Concgramming: A computer driven approach to learning the phraseology of English

Chris Greaves and Martin Warren

ReCALL / Volume 19 / Issue 03 / September 2007, pp 287 - 306
DOI: 10.1017/S0958344007000432, Published online: 24 August 2007

Link to this article: http://journals.cambridge.org/abstract_S0958344007000432

How to cite this article:
Chris Greaves and Martin Warren (2007). Concgramming: A computer driven approach to learning the phraseology of English. ReCALL, 19, pp 287-306 doi:10.1017/S0958344007000432

doi:10.1017/S0958344007000432 Printed in the United Kingdom

Concgramming:
A computer driven approach to learning the phraseology of English

CHRIS GREAVES AND MARTIN WARREN
English Department, The Hong Kong Polytechnic University
(*email: ecchgr@yahoo.com,egwarren@polyu.edu.hk*)

Abstract

This study introduces a new computer-based methodology, 'concgramming', that has as its primary aim the automatic identification of the phraseological profile and hence the 'aboutness', of a text or corpus. It is argued that this methodology can be employed by language learners and teachers to raise awareness of the importance of the phraseological tendency in language. The methodology is outlined, and examples of its potential for use by language learners in a data-driven learning mode are described. The wider implications of concgramming, and the concgrams so generated, are also discussed with regard to CALL.

Keywords: aboutness, concgramming, concgrams, data-driven language learning, phraseological profile, phraseology

1 Introduction

Over the last twenty years there has been an increasing awareness of the importance of phraseology in English language description and researchers in the field have produced significant results. Here the term 'phraseology' is used broadly and refers to what Clear (1993: 277) terms "the recurrent co-occurrence of words". In other words, the phraseological character of natural language refers to the more-or-less fixed co-occurrence of linguistic elements (Hunston, 1995). Corpus linguists examining co-occurrences found in linguistic patterns have contributed to our understanding of, for example, pattern grammar (see, for example, Hunston & Francis, 2000), phraseology (see, for example, Sinclair, 1987; Sinclair, 1996; Sinclair, 2004a; Stubbs, 2001; Tognini-Bonelli, 2001) and semantic prosody (see, for example, Louw, 1993; Sinclair, 1991).

We now better understand that when we speak and write, on most occasions, we select words in combination. This is termed the 'the idiom principle' (Sinclair, 2004a: 29), i.e.

the phraseological tendency, whereby words are co-selected rather than being selected separately constrained only by grammar. The process of co-selection involves a combination of five categories of co-selection (Sinclair, 1996), i.e. semantic prosody, semantic preference, colligation, collocation and the invariant core word(s), and the outcome of these co-selections is a 'lexical item' (Sinclair, 1996). These co-selections are now starting to be given space in new corpus-based grammars of the English language (see, for example, Biber *et al.*, 1999; Carter & McCarthy, 2006), but have yet to be foregrounded. The attention has been on the most frequently occurring contiguous word associations, which are also termed 'clusters' or 'bundles'. Biber *et al.* (1999), for instance, identify the most frequent 'lexical bundles' in their data, and classify them in terms of the structural patterns that they encompass as well as the grammatical category of the end word of a lexical bundle (*ibid*: 996-997). Carter and McCarthy (2006: 504-505) also focus on the structure of 'clusters' along with their functions across different genres.

Phraseology has not only become a new frontier which is of interest to researchers, but it also presents new and exciting challenges for learners and teachers of the English language. For while the importance of phraseology is not contested, currently this is an area which is relatively neglected in the learning and teaching of applied English language studies, English language proficiency, and ESP. Exceptions to this general observation are recent textbooks on phraseology and collocation (see for example, McCarthy, 2005; Sinclair, 2003 and Stubbs, 2002). This paper argues that greater emphasis should be placed on the learning and teaching of phraseology and applies a new computer-mediated research methodology to introduce and promote the learning and teaching of phraseology in CALL. In other words it argues that relatively simple developments in the currently available technology can assist teachers and learners in this important area of language learning and teaching. It supports this proposal by discussing replicable learning and teaching activities which enable learners and teachers to raise their awareness of patterns of phraseology. This study therefore builds on the work of others who have advocated the use of corpora and corpus linguistics in language learning in general (see for example, Aston, 1997; Bernadini, 2002; Braun, 2005; Kennedy & Miceli, 2002 and Sinclair, 2004b) and the use of concordancing in particular (see for example, Bernadini, 2000; Cobb, 1997; Gaskell & Cobb, 2004; Johns, 1991; Sinclair, 2003 and Stevens, 1991).

2 Concgrams

2.1 Background

Uncovering the extent of word associations and how they are manifested in collocations has been an important area of study in corpus linguistics since the 1960s (Sinclair et al., 1970), but finding them all has posed problems in the past. Those working in the fields of natural language processing (NLP), computational linguistics and corpus linguistics are familiar with 'n-grams', sometimes termed 'word clusters', 'lexical clusters' or 'bundles' (see, for example, Biber *et al.*, 1999; Biber, Conrad & Cortes, 2004; Carter & McCarthy, 2006), which are contiguous words that constitute a phrase, or a pattern of use, and that recur in a corpus. Instances of n-grams are in the form of bi-grams, tri-grams, and so on, depending on the number of words in the phrase. Current searches for n-grams generate phrases made up of contiguous word associations, such as 'different people', but would miss instances of the same phraseological pattern when it is realised in instances such as 'different kinds of people' or 'different types of people'. In other words, n-gram searches are only helpful in finding instances of collocation that are strictly contiguous in sequence. The result is that many instances of word associations may be overlooked, and phrases that typically, or on occasion, occur in non-contiguous sequences (i.e. AB, A*B, where '*' represents an intervening word) risk going undiscovered.

These limitations of n-gram searches have led to the recent development of searches for gapped n-grams or 'skipgrams' in NLP (see Wilks, 2005). These skipgrams can include a certain amount of constituency variation (i.e. AB and A*B). Skipgram searches also have their limitations, however. They are currently limited to 3-word skipgrams and four 'skips' (Wilks, 2005), meaning that associated words that are more than four words apart are not found. Skipgram searches have two more limitations: they cannot handle positional variation (i.e. AB, BA), and they are limited with regard to either the size or the kinds of skipgrams found. In addition, many existing searches for non-contiguous word associations may require the input of a formula which can be user-unfriendly.

An example of an automated non-contiguous word association (i.e. skipgram) search is Fletcher's (2006) 'phrase frames' which does not require a user-nominated search query. Phrase frames are based on an initial automated search for n-grams up to eight words long (*ibid*, 2006). Based on these n-grams, another automated search finds phrase-frames which are "sets of variants of an n-gram identical except for one word' (*ibid*, 2006). Phrase frames, then, are a form of skipgram constrained by narrow search parameters, with the result that other non-contiguous associations made up of the same words remain undiscovered if they differ by more than one word, as do patterns with positional variation.

