Taiwan Journal of TESOL
Vol. 5.2, 1-28, 2008

WEB AS CORPUS, GOOGLE, AND TESOL: A NEW TRILOGY

Chris Shei

ABSTRACT

There is a kind of formulaic sequence pervasive in most genres of texts which has not been properly recognised and studied. They are not fixed in form and are usually lengthier than the relatively well studied two-word collocations and more irregular than fixed idioms. They appear in text as a package with a core collocation and some accompanying semantic, syntactic, or pragmatic features that influence the choices of words around the core. Stubbs (2002) proposed the term e*xtended lexical unit* for this kind of structure. In this article, the term *extended collocation* is used instead to better reflect its phraseological nature. It is argued that only the web itself is large enough to provide adequate instances for the investigation of extended collocations, and therefore the integration of phraseology into TESOL research and practice. Arguably the most difficult aspect of studying any formulaic sequences is the initial identification and validation of their phraseological status. This article proposes a reliable method of identifying extended collocations and other phraseological units via Google search. This method will be useful not only for research of phraseology but also for the teaching of English phraseology to speakers of other languages.

Key Words: Google, corpus, condordancing, extended collocation, phraseology

INTRODUCTION

It is now generally acknowledged that grammar and lexis are not the only two components of language worthy of research and teaching, and that there is a level of organization called phraseology. Unlike the relatively clear concept of ‘grammar’ or ‘vocabulary’, however, the concept of ‘phraseology’ is highly inclusive and inherently fuzzy. First of all, the concept itself maps into several different terms in the field:

formulaic speech, prefabricated routine, multi-word sequence, lexical phrase, to name just a few. Inevitably, the elements included in this fuzzy concept tend to be arbitrarily assigned and variously defined: *collocation, chunk, fossilized form, phraseme, stereotyped phrase*, and so on (See Wray, 2002, p. 9 for a comprehensive list). In this article, I shall use *phraseology* to refer to the level of structure which consists of word combinations that recur in text either in a prototypical form or as a recognizable variant. I also use *phraseological unit* as a general term for all the entities that are normally treated as or can be reasonably assumed to represent an instance of phraseology. Both collocations (*serious injury*) and idioms (*rain cats and dogs*) are a kind of phraseological unit.

Without attempting to offer a new typology of any kind for phraseology, I shall focus on a phraseological unit which I think so far has not been adequately recognized and investigated. I call this phraseological unit **extended collocation**. The following text includes at least one example of extended collocation.

(1)
The study, published in the journal Neurology, raises the possibility that caffeine may even protect against the development of dementia. (BBC News, 2007)

The fragment *protect against the development of* is a good example of extended collocation. Although Google returns 87,100 hits for this string, such a long expression cannot be found in either British National Corpus or the Bank of English in its entirety through their respective web interfaces. Here we are drawn to Sinclair's insight that "we have to have very large corpora indeed, in order to look at phraseology in any systematic way" (2004, p. 189). Research relying on the web as corpus is already beginning to thrive, from collecting instances of collocation (Guo & Zhang, 2007) to the development of linguistically tailored web search engines (Renouf, Kehoe, & Banerjee, 2005). If there are worries as to the usability of the relatively disorganized web in comparison to a well-designed corpus in linguistics research, they should be considerably eased by works done for example by Keller and Lapata (2003), who found "a high correlation between Web frequencies and corpus frequencies" (p. 459), among other things.

It is the intention of this article to support the view of the web as a resource for linguistic research and language teaching. The vehicle used

for this purpose is the search engine Google, which is used here for identifying phraseological units in text, including extended collocations. Detailed analyses of the extended collocation *protect against the development of* and other phraseological units will be conducted in the following sections based on the web as a corpus and Google as a tool. Overall, I wish to establish a web-based machinery for validating the phraseological status of word sequences, and point out how useful the Internet could be as a large and free corpus for linguistic circles in general and the TESOL profession in particular.

PROPERTY OF EXTENDED COLLOCATION

Research on phraseology seems to have gained momentum and increased its pace since the 1970s with the availability of computers and large corpora. By now a great deal of work has been done on well-known phraseological units such as collocations and idioms, exploring both their linguistic nature and pedagogical implications. Functions of formulaic sequences have been proposed, such as concisely expressing complicated ideas (Hill, 2000, p. 55), saving processing effort, signifying group membership and individual identity (Wray, 2002), and so on. The 'psycholinguistic reality' of formulaic language is also being examined by experiments like those conducted by Schmitt, Grandage, and Adolphs (2004) and Schmitt and Underwood (2004). Despite all the progress made, however, the fundamental question of what exactly phraseology is and what 'counts as' a phraseological unit, remains open.

The most frequently adopted criteria for phraseology include the semantic opaqueness of a formulaic sequence (i.e. its overall meaning cannot be inferred from the sum of its components), the fixedness of its form (i.e. neither overall word order nor individual items can be changed), and perhaps its deviant grammatical properties (e.g. *long time no see*). However, there is a kind of phraseological unit which is not so fixed in form and whose meaning is entirely compositional. We already saw an example in text (1) above: *protect against the development of*. This kind of phraseological unit is notably different from what Biber, Conrad, and Cortes (2004) called 'lexical bundles', or what Scott and Tribble (2006) termed 'word clusters', in that their units consist mostly of high frequency words, for example, *one of the, a number of, there was a* (Scott & Tribble, 2006, p. 132), and so on. The word bundles or clusters they describe tend to have more to do with 'discourse

mechanics' rather than with the information content itself. The phraseological units I am concerned with in this article, on the other hand, must have at least one content word in them. To apply a pair of conventional terms, Biber and Scott and their colleagues dealt with extended versions of *grammatical collocation*, while I deal with extended versions of *lexical collocation.*

Both Biber et al. (2004) and Scott and Tribble (2006) worked on corpora of limited size, and when this constraint is in force, it is not easy to discover frequently occurring clusters of more than three words' length which contain a content word. In contrast, we stand a better chance of finding larger clusters containing content words if we work on the Internet, which currently contains at least 3,033 million web pages, according to Keller and Lapata (2003). The sequence *protect against the development of* from (1) is a relevant example. This is a complex kind of lexical collocation which I will conveniently call *extended collocation* in this paper. Stubbs (2002) offers a cogent analysis of this kind of structure which he calls **extended lexical unit**.

My understanding of Stubbs (2002) is that there is a level of organization in language which involves a node word or phrase and their fuzzy extensions. This unit of language usually demonstrates four kinds of relations:

- Collocation
- Colligation
- Semantic preferences
- Discourse prosody

For example, we can take *naked eye* as the node phrase, which itself is an attested collocation. This node collocation often co-occurs with adjectives (colligation) denoting 'size' (semantic preference), e.g. *large enough to see with the naked eye* (Stubbs, 2002, p. 112). The entire unit when used in discourse often denotes a pragmatic meaning of 'visual difficulty', for example, in a negative sentence like *Mars did not appear large to the naked eye* (Google hit). Thus, this kind of multi-dimensional analysis restores life, so to speak, for word sequences which may seem unimpressive at first glance. Let us now apply this analytical framework to the word sequence *protect against the development of* by firstly observing more examples retrieved by Google.

