To appear in International Journal of Corpus Linguistics

From n-gram to skipgram to concgram

1. Introduction

One of the most important findings, if not the most important finding, to come out of corpus linguistics has been what Sinclair (1987) terms 'the idiom principle ', i.e. the phraseological tendency, whereby words are co-selected by speakers and writers which gives rise to collocation and other features of idiomaticity. More recently, Sinclair (1996) uses the term 'lexical item' to describe the outcome of the combination of five categories of co-selection (i.e. semantic prosody, semantic preference, colligation, collocation and the invariable core word/s). Today, nobody seriously interested in the meaning and use of language can ignore tendencies of word co-selection which are evident in linguistic patterns. Researchers have uncovered significant findings in, for example, pattern grammar (see, for example, Hunston and Francis, 2000; Hunston, 2002; Partington, 1998), phraseology (see, for example, Hoey, 2005; Tognini-Bonelli, 2001; Sinclair, 1987; Sinclair, 1996; Sinclair et al, 2004; Stubbs, 2001; Halliday, Teubert and Yallop, 2002; Teubert, 2005) and semantic prosody (Louw, 1993; Sinclair, 1991; Sinclair, 2004).

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 how are we to find them all? Those working in the fields of NLP, computational linguistics and corpus linguistics are familiar with 'n-grams' which are contiguous words that constitute a phrase, or a pattern of use, and that recur in a corpus. Actual realisations of n-grams come in the form of bi-grams, tri-grams, and so on, indicating the number of words in the phrase. Current searches for n-grams, sometimes termed 'word clusters', 'lexical clusters' or 'bundles' (see, for example, Biber, Conrad

and Cortes, 2004; Carter and McCarthy, 2006), generate phrases such as 'a lot of people', but would miss instances of the same pattern 'a lot of people' when it is realised in instances such as 'a lot of local people' or 'a lot of different 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 association may be overlooked, and that collocations that typically occur in non-contiguous sequences (i.e. AB, ACB) risk going undiscovered.

The limitations that are a product of n-gram searches have led to the recent development of gapped n-grams or 'skipgram' searches. 'Skipgram' is used in NLP (see, for example, Wilks, 2005) to describe non-contiguous word associations. In other words, skipgrams can handle constituency variation. The work on skipgrams is still at an early stage, but skipgrams are already seen as a means to do more with less, according to Wilks (2005) who claims that a 3-word skipgram search of a 50-million-word corpus will reveal all of the trigrams found in a 200 million-word corpus. However, as a skipgram search also includes all contiguous word associations, and so subsumes n-grams found in the same span (Wilks, 2005), its name is potentially misleading as one might suppose that it only locates non-contiguous associations. Skipgram searches, however, are not without limitations. They are currently limited to 3-word skipgrams and four 'skips' (Wilks, 2005), meaning that any two associated words that are more than four words apart stay undiscovered. With a total window of usually 11 tokens, the cut-off is bound to miss instances. In addition, existing skipgrams searches may require the input of a formula which can be cumbersome.

Similar to n-gram searches, 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. An example of an automated skipgram search is Fletcher's 'phrase frames' (2006) which does not require prior nomination of a search query by the user. Phrase frames are based on an initial automated search for n-grams, 'where *n* falls in the range 1-8' (Fletcher, 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' (Fletcher, 2006). Thus phrase frames are one restricted form of skipgram constrained by narrow search parameters, with the result that other non-contiguous associations of the same words remain undiscovered, as is any pattern with positional variation. The driving force behind existing skipgram searches seems be the perceived primacy of n-grams. Skipgrams are viewed as another means of revealing n-grams, or variants of n-grams, rather than an end in themselves.

2. ConcGram©[1]

Given the limitations of the existing search engines that generate n-grams and skipgrams, what is needed is a search engine which, on top of the capability to handle constituency variation (i.e. AB, ACB), also handle positional variation (i.e. AB, BA), conduct fully automated searches, and search for word associations of any size. The program ConcGram© developed by Greaves (2005), who works concurrently with those in NLP, is designed with the goal of meeting all of the requirements of such a search engine. ConcGram© can identify all the potential configurations of between 2 and 5 words in any

[1] ConcGram© is a search engine designed and implemented by Chris Greaves, Senior Project Fellow, English Department, The Hong Kong Polytechnic University, specifically to perform fully automated concgram searches.

corpus, based on a window of any size, to include the associated words even if they occur in different positions relative to one another (i.e. positional variation) and even when one or more words occur in between the associated words (i.e. constituency variation). Most important of all, this search engine can conduct fully automated searches throughout the data with no prior nomination of any parameters from the researcher; in other words, it will nominate the groupings itself.

This paper describes the development and implementation of ConcGram©, delineates its unique features and functions, and explores its implications. ConcGram© was piloted on a one-million-word sample of the Hong Kong Corpus of Spoken English (HKCSE) (see Cheng, Greaves and Warren, 2005 for details of this corpus). The paper discusses the concgram search results to demonstrate the potential of ConcGram© to corpus linguists. Following the position of Stubbs (1995), the paper also discusses and compares various *t*-test scores and MI values and raises questions about the value and importance of these statistical tests in corpus linguistic studies.

3. Defining concgrams

For our purposes, a 'concgram' is all of the permutations of constituency variation and positional variation generated by the association of two or more words. This means that the associated words comprising a particular concgram may be the source of a number of 'collocational patterns' (Sinclair, 2004: xxvii). In fact, the hunt for what we term 'concgrams' has a fairly long history dating back to the 1980s (Sinclair, 2005, personal communication) when the Cobuild team at the University of Birmingham led by

Professor John Sinclair attempted, with limited success, to devise the means to automatically search for non-contiguous sequences of associated words.

