ConcGram 1.0

a phraseological search engine

by Chris Greaves

Dedicated to the memory of John McHardy Sinclair, a great teacher, colleague and friend whose ideas, suggestions and support were invaluable in the design and development of this program.

Contents

CHAPTER 1

ConcGram List Builder

CHAPTER 2

Selecting a text or corpus to interrogate

CHAPTER 3

Getting started with ConcGram

CHAPTER 4

Managing the lists

CHAPTER 5
Using the Configuration List Boxes

CHAPTER 6
Automatic searches from specified words

CHAPTER 7
Statistical tests

CHAPTER 8
User nominated searches with ConcGram

APPENDIX
ConcGram Tutorial

Acknowledgement

The organisation of this manual reflects the basic design principle behind the software. The original idea for ConcGram was that it would be the first program able to identify up to five co-occurring words, irrespective of either constituency or positional variation, in a text or corpus fully automatically. The automatic identification of concgrams remains its main distinguishing feature and explains why this key function is the first to be covered in the manual. ConcGram has many other functions, in addition to this central one, and these are explained later in the manual. ConcGram also has a comprehensive help file which can be accessed via the program.

I would like to acknowledge the input of my co-researchers in the process of writing the program and here John Sinclair deserves a very special mention. John was a very good friend over very many years and he was tremendously supportive of ConcGram, concgramming and concgrams. Many of the terms used in the program, and many of its functions, came from John and his support was very important at every stage of its development. I am also very grateful to my friends and colleagues Elena Tognini-Bonelli, Winnie Cheng and Martin Warren for their input and support.

The work that led to the design of ConcGram was substantially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region (Project Nos. G-YF39 and PolyU 5459/08H) and the Siena University Project:

TITLE: The pragmatic and phraseological dimension of topic "aboutness": Key-words and the intercollocation of collocates in the language of business and economics.

PRINCIPAL INVESTIGATOR: Elena Tognini Bonelli.

FUNDED JOINTLY BY: The University of Siena and the Ministry of Education, University and Research (Year 2005 – prot. 2005107939_005).

I hope you find the program a useful resource and I look forward to receiving users' feedback.

Chris Greaves

Introduction

Why concgram?

Introduction

This manual aims to assist the user in getting the most out of a highly innovative corpus linguistics software program: ConcGram. The obvious question to ask of any program is: 'What can it do that other programs can't do?' and, if the program really does have different functions, then the next question is: 'So what?'. This introduction aims to answer these questions by outlining how ConcGram can make a significant contribution to the ongoing search for a better understanding of phraseology, especially phraseological variation, which is at the heart of Sinclair's (1987) idiom principle.

Background

The idea to explore phraseological variation by means of concgramming can be directly traced back to the work of John Sinclair (1996 and 1998) in which he further expounds his idiom principle and describes the model of the five categories of co-selection which comprise a lexical item:[1] the obligatory semantic prosody and invariable core word or words, plus the optional collocates, colligates and semantic preference. Back in 2005, a small team of researchers[2] based in Hong Kong was interested in identifying lexical items in corpora, but we were faced with a major problem — how could we fully automatically retrieve the co-selections which comprise lexical items from a corpus? In other words, how could we identify lexical items without relying on potentially misleading clues in single word frequency lists, or lists of n-grams (contiguous groupings of words, sometimes termed clusters or bundles), or some form of user-nominated search? We felt that single word frequencies are not a reliable guide to frequent phraseologies in a corpus, and we were concerned that n-grams miss countless instances of phraseology that have constituency (AB, A*B, A**B, etc.) and/or positional (AB, BA, B*A, etc.) varia-

[1] Sinclair later adopts the term 'meaning shift unit' (2007a) in preference to 'lexical item'.

[2] See Cheng, Greaves and Warren (2006).

tion. We decided that what was needed was a program that could identify all of the co-occurrences of two or more words irrespective of constituency and/or positional variation in order to more fully account for phraseological variation and provide the raw data for identifying lexical items and other forms of phraseology. In addition, we wanted the program to be able to identify these co-occurrences fully automatically in order to support corpus-driven research (Tognini-Bonelli, 2001), by making it unnecessary for any prior search parameters to be inputted by the user. Luckily, Chris Greaves, who is well-known for his corpus linguistics programs,[3] is a member of the team and he took on the difficult task of designing and implementing such a program. It was suggested that Chris should wait until after I had spoken with John Sinclair at the AACL/ICAME conference at Michigan University in May 2005. John was very supportive of the idea and, in fact, he had tried to develop just such a program in the 1980s. Greatly encouraged, I returned to Hong Kong to discover that Chris had already written a prototype program which he had named ConcGram (the products of the searches are, of course, termed 'concgrams'). Those familiar with Chris' work ethic and programming expertise will not be surprised that he had developed a prototype so quickly. Importantly, from this point on, John was very much a part of the team working on both the program and how to analyse the outputs (see, for example, Cheng, Greaves, Sinclair and Warren, in press, 2009). Many of the functions, terms and concepts used in the program, and in our analyses of concgrams, are thanks to John, and we will always be extremely grateful to him for sharing his ideas with us. The team also grew with the addition of Elena Tognini-Bonelli who, since 2005, has been involved in two research projects[4] looking at ways to determine the aboutness of texts by analysing both keywords and phraseology which have resulted in the exploration of new applications for concgramming and concgrams, and some of this work is briefly described below.

What can ConcGram do?

ConcGram was specifically designed to fully automatically uncover instances of phraseology, where phraseology is defined as the co-selection of words. Accepting the default setting for the identification of concgrams, comprised of between two and five words, means that the program will find *all* word co-occurrences within a limit set by the user. The default is 50 characters (i.e. approximately 12 words) either side of the centred word

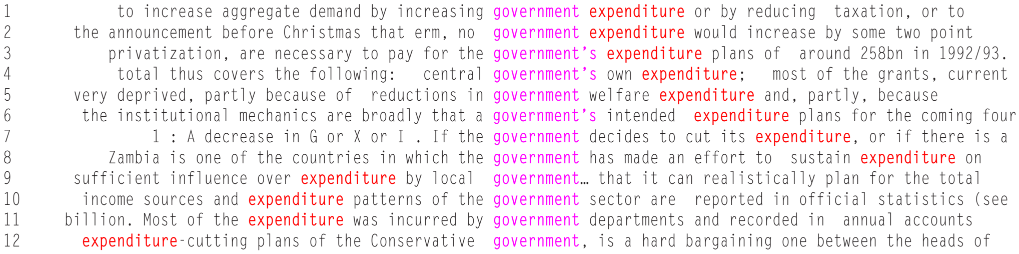
[3] Chris has written, for example, ConcApp, iConc and ecConcord.

[4] These projects, financed by the Italian Ministry of Education and Siena University, contributed to and helped to sponsor the development of ConcGram to identify 'aboutness' in a text under John Sinclair's guidance.

fully automatically. This is a unique attribute of ConcGram which makes the program truly corpus-driven as there is no prior intervention by the user of any kind. Using ConcGram in this unfettered mode guarantees a comprehensive search for all word co-occurrences which can then be listed and the concordances examined. It is important to note that in its default mode, ConcGram finds the co-occurrences of words in a wide span, and not all of these instances are necessarily meaningfully associated. As a result, we have found it useful to distinguish between 'co-occurring' words (i.e. concgrams) and 'associated' words (i.e. phraseology). In other words, concgrams are objective, automatically generated data which then need to be interpreted as being meaningfully associated or not. At this stage, we think that phraseological associations can be grouped under three main categories: meaning shift unit, collocational framework and organisational framework. These categories are described briefly later.

In order to convey the uniqueness of ConcGram's outputs, it is helpful to look at some concgram concordance lines. A sample of concordance lines for the two-word concgram 'expenditure/government'[5] is given below and it illustrates the potential of ConcGram to uncover the full range of phraseological variation.

Figure 1. Sample concordance lines for 'expenditure/government'[6]



```
1           to increase aggregate demand by increasing government expenditure or by reducing  taxation, or to
2        the announcement before Christmas that erm, no  government expenditure would increase by some two point
3            privatization, are necessary to pay for the government's expenditure plans of  around 258bn in 1992/93.
4             total thus covers the following:   central government's own expenditure;   most of the grants, current
5        very deprived, partly because of  reductions in government welfare expenditure and, partly, because
6         the institutional mechanics are broadly that a government's intended  expenditure plans for the coming four
7                 1 : A decrease in G or X or I . If the government decides to cut its expenditure, or if there is a
8            Zambia is one of the countries in which the government has made an effort to  sustain expenditure on
9        sufficient influence over expenditure by local  government… that it can realistically plan for the total
10        income sources and expenditure patterns of the government sector are  reported in official statistics (see
11      billion. Most of the expenditure was incurred by government departments and recorded in  annual accounts
12        expenditure-cutting plans of the Conservative  government, is a hard bargaining one between the heads of
```

The display of concordances for concgrams is designed to be very reader-friendly. The various concgram configurations for 'expenditure/government', when 'government' is centred, begin with contiguous words to the right, and subsequent lines show instances ordered according to the distance between the words in the concgram. Once all of the instances to the right of the centred word have been displayed, those to the left are displayed following the same principle. How best to display concordance lines is crucial for a program specifically designed to uncover all instances of phraseology irrespective of variation because the variation uncovered needs to be displayed in a way that makes

[5] Concgrams are written alphabetically separated by a forward slash.

[6] All the examples of concgrams presented here are from a five-million word sample of the British National Corpus (three-million written and two-million spoken).

it manageable for the user to then study. On the computer screen, the use of different colours to represent each of the words in a concgram is another significant display feature which helps the user to identify them instantly.

An important feature of ConcGram, which becomes apparent when one first views the concordance for a concgram, is that concgrams represent a serious challenge to the current view about word co-occurrences that underpins the KWIC[7] display. Studying KWIC displays, which only highlight the node (i.e. the centred word), has unintentionally created, in the minds of some users, a hierarchical approach which puts the node as the centre of attention and the words associated with the node as subordinate to it. It is worth restating the point made by Sinclair, et al. (1970: 10), that while 'node' and 'collocate' are convenient terms to use, the term 'node' does not imply a hierarchical relationship between a node and its collocate, and that a node word which has a collocate becomes a collocate itself when the collocate is selected as the node.

By not simply focusing on the node, ConcGram highlights all of the co-occurring words in a concgram in each concordance line. This unique feature then has the benefit of shifting the user's focus of attention away from the node to all of the words in the concgram. Thus, word co-occurrences become the focus of study. It is for this reason that the term 'origin' is used for the word or words which are the source of automated concgram searches in order to emphasise the important difference between ConcGram and KWIC displays. Since ConcGram, in its fully automatic mode, begins by finding all the two-word concgrams, and then builds up iteratively to five-word concgrams, the notion of a 'node' is redundant. Instead the notion of 'origin' (one-word, two-word, three-word or four-word) is used to better foreground the central design feature that co-occurring words are the target of every search. For purely display layout purposes, the on-screen view of concgram concordance lines requires a sort-point simply to present a visually intelligible page, but a simple click on the 'switch centred word' button enables the user to centre any word in a concgram.

One important point to be borne in mind is that when you use ConcGram in the fully automatic default mode, as described above, you are setting the computer a complex computing task, and this may take some time to accomplish depending on the specifications of your computer and the size of the corpus.[8] For those who want to use ConcGram in user-constrained search modes, you will find that ConcGram has a wide array of options for users to limit their searches for concgrams (see Chapter 4), such as selecting a smaller span size, using an exclusion ('stop') list, setting a cut-off based on frequency and so on.

[7] KWIC = key word in context

[8] Approximations on computing times are provided in the manual (see Chapter 1 paragraph 4).

With ConcGram you can also conduct user-nominated searches for particular combinations of up to five words which by-pass the fully automatic generation of concgrams lists. In addition, ConcGram has all of the functions usually associated with traditional corpus linguistics software, such as the generation of word frequency lists, the determination of the specificity (i.e. 'keyness') of single words (plus two-word concgrams), the generation of single origin concordances, mutual information values, t-scores, and so on. Lastly, the program should be able to handle any language which has spaces as word delimiters, and it has been used successfully by colleagues working with German, Italian, and Spanish corpora.

What can we learn from the study of concgrams?

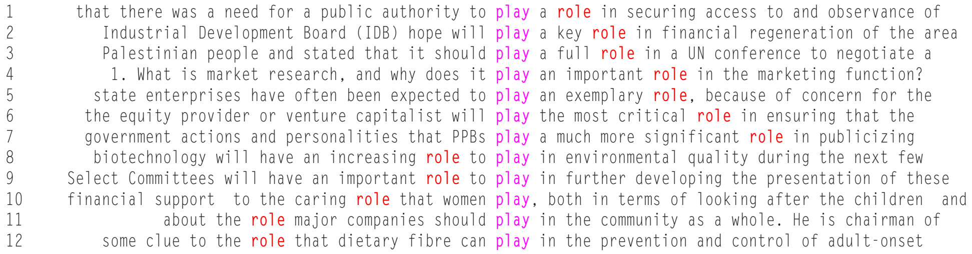
Having outlined the significant ways in which ConcGram adds to a corpus linguist's software resources and corpus analysis methods, here I will make the case for concgramming by summarising some of the ways in which it is currently being put to use, namely analysing phraseological variation and phraseological profiles of texts and corpora.

To date, three major categories of phraseology, which were originally set out by Sinclair (Greaves and Warren, 2008; Sinclair and Tognini-Bonelli in press), are beginning to be more exhaustively identified and described thanks to ConcGram. These are meaning shift units, collocational frameworks and organisational frameworks. Congramming has also been found to be a very useful means of identifying the phraseological profile of texts and corpora and hence their aboutgrams.

Meaning shift units

As mentioned above, the driving force behind the design of ConcGram was that it should enable corpus linguists to more fully identify and describe meaning shift units (MSUs) (Sinclair, 2007a) which Sinclair formerly termed 'lexical items'. Central to a description of phraseology is the identification of MSUs (Sinclair 2007a and 2007b) and Cheng et al. (in press, 2009) outline an analytical procedure for handling concgrams which can lead to their identification. In the latter study a two-word concgram, 'play role', is analysed and a sample of its concordance lines is given below.

Figure 2. Sample concordance lines for the two-word concgram 'play/role'



```
1       that there was a need for a public authority to play a role in securing access to and observance of
2          Industrial Development Board (IDB) hope will play a key role in financial regeneration of the area
3          Palestinian people and stated that it should play a full role in a UN conference to negotiate a
4           1. What is market research, and why does it play an important role in the marketing function?
5         state enterprises have often been expected to play an exemplary role, because of concern for the
6        the equity provider or venture capitalist will play the most critical role in ensuring that the
7        government actions and personalities that PPBs play a much more significant role in publicizing
8         biotechnology will have an increasing role to play in environmental quality during the next few
9      Select Committees will have an important role to play in further developing the presentation of these
10     financial support  to the caring role that women play, both in terms of looking after the children  and
11                about the role major companies should play in the community as a whole. He is chairman of
12         some clue to the role that dietary fibre can play in the prevention and control of adult-onset
```

All of the concordance lines of 'play/role' are studied by Cheng et al. (in press, 2009) and all of the possible concgram configurations and their frequencies of occurrence are identified. Based on frequency, the canonical form is identified and its meaning described (in the above sample the canonical form is exemplified in lines 2–5). The canonical form then becomes the benchmark for all of the other concgram configurations, and the end result is a ranking of the concgram configurations relative to their adherence to the canonical form. At the end of this process, an MSU is identified and described along with its potential variations which together comprise a paraphrasable family with a canonical form and different patterns of co-selection.