(2)

Coffee Consumption May **Protect Against the Development of** Gout

(vitamin B3) may **protect against the development of** Alzheimer's disease

in the first year of life **protect against the development of** atopy in children

OC may **protect against the development of** rheumatoid factor

Does coffee **protect against the development of** Parkinson disease?

physical activity may **protect against the development of** depression

Deodorant composition to **protect against the development of** body odor

Does a higher number of siblings **protect against the development of** allergy and consumption of fruit and vegetables **protect against the development of** cancer?

breast feeding **protect against the development of** clinical symptoms of celiac disease

What we see in (2) are ten concordance lines extracted from the top ten of the 87,400 documents indexed by Google which match the search string *protect against the development of*. I would like to emphasize again that these concordance lines can only be drawn from the web, as no large corpus that I know of such as BNC or COBUILD Bank of English returns any solution to this search (although COBUILD does offer an example containing the cluster *protect against the potential development of*).

The first thing we notice about the concordance lines in (2) is that they all carry some kind of medical implication. In fact, this extended collocation seems always followed by the name of a disease. This will be its semantic preference according to Stubbs' scheme. Pragmatically, it seems to be used predominantly in a consultation setting, as can be gathered from the brief examples in (2). To use the terminology of register analysis (e.g. Eggins & Martin, 1997), the Field of the discourse *protect against the development of* finds itself in is medical advice regarding the prevention of a disease using a certain ingredient or method. The Tenor is that of a well-informed medical professional addressing the general public in a reserved manner, hence the use of modal auxiliary *may*, and plenty of topic-initiation questions. The Mode

of the text this extended collocation resides in is decidedly formal as the unit itself already contains a nominalization (i.e. *development*), which seems to influence the grammatical choice of other words, such as *consumption* towards the more formal side. Overall, it can be inferred that the cluster *protect against the development of* is indeed a phraseological unit or part of or a variant of a unit with distinct grammatical, semantic and pragmatic associations. More importantly, this multi-dimensional relation seems to indicate the conceptual foundation of the phraseology, which dictates a qualified speaker (e.g. a doctor) to address a layman with this unit or a variant when a conceptual structure habitually projected into it is activated.

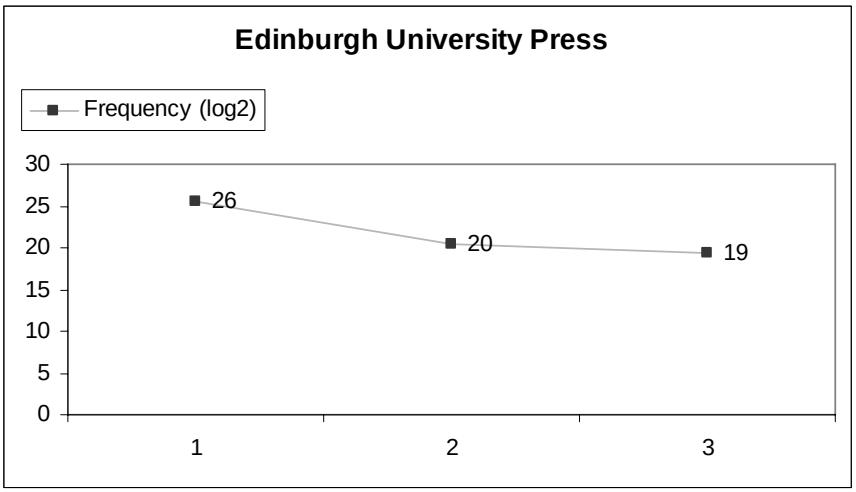
It is useful here to note Hunston and Francis' (2000) findings that pattern and meaning are strongly associated, and that grammar and lexis are inseparable, which I interpret to mean that both grammar and lexis are conceptually grounded. That is, grammatical patterns are not as 'lifeless' as people tend to think; nor are they so boundlessly creative. We could potentially say *This coat of painting can protect against the development of rust*, for example, but in effect, no similar instances are found on the web. This phraseology with its particular pattern and associated lexical items are used only in the medical setting that we discussed. People seem to prefer to use set phraseological units to express their associated ideas in a tightly projected manner whenever such links exist. Stubbs (2002) notices that the word *doses* often appears in a lexico-syntactic frame consisting of a verb denoting the meaning of 'give' or 'take', a size adjective such as *massive*, and a medical term, for example, *take high doses of vitamin E* (Google hit). He concludes that "when people talk about doses of something, then there are the meanings which frequently get expressed" (p. 87). This proposed link between phraseology and conceptual structures cannot be overlooked.

I hope I have drawn the reader's attention to the existence of extended collocations in text which can only be adequately observed and analyzed via the use of the Internet. In the next section, a Google-based phraseology frequency analysis shall be offered as a complementary method for identifying phraseological units on the web. Hopefully, this will help establish the importance of the web as a corpus for linguistic research, and for the application of results and methodology to the TESOL profession.

GOOGLE-BASED PHRASEOLOGY RESEARCH

I shall now explain the rationale for the Google-based phraseology identification procedure. A search engine like Google allows the user to query a single word (*moon*), a group of words without particular order (*free online dictionary*) or an exact phrase ("*protests at Heathrow Airport*"). If we start from a single word and stick to the exact phrase method and increase the number of words in subsequent searches, the numbers of resultant hits will decrease at a certain rate, depending on how close the relationship is between the last-added word and the ongoing fragment. For example, we can expect the addition of *University* to *Edinburgh*, i.e. "*Edinburgh University"* to generate more Google hits than the addition of *market*, i.e. "*Edinburgh market"*, because *Edinburgh University* is presumably a much more familiar and frequently used phrase than *Edinburgh market* is. That is, the difference between the frequency of *Edinburgh* (52,200,000) and that of "*Edinburgh market"* (806) is much greater than that between *Edinburgh* (52,200,000) and "*Edinburgh University"* (1,370,000). Similarly, if we keep adding words to *Edinburgh University*, such as "*Edinburgh University Press"* (711,000) or *Edinburgh University branch* (90), then we will be able to see how strong the phraseological property of the sequence is by the extent of the frequency gap created by the new addition. To make the observation easier, we can apply a smoothing procedure to the frequencies and draw a 'frequency descending' chart that is neither too dramatic nor too flat and unrevealing. I have found log base 2 to be a good measure which seems to produce results corresponding well to human intuitions about idiomaticity, as Figure 1 and Figure 2 show.