The development of the notion of a concgram challenges the current view about word co-occurrences that underpins the KWIC display. Years of studying KWIC displays have perhaps unintentionally created, in the minds of some users, a hierarchical approach which regards the node as the centre of attention and the words associated with the node as being in a subordinate relationship to it. It is worth restating, as was first done in the work of Sinclair, et al. (1970: 10), that although these are convenient terms to use, the term 'node' does not imply a hierarchy between it and its 'collocate', and that 'node' words that have 'collocates' are themselves collocates if the collocate is studied as the node.

Rather than focusing on the node, ConcGram© highlights all of the associated words of a concgram in each concordance line. This unique feature shifts the user's focus of attention from the node to the concgram. In other words, word associations become the focus of attention, and a 'node' is not the 'sun' around which collocates orbit in a subordinate relationship. For this reason, the term 'origin' is used for the word or words that form the basis of the automated concgram search to emphasise the difference between ConcGram© and KWIC. For purely display layout purposes, the on-screen view of congram concordance lines needs a sort-point simply to present a visually intelligible page. Since the automatic mode of ConcGram© begins with the creation of 2-word concgrams, and then builds up iteratively[2] to 5-word concgrams, the notion of a

[2] i.e. a double-origin search based on a 2-word concgram finds the third member of a 3-word concgram, which then becomes a triple-origin search which finds the fourth member of a 4-word concgram, and so on

'node' is redundant and the notion of 'origin' (1-word, 2-word, 3-word or 4-word) better foregrounds the fact that associated words are at the heart of every search.

The primary function of ConcGram© is to perform fully automated concgram searches, but it is also possible for the user to specify a word or words as a concgram search query. When user-nominated words are performed, the choice of which word is to be in central position is decided alphabetically.

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), and so further increases the likelihood that the concgram searches will enable the researcher to discover not only a more extensive description of patterns of collocation and their meanings, but also, and more importantly, new patterns of language use. That the researcher does not have to have specific words in mind means that studies are corpus-based rather than corpus-driven. Identification of all the potential patterns of collocation contributes not only to the co-selections that constitute extended units of meaning, i.e. 'lexical items' (Sinclair, 1996), but also enhances our attempts to understand 'intertextuality', 'intercollocability' and 'interparaphrasability', all of which are fundamental to our understanding of language (Sinclair, 2005).

4. The concgram search

The product of the concgram search is the identification of the associated words and their configurations in a corpus within a given span, and most useful of all, this span can be tailored to suit the needs of the user. The process of creating the initial 2-word concgram list can be summarised as follows:

Step 1: All the unique words (i.e. types) in a text are identified and listed.

Step 2: With this list concordance searches are made, with each unique word acting as the single origin for the search.

Step 3: All co-occurring words in the concordance lines are then listed for each single origin.

From this initial 2-word concgram list, the user can go on to build a 3-word concgram list, then a 4-word list, and finally a 5-word concgram list, all derived from fully automated searches. For example, the 3-word concgram list is created by performing double-origin searches based on the 2-word concgram list, taking the resulting concordance lines and listing each associated word found in them together with each double origin searched.

To illustrate what a concgram is, Figure 1 shows sample concordance lines of the result of an automated 3-word concgram search. The concgram is 'Asia/world/city', from a search with 'Asia/world' as the double origin.

[Insert Figure 1: Results of a 3-word concgram search]

Figure 1 shows the 3-word concgram (aisa/world/city) sorted by configuration, and illustrates positional variation (ABC, CAB). Two configurations are found: *world city of Asia* (lines 3-7) and *Asia/Asia's world city* (lines 8-16). A conventional tri-gram search would not have found the first non-contiguous configuration, with the associated words ('world' and 'city') at N- positions. Figure 1 also shows that the concgram search is not fazed by features of spoken discourse corpora such as repetition, pauses or fillers (*er*, *um*,

etc.). On line 2, for example, the 3-word concgram is still revealed when the speaker says *world city of of of Asia*.

For a word span to be applied, there must be at least a double-origin search. The reason that the initial 2-word concgram list (with a single origin) has no word span is because it starts with a single origin search for all the unique words in the text. A single origin search can only find all the instances of that word, and the span is what the user has set in characters, the default width being 50 characters on each side of the node word. The number of words on either side of the single origin word is variable, but probably averages at 9 or 10 words in each 50 characters. The associated words can then be identified and listed for each single origin word search.

Figure 2 below shows both 2- and 3-word concgram lists for words starting with 'c' automatically generated from one million words of the HKCSE.

[Insert Figure 2: 2- and 3-word concgrams for words beginning with letter 'c')]

5. Concgram lists and ways of determining significance

The reason for administering statistical tests is to attempt to calculate the significance of word associations in context. While the fully automated concgram search will find all of the contiguous and non-contiguous collocations that constitute 2-word, 3-word, 4-word and 5-word concgrams, including positional variation, the search will also list word co-occurrences that may not prove to be meaningfully associated when examined in context.

For these reasons, in ConcGram©, statistical tests can be run to generate *t*-scores and MI values to find out the statistically significant cut-offs for concgram lists and to

provide the user with an indication as to which word associations are more likely to prove to be meaningful and which ones the user can reasonably afford to ignore. More statistical tests could be added in the future, but these two tests have been chosen initially because they are widely used in corpus linguistics (see, for example, Barnbrook, 1996; Clear, 1993; Stubbs, 1995). However, the extent to which *t*-scores and MI values are useful will be discussed later in this paper, and some users may wish to access the concgram lists without the intervention of one or more statistical tests.