Cheng, Greaves and Warren (2006) describe the contribution that computer-mediated software can make to identifying units of meaning in naturally occurring text. The paper describes a search engine, ConcGram,[1] which is able to extract recurrent concgrams (i.e. sets of between two and five co-occurring words) fully automatically, within a wide span (up to twelve words on either side of the origin),[2] and which include all of a concgram's configurations irrespective of any constituency (e.g. AB and A*B) and positional variation (e.g. AB and BA) present. Cheng *et al.* (2006) argue that the identification of concgrams facilitates a fuller appreciation and understanding of Sinclair's (2004a) idiom principle. This is because concgrams are a useful source of raw data to reveal the co-selections made by the speakers and writers represented in a text or corpus. They are thus a potential starting point for quantifying the extent of phraseology in a text or corpus and hence determining the phraseological profile of the language contained within it. By 'phraseological profile' we mean the identification of the meaningful word associations in a text or corpus.

Phillips (1983, 1989) offers a rationale for the determination of the topic of a text through an objective, quantitative, distributional methodology. What we here are referring to as "phraselogical profile" is linked to what Phillips refers to as the "aboutness" of a text. While 'phraseological profile' refers to the meaningful word associations in a text or corpus, 'aboutness' refers to the meaningful word associations

[1] ConcGram is a search engine written and developed by Chris Greaves, Senior Project Fellow, English Department, The Hong Kong Polytechnic University.

[2] The term 'origin' is used rather than 'node'. The reasons for this distinction are given later in the paper.

that are specific to that text or corpus. Phillips claims that “aboutness” is a product of the global patternings in the text, or what Phillips terms the text's “macrostructure”. Importantly, Phillips argues that one should determine the macrostructure of texts by computational means, to ensure that the results are derived from the text itself and not from external features. The basic assumption of this position is that meanings in language are ultimately constructed by lexical items or the associations of lexical items. This basic assumption also underpins ‘concgramming’ and the activities described later in this study.

2.2 What is a concgram?

A ‘concgram’ is all of the permutations of constituency variation and positional variation generated by the association of two or more words (Cheng *et al*., 2006). As a result, the associated words of a concgram may be the source of a number of ‘collocational patterns’ (Sinclair, 2004a: xxvii). In fact, attempts to identify what we term ‘concgrams’ can be traced back to the 1980s (Sinclair, 2005, personal communication) when the Cobuild team based at the University of Birmingham tried, with limited success, to develop a way to automatically search for non-contiguous sequences of associated words.

The development of a concgram search represents an important shift in perspective from that which underlies the KWIC (i.e. key word in context) display of concordance lines which has long been associated with corpus linguistics. The study of KWIC displays has unintentionally led users to regard the node (i.e. the search word) as the centre of attention and the words associated with the node as being subordinate to it.

Rather than focusing on the node, ConcGram, highlights in colour all of the associated words of a concgram in each concordance line. This then means that the user focuses not on the node, but on the concgram and, therefore, word associations become the focus of attention. It is for this reason that the term ‘origin’ is used for the word or words that form the basis of the automated concgram search in place of ‘node’, in order to underscore the difference between a concgram search and its display and the traditional KWIC display. The automatic mode of the search engine begins with a search for 2-word concgrams, and then builds up iteratively to 5-word concgrams. The concept of a ‘node’ is therefore irrelevant and the notion of ‘origin’ better conveys that associated words are the focus of every concgram search. Currently, the search engine operates with 1-word, 2-word, 3-word or 4-word origins. The necessities of display layout mean that the on-screen view of concgram concordance lines requires a sort-point in order to have a visually intelligible page. In other words, a single origin is centred, or the word to be centred in a multi-word origin can be determined alphabetically by the user, but any word in a concgram can be centred if the user wishes to switch the centred word.

The search engine has been designed to perform fully automated concgram searches, but the user can override the default automatic search function and enter a word or up to five words as a user-nominated concgram search query. When user-nominated searches are performed, the choice of which word is to be in central position in the on-screen display is decided by the user.

The fully automated capability of the search engine, i.e. the absence of any form of prior intervention by the user, makes it a truly ‘corpus-driven’ methodology (Tognini-Bonelli, 2001: 11), and so further increases the likelihood that the concgram searches enable the user to discover not only a more extensive description of known patterns of collocation and their meanings, but also, and more importantly, new phraseological patterns of language use. The potential of the search engine to be used by teachers and learners in CALL is illustrated later in this study. First, the process of the automatic concgram search is described.

2.3 The concgram search

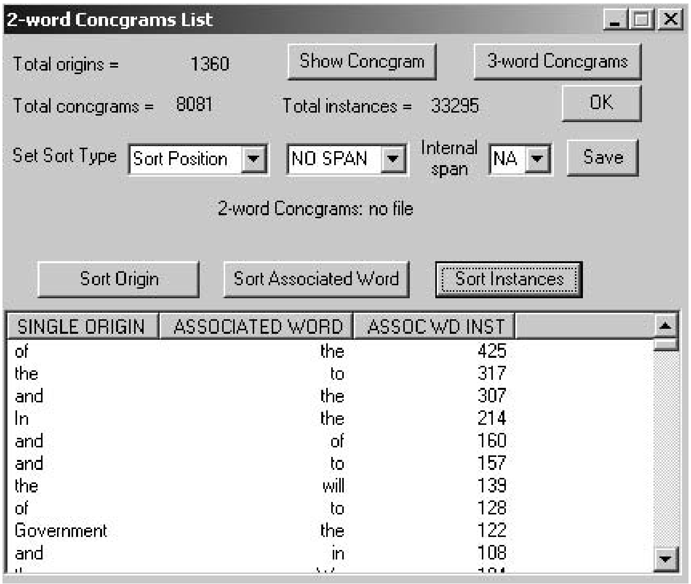
The concgram search identifies all of the co-occurrences of words regardless of their various configurations (i.e. the constituency and positional variations that may be realised within a concgram) in a text or corpus within a given span set by the user. The process of creating the initial 2-word concgram list can be summarised as follows:

- All the unique words (i.e. 'types') in a text or corpus are identified and listed.
- Based on this unique word list, 2-word concgram searches are made with each unique word acting as a single origin in each search.
- All co-occurring words are then listed for each single origin (see Figures 1 and 2).
- Each concgram can be displayed in terms of its configurations which denote either constituency variation, positional variation, or both (see Figures 3 and 4).
- Each concgram may be viewed in its concordance lines (see Figure 5).

Once a 2-word concgram list has been created, the search engine proceeds to fully automatically search for all of the 3-word concgrams, all the way to 5-word concgrams. For example, the 3-word concgrams are found by the search engine performing double-origin searches based on all of the 2-word concgrams found. All of the resulting concgrams can then be viewed either in a list format or in their concordance lines.

