

Multiword Units for Pre-intermediate Students: An Experimental Basic English Corpus

プリ・インターミィディエイトの学生のための熟語研究：基礎英語コーパス

Sharif MEBED

シェリフ メベッド

Key words : Multiword Units(MWUs), Corpus, N-grams, Word Frequency

キーワード：熟語、コーパス、Nグラム、語彙頻度

Abstract

In the past 20 years, a number of research projects have endeavored to discover which vocabulary words learners of English as a foreign language should study. Researchers have discovered which English words appear with greatest frequency in English language texts using the British National Corpus (BNC: 10 million words) and the American National Corpus (ANC: 2 million words) among others. Words with higher frequency by definition are used more often and therefore are more useful to learners, as they are more likely to be encountered. However, semantic information is not limited to the single word level, but exists within phrases or multi-word units (MWUs: idioms, and expressions which are made up of two or more words). The present article attempts to discover which English MWUs are most frequent and of most use to English as a foreign language learners. However, conducting a search of existing corpuses may not be the solution for pre-intermediate EFL students, as they are based on news media, medical research articles, papers on engineering and court proceedings, among others. For that reason, the researcher has prepared a corpus composed of children's literature, and language aimed at linguistically less sophisticated audiences. In this way the researcher hopes to find the most basic and essential MWUs, or *must-learn* expressions. The search of this corpus returned 64 very high frequency MWUs, which could comprise a list of expressions taught to secondary and university students. The results are presented with a commentary and analysis.

要約

英語学習において学習者が学ぶべき単語に関して、過去20年間、様々な研究が行われている。BNC（1億語のイギリス英語のコーパス）やANC（2千万語以上のアメリカ英語のコーパス）の研究によって最も頻度の高い英単語は判明している。頻度の高い語は最も普段に使われている単語であり、学習者が学ぶべきであろうと考えている。しかし英語のみならず、人間の言語は単語単位だけではなく、句(phrase)のレベルが存在する。本研究では、学習者が覚えるべき最も頻度の高いMWU（慣用語等・複数の単語によって成り立つ表現）をコーパスから探し出すことを試みる。しかし、BNCなどに現れるMWUはニュース・メディア、医学書、工学研究の論文、法廷の記録など様々な高レベルの文章を元に作られている。そのため、英語を外国語とし学ぶ日本の大学で学ぶ学習者に適切なMWUであると限らない。そのため、本研究で新しい英語学習者のためのコーパスを作って、その中に学習者が学ぶべき基礎のMWUを探し出した。結果として、学習者が学ぶべき約64組のMWUを特定した。これらのMWUは、かならず学習者紹介する英語の慣用語であると論じる。

Introduction

The goal of this article is to consider which multiword units (MWUs) should be introduced to pre-intermediate level students at an early stage in their development. The current research attempts to discover the most common multiword units that students of English as a foreign language (EFL) are likely to encounter. Since the late 1980s, and the advent of corpus linguistics into foreign language education, a great emphasis has been placed on introducing higher frequency vocabulary before low frequency. Nation (2001, p.13) argues that the most frequent 2000 words on the General Service List (West, 1953) account for 80% of language in most texts. The importance of learning these words is then self-evident. Some textbook publishers have not ignored the situation and advertise the fact that their texts are corpus based, suggesting that they are more effective in improving communicative skill. One example of this is Cambridge University Press's "Real English Guarantee" appearing on the back of Redman's *English Vocabulary in Use* (2003).

Although finding the most frequent vocabulary words and introducing them to students enables the highest pay-off in terms of study-time investment, there are a number of problems associated with a frequency-centered approach. The biggest concern I have is that frequency of single words may hide the need to consider multiword units (MWUs) at an earlier point in language education. Multiword units

should be taking on a more important role in the eyes of EFL teachers. For example, Rogers (2000) suggested that the study of phraseology would be the next important phase in EFL learning. Lewis (2008, p.95) takes the argument a step further, suggesting a critical role for MWUs:

It now seems plausible that an important part of language acquisition is the ability to produce lexical phrases as unanalyzed wholes or 'chunks', and that these chunks become the raw data by which the learner begins to perceive patterns, morphology, and those other features of language traditionally thought of as 'grammar'.