For illustration purposes, the same file generated for all words starting with the letter 'c' is used (Figure 2). Figure 2 shows that to create the 3-word concgram list, 56,739 searches were performed, based on a double-origin search (i.e. the search takes each 2-word concgram and looks for an associated word). This list of 56,739 instances of 2-word concgrams for 1,209 single origin words resulted in a list of 385,746 instances of 3-word concgrams with a span of 5 words (Table 2), and 397,822 instances with a span of 10 words (Table 3). The difference between the instances of 4- and 5-word concgrams is much greater when the word span is increased from 5 to 10 (see Tables 2 and 5). While the size of the lists increases dramatically[3], measures can be taken to reduce the size of the lists by using statistical cut-offs. Four tests were conducted:

Test 1: with no *t*-score or MI cut-off set

Test 2: with *t*-score cut-off set at 2.0

Test 3: with an MI cut-off set at 3.0

Test 4: with both MI and *t*-score cut-offs used

[3] The long lists generated by the searches take time to process. List management options are built into ConcGram©. A combination of list management (i.e. splitting the lists and later re-merging them) and using several computers to process searches based on the split lists saves time. 50,000 searches in a one-million-word corpus take approximately 24 hours on a typical PC.

The formulas used for calculating both *t*-scores and MI values and the cut-offs employed are those given by Barnbrook (1996), and Tables 1 and 2 show the figures which resulted. Currently there are no figures available for implementing *t*-scores and MI values for 3-, 4- and 5-word concgrams because these tests are designed to determine only the significance of the associations of two words. Table 1 shows that for the 2-word concgram list set at a span of 5 words, using an MI cut-off reduces the size of the no cut-off concgram list from 35,310 to 9,449 concgrams, and the *t*-score only cut-off resulted in the smallest list of the three with 2,959 concgrams. A similar phenomenon is observed for the 10-word span (Table 2). As for the 3-word concgram lists, the lists with no cut-off, set at 5- and 10-word spans, were similar in size: 385,746 for the 5-word span and 397,822 for the 10-word span.

Table 1. 5-word span – total number of concgrams plus overall concgram frequencies for 'c' words

Concgram list	With no cut-off (number of concgrams)	With no cut-off (total instances)	With *t*-score cut-off (number of concgrams)	With *t*-score cut-off (total instances)	MI cut-off (number of concgrams)	MI cut-off (total instances)
2-word	35,310	219,264	2,959	65,845	9,449	34,491
3-word	385,746	1,496,542				
4-word	138,926	365,064				
5 word	103,234	239,096				

Table 2 10-word span – total number of concgrams plus overall concgram frequencies for 'c' words

Concgram list	With no cut-off (number of concgrams)	With no cut-off (total instances)	With *t*-score cut-off (number of concgrams)	With *t*-score cut-off (total instances)	MI cut-off (number of concgrams)	MI cut-off (total instances)
2-word	55,888	372,989	3,751	81,033	12,650	43,807
3-word	397,822	1,521,550				
4-word	451,094	1,142,739				

5 word	855,740	2,101,963

Stubbs (1995) suggests that one way to use the MI and *t*-score cut-offs is to use them iteratively on the same list. In this case, the list should first be created with an MI cut-off of 3.00, and then sorted with a *t*-score cut-off of 2.00. Accordingly, ConcGram© also provides for these iterative cut-offs, and the same list was used as the basis for these iterative searches. In practice, the same list results whether the user starts with an MI cut-off, and then applies a *t*-score cut-off on the resulting list, or if the user starts with a *t*-score cut-off and then applies an MI cut-off to the list. The resulting list is the same and is shorter than that obtained by either MI cut-off only or *t*-score cut-off only (Table 3).

Table 3. Concgrams and frequencies for iterative search

Span	Number of concgrams	Total instances
5	1,456	15,848
10	1,869	18,998

6. Some limitations of *t*-score and MI tests

As is well-known, all statistical measures have their limitations, and those of the *t*-score and MI value are well documented (see, for example, Stubbs, 1995). Consequently, users of ConcGram© may wish to adopt a more 'purist' or 'unadulterated' approach to the concgram lists and not apply any statistical measure of significance to them. While this means that no collocational patterns are inadvertently dropped from the list, it also means the user is left to face very, very long lists to examine.

Table 4 shows the top words in the list with *t*-score cut-off set at 2.00. The *t*-score seems to place high significance on frequency of occurrence, whereas all of the top

ten collocates in the *t*-score sorted list would be discarded by an MI cut-off set to 3.00 (Table 5).

Table 4. 2-word concgrams with t-score cut-off set at 2.00

Single origin	Associated word	Instances of single origins	Instances of associated words	*t*-score	MI value
can	you	2005	2635	32.864986	1.474895
course	of	557	643	19.349775	2.077532
can	we	839	1060	18.719505	1.234347
can	I	1049	1433	16.754930	0.843237
can	see	342	352	14.195622	2.038772
come	to	416	573	13.528072	1.201388
can	it	735	838	13.237874	0.881759
continue	to	173	243	11.988043	2.114243
China	in	288	355	11.921243	1.445024
can	that	870	959	11.908827	0.700301

Table 5 below shows that the 2-word concgram list created with MI cut-off is very different from that created with *t*-score cut-off in that all the most frequent 'grammatical' words have been dropped from the list, leaving 'lexical' words such as 'coca/cola', 'connectors/misuse' and 'canal/flyover'.

