرستان به ترتیب ۱۵/۷۶، ۱۶/۷۲ و ۱۸/۲ درصد مجموع K1 و K2 را پوشش داده‌اند. مقایسه‌ی هم‌نشینی لغات جدید توسط مؤلفین با نتایج یا برودادهای برنامه‌های تطبیقی مشخص کرد که هماهنگی حداقل ۳۶/۹۲ و حداکثر ۵۸/۹۹ درصد وجود دارد. بالاخره این که رایج‌ترین ۱۲ فعل مکالمه‌ای فقط ۹/۰۵ درصد پوشش داده شده‌اند. بر این اساس در مقاله حاضر یافته‌های پیکره محور به عنوان منبعی مطمئن‌تر و موجه‌تر در تهیه و تدوین کتب درسی زبان انگلیسی مورد بحث و بررسی قرار می‌گیرند. یافته‌های این مطالعه می‌تواند علاوه بر مؤلفین کتب درسی، مورد استفاده‌ی معلمان زبان نیز قرار گیرد. در هنگام تدریس لغات جدید، با معرفی خانواده‌ی کلمات دانش‌آموزان و در نتیجه توانایی خواندن آن‌ها افزایش خواهد یافت. برای مثال، هنگام تدریس «understand»، دبیران می‌توانند با معرفی خانواده‌ی آن یعنی (understanding, understood, understandable, misunderstand, misunderstandings, misunderstood) به‌طور قابل توجهی به دانش‌آموزان و توانایی خواندن آن‌ها کمک کنند. دبیران همچنین می‌توانند با مراجعه به نرم‌افزارهای رایگان موجود در اینترنت، لغات و کلمات جدید درس‌ها را پردازش و هم‌نشینی آن‌ها را مطالعه کنند و مثال‌های مناسبی را در کلاس ارائه دهند. مزیت این مثال‌ها آن است که براساس پیکره‌های بسیار بزرگی ارائه می‌شوند و الگوهای واقعی و کاربردی کلمات را معرفی می‌کنند. همچنین، دبیران زبان می‌توانند پس از آشناکردن مختصر دانش‌آموزان با این برنامه‌ها و به عنوان کار در خانه، از آن‌ها بخواهند با استفاده از برودادهای این برنامه‌ها و با توجه به هم‌نشینی کلمات جمله‌سازی کنند.

کلیدواژه‌ها: مطالعه‌ی پیکره محور، مطالعات فراوانی، برنامه‌های تطبیقی، متون خواندن، کتب درسی ایران.

Abstract

This study aimed at investigating reading texts of Iranian high school English books in terms of their use of the first (K1) and the second (K2) most common 1000 words, comparing collocations of the vocabulary items with those of concordancing outputs, and researching conversational English with regard to coverage they give to Biber and Conrad's 12 most common verbs of English conversations. Token-based calculation of VocabProfile outputs of the reading texts suggest that the passages go with the typical profile. Calculation on a word family basis, however, demonstrates that the reading texts cover 15.76%, 16.72% and 18.2% of K1 + K2. Comparing text developers' collocations with those of concordancing outputs revealed a 36.92% to 58.99% compatibility. Lastly, only 9.05% coverage has been given to the 12 most common conversational verbs-verbs that account for more than 45% of conversational English. Corpus-based findings are advocated as a more reliable source of material development than intuition.

Key Words: corpus-based analysis, frequency studies, concordancing programs, reading texts, Iranian textbooks.

Introduction

Certainly one of the most important elements of language classes is textbooks. Few teachers enter class without a textbook- often a required one- that provides content and teaching/learning activities that shape much of what happens in that classroom. 'Researching' textbooks is a multi-dimensional process involving analyzing subject matter, vocabulary and structure, exercises, illustrations, and even physical make-up (Byrd, 2000).