Collocational frameworks

Just as so-called 'grammatical' words dominate single word frequency lists, so co-occurrences of these same words dominate concgram frequency lists. Renouf and Sinclair (1991) call the co-selections of grammatical words 'collocational frameworks' and, despite the prevalence of this form of phraseology in the language, they still remain under-researched. The use of ConcGram makes the study of collocational frameworks much easier because the constituency and positional variation typically exhibited by them present no problem for the software. Initial studies of concgrams in pursuit of collocational frameworks (Greaves and Warren, 2008; Li and Warren, 2008) show that the five most frequent are 'the … of', 'of … the', 'in … the', 'a/an … of', and 'the … of … the' in a five-million word sample of the British National Corpus (Li and Warren, 2008). Examples of the top two are given below.

Figure 3. Sample concordance lines for the two-word concgram 'of/the'

```
1            a Japanese motor manufacturer by a member of the public (ASA Case Report 119, 1985). In this
2              <u who=PS1HH> Ah!  <u who=PS000> First of all the [unclear] coming down.  <u who=PS1HH>
3       she had a house to start with. When you think of all the families on the waiting list <u
4       that’s just enforcing  just merely a question of enforcing the law. But if you turned round and
5     serious  implications as to what we do in terms of improving the highway infrastructure  at
6    to get there)  thoroughly. There is no other way of anticipating the potential and the problems or
7      careful planning, especially in the allocation of committee rooms  for use by committees which are
7    law. Sections “B” and “C” outline the provisions of the civil and  criminal law respectively and
9       gradually increased, including the resumption of coffee cultivation but on a  much more limited
10     of advice in recent years about the importance of a healthy diet,  according to a report by market
11    secondary, day or boarding.   The vast majority of schools are “government-aided” and are run by
12   shareholders as dividends then the capital base  of the business has been eroded.   However, under
```

The widespread use of collocational frameworks such as those in Figure 3 suggests that they deserve greater attention from researchers, teachers and learners. As long ago as 1988, Sinclair and Renouf argued that they should be included in a lexical syllabus, but to date they remain overlooked. Currently, new grammars have begun to list and describe n-grams (clusters or bundles). For example, Carter and McCarthy (2006: 503–505) list four-word clusters in written texts, and the list includes the following instances: *the end of the*, *the side of the*, *the edge of the*, *the middle of the*, *the back of the*, *the top of the*, and *the bottom of the*. Now that we have ConcGram, which is able to more fully uncover collocational frameworks, these n-grams can be preceded by a description of the three-word collocational framework common to them all, i.e. *the * of the*. Lists solely comprised of n-grams will also need to be expanded to include instances of phraseological variation uncovered by ConcGram.

Organisational frameworks

Hunston (2002: 75) briefly describes what she tentatively terms 'clause collocation' which refers to the tendency for particular types of clause to co-occur. She provides one example of a clause collocation, 'I wonder … because', where 'I wonder' and 'because' function to link clauses in the discourse (ibid: 75). She also notes that such collocations are hard to find because the size of the 'I wonder' clause is indeterminate (ibid: 75).

Adopting the distinction between organisation-oriented elements and message-oriented elements used in linear unit grammar (Sinclair and Mauranen, 2006), Greaves and Warren (2008) term this form of phraseology an 'organisational framework' to denote the ways in which organisational elements in the discourse, such as conjunctions, connectives and discourse particles, may be co-selected by speakers and writers. ConcGram makes the study of these organisational frameworks possible because of its ability to retrieve co-occurring words across a wide span. Sample concordance lines of the organisational framework 'because/so' are shown in Figure 4 below.

Figure 4. Sample concordance lines for the two-word concgram ‘because/so’

```
1     assume that it’s constant it will be Q over two because that takes you half way up so the holding cost
2      small pixels the colour doesn’t change any but because the size is become smaller so the resolution
3    didn’t get any nasal pharyngeal aspirate anymore because the nurses refuse to do it so we in- instead we
4       both by nominative and nominative case (.) er because this is the accusative (.) so this is ruled out
5    appropriate for the items that you’re looking at because they won’t all be the same so we can attempt try
6     that question I think it’s highly complementary because China is such a big country so we’ll be doing
7      to Admiralty that often so that’s okay with me because I don’t have to see it that often erm but my
8    whenever you can do it so reciprocity isn’t rare because it’s Asian but blue is m- is a more western
9    types of situations so we deal with these things because there’s a very good chance that you’ll be caught
10    River Delta cities so you’re welcome to join us because I hope you would find it useful to you in terms
11     the tourist group so he goes as fast as he can because the dolphins are like chasing the boat [(.) and
12    citizen’s budget so they didn’t get into Todd’s because they thought the prices were expensive (.) P_ R_
```

Some instances of organisational frameworks are well-known, and are sometimes listed in grammars as ‘correlative conjunctions’, for example, ‘either … or’, ‘both … and’ and ‘whether … or’ . There are others, however, such as ‘because/so’ in Figure 4 and Hunston’s ‘I wonder … because’, which are not so familiar, or are even currently unknown, which deserve more attention. It is also of interest to note that the organisational framework ‘because/so’ exhibits not only constituency but also positional variation which, again, ConcGram is particularly well-suited to uncovering.

Phraseological profiles and aboutgrams

There has been considerable interest in keywords and the notion of keyness in corpus linguistics (see, for example, Scott and Tribble, 2006). Given that phraseology is all pervasive in language, ConcGram can be used to extend the notion of keyness beyond keywords to include the full range of phraseology. Concgrams provide a useful source of raw data which, when analysed, can reveal the co-selections made by the speakers and writers represented in a text or corpus. They are a starting point for quantifying the extent of phraseology in a text or corpus and determining the phraseological profile of the language contained within it. There is evidence to suggest that n-grams, including those made up entirely of grammatical words, can be genre-sensitive (Carter and McCarthy 2006: 828–837; Scott and Tribble 2006: 131–159; O’Keeffe *et al*. 2007: 68), and there is evidence that this is also the case for concgrams. Early studies using concgrams to examine the aboutness of texts and corpora (see Cheng, 2009 and 2008; Greaves and Warren, 2007; Melizia and Spinzi, 2008; O’Donnell, Scott and Mahlberg, 2008; Tognini Bonelli, 2006) suggest that this is fertile ground for further research.

According to Phillips (1989), aboutness is a product of the global patternings of a text. He argues that it should be possible to identify them by computational means, so that they are derived from the text rather than external features. The identification of the phraselogical profile of a text is linked to what Phillips refers to as the aboutness of a text. The phraseological profile is all of the word associations in a text or corpus, and the aboutness of the text or corpus can be determined from the word associations that are specific to that particular text or corpus. Word associations which are specific to a text or corpus are termed 'aboutgrams' (Sinclair, personal communication).

Sinclair (2006) outlines a procedure for using concgrams to produce a list of aboutgrams representing the aboutness of a text which differs from the procedure for determining its phraseological profile. The aboutness of a text is determined by an iterative process, whereby the most frequently occurring lexical phrases in, for example, an engineering text are put on a provisional aboutgram list and are then searched for in a specialised corpus of engineering texts. Those found to occur equally frequently in both the text and the specialised corpus, or more frequently in the specialised corpus, are removed from the list. The process is then repeated using a general corpus and, once again, the list is further revised. Finally, a list of aboutgrams is confirmed which together represents the aboutness of the text. The same methodology can also be employed to uncover the aboutgrams of specialised corpora.

The specificity of two-word concgrams in a text or corpus can also be measured using t-scores or mutual information values. In other words, ConcGram can indicate whether the frequency of occurrence of a two-word concgram is significant relative to its occurrence in a different text or corpus. This facility is particularly helpful for those interested in studying aboutgrams and aboutness.

Summary

Above, I have outlined the unique features of ConcGram and the ways in which its products — concgrams — are proving to be invaluable in studies of phraseology, especially phraseological variation. Congramming also has implications for those of us involved in the learning and teaching of applied linguistics, language studies and Languages for Specific Purposes. Concgrams clearly have a role to play in data-driven learning (DDL) activities (Johns, 1991), and should further advance the learning and teaching of phraseology.

Whatever the source of your interest in phraseology, all researchers, teachers and learners can benefit from the new insights into what Sinclair (1987) terms 'the phraseological tendency of language' which ConcGram makes possible.

Martin Warren
Research Centre for Professional Communication in English, Department of English,
Faculty of Humanities, The Hong Kong Polytechnic University
October, 2008

References

Cheng, W. 2009. *income/interest/net*: Using internal criteria to determine the aboutness of a text in business and financial services English. In *Corpora and Language Teaching* [Studies in Corpus Linguistics 33], K. Aijmer (ed.), 157–177. Amsterdam: John Benjamins.

Cheng, W. 2008. Concgramming: A corpus-driven approach to learning the phraseology of discipline-specific texts. *CORELL: Computer Resources for Language Learning* 1: 22–35.

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

Cheng, W., Greaves, C., Sinclair, J. McH. & Warren, M. In press, 2009. Uncovering the extent of the phraseological tendency: Towards a systematic analysis of concgrams. *Applied Linguistics*.

Greaves, C. & Warren, M. 2007. Concgramming: A computer-driven approach to learning the phraseology of English. *ReCALL Journal* 17(3): 287–306.

Greaves, C. & Warren, M. 2008. Beyond clusters: A new look at word associations. IVACS 4, 4th International Conference: Applying Corpus Linguistics. University of Limerick, Ireland, 13–14 June 2008.

Carter R. & McCarthy, M. 2006. *Cambridge Grammar of English*. Cambridge: CUP.

Hunston, S. 2002. *Corpora in Applied Linguistics*. Cambridge: CUP.

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

Li, Y. & Warren, M. "in … of": What are collocational frameworks and should we be teaching them? 4th International Conference on Teaching English at Tertiary Level. Zhejiang, China 11–12 October, 2008.

Milizia, D. & Spinzi, C. 2008. The 'terroridiom' principle between spoken and written discourse. *International Journal of Corpus Linguistics* 13(3): 322–350.

O'Donnell, M.B., Scott, M. & Mahlberg, M. Exploring text-initial concgrams in a newspaper corpus. 7th International Conference of the American Association of Corpus Linguistics, Brigham Young University, Provo, Utah, USA, 12–15 March, 2008.

O'Keefe, A., Carter, R. & M. McCarthy, M. 2007. *From Corpus to Classroom: Language Use and Language Teaching*. Cambridge: CUP.

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

Renouf, A.J. & Sinclair, J.McH. 1991. Collocational frameworks in English. In *English Corpus Linguistics. Studies in Honour of Jan Svartvik,* K. Ajimer & B. Altenberg (eds), 128–43. Harlow: Longman.

Scott, M. & Tribble, C. 2006. *Textual Patterns: Key Words and Corpus Analysis in Language Education* [Studies in Corpus Linguistics 22]. Amsterdam: John Benjamins.

Sinclair, J. McH. 1987. Collocation: A progress report. In *Language Topics: Essays in Honour of Michael Halliday,* R. Steele & T. Threadgold (eds), 319–331. Amsterdam: John Benjamins.

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

Sinclair, J. McH. 1998. The lexical item. In. *Contrastive Lexical Semantics* [Current Issues in Linguistic Theory 171], E. Weigand (ed.), 1–24. Amsterdam: John Benjamins.

Sinclair, J. McH. 2006. Aboutness 2. Ms, Tuscan Word Centre, Italy.

Sinclair, J. McH. 2007a. Collocation reviewed. Ms, Tuscan Word Centre, Italy.

Sinclair, J. McH. 2007b. Defining the definiendom — new. Ms, Tuscan Word Centre, Italy.

Sinclair, J. McH., Jones, S. & R. Daley, R. 1970. English lexical studies. Report to the Office of Scientific and Technical Information.

Sinclair, J. McH. & Mauranen, A. 2006. *Linear Unit Grammar* [Studies in Corpus Linguistics 25]. Amsterdam: John Benjamins.

Sinclair, J., McH. & Renouf, A. 1988. A lexical syllabus for language learning. In *Vocabulary and Language Teaching*. R. Carter & M. McCarthy (eds), 140–160. London: Longman.

Tognini-Bonelli, E. 2001. *Corpus Linguistics at Work* [Studies in Corpus Linguistics 6]. Amsterdam: John Benjamins.

Tognini-Bonelli, E. 2006. The corpus as an onion: The CÆT Corpus Siena (a corpus of academic economics texts). International Seminar: Special and Varied Corpora. Tuscan Word Centre. October, 2006.

Tognini-Bonelli, E. (ed.) Forthcoming. *John Sinclair on Essential Corpus Linguistics.* London: Routledge.

Chapter One

ConcGram List Builder ©

ConcGram is a program for the automatic identification of phraseological variation. It has been designed, based on an inclusive view of phraseology, to find all the word co-occurrences, called 'concgrams', in a text, and it is left to the user to determine from the context in which those co-occurrences are found whether or not they constitute meaningful word associations. (The term 'co-occurrence' is used here to mean any word which occurs within a vicinity set by the user of another word, and which may or may not be by chance; an 'associated' word is one whose co-occurrence is not accidental but forms a collocate of the other word.)

The search engine identifies these co-occurrences by fully automatically finding and listing concgrams. While this automatic functionality is its primary purpose, it also has the same user nominated functionality as *ConcApp* for languages which use the ASCII character set and have words separated by spaces. (The ConcApp program performs various concordance searches, for both single origin and with co-occurring words, but is limited to user nominated searches only, and is available for download from http://www.edict.com.hk/pub/concapp/.)

Concgrams are instances of word co-occurrence. You can use *ConcGram* to find all word co-occurrences, both grammatical and lexical, and only lexical. These co-occurrences are listed by frequency and can be sorted by character position, i.e. the position in the line determined by character distance from the centred word, not the number of intervening words. The co-occurrences are alphabetically sorted by words occurring to the right or left of the origin, or simply left unsorted. Concgrams may be 2, 3, 4 or 5-word instances of word co-occurrence. (The tutorial in the appendix demonstrates how ConcGram can be used to determine the 'aboutness' of two small files.)

This program is intended as a tool for text analysis, and automatic searches are best conducted on files which do not have more than 5 million words, and are faster on smaller files. For example, a corpus file of about 1 million words with 18,000 unique words will take about a day to create the initial 2-Word Concgram List on a PC running Windows XP with a Pentium 4 3 GHz CPU and 2 gigabytes RAM. If you intend to concgram large corpora you should read Chapter 5 which describes the logistics of doing this and has suggestions for reducing the concgram lists generated by the automatic searches.