Table 5. 2-word concgrams with MI cut-off set at 3.00

Single origin	Associated word	Instances of single origins	Instances of associated words	*t*-score	MI value
cola	coca	7	9	2.999980	17.221076
connecters	misuse	2	2	1.414197	16.373079
cure	counseling	2	2	1.414197	16.373079
canal	flyover	2	3	1.732024	15.958042
climbing	Angel	2	3	1.732024	15.958042
clockwise	anti	8	14	3.741594	15.858506
coca	Motorola	3	4	1.999959	15.565724
callipers	verniar	3	6	2.449432	15.373079
candidacy	endorse	2	2	1.414180	15.373079
cerebral	palsy	2	2	1.414180	15.373079

As shown in Table 6 below, using the two statistical tests together does not really help our purposes, as the MI values do not reflect the number of instances which receive high *t*-scores and they will be discarded. When both tests are used, the result is a list which is similar to that obtained by using MI cut-off alone.

Table 6 2-word concgrams with both t-score cut-off set at 2.00 and MI cut-off set at 3.00

Single origin	**Associated word**	**Instances of single origins**	**Instances of associated words**	***t*-score**	**MI value**
cola	coca	7	9	2.999980	17.221076
clockwise	anti	8	14	3.741594	15.858506
callipers	verniar	3	6	2.449432	15.373079
Conduit	Connaught	3	7	2.645618	14.273544
correctness	incorrectness	5	5	2.235910	13.788117
Chee	Hwa	25	25	4.999602	13.616056
curricular	extra	7	9	2.999760	13.612267
clogged	feedbox	7	10	3.162017	13.565724
Chau	Cheung	3	5	2.235878	13.525082
counsellor	sports	2	5	2.235847	13.302690

A list purely sorted on frequency of occurrence produces the following words at the top of the list (Table 7).

Table 7 2-word concgrams sorted by frequency of occurrence

Single origin	**Collocate**	**Frequency**
can	you	2,248
can	the	2,224
can	and	1,419
can	I	1,390
can	to	1,336
can	that	1,282
can	it	1,150
can	a	1,118
can	we	1,008

This list is much closer to that sorted by *t*-scores than to that sorted by MI values. In conclusion, for the purpose of studying a corpus of spoken English at least, we are reluctant to fully endorse either the *t*-score or the MI-value. Also, setting the span to 5 words, which may be the optimum value when studying a corpus of written language, is not the optimum value for spoken data as it misses concgrams.

7. Concgram examples from the HKCSE

Concgrams show both positional and constituency variation, which can be calculated and sorted by the computer. All of the examples given here have been sorted by configuration variation (-5 to +5 with a span set to 5). This means that the most common positional and constituency variations are immediately apparent, and the computer can thus calculate the configurations' frequencies which may be listed in a separate listbox. One of the reasons for a researcher, perhaps, not wanting to rely on statistically determined cut-offs is the risk of losing patterns of collocation that are of interest, due to the unreliability of the tests in determining meaningful word associations. An example of this is examined for the 2-word concgram 'alright/so' (Table 8). The list below gives some of the figures for collocates with 'alright'. These are all commonly found words, sorted by instances, and whereas 'so' gets the highest significance value with *t*-score of 4.26, and would not be discarded from a list with *t*-score cut-off set at 2.00, it only gets 0.48 with an MI test, and would be discarded.

Table 8. t-scores and MI values for 2-word concgrams with 'alright' as single origin

Single origin	Associated word	Instances of single origins	Instances of associated words	*t*-score	MI value
alright	the	303	453	-20.947807	-0.988567

alright	you	305	424	0.554295	0.039368
alright	to	233	318	-8.988157	-0.588834
alright	I	205	292	-3.256034	-0.251622
alright	and	194	249	-13.778091	-0.905467
alright	it	182	224	1.741316	0.178447
alright	so	192	224	4.265845	0.484033
alright	that	165	189	-4.937490	-0.442704
alright	a	150	184	-9.232223	-0.748984

Example 1 below shows that 'alright' and 'so' are associated in spoken English, although the MI value would result in it being discarded. Setting the span to 5 would also discard at least half the examples because the distance between 'alright' and 'so' is more than 5 words in 7 of the 19 instances.

Example 1: 2-word concgram 'alright/so'

```
1          more so this should be emphasized a lot ( ) alright ( ) and of course this
2       the middle so the customer would say I want this alright so we design this and
3        two doors so if you haven't yet had one get one alright pick up one on your way
4                  so you don't get that matching loan alright (everybody understand)
5           ar [( ) so ( ) so that's how it works B: [mm alright a2: mm B: mhm [so you
6             okay so so they are implemented correctly alright ( ) again we can we can
7       school teacher so that's phonetics and phonology alright erm morphology and
8         okay alright so we need to to do some more ( ) alright B061 a1: yeah a2: not
9        [so you can manipulate we can retry B:          [alright yea yea yea twenty four
10            ( ) how about SO is it a good strategy ( ) alright how about ( ) let's say
11      think you need to do so many things b:   [mm b: [alright y: [you y: what your
12         okay okay okay I see so you got some exposure alright a1: mm a1:  okay have
13               you will obtain so called the resonance alright this is the resonance
14       and it goes (inaudible) so what I saw like this alright so all these things can
15      Four Seasons Regent Hotel so you're um a: yea B: alright B: so um you're in your
16       sections i:  yeah sections so you see the links alright so cohesive I items
17          want to sleep b: yea B: so it's very hard b: alright B: okay signature b: yes
18       we ( ) sit in a group okay so that you can talk alright er let's let us have ( )
19          Personnel b:   okay B:   so that B070 (1) B: alright and you're um you are
```

The relationship between 'alright' and 'so' is an interesting one in spoken English. A typical example is on line 1. The speaker says 'so' to conclude that what he has been talking about has to be emphasized and then rounds off this section of the talk with 'alright' (after 6 intervening words) before continuing. Again, on line 2, the speaker begins a concluding comment with 'so' and ends the comment with 'alright' (after 7 intervening words). These two words are associated, and achieve an important discourse

organisational function. The nature of this function means that the positional variant 'so ... alright' is typically an indeterminate number of words apart, depending on the length of the speaker's comment, the start and end of which is marked by 'so' and 'alright'. Also, for those who would end searches for non-contiguous associated words at the end of sentences, or, in the case of spoken discourse, utterance boundaries, there is evidence of the association of 'alright' and 'so' across utterance boundaries. On lines 5 and 17, speakers begin concluding comments with 'so' and it is the hearers who supply 'alright' which serves to round off the comment by the previous speaker.