The data used in this study are two political speeches. They are both "Policy Addresses" given by the Chief Executive of Hong Kong (the holder of this post is the head of the Government of the Hong Kong Special Administrative Region) in October 2006 and October 2005, in which he outlined the political agenda of the government for the coming twelve months. These speeches are much anticipated in Hong Kong and they are the subject of considerable speculation before they are given and much analysis afterwards. It was therefore decided that these two texts would be of interest to the students in Hong Kong to analyse in terms of their respective phraseological profiles and, from these, their respective aboutness.

In Figure 1, we have the products of a search for 2-word concgrams in a list format based on the frequency of instances in the 2006 Policy Address. On the left are listed the origins, in the middle the co-occurring words, and on the right the frequencies. It can be seen at the top of the list that 'the' is listed as single origin with 'of' as the associated word 430 times and then the next concgram on the list shows them in the reverse order 425 times. The difference in totals, i.e. 5, is a result of where an associated word occurs relative to the search window which is determined by the span set by the user. The concgrams populating the top of the list are all made up of 'grammatical' words and constitute what Sinclair and Renouf (1991: 57) term 'collocational frameworks'. Although these word associations are not the focus of this paper, their prevalence may suggest that they should be given far more attention in English language learning and deserve more time and space in the curriculum.



Total origins = 1360

Total concgrams = 8081 Total instances = 33295

Set Sort Type: Sort Position; NO SPAN; Internal span: NA

2-word Concgrams: no file

SINGLE ORIGIN	ASSOCIATED WORD	ASSOC WD INST
of	the	425
the	to	317
and	the	307
In	the	214
and	of	160
and	to	157
the	will	139
of	to	128
Government	the	122
and	in	108

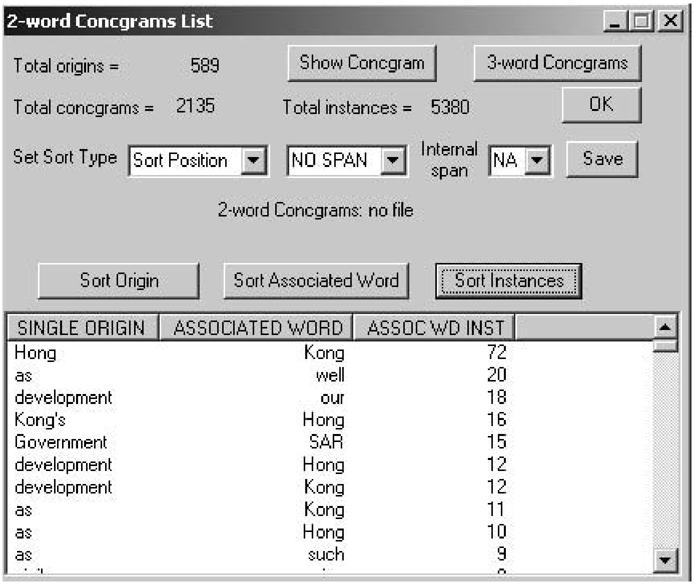
Fig. 1. Two-word concgram list.

If the user is interested in focusing on the more 'lexically-rich' concgrams, it is possible to reduce the length of the list of concgrams by implementing the exclusion list[3] provided, or one of the user's own choice. In Figure 2, we now find the list of concgrams is more manageable for analysing the 'aboutness' (Phillips, 1983 and 1989) of the text or corpus and this is described in more detail later in this study. The exclusion of the fifty most frequent words in the English language means that the more lexically-rich concgrams now populate the top of the frequency lists and time does not have to be spent locating them in much longer lists dominated by concgrams which are collocational frameworks. We have found that excluding the top fifty words is sufficient because typically they make up nearly 40% (Ahmad, 2005) of all of the language in a text or corpus. As mentioned earlier, it is important not to discard these collocational frameworks altogether because they are the essential building blocks in phraseology and need to be better understood and learned in their own right.

In Figure 3, we see the constituency variation of a concgram. The variation in this example is of the concgram 'development/our' with 'development' as the origin and 'our' as the associated word. The most common configurations in this 2-word concgram are for 'our' to occur two words before (5 times), or two words after (6 times), the origin, 'development'. This concgram does not have a contiguous configuration. It always contains at least one intervening word and this is typically 'of' when 'our' comes after 'development', and it is usually a modifier when 'our' precedes 'development' (see Figure 5).

It is possible for the user to view a summary of the positional variation that exists within a concgram. In Figure 4, we can immediately see the spread of 'development'

[3] The exclusion list is based on Ahmad's (2005) list of the fifty most frequent words in the British National Corpus which are, of course, all 'grammatical' words.



Total origins = 589

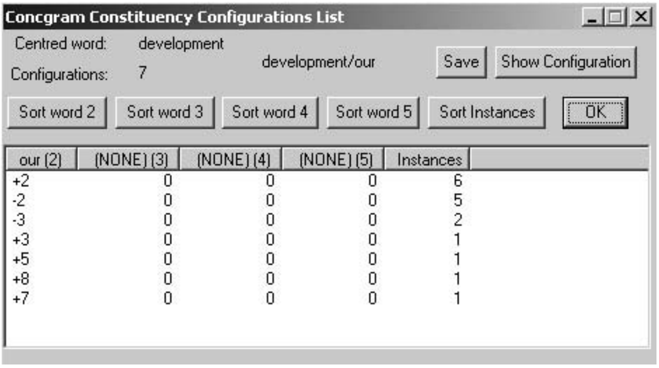
Total concgrams = 2135 Total instances = 5380

Set Sort Type: Sort Position; NO SPAN; Internal span: NA

2-word Concgrams: no file

SINGLE ORIGIN	ASSOCIATED WORD	ASSOC WD INST
Hong	Kong	72
as	well	20
development	our	18
Kong's	Hong	16
Government	SAR	15
development	Hong	12
development	Kong	12
as	Kong	11
as	Hong	10
as	such	9

Fig.2. Two-word concgram list using exclusion list.



Centred word: development

Configurations: 7

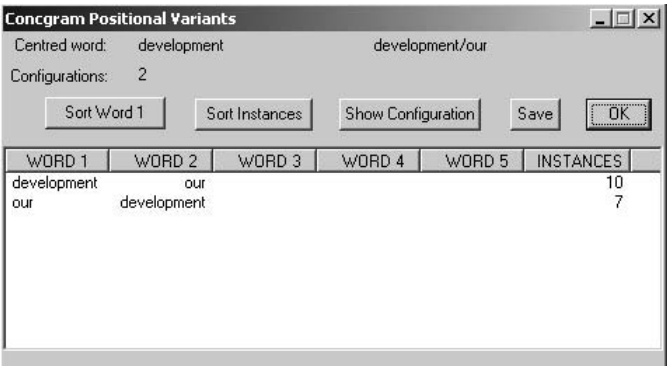
development/our

our (2)	(NONE) (3)	(NONE) (4)	(NONE) (5)	Instances
+2	0	0	0	6
-2	0	0	0	5
-3	0	0	0	2
+3	0	0	0	1
+5	0	0	0	1
+8	0	0	0	1
+7	0	0	0	1

Fig. 3. Constituency variation.

and 'our' in terms of which word precedes the other, and, in this particular concgram, the spread across the two possible positions is fairly even.