According to Lewis, unanalyzed chunks, of which MWUs are a part, are not just good to learn, but they are the basis for coming into contact with patterns in language. For this reason, I argue that elementary-level students need explicit instruction in MWUs at early stages. Currently, MWUs are taught randomly at best, and at worst they are totally ignored. However, as Lewis suggests, learners may benefit from early explicit introduction of MWUs. Not only learning particular MWUs, but students and teachers need to be aware of MWUs as a central aspect of communication and language organization. Likewise, Willis (1990, p.38) argues the need to research the most common words and the most common patterns in the language, and expose students to them. Thanks in part to projects like Cobuild and the BNC (British National Corpus), it is clear which single words are most frequent, but there seems to be no consensus about multiword units.

What I suggest in the present paper is that like word frequency lists, we can develop a list of frequent MWUs that are "must-learn" for students. Below, I will focus on the most basic MWUs with the goal of making a foundation for materials intended for pre-intermediate English language learners.

Literature Review

There are a number of scholarly works of immediate pertinence to this study which I would like to discuss briefly at this point. Hsu (2006) compares three textbooks that claim to introduce multiword units to find which are the most helpful for students to learn. His results, however, indicate that there is no consensus among

the small sample investigated. He suggests that the reason for the disagreement is that two of the three textbooks did not base their choice of MWUs on corpus evidence. Another investigation into a similar field is Grant & Bauer (2004). In that study the researchers try to reanalyze the categorizations of MWUs for the purpose of teaching to non-native speakers. Although Grant and Bauer do not use a corpus to investigate, their paper addresses the concerns of MWUs for EFL/ESL students and therefore provides a number of hints that I will mention below. Simpson and Mendis (2003) investigate the appearance of idioms in the Michigan Corpus of Academic Spoken English (MICASE), a corpus made from recordings of classroom interaction and lectures recorded between 1997 and 2000. The investigation found a great deal of idioms used in both humanities and science lectures, as well as discussions. The investigators report repeated use of many opaque and figurative expressions like “in a nutshell”, “get a handle on”, and “on the same page as”. Their main finding was that highly opaque expressions are used very often in the classroom, and also that it is possible to mine a narrow corpus for a number of helpful idioms. They also produced a list of idioms that one might expect to encounter in a university classroom. Simpson and Mendis’s research may have some similarities with the current research in terms of its goals and the use of a specialized corpus.

Moon (1997) noted the difficulty in finding the most common MWUs. She also pointed out that the occurrence of colorful idioms (“the top dog”, “pick of the litter”) have very low frequencies. However there are a large number of MWUs that appear with high frequencies yet may be unknown to many pre-intermediate students, since they are often missed by many texts and syllabi that concentrate on teaching rule-generated language. These high frequency MWUs will be the focus of this article. Finally, Shin and Nation (2008) identify the most common MWUs in the BNC spoken corpus. This list is an excellent guide for conversation teachers. However, it may not be enough for reading and writing classes. For that reason, I will analyze a different corpus to find common MWUs that will best assist teachers and materials creators in Japan.

The problem of identifying multiword units

One major obstacle to this research and to the research of others mentioned above is the definition of a *multiword unit*. There is a range of opinion on what constitutes

a MWU. In their review of the literature, Grant & Bauer (2004) note the differences between categorizations of major researchers, but find three basic aspects of MWU definition that run through most of the categorization attempts. These are: institutionalization, non-compositionality and frozenness.

Institutionalization refers to whether or not the language community considers a particular MWU to be one unit. The key test for this is whether or not it recurs. Non-compositionality refers to whether or not a MWU can be understood by understanding each item in the unit. For example, even if a learner knows the meaning of the elements in “living hand to mouth,” he or she will not be able to determine the idiomatic meaning without knowing the socially defined multiword signification. Finally, there is the aspect of frozenness or fixedness. An example of this is, ‘run’ in the idiom ‘run the gauntlet’ which cannot be rewritten as “gauntlet running” or “sprint the gauntlet”. For a phrase to be an MWU, all three of these characteristics should be present.

