Creating a Semantically Non-transparent Formulas List for EFL Students

Wenhua Hsu

***Abstract*—**This paper describes the attempt to establish a pedagogically useful list of the most frequent semantically non-transparent formulaic sequences for EFL students. The list was compiled from the academic section of the Corpus of Contemporary American English (COCA). The academic sub-corpus of the COCA contains 120 million running words across nine academic domains. Through the program *Collocate* and manual checking, a set of criteria were applied: frequency, range, dispersion, meaningfulness, grammatical well-formedness and semantic non-transparency. A total of 331 items of 2 to 5-word sequences were selected and they accounted for approximately 1.18% of the running words in the COCA-academic. As with other wordlists, it is hoped that this phrase list may serve as a reference for ELT teaching materials development.

***Keywords*—**formulaic sequences, semantic non-transparency

I. Introduction

VOVABULARY may be a good predictor of reading comprehension [1]. A rich vocabulary makes a reading task easier to perform. According to Nation, lexical coverage is defined as ‘‘the percentage of running words in the text known by the reader’’ and generally regarded as a gauge of whether a text is likely to be adequately understood [2]. Running words here refer to individual words. Nation and Waring reported that the first 2,000 most frequent words of English provided a lexical coverage of 78% to 92% in all sorts of written texts [3].

While lexical coverage with an emphasis on individual words is calculated, multi-word sequences are not taken into account. Among a plethora of multi-word combinations, not all of them are semantically transparent. Semantic transparency denotes how easily a multi-word sequence can be interpreted from its component words. Conversely, semantic non-transparency signifies that the individual words of an expression do not help each other to reveal the meaning as a whole.

Martinez and Murphy pointed out that non-transparent multi-word sequences may cause deceptive comprehension, especially when they are composed of the most frequent general words and concealed among them [4]. Students may presume that they are familiar with these very common words (e.g., *as, in, of, that, well*) but actually they are not acquainted with the words in combination (e.g., *as of, in that, as well, as well as*) and may even deduce a wrong meaning. Multi-word sequences of general use may traverse various academic domains along with high-frequency words. If no distinction is made between individual high-frequency words and multi-word sequences made up of high-frequency words, the latter may be overlooked or misinterpreted.

As such, the lexical coverage of a text may be overestimated when non-transparent multi-word sequences are hidden in known words and their meanings as a whole happen to be unknown to learners. This research targeted recurrent multi-word sequences (so called formulaic sequences/formulas) at the 2K word level with a focus on semantic non-transparency. Compiling a semantically opaque formulas list may contribute to filling the chasm of lexical coverage that the first 2,000 most frequent words fail to account for. This research sought to answer the following two questions.

1. What are the most frequently-occurring non-transparent formulaic sequences at the 2K word level?
2. How important are the most frequent non-transparent formulaic sequences in academic texts (specifically, % lexical coverage in tokens)?

II. Literature Review

A. Approaches to Retrieving Multi-word Units

A text of any genre is not only made up of individual words but also a large number of multi-word sequences. In language use, some words co-occur with other words with greater than random frequency and may constitute a large portion of discourse [5]. This phenomenon is generally referred to as formulaic language and each individual case of formulaic language is called a formulaic sequence [6].

In the literature, a variety of terms have been used to refer to frequent multi-word sequences and have been studied under different rubrics, such as collocations [7], formulaic sequences/formulas [8]–[10], lexical bundles [11]–[14] and *n-grams* [15].

Although researchers give different definitions to recurrent multi-word sequences, there are two fundamental approaches used to retrieve them: a frequency-based approach and a phraseological approach [16]. Through computer software with statistical measures installed, automatic searches to extract recurrent word strings rely on frequency, dispersion, range and collocational strength as screening criteria, whereas the phraseological approach primarily resorts to semantics and grammar, and hence manual judgment is indispensable.

The pre-determined cut-off points in the literature for frequency and dispersion have been arbitrary, subject to researchers’ goals. Biber and his colleagues adopted a very flexible cut-off point at a minimum of ten times across five or

Wenhua Hsu is with I-Shou University, Kaohsiung 84001 Taiwan (e-mail: whh@isu.edu.tw).

http://dx.doi.org/10.15242/ICEHM.ED0115019

more texts per million words [17]. Cortes was more conservative and opted for 20 times, when conducting a survey on a comparison of the frequency and function of lexical bundles used in published writing and student disciplinary writing in history and biology [13]. Biber, Conrad and Cortes were even more cautious in choosing lexical bundles from their corpora by setting a relatively high frequency cut-off at 40 times per million words [12]. Following Biber et al.'s approach, Hyland increased the cut-off value from a minimum of 10 times to 20 times per million words and decided on the breadth of lexical bundles at occurring in at least 10% of the texts, when selecting lexical bundles in his 3.5-millon-word corpus of academic writing in articles, PhD dissertations and Master's theses [14]. Hyland found that as strings are extended to five or more words, their frequencies drop dramatically.