1.1 Why concgrams?

You may know the terms 'n-gram', 'cluster', 'bundle', or 'chunk' which are used to refer to adjacent or contiguous words which recur in language. Actual n-grams come in the form of 'bi-grams' (e.g. 'of the'), 'tri-grams' (e.g. 'I don't know'), and so on, indicating the number of adjacent words.

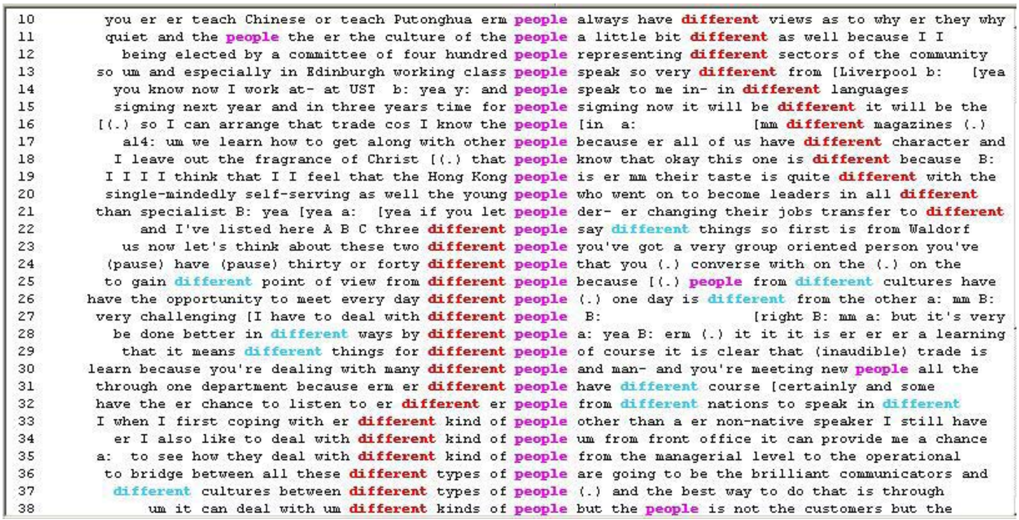
But there are many word co-occurrences which do not occur in one fixed configuration. The relationships of verbs-adverbs, verbs-nouns, nouns-adverbs, noun phrase constituents, quantifier-noun, to name but a few, are flexible, and may occur in non-fixed patterns. For instance, most adjectives can be used both attributively and predicatively. The bi-gram 'challenging exercise' would show in an n-gram search, but when the adjective is used predicatively as in 'the exercise turned out to be quite challenging', it would not. The positions for 'challenging' in this case would be -1 and +6 from the search word 'exercise' respectively. But in both cases, the word 'co-occurrence' is significant.

The terms 'skipgram' (Wilks, 2005) or 'phrase frame' (Fletcher, 2006) are used to describe non-contiguous word co-occurrences of limited membership, and which occur in a fixed pattern of use, for example 'a lot of people' in instances such as 'a lot of business people' and 'a lot of different types of people', but the term 'skipgram' also includes all contiguous co-occurrences and so subsumes n-grams. All these searches require that the words are in the same order (Cheng, Greaves and Warren, 2006). This means that many instances of co-occurrence that typically occur in non-contiguous sequences may not be discovered.

Searches that are user-nominated are also limited by the requirement that the user must enter, and therefore know, items to enable the search to take place. The automatic concgram search provided by ConcGram is able to reveal all word co-occurrences, both contiguous and non-contiguous in a corpus, with both positional (AB, BA) and constituent (ACB) variation. Since it is automatic, the user does not have to first enter one or more search items.

Below is an example of how a concgram search can reveal all the potential instances of co-occurring words, showing both positional and constituent variations.

Figure 1. A concgram search for 'people/different' (2-word concgram)



```
10        you er er teach Chinese or teach Putonghua erm people always have different views as to why er they why
11        quiet and the people the er the culture of the people a little bit different as well because I I
12          being elected by a committee of four hundred people representing different sectors of the community
13       so um and especially in Edinburgh working class people speak so very different from [Liverpool b:   [yea
14         you know now I work at- at UST  b: yea y: and people speak to me in- in different languages
15         signing next year and in three years time for people signing now it will be different it will be the
16      [(.) so I can arrange that trade cos I know the people [in  a:            [mm different magazines (.)
17          al4: um we learn how to get along with other people because er all of us have different character and
18         I leave out the fragrance of Christ [(.) that people know that okay this one is different because  B:
19        I I I I think that I I feel that the Hong Kong people is er mm their taste is quite different with the
20        single-mindedly self-serving as well the young people who went on to become leaders in all different
21       than specialist B: yea [yea a:  [yea if you let people der- er changing their jobs transfer to different
22            and I've listed here A B C three different people say different things so first is from Waldorf
23          us now let's think about these two different people you've got a very group oriented person you've
24        (pause) have (pause) thirty or forty different people that you (.) converse with on the (.) on the
25        to gain different point of view from different people because [(.) people from different cultures have
26     have the opportunity to meet every day different people (.) one day is different from the other a: mm B:
27       very challenging [I have to deal with different people  B:                [right B: mm a: but it's very
28         be done better in different ways by different people a: yea B: erm (.) it it it is er er er a learning
29          that it means different things for different people of course it is clear that (inaudible) trade is
30     learn because you're dealing with many different people and man- and you're meeting new people all the
31       through one department because erm er different people have different course [certainly and some
32       have the er chance to listen to er different er people from different nations to speak in different
33       I when I first coping with er different kind of people other than a er non-native speaker I still have
34          er I also like to deal with different kind of people um from front office it can provide me a chance
35       a:  to see how they deal with different kind of people from the managerial level to the operational
36        to bridge between all these different types of people are going to be the brilliant communicators and
37         different cultures between different types of people (.) and the best way to do that is through
38              um it can deal with um different kinds of people but the people is not the customers but the
```

Figure 1 illustrates a 2-word concgram from search results which has 'people/different' as the 2 co-occurring words. In lines 1–21 we find the configuration 'people *n different' (where *n represents a number of intervening words), and from line 22 this becomes 'different (*n) people'. A bi-gram search would not have found the first configuration because it does not occur contiguously. (The different configurations are listed separately under the Statistics Menu, as explained in Chapter 5.) The concgram search also works when there are common features of spoken language such as repetition, pauses or fillers (i.e. the use of 'er', 'um', etc).

We can also see from Figure 1 that the concgram is sorted by 'character position', not by intervening words, starting with the position which is closest to the origin on the right and becoming more distant, and then (starting in line 22) showing words closest in position to the origin to the left of the origin and becoming more distant. The position is not determined by the number of intervening words, as line 10 is 'people always have different' whereas line 12 is 'people representing different'. In line 12, although there is only one intervening word, the character position is greater than that for line 10 which has 2 intervening words.

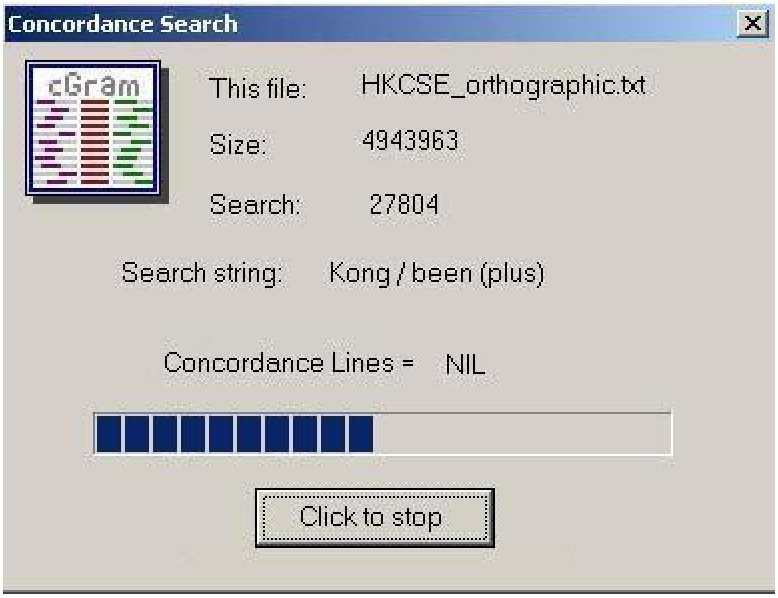
Figure 1 shows only the appearance of a congram display, with sorting by character position (the default). However, the concgram search will find all of the word co-occurrences within a given span, i.e. a distance measured in words between the outer word and the centred word of the concgram. The search will also list co-occurrences that may not prove to be related when examined in context. For this reason ConcGram also provides statistical tests (see Chapter 7) that can provide an indication as to which word co-occurrences are likely to prove to be significant and which ones the user can reasonably afford to ignore.

By creating automatic concgram lists, ConcGram can be used to identify all word co-occurrences that may occur in a corpus within a certain span, and this span can be tailored to suit the needs of the user. The searches can create 2, 3, 4 and 5-word concgrams, but if you are using the automatic search, you must first start with a 2-Word Concgrams list. From this initial 2-word list, you 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 automatic searches.

1.2 Origins and co-occurring words

It is important to understand the difference between 'origins' and 'co-occurring words'. 'Origins' are the source of the search, which may be single origins, double, triple or quadruple origins, and relate to the process of creating automatic 2, 3, 4 and 5-word concgram lists. Co-occurring words are found as a result of the origin searches, and are the words which co-occur with the origin. Origins are indicated in the 'Search Progress' Dialog Box as words separated by slashes, and co-occurring words are enclosed in parenthesis brackets. Figure 2 shows the search string resulting from an origin 'Kong/ been', with a co-occurring word 'plus'. Together they make a 3-word concgram.

Figure 2. Origins and co-occurring words



This file: HKCSE_orthographic.txt

Size: 4943963

Search: 27804

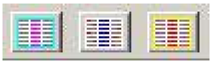
Search string: Kong / been (plus)

Concordance Lines = NIL

1.3 Switching the centred word

The first word in the origin is always centred, but any word in the concgram can be centred by simply highlighting the word and selecting the 'Switch Centred Word' button shown in Figure 3.

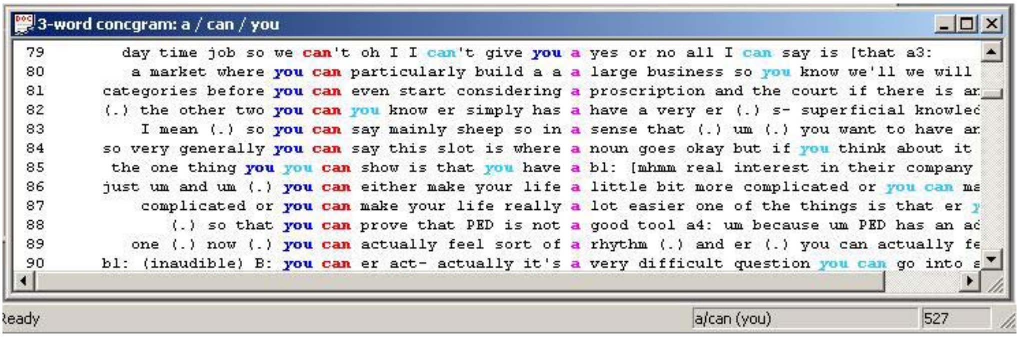
Figure 3. Search buttons on the toolbar



The toolbar features 3 buttons which initiate searches. They look similar except for the colours used. The left button is for single word searches, the centre button for concgram searches, and the right button is for switching the centred word (Figures 4–6).

The concordance for the 3-word concgram 'a/can/you' is shown in Figure 4:

Figure 4. Initial concgram search for 'a/can/you' (3-word concgram)



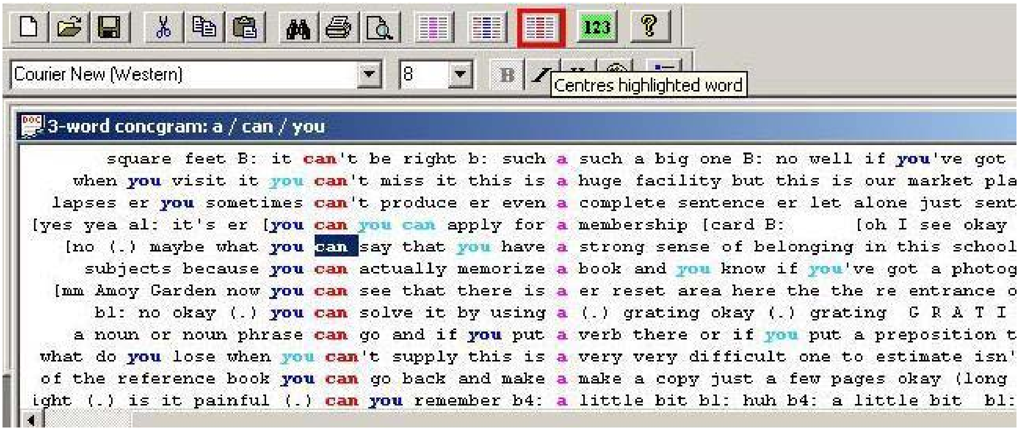
3-word concgram: a / can / you

```
79        day time job so we can't oh I I can't give you a yes or no all I can say is [that a3:
80         a market where you can particularly build a a a large business so you know we'll we will
81     categories before you can even start considering a proscription and the court if there is ar
82     (.) the other two you can you know er simply has a have a very er (.) s- superficial knowlec
83         I mean (.) so you can say mainly sheep so in a sense that (.) um (.) you want to have ar
84     so very generally you can say this slot is where a noun goes okay but if you think about it
85      the one thing you you can show is that you have a bl: [mhmm real interest in their company
86     just um and um (.) you can either make your life a little bit more complicated or you can ma
87         complicated or you can make your life really a lot easier one of the things is that er y
88            (.) so that you can prove that PED is not a good tool a4: um because um PED has an ac
89         one (.) now (.) you can actually feel sort of a rhythm (.) and er (.) you can actually fe
90     bl: (inaudible) B: you can er act- actually it's a very difficult question you can go into s
```

Ready   a/can (you)   527

The concgram has 'a' centred. The use of light blue is to show words which are repeated in the concgram. To make another word in the concgram the centred word, simply select the word (the easiest way is to double click the left mouse button over the word) and then select the 'Switch Centred Word' button from the toolbar as shown in Figure 5:

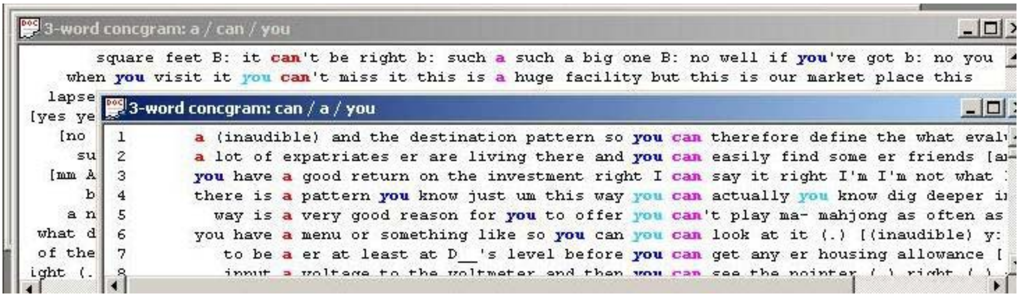
Figure 5. Changing the centred word



```
      square feet B: it can't be right b: such a such a big one B: no well if you've got
    when you visit it you can't miss it this is a huge facility but this is our market pla
  lapses er you sometimes can't produce er even a complete sentence er let alone just sent
[yes yea al: it's er [you can you can apply for a membership [card B:      [oh I see okay
   [no (.) maybe what you can say that you have a strong sense of belonging in this school
     subjects because you can actually memorize a book and you know if you've got a photog
  [mm Amoy Garden now you can see that there is a er reset area here the the re entrance o
     bl: no okay (.) you can solve it by using a (.) grating okay (.) grating  G R A T I
    a noun or noun phrase can go and if you put a verb there or if you put a preposition t
 what do you lose when you can't supply this is a very very difficult one to estimate isn'
 of the reference book you can go back and make a make a copy just a few pages okay (long
ight (.) is it painful (.) can you remember b4: a little bit bl: huh b4: a little bit  bl:
```

Clicking the indicated button results in ‘can’ being the centred word as in Figure 6:

Figure 6. The same concgram ‘a/can/you’ with ‘can’ centred


3-word concgram: a / can / you
3-word concgram: can / a / you

```
      square feet B: it can't be right b: such a such a big one B: no well if you've got b: no you
    when you visit it you can't miss it this is a huge facility but this is our market place this
```

```
1        a (inaudible) and the destination pattern so you can therefore define the what eval
2        a lot of expatriates er are living there and you can easily find some er friends [a
3        you have a good return on the investment right I can say it right I'm I'm not what
4        there is a pattern you know just um this way you can actually you know dig deeper i
5          way is a very good reason for you to offer you can't play ma- mahjong as often as
6        you have a menu or something like so you can you can look at it (.) [(inaudible) y:
7           to be a er at least at D__'s level before you can get any er housing allowance [
8           input a voltage to the voltmeter and then you can see the pointer (.) right (.)
```

Chapter Two:

Selecting a text or corpus to interrogate

2.1 Opening the text or corpus

The first thing to do is open a corpus file to interrogate. If this is an MS Word document (i.e. ____.DOC). It must be saved by MS Word as a 'Plain Text' file (i.e. ____.TXT). Use the 'Save As' menu option from the 'File' menu in Word. Files which have been tagged (such as files saved in HTML or XML format) should have the tagging removed, and there is a function to remove tagging under the Tools Menu. Open the file you want to convert to text only, and select 'Tools ➡ Remove Tagging', and the file with tags removed will be opened in a new window.

2.2 Saving the file in MS Word

When you click the 'Save' button after selecting 'Plain Text' from the dropdown list, a 'File Conversion' Dialog appears. You should select the 'Insert Line Breaks' option, and leave everything else to the default, then click 'OK'.

2.3 Merging several smaller files into one larger corpus file

If you have a number of small corpus files, they need to be merged into one larger file. The concgram automatic list ***only*** works with a single corpus file which must be opened first. You can create a corpus from several smaller files either by using the operating system directly from the command line, or by using the program function to do this under the 'Tools Menu'. The latter function copies all the selected files into a new file, and is suitable for text files.

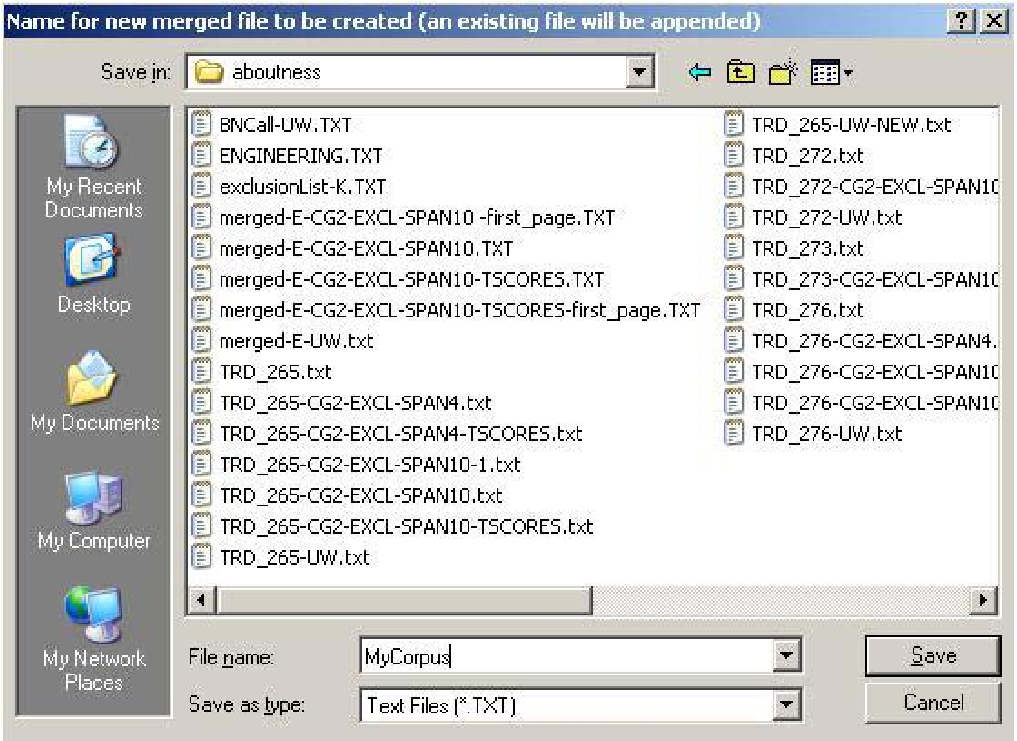
To create the corpus file from the command prompt:

- Navigate to where the files are, for instance a directory named C:\CONCGRAM\CORPUS.
- To get there, open the command prompt (START → RUN → CMD in Windows XP, in earlier versions of the operating system PROGRAMS → ACCESSORIES → MS DOS PROMPT), then type CD C:\ and press ENTER to change to the root directory of the C drive.
- Now type CD C:\CONCGRAM\CORPUS and press ENTER.
- Type DIR to list all the files.
- When you have verified that you are in the correct directory, type COPY *.* MYCORPUS.TXT where MYCORPUS.TXT is the name of the file that you will create for your merged corpus.

There is a function under the 'Tools Menu' for merging files. To use this function, follow the instructions below:

- All the files to be merged must first be copied into one folder.
- Select 'Tools → Merge Files' from the menu.
- A 'Files Dialog' appears first asking for the name of the new Merged File.

Figure 7. The Merged File Dialog



Name for new merged file to be created (an existing file will be appended)

Save in: aboutness

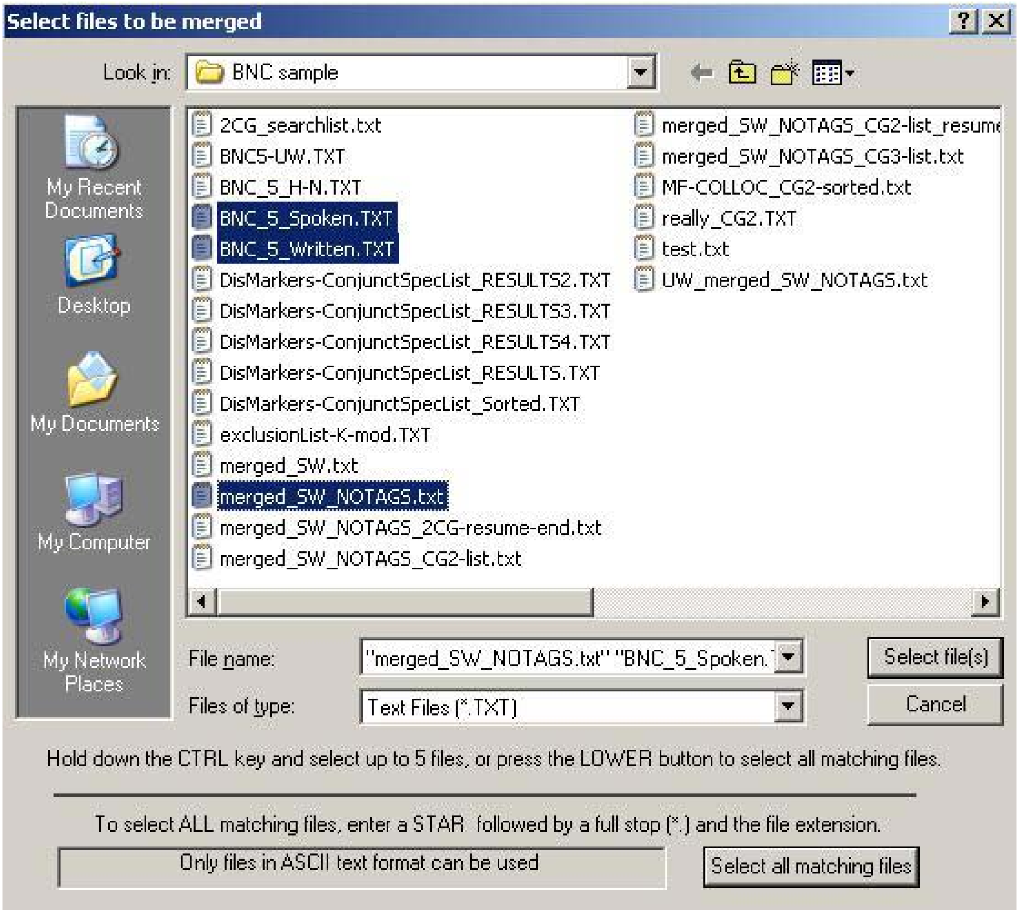
File name: MyCorpus

Save as type: Text Files (*.TXT)

You can select any path for this file. It does not need to be in the same directory as the files you select for merging will all be appended to this file. If you select a file which already exists, the contents of the file will not be overwritten, but the files you choose to merge will be appended to it.

- Click 'Save' when you have specified the name of the new 'Merged File', e.g. MyCorpus.txt.
- Another 'File' Dialog appears, asking you to select the files to merge.

Figure 8. The Select Merge Files Dialog


Select files to be merged

Look in: BNC sample

File name: "merged_SW_NOTAGS.txt" "BNC_5_Spoken.T

Files of type: Text Files (*.TXT)

Hold down the CTRL key and select up to 5 files, or press the LOWER button to select all matching files.

To select ALL matching files, enter a STAR followed by a full stop (*.) and the file extension.

Only files in ASCII text format can be used

My Recent Documents
Desktop
My Documents
My Computer
My Network Places
2CG_searchlist.txt
BNC5-UW.TXT
BNC_5_H-N.TXT
BNC_5_Spoken.TXT
BNC_5_Written.TXT
DisMarkers-ConjunctSpecList_RESULTS2.TXT
DisMarkers-ConjunctSpecList_RESULTS3.TXT
DisMarkers-ConjunctSpecList_RESULTS4.TXT
DisMarkers-ConjunctSpecList_RESULTS.TXT
DisMarkers-ConjunctSpecList_Sorted.TXT
exclusionList-K-mod.TXT
merged_SW.txt
merged_SW_NOTAGS.txt
merged_SW_NOTAGS_2CG-resume-end.txt
merged_SW_NOTAGS_CG2-list.txt
merged_SW_NOTAGS_CG2-list_resume
merged_SW_NOTAGS_CG3-list.txt
MF-COLLOC_CG2-sorted.txt
really_CG2.TXT
test.txt
UW_merged_SW_NOTAGS.txt
Select file(s)
Cancel
Select all matching files

- Navigate to where the files are, for instance a directory named C:\Congram\Corpus.

- Either choose ‘Select Files’ (hold down the Ctrl key and select with the mouse for up to 5 files), or the lower button ‘Select All Matching Files’ (i.e. all files in the directory with the same extension).
- The new ‘Merged File’ that you named before will be created and opened in a new window with all the files you specified copied into it.

Chapter Three:

Getting started with Concgram: Automatic searches

3.1 To Open ConcGram

Open ConcGram by clicking on the 'Start' button, go to 'All Programs' and select 'ConcGram'.

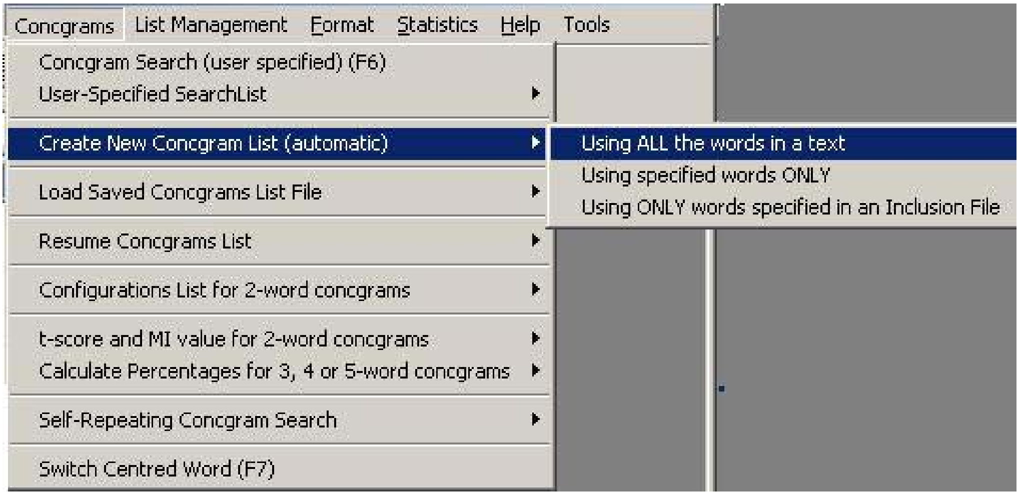
3.2 Open the text or corpus file

To open the file click on the 'File >> Open' menu option, and select from the 'File' Dialog.