Traditionally, analysis of vocabulary section consisted of answering some items in a checklist. In one such checklist, Donald and Celce-Murcia (1979) ask the respondents to answer questions like 'Does the vocabulary load (the number of new words introduced in every lesson) seem to be reasonable for the students of that level? (emphasis added)' or 'Are the new vocabulary items controlled to ensure systematic gradation from simple

to complex items?'. The answers to such questions, e.g. excellent, good, adequate, weak, and totally lacking, were coded numerically and the analysis followed.

With the advent of sophisticated computer programs, invaluable rich databases of naturally-occurring language use and clever concordancers, researchers and practitioners have shifted markedly in approach and practice. The electronically stored collections of written and spoken naturally occurring language, i.e. corpora, have provided researchers, teachers, material developers, and language learners with useful information on word frequency lists and collocations. The analysis of language teaching materials has also been enlightened by such findings. McKay (2006) believes that corpus-based research is one of the best ways to research the language of textbooks. For her, the language of the textbooks means the grammatical structures lexicon and their use in dialogues and reading texts.



Two areas of corpus-based research are especially valuable for developing and assessing L2 textbooks. The first is the calculation of *word frequency lists*. A frequency list is simply a list of all the types of words that appear in a corpus, along with the number of occurrences of each word. A second type of analysis that can be done with a corpus is *concordancing*. Concordancing programs allow the user to bring together all the instances of a particular word along with the words that surround it. The selected word is referred to as the node word/phrase. The best way to read concordance lines is to skim them initially from top to bottom, looking for central patterns (McKay, 2006).

□ Lexical Text Analysis by Computer Programs

Lexical text analysis is currently conducted by the easy-to-use computer programs. One such program is called Vocab Profile (VP). VP takes any text and divides its words into four categories by frequency: (1) the most frequent 1000 words of English-known as K1, (2) the second most frequent 1000 words of English, i.e. 1001 to 2000 – known as K2, (3) the academic words of English Coxhead's (2000) Academic Word List, 550 words that are frequent in academic texts across subjects), and (4) the NIL (Not In the List) or off-list words which are not found on the other lists. They may include proper nouns, unusual words, specialist vocabulary, acronyms, abbreviations, and misspellings. The important point is that if

someone knows K1, he would know 72% of the running words in a text. Knowing about 2000 word families, i.e. K1+K2, gives near to 80% coverage of the written text. Research by Liu Na and Nation (1985) has shown that this is not sufficient to allow reasonably successful guessing of the meaning of the unknown words. At least 95% coverage is needed for that. Research by Laufer (1989) suggests that 95% coverage is sufficient to allow reasonable comprehension of a text. Nation and Waring (1997) believe that the second language learner needs to know the 3000 or so high frequency words of the language.

The Present Study

□ Purposes

This study is an attempt to investigate reading texts of Iranian high school English books in terms of their use of the first (K1) and the second (K2) most common 1000 words, to compare collocations of the vocabulary items with those of concordancing outputs, and to study conversational English of mainstream textbooks in terms of the coverage they give to Biber and Conrad's 12 most common verbs of English conversations.

□ Procedures

To find the answer for the first question of the study, all the reading passages in high school English textbooks were fed into the VP program. They were processed first grade by grade and then all passages of the three textbooks were combined as a single

file and analysed as such. All proper nouns were excluded when processing the files. As for the second question, the words in New Words section of the three textbooks were entered into the concordancing program. According to Waring (personal communication), the most pedagogically-desirable strategy is to seek and take into account both 'left' and 'right' collocations of a given word. As a result, after feeding the new words of the textbooks into the concordancing program, the 'left' and 'right' collocations of the highest frequency in the corpus were sought and compared with the collocations in the textbooks. As far as the third purpose is concerned, the whole analysis procedure was undertaken with 6 textbooks. Since Iranian EFL learners begin practicing/learning English conversations in junior high schools, their three textbooks at this level were also considered in the study. The 'dialogue' sections of these books and the 'conversation in context' and 'language function' sections of the three high school textbooks were studied. Biber and Conrad's (2001) 12 key verbs make up the criterion.

Results and Discussion

□ K1 and K2 word families

Table 1 shows the VP analyses of the reading passages in Iranian high school English textbooks. Eyeballing the VP outputs for each textbook demonstrates the following points about each frequency zone.