Present-day n-gram programs ensure the properties of frequency and multi-text occurrences. Nevertheless, they do not adequately deal with meaningful retrievals. Purely based on statistical measures, a phrase extractor may generate a long list of multi-word sequences, part of which have little meanings (e.g., *that do not, which is the*) or part of which are grammatically ill-formed as in the examples of '*is one of the*', '*was found in the*' and '*of the distribution of*'. These instances all meet the selection principles, i.e. frequency and breadth of use. Though being frequent and widespread, such lexical bundles may not, however, be "pedagogically compelling" [9].

B. Past Studies

To identify the most frequent collocations in spoken English from the British National Corpus, which need to be meaningful and comprehensible for deliberate learning, Shin and Nation proposed a set of criteria, one of which was "grammatical well-formedness" and involved a great deal of manual checking [18]. They targeted a sequence of words which do not span "immediate constituents" [19] (i.e. two neighboring phrases/clauses), because a grammatical well-formed word sequence is a comprehensible unit. For instance, '*the fact that'* is more understandable than '*fact that the'*.

To tackle the problem of teachability, Simpson-Vlach & Ellis put forward the notion of Formula Teaching Worth (FTW) by incorporating mutual information (MI) into their weeding procedure in lieu of a purely frequency-based approach [9]. MI is a statistical measure of the cohesiveness of words, which signifies collocational strength and a degree of idiomaticity [20]. Therefore, recurrent multi-word combinations with a high MI score are more likely to be meaningful and hence pedagogically useful. In one of their cases, the sequence of words '*and of the'* occurred more frequently than expected (passing a certain threshold of both frequency and distributional range); however it does not seem to be pedagogically useful. On the other hand, the expression '*on the other hand*' cohered much more than would be expected by chance based on the high mutual information score and is more likely to be pedagogically relevant.

In the above-mentioned studies, semantic opacity was not considered. Instead, Martinez and Schmitt sought to identify the most frequent non-transparent phrasal expressions that are compatible with the BNC word-frequency wordlists [8]. Referring to Wray and Namba's eleven criteria regarding whether a word string is a formulaic sequence [21], Martinez and Schmitt established six post-hoc criteria for use after a frequency-based n-gram search to minimize intuitions. They were mainly related to the judgment of whether an expression is a Morpheme Equivalent Unit and semantically opaque.

The review of previous studies has helped to shape our own approach to selecting the most frequent non-transparent formulaic sequences, which will be specified as follows.

III. RESEARCH METHD

A. The Corpus

The present data for research was downloaded from the Corpus of Contemporary American English. The COCA contains more than 450 million words of text and is equally divided among spoken, fiction, popular magazines, newspapers, and academic texts The COCA-academic was chosen because of its academic nature, large size (120 million words), contemporariness (involving the years from 1990 to the present) and wide coverage (encompassing nine academic fields: business, medicine, humanities, history, education, science & technology, law & political science, social sciences & geography, and philosophy, religion & psychology). The most commonly-used multi-word sequences retrieved from this academic corpus can cater to our students' needs, since they need to read English-medium academic materials. Regardless of major, they may encounter these frequent phrasal expressions very often while reading texts in their fields.

B. The Instruments

The n-gram software *Collocate* was used to retrieve multi-word sequences from the corpus [22]. The span parameter for word length was set from 2 to 5, because frequencies drop drastically as word sequences are extended to five words or beyond [14]. Though recurrent 5-word sequences may be relatively rare, they were included in the initial screening for the sake of completeness.

Heatley, Nation and Coxhead created the RANGE program, equipped with large-scale word family lists derived from the British National Corpus (BNC) and the COCA to measure the vocabulary of a text in frequency, word types and word families [23]. Using the BNC and the COCA, Nation and his colleagues compiled BNC/COCA 25,000 English word families and classified them into twenty-five ranked 1,000-word-family lists according to their occurring frequency, range and dispersion in the corpora [24]. In this research, in order to measure the vocabulary level of frequently-occurring non-transparent multi-word sequences, RANGE was utilized.