3.3 Run an automatic search

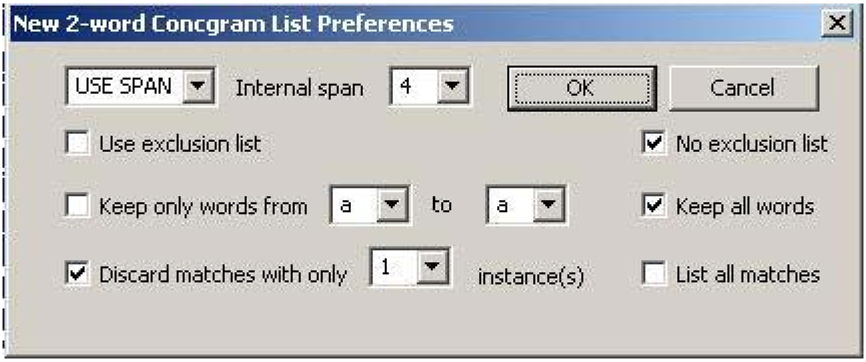
This may take some time to execute, depending on how large your corpus file is. To run an automatic search, select 'Congrams >> Create New Congram List (Automatic)', and choose one of the options (Figure 9):

Figure 9. Running an automatic search


Concgrams
List Management
Format
Statistics
Help
Tools
Concgram Search (user specified) (F6)
User-Specified SearchList
Create New Concgram List (automatic)
Load Saved Concgrams List File
Resume Concgrams List
Configurations List for 2-word concgrams
t-score and MI value for 2-word concgrams
Calculate Percentages for 3, 4 or 5-word concgrams
Self-Repeating Concgram Search
Switch Centred Word (F7)
Using ALL the words in a text
Using specified words ONLY
Using ONLY words specified in an Inclusion File

You will then be asked to choose options from the 'Concgram List Preferences' Dialog Box which has the following choices (Figure 10):

Figure 10. Concgram List Preferences



New 2-word Concgram List Preferences

USE SPAN ▾ Internal span 4 ▾ OK Cancel

□ Use exclusion list ☑ No exclusion list

□ Keep only words from a ▾ to a ▾ ☑ Keep all words

☑ Discard matches with only 1 ▾ instance(s) □ List all matches

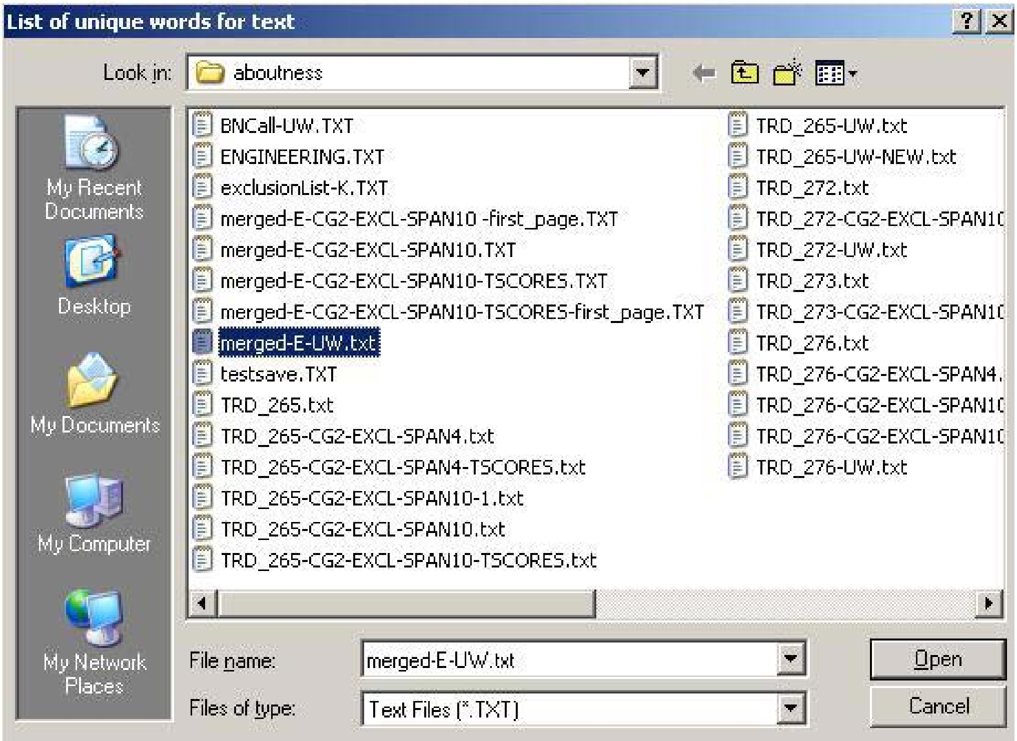
What preferences you select may depend upon the size of the corpus you are interrogating. Remember a corpus file of spoken English of about 1 million words and 18,000 unique words will take about a day to create the initial 2-word concgram list on a PC running Windows XP with a Pentium 4 3 GHz CPU and 2 gigabytes RAM. If you decide only to search for words beginning with a single letter, the initial list will be completed much more quickly.

3.4 Creating the Initial Concgram List

To create the Initial Concgram List, first select 'Create New Concgram List (Automatic) >> Using all the Words in a Text' from the Concgrams Menu. You will be prompted to select from the 'Concgram Preferences' Dialog if you wish to modify the size of the lists created, for example, more than half of all searches will result in only one match being found. If you wish, these single instances can be discarded, and the resulting list will be shortened. Words such as 'the', 'of', 'to', and 'and' occur very frequently, and can be listed in an Exclusion List. (For a more detailed explanation of the functions for reducing the size of the lists and for an example of Exclusion List, see Chapter 4.) A word of caution is appropriate here. While grammatical words make up a very large part of English (almost 40% of the English Language is made up of 50 grammatical words (Ahmad 2005)), if you want to study combinations of these words, you will not include them in an Exclusion List. John Sinclair referred to these combinations as 'collocational frameworks', by which he meant combinations such as 'a/of' in '**a lot of**'. If you want to learn more about collocational frameworks, a good place to start is Renouf and Sinclair (1991).

There are two steps in creating an initial automatic concgram list. The first step is to open a list of 'Unique Words' in the text, as shown in Figure 11. A word may be used many times in a text, for instance, 'and', but it will be counted only once in the list of Unique Words. To create and save a list of unique words, use the Statistics ➡ Unique Words menu function as explained in the Help File.

Figure 11. Opening a list of Unique Words


List of unique words for text

Look in: aboutness

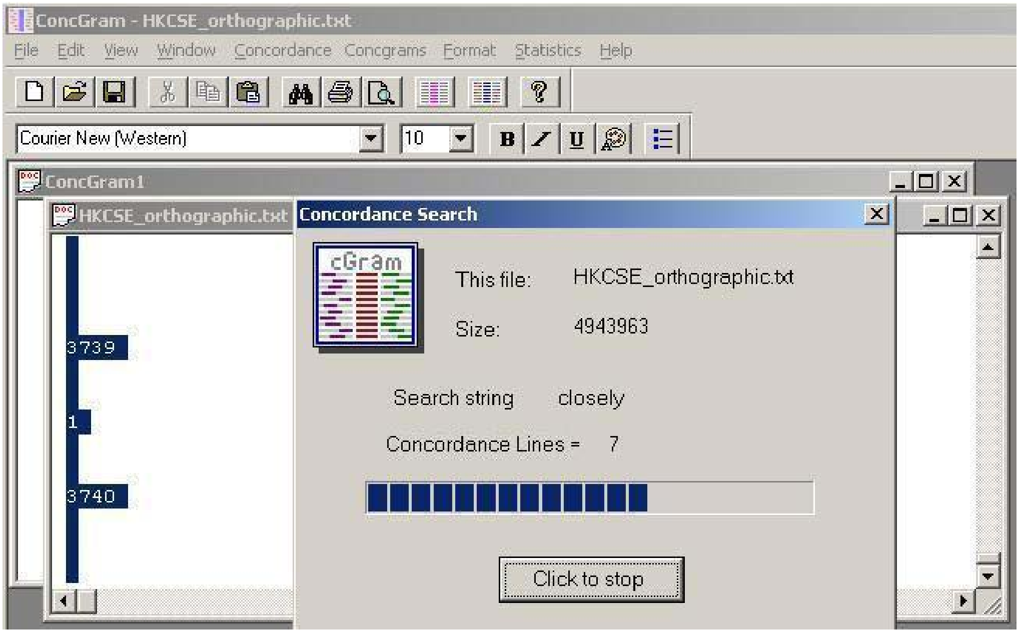
BNCall-UW.TXT
ENGINEERING.TXT
exclusionList-K.TXT
merged-E-CG2-EXCL-SPAN10 -first_page.TXT
merged-E-CG2-EXCL-SPAN10.TXT
merged-E-CG2-EXCL-SPAN10-TSCORES.TXT
merged-E-CG2-EXCL-SPAN10-TSCORES-first_page.TXT
merged-E-UW.txt
testsave.TXT
TRD_265.txt
TRD_265-CG2-EXCL-SPAN4.txt
TRD_265-CG2-EXCL-SPAN4-TSCORES.txt
TRD_265-CG2-EXCL-SPAN10-1.txt
TRD_265-CG2-EXCL-SPAN10.txt
TRD_265-CG2-EXCL-SPAN10-TSCORES.txt
TRD_265-UW.txt
TRD_265-UW-NEW.txt
TRD_272.txt
TRD_272-CG2-EXCL-SPAN1
TRD_272-UW.txt
TRD_273.txt
TRD_273-CG2-EXCL-SPAN1
TRD_276.txt
TRD_276-CG2-EXCL-SPAN4.
TRD_276-CG2-EXCL-SPAN1
TRD_276-CG2-EXCL-SPAN1
TRD_276-UW.txt

File name: merged-E-UW.txt

Files of type: Text Files (*.TXT)

The second step is that the program uses the list of Unique Words to search the whole file, based on each word in the list as a single origin, as shown in Figure 12. These searches create the initial 2-word concgram list.

Figure 12. Initial search using the Unique Words list



ConcGram - HKCSE_orthographic.txt
File Edit View Window Concordance Concgrams Format Statistics Help
Courier New (Western)
10
ConcGram1
HKCSE_orthographic.txt
3739
1
3740
Click to stop

Concordance Search

This file: HKCSE_orthographic.txt

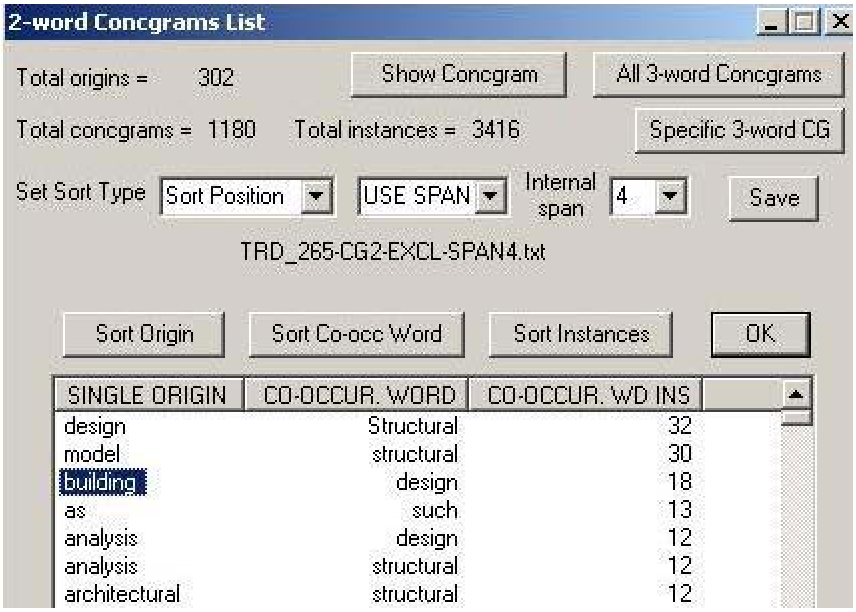
Size: 4943963

Search string closely

Concordance Lines = 7

For each concordance resulting from the search, a new list of co-occurring words is generated and displayed in the List Box for 2-word concgrams (Figure 13).

Figure 13. The 2-Word Concgrams List Box



Total origins = 302

Total concgrams = 1180 Total instances = 3416

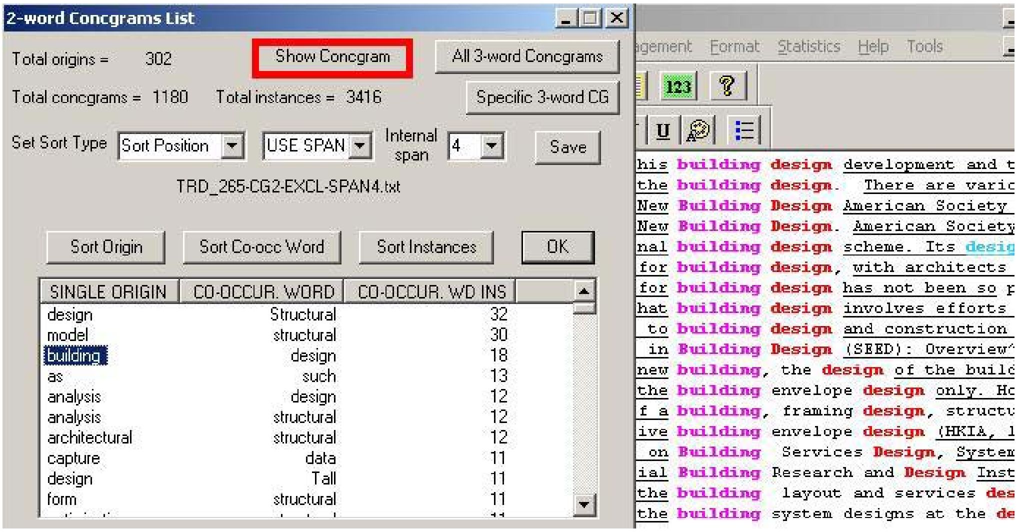
Set Sort Type: Sort Position | USE SPAN | Internal span: 4

TRD_265-CG2-EXCL-SPAN4.txt

SINGLE ORIGIN	CO-OCCUR. WORD	CO-OCCUR. WD INS
design	Structural	32
model	structural	30
building	design	18
as	such	13
analysis	design	12
analysis	structural	12
architectural	structural	12

Once you have created the first list, you can either generate an individual 2-word concgram search by selecting any item from the list and selecting the 'Show Concgram' button (Figure 14), or a completely new list from all the items in the List Box by selecting the '3-Word Concgrams' button. Note that the numbers shown in the 'Sort Instances' column give the number of occurrences of the co-occurring word with the single origin. This may not be the same as the number found in the concgram searches, which includes repeats of both co-occurring word and origin word(s) in the same concordance line. Nevertheless, these numbers give a good idea of how many concordance will been found.