Using these three conditions, I attempted to identify MWUs within a corpus of simplified English, which I will describe below. Many writers have made efforts to distinguish the various types of MWUs, Moon (1997, p.44) lists seven types: compounds, phrasal verbs; opaque idioms; fixed phrases; and prefabs, while Yorio (1980) defines only three. For the purpose of this research, my main concern will not be defining the type of MWUs that appear in the corpus. Rather, I will only be concerned with identifying MWUs as opposed to common collocations based on the guidelines above. Questions concerning what kind of MWUs will be touched upon only when it directly affects the results of the current research.

Methodology I : Finding the multiword-units in a corpus (n-grams)

This search for the most frequent MWUs centers on a study of a language corpus to find which MWUs are of high frequency. However since there are so many known MWUs, (The Oxford Idioms Dictionary (Parkinson D. & Francis B., 2006) claims to have 10,000), a search for all of them would be extremely time-consuming. The other option is to ascertain what word combinations exist in the corpus and then from that list try to discover MWUs. To accomplish that, I made use of a software tool that searches for n-grams (common word combinations). N-grams have been the topic of much research, especially by computer scientists working on translation and voice

recognition software. Among these Cheng et. al. (2006) discuss n-grams as well as some of their drawbacks and the implementation of improvements to computerized n-grams searches. Although there are some problems related to the use of n-gram searches, which I will take up in the discussion section below, it remains an efficient means for use in discovering MWUs in a corpus.

Methodology II : A self-made corpus -- The Simple English Corpus

The present research is intended to discover multiword units that are appropriate for pre-intermediate students to learn. For this reason, I have chosen to create a small corpus of relatively non-complex English. The corpus used here is composed from three different sources. The first of these is the Simple English Wikipedia (http://simple.wikipedia.org/wiki/Wikipedia:Simple_English_Wikipedia). The writers of Wikipedia's Simple English articles are required to use more accessible language than a typical English Wikipedia article, be it with simpler words, or fewer idiomatic expressions. Furthermore, sentences are markedly shorter in the Simple English section. Together the 35 articles taken from Wikipedia for this study have approximately 57,000 running words.

From the articles on the Wikipedia Simple English website, I have elected to use entries that have been chosen by the Wikipedia editors as 'well written', appearing in the list of 'good articles' (http://simple.wikipedia.org/wiki/Wikipedia:Very_good_articles). After the file was created, I used a vocabulary profiler available at Compleat Vocabulary Tutor website (<http://www.lex Tutor.ca/>) to analyze the 35 Simple Wikipedia articles included in the corpus. That analysis returned results indicating that 75.58% of the tokens in the Wikipedia sub-corpus were in the most frequent 1000 words of the BNC. Approximately 7.90% of the tokens were found in the range of 1001-2000, and 3.33% were from the list of 2001 to 3000. This means 83% of tokens in the file can be found in the 1-1000 and 1001-2000 levels of the BNC, and nearly 87% of the tokens appear within the most frequent 3000 words of the BNC. All frequency lists are lemmatized.

The next section of the self-developed corpus is the children's literature section. This section is composed of children's books available on the Internet, including *Wind in the Willows*, *Pinocchio*, and *Peter Pan* from sites like the Guttenberg Project. The portion of the corpus from these sources had 250,000 running words. 80.29% of the

tokens in this section were found to be in the list of the BNC's most frequent 1000. 6.17% were found in the 1001 to 2000 range. A further 3.04% were found in the 2001-3000 range. In all, approximately 89% of the Children's Literature sub-corpus was within range of the most frequent 3000 items of the BNC.