C. The Procedures

Referring to prior studies on multi-word sequences, the present research adopted a mixed approach. The present method involved several stages, revolving around frequency, range, dispersion, meaningfulness, grammatical well-formedness and semantic non-transparency. The frequency and range measures resembled those of lexical bundles used in past studies in some ways. Apart from quantitative measures, qualitative measures were further applied in this research. To lessen subjectivity, we referred to Martinez and Schmitt [8] as well as Shin and Nation

http://dx.doi.org/10.15242/ICEHM.ED0115019

[18] and thereby formulated three questions to guide the decision of candidate multi-word units for inclusion in the list. They were used to gauge meaningfulness (Q1), grammatical well-formedness (Q2), semantic non-transparency (Q3), after potential formulaic sequences were initially identified by the *Collocate* program. The three post-hoc questions were:

Q1. Does the candidate multi-word sequence convey a meaning?

Q2. Does the candidate multi-word sequence cross the boundary of an immediate constituent/phrase?

Q3. Is the candidate multi-word sequence semantically non-compositional? That is, the meaning as a whole does not remain when each component word is decoded with its core meaning.

Since this research used other measures to sift out multi-word sequences across sub-disciplines, a less rigorous criterion was set to begin with, namely once to four times per million words (a minimum of an occurrence per million words for 5-word sequences, twice for 4-word sequences, three times for 3-word sequences and four times for 2-word sequences), in order to prevent potential multi-word sequences from being excluded at the onset. As far as the 120 million words in the present corpus were concerned, appearing 120 to 480 times at least for the extraction of 5 to 2 words respectively were the selection threshold.

The next goal was to identify the formulaic sequences that appear across a wide range of academic domains. In consideration of formulaic sequences in widespread use, two decisions were made: 100% distributional range across nine academic domains and 50% dispersion across texts of the same subject domain as selection criteria. The decisions were admittedly arbitrary but relatively in agreement with the present goal (wide-range and even dispersion across all the academic disciplines) and more rigorous than the practice in the literature, to guard against idiosyncratic uses.

Another consideration given to the formulaic sequences for inclusion in the list was meaningfulness. The multi-word sequences to be selected must have meaning(s) and can be learned as a whole. This criterion would help to select formulas that can be comparable with the individual words in a frequency-based wordlist. Before manual checking for meaningful units, the measure *Mutual Information* did the initial screening.

High MI multi-word sequences are those with much greater coherence, which may have more easily identifiable, distinctive functions and meanings, and may thus be suitable for teaching. According to Hunston, collocations with an MI score no less than 3 are considered strong [25]. Multi-word sequences with high MI are those with much greater coherence and may thereby have more easily identifiable meanings. Multi-word sequences with the MI score lower than the default value (=3) were eliminated at this phase. They were, for example, '*with which the'* and *'to that of'*.

Subsequently, meaningfulness, grammatical well-formedness and semantic non-transparency guided manual checking. To lessen subjectivity, three post-hoc questions (see above) were used to guide the judgment.

For the post-hoc questions, the researcher and her colleague (associate professor of English linguistics) made an independent judgment on candidate multi-word sequences with an occurrence passing the frequency threshold and MI>=3. The 3-point scale was used and the responses of *yes*, *not sure* and *no* were coded as 1, 0.5 and 0 respectively. When the answers of both raters were the same, which shows a clear-cut decision, the entry was either excluded from or included for further analysis. When there was no agreement between the two raters or the answer was 'not sure', the entry was decided for tentative inclusion in the list.

It is worth mentioning here that in the process of comparing the two raters' judgment in deciding non-transparent formulas, disagreement often occurred in a polysemous formula or in a formula with one of its component words having multiple meanings. The cases in point of the former are *account for*, *make up*, and *be used to*, while the instances for the latter are '*view'* in '*point of view'* and *'in view of'*. One rater considered a formulaic sequence involving polysemy as opaque, because she presumed that students may not know all of the core meanings of a polysemous word, leading to an inaccurate interpretation of the entire word combination.

Additionally, we spotted that some candidate multi-word sequences having a word with a derivational affix may mislead learners into making a wrong form-meaning connection of the whole. This is because our learners think they know the base form of the word but they are unaware that the meaning of its derivational form has been altered (e.g., like, alike and likely; respect, respective, irrespective; verse and versed). Therefore, we considered a potential formulaic sequence involving such a case as non-transparent.