Figure 14. Result of a 2-Word Concgrams search by selecting from the list and selecting the 'Show Congram' button



Total origins = 302

Total concgrams = 1180 Total instances = 3416

Set Sort Type: Sort Position | USE SPAN | Internal span: 4

TRD_265-CG2-EXCL-SPAN4.txt

SINGLE ORIGIN	CO-OCCUR. WORD	CO-OCCUR. WD INS
design	Structural	32
model	structural	30
building	design	18
as	such	13
analysis	design	12
analysis	structural	12
architectural	structural	12
capture	data	11
design	Tall	11
form	structural	11

```
his building design development and t
the building design.  There are varic
New Building Design American Society
New Building Design. American Society
nal building design scheme. Its desig
for building design, with architects
for building design has not been so p
hat building design involves efforts
 to building design and construction
 in Building Design (SEED): Overview^
new building, the design of the buil
the building envelope design only. Hc
f a building, framing design, structu
ive building envelope design (HKIA, l
 on Building  Services Design, Syster
ial Building Research and Design Inst
the building  layout and services des
the building system designs at the de
```

3.5 Saving the Lists

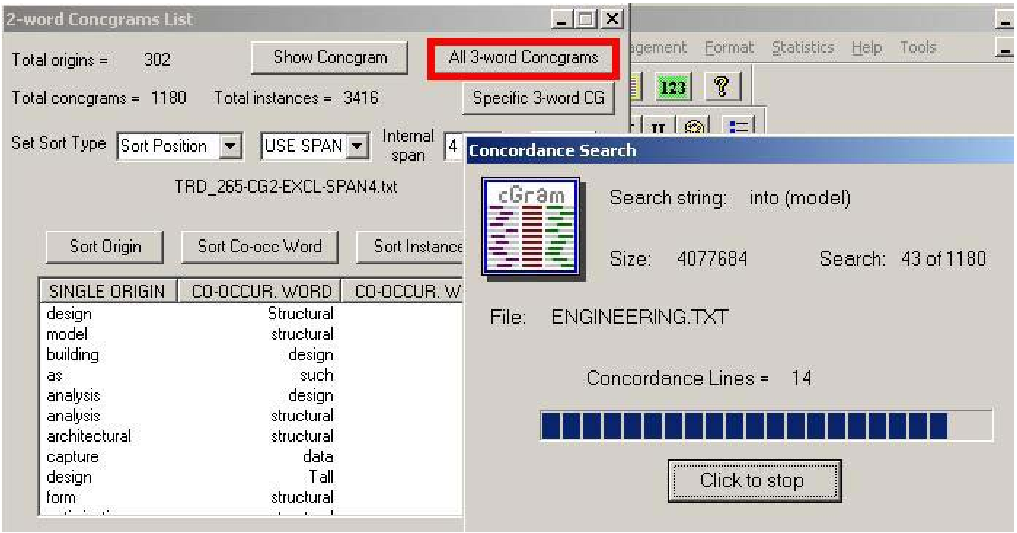
All the Concgram List Boxes have a 'Save' button which enables the user to save the list as a text file which can then be opened under item 4 'Concgrams >> Load Saved Concgrams List File' of the 'Concgrams' menu. This loads the concgrams list in a Concgram List Box so that you can search for concordances or use any of the other functions in the list box.

When the list is first saved as a text file it is opened in a new window. It is important that if you want to perform concordance searches you must first close this text file as concordance searches will be performed on ***all files which are open*** as well as the corpus file itself, and if the list is also open as a text file then it will also be searched and the search will produce meaningless results. Concordance files can be left open as they will not be searched, but a list which has been saved as a text file ***must first be closed*** before performing any searches.

3.6 Creating a 3-Word Concgrams List

The 3-word Concgrams List is created by selecting the 'All 3-Word Concgrams' button in the '2-Word Concgrams List' Dialog Box, which then performs an automatic concgram search on every pair of words generated by the initial search. Figure 15 below shows the search string as the 'Single Origin' with the 'Co-occurring Word' given in parenthesis following it:

Figure 15. Automatic 3-Word Concgrams search using all the origins with co-occurring words listed in the 2-Word Concgrams list, created by selecting the All 3-Word Concgrams button



Total origins = 302

Total concgrams = 1180 Total instances = 3416

Set Sort Type: Sort Position; USE SPAN; Internal span: 4

TRD_265-CG2-EXCL-SPAN4.txt

SINGLE ORIGIN	CO-OCCUR. WORD	CO-OCCUR. W
design	Structural	
model	structural	
building	design	
as	such	
analysis	design	
analysis	structural	
architectural	structural	
capture	data	
design	Tall	
form	structural	

Search string: into (model)

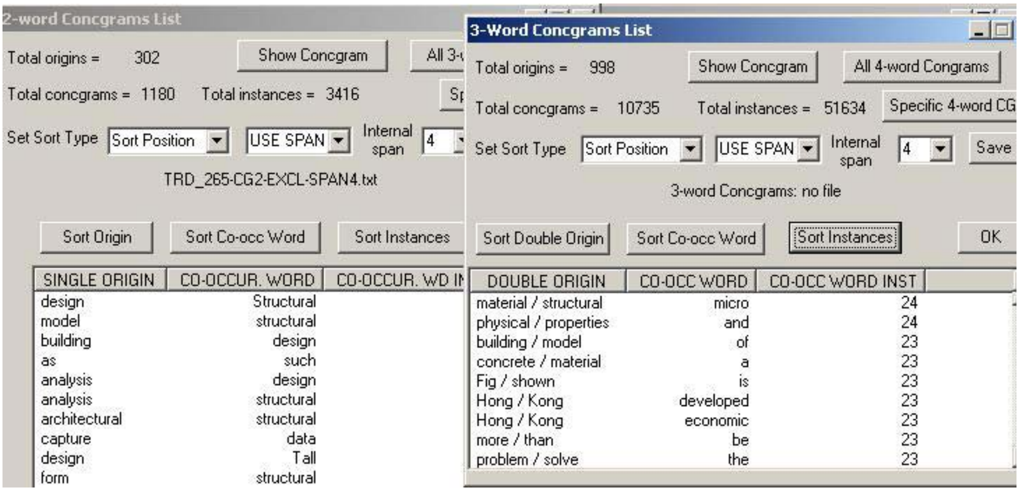
Size: 4077684 Search: 43 of 1180

File: ENGINEERING.TXT

Concordance Lines = 14

For each string searched, all the co-occurring words are subsequently displayed in the '3-Word Concgrams List' Box (Figure 16):

Figure 16. Results shown in a 3-Word Concgrams List Box



Total origins = 302

Total concgrams = 1180 Total instances = 3416

Set Sort Type Sort Position USE SPAN Internal span 4

TRD_265-CG2-EXCL-SPAN4.txt

SINGLE ORIGIN	CO-OCCUR. WORD	CO-OCCUR. WD IN
design	Structural	
model	structural	
building	design	
as	such	
analysis	design	
analysis	structural	
architectural	structural	
capture	data	
design	Tall	
form	structural	

Total origins = 998

Total concgrams = 10735 Total instances = 51634

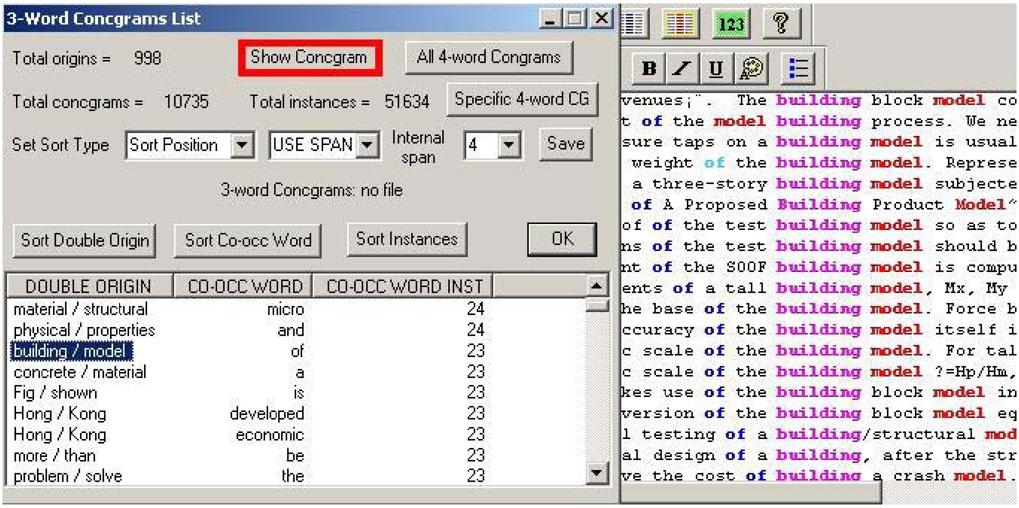
Set Sort Type Sort Position USE SPAN Internal span 4

3-word Concgrams: no file

DOUBLE ORIGIN	CO-OCC WORD	CO-OCC WORD INST
material / structural	micro	24
physical / properties	and	24
building / model	of	23
concrete / material	a	23
Fig / shown	is	23
Hong / Kong	developed	23
Hong / Kong	economic	23
more / than	be	23
problem / solve	the	23

Double Origins and Co-occurring Words are listed, and an individual '3-Word Concgrams' search (Double Origin + Co-occurring Word) can be performed by selecting any item from the list and selecting the 'Show Concgram' button (Figure 17).

Figure 17. Results of selecting the 'Show Concgram' button in a 3-Word Concgrams List Box



Total origins = 998

Total concgrams = 10735 Total instances = 51634

Set Sort Type Sort Position USE SPAN Internal span 4

3-word Concgrams: no file

DOUBLE ORIGIN	CO-OCC WORD	CO-OCC WORD INST
material / structural	micro	24
physical / properties	and	24
building / model	of	23
concrete / material	a	23
Fig / shown	is	23
Hong / Kong	developed	23
Hong / Kong	economic	23
more / than	be	23
problem / solve	the	23

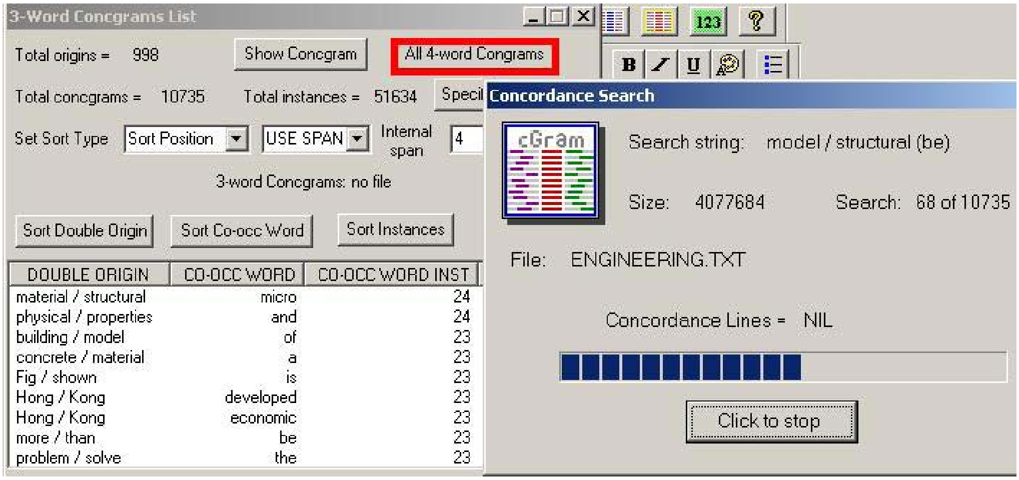
```
venues;¨.  The building block model co
t of the model building process. We ne
sure taps on a building model is usual
 weight of the building model. Represe
 a three-story building model subjecte
 of A Proposed Building Product Model"
of of the test building model so as to
ns of the test building model should b
nt of the SOOF building model is compu
ents of a tall building model, Mx, My
he base of the building model. Force b
ccuracy of the building model itself i
c scale of the building model. For tal
c scale of the building model ?=Hp/Hm,
kes use of the building block model in
version of the building block model eq
l testing of a building/structural mod
al design of a building, after the str
ve the cost of building a crash model.
```

You can then generate a completely new list from all the items in the List Box by selecting the '4-Word Concgrams' button.

3.7 Creating 4-Word Concgrams List

The '4-Word Concgrams List' is created by selecting the 'All 4-Word Concgrams' button in the Dialog Box, which initiates a Co-occurring Word Concgram search on every group of 3 words in the list generated for 3-word concgrams (Figure 18).

Figure 18. Automatic 4-Word Concgrams search generated from the All 4-Word Concgrams button



Total origins = 998

Total concgrams = 10735

Total instances = 51634

Set Sort Type: Sort Position, USE SPAN

Internal span: 4

3-word Concgrams: no file

DOUBLE ORIGIN	CO-OCC WORD	CO-OCC WORD INST
material / structural	micro	24
physical / properties	and	24
building / model	of	23
concrete / material	a	23
Fig / shown	is	23
Hong / Kong	developed	23
Hong / Kong	economic	23
more / than	be	23
problem / solve	the	23

Search string: model / structural (be)

Size: 4077684

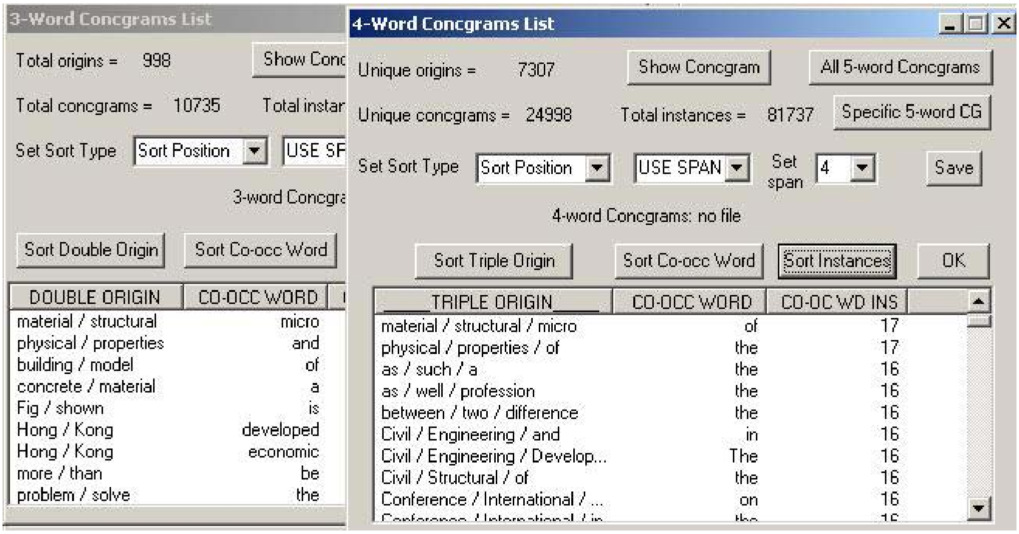
Search: 68 of 10735

File: ENGINEERING.TXT

Concordance Lines = NIL

Figure 18 shows the Double Origin and Co-occurring Word given in parenthesis following the Double Origin. The results of this 4-Word Concgrams search are displayed in the List Box shown in Figure 19 below.

