Investigating Multi-Word Items in a Contemporary ELT Course Book

Paul McAleese
Momoyama Gakuin University

Reference Data:
Mcaleese, P. (2013). Investigating multi-word items in a contemporary ELT course book. In N. Sonda & A. Krause (Eds.), *JALT2012 Conference Proceedings*. Tokyo: JALT.

Research in corpus linguistics and language acquisition has resulted in an increasing awareness that a significant amount of authentic language is made up of and processed as multi-word "chunks" (Sinclair, 1991; Biber & Conrad, 1999). It has become clear that awareness of such multi-word items (MWIs) can enhance language proficiency for learners in areas such as processing speed and pragmatic competence. However, only a small number of studies have investigated MWIs represented in language teaching materials. This exploratory study investigated MWIs in the course book *English Firsthand 1, 4th edition* (Helgesen, Brown, & Wiltshier, 2010) by using a large-scale corpus to determine the frequency at which these items exist in samples of authentic language. The study showed that while this course book incorporates many MWIs, a large proportion of them may be unrepresentative of authentic language and therefore have limited value to the learners in question.

コーパスと言語習得の研究によって、多くの一般的な言いまわしは複数の単語で構成されている定型文（multi-word items）であるという認識が高まってきている（Sinclair, 1991; Biber & Conrad, 1999）。その定型文の利用によって、学習者の流暢さや語用的な能力などを高めるということが明らかになっている。しかし、言語教育の教材に使用されている定型文を分析する研究はまだ少ない。本研究は大規模なコーパスを用いてEnglish Firsthand 1, 4th edition（Helgesen, Brown, & Wiltshier, 2010）という教科書に使用されている語彙的定型文はどこまで自然で代表的な英語であるかを検証する。その結果は、本教科書において定型文が数多く載っているが、対象の学習者にとって教育的に役に立たない定型文も数多く載っているということを表している。

In recent years there have been a large number of studies in the areas of phraseology and collocation. However, the definitions and terminology used to describe such language have varied due to their range of structural fixedness. In the literature these include *lexical phrases* (Nattinger & DeCarrico, 1992), *lexical bundles* (Biber & Conrad, 1999) and *formulaic language* (Wray, 2008). For this study it was decided to adopt the terminology *multi-word items* (MWIs) provided by Moon (1997), who described them as sequences of two or more words occurring together with a high degree of regularity and limited degree of structural variation. Examples of such language range from multi-word compounds to phrasal verbs and fixed or semi-fixed expressions. Although MWIs contrast with language that is syntactically constructed, their degree of fixedness allows for limited variation such as verb inflections, pluralization, or substitution of a single noun or up to two intervening words after a verb (see Table 1).

Table 1. Differentiating MWIs and Syntactically Constructed Language

MWIs	Syntactically constructed
[fixed] of course [semi-fixed] a lot of (money) [semi-fixed] take (good) care of	I'm going to the park today. Are you feeling alright?

Benefits of MWIs

A range of research has shown that recognition of MWIs can provide a number of benefits to second language learners. Recent corpus studies have suggested that MWIs are more prevalent in language than previously thought, with MWIs comprising from 20-50% of all language (Erman & Warran, 2000; Biber & Conrad, 1999). A number of studies on second language learners have also shown that knowledge of MWIs enhances the speed of encoding and decoding language and hence fluency (Ellis, 1996; Wood, 2007). Lewis (1993) and Nattinger and DeCarrico (1992) also argued that knowledge of functional MWI expressions, such as those used to clarify meaning or manage discourse, enhances pragmatic competence, giving immediate communicative benefits even to lower proficiency-level learners.

Evaluation of MWIs

To investigate MWIs in the course book in question, firstly criteria for investigation needed to be established. Clearly, a number of considerations such as frequency, range, learnability, and learner interests can be taken into account when evaluating the potential usefulness of lexis for syllabus design and materials (Mackey, 1965; White, 1988). For the purpose of this study, frequency and range were selected to evaluate the usefulness of MWIs to learners. High frequency language provides lexis that the learner is most likely to reencounter and therefore most representative of authentic language. Range ensures that the lexis in question is used across a variety of language registers and genres such as casual conversation, news programs, and magazines. Recent developments in the size of large scale, computer-based corpora provide much larger samples of language, and accordingly frequency and range can now be estimated with increasing degrees of accuracy. Although it is accepted that other considerations also need to be made when evaluating the pedagogical usefulness of MWIs to learners, frequency and range are generally considered to be primary criteria (Nation & Waring, 1997; Sinclair, 1991).