For manual checking, the Cohen's Kappa statistic was repeatedly used to test the inter-rater reliability. The k values were 0.95, 0.91 and 0.88 (>0.80) for Q1, Q2 and Q3 respectively, reaching a substantial level of agreement between the two raters.

IV. Results And Discussion

A. The Most Frequent Non-transparent Formulaic Sequences at the 2K Word Level

Large A total of 311 non-compositional multi-word expressions of 2 to 5 words were ultimately chosen and formed the Non-Transparent Formulas List. The list encompassed 190 two-word, 104 three-word, 35 four-word and 2 five-word interdisciplinary opaque formulaic sequences commonly used in academic genre.

Table 1 presents a full picture of the percentage coverage of the non-transparent formulas in the BNC/COCA 2,000 word families. The non-transparent formulas list consisted of 1,083 running words and involved 416 word types as well as 291 word families. The BNC/COCA first 1,000 word families accounted for 93.49% of the total words in the list and the second 1,000 made up 6.51%.

TABLE I
TOKENS AND LEXICAL COVERAGE AT THE BNC/COCA 2,000 BASE WORD LISTS FOR THE NON-TRANSPARENT FORMULAS LIST

BNC/COCA base word lists	Tokens (running words)	% coverage in tokens	Number of word types	Number of word families
1st 1,000	1,006	92.89%	290	217
2nd 1,000	77	7.11%	126	74

The pairings or strings of content words (nouns, lexical verbs, adjectives or adverbs) and function words (determiners, conjunctions, prepositions, pronouns, auxiliary verbs, modals and quantifiers) form a common pattern in the current list. For example, *much as* (=though), *as well as*, *in order to*, *there + be*, and *to do with*. Among the instances, the everyday words *as*, *well*, *order*, *do*, *much* and *there* do not have an independent meaning but are a component of a repertoire of multi-word combinations that make up a text, as Sinclair has claimed [26]. Without specialist knowledge involved, these semantically non-transparent word sequences occur across a wide range of subject areas with their high-frequency component words.

The length of multi-word units has some influence on semantic transparency. When formulaic sequences become longer, their potential for ambiguity and polysemy will decrease. As for how long non-transparent formulaic sequences can be, the present data shows that two 5-word sequences extended from three words can still be semantically opaque while retaining a cross-disciplinary attribute, as shown in the instances of *have ~ to do with* and *as far as ~ be concerned.*

Concerning the structure of 2-word sequences, a vast majority of them (175 out of 190) were grammatically-conditioned pairs, namely a content word combined with a function word, as opposed to only 15 lexical collocations, a content word tied with a content word (e.g., *simply put, no matter, so far, very few*). Amid grammatical collocations, phrasal verbs were in the majority (69/190=36.3%) (e.g., *account for, cope with, carry out*) and phrasal prepositions came second (18/190=9.47%) (e.g., *as for, apart from, as per, according to*), followed by the pattern *a preposition + a noun* (13/190=6.84%) (e.g., *at once, at times, in place, in question*), being the third.

The most common pattern in the 3-word sequences list was a passive verb followed by a preposition requiring a noun phrase or by an infinitive-*to* for completion. In the present academic corpus, past participle phrases came from a reduction of an adjective clause by omitting the relative pronoun and the verb-be form and used as a post-nominal adjective phrase to modify the preceding noun. For the sake of thoroughness and flexibility, they are presented as *(be) + past participle + preposition or infinitive-to,* as in the cases of *(be) bound to* and *(be) concerned with.* When the verb *be* is added, they form the passive and can stand alone appearing in an independent clause/sentence. Moreover, it should be noted that the frequent use of the passive voice without a *by*-phrase seems to be one of the grammatical features in academic prose. This also reveals a different picture of how we do the passive exercises from a grammar textbook in a General English class (the passive followed by a *by*-phrase) and how the passive is used in authentic discourse (i.e., the passive followed by a preposition other than *by* or followed by an infinitive *to*).

The three patterns *as ~ as*, *a ~ of*, and by + noun phrase were also productive among the 3-word units, as in the cases of *as much as, as far as, as soon as, a range of, a couple of, by means of,* and *by way of.* As enumerated, these three patterns contribute to the description of quantity, the coverage of a subject or an approach.