Figure 1 shows a moderate dropping in frequency from *Edinburgh* to *University*, which is only to be expected as there are plenty of possibilities after *Edinburgh*. However, when *Edinburgh University* is in order, the addition of *Press* does not cause the logarithmic line to slop down too much—in fact, it seems to stop dropping at any perceptible ratio and remains almost horizontal. I propose to take this horizontalness as a rough measure of seeing the word sequence, in this case, *Edinburgh University Press*, as an acceptable phraseological unit. Compare Figure 1 with Figure 2, where the addition of the word *branch* causes the frequency line to plummet—at a rate much greater than the previous drop. This is a good indication that the sequence *Edinburgh University branch* as a whole does not constitute a phraseological unit.

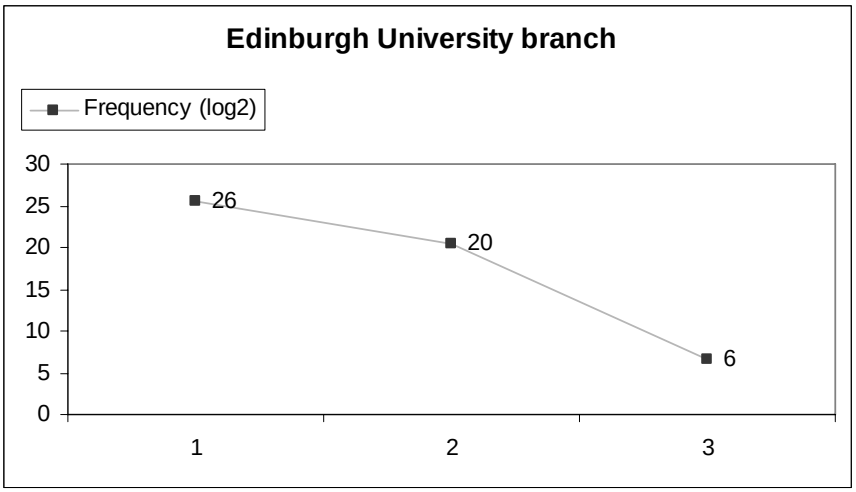


No. of Words	Word Cluster	Google Hits	Log_2
1	Edinburgh	52,200,000	26
2	Edinburgh University	1,370,000	20
3	Edinburgh University Press	711,000	19

Figure 1. *Edinburgh University Press* Frequency Descending Chart with Table

Let us look at some more examples to appreciate how logarithmic frequency lines could be a good indication of phraseology. We will make frequency descending charts for three word clusters: a known idiom and two possible formulaic sequences.

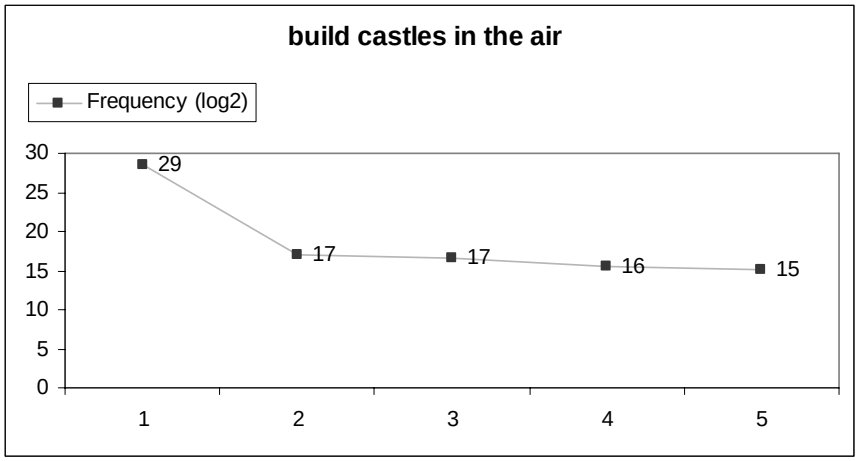
- Idiom: *build castles in the air*
- Possible phrase: *build your own social network*
- Possible phrase: *build cars with better mileage*



No. of Words	Word Cluster	Google Hits	Log_2
1	Edinburgh	52,200,000	26
2	Edinburgh University	1,370,000	20
3	Edinburgh University branch	90	6

Figure 2. *Edinburgh University branch* Frequency Descending Chart with Table

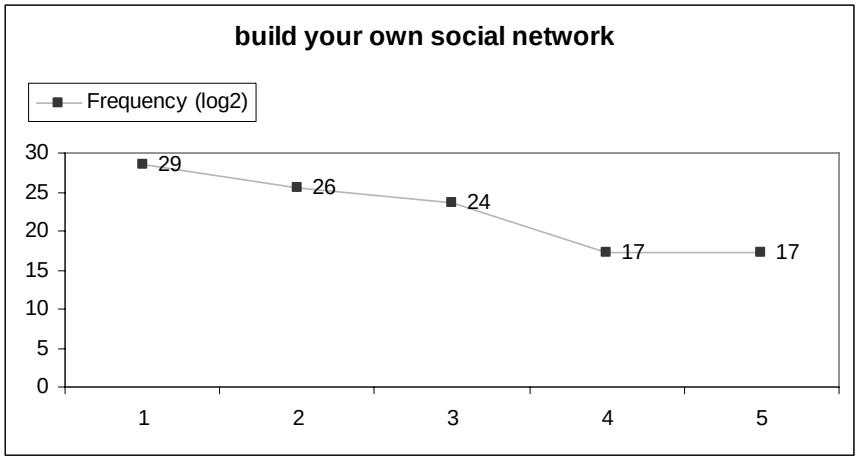
Figure 3 presents a graphic analysis of the frequency decreasing rate for the known idiom *build castles in the air*. It displays a fantastically horizontal line from the second word (*castles*) onward, which shows how tightly 'glued' to each other these four words (*castles in the air*) are when they follow the word *build* to form a phraseological unit. Although understandably the number of Google hits drops considerably from *build* to "*build castles*", since there are too many possible words after *build*, yet when *build castles* is selected, the frequency counts very much stick on a logarithmic scale when the following three words are subsequently added, thereby creating the horizontal line. Thus, after applying logarithmic smoothing to decreasing frequency counts of a word sequence, the web as corpus seems to unveil the phraseology level of organization, which simply cannot be detected by the naked eye.