Previous Studies

It is clear from the literature that MWIs have a more prominent role in language than previously thought. Despite this, there have been only a small number of studies investigating the use of MWIs in ELT course books to date, and all have generally identified deficiencies in course book treatment of MWIs (Hsu, 2008; Koprowski, 2005; Meunier & Gouverneur, 2007).

Among the studies, Koprowski's (2005) stood out as particularly relevant to the current study. In Koprowski's study, MWIs were investigated in a small number of upper intermediate proficiency-level ELT course books. Using corpus frequency and range data, Koprowski concluded that not only did there appear to be no standardized criteria for MWI selection in the course books, but a large percentage of the MWIs selected had disproportionately low frequency and range values, suggesting limited pedagogical value to the target learners.

The studies identified to date have all focused on intermediate to advanced proficiency-level course books. In this study it was decided to adopt a methodology similar to that of Kowproski's

(2005) study, but apply it to lower proficiency-level teaching materials.

Method

The course book selected for this study was *English Firsthand 1, 4th Edition* (Helgesen, Brown, & Wiltshier, 2010). This book was used by this writer and a number of colleagues who teach EGP (English for General Purposes) courses in private Japanese university contexts. The learners were non-English majors from upper beginner (false beginner) to preintermediate proficiency levels. Published in 2010, this course book is contemporary and could potentially have accommodated the more recent developments in the understanding of MWIs outlined earlier.

All MWIs included in the course book were identified from the individual units' target vocabulary summaries in the course book appendix. These vocabulary lists were summaries of what the writers considered key words and phrases that were explicitly covered in the units. To differentiate the MWIs from more syntactically generated language covered in the units, care was taken to include only items that were explicitly introduced by the course book as complete units or chunks of vocabulary. Some examples of these were *alarm clock* (unit 4) and *excuse me?* (unit 5).

Determining Frequency and Range

A large-scale corpus was used to identify the frequency and range of each MWI. The corpus chosen for this study was The Bank of English Corpus (BOE; see Appendix A). This corpus was selected because it contained only language samples taken from countries where English is spoken as a first language (for example the UK and the US) and could therefore be considered a reasonably representative sample of authentic English. Also, as of 2008, this corpus consisted of approximately 450 million words including 20 subcorpora that cover a wide range of both written and transcribed spoken language, giving a large and wide-ranging sample size.

For the purpose of this study, it was decided to adopt the methodology developed by Koprowski (2005). This allowed for single numerical values to be calculated for each MWI, incorporating both corpus frequency and range data, allowing the MWIs to be easily compared and ranked. It also provided objective criteria for dealing with the structural variations of the MWIs outlined earlier. In accordance with Koprowski's methodology, R-scores for all MWIs identified were calculated by averaging the individual MWI corpus frequency values over the five subcorpora in which the items occurred most frequently. In other words, to get the R-score for each MWI, the frequencies (words per million) of the five subcorpora where the MWI occurred the most were totaled and divided by five. For example, for the MWI *healthy lifestyle*, the top five subcorpora were UK ephemera, US ephemera, UK books, Oz (Australian) papers, and UK magazines. By averaging these respective frequencies an R-score of 1.02 was calculated (see Table 2). Accordingly, a higher R-score suggests the MWI in question is more representative of authentic language.

Table 2. Calculation of R-Score for *Healthy Lifestyle*

MWI	UK ephemera	US ephemera	UK books	Oz papers	UK magazines	R-score
healthy lifestyle	1.9	1.1	0.8	0.7	0.6	1.02

In the case of MWIs exhibiting structural variations or polysemy, only the meanings in the context of the course book were considered. Pluralization, verb inflections, intervening words, and spelling variations were also taken into account (see Appendix B).

Results and Discussion

First, the total number of MWIs in the course book was tallied and calculated as a percentage of the total course book lexis (see Table 3). Next, R-scores were calculated for all MWIs identified, and then mean, median, and statistical range values were calculated for the course book as a whole (see Table 4).

Table 3. Number of MWIs Compared to Total Lexis

Total MWIs	Total lexis
220 (31.7%)	693 (100%)

Table 4. Total MWI R-Scores

Mean	Median	Statistical range
12.3	3.2	0.0-340.0

It is clear from the results that a significant proportion (31.7%) of the course book lexis is devoted to MWIs and there is a very wide statistical range of R-scores. Additionally, the proportionally low median value indicates that a significant number of MWIs have comparatively very low R-scores.