The sequencing of the concordance lines in the concgram display is designed to facilitate the identification of patterns of constituency and positional variation in a concgram. In Figure 5, the lines begin with 'our' positioned to the right of the origin, 'development', and they also start with instances with no intervening words, if any, and then one intervening word, two words, and so on. In this example, there is always at least one intervening word. Once all of the constituency variation to the right of the origin has been listed, the same system is followed for ordering the associated word(s) to the left of the origin and, in this example, these begin on line 11.

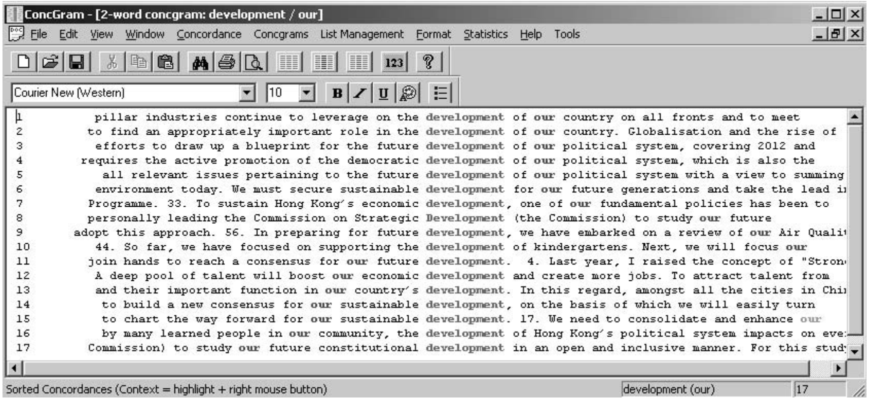


Centred word: development development/our

Configurations: 2

WORD 1	WORD 2	WORD 3	WORD 4	WORD 5	INSTANCES
development	our				10
our	development				7

Fig. 4. Positional variation.



```
1         pillar industries continue to leverage on the development of our country on all fronts and to meet
2        to find an appropriately important role in the development of our country. Globalisation and the rise of
3         efforts to draw up a blueprint for the future development of our political system, covering 2012 and
4       requires the active promotion of the democratic development of our political system, which is also the
5          all relevant issues pertaining to the future development of our political system with a view to summing
6         environment today. We must secure sustainable development for our future generations and take the lead i
7        Programme. 33. To sustain Hong Kong's economic development, one of our fundamental policies has been to
8        personally leading the Commission on Strategic Development (the Commission) to study our future
9      adopt this approach. 56. In preparing for future development, we have embarked on a review of our Air Qualit
10        44. So far, we have focused on supporting the development of kindergartens. Next, we will focus our
11       join hands to reach a consensus for our future development.  4. Last year, I raised the concept of "Stron
12        A deep pool of talent will boost our economic development and create more jobs. To attract talent from
13        and their important function in our country's development. In this regard, amongst all the cities in Chi
14         to build a new consensus for our sustainable development, on the basis of which we will easily turn
15         to chart the way forward for our sustainable development. 17. We need to consolidate and enhance our
16         by many learned people in our community, the development of Hong Kong's political system impacts on eve
17       Commission) to study our future constitutional development in an open and inclusive manner. For this stud
```

Sorted Concordances (Context = highlight + right mouse button) development (our) 17

Fig. 5. Concgram concordance lines.

3 Concgramming and CALL

3.1 Potential contribution of concgramming

There are at least three areas where concgramming could make a significant contribution in CALL. First, it can be used as a tool for textual analysis (see, for example, Carter & McCarthy, 1994; Stubbs, 2002). Second, it can be used to help raise learners' awareness of the idiom principle, in that it helps learners to find associated words and chunks in general (Sinclair & Mauranen, 2006). The third area is related to the second, and is the use of concgramming to enable learners to master the discourse of specialised areas and their specific genres (see, for example, Bhatia, 2004; Swales, 2004). In the description of concgramming activities that follows, the first area is dealt with and the second area is also present as learners will acquire aspects of the phraseology as they analyse the texts.

3.2 Student preparation

In this section, we outline replicable language learning activities that raise learners' awareness of the prevalence and importance of phraseology. The activities help to develop in learners the computational and analytical skills needed to conduct an initial study of the phraselogical profile of a text. Before doing the activities, the students need to be introduced to the broad notion of phraseology with examples to illustrate its forms and functions at the lexical-grammatical, discoursal and pragmatic levels, and also how the phraseology of English can vary for specific purposes compared to that of general English. Once this introduction to phraseology is complete, students can be introduced to the Concgram© search engine, and trained how to use its concgramming functions.

3.3 Concgramming activities

We describe the results of concgramming two Policy Addresses given by the Chief Executive of Hong Kong in October 2006 and October 2005. They are both texts which are of inherent interest to students in Hong Kong and they are also texts which are long enough (the 2005 Policy Address is 12,811 words and the 2006 Policy Address is 8,251 words) to yield sufficient instances of patterning, but not so long that there are lengthy delays while the program conducts its exhaustive fully automated searches. Texts of this kind of length can be concgrammed by students up to 3-word concgrams in under one hour on a regular desktop computer, and even faster if an exclusion list is used. This means that students can concgram the texts outside of class so that in class they can concentrate on working in small groups analysing and discussing their findings. We have found that lists of 2-word and 3-word concgrams are usually sufficient to yield an initial phraseological profile of a text and the amount of data for the students to analyse in one to two hours is also manageable. If students and teachers are interested, and there is time, the activity can be extended to include up to 5-word concgrams. Of course, some 4- and 5-word concgrams will be found when the concordance lines of the 2- and 3-word concgrams are studied.

The suggested activities are all concerned with working towards an initial determination of the aboutness of the two texts extracted from their phraseological profiles and the specifics of the activities are detailed below.

- Compile a list of the ten most frequent words in each text (combine inflected forms when appropriate).
- Compile a list of the twenty most frequent phrases in each text (again, combine inflected forms when appropriate).
- Monitor and record the frequency with which the most frequent words and phrases found in 2005 Policy Address occur in the 2006 Policy Address and vice versa.
- Discuss your findings from the two texts.
- Throughout your analysis of the two Policy Addresses, remember that the two Policy Addresses are of differing lengths and so direct comparison of frequencies need to take this into account.