The third source for this corpus was Time Magazine, which offers a version of its articles on-line rewritten for children. Some articles in the corpus are aimed at native elementary school first and second graders, and other articles are aimed at third and fourth graders. The level of difficulty of vocabulary is comparable to the Simple Wikipedia articles. 86% of tokens are covered within the list of most frequent 3000 items appearing in the BNC.

The Simple-English Corpus that I have developed is not necessarily simple to read for EFL learners. I have found that the pre-intermediate students that I teach have a lot of trouble understanding much of what is in the corpus. However the writers who produced the texts had a target audience with a less sophisticated knowledge of the language than the writers of general texts like those composing the BNC. If we agree with Nation's (2001, 148) claim that students should be reading texts that are just a little higher than their current level, then the English appearing in less sophisticated texts with more high frequency words are more appropriate for these learners than the BNC. The introduction of MWUs appearing within this simplified corpus may be more necessary for pre-intermediate students than many that appear in the BNC. It might be helpful to compare them to articles from the London Times, which tend to have only approximately 81% of words appearing in the list of most frequent 3000 words. Of course this rating does not take into account complexity of sentences and is only an indirect indicator.

Further considerations of the Simple English Corpus

Before I move on to a discussion of the MWUs found in the corpus, I would like to discuss a number of weaknesses of the corpus I have created. Firstly, as compared with existing corpora like the BNC and the Bank of English, the Simple English Corpus is not based on a wide variety of texts. There are only three different types of texts as noted above. Secondly, the children's literature section is dependent on works that are available freely online, which means that they have entered the public

domain 50 years after the writer's death. The result of this is some works may include expressions that are outdated. Thirdly, it does not contain a spoken component. All texts in the corpus are examples of written English. For these reasons, we must look critically at the results that the corpus produces.

In order to gauge how the Simple English Corpus differs from the BNC, I have listed the most frequent 4-grams (four word combinations) of both the BNC and the Simple-English Corpus.

BNC 4-grams	Simple-English Corpus 4-grams
I don't know	i don t know
the end of the	the old lady is
at the end of	(proper name)
at the same time	i don t want
i do n't think	as soon as he
for the first time	don t want to
on the other hand	for the first time
between # and #	i am going to
the rest of the	in the middle of
as a result of	and the old lady
in the case of	i m going to
one of the most	in the united states
# per_percent of the	at the same time
the secretary of state	in the solar system

In this comparison, we see that a few expressions appear in both lists, including the number one most frequent 4-gram of both. Even with this small sample, the print-media news component of the BNC is conspicuous. For example, “the secretary of state”, appears frequently in the BNC. The Simple-English Corpus retains many of the n-grams found in the BNC, but may yield more basic phrases which I argue need to be mastered by pre-intermediate students before they can move up to the intermediate level. However, the small size of the Simple-English Corpus causes some non-typical examples to occur as well, like “in the solar system” (not a MWU).

Methodology: Discovering MWUs among the n-grams

In order to use the Basic English Corpus to do an n-gram search, I employed a software program developed by Laurence Anthony at Waseda University in Tokyo entitled AntConc (<http://www.antlab.sci.waseda.ac.jp/software.html>). This program

allows users to search for n-grams in data files. I performed searches for 2-grams, 3-grams, 4-grams and 5-grams. Once those lists were produced, they were recorded in a Microsoft Excel file. Next, those lists were searched manually to determine whether each n-gram was a MWU or not. Starting with the 2-grams, the most frequent two-word combinations, the AntConc software analyzed the corpus returning all combinations that appeared six or more times. There were 6951 of those. The process was repeated with the 3-grams, of which there were 2071, the 4-grams numbering 273. Finally I analyzed the 5-grams, of which there were only 17 examples.

Based on the three characteristics of MWUs discussed earlier, I used the following test to determine which n-grams were MWUs. In order to pass this test each question should return the answer 'yes'.

- Institutionalality: Does the string have a robust and relatively static meaning, and is that the meaning for which it is used in the corpus at least 6 times?
- Non-compositionality: Is it impossible to understand the expression by checking the constituent words' dictionary definition?
- Frozenness: Does the meaning of the word combination change when the order of components is adjusted, or when synonyms of particular components are substituted?