The following list of 2-word concgrams with 'call' as the single origin (Table 9), which has been sorted by *t*-scores, shows that the two highest scores are for the concgrams 'call/we' and 'call/what'. Both would be discarded on the criterion of MI values with a cut-off of 3.00.

Table 9. t-scores and MI values for 2-word concgrams with for 'call' as single origin

Single origin	**Associated word**	**Instances of single origins**	**Instances of associated words**	***t*-score**	**MI value**
call	we	107	124	8.010980	1.833450
call	what	85	89	7.825053	2.551767
call	it	77	87	5.561935	1.308652
call	a	73	91	3.787517	0.729865
call	this	46	49	3.352858	0.940589
call	I	74	93	3.247290	0.592328
call	you	71	84	1.177465	0.198381
call	and	70	81	-0.195501	-0.031003
call	is	46	48	-0.270278	-0.055211

As with the 'alright/so' 2-word concgram, some of the concordance results for the concgram 'call/what' (Example 2) contribute to an extended unit of meaning, even though based on the MI value criterion, it would also be discarded.

Example 2. 2-word concgram 'call/what'

```
1        his is the district work or what I would rather call the community work a lawyer
2         will go in for a few weeks into er what we now call industrial engineering and
3         know A:    [have you do you like er what do you call it I am going you know er
4             not actress ((laugh)) [yea ( ) what do you call this to write the scripts
5          to love that B:           [but er what do you call the er A: there's a certain
6        so I started er from er from I guess what you'd call white collar work I started
7         are slackening this together with er what am I call the Enron effect has
8         in this there is a lot of data about what they call emotional intelligence I
9                 [er which is er you know the what they call the er the new technology
10         there have been regular meetings er what they call at that time er er Cross
11         percent of the time so we want to set what we call ninety-five percent
12         we have to do is to try and determine what we call the optimum order cycle and
13       dominant values and let's just look at what we call the ( ) key themes or
14         ( ) now these has to be tied in with what we call the competitive advantage
15         content and context ( ) and visual er what I call behaviour a- acuity not now
16        through these bad times ( ) now this is what I call use the strength ( )
17       work through the China market place it's what I call a triple play that is
```

This is a good example of a concgram with only non-contiguous variation, with anything from 1 to 3 intervening words. Interestingly, despite the selection of 'what' to the left of 'call', most of the instances are not questions (lines 3, 4 and 5 are examples of 'what' and 'call' being co-selected to ask a question). The association of 'what' and 'call' seems to be mainly in relation to a speaker re-formulating what he/she has just said (e.g. lines 1, 2, 6, 7, 8, 9, 10, 15, 16 and 17), or to introduce something based on what has just been said (lines 11, 12 and 13).

Example 3 below shows all of the concordance lines for the concgram 'high/low', with 'high' as the single origin.

Example 3. 2-word concgram 'high/low'

```
1    proved my er hypothesis is correct it's um high proficiency students got better
results than low
2  low profit and low earnings and and not very high stock prices and that makes people
3 authority versus the low authority versus the high structure versus the low structure
4        people that's a low authority society a high authority society is just the
5  try and buy when it's low and sell when it's high otherwise doesn't matter how
6  they're taking advantage of the low cost and high quality of production facilities in
7a low individualism society or or or if you're high collectivist say Hofstede
8    okay can you (inaudible) ( ) individualism high and low individualism if you are low
9   because it is too erm the whatever it's too high and too [low then erm this is
10 ays and bad the great changes the moments of high peaks and low troughs but I always
11 lity to be able to feel that at all that's a high EQ person a low EQ person's one
12     period of over fifty months of deflation high unemployment low levels of consumer
13 ividual er relationships are emphasized in a high individ- in a low individualism
14   students in order to know whether they are high proficiency or low proficiency
15 tside and there's imbalance because it's too high and this is too low ( ) and that
16 del is applicable for any company dealing in high tech middle tech low tech and even
```

```
17 business model applies for any company doing high tech middle tech low tech and even
18individualist versus the collectivist between high authority versus the low authority
19  it's the group will take care (.) whereas a high individualism society or a low
20 wer-fixed correction if the voltage ratio is high in the case of the boost or low in
21 he buckle converter is not preferred for its high peak current especially when low
22 (laugh)) B: but it needs to doesn't [come up high enough a1: [well it's too low it
```

Example 3 shows that the concgram 'high/low' has both constituency and positional variations. All instances are non-contiguous. The positional variant 'low ... high' on lines 2-7 has between 2 and 7 intervening words. On lines 1 and 8-22, the other positional variant 'high ... low' has between 1 and 7 intervening words. Three uses are observed. First, speakers typically juxtapose points on a scale of 'high <--> low' presumed to be shared with the hearer, the item or attribute being juxtaposed include proficiency (line 1, 14), authority society (line 4), individualism (line 8, 13), EQ person (line 11), tech(nology) (line 16, 17), authority (line 18), and voltage ratio (line 20). Second, speakers present a relationship between two related items or qualities, for instance, 'low earnings' and 'not very high stock prices' (line 2), 'low cost and high quality of production facilities' (line 6), 'low individualism society' and 'high collectivist' (line 7), 'high peaks and 'low troughs' (line 10), 'high unemployment and low levels of consumer confidence' (line 12), and 'high peak current' and 'low voltage ration' (line 21). The last usage is that the concgram of 'high/low' extends across two speakers, and is an example of paraphrasing in which one speaker's 'doesn't come up high enough' is another speaker's 'it's too low' (line 22).