Number of MWIs and R-scores

A significant proportion of course book lexis is devoted to MWIs. Furthermore, with R-scores as high as 340, it is clear that some items have very high frequencies and ranges (see Table 5).

Table 5. Top 10 MWI R-scores

MWI	R-score
about (170 cm)	340.04
pay for	131.7
set up	109.98
credit card	97.22
pick up	93.92
in front of	83.48
on (that) street	82.06
I hope . . .	81.82
is born	67.85
every day	66.82

However, the widely varying R-scores suggest the course book writers are not consistently referring to corpus-based MWI frequency and range lists when making selections. This in turn suggests they are not consistently presenting language that is representative of authentic English.

Further Investigation of Disproportionately Low R-scores

As outlined above, the R-scores exhibit very wide statistical ranges with comparatively low median values, suggesting a large proportion of R-scores are disproportionately low. Accordingly, it was decided to further investigate the proportion

of MWIs with comparatively very low R-scores. There appears to be no external criteria in the literature that can be used to determine what is a suitable minimum frequency or range for different learner proficiency levels. However, it was decided to investigate the proportion of R-scores under 0.5. This value is equivalent to one time per two million words, and refers to words occurring fewer than approximately 225 times in the 450 million word BOE corpus (0.00005% of the corpus), Examples of single words with similar R-scores are *rejectionist* (64 occurrences, R-score = 0.5) and *microflora* (22 occurrences, R-score = 0.3). Clearly, equivalent MWIs would have limited value to an upper beginner proficiency-level learner, even for purely receptive purposes. Accordingly, all MWIs with R-scores under 0.5 were identified (see Table 6).

Table 6. Total MWIs With R-scores Under 0.5

Total MWIs	MWIs under 0.5
220	73 (25.9%)

The results show that over a quarter (25.9%) of all MWIs identified have R-scores under 0.5. It was also discovered that, of these MWIs, the lowest 20 MWIs all have R-scores under 0.1, with five not occurring in the 450 million word corpus even one time (see Table 7).

Table 7. 20 Lowest R-Scores

MWI	R-score
has (a) round face	0.08
phone store	0.08
game center	0.06
import company	0.06
post office clerk	0.06
TV anchor	0.06
do fun stuff	0.04
computer table	0.04
rainbow-striped	0.04
portable DVD player	0.04
late 20s	0.02
poetry slam	0.02
mini-notebook computer	0
How do you say (that) in English?	0
culture festival	0
close your book	0
doing fingernail art	!
magic club	!
don't do anything special	!
I don't understand yet	!
karaoke place	!

Note. ! = no occurances in corpus

Issues Concerning Disproportionately Low R-scores

The results show approximately 25% of the MWIs studied have R-scores under 0.5, indicating that these items have very low frequency and range values. This suggests that these items are unrepresentative of authentic language and therefore of limited value to the learners in question. This is particularly worrying for upper beginner proficiency-level learners, whose materials would be expected to start with the most commonly used MWIs.

In many cases, the MWIs in question could be easily replaced with MWIs that have similar meanings but significantly higher R-scores. For example, *mini-notebook computer* has an R-score of 0; however, simply replacing the item with *laptop computer* would increase the R-score to 2.5. Another example would be replacing *computer table* (R-score = 0.04) with *computer desk* (R-score = 0.12).

As mentioned earlier, course book writers also consider factors other than representativeness when selecting MWIs. Some MWIs with low frequency and range values may have other pedagogical value such as relevance to learners or learnability. In this course book, there appear to be a small number of MWIs that fit into this category. For example, of the MWIs identified with low R-scores, *How do you say (that) in English?* (R-score = 0) would likely have particular pragmatic value to a second language learner. Also, an MWI like *close your book* (R-score = 0) may be useful in classroom management.

However, most of the MWIs with low R-scores would appear to have very limited value to the learners in question. The use of corpus-derived frequency and range data, while not the only pedagogical consideration, can at least provide an empirically based starting point for MWI selection. Failure to consider language authenticity, to at least remove questionable material, is doing a disservice to the students.

Conclusion

This study showed that a large number of MWIs are presented in the course book investigated, with MWIs making up over 30% of the total lexis covered. However, although it is generally accepted in the literature that frequency and range should be primary criteria for lexis selection (Nation & Waring, 1997; White 1988), the MWIs identified have significant variations in these values.