No. of Words	Word Cluster	Google Hits	Log_2
1	build	380,000,000	29
2	build castles	139,000	17
3	build castles in	101,000	17
4	build castles in the	50,800	16
5	build castles in the air	36,400	15

Figure 3. *build castles in the air* Frequency Descending Chart with Table

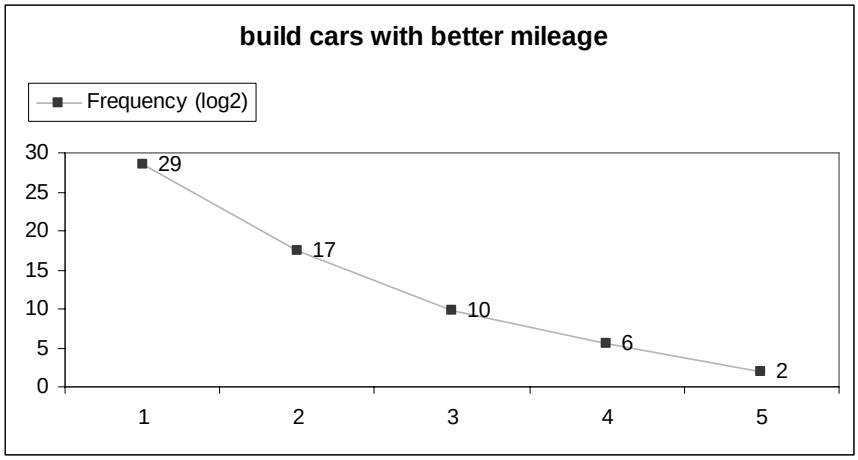
The horizontal line we saw in Figure 3 is partially repeated in Figure 4. The phrase *build your own social network* actually creates two smaller sections of horizontal lines, corresponding to *build your own* and *social network*. While these two groups of words show very strong internal bonds, there is nevertheless a sharp fall in frequency from the first group to the second, signifying there are plenty of other choices after *build your own*, for example, *build your own home*, *build your own PC*. That being said, the actual frequency (158,000) of the entire phrase *build your own social network* cannot be overlooked, either. A closer examination of the Google hits shows that this particular variant is mostly used in advertisement of web technology software.



No. of Words	Word Cluster	Google Hits	Log_2
1	build	380,000,000	29
2	build your	47,900,000	26
3	build your own	13,300,000	24
4	build your own social	164,000	17
5	build your own social network	158,000	17

Figure 4. *build your own social network* Frequency Descending Chart with Table

Finally, the third phrase *build cars with better mileage* does not generates any section of horizontal line, and each section drops to the next at almost the same angle, signifying there being no close bond whatsoever between any two of the component words. This word sequence is thus a good example of 'free combination' in that its formation is rarely repeated on the web as currently indexed by Google.



No. of Words	Word Cluster	Google Hits	Log_2
1	build	380,000,000	29
2	build cars	182,000	17
3	build cars with	837	10
4	build cars with better	49	6
5	build cars with better mileage	4	2

Figure 5. *build cars with better mileage* Frequency Descending Chart with Table

I have introduced a method to identify possible phraseological units (or free combinations) on the web based on the search results of Google. This method could be applied to linguistic research, second language acquisition research, and language teaching, especially the TESOL profession. The phraseological units identified through this route can serve as starting points for further research and language learning, by going into the actual documents and retrieve the context of the linguistic construct. Admittedly, as Wray (2002) cogently remarked, frequency cannot be the only criterion for judging idiomaticity. However, there have been several recent studies of phraseology based solely on frequency, for example, Biber's well-known research on lexical bundles (Biber, Johansson, Leech, Conrad, & Finegan, 1999; Biber et al., 2004). Arguably, the web as a large corpus does offer a comprehensive range of observable phraseological units as well as more substantial frequency counts for meaningful generalizations. Web search can always act as an

effective first line of offence, ensued by other criteria such as semantic preference, colligational formation, and pragmatic environment which together should firmly establish the phraseological status of the unit.

NATIVENESS AND PHRASEOLOGY

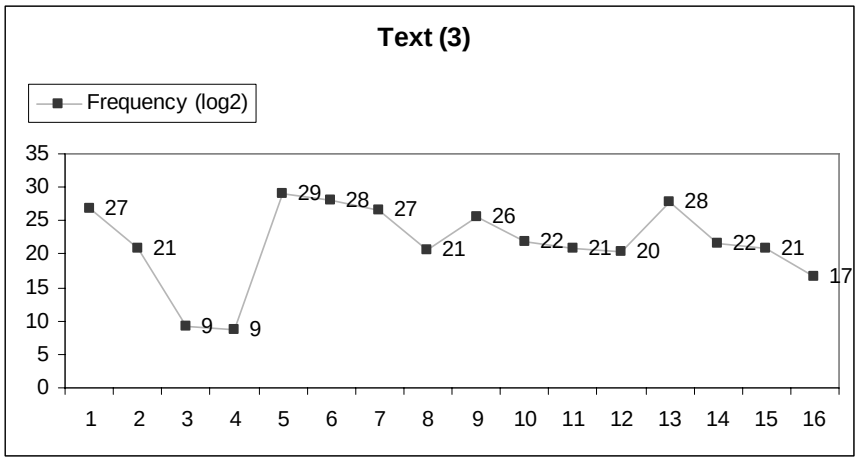
Numerous researches have pointed out the important function of phraseology to differentiate between native speaker (NS) and non-native speaker (NNS) performances (Pawley & Syder, 1983; Pawley, 2007; Simpson & Mendis, 2003). It has not been easy, however, to see this widely acknowledged gap in a somewhat clearer qualitative or quantitative manner. In this section, I will show how the methodology based on web frequency explained above can bring the NS-NNS differences in phraseological performance to the spotlight. Let us examine the BBC text in (1) again here expanded as (3) below.

(3)
Caffeine may help older women ward off mental decline, research suggests…. The study, published in the journal Neurology, raises the possibility that caffeine may even protect against the development of dementia. (BBC News, 2007)

We can create a 'phraseology profile' for the text by slicing it into several fragments. For ease of comparison with NNS text, we will focus on the verb phrase in our texts only, taking any verb from the text and three ensuing words which mostly (but not all) belong to the same VP structure. This creates the set of fragments in (4) from text (3).

(4)
a. ward off mental decline
b. published in the journal
c. raises the possibility that
d. protect against the development

We can now build a VP-based phraseology profile for text (3) in a consecutive frequency fluctuation chart, as Figure 6 shows.



No. of Words	Word Cluster	Google Hits	Log_2
1	ward	117,000,000	27
2	ward off	1,910,000	21
3	ward off mental	568	9
4	ward off mental decline	443	9
5	published	555,000,000	29
6	published in	281,000,000	28
7	published in the	96,600,000	27
8	published in the journal	1,720,000	21
9	raises	53,600,000	26
10	raises the	4,030,000	22
11	raises the possibility	1,760,000	21
12	raises the possibility that	1,240,000	20
13	protect	230,000,000	28
14	protect against	3,240,000	22
15	protect against the	1,890,000	21
16	protect against the development	97,400	17

Figure 6. Frequency Fluctuation Chart for Text (3) with Table

The frequency line overall is surprisingly flat, the only noticeable gap being that created by the fragment of the first VP, *ward off mental decline*, which touches upon a less popular sub-topic. This means that up to the fourth word in each VP construct, the native speaker's word