Table 10 shows that both 'low/high' and 'high/low' ca be said to be significant concgrams, based on the *t*-scores and MI values.

Table 10. t-scores and MI values for 2-word concgrams with 'high' or 'low' as single origin

Single Origin	Associated word	Instances of single origins	Instances of associated words	*t*-score	MI value
low	high	22	23	4.604069	4.644390
high	Hong	27	31	2.196460	0.723791
high	individualism	4	5	2.217049	6.877384
low	individualism	5	6	2.438839	7.845342
low	inventory	3	5	2.180647	5.334380
high	is	79	100	2.810706	0.476078
high	it	49	72	2.518581	0.508029
low	it	36	50	2.678560	0.686883
high	Kong	27	31	2.394098	0.810948
low	labour	7	7	2.590283	5.575881
high	level	20	21	4.247325	3.772848
low	level	11	13	3.344152	3.785893
high	levels	6	6	2.275872	3.818490
high	low	22	24	4.711255	4.705790

Example 4 below shows all of the concordance lines for the concgram 'correctness/incorrectness', with 'correctness' as the single origin.

Example 4: 2-word concgram 'correctness/incorrectness'

```
1 the percentage of incorrectness is higher than correctness um but er you will find that
2  test one it is the overall result to show the correctness and incorrectness done by er
3  look up the meaning rather than usage and the correctness and incorrectness er
4  got more correct than incorrect got er answer correctness more than incorrectness so
5     um but it you can see that um test one the correctness is higher than incorrectness
```

The concgram of 'correctness/incorrectness' has both constituency and positional variations. All instances are non-contiguous. The first positional variant is 'correctness ... incorrectness' (lines 2-5) with from 1 to 3 intervening words; the second is 'incorrectness ... correctness' with only one example (line 1) with 3 intervening words. All of the instances come from the same lecture in which the two members of this concgram (i.e. 'correctness' and 'incorrectness') are associated in a relationship of antonymy. Table 7 shows that 'correctness/incorrectness' has a high MI value (13.79) and a *t*-score above the 2.0 cut-off point (2.23).

Example 5 is a 3-word concgram 'challenges/facing/we' with a double origin 'challenges/facing'.

Example 5: 3-word concgram 'challenges/facing/we'

```
1       at the moment we are facing tremendous challenges as our economy grapples with
2        at the moment we're facing tremendous challenges as our economy grapples with
3            based economy we are facing major challenges indeed difficulties may be with
4         we plan to overcome the numerous new challenges facing us but before I launch
5       I think that we're now looking at many challenges facing business community and
6 we are doing to tackle the difficulties and challenges facing us to lay the foundations
7       the one issue and that is the economic challenges facing Hong Kong and what we're
8        them what we're going to do about the challenges we're facing and where our
9      prolonged (pause) let me go back to the challenges we are facing in Hong Kong on
10   talking about I I think the the the most challenges that we're facing for FB
11     and shared pain to really resolve the challenges that we are facing these
12          er the both the opportunities and challenges that er er we are facing er
```

This concgram illustrates that while the association of 'challenges' and 'facing' is well-known, variation exists. When 'facing' precedes 'challenges' (lines 1-3), the two words are non-contiguous (although a larger corpus would probably find a contiguous variation). When the positions of 'challenges' and 'facing' are reversed (lines 4-12), there is constituency variation: contiguous (lines 4-7) and non-contiguous (lines 8-12). The associated word 'we' occurs in different positions in relation to 'challenges' and 'facing': 'we/facing/challenges' (lines 1-3), 'we/challenges/facing' (lines 4-8), and 'challenges/we/facing' (lines 9-12). The number of intervening words ranges from 1-7.

As yet, ConcGram© has no inbuilt statistical measure to determine the significance of 3-, 4- and 5-word concgrams. However, below examples of 4-word and 5-word concgrams are discussed. Figure 3 shows the frequencies of the 4-word concgram list for 'case/the/is/this', with 'case/the/is' as the triple origin.

[Insert Figure 3: The 4-word concgram 'case/the/is/this']

Example 6 4-word concgram 'case/the/is/this'

1 rent (pause) the rent (.) is the killer in this case the rent (.) which is a
2 passage er so er er ( ) this is the most special case and um the others are not
3 (.) you're supposed to prepare the (inaudible) case and this is the er case
4 but if this is the case this is the genuine case for the whole industry ( )
5 your parents they won't do it it's the reverse case so if this is the case look
6 prepare the (inaudible) case and this is the er case that I would like you to
7 only in Hong Kong b2: well this is always the case and and I'm sure er Mister
8 it's constant and in reality this is seldom the case we have to assume that the
9 the way back down to here and this is often the case and if you if you look at a
10 the room I don't want to (inaudible) is this the case (pause) do we do we agree
11 and desires will be done erm (.) if this is the case (.) I just use the McDonald
12 shops in those (inaudible) okay if this is the case now the rent kill them (.)
13 do it it's the reverse case so if this is the case look at the economy it's a
14 is not impeded ( ) to ensure this is the case we propose to narrow the
15 so you are just wondering whether this is the case (.) alright er okay now let
16 job ( ) if (.) if and only if it this is the case if they erm of you go back
17 real child ((laugh)) and certainly this is the case [of the supreme court [(.)
18 cash flow problems for Yaohan but if this is the case this is the genuine case
19 the impression and I hope very much this is the case that when I was in Japan er
20 input and the output which is actually in this case the sine square minus the P
21 factor is rent now if rent is is true in this case the question come up with
22 share prices have risen this is a classical case of the end justifying the
23 of heat resistant material er in er in this case er the water is er
24 like to use so you can say person people in that case this is the plural form of
25 case you have to do this (.) now this is a local case when we have the case those
26 could be if you talk about competitors in this case that that is the threat
27 b: mhmm A: erm ( ) I mean in fact in this case what we're after is the