Most importantly, the corpus analysis revealed that over 25% of the MWIs investigated have extremely low frequency and range values, to the extent that a number of them do not even occur one time in a 450 million word corpus. This suggests a significant proportion of MWIs investigated are unrepresentative of real-life English, making them of limited value to the learners in question, at least on the basis of language authenticity. Furthermore, a number of the low-frequency items could be easily replaced with alternative or shorter MWIs, resulting in significantly higher frequencies and ranges.

It is clear that frequency and range data need to be given more consideration in MWI selection for this course book. At the very least, MWI frequency and ranges need to be checked, and those with very low values omitted or annotated in some way. Large-scale corpora and corpus-based materials are now widely available and can be easily accessed by not only course book writers but also educators and learners. Examples of these are the BNC (British National Corpus) and the COCA (Corpus of Contemporary American English), both of which offer free online access. Pedagogically-based, corpus-derived materials such as the *Phrasal Expressions List* (Martinez & Schmitt, 2012) and lists of frequently spoken collocations (Shin & Nation, 2008) are also useful starting points when deciding which MWIs to select for course books.

It is also important to remember a number of limitations when considering the results of this exploratory study. Firstly, course book writers are likely faced with a number of other issues when selecting MWIs. In addition to representativeness, factors such as relevance to learners and learnability also need to be taken into account. As a result, even inclusion of very low frequency and range MWIs may be pedagogically justified. Another limitation is the use of a corpus data to determine representativeness. A corpus can only provide an approximation of authentic language and will always have limits on amount

and type of language. Although the BOE corpus used in this study has approximately 450 million words and includes a wide variety of language, it has more written language and spoken British English, which could potentially distort the results of this study. However, even with such limitations in mind, use of corpora can provide empirically based data upon which to base course book language selection.

This is an exploratory study on a single course book, so clearly further research is needed before more concrete recommendations can be made. Repeating the study to check for frequency and range using other corpora or combination of copora, or even a different methodology for determining representativeness using frequency and range, may provide more support to the findings of this study. Also, further investigation of the course book writers' motivations for including such a significant number of unrepresentative MWIs may help clarify any possible alternative pedagogical justifications used. It would also be interesting to widen the study to investigate the representativeness of MWIs in other course books at the same proficiency level to see if similar results are obtained.

Bio Data

Paul McAleese began his teaching career in Japan after graduating from Waikato University in New Zealand. He began working at conversation schools and then, after completing the CELTA course, became active in ESP courses and testing. He later developed an interest in pragmatics and spoken discourse, and last year completed an MA in applied linguistics at the University of Birmingham. He is now teaching at a number of universities in the Kansai area. <paul@pomaka.com>

References

Biber, D., & Conrad, S. (1999). Lexical bundles in conversation and academic prose. In H. Hasselgard & S. Oksefjell (Eds.), *Out of corpora: Studies in honor of Stig Johansson* (pp.181-189). Amsterdam: Rodopi.

Ellis, N. C. (1996). Sequencing in SLA: Phonological memory, chunking, and points of order. *Studies in Second Language Acquisition, 18*, 91-126.

Erman, B., & Warran, B. (2000). The idiom principle and the open choice principle. *Text, 20*, 29-62.

Helgesen, M., Brown, S., & Wiltshier, J. (2010). *English firsthand 1 student's book* (4th ed.). Hong Kong: Pearson Longman.

Hsu, J. T. (2008). Role of the multi-word lexical units in current EFL/ESL course books. *US-China Foreign Language, 6*, 27-39.

Koprowski, M. (2005). Investigating the usefulness of lexical phrases in contemporary course books. *ELT Journal, 54*, 322-332.

Lewis, M. (1993). *The lexical approach: The state of ELT and a way forward*. London: Language Teaching Publications.

Meunier, F., & Gouverneur, C. (2007). The treatment of phraseology in ELT course books. In L. Hildalgo, L. Querera, & J. Santana (Eds.), *Language and computers: Corpora in the foreign language classroom, 61*, 119-139. Amsterdam: Rodopi.

Mackey, W. F. (1965). *Language teaching analysis*. London: Longman.

Moon, R. (1997). Vocabulary connections: Multi-word items in English. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 20-63). Cambridge: Cambridge University Press.

Martinez, R., & Schmitt, N. (2012). A phrasal expression list. *Applied Linguistics, 33*, 299-320.

Nation, P., & Waring, R. (1997). Vocabulary size, text coverage and word lists. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 6-19). Cambridge: Cambridge University Press.