Table 1: VP outputs for the reading passages of high school textbooks

	Textbook1	Textbook2	Textbook3	TPTAL
K1 words	89.86%	89.09%	87.9%	88.98%
K2 words	7.12%	7.34%	6.00%	6.85%
AWL words	0.54%	0.39%	2.52%	1.10%
Off-list words	2.48%	3.19%	3.53%	3.07%

Firstly, the high percentages of K1/s usually occur with texts that are either conversational English or simplified English (Cobb, personal communication). So high K1 percentages are reasonably justified since Iranian high school students are exposed to what is called 'simplified' English. Cobb (personal communication) believes that such high K1 percentages are good provided that the learners are beginners.

Secondly, and as far as K2 word-base is concerned, the reading passages go with the typical profile—which is K1=70, K2=10, AWL=10, and Off-list 10. According to Cobb, about 7%-10% is a normal K2 output. Yet since the texts contain simplified English for learning/teaching purposes, AWL and Off-list words are lower compared with the typical of 10 each. The very low index of AWL output is also expected since the texts have been developed for beginning EFL learners.

Thirdly, and most importantly, is the question of 'coverage' of K1 and K2 word bases in these texts. The answer to this question pertains to the first research question. What do 89.86%, 89.09%, and 87.94% mean? Does the total of 88.98% mean that 88.98% of the K1 list has been

used and covered? The following VP output for the following sentence is quite revealing and useful in answering this tricky question.

The sentence: I go to school by bus every day but my friend rides on his bike.

VP Output: K1: 86.67% K2: 6.67%
AWL: 0.00% Off-list: 6.67%

Does the K1 of 86% mean that 86% of the words in this zone (the first 1-1000 most frequent words) have been used or covered? The same questions can be raised for the subsequent zones as well. Clearly this is not the case.

When it is argued that 2000 words (K1+K2) provide 80% coverage, it is 2000 word *families* that does so. For example the list of *understand, understanding, understood, understandable, misunderstand, misunderstanding, misunderstandings, misunderstood* makes up a single unit or a family. Therefore, the first research question cannot be answered on the basis of K1 and K2 percentages given above. The 89% of K1 for textbook one means that 89% of the reading words are from the first 1000 words and 7% of K2 means 7% are from the second 1000 words... The positive point about such high percentages-as went above- is the fact that the output text would be 'easy' or what Cobb calls 'simplified' English. Since these texts aim beginning EFL learners it might be safely argued that text developers have been legitimately bound to K1 and K2 zones. Coverage in this

sense, i.e. on a token basis, is excellent.

Table 2 demonstrates the number of word families used in each textbook in three zones of K1, K2 and AWL. Off-list word families are not given since they are not available in VP output and have to be pre-calculated. Word families are needed to investigate what percentage of K1, K2 and AWL lists have been used or covered in each text (Waring, personal correspondence). Cobb notes that coverage is usually calculated in tokens—all the running words. Yet he goes on and indicates that 'families against all tokens' is the best way to decide on coverage. Nation and Laufer (2000) note 2000 families give 80% (token) coverage in average texts. If this is the case, putting K1 and K2 word families together would give the total number of word families fitting in K1+L2 domain. This, in turn, would give an indication of how much these reading texts prepare EFL learners with tackling average English texts.

Table 2. word families used in reading texts of each textbook

	Grade 1			Grade 2			Grade 3			Total		
	K1	K2	AWL	K2	K2	AWL	K1	K2	AWL	K1	K2	AWL
word families	310	66	4	319	71	6	332	61	25	523	171	35

Of course, not too much is expected in this regard since Iranian high school students are beginning students after all. If Nation and Waring's (1997) 80% coverage is set as a criterion. Iranian reading passages have

helped them with 15.76%, 16.72% and 18.2% (moving from grade 1 to 3) coverage of tokens in *average texts*. Putting three books together, the reading passages would help them with 27.76% of tokens in average texts. Whether this amount of preparation is adequate and satisfactory cannot be determined accurately. The percentages, however, might shed some light on the gap between the same learners' General English knowledge acquired during high school and the knowledge of General English expected in higher education programs. A good part of the learners' problems in English reading classes might be attributed to the fact that the learners' problems in English reading classes might be attributed to the fact that the learners' have not been exposed to the first and second 1000 word lists adequately, meaningfully and comprehensively. Of course, the same students study one more text book at pre-university level during which they naturally enrich their vocabulary repertoire to some further extent.