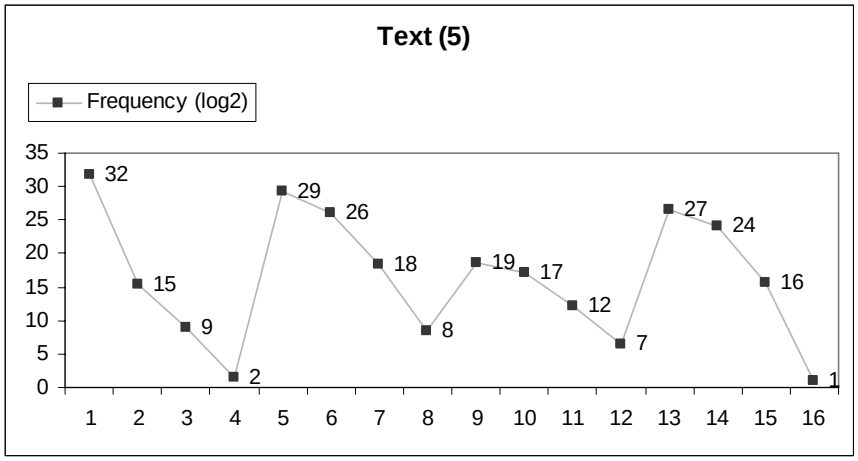
sequences show very good resistance to frequency decline. Let us now compare this native speaker text with one retrieved from a non-native English web site.

(5)
When patients are uncooperative in taking drugs and seldom tell the doctor when revisiting, this will make the doctor misjudge the effect of the drug and cannot adjust the dosage correctly. (Taipei Gov, 2007)

In order to make the two frequency fluctuation lines comparable in terms of the width of the chart and the number of words, only four verb phrases are taken from (5) to form a set of four fragments, each consisting of four words, as shown in (6) below, making a total of 20 words, which is exactly the same as fragments extracted from text (3) in terms of word count.

(6)
a. are uncooperative in taking
b. tell the doctor when
c. misjudge the effect of
d. adjust the dosage correctly

As Figure 7 shows, the four sections of the logarithmic frequency line are generally much steeper than those in Figure 6, creating large gaps between the sections, showing that the frequency of the overall fragment consistently drops sharply with the addition of a new word. This can only mean that the choice the non-native writer makes at each stage does not really conform to the native-speaker phraseological convention. Also, all four sections of the frequency line ultimately sink to the bottom or nearly the bottom, of the chart, meaning the word sequences taken from NNS text have extremely low frequencies when taken as a whole.

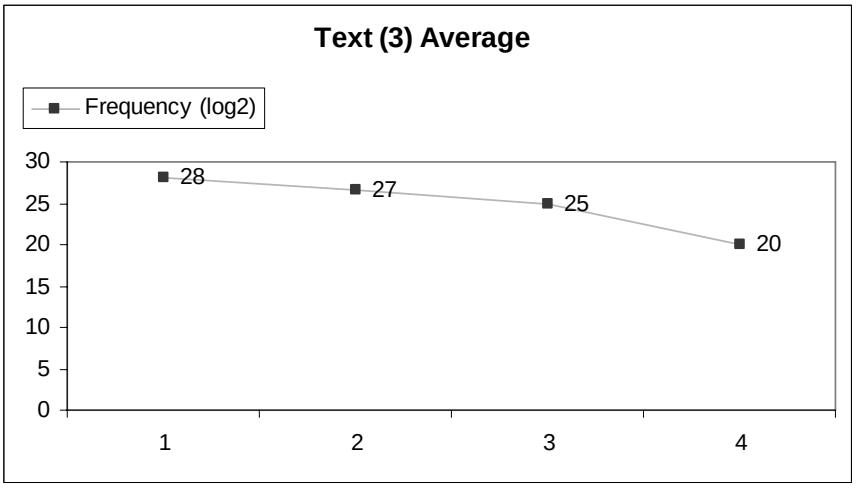


No. of Words	Word Cluster	Google Hits	Log_2
1	are	3,570,000,000	32
2	are uncooperative	41,700	15
3	are uncooperative in	496	9
4	are uncooperative in taking	3	2
5	tell	610,000,000	29
6	tell the	68,200,000	26
7	tell the doctor	327,000	18
8	tell the doctor when	328	8
9	misjudge	419,000	19
10	misjudge the	143,000	17
11	misjudge the effect	4,290	12
12	misjudge the effect of	94	7
13	adjust	96,800,000	27
14	adjust the	16,800,000	24
15	adjust the dosage	50,500	16
16	adjust the dosage correctly	2	1

Figure 7. Frequency Fluctuation Chart for Text (5) with Table

Another thing we can do with these sets of web-based frequency data is to conflate the frequencies for the initial four-word fragments from each text into a general frequency ascending line and compare the two lines generated by the NS and the NNS text respectively. This produces two frequency ascending charts for text (3) and text (5) in Figure 8 and

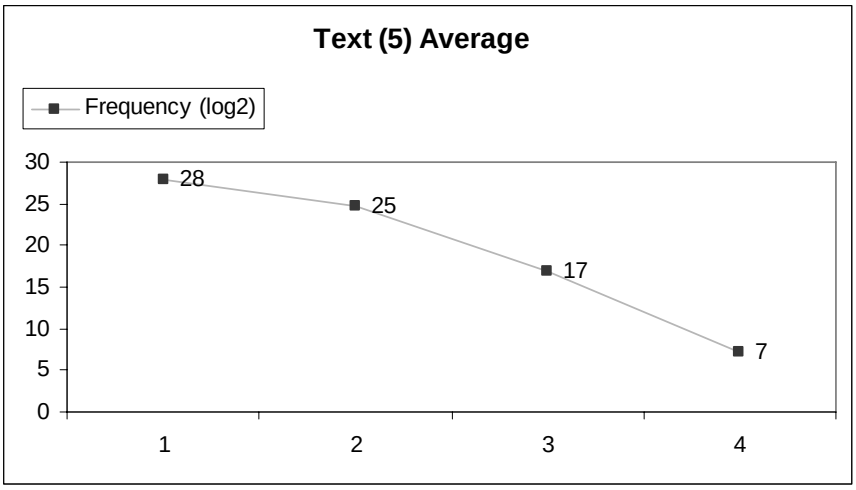
Figure 9 respectively.



	1 Word	2 Words	3 Words	4 Words
ward off mental decline	117,000,000	1,910,000	568	443
published in the journal	555,000,000	281,000,000	96,600,000	1,720,000
raises the possibility that	53,600,000	4,030,000	1,760,000	1,240,000
protect against the development	230,000,000	3,240,000	1,890,000	97,400
Average Frequency	279,533,333	96,090,000	33,416,667	1,019,133
Log_2	28	27	25	20

Figure 8. Average Four-word VP Cluster Frequency for Text (3)

As can be seen, the overall frequency line in Figure 8 is much more horizontal than that in Figure 9. This provides another prospective to see how native speakers consistently use phraseology recognizable by the English-speaking web communities, while non-native speakers choose grammatical but relatively novel word sequences.