For 4-word sequences, the prepositional phrase was the most common structure, comprising about 56% of all forms in the category of 4-word sequences (20/35 items). They were, for instance, *in the light of, in the wake of, in the event of/that, on the grounds of/that, on one's own account.*

As can be seen, the structural types of the most frequent non-transparent formulas are proliferous and it may not be easy to fold them into a compact categorization. However, if applying Biber, Conrad and Cortes's functional taxonomy to the present non-transparent formulaic sequences, they can be divided into three types: referential expressions, stance expressions and discourse organizers. According to Biber, Conrad and Cortes, referential bundles make direct reference to physical or abstract entities or to the textual context. Stance bundles express attitudes or assessments of certainty that provide a frame for the interpretation of the subsequent proposition. Discourse organizers reflect relationships between prior and coming discourses [12]. Despite multiple functions depending on the context, the frequent non-transparent formulas were classified based on their most common use. Table 2 provides an overall distribution of the non-transparent formulaic sequences across the three primary functions.

TABELE II
DISTRIBUTION OF THE MOST NON-TRANSPARENT FORMULAS ACROSS FUNCTIONS

Non-transparent formulas	Number	Instances
Referential expressions	236	according to; by means of; such as; in terms of
Discourse organizers	75	so that; in order to; as well as; on the other hand
Stance expressions	20	assuming that; appear to; be likely to; in a sense

Among 331 non-transparent formulaic sequences, there were 236 referential formulaic sequences plus 75 discourse organizers, and 20 opaque formulas serve as stance expressions. It may be challenged on the precision of categorization as a result of the multilayered functions of some formulaic sequences. Nevertheless, the preliminary typology may display a general pattern concerning the usage of these frequently-occurring non-transparent formulaic sequences. The results show that a very high proportion of opaque formulaic sequences in academic genre were referential expressions, accounting for 71.3% of the total non-transparent formulas (=236/331). Discourse organizers were the second dominant (22.66%) while stance formulas were far less common (6.04%) in academic texts. This may be ascribed to the reason that the nature of academic prose is mostly expository, aiming to explain or illustrate theories and hypotheses rather than stating a personal stance on a topic.

B. The Lexical Coverage of the Most Frequent Non-transparent Formulas in Academic Texts

Number Research Question Two 'How important are the most frequent non-transparent formulaic sequences in academic texts (specifically, % lexical coverage in tokens)?' can be reformulated as "What is the text coverage (%) of the most frequent non-transparent formulas in the COCA-academic?" The present list contains a total of 331 formulaic phrases of 2 to 5 words with an accumulation of 367,716 individual instances and 1,416,000 running words, which makes up 1.18 % of the tokens in the COCA-academic.

V. CONCLUSION

The principal concern of this study was to create a semantically non-transparent subset of formulaic language for EFL students for receptive use. By means of a set of criteria, a total of 331 items of 2 to 5-word non-transparent formulaic sequences were selected and they made up 1.18 % of the running words in the COCA-academic. The present list contains the most widely-used phrases across various academic fields. It is made up of the BNC/COCA top 2,000 word families. Accordingly, the non-transparent phrasal expressions can bridge the gap between the lexical coverage that the most general words can and cannot account for in a text. Irrespective of their majors, EFL students may come across these opaque formulaic sequences while reading texts in their fields. The current list is short and may be a viable option for all fields of students to learn in a short time.

Despite arbitrary decisions on cut-off values in the compilation of a list of the most frequent non-transparent formulaic sequences, there may be some advantages to overt instruction of these frequent expressions. The effectiveness of learning opaque formulaic sequences is worth investigation but beyond the present focus. It is hoped that the list may provide some inspiration for future empirical studies and ELT materials development.

ACKNOWLEDGMENT

This study was supported by a Grant (MOST 103-2410-H-214-002) from the Ministry of Science and Technology in Taiwan.

REFERENCES

[1] M. Hu, and I.S.P. Nation, "Unknown vocabulary density and reading comprehension," *Reading in a Foreign Language*, vol. 13, no. 1, pp. 403–430, Jan. 2000.

[2] I.S.P. Nation, "How large a vocabulary is needed for reading and listening?" *The Canadian Modern Language Review*, vol. 63, no. 1, pp. 59–82, Jan. 2006. http://dx.doi.org/10.3138/cmlr.63.1.59

[3] I.S.P. Nation, and R. Waring, "Vocabulary size, text coverage and word lists," in *Vocabulary: Description, Acquisition and Pedagogy,* N. Schmitt, and M. McCarthy, Eds. Cambridge, England: Cambridge University Press, 1997, pp. 6-19.