However, determining whether or not words fulfilled the criteria was not always a simple task. For instance, the question of non-compositionality is extremely subjective. It could be argued that in the list below some examples are in fact understandable from the dictionary definition of the constituent words. I attempted to include words that were not totally clear from the constituent words, especially considering that the target audience for the MWU list has an unsophisticated knowledge of English, and may not be able to extrapolate abstract meanings as well as native speakers.

A further problem related to distinguishing MWU's from mere common collocations was the polysemous nature of the phrases. For example, the 2-gram "up to" appeared very high on the list of word combinations in the Basic English Corpus. It could be categorized as a MWU in the case that it meant, "you decide." However, it

may be more appropriate to look at this as simply a two-word preposition that can be decoded literally like in the following:

I walked right up to him and said, "who do you think you are?"

The question then is how is the word used in the corpus? Using the same software AntConc, I was able to do a KWIC concordance search of the Basic English Corpus, a section of which appears below.

22 a blaze of flowers; the creek that led up to the boat-house, the little wooden
23 first, he thought. Very warily he paddled up to the mouth of the creek, and was just passing
24 mysteriously at him, went straight up to the door and opened it, and in walked
25 the best of times. He came solemnly up to Toad, shook him by the paw, and
26 Then the Mole pulled his chair up to the table, and pitched into the cold
27 he floor, against the walls of the room, and even up to the ceiling. He listened for the
28 lage of the Dead. Pinocchio, in desperation, ran up to a doorway, threw himself
29 for Pinocchio!" "Pinocchio, come up to me!" shouted Harlequin. "Come to the arms
30 meet them on the road, what matter? I'll just run up to them, and say, 'Well,
31 He stepped into the field. He went up to the place where he had dug the hole

From this KWIC concordance, it is clear that the overwhelming majority of corpus evidence for "up to" indicates *direction of movement*, and not the *right to choose*. Therefore, as far as this particular corpus is concerned "up to" fails to qualify as a frequent MWU, since in most cases the meaning of 'up to' can be understood by analyzing the meaning of its constituent words. A number of other word combinations in the n-gram lists had various possibilities of interpretation and had to be investigated similarly with concordance software.

Results of corpus analysis

The sort of MWUs resulted in the following table, listing from highest to lowest frequency.

2-grams

of course
 at last
 at once
 instead of
 because of
 so much
 made of
 a lot
 far away
 away from
 filled with
 no longer
 long ago
 as for
 in time
 all over
 lots of
 if only
 even though
 never mind
 no doubt
 not yet
 fall asleep
 good deal
 look like
 pick up
 sitting room
 stretch out
 run away
 tired of
 all night
 set out (to depart)