	1 Word	2 Words	3 Words	4 Words
are uncooperative in taking	3,570,000,000	41,700	496	3
tell the doctor when	610,000,000	68,200,000	327,000	328
misjudge the effect of	419,000	143,000	4,290	94
adjust the dosage correctly	96,800,000	16,800,000	50,500	2
Average Frequency	235,739,667	28,381,000	127,263	141
Log_2	28	25	17	7

Figure 9. Average Four-word VP Cluster Frequency for Text (5)

PEDAGOGICAL IMPLICATIONS

The place of phraseology in the research and teaching of language has never been higher. There is a growing feeling that some properties and functions previously attributed to grammar and the lexicon can be reassigned to phraseology. Existing speech production models based on the lexicon and grammar (Levelt, 1989; Levelt, Roelofs, & Meyer, 1999) is implicitly and increasingly being challenged by theories that view

phraseology as the frontrunner in language production. While summarizing Sinclair's widely acknowledged open choice principle, O'Keeffe, McCarthy, and Carter (2007) say: "Syntax… far from being primary, is only brought into service occasionally, as a kind of 'glue' to cement the lexical chunks together" (p. 60). Wray (2002) notices that by using prefabricated material, the speaker can reduce his or her processing load. Cowie (1992) comments that any speech or writing is hardly acceptable to native speakers "without controlling an appropriate range of multiword units" (p. 10). It is no longer sustainable to insist that we map concepts to individual lexical items and generate a sentence based on the grammatical properties of these items when we speak (Bock & Levelt, 1994). Instead, words are found to be happy in particular patterns and phraseologies, which evoke meanings different from or in addition to the sum of individual words. We do not create novel sentences every time we speak. Instead, we might be 'thinking in phraseology'.

The relationship between prefabricated patterns and social conventions and discourse functions has long been acknowledged by the linguistic and TESOL circles. Biber et al. (2004), for example, suggest that lexical bundles like *well I don't know* construct a frame in discourse to express stance, discourse organization, or referential status. Stubbs (2002), Hoey (2005), Hunston and Fransis (2000) and other works seem to point towards a conceptual framework where lexical items in their particular phraseological environment or patterns are at the centre of language production and comprehension, intimately related to the component of meaning. Thus a direct connection between text, phraseology, social interaction, conceptual representation and interpretation seems an inevitable conclusion. As Sinclair puts it:

> We understand text by relating the phrasings to our stored experience of recurrent meaningful patterns, and interpreting those that vary from what is stored. (Sinclair, 2004, p. 288)

If it is true that we think and speak in chunks of language, then it is only logical that we teach all kinds of phraseology as a top priority, in a comprehensive and profound manner. Despite the general success of communicative language teaching, however, classroom practice on phraseology to date still remains at the level of mapping fixed expressions to social functions. Any attentions paid to explicit instruction are normally directed to the two traditional components of grammar and

vocabulary. No complicated view that I am aware of is offered to the learner regarding the kind of extended collocation discussed in the previous section, for example, and their rich semantic and pragmatic implications. Jones and Haywood (2004) are alarmingly correct when they say:

> In spite of the increasing interest in and knowledge about phraseological development amongst L1 and L2 speakers, little progress has been made when it comes to applying the new insights to the EFL classroom. (p. 271)

Jones and Haywood (2004) also usefully remind us: "If coursebooks fail to give due attention to the teaching of formulaic sequences in academic discourse, then it is up to the teacher to do so" (p. 271). Teacher-directed discovery learning is in fact a dominant mode in corpus-based or so-called data-driven learning, in which "the computer is used as a special type of informant and the students are led through a process of self-discovery" (Tognini-Bonelli, 2001, pp. 43-44). Johns (1991) also points out that "the task of the language teacher is to provide a context in which the learner can develop strategies for discovery" (p. 1). However, despite all the talk about the importance of corpora to language learning, there is no well-known concordancing software specifically designed for this purpose to date. Popular concordancers like WordSmith, Sara (BNC), COBUILD Collocation Sampler and so on, are linguistics research tools and are not equipped with any didactic feature. Moreover, as noted by Guo and Zhang (2007), these kinds of software are expensive to buy or subscribe to. It seems welcoming news then, that a web search engine like Google, which is free and familiar to any Internet user, can be useful for language teaching and learning, if only for verifying the phraseological status of a word sequence as demonstrated in the previous section. Other examples of using Google for language learning purposes are available, for example, Mills and Salzmann (1995) for finding instances of given grammatical structures, and Robb (2003) for comparing the usage of phrases from different web domains, and so on.

Google has been used in this study as a vehicle to show how phraseology research can benefit from the web as corpus, especially regarding the identification of phraseological units via frequency-based profiling of word sequences. Admittedly, Google is not specifically

designed for academic purposes. In fact, researchers like Leech (2007) have criticized the use of Google by attacking the reliability of its frequency counts and the eligibility of the web to represent English language use. On the other hand, more and more researchers find the web being the only corpus large enough to provide adequate instances of the linguistic phenomena they are investigating (Keller & Lapata, 2003; Rigau, Magnini, Agirre, Vossen, & Carroll, 2002; Rosenbach, 2007; see also Hundt, Nesselhauf, & Biewer, 2007), and research has confirmed that web-based frequencies correlate well with frequencies obtained from well-designed large corpora (Keller & Lapata, 2003; Mair, 2007; Rohdenburg, 2007). Many researchers remark on the merits of using the web as corpus, of which Fletcher's (2007, p. 27) list seems representative:

- Freshness and spontaneity
- Scope and completeness
- Linguistic diversity
- Cost and convenience
- Representativeness

Thus, the web as corpus constantly offers renewed language material without our maintenance, includes many genres and domains of interest and rarer expressions, covers many languages and language varieties, is free and convenient to access, and is representative of what people speak and write in the modern information age.

Researchers have pointed out various ways of using the web as corpus. For example, Hundt et al. (2007) distinguish between the "Web as corpus" and the "Web for corpus building" approach (p. 2). Lüdeling, Evert, and Baroni (2007) further distinguish two methods from the first category: using the commercial engine directly, or pre-processing the query or post-processing the result. As for the second category, they also distinguish between automatic compilation of corpus by using "web crawler" programs and manually downloading web pages. For language learning purposes, the first category seems to be most convenient and practical; that is, using the commercial search engine such as Google directly, perhaps supplemented by advanced search skills such as using the wildcard. The key idea for the acquisition of formulaic sequences is, according to Jones and Haywood (2004), "repeated exposure and discussion", which enables the students to notice and retrieve the phraseological units in question (p. 290). Thus, the most convenient

approach to Google-based language learning seems for the teacher to design activities for students to make self-discoveries, and to hold discussion sessions afterward to facilitate the retrieval of the phraseological units they have noticed from web searching. It is also possible to design separate programs which process Google searching at the background and present the result of the search in a separate interface. Guo and Zhang (2007) take this approach and use Google as an engine to extract English collocations from the web for language learning. The Google results are presented in concordance lines rather than in the original document extract formats for easy scrutiny. Similar interface can be developed for extracting phraseological units from the web using the frequency calculation method introduced in this article and the application programming interfaces (APIs—see Fletcher, 2007, p. 32).