□ Collocations of the New Words and Concordancing Programs

The second purpose of the study was to investigate the way New Words of the lessons have been collocated. In other

words, it was intended to see how much the collocations of the words in New Words section go with the concordancing outputs. The New Words section was selected since the text writers have tried to contextualize the new items through using them in one, two or three sentences along with pictures if appropriate. In so doing, they have naturally got involved in combining the new items with some other words; i.e. collocates.

To study the collocations, it is more advisable and meaningful to check both right and left collocates (Waring, personal correspondence). This, in turn, lessened the number of words to be checked out since not all new words had been collocated on both sides. One hundred twenty new words lent themselves for such analysis (grade 1: 65, grade 2: 36, grade 3: 19). The corpus was again Brown's. The concordancing program allows the user to bring together all the instances of a particular word along with the words that surround it. Also the program gives a table in which the number of instances tells which surrounding word appears how many times. For instance the concordancing output for 'earn' is given below- a node word with only 16 cases is given as example just to save space. Table 3



demonstrates the number of instances of left and right collocates for 'earn' respectively.

Table 3: left and right collocates for 'earn' and the number of instances

Left collocates for 'earn'		Right collocates for 'earn'	
to	7	a	3
may	1	an	2
might	1	exemption	1
must	1	her	1
should	1	in	1
don't	1	or	1
they	1	our	1
countries	1	ten	1
who	1	the	1
would	1	their	1

Table 3 tends to suggest that the most frequent and at the same time naturally - occurring use of 'earn' is 'TO EARN A' in Brown's corpus. This pattern, then, should be prioritized when developing textbooks simply because it is the most frequent pattern in natural language use and is more reliable than intuition (Biber and Conrad, 2001). McKay (2006) believes that concordancing outputs can also help with looking for the common syntactic pattern, such as 'earn' that tends to be used in infinitive form. Of course, it is suggested to check a node word in more than one corpus (McKay, 2006).

All 120 words that appear in New Words sections were processed using the concordancing program. Since there were cases in which the number of instances (given in collocate tables) were the same or close, e.g. a = 3 and an = 2 above, it was decided to set the first two instances as the

norm patterns. The percentages in table 4 show how much the left and right collocates used by text – developers match with the corpus – based concordancing outputs.

Table 4: percentages of compatibility of left and right collocations in New Words with concordancing outputs

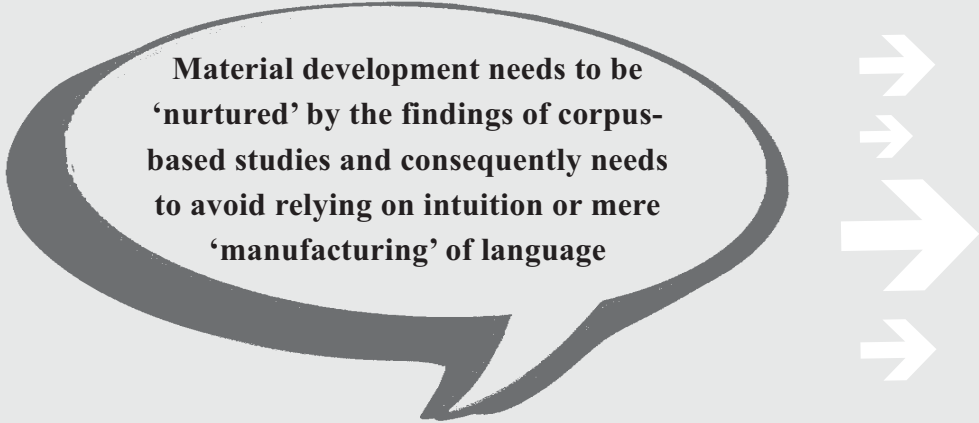
TEXTBOOKS	Collocates	
	Left	Right
Grade 1	36.92%	36.92%
Grade 2	58.99%	41.66%
Grade 3	42.10%	52.69%