The activities are influenced by data-driven learning (DDL, see Johns, 1991) that encourages teachers to give students direct access to data based on the assumption that "effective learning is a form of linguistic research" (*op. cit.*: 30). Ideally, the concgramming of the texts, or corpora, should take place outside of the classroom and teachers should divide up the students into small groups to make the activities both more interactive and collaborative, whether they are conducted face-to-face in the classroom or online. The activities have all been conducted with undergraduate language major students and are suitable for students with this background. It will be noticed that the activities as detailed above, and described below, require students to consider the inclusion, or exclusion of, inflected forms when compiling their lists. This is by no means an easy task and this requirement can be omitted to make the introduction to phraseology and aboutness more accessible for students from a different academic background.

3.4 Discussion of preliminary findings from the activities

Compiling a single word frequency list is in itself an interesting exercise in that it raises issues regarding inflected forms and whether or not they should be compiled together when determining the aboutness of a text. There are studies (see, for example, Tognini-Bonelli, 2001) which have shown that inflected forms tend to be associated with different meanings and functions, and so careful analysis of the concordance lines needs to be carried out and this can produce a lot of discussion and promote language awareness. When groups of students present their lists to the rest of the class, they need to be prepared to make a case for what they have decided to include.

Tables 1 and 2 show the ten most frequent words in the two Policy Addresses, including inflected forms when considered appropriate. In the 2005 Policy Address list (Table 1), there are four ‘words’ (nos. 3, 6, 7 and 9) which are combined totals. One is a combination of singular and plural forms, ‘policy(ies)’ (no. 6), and another is a combination of ‘develop’ and ‘development’. The latter is different in that it is a combination of two different word classes derived from the same root form: noun and verb. Again, the lines were studied and it was determined that in these instances the meanings were very similar and so justified combining them together. It should be emphasised that this is not always the case, for example, the words ‘govern’, ‘government’ and ‘governance’ in the 2006 Policy Address, which all share the same root form, were found to have different meanings and so were not combined. It can be seen throughout this study that our notion of phraseology challenges a categorisation

Table 1 *2005 Policy Address: Top 10 most frequent words*
(Total words spoken: 12,811)

Ranking	Word	Frequency	(Frequency in 2006)
1	Hong Kong	133	(72)
2	government	118	(71)
3	development, develop	73	(76)
4	public	66	(33)
5	community	60	(38)
6	policy(ies)	58	(23)
7	work, works, working	58	(17)
8	people	57	(17)
9	social, society	53	(26)
10	Mainland	46	(13)

Table 2 *2006 Policy Address: Top 10 most frequent words*
(Total words spoken: 8,251, i.e. 36% shorter than the 2005 Policy Address)

Ranking	**Word**	**Frequency**	**(Frequency in 2005)**
1	development, develop	76	(73)
2	Hong Kong	72	(133)
3	government	71	(118)
4	support	48	(33)
5	year(s)	47	(36)
6	family(ies)	43	(33)
7	community(ies)	38	(60)
8	public	33	(66)
9	service(s)	31	(41)
10	provide	27	(36)

system based on word class; rather, what is driving the analysis and subsequent categorisation is the meaning in context of the word(s). Of the other two instances, one is the inflected forms of the verb ‘work’ (no. 7) which in this speech share a similar meaning, and the other is ‘social, society’ (no. 9) which is an interesting example and is discussed later in the paper.

In the list based on the 2006 Policy Address (Table 2), we can see that five ‘words’ (nos. 1, 5, 6, 7 and 9) are combined totals made up of two inflected forms. Four of these are a result of combining singular and plural forms (nos. 5, 6, 7 and 9), where the two forms were judged to convey similar meanings. The fifth one is ‘development, develop’, also found in the 2005 list and discussed earlier.

Tables 3 and 4 show lists of the twenty most frequent phrases in the two Policy Addresses, including inflected forms when considered appropriate. A forward slash indicates that there is variation, either constituency, positional or both, in the phrase.

Before discussing individual phrases in the above lists, two interesting points are worth making. First, there are differences between the single words which appear in the word frequency lists and those words which are associated together in the Policy Addresses. This suggests that single word frequencies are not necessarily good indicators of the phraseological profile of a text and hence its aboutness. For example, in the 2005 list, ‘public’, ‘policy(ies)’, ‘work, works, working’ and ‘Mainland’ are among the most frequent words but are not to be found in any of the most frequent phrases. Similarly, we can see that while ‘service(s)’, ‘public’, and ‘community(ies)’ are all in the 2006 word frequency list, they do not form part of the most frequent phrases. Conversely, there are words which are associated in some of the most frequent phrases, but none of which are among the most frequent single ‘words’, for example, in the 2005 list ‘Chief Executive’, ‘Legislative Council’, ‘air quality’, ‘service(s)’ and ‘the Central Authorities’. Examples from the 2006 Policy Address include ‘strong governance’, ‘film industry’, and ‘protect/environment’. These observations are useful for students to discover for themselves and are a good source of discussion and language awareness raising. Second, even when the single ‘words’ recur in the list of the most frequent phrases, we find that the lists of phrases are usually much better at revealing the

Table 3 *2005 Policy Address: 20 most frequent phrases (Total words spoken: 12,811)*

Ranking	Phrase or Word Combination	Frequency	(Frequency in 2006)
1	Chief Executive	18	(8)
2	the SAR Government	17	(14)
2	Legislative Council	17	(8)
3	the Central Authorities	1	(2)
4	Hong Kong/people	14	(2)
5	social harmony/harmonious society	13	(4)
6	Hong Kong/development/develop	11	(10)
6	food safety	11	(0)
7	the Basic Law	10	(2)
8	community/support	9	(8)
8	powers and functions	9	(0)
8	the/government/continue	9	(0)
9	economic/economy/development	8	(7)
9	government/support	8	(4)
9	air quality	8	(8)
10	world city	7	(1)
10	one country two systems	7	(0)
10	principal officials	7	(0)
10	emissions reduction	7	(3)
11	Commission on Strategic Development	6	(1)

Table 4 *2006 Policy Address: 20 most frequent phrases (Total words spoken: 8,251, i.e. 36% shorter than the 2005 Policy Address)*

Ranking	Phrase or Word Combination	Frequency	(Frequency in 2005)
1	family/support	15	(1)
2	the SAR Government	14	(17)
3	Hong Kong/development	10	(11)
4	support/development	8	(1)
4	air quality	8	(8)
4	Chief Executive	8	(18)
4	family members	8	(5)
4	Legislative Council	8	(17)
5	economic/development	7	(8)
5	sustain/development	7	(2)
5	last year	7	(5)
6	future/development	6	(3)
6	development of/political system	6	(0)
6	mutual/support	6	(1)
6	strong governance	6	(2)
6	protect/environment	6	(0)
7	provide/support	5	(5)
7	foster/family	5	(0)
7	film industry	5	(2)
7	provide/parents	5	(0)

aboutness of the text than the single word list. For example, 'develop, development' is in the 2005 most frequent word list, but it is only when we study the associated words in the most frequent phrase list, 'Hong Kong/development/develop' and 'economic/economy/development', that we find out what is being developed or what the development is. Another example is 'family(ies)' which is in the 2006 single word list but does not tell us much in terms of aboutness. However, when we look at its phraselogical profile we find out much more in terms of what it is 'about' from the phrases that it belongs to in this text: 'family/support', 'family members', and 'foster/family'.