look for

look out

look around

no sooner

noted for

get home

if ever

in hand

3-grams

as soon as

as well as

the end of

by this time

a bit of

the rest of

in the middle

a lot of

be able to

in order to

in spite of

as far as

as long as

the idea of

the next day

4-Grams

at the same time

one of the most

out of the way

for the last time

he could not help

I beg your pardon

5-grams

no sooner said than done

as hard as SB could

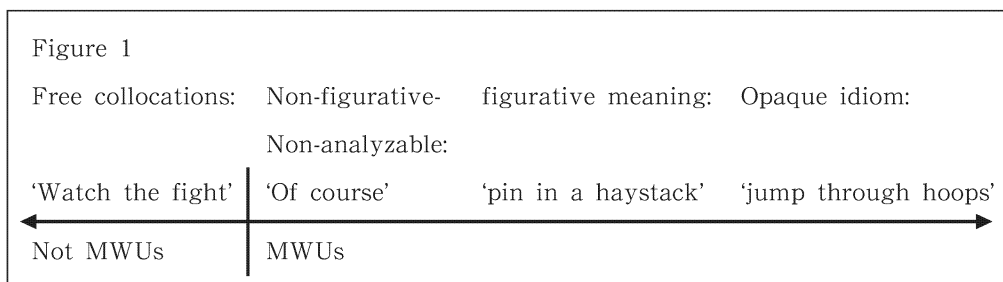
Commentary

A number of important pieces of information can be derived from the lists of MWUs that were produced. First of all, the MWUs listed above have high frequency, and the institutionalization is well attested. The full length of the corpus is approximately 312,000 running words. Therefore 6 instances would indicate a frequency of approximately 15 per one million running words, which would have a comparable frequency to single words like 'log' and 'advisory' on the BNC un-lemmatized frequency list (both ranked at approximately 4200). However, most of the expressions appeared more frequently than 6 times. For example "as soon as" appeared in the corpus 78 times. This is comparable to the single word "increase" ranked 430 on the same BNC frequency list.

In all, 63 MWUs were culled from approximately 9282 candidates. These 63 MWUs were composed of approximately 160 tokens, of which all but seven were in the British National Corpus most common 1000 word families (BNC-2000: asleep, beg, stretch, tired; BNC-3000: spite). The high frequency of the constituent words attests to the importance of these MWUs for early introduction to pre-intermediate students.

Basic nature of the MWUs found

We can think of the n-grams appearing in the search as set on a cline from free collocations on one side of the spectrum and opaque idioms on the other. This is illustrated in Figure 1.



In the previous literature, opaque idioms tend to be the focus. For example, Howarth (1998) discusses examples including "blow your horn", and "blow the gaff"; Grant (2003) discusses "a red herring", and "kick the bucket" among others. These are expressions that cannot be decoded by understanding their components alone. The present research is in contrast to those in that it focuses on much more basic MWUs

almost completely appearing on the left side of the chart in the region marked non-figurative non-analyzable. Although the MWUs discovered in the simple corpus lack the intrigue and humor of many MWUs taken up in the previous literature, their frequency indicates greater usefulness. Also, it is clear from classroom observation, that many Japanese university students are unfamiliar with a large number of these more common examples.

Of the 2-grams that were judged to be MWUs, the most frequent were almost exclusively adverbials. The second most common two-word MWUs were multiword verbs. The large number of multiword verbs indicates the great importance of this aspect of the English language. Despite that fact, many textbooks include far too few multiword verbs, and in casual observation, I have noticed that large numbers of university students learn low frequency verbs (like “expose”) before they learn semantically similar high frequency multiword verbs (like “find out”), though the reason for this is not clear.

Although there were many noun phrases appearing in the corpus, as a whole nearly all failed the test of non-compositionality. Only a few noun combinations were both high in frequency and non-compositional. One example in the list is “sitting room”, which is non-analyzable since it sounds as though a room is sitting rather than a room where people sit. Adding to this ambiguity is the fact that we sit in nearly every room of the house including the toilet. Therefore it is indeed a MWU.

Among the 3-grams, there are four examples of ‘as+adjective+as’ patterns: “as soon as”, “as well as”, “as far as” and “as long as”. The relative prevalence indicates the need not only to explicitly teach these examples, but also to raise the attention of students to the pattern itself. Finally, I would like to note the two idiomatic speech routines appearing in the data: ‘I beg your pardon’, and ‘no sooner said than done’. The former has a very pragmatic utility and should be memorized for production. The latter, lacks the usefulness of ‘I beg your pardon’ and appears only in one book of the children’s literature sub-corpus (Pinocchio). This limited range of appearance may warrant removal from the list.

Finally, one last point must be raised about the corpus search that generated the MWU list. It may be possible that the simplicity of the search technique has contributed to an elision of some frequent combinations that manifest variation. Cheng et. al. (2006) discuss the statistical evaluation of expression like “I can give you a ride”

and “I could give you a ride.” It would be possible that these combinations with one semantically equivalent difference could be lost in the n-gram search because they are counted as two different expressions. Future attempts to make lists of the most frequent MWUs would benefit from a consideration of variation.