In Example 6, the positional variant of 'this/is/the/case' is the most prevalent with 15 instances. Contiguous variants are found on lines 11-19 and non-contiguous ones on lines 2, 4, 6, 7, 8 and 9. In all of these examples, 'case' is modified (e.g. 'this is the most special case' on line 2).

Another case of a 4-word concgram, 'come/to/the/we', with 'come/to/the' as the triple origin, is found in Figure 4.

[Insert Figure 4: The 4-word concgram 'come/to/the/we']

Example 7 below shows all of the instances of the concgram 'come/to/the/we' with 'come/to/the' as the triple origin.

Example 7. 4-word concgram 'come/to/the/we'

1	should organize events to attract the public	to	come	to Hong Kong [we should
2	[( ) mm and those are the things that have	to	come	first then we are open-
3	very difficult decision for the government	to	come	to but faced as we were
4	well-educated er people from the mainland	to	come	can we er let them come er
5	that before we get people from the outside	to	come	[we've got to build up
6	funny question would be do we have the people	to	come	to Hong Kong a: er b:
7	we would continue to do so er in the months	to	come	er we also er in order to
8	success is the sign of even better things	to	come	in the future we'll see an
9	er Commissions and so on then we will be able	to	come	up with I think the best
10	time for each of the sub-categories and when we		come	to that statistical
11	that we can meet the target now we have		come	to the most interesting
12	ranges okay (.) this is the range and then we		come	to the hysteresis (.)
13	er have access to er health care b2: we have er		come	to the conclusion that um
14	((applause)) B1: thank you Missus Chan er we now		come	to the time when we take
15	are being renovated at the moment b: oh B: we've		come	to the end b: no no no I
16	(MD AC9 (55')) a1: er okay very quickly we		come	to the concepts of langue
17	we can prevent the wastage (pause) okay now we		come	to the billing system (long
18	have use it for our own use and ( ) and now we		come	to the selling checklist
19	well thank you and um ((applause)) B1: and we		come	to the last speaker in the
20	know the problems worldwide and then before we		come	to the Pearl River Delta
21	er perfective aspect (pause) b2: and then we		come	to the mortho- er
22	you are receptive then that (inaudible) okay we		come	to the next group which is
23	a woman [much more [((laugh)) feminine if we		come	to the second point if a
24	er perfective indicator of Chinese (.) and we		come	to the previous studies and
25	sum game if conducted appropriately we have		come	to a stage where the
26	have presented this snapshots of how far we have		come	not to blow the trumpet for
27	need to change all the er compo[nents then [we		come	back to B:
28	later so let's look at the tasks first and we'll		come	back to this mm (
29	slide please (.) so now we talk about now we		come	back to the macro- scope of
30	the intender and through the intender while we		come	into this I'm going to

Examination of the lines in Example 7 reveals several positional variants, namely 'we/come/to/the' (15 instances), 'the/we/come/to' (6), 'the/to/come/we' (6), 'we/the/to/come' (1), 'to/come/the/we' (1), and 'we/to/come/the' (1). The predominant variant is therefore 'we/come/to/the', with 9 contiguous (lines 16-24) and 6 non-contiguous (lines 13-15, 25, 26 and 29) instances. The non-contiguous variant can be quite complex (e.g. 'we have come not to blow the trumpet of ...') with none of the words being adjacent to one another. All of the instances of 'the/to/come/we' are non-contiguous. Example 7 serves to underline the power of the fully automated search engine to reveal a full range of constituency and positional variations.

Figure 5 shows the instances of the 5-word concgram 'can/you/I/know/mean'.

[Insert Figure 5: 5-word concgram 'can/you/I/know/mean']

Below, Example 8 also illustrates the desirability of setting the span higher than 5, at least when studying spoken English data.

Example 8. 5-word concgram 'can/you/I/know/mean'

```
1          job do you know what I mean [so that you can do it for twenty minutes but
2      ( ) remem- you know what I mean so like ( ) S can be add to verb ( ) ING for
3      cause I once you know the difference then you can work on the similarities you see what I mean y:
4      nd then like you just spend it I mean you you can put towards your hi-fi you know the your your
5      as the singu- you know what I mean ( ) so you can say mainly sheep so in a
6        between them you know B: alright but I mean can we now that we've got this
7       how about it B: [you know B: well you've you can -s you I mean you you are
8      company or I mean or you can you can well you can you can say you know the the
9     I absolutely know what you mean I have you you can benefit in two ways one way
10              [yea history you can't [you know you can't blame anyone I mean this is this is er I
11 goody she's such a sweet [you know I mean if you can't ride  a:                [mm
12      y: uhuh b: and I I mean you know where where can all these people get the
13     ten minutes a: yea A: ( ) but you know er she can't I mean ( ) you know
14    control a big company or I mean or you can you can well you can you can say you know the the the
15 is you know  B:                  [yea history you can't [you know you can't blame anyone I mean this
16         to control a big company or I mean or you can you can well you can you can say you know the
17      a2: yea right I mean the er the er let's can can I [you know deal with Sing-
18    China a2: yea right I mean the er the er let's can can I [you know deal with
19    course language learning can't I mean language can't be learnt in you know sort of it's not like
20  oh well yea we ate there yesterday B1: I I just can't eat that mu- much you know what I mean B2:
21      B: you know what I mean I mean sometimes you can do something outrageous just to give people
22      I know there are I mean at least the pitches can be outdated already so have you checked
23  a1: so what I'm looking at is ( ) suffixes that can add to adjectives remem- you know what I mean
24    some kind of  B:                [yea yea no it can no yea I I know  privacy what you mean ( ) yea
25 I mean [you shouldn't have it you know [( ) oh I can't believe it y: [((laugh))          [oh that
26    mistakes b: alright ( ) I don't know that er I can't believe it yea [( ) but I mean now you are
27      a short time and of course language learning can't I mean language can't be learnt in you know
```