Another point to make is that the words in those phrases which do not contain a forward slash are contiguous. For example, 'the SAR Government' (14 instances) is an invariable contiguous word association. Those that do contain a forward slash are phrases which contain either constituency or positional variation, or both. It should be noted that the non-contiguous forms of phrases make up more than one third (7 instances) of the 2005 list and more than half (11 instances) of the 2006 list. This fact underlines the importance of studying the variation that can exist in concgram configurations when compiling the phraseological profiles of a text or corpus.

We now look at some examples of the kinds of concgram concordance lines which have to be analysed by students when compiling the lists of the most frequent phrases. They have been chosen to illustrate the potential variety to be found within a concgram and the kinds of learning experience afforded from critically studying them.

economic/economy/development (2006 Policy Address)

1. growth. Strong government is a prerequisite for **economic development.** A harmonious society, itself
2. society, itself founded on strong government and **economic development,** will create a favourable
3. workforce is more than a deciding factor in **economic development.** It also helps create social
4. 71. We have a steadfast commitment to promoting **economic development.** Following a strong rebound last
5. Although there will be various risks in global **economic development** in the coming year, the recovery of
6. set up under the Commission to study political, **economic** and social **development.** The Central Policy Unit
7. Hong Kong has **development** into a services-oriented **economic** that relies on the vast Mainland market. The

These lines include five instances of 'economic development' (lines 1-5) and one non-contiguous instance, 'political, *economic* and social *development*' (line 6) which is considered to have essentially the same meaning as those on lines 1-5. The last concordance line appears to be quite different from the others. On line 7 two different forms, 'developed' and 'economy' are associated and one must argue it shares the same canonical meaning as the rest to justify compiling it alongside the others, and this makes it an interesting case to discuss.

family(ies)/support (2006 Policy Address)

1. for policies and initiatives relating to **family support.** The

Commission would bring under one

2. the functioning of **family** and provide various **family**-based **support** as well as fostering close and
3. social problems, the key lies in establishing a **family**-based **support** network and forging closer and
4. aid among neighbours is a strong **support** for **family.** We encourage community building and friendly
5. of life, in particular, strengthened **support** for **family** and intensified efforts in our anti-pollution
6. Next, we will focus our resources to **support** the **family** by easing the financial burden of parents. We
7. We will also continue to provide **support** for **family** with disabled members. The Government will
8. will also promote mutual **support** among **family** through our community networks. 39. The
9. an integral part of government **support** for the **family.** The Education and Manpower Bureau issued the
10. a number of policies in **support** of extended **family.** For example, under the public housing
11. further ways to enhance **support** for extended **family.** 40. The Government implemented the first
12. violence, as well as **support** services for the **family** members of victims. We will strengthen the
13. geared towards **support**ing and consolidating the **family**, and fostering the well-being of **family** members.
14. to start with **support**ing and strengthening the **family:** fostering a sense of responsibility and
15. 46. The **support** rendered by the Government to **families** is not just confined to pre-primary education.

The above phrase ‘family(ies)/support’ is a very good example of the extensive range of constituency and positional variation that can be found in the configurations of concgrams. Combined, this paraphrasable family has positional variation with ‘family/support’ (lines 1-3) and ‘support/family’ (lines 4-15). There is also considerable constituency variation with one contiguous instance (line 1) and then a range of intervening words ranging from one intervening word (lines 2-8) up to five intervening words (line 15). Despite the wide range of variation that exists, these different configurations all share a similar meaning in this text and demonstrate the broad notion of phraseology adopted in this study.

environment/protect/protection (2006 Policy Address)

1. ed on a review of our Air Quality Objectives. The **Environment**al **Protection** Department, in light of the World Hea
2. icy initiatives for environmental protection. 52. **Environment**al **protection** is a long-term undertaking. First, we
3. le engaging the public to formulate and implement **environment**al **protection** policies and measures. I encourage al
4. les and introduce specific policy initiatives for **environment**al **protection.** 52. **Environment**al protection is a lo

5. e best practicable means. The need to **protect** our **environment** will be the focus of our negotiations with the pow
6. d industry sectors to do their bit to **protect** the **environment** by means such as adopting comprehensive clean prod

In the above lines, there are four instances of the contiguous 'environmental protection', plus two instances of 'protect' plus 'environment' (lines 5 and 6) which belong to different word classes compared to 'environmental' and 'protection', but have a similar meaning in this text. The latter have positional and constituency variation compared to the more common form and share a similar structure, '*to protect* + determiner ('the' or 'our') *environment*'.

development/future (2006 Policy Address)

1. environment today. We must secure sustainable **development** for our **future** generations and take the lead in
2. personally leading the Commission on Strategic **Development** (the Commission) to study our **future**
3. join hands to reach a consensus for our **future development**. 4. Last year, I raised the concept of "Strong
4. adopt this approach. 56. In preparing for **future development**, we have embarked on a review of our Air Quality
5. all relevant issues pertaining to the **future development** of our political system with a view to summing up
6. efforts to draw up a blueprint for the **future development** of our political system, covering 2012 and
7. Commission) to study our **future** constitutional **development** in an open and inclusive manner. For this study,

This is another example of a concgram which has considerable variation across its seven instances. There are four contiguous instances (lines 3-6) and constituency variation of between one and three words. This example also serves to illustrate the kinds of decision that need to be made by students when compiling their frequent phrase lists. Line 2 was not counted because it was judged that this instance of co-occurrence of the words 'development' and 'future' were not associated in the way that these words are associated in the other lines. In fact, we can see that 'future' in line 2 is associated with a different instance of 'development' in line 7.

continue/government (2005 Policy Address)

1. multi-disciplinary system. In response, the **Government continue**s to work with the whole spectrum of
2. the effective implementation of CEPA, the SAR **Government continue**s to work closely with the Mainland
3. are required to play an important role. The **Government** will **continue** its partnership with the social
4. left school to help them find employment. The **Government** will **continue** to study the needs of the ethnic
5. harmony is the foundation of social harmony. The **Government** will **continue** to enhance family cohesion with
6. another to tie in with this development. 42. The **Government** will

continue to implement various social

7. development for our children and youth. The **Government** will **continue** to provide extra support to
8. for Teaching Excellence and Teachers?Day, the **Government** will **continue** to express its regards and respect
9. and interplay among creative talent. The **Government** will **continue** to allocate resources to foster a

These lines include an inflected form of 'continue', 'continues' (lines 1 and 2), and it is the use of the modal 'will' which accounts for much of the constituency variation found (lines 3-9). This paraphrasable group can also be seen to typically precede a *to*-infinitive, with the exception of line 3. This example shows how students can be encouraged to look beyond the associated words in the concgram to uncover other patterns of co-selection.