[4] R. Martinez, and V. A. Murphy, "Effect of frequency and idiomaticity on second language reading comprehension," *TESOL Quarterly*, vol. *45*, no. 2, pp. 267-290, Apr. 2011. http://dx.doi.org/10.5054/tq.2011.247708

[5] N. Schmitt, and R. Carter, *Formulaic Sequences: Acquisition, Processing, and Use.* Amsterdam: John Benjamins, 2004. http://dx.doi.org/10.1075/lllt.9

[6] N. Schmitt, *Researching Vocabulary: A Vocabulary Research Manual.* Hampshire, England: Palgrave Macmillan, 2010. http://dx.doi.org/10.1057/9780230293977

[7] M. Lewis, *The Lexical Approach: The State of ELT and the Way Forward*. Hove, England: Language Teaching, 1993.

[8] R. Martinez, and N. Schmitt, "A phrasal expressions list," *Applied Linguistics*, vol. *33*, no. 3, pp. 299-320, July 2012. http://dx.doi.org/10.1093/applin/ams010

[9] R. Simpson-Valch, and N. Ellis, "An academic formulas list: New methods in phraseology Research," *Applied Linguistics*, vol. 31, no. 4, pp. 487-512, Oct. 2010. http://dx.doi.org/10.1093/applin/amp058

[10] A. Wray, *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press, 2002. http://dx.doi.org/10.1017/CBO9780511519772

[11] D. Biber, S. Conrad, and V. Cortes, "Lexical bundles in speech and writing: An initial taxonomy," in *Corpus linguistics by the Lune: A festschrift for Georffrey Leech,* A. Wilson, P. Rayson, and T. McEnery Eds. Fankfurt: Peter Lang, 2003, pp. 71-93.

[12] D. Biber, S. Conrad, and V. Cortes, "*If you look at* …: Lexical bundles in university teaching and textbooks," *Applied Linguistics*, vol. 25, no. 3, pp. 371-405, July 2004. http://dx.doi.org/10.1093/applin/25.3.371

[13] V. Cortes, "Lexical bundles in published and student disciplinary writing: Examples from history and biology," *English for Specific Purposes,* vol. *23,* no. 1, pp. 397-423, Jan. 2004. http://dx.doi.org/10.1016/j.esp.2003.12.001

[14] K. Hyland, "As can be seen: Lexical bundles and disciplinary variation," *English for Specific Purposes*, vol. 27, no. 1, pp. 4-21, Jan. 2008. http://dx.doi.org/10.1016/j.esp.2007.06.001

[15] M. Stubbs, "An example of frequent English phraseology: Distribution, structures and functions," in *Corpus Linguistics 25 years on,* R. Facchinetti, Ed. Amsterdam: Radopi, 2007, pp. 89-105.

[16] N. Nesselhauf, *Collocations in a Learner Corpus*. Amsterdam: John Benjamins, 2005. http://dx.doi.org/10.1075/scl.14

[17] D. Biber, S. Johansson, G. Leech, S. Conrad, and E. Finegan, *Longman Grammar of Spoken and Written English.* Harlow, England: Pearson, 1999.

[18] D. Shin, and I.S.P. Nation, "Beyond single words: The most frequent collocations in spoken English," *ELT Journal,* vol. *62*, no. 4, pp. 339-348, Oct. 2008. http://dx.doi.org/10.1093/elt/ccm091

[19] L. Bloomfield, *Language.* New York: Henry Holt, 1933.

[20] M. Stubbs, "Collocations and semantic profiles: on the cause of the trouble with quantitative studies," *Functions of Language*, vol. *2*, no. 1, 23-55, Jan. 1995. http://dx.doi.org/10.1075/fol.2.1.03stu

[21] A. Wray, and K. Namba, "Formulaic language in a Japanese-English bilingual child: A practical approach to data analysis," *Japan Journal for Multilingualism and Multiculturalism,* vol. 9, no. 1, pp. 24-51, Jan. 2003.

[22] M. Barlow, *Collocate* (Computer software). Houston, TX: Athelstan, 2004.

[23] A. Heatley, I.S.P. Nation, and A. Coxhead, *RANGE* (Computer software). Wellington, NZ: Victoria University of Wellington, 2002.

[24] I. S. P. Nation, "The BNC/COCA word family lists 25000," in *RANGE*. Wellington, NZ: Victoria University of Wellington, 2012.

[25] S. Hunston, *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press, 2002. http://dx.doi.org/10.1017/CBO9781139524773

[26] J. Sinclair, *Corpus, Concordance, Collocation*. Oxford: Oxford University Press, 1991.