Phraseology is obviously not the only component of language teaching which can be explored by commercial search engines and the web. I have mentioned the example of hunting for grammatical structures using Google—the Grammar Safari project (Mills & Salzmann, 1995). Some research boarders on the intersection between grammar and phraseology. Rosenbach (2007), for example, compares the N+N structure with the *s*-genitives from both the web and the traditional corpora and finds animate modifiers to prefer the *s*-genitive (*lawyer's fees*), while the inanimate modifiers favour the N+N construction (*car engine*). Rosenbach uses the key phrase *grammatical variation* in her abstract, but arguably, the issue could also be discussed under the rubric of phraseology, depending on which theoretical convention one is evoking. As Hoey (2005) cogently argues, grammar is the result of lexical priming. Like Stubbs (2002), Hoey holds that lexical units decide not only on the choice of neighboring words, but also their colligational, semantic and pragmatic associations. Like Hunston and Francis (2000), Hoey thinks words have their preferred structures or patterns. This means words, phrases, and grammar all work together in an integrated fashion to produce language. Hundt et al. (2007) usefully point out: "For a lot of interesting research questions, carefully compiled corpora offer either very limited information or no information at all" (p. 2). I have noted in this article how recognized large corpora fail to produce any examples of extended collocations, and that the web may be the only corpus large enough to offer adequate instances of phraseology for fruitful linguistic enquiries and their applications to language teaching. This could also be true with grammar and other levels of linguistic

analysis. It is time for the TESOL profession to start drawing a blueprint for exploiting the web as a corpus for extracting examples and generalizations of language usage.

As a first step towards utilizing Google for classroom instruction, Chen (2007) conducted a teaching experiment. She told 20 Chinese students in Taiwan to write an English composition each, and analyzed their errors in terms of word choice, collocation, and grammar. The compositions were also graded by two different tutors. She then compiled a set of Google search strategies designed to help students find correct words, collocations, and grammatical patterns. Students were told to write a second composition each using those Google strategies. Post-analysis shows errors are fewer and marks for compositions are higher. Also, the result of a questionnaire investigation shows students are in favour of those Google search strategies provided in their composition processes. By the same token, Google strategies can be designed by language teachers whether they are teaching words or grammar, speaking or writing, or stylistics and literature. Google search offers infinite possibilities for teachers to integrate self-studies into their curricula and it remains a constant companion to the learner in the absence of the tutor. All the TESOL teacher has to do is to show the learner how to use this versatile tool.

CONCLUSION

I have set out in this article to prove the existence of a complex kind of phraseological unit which I call extended collocation. In doing so, I introduced a web-based frequency verification method for objective evaluation of phraseological units on the web. Researchers have noticed that numerous formulaic expressions known to native speakers simply cannot be found "even in the mega-corpora" (Read & Nation, 2004, p. 32). Arguably, the web is the only corpus large enough to provide adequate instances for observation and computation of a fuller range of phraseology. I have shown how this can be done by a simple frequency-based computational method in this article.

I have also argued briefly about the importance of phraseology to language processing, especially in terms of its connections to the conceptual and social interactional domains, and therefore its overall importance to second language learning. Designing class sessions using a search engine like Google to help students notice and retrieve phraseology

on the web, I suggest, is a good way to sharpen their social-conventional awareness and establish their conceptual-phraseological mappings in English as a second language. This link has to be set up as a matter of urgency by the TESOL profession.

The use of the web as corpus and its application to language teaching no doubt causes some concerns and debates. The key problems identified for this approach seem to be: the absence of annotation, misleading frequency and doubtful representativeness (Leech, 2007), authorship issue (Thelwall, 2005), and the problem of reproducibility (Lüdeling et al., 2007). However, these alleged problems can be overcome by pre-processing the queries, post-processing the results when using a commercial search engine (Guo & Zhang, 2007; Lüdeling et al., 2007), by using different search engines and repeating searches periodically (Fletcher, 2007), by designing linguistic search engines (Fairon 2000; Fletcher, 2007; Renouf et al., 2005; Renouf, Kehoe, & Banerjee, 2007), or by building own web-based corpora (Thelwall, 2005). Renowned scholar such as Sinclair actually considered untagged corpora more suitable for data-driven research (Sinclair, 2004). The bottom line is: the web has become arguably the most important discourse community for the next generation. English on the web may contain over-represented varieties (blogs, wikis, academic and commercial web pages), machine-generated materials, ill-formed language or texts written by non-native speakers, but if this is where the future English is heading, then it needs to be accepted and understood. I have introduced a way of understanding an important aspect of language on the web—the English phraseology.

Although the phrase-identification procedure introduced in this article is reasonably reliable and useful, it lacks an accurate definition (in numerical terms) as to what kind of frequency profile 'counts as' a phraseological unit. Future research by computational linguists can perhaps bridge this gap by developing more accurate rationales for automatically and unambiguously identifying a fragment as a phraseological unit. Also, the methodology introduced in this article can deal with a series of consecutive but not discontinuous items. Other methods need to be developed for identifying a phrasal structure which is not continuous (e.g. *killed…after being* in *A school student has become the 17th teenager to be killed in London this year after being fatally stabbed*). Finally, in terms of TESOL pedagogy, this article did not consider the details of incorporating Google search procedures into

syllabus, textbook, or lesson plans. A good line of future research is to investigate how Google-based learning can best promote the principles of communicative language teaching, how Google activities can be included in task-based learning, and so on.

REFERENCES

BBC News. (2007). *Coffee 'protects female memory'*. Retrieved August 13, 2007, from http://news.bbc.co.uk/1/hi/health/6930114.stm

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at…: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, *25*(3), 371-405.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English.* Harlow, UK: Longman.

Bock, J. K., & Levelt, W. J. M. (1994). Language production: Grammatical encoding. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 945-984). Orlando, FL: Academic Press.

Chen, Y-S. (2007). *A New kind of CALL: Using GOOGLE as a tool for EFL learning*. Unpublished master's thesis, Swansea University, Swansea, UK.

Cowie, A. (1992). Multi-word lexical units and communicative language teaching. In P. Arnaud & H. Béjoint (Eds.), *Vocabulary and applied linguistics* (pp. 1-12). London: Macmillan.

Eggins, S., & Martin, J. R. (1997). Genres and registers of discourse. In T. A. van Dijk (Ed.), *Discourse as structure and process* (pp. 230-256). London: Sage.