As an example, about 59% of the left collocates in the second textbook (and in New Words section) go with the first or second pattern outputs while it is about 37% and 42% for textbooks 1 and 3 respectively. When it goes to the right collocates, textbook 3 seems to enjoy a more naturally – collocated language compared with textbooks 2 and There are just two above-fifty percentages. The low percentages might be attributed to the old practice of materials development; intuition.

The following examples and comparisons might prove helpful.

Example 1 (1st grade): The **distance** from our house to my school is one kilometer.

While the node word 'distance' has been collocated by 'the' on the left and 'from' on the right, the concordancing output suggests THE (34) **DISTANCE OF** (22) as the most frequent pattern in natural use



Material development needs to be ‘nurtured’ by the findings of corpus-based studies and consequently needs to avoid relying on intuition or mere ‘manufacturing’ of language

(numbers within parentheses show the number of instances in the corpus). It might be argued, then, that text developers have done a good job on the left side but not on the right side. Of as the right collocates is more frequent than FROM that happens to be 10.

Example 2 (2nd grade): What do you **require**? He **requires** peace and quiet. I **require** help.

The same node word collocates with WOULD (13 times) and WILL (11 times) MORE (8 times) on the right. These right and left collocations in the corpus tend to suggest the patterns like WOULD **REQUIRE** THE or WILL **REQUIRE** MORE are better candidates when teaching and contextualizing ‘require’.

Example 3 (3rd grade): The moon **orbits** round the Earth and the Earth **orbits** round the Sun.

The output for ‘orbit’ suggests that as a verb, a pattern such as PUT INTO **ORBIT** is more natural and frequent. When it is a noun, it is usually followed by [s], e.g. Earth’s orbit. These corpus-based findings could have been usefully applied in collocating new vocabulary items. For instance, the students could have been

exposed to sentences like *The new rocket was put into orbit* or *The Earth’s orbit takes 24 hours*.

□ Dialogues and Conversations in Textbooks and Biber and Conrad’s 12 Key Words

After a corpus - based study of English conversations, Biber and Conrad (2001) pointed out that ‘only 63 verbs occur more than 500 times and only 12 verbs occur more than 1,000 times per million words. These 12 verbs are: *say, get, go, know, think, see, make, come, take, want, give, and mean*. With this finding, they believe that text writers would clearly want to include the 12 most common verbs in beginning level materials. Put it differently, these key verbs account for over 45% of all verbs in conversational English.

Therefore, the third research purpose is restricted in its focus since conversation and dialogues of 6 textbooks are analysed with respect the above 12 *verbs*. Table 5 illustrates how much coverage has been given to these 12 most common verbs in ‘dialogue’ and ‘language function’ sections of the 6 text books. Totally, there are about 562 different verbs in dialogue

and language function sections of the textbooks. And the 12 most common verbs have appeared only 51 times.

Table 5: The coverage of the 12 most common verbs in conversation sections

Number of verbs	12 key verbs	Percentage
562	51	9.05%

Biber and Conrad's (2001) findings confirm the fact that the 12 key verbs account for more than 45% of all verbs in conversation. That is why these verbs are argued to be of priority and significance in developing beginning level materials. A comparison between 45% and 9.05% demonstrates how big the gap is between the use of these 12 verbs in natural conversations and the developed ones. Among the 12 key verbs, the mostly used verb was 'go' with 16 number of occurrences and the verbs 'say', 'make', and 'mean' have not been used at all! The wide gap affirms again the fact that material development should be based on natural use rather than on intuition per se.