Figure 19. Results shown in a 4-Word Concgrams List



Total origins = 998

Total concgrams = 10735 Total instar

Set Sort Type: Sort Position USE SF

3-word Concgra

DOUBLE ORIGIN	CO-OCC WORD
material / structural	micro
physical / properties	and
building / model	of
concrete / material	a
Fig / shown	is
Hong / Kong	developed
Hong / Kong	economic
more / than	be
problem / solve	the

Unique origins = 7307

Unique concgrams = 24998 Total instances = 81737

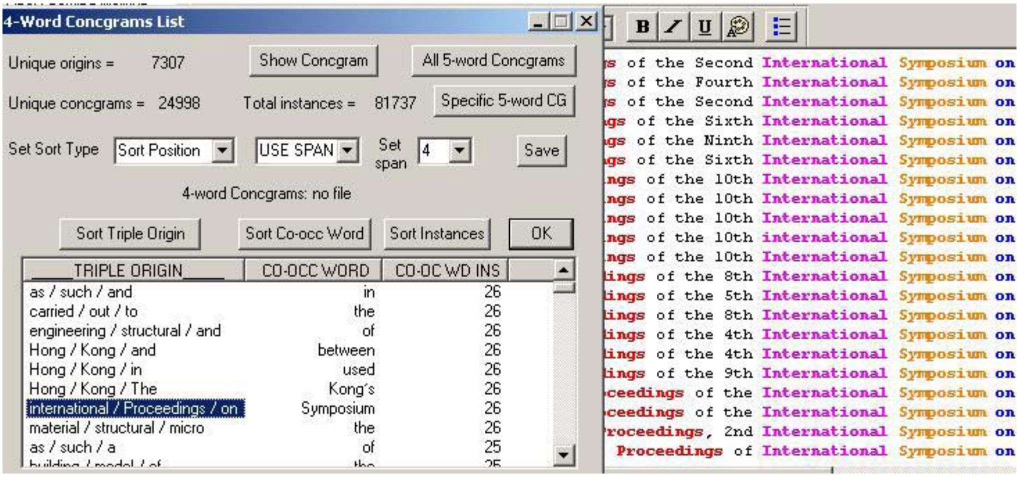
Set Sort Type: Sort Position USE SPAN Set span: 4

4-word Concgrams: no file

TRIPLE ORIGIN	CO-OCC WORD	CO-OC WD INS
material / structural / micro	of	17
physical / properties / of	the	17
as / such / a	the	16
as / well / profession	the	16
between / two / difference	the	16
Civil / Engineering / and	in	16
Civil / Engineering / Develop...	The	16
Civil / Structural / of	the	16
Conference / International / ...	on	16

The Triple Origin is listed by using the concordance strings resulting from the searches. An individual '4-Word Concgrams' search (Triple Origin + Co-occurring Word) can be performed by selecting any item from the list and selecting the 'Show Concgram' button. Alternatively, you can generate a completely new list from all the items in the List Box by selecting the '5-word Concgrams' button (Figure 20).

Figure 20. Results of an automatic search by selecting 'Show Concgram' from the 4-Word Concgrams List Box



Unique origins = 7307

Unique concgrams = 24998 Total instances = 81737

Set Sort Type: Sort Position USE SPAN Set span: 4

4-word Concgrams: no file

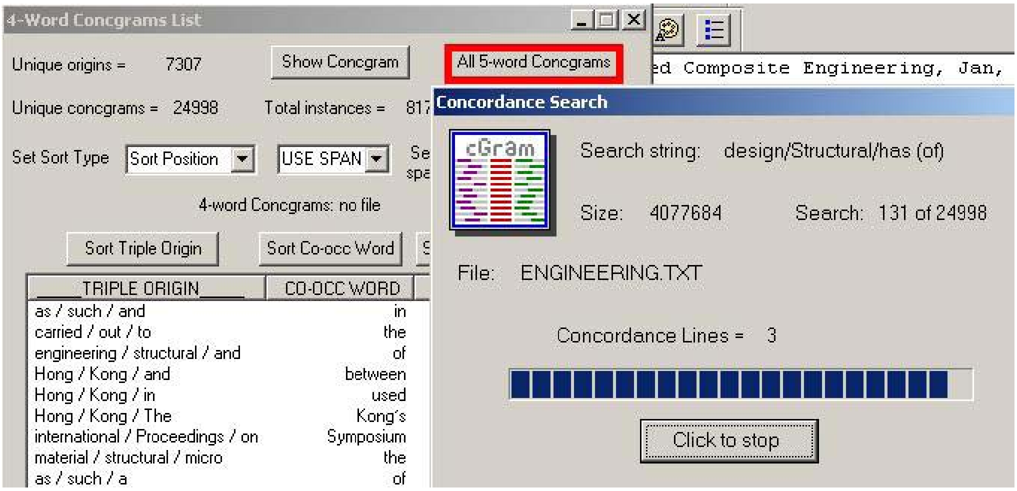
TRIPLE ORIGIN	CO-OCC WORD	CO-OC WD INS
as / such / and	in	26
carried / out / to	the	26
engineering / structural / and	of	26
Hong / Kong / and	between	26
Hong / Kong / in	used	26
Hong / Kong / The	Kong's	26
international / Proceedings / on	Symposium	26
material / structural / micro	the	26
as / such / a	of	25

```
s of the Second International Symposium on
s of the Fourth International Symposium on
s of the Second International Symposium on
gs of the Sixth International Symposium on
gs of the Ninth International Symposium on
gs of the Sixth International Symposium on
ngs of the 10th International Symposium on
ngs of the 10th International Symposium on
ngs of the 10th International Symposium on
ngs of the 10th international Symposium on
ngs of the 10th International Symposium on
ings of the 8th International Symposium on
ings of the 5th International Symposium on
ings of the 8th International Symposium on
ings of the 4th International Symposium on
ings of the 4th International Symposium on
ings of the 9th International Symposium on
ceedings of the International Symposium on
ceedings of the International Symposium on
roceedings, 2nd International Symposium on
 Proceedings of International Symposium on
```

3.8 Creating 5-Word Concgrams Lists

Finally, you can automatically create a 5-Word Concgrams List by selecting the button from the '4-Word Concgrams' Dialog Box (Figure 21).

Figure 21. Automatic 5-Word Concgrams search generated from the 5-Word Concgrams button



4-Word Concgrams List

Unique origins = 7307

Unique concgrams = 24998 Total instances = 817

Set Sort Type Sort Position USE SPAN

4-word Concgrams: no file

TRIPLE ORIGIN	CO-OCC WORD
as / such / and	in
carried / out / to	the
engineering / structural / and	of
Hong / Kong / and	between
Hong / Kong / in	used
Hong / Kong / The	Kong's
international / Proceedings / on	Symposium
material / structural / micro	the
as / such / a	of

ed Composite Engineering, Jan,

Concordance Search

Search string: design/Structural/has (of)

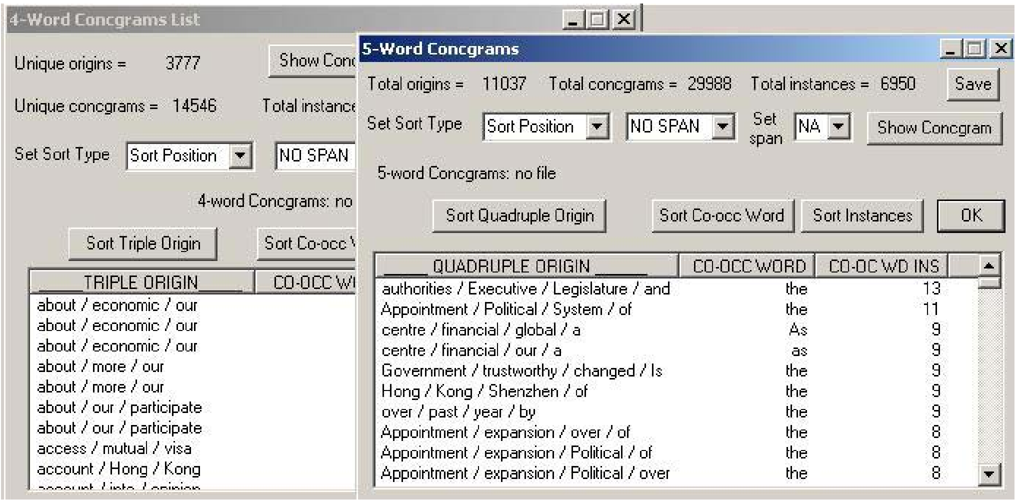
Size: 4077684 Search: 131 of 24998

File: ENGINEERING.TXT

Concordance Lines = 3

Figure 21 shows the Triple Origin with the Co-occurring Word given in parenthesis following the Triple Origin. The results of this 5-Word Concgrams search are displayed in the List Box shown in Figure 22 below.

Figure 22. Results shown in a 5-Word Concgrams List Box



Unique origins = 3777

Unique concgrams = 14546 Total instance

Set Sort Type: Sort Position; NO SPAN

4-word Concgrams: no

TRIPLE ORIGIN	CO-OCC WI
about / economic / our	
about / economic / our	
about / economic / our	
about / more / our	
about / more / our	
about / our / participate	
about / our / participate	
access / mutual / visa	
account / Hong / Kong	

Total origins = 11037 Total concgrams = 29988 Total instances = 6950

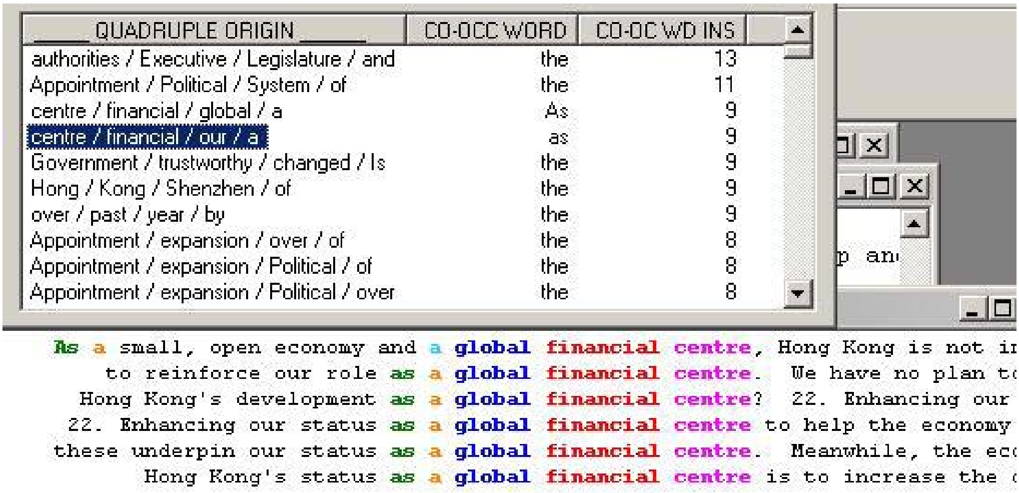
Set Sort Type: Sort Position; NO SPAN; Set span: NA

5-word Concgrams: no file

QUADRUPLE ORIGIN	CO-OCC WORD	CO-OC WD INS
authorities / Executive / Legislature / and	the	13
Appointment / Political / System / of	the	11
centre / financial / global / a	As	9
centre / financial / our / a	as	9
Government / trustworthy / changed / Is	the	9
Hong / Kong / Shenzhen / of	the	9
over / past / year / by	the	9
Appointment / expansion / over / of	the	8
Appointment / expansion / Political / of	the	8
Appointment / expansion / Political / over	the	8

The Quadruple Origin is listed using the concordance lines resulting from the searches, and an individual '5-Word Concgram' can be displayed (Quadruple Origin + Co-occurring Word) can be performed by selecting any item from the list and selecting the 'Show Concgram' button (Figure 23).

Figure 23. Results of an automatic search by selecting 'Show Concgram' from the 5-word Concgrams List Box



QUADRUPLE ORIGIN	CO-OCC WORD	CO-OC WD INS
authorities / Executive / Legislature / and	the	13
Appointment / Political / System / of	the	11
centre / financial / global / a	As	9
centre / financial / our / a	as	9
Government / trustworthy / changed / Is	the	9
Hong / Kong / Shenzhen / of	the	9
over / past / year / by	the	9
Appointment / expansion / over / of	the	8
Appointment / expansion / Political / of	the	8
Appointment / expansion / Political / over	the	8

```
                    As a small, open economy and a global financial centre, Hong Kong is not i
              to reinforce our role as a global financial centre.  We have no plan t
   Hong Kong's development as a global financial centre?  22. Enhancing our
22. Enhancing our status as a global financial centre to help the economy
  these underpin our status as a global financial centre.  Meanwhile, the ec
           Hong Kong's status as a global financial centre is to increase the
```

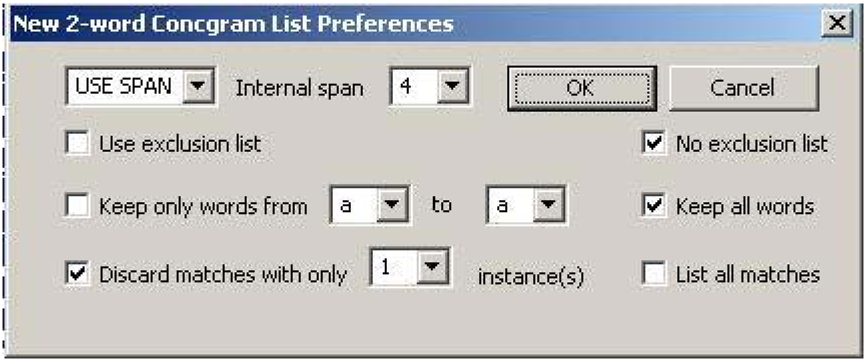
Chapter Four

Managing the lists

4.1 Making the lists shorter

When you generate the concgram lists using the automatic concgram search functions, you will find that the lists can be very long. As a result, the program has a variety of ways to reduce the length of the resulting lists. What method you choose depends on the outcome that you are seeking and the size of the file that you are using initially to create the lists. Preferences can be selected using the 'Concgram List Preferences' Dialog which pops up every time you select 'Concgrams >> Create New Concgram List >> Using ALL the Words in a Text'.

Figure 24. Concgram List Preferences Dialog



New 2-word Concgram List Preferences

USE SPAN ▾ Internal span 4 ▾

□ Use exclusion list ☑ No exclusion list

□ Keep only words from a ▾ to a ▾ ☑ Keep all words

☑ Discard matches with only 1 ▾ instance(s) □ List all matches

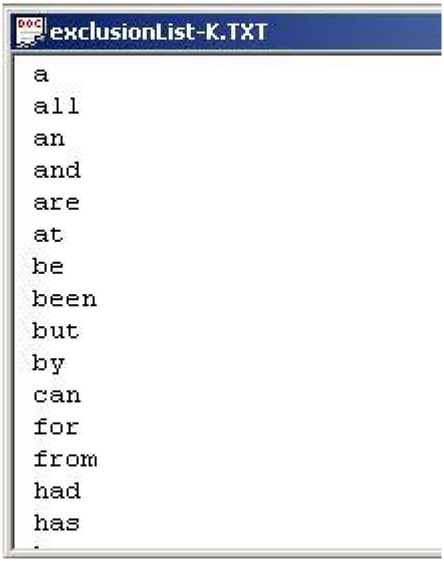
There are several ways of controlling list length:

1. **Internal span for the searches**. 'Internal span' refers to the number of intervening words between the centred word and outer co-occurring word in a concgram. The smaller the span, the shorter the list.
2. **Exclusion List**. This is a text file containing a list of words which the search engine will ignore. Grammatical words such as 'the, of, to, and' occur very frequently, and can be studied independently by means of an 'Inclusion List', but can first be listed

in an 'Exclusion List'. This will greatly reduce the size of the resulting lists, as these words will not be included as an origin. A sample exclusion list is shown in Figure 25 below.