The 5-word non-contiguous concgram 'can/you/I/know/mean', with the quadruple origin 'can/you/I/know', would not be found with a span set to 5. Two positional variants are found in this concgram, the most common of which is 'I mean ... can ... you know' (lines 4, 8, 14, 16, 17, 18, 19, and 25). In all of these, the speaker introduces a suggestion with 'I mean' which contains 'can', and later in the utterance says 'you know'. This co-selection of 'I mean' and 'you know' has the effect of drawing the hearer closer to agreeing to the suggestion and acts as an appealer. Again, this shows the advantage of using a wide span to capture non-contiguous variants such as this. A lesser positional variant is seen on lines 1, 2 and 5, with 'you know what I mean' followed by 'so ... can'. This variant is perhaps less interesting as much of it is a well-known contiguous variant.

8. Conclusions

This paper has described and defined a new way of identifying and categorising word associations, the concgram, which is all of the permutations of constituency variation and positional variation generated by the association of two or more words. The concgrams of a corpus are preferably identified and generated without prior input from the user, other than to set the size of the span, as it is only a fully automated concgram search that can reveal all of the possible collocational patterns that exist in a corpus.

Studying concgram search results can reveal word associations in a way that other searches do not. In the case of the latter, attention is primarily drawn to the user-nominated node word(s), a popular and traditional starting point for corpus queries which is replaced by the notion of 'origin' in concgram searches where the focus of attention is on word associations and their constituency and positional variations. Concgram searches begin with an origin (single, double, triple or quadruple) and have the central aim of uncovering the phraseological patterns in the language.

Preliminary searches on the one-million-word HKCSE have found that the majority of concgrams seem to be made up of non-contiguous collocations, and show both constituency (AB, ACB) and positional (AB, BA) variations which can be calculated and sorted by frequency. Although contiguous collocations are also found in concgram searches, since many collocational patterns never occur contiguously, searches which focus on contiguous collocations present an incomplete picture of the word associations that exist. Many concgrams reveal patterns of collocation which would not have been uncovered, relying on intuition alone or other search engines.

Concgram searches, by their very nature, emphasise the prevalence of word associations in language use, and diminish the attention that may be unduly paid to the node word(s) in user nominated queries in KWIC display. Such searches, we believe, will aid corpus linguists, and others in related fields, to uncover the full extent of the idiom principle (Sinclair, 1987).

Aside from the above major conclusions, we also have outlined our reservations relating to the use of *t*-scores, MI values, and a combination of these two measures. Future studies of concgrams need to bear in mind these reservations and may prefer to use the original concgram results without the intervention of statistical tests.

Acknowledgements

The authors are very grateful to John Sinclair for his many insightful comments on earlier versions of the paper which proved to be invaluable. Problems which may persist in the paper are very much our own. 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

Barnbrook, G. (1996). *Language and Computers: A Practical Introduction to the Computer Analysis of Language*. Edinburgh: Edinburgh University Press, pp 88-106.

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

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

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

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

Fletcher, W. H. (2006) "Phrases in English" Home. Retrieved 15 February 2006, from /http://pie.usna.edu/.

Greaves, C. (2005). Introduction to ConcGram©. Tuscan Word Centre International Workshop. Certosa di Pontignano, Tuscany, Italy, 25-29 June 2005.

Halliday, M.A.K., Teubert, W. and Yallop, C. (2002). *Perspectives in Lexicology and Corpus Linguistics*. London: Continuum.

Hoey, M. (2005). *Lexical Priming. A new theory of language*. London: Routledge.

Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

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

Louw, B. (1993). Irony in the Text or Insincerity in the Writer? The Diagnostic Potential of Semantic Prosodies. In M. Baker, G. Francis and E. Tognini- Bonelli (eds) *Text and Technology: In Honour of John Sinclair*, (pp. 157-176). Amsterdam/Philadelphia: John

Benjamins.

Partington, A. (1998). Patterns *and Meanings*. Amsterdam: John Benjamins.

Sinclair, J. McH. (1987). The nature of the evidence. In J. McH. Sinclair (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. (2004). *Trust the Text*. London: Routledge.

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

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

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

Stubbs, M. (1995). 'Collocations and semantic profiles: on the cause of the trouble with quantitative methods', *Functions of Language* , 2(1), pp. 23-55.

Stubbs, M. (2001). *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford: Blackwell.

Teubert, W. (ed). (2005). *Corpus Linguistics*-Critical Concepts in Linguistics. London: Routledge.

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. Certosa di Pontignano, Tuscany, Italy, 31

To appear in International Journal of Corpus Linguistics

June – 3 July 2005.