Conclusion and Implications

The findings of this study might be helpful both to material developers and teachers. It was argued that the gap between Iranian EFL learners' knowledge acquired during high school programs and the knowledge required in universities and colleges might be

partially attributable to the low coverage given to K1 and K2 in mainstream English textbooks. Also it was maintained that naturally occurring language can be easily retrieved from concordancing programs and taken into consideration when collocating and contextualizing new words. Furthermore, material developers may need to reconsider the conversational English of Iranian textbooks regarding what frequency-based studies suggest. In short, material development needs to be 'nurtured' by the findings of corpus-based studies and consequently needs to avoid relying on intuition or mere 'manufacturing' of language.

English teachers can also benefit from the findings of this study. Firstly, a pedagogically important issue is the practice of introducing a new word's *family* when teaching vocabulary items. As went earlier, this would significantly help with the enrichment of the learners' vocabulary repertoire and the improvement of their reading comprehension. For instance it is a good idea to teach "able, ability, unable" or "usually, usual, unusual" as families. Secondly, teachers can easily benefit from free concordancing programs. Such programs, (retrievable at, www.collinswordbanks.co.uk and www.edict.com.hk/concordance.) would provide them with useful information particularly of word collocations. The most frequent patterns, then, can be introduced to the students. Teachers may require their

**A
good part of the learners’
problems in English reading classes
might be attributed to the fact that the
learners’ problems in English reading classes might
be attributed to the fact that the learners’ have
not been exposed to the first and second 1000
word lists adequately, meaningfully and
comprehensively**

students to do a similar job, i.e. process the new words in concordancing programs and find the most frequent patterns, and later make sentences according to their findings. Teaching/learning new words through collocations is also strongly supported by those researchers who believe that a good part of language learning is learning different ‘chunks’ of language (see Schmitt, 2002).

Additionally, teachers might consider as necessary to assign some supplementary reading materials for their students in order to help minimize the gap referred to above. Fortunately, the list of K1 and K2 (and Academic Words) is easily available and free. Teachers can study the lists and think of introducing those highly frequent items that have been totally ignored by textbook developers. Although vocabulary learning/teaching needs to be done through meaning-focused input (listening and reading) and meaning-focused output (speaking and writing), deliberate vocabulary learning is also effective (Nation and Meara, 2002).

References

- Biber, D. & Conrad, S. (2001) Quantitative Corpus-based Research: Much More Than Bean Counting. *TESOL Quarterly*, 35(2), 331-336.
- Byrd, P. (2000) Textbooks: Evaluation for Selection and Analysis for Implementation. In M. Celce-Murcia (Ed.), *Teaching English as a Second or Foreign Language* (pp. 415-427). Heinle & Heinle.
- Coxhead, A. (2000). A New Academic Word List. *TESOL Quarterly*, 34, 213-238
- Donald, A. M., and Celce-Murcia, M. (1979) Selecting and Evaluating a Textbook. In M. Celce-Murcia and McIntosh (Eds.) *Teaching English as a Second or Foreign Language* (pp. 302-307) NY: Newbury House.
- Laufer, B. (1989). What Percentage of Text-Lexis Is Essential for Comprehension? In C. Lauren & M. Nordmann (Eds.), *From Humans Thinking to Thinking Machines* (pp. 316-323). Clevedon: Multilingual Matters.
- Liu, N., & Nation, I.S.P. (1985). Factors Affecting Guessing Vocabulary in Context. *RELC Journal*, 16, 33-42.
- McKay, L. S. (2006) *Researching Second Language Classrooms*. Lawrence Erlbaum Associates, Inc., Publishers, New Jersey.
- Nation, P. & Meara, P. (2002). Vocabulary. In Schmitt, N. (ed.) *An introduction to Applied Linguistics*. Arnold: London.
- Nation, I.S.P. & Waring, R. (1997). Vocabulary Size, Text Coverage & Word Lists. In Schmitt, N., & McCarthy, M.(Eds.) *Vocabulary: Description, Acquisition, Pedagogy* (pp. 6-19). New York: Cambridge University Press.
- Schmitt, N. (2002) *An Introduction to Applied Linguistics*. Arnold London.
- West, M. (1953). *A General Service List of English Words*.