harmony/harmonious/social/society (2005 Policy Address)

1. to implement 'One Country, Two Systems', promote **soci**al **harmon**y, and enhance economic growth. Strong
2. helping the economy power ahead at full steam. **Soci**al **harmon**y includes harmony between humankind and
3. its five-year objective of "Working Together for **Soci**al **Harmon**y". On the basis of joint responsibility,
4. right time, we now have the right climate for **soci**al **harmon**y and good governance. This is an epochal
5. themselves, and create conditions for fostering **soci**al **harmon**y. 40. To help the needy we have an
6. in economic development. It also helps create **soci**al **harmon**y. We place special emphasis on
7. community. Family harmony is the foundation of **soci**al **harmon**y. The Government will continue to
8. for economic development. A **harmon**ious **soci**ety, itself founded on strong government and
9. for sustaining the vitality and **harmon**y of **soci**ety. Hong Kong has long been recognised as the
10. are striving to foster a **harmon**ious **soci**ety. For example, the Women's Commission has
11. They are also the foundations of a **harmon**ious **soci**ety. The SAR Government is determined to safeguard
12. relations are essential in building a **harmon**ious **soci**ety. 53. The Labour Advisory Board is now
13. those factors that threaten long-term **harmon**y in **soci**ety. These include: employment difficulties for

The above concordance lines illustrate the interparaphrasability (Sinclair, 2005: 4) employed by speakers and writers. Students need to be trained and encouraged to search for the possible presence of inflected forms to see whether the different forms have similar or different meanings. 'Social harmony' (seven instances) and 'harmonious society' (four instances) are two phrases, each of which contains the inflected forms of the other, and 'harmony of society' (one instance) and 'harmony in society' (one instance) both combine words from the two main forms. All the forms are functioning as paraphrases in this discourse and are an informative example of a paraphrasable family of word associations.

3.5 Discussion of findings from the lists

Once the students have compiled their frequency lists, they can be encouraged to compare the contents of the lists and compare the frequencies of phrases in one text against the number of instances, if any, in the other text. The similarities and differences found when comparing the lists are of interest and can shed light on the genre-specificity of some phrases and the transient nature of others. For example, it would seem that there are some words and phrases that are more predictable in such texts as they are present in both texts. In the single word lists, we find that half of the words are common to both lists, 'Hong Kong', 'government', 'development/develop', 'community(ies)' and 'public'. This might be a case of genre-specific usage which could be further investigated by students comparing the frequencies in other Policy Addresses, or in political speeches in general.

In terms of phrases, six are found in both lists, 'Chief Executive', 'the SAR Government', 'Legislative Council', Hong Kong/development/develop', 'air quality' and 'economic/economy/development' and, if more Policy Addresses were studied, it might well be the case that these are found to be essential phrases in any Hong Kong Policy Address, although in the case of the phrase 'air quality' this is more likely to signal a longer-term problem and might eventually disappear from future Policy Addresses. Thus commonality can be explored further if corpora of Policy Addresses and political speeches are available to enable comparative studies (and they are in Hong Kong)[4] and could lead to an enhanced awareness of genre-specific usage; but then there are the differences to be discussed.

There are two kinds of differences for students to ponder. The first is when students find a relative decline, or relative increase, in the use of certain words and phrases. For example in the single word lists, there are no words in one top ten list that do not occur at all in the other. The differences here are relative and might reflect different political priorities, or political expediency. For example, 'Mainland' (i.e. the China Mainland as opposed to Hong Kong) has 46 instances in 2005, but drops to only 13 instances in 2006. This could be the result of a shift in focus between the Policy Addresses or a desire to play down the extent of the interrelationship between the Hong Kong and the Mainland. A similar example, can be found in the lists of phrases. In 2005, 'Central Authorities' (i.e. the central government in Beijing) has 15 instances and in 2006 just 2 and, again, the reasons behind this can be interesting to examine further.

The second difference is when a phrase is used in one text, but not in the other. As mentioned above, there are no instances of this kind of difference for the most frequent single words. Examples are 'food safety' (11 instances in 2005, 0 in 2006), 'one country two systems' (7 instances in 2005, 0 in 2006). In these cases, one might presume that

[4] The Hong Kong Corpus of Spoken English is a 2m word corpus which contains a 0.5m word sub-corpus of public discourses many of which are political speeches (for details of the corpus see Cheng, Greaves & Warren, 2005).

food safety was prioritised in 2005 and no longer an issue in 2006. However, when it comes to the decline in the usage of 'one country two systems', which is fundamental to Hong Kong's status as a Special Administrative region enjoying a high degree of autonomy from China mainland, this might signal a change in priorities of a different order. In 2006, again, we find phrases that were not in the 2005 Policy Address, for example, 'development of /political system' (6 instances) and 'protect/environment' (6 instances). These both reflect the changing priorities of the government in Hong Kong.

Based on these kinds of discussion, students can critically reflect on their findings and discuss which words and phrases might illustrate the aboutness of this genre and which ones might illustrate the aboutness of a particular Policy Address. There are probably 'aboutgrams' (Sinclair, 2006, personal communication) which are genre-specific and others which are text-specific among the many concgrams in a text and, if time and corpora are available, these are very useful aspects to explore more fully. This further illustrates that phraseology is not fixed and, in political discourse in particular, as has been observed by others (see, for example, Cheng, 2004), some phrases have a relatively short shelf life compared to others.

4 Conclusions and implications

This study has introduced a new computer-mediated methodology, concgramming, which aims to facilitate the introduction of phraseology to language learners. It has been argued that language learners need to be aware of and understand both the extent and the importance of phraseology in the English language. The basic principles of the proposed contribution to learning and teaching methodology are aimed to ensure that the learning process is both interactive and collaborative in nature and is derived from DDL (see Johns, 1991 and Cheng, Warren & Xu, 2003) which casts the language learner in the role of language researcher (Johns, 1991: 2).

Concgramming learning and teaching activities have been described that highlight key elements in the understanding and production of phraseology in English and which can be replicated in applied language studies and English proficiency courses using texts and/or corpora. In addition to enhancing teachers' and students' critical awareness of the nature and role of phraseology in the English language, the activities also enhance students' critical and creative thinking through the understanding, analysis, comparison and application of phraseology that is specific to individual text types. Concgramming also has the potential to become a useful additional tool in LSP because it provides the means for identifying genre-specific uses of phraseology. This methodology and these kinds of activities with respect to English phraseology are proposed because it has been argued that phraseology is a major area of English language study that is currently given insufficient, or no attention, and this imbalance should begin to be rectified.

Acknowledgements

In March this year the world of linguistics lost its most original thinker, Professor John McHardy Sinclair. We are greatly saddened by John's passing and we extend our deepest sympathies to his family. We were privileged to have had John as the consultant to the project from which this paper originates. His groundbreaking ideas and generous support have been invaluable to us.