3. **Discarding matches with only 1 or 2 instances**. If you drop single instance matches, you can reduce the size of the list dramatically. Depending on your data file, single instance matches may comprise more than half of the total number, and will probably not be significant. Unless of course you are looking specifically for more general patterns, in which case you may not want to discard single instances.
4. **Searching only for words starting with a particular letter or letters.** If you do not want to use an Exclusion List, you can create Concgram Lists for all the letters of the alphabet separately.

Figure 25. An Exclusion List

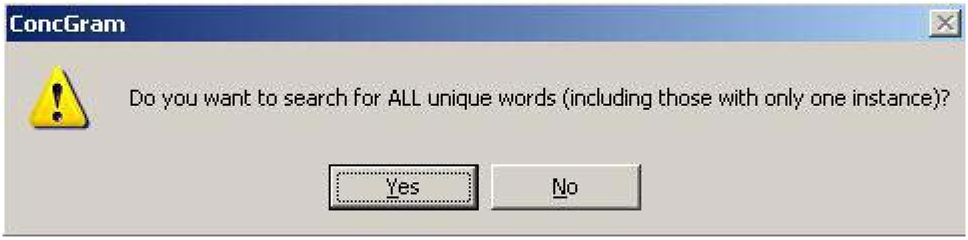


```
a
all
an
and
are
at
be
been
but
by
can
for
from
had
has
```

This list has been created using the 50 grammatical words which, according to Ahmad (2005), make up almost 40% of the English Language.

Immediately following the Concgram Preferences Dialog the user is asked whether to include ALL the unique words or only those which occur more than once in the text. About 40% or more of the unique words in a list typically occur only once in a text, these words can safely be discarded, and this will save time doing the searches. To discard these words the user must select "NO".

Figure 26. Choosing whether to search for all unique words

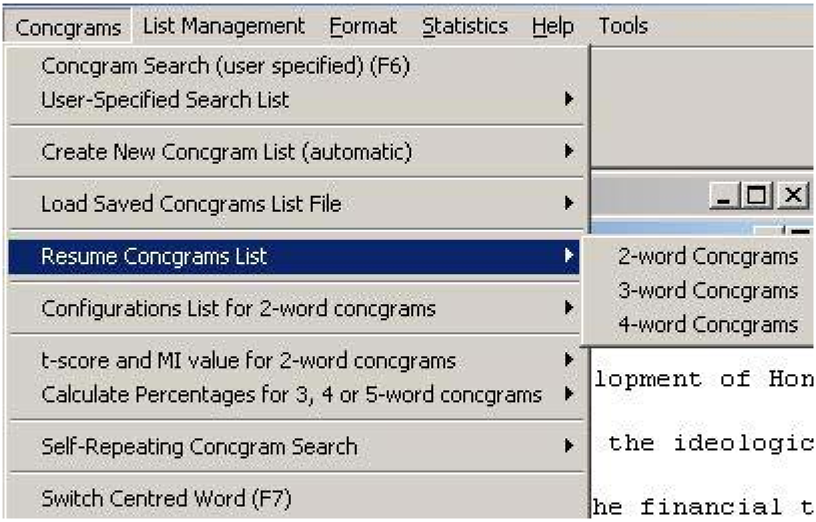


Do you want to search for ALL unique words (including those with only one instance)?

4.2 Stop and Resume Creating a List

Another way to control the lists and also to bring down the virtual memory consumed in doing the searches is to stop and resume an automated search. This function is available by selecting item 5 from the 'Concgrams' menu.

Figure 27. Resume creating a list

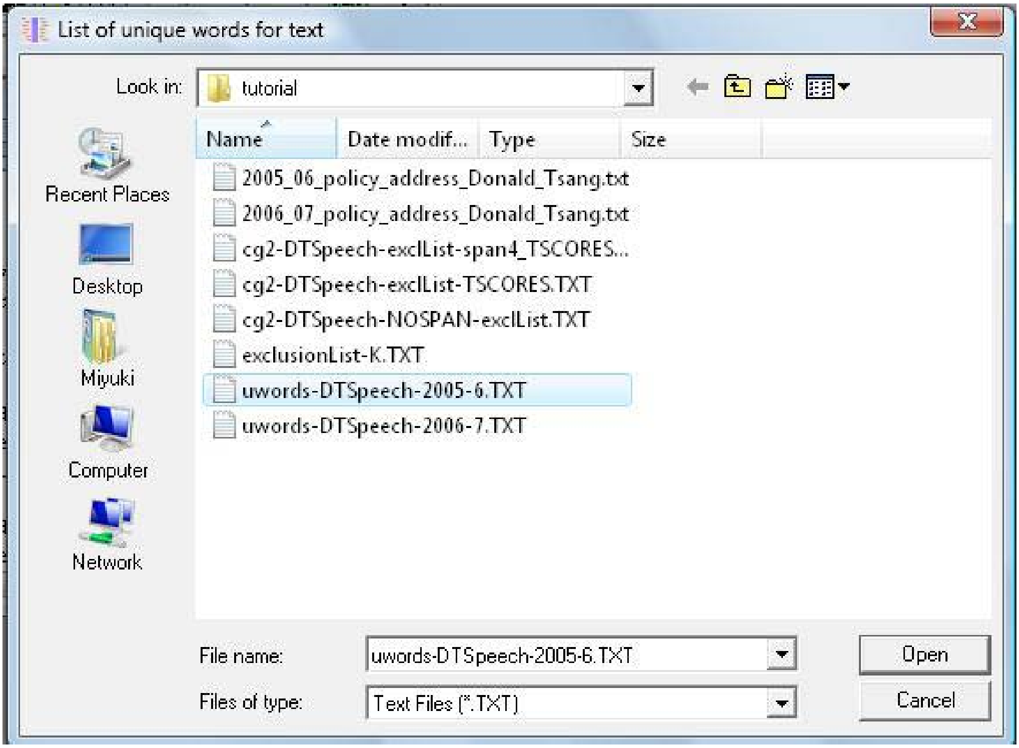

Concgrams
List Management
Format
Statistics
Help
Tools
Concgram Search (user specified) (F6)
User-Specified Search List
Create New Concgram List (automatic)
Load Saved Concgrams List File
Resume Concgrams List
Configurations List for 2-word concgrams
t-score and MI value for 2-word concgrams
Calculate Percentages for 3, 4 or 5-word concgrams
Self-Repeating Concgram Search
Switch Centred Word (F7)
2-word Concgrams
3-word Concgrams
4-word Concgrams
lopment of Hon
the ideologic
he financial t

This function to stop and resume creating a concgrams list can be for 2, 3 and 4 word lists, but there are some differences in the way they are implemented and behave. This is particularly useful for creating 2-word concgrams, which are created using a list of unique words. If you stop a 2-word concgram list to resume later, resumption of the list is at the letter which was being processed when you stopped. For example, if the original list started with letter 'A', you must complete all the 'A' words in the list before stopping, and at least be on the words beginning with letter 'B' when you stop. If you subsequently resume the list, it will start at letter 'B' words, regardless of where you were when you stopped. Similarly, if you were on letter 'E' words when you stopped, the program will resume at the first word starting with letter 'E'.

Resuming 3 and 4-word concgram lists is a little different, as double and triple origins are used in the searches. Resumption therefore occurs at the double or triple origin before the one which was the focus of the search when you stopped.

If you want to resume a 2-word concgram list, after the Concgram Preferences Dialog appears you will be asked to first load the unique word list you used when you first created the list.

Figure 28. To open a 2-word concgrams list, first open the Unique Word List



List of unique words for text

Look in: tutorial

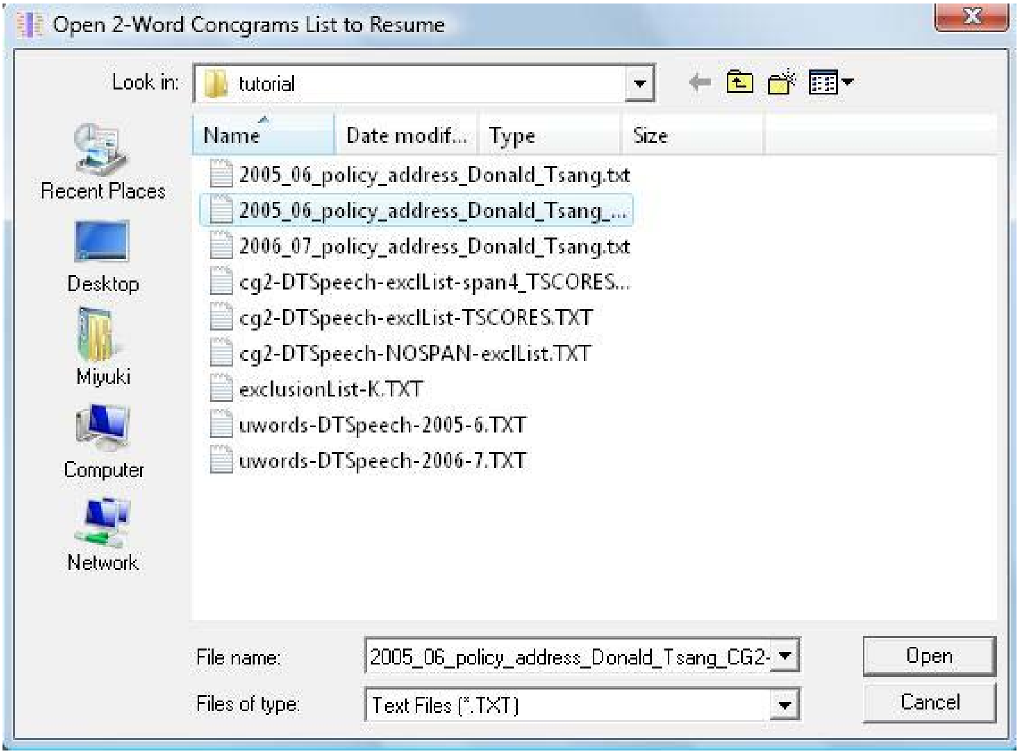
Name	Date modif...	Type	Size
2005_06_policy_address_Donald_Tsang.txt			
2006_07_policy_address_Donald_Tsang.txt			
cg2-DTSpeech-exclList-span4_TSCORES...			
cg2-DTSpeech-exclList-TSCORES.TXT			
cg2-DTSpeech-NOSPAN-exclList.TXT			
exclusionList-K.TXT			
uwords-DTSpeech-2005-6.TXT			
uwords-DTSpeech-2006-7.TXT			

File name: uwords-DTSpeech-2005-6.TXT

Files of type: Text Files (*.TXT)

Next you will be asked for the 2-word concgram list to resume:

Figure 29. Open the 2-word concgram list to resume



Open 2-Word Concgrams List to Resume

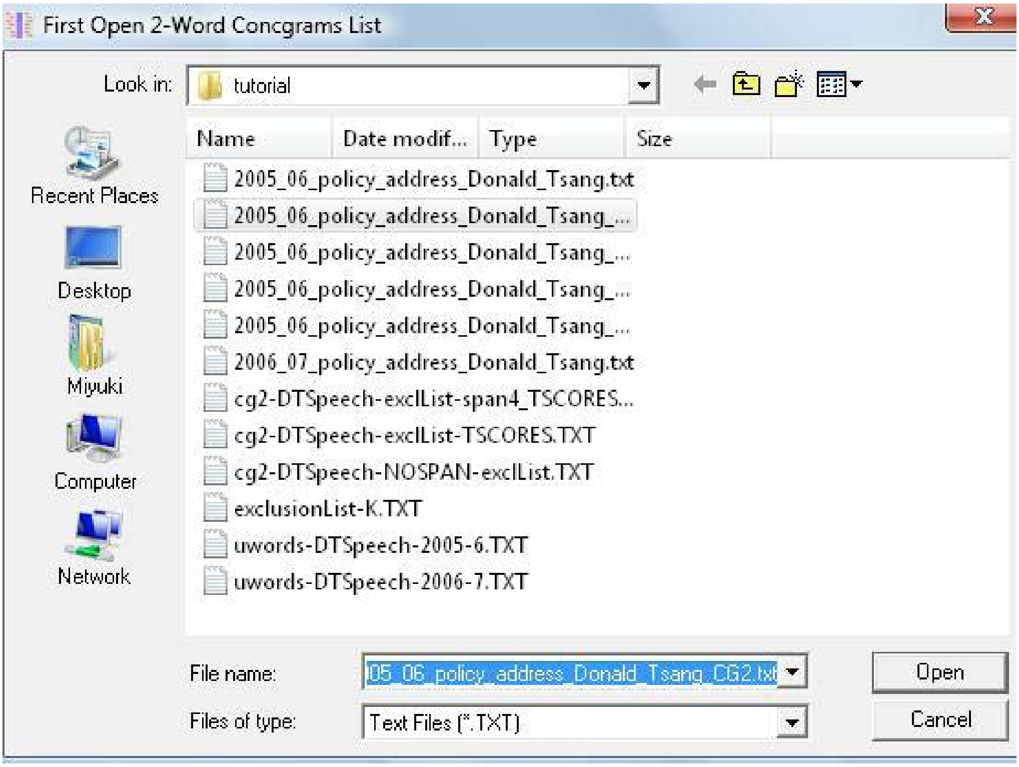
Look in: tutorial

Name	Date modif...	Type	Size
2005_06_policy_address_Donald_Tsang.txt			
2005_06_policy_address_Donald_Tsang_...			
2006_07_policy_address_Donald_Tsang.txt			
cg2-DTSpeech-exclList-span4_TSCORES...			
cg2-DTSpeech-exclList-TSCORES.TXT			
cg2-DTSpeech-NOSPAN-exclList.TXT			
exclusionList-K.TXT			
uwords-DTSpeech-2005-6.TXT			
uwords-DTSpeech-2006-7.TXT			

File name: 2005_06_policy_address_Donald_Tsang_CG2-

Files of type: Text Files (*.TXT)

Figure 30. First open the 2-word concgram list to resume



Look in: tutorial

2005_06_policy_address_Donald_Tsang.txt
2005_06_policy_address_Donald_Tsang_...
2005_06_policy_address_Donald_Tsang_...
2005_06_policy_address_Donald_Tsang_...
2005_06_policy_address_Donald_Tsang_...
2006_07_policy_address_Donald_Tsang.txt
cg2-DTSpeech-exclList-span4_TSCORES...
cg2-DTSpeech-exclList-TSCORES.TXT
cg2-DTSpeech-NOSPAN-exclList