Fairon, C. (2000). GlossaNet: Parsing a web site as a corpus. *Linguisticae Investigationes*, *22*(2), 327-340.

Fletcher, W. H. (2007). Concordancing the web: Promise and problems, tools and techniques. In M. Hundt, N. Nesselhauf, & C. Biewer (Eds.), *Corpus linguistics and the Web* (pp. 25-45). Amsterdam: Rodopi.

Guo, S., & Zhang, G. (2007). Building a customised Google-based collocation collector to enhance language learning. *British Journal of Educational Technology*, *38*(4), 747-750.

Hill, J. (2000). Revising priorities: From grammatical failure to collocational success. In M. Lewis (Ed.), *Teaching collocation* (pp. 47-70). Hove, England: Language Teaching Publications.

Hoey, M. (2005). *Lexical priming: A new theory of words and language*. London/New York: Routledge.

Hundt, M., Nesselhauf, N., & Biewer, C. (2007). Corpus linguistics and the web. In M. Hundt, N. Nesselhauf, & C. Biewer (Eds.), *Corpus linguistics and the Web* (pp. 1-5). Amsterdam: Rodopi.

Hunston, S., & Francis, G. (2000). *Pattern grammar—A corpus-driven approach to the lexical grammar of English*. Amsterdam/Philadelphia: John Benjamins.

Johns, T. (1991). Should you be persuaded—Two samples of data-driven learning materials. *English Language Research Journal*, *4*, 1-13.

Jones, M., & Haywood, S. (2004). Facilitating the acquisition of formulaic sequences. In N. Schmitt (Ed.), *Formulaic sequences*: *Acquisition, processing and use* (pp. 269-300). Amsterdam: John Benjamins.

Keller, F., & Lapata, M. (2003). Using the Web to obtain frequencies for unseen bigrams. *Computational Linguistics*, *29*(3), 459-484.

Leech, G. (2007). New resources, or just better old ones? The Holy Grail of representativeness. In M. Hundt, N. Nesselhauf, & C. Biewer (Eds.), *Corpus linguistics and the Web* (pp. 133-149). Amsterdam: Rodopi.

Levelt, W. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*, 1-75.

Lüdeling, A., Evert, S., & Baroni, M. (2007). Using Web data for linguistic purposes. In M. Hundt, N. Nesselhauf, & C. Biewer (Eds.), *Corpus linguistics and the Web* (pp. 7-24). Amsterdam: Rodopi.

Mair, C. (2007). Change and Variation in Present-Day English: Integrating the Analysis of Closed Corpora and Web-Based Monitoring. In M. Hundt, N. Nesselhauf, & C. Biewer (Eds.), *Corpus linguistics and the Web* (pp. 233-248). Amsterdam: Rodopi.

Mills, D., & Salzmann, A. (1995). Grammar safari. Retrieved August 28, 2007, from http://www.iei.uiuc.edu/student_grammarsafari.html

O'Keeffe, A., McCarthy, M., & Carter, R. (2007). *From corpus to classroom: Language use and language teaching*. Cambridge: Cambridge University Press.

Pawley, A. (2007). Developments in the study of formulaic language since 1970: A personal view. In P. Skandera (Ed.), *Phraseology and culture in English* (pp. 3-45). Berlin: Mouton de Gruyter.

Pawley, A., & Syder, F. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. Richards, & R. Schmidt (Eds.), *Language and communication* (pp. 191-226). London: Longman.

Read, J., & Nation, I. S. P. (2004). Measurement of formulaic sequences. In N. Schmitt (Ed.), *Formulaic sequences: Acquisition, processing and use* (pp. 23-35). Amsterdam: John Benjamins.

Renouf, A., Kehoe, A., & Banerjee, J. (2005). The WebCorp search engine: A holistic approach to web text search. In *Proceedings of the Corpus Linguistics 2005 Conference*. Birmingham, UK: University of Birmingham. Retrieved October 30, 2008, from http://www.corpus.bham.ac.uk/PCLC/cl2005-SE-pap-final-050705.doc

Renouf, A., Kehoe, A., & Banerjee, J. (2007). WebCorp: An integrated system for web text search. In M. Hundt, N. Nesselhauf, & C. Biewer (Eds.), *Corpus linguistics and the Web* (pp. 47-68). Amsterdam: Rodopi.

Rigau, G., Magnini, B., Agirre, E., Vossen, P., & Carroll, J. (2002). MEANING: A roadmap to knowledge technologies. In *Coling-02 on A Roadmap For Computational Linguistics: Vol. 13 International Conference On Computational Linguistics* (pp.1-7). Morristown, NJ: Association for Computational Linguistics.

Robb, T. (2003). Google as a quick 'n dirty corpus tool. *TESL-EJ*, *7*(2). Retrieved August 29, 2007, from http://www-writing.berkeley.edu/TESL-EJ/ej26/int.html

Rohdenburg, G. (2007). Determinants of grammatical variation in English and the formation/confirmation of linguistic hypotheses by means of Internet data. In M. Hundt, N. Nesselhauf, & C. Biewer (Eds.), *Corpus linguistics and the Web* (pp. 191-209). Amsterdam: Rodopi.

Rosenbach, A. (2007). Exploring constructions on the web: a case study. In M. Hundt, N.

Nesselhauf, & C. Biewer (Eds.), *Corpus linguistics and the Web* (pp. 167-190). Amsterdam: Rodopi.
Schmitt, N., & Underwood, G. (2004). Exploring the processing of formulaic sequences through a self-paced reading task. In N. Schmitt (Ed.), *Formulaic sequences: Acquisition, processing and use* (pp. 173-268). Amsterdam: John Benjamins.
Schmitt, N., Grandage, S., & Adolphs, S. (2004). Are corpus-derived recurrent clusters psycholinguistically valid?. In N. Schmitt (Ed.), *Formulaic sequences: Acquisition, processing and use* (pp. 127-151). Amsterdam: John Benjamins.
Scott, M., & Tribble, C. (2006). *Textual patterns: Key words and corpus analysis in language education*. Amsterdam: John Benjamins.
Simpson, R., & Mendis, D. (2003). A corpus-based study of idioms in academic speech. *TESOL Quarterly*, *37*(3), 419-441.
Sinclair, J. (2004). *Trust the text: Language corpus and discourse*. London: Routledge.
Stubbs, M. (2002). *Words and phrases*. Oxford: Blackwells.
Taipei Gov. (2007). *Health education on sleeping pills*. Retrieved August 18, 2007, from http://english.taipei.gov.tw/docms/index.jsp?categid=641&recordid=9882
Thelwall, M. (2005). Creating and using web corpora. *International Journal of Corpus Linguistics*, *10*(4), 517-541.
Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. Amsterdam/Philadelphia: John Benjamins.
Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

CORRESPONDENCE

Chris Shei, School of Arts, Swansea University, UK
E-mail address: C-C.Shei@swansea.ac.uk