We would also like to acknowledge the very useful feedback provided by the three anonymous reviewers. The work described in this paper was substantially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region (Project No. G-YE86).

References

Ahmad, K. (2005) *Terminology in Text. Paper presented at the Tuscan Word Centre International Workshop: Dial a Corpus*. Certosa di Pontignano, Italy, June, 2005.

Aston, G. (1997) Small and large corpora in language learning. In: Lewandowska-Tomaszczyk, B. and Melia, J. P. (eds.) *Practical applications in language corpora*. Lodz: Lodz University Press, 51–62.

Bernardini, S. (2000) Systematising serendipity: Proposals for concordancing large corpora with language learners. In: Burnard, L. and McEnery, T. (eds.) *Rethinking language pedagogy from a corpus perspective*. Frankfurt: Peter Lang, 225–234.

Bernardini, S. (2002) Exploring new directions for discovery learning. In: Kettemann, B. and Marko, G. (eds.) *Teaching and learning by doing corpus analysis*. New York: The Edwin Mellen Press, 165–182.

Bhatia, V. (2004) *Worlds of Written Discourse*. London: Continuum.

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

Biber, D., Johansson, S., Leech, G. Conrad, S. and Edward, F. (1999) *Longman Grammar of Spoken and Written English*. Harlow: Longman.

Braun, S. (2005) From pedagogically relevant corpora to authentic language learning contents. *ReCALL*, **17**(1): 47–64.

Carter, R. and McCarthy, M. (1994) *Language as Discourse: Perspectives for Language Teaching*. London and New York: Longman.

Carter, R. and McCarthy, M. (2006) *Cambridge Grammar of English*. Cambridge: Cambridge University Press.

Cheng, W. (2004) // FRIENDS // LAdies and GENtlemen //: Some preliminary findings from a corpus of spoken public discourses in Hong Kong. In: Connor, U. and Upton, T. A. (eds.), *Applied Corpus Linguistics: A Multidimensional Perspective*. Amsterdam/New York: Rodopi, 35–50.

Cheng, W., Greaves, C. and Warren, M. (2005) The creation of a prosodically transcribed intercultural corpus: The Hong Kong Corpus of Spoken English (prosodic). *ICAME Journal*, **29**: 47–68.

Cheng, W., Greaves, C. and Warren, M. (2006) From n-gram to skipgram to concgram. *International Journal of Corpus Linguistics,* **11**/4: 411–433.

Cheng, W., Warren, M. and Xu, X. (2003) The language learner as language researcher: Putting corpus linguistics on the timetable. *System*, **31** (2): 173–186.

Clear, J. (1993) From Firth principles: computational tools for the study of collocation. In: Baker, M., Francis, G. and Tognini-Bonelli, E. (eds.) *Text and Technology*. Amsterdam: John Benjamins, 271–92.

Cobb, T. (1997) Is there any measurable learning from hands on concordancing? *System*, **25** (3): 301–315.

Fletcher, W. H. (2006) "Phrases in English" Home. http://pie.usna.edu/.

Gaskell, D. and Cobb, T. (2004) Can learners use concordance feedback for writing errors? *System*, **32** (3): 301–319.

Hunston, S. (1995) A corpus study of some English verbs of attribution. *Functions of Language*, **2** (2): 133–158.

Hunston, S. and Francis, G. (2000) *Pattern Grammar: A Corpus-driven Approach to the Lexical*

Grammar of English. Amsterdam: John Benjamins.

Johns, T. (1991) Should you be persuaded: two samples of data-driven learning materials. In: Johns, T. and King, P. (eds.) *Classroom Concordancing*. English Language Research: Birmingham University, 1-16.

Kennedy, C. and Miceli, T. (2002) The *CWIC* project: Developing and using a corpus for intermediate Italian students. In: Kettemann, B. and Marko, G. (eds.) *op. cit.*, 183–192.

Louw, B. (1993) Irony in the Text or Insincerity in the Writer? The Diagnostic Potential of Semantic Prosodies. In: Baker, M., Francis, G. and Tognini-Bonelli, E. (eds.) *Text and Technology: In Honour of John Sinclair*. Amsterdam/Philadelphia: John Benjamins.

McCarthy, M. (2005) *English Collocations in Use*. Cambridge: Cambridge University Press.

Phillips, M. (1983) *Lexical Macrostructure in Science Text*. (Unpublished PhD thesis, Department of English, Faculty of Arts, University of Birmingham).

Phillips, M. (1989) Lexical Structure of Text. Discourse analysis monographs: 12. *English Language Research*: University of Birmingham.

Policy Address (2005/2006) http://www.policyaddress.gov.hk/05-06/eng/index.htm

Policy Address (2006/2007) htttp://www.policyaddress.gov.hk/06-07/eng/pdf/speech.pdf

Sinclair, J. McH. (1987) The nature of the evidence. In: Sinclair, J. McH. (ed.) *Looking Up: An Account of the COBUILD Project in Lexical Computing*. London: Collins, 150–159.

Sinclair, J. McH. (1991) *Corpus Concordance Collocation*. Oxford: Oxford University Press.

Sinclair, J. McH. (1996) The search for units of meaning. *Textus*, **9** (1): 75–106.

Sinclair, J. McH.(2003) *Reading Concordances*. London: Pearson Longman.

Sinclair, J. McH. (2004a) *Trust the Text*. London: Routledge.

Sinclair, J. McH. (ed.) (2004b) *How to Use Corpora in Language Teaching*. Amsterdam: John Benjamins.

Sinclair, J. McH. (2005) *Document Relativity* (manuscript). Tuscan Word Centre, Italy.

Sinclair, J. McH., Jones, S. and Daley, R. (1970) *English Lexical Studies*. Report to the Office of Scientific and Technical Information.

Sinclair, J. McH., Jones, S. and Daley, R. (2004) *English Collocation Studies: the OSTI Report*. London: Continuum.

Sinclair, J. McH. and Mauranen, A. (2006) *Linear Unit Grammar*. Amsterdam: JohnBenjamins.

Sinclair, J.McH. and Renouf, A. (1991) Collocational Frameworks in English. Reprinted in Sinclair, J. McH. *Lexis and Lexicography*. National University of Singapore: Unipress, 55–71.

Stevens, V. (1991) Concordance-based vocabulary exercises: a viable alternative to gap-fillers. In: Johns, T. and King, P. (eds.) *op. cit.*, 47–63.

Stubbs, M. (2002) *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford: Blackwell.

Swales, J. (2004) *Research Genres: Explorations and Applications*. Cambridge: Cambridge University Press.

Tognini-Bonelli, E. (2001) *Corpus Linguistics at Work*. Amsterdam: John Benjamins.

Wilks, Y. (2005) REVEAL: the notion of anomalous texts in a very large corpus. *Tuscan Word Centre International Workshop: Dial a Corpus*. Certosa di Pontignano, Tuscany, Italy, 31 June–3 July 2005.