Conclusion

The present study uses a custom-made corpus with the goal of uncovering some of the most common and useful MWUs for students who study English as a foreign language. Despite the fact that there are a number of problems associated with the corpus, including the corpus’s narrowness in terms of genre and size and the possibility of variant idioms escaping the search (like in Cheng et. al.’s study), the results do highlight a large number of core MWUs or phrases, that should be brought to the attention of pre-intermediate students. The MWUs found here are combinations of single-words that are almost entirely drawn from the 1000 most frequent words of the BNC. This fact indicates the very basic nature of the list. Nearly all of the MWU’s found in the corpus have semantic values that serve a particular pragmatic or communicative goal in a variety of situations, making them extremely useful for learners. They are not merely free or common collocations. Students need to study them as complete units, and cannot comprehend them based on their components. Additionally, the 2-gram section attests to the high frequency of multi-word verbs, and the 3-gram section highlighted the common pattern of *as + adjective + as*. With only a few exceptions, all of the MWUs extracted from the Simple English Corpus appear to be absolute must-learns for pre-intermediates and intermediate level students.

From here a number of possible refinements and augmentations of this research are possible. The Simple English Corpus needs to be refined with an expansion of the number of text genres. Also, a further consideration of word class patterns may help students and teachers to improve their understanding of MWUs. Finally, there needs to be an investigation into what degree students currently know these expressions in receptive and productive modes, followed by considerations of how to teach the unknown members of the list effectively.

References:

- Cheng, W., Greaves, C., & Warre, M. (2006). From n-gram to skipgram to congram. *International Journal of Corpus Linguistics*, 11(4), 411-433.
- Grant, L., & Bauer, L. (2004). Criteria for Re-defining Idioms: Are we barking up the wrong tree? *Applied Linguistics*, 25(1), 38-61.
- Grant L. (2003). *A Corpus-based investigation of idiomatic multiword units*. Unpublished Ph.D. thesis. Victoria University of Wellington.
- Howarth, P. (1998). Phraseology and Second Language Proficiency. *Applied Linguistics*, 19(1), 24-44.
- Hsu, J. (2006). An analysis of the multiword lexical units in contemporary ELT textbooks. Paper presented at the International Conference on English Teaching and Learning in the Republic of China (23rd, Kaohsiung, Taiwan, May 27-28, 2006). Retrieved from <http://www.eric.ed.gov/PDFS/ED497440.pdf>
- Lewis, M. (2008). *The lexical approach—The state of ELT and a way forward*. Andover: Heinle, Cengage Learning.
- Moon, R. (1997). Vocabulary connections multi-word items in English. In Schmitt, N. & M. McCarthy (Eds.), *Vocabulary description, acquisition and pedagogy*. (pp.40-63). Cambridge: Cambridge University Press.
- Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.
- Nation, I. S. P. (2004). A study of the most frequent word families in the British National Corpus. In P. Bogaards & B. Laufer-Dvorkin (Eds.), *Vocabulary in a second language: selection, acquisition, and testing*. (pp. 3-14). Philadelphia: John Benjamins Publishing Company.
- Parkinson, D. & Francis B. (2006). *Oxford idioms dictionary*. Oxford: Oxford University Press.
- Redman, S. (2003). *English vocabulary in use 2nd edition*. Cambridge: Cambridge University Press.
- Rogers, T. (2000). Methodology in the new millennium. *Forum*, 38(2), 2-16.
- Shin, D., & Nation, I. S. P. (2008). Beyond single words: The most frequent collocations in spoken English. *ELT Journal*, 62(4), 339-348.
- Simpson, R., & Mendis, D. (2003). A corpus-based study of idioms in academic speech. *TESOL Quarterly*, 37(3), 419-441.

Willis, D. (1990). *The lexical syllabus*. London: Collins ELT.

West, M. (1953). *A general service list of English words*. London: Longman, Green & Co..

Yorio, C. (1980). Conventionalized language forms and the development of communicative competence. *TESOL Quarterly*, 14, 433-442.