Nattinger, J., & DeCarrico, J. (1992). *Lexical phrases and language teaching*. New York: Oxford University Press.

Shin, D., & Nation, P. (2008). Beyond single words: The most frequent collocations in spoken English. *ELT Journal, 62*, 339-348.

Sinclair, J. M. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

White, R.V. (1988). *The ELT curriculum: Design, innovation and management*. Oxford: Basil Blackwell.

Wood, D. (2007). Mastering the English formula: Fluency development of Japanese learners in a study abroad context. *JALT Journal, 29*, 209-230.

Wray, A. (2008). *Formulaic language: Pushing the boundaries*. New York: Oxford University Press.

Appendix A

The Bank of English Corpus (BOE)

The BOE is a computer-based collection of authentic English that, as of 2008, consists of approximately 450 million words. It was developed jointly by COBUILD (a division of Harper Collins Publishing) and the University of Birmingham, and is made up of 20 different subcorpora covering a wide range of both written and transcribed spoken language. The written subcorpora include samples from material such as magazines, newspapers, letters, nonfiction, and fiction books. The spoken subcorpora include transcribed samples from material such as casual conversation, meetings, radio, and discussions (see below for corpus profile). In order for the corpus to be as representative of authentic English as possible, all the language samples were taken from countries where English is used as a first language (for example the UK and the US). Most of the samples were entered into the BOE after 1990 and the BOE is periodically reviewed for language variety and balance to have the corpus continue to reasonably reflect contemporary English.

Table A1. Profile of BOE Corpus

Subcorpora	Number of words
US Academic books	6,341,888
US Ephemera	3,506,272
UK New Scientist	7,894,959
US Public radio	22,232,422
UK Sun/NoW	44,756,902
UK Books	43,367,592
UK Magazines	44,150,323
UK Guardian	32,274,484
UK Economist	15,716,140
UK BBC radio	18,604,882
US Spoken	2,023,482
UK Business	9,648,371
CA Canadian mixed corpus	15,920,137
OZ Papers	34,940,271
UK Ephemera	4,640,529
US Books	32,437,160
US Papers	10,002,620
UK Independent	28,075,280
UK Times	51,884,209
UK Spoken	20,078,901
TOTAL	448,496,824

Appendix B

Dealing with MWI Polysemy and Structural Variations

Polysemy

In cases where MWIs had potential multiple meanings, a random sample of 100 concordance lines containing the MWI was queried, the number having the original course book meaning was tallied, and then the original R-score was multiplied by this ratio. For example, *fill in* resulted in an initial R-score of 33.1, however only 82/100 of the random corpus concordance lines represented the same meaning as the course book, so the R-score was multiplied by 0.82 to give an adjusted R-score (see Table B1).

Table B1. Adjusting R-scores for Multiple Meanings

MWI	UK ephemera	US ephemera	UK books	Oz papers	UK magazines	R-score	Adjustment
fill in	68.3	26.7	25.8	23.1	21.6	27.14	x 0.82

My native speaker intuition was initially used to determine whether an MWI should be checked for multiple meanings. It is possible that some were not detected. However, due to the large number of MWIs investigated it was considered this would not significantly influence the overall results.

Singular and Plural Nouns

MWI *compounds* and *collocations* identified include countable nouns introduced in only their singular or plural form. In such cases, both the singular and plural forms of the items were also considered in the R-scores. For example, in the case of the MWI *love story*, the frequencies were combined with *love stories* before the R-score was calculated.

Verb Inflections

MWI phrasal verbs and collocations identified include verbs introduced in only one form. However, to gain accurate R-scores these were investigated in all their inflected forms. For example, in the case of the MWI *play tennis*, the frequencies were combined with *plays tennis*, *playing tennis* and *played tennis* before the R-score was calculated.

Intervening Words

MWIs containing verbs (not fixed expressions) were queried for 0-2 intervening words after the verb. A limit of two words was chosen because most collocations appeared to have no more than two modifiers. Additionally, allowing more than two words would potentially distort the results by possibly exceeding phrase or sentence boundaries. Intervening words were also considered in other constructions when a particular slot could obviously be filled by paradigmatic substitution. For example, in the MWI *in 5 years time*, the *5 years* would be considered a slot where the time value would be variable. In such cases, from 1-2 variable intervening words were also taken into account.

British and North American Spellings

As the course books sometimes introduced MWIs in British or North American spelling, both variants were included. For example, the MWI *movie theatre* was also queried, as well as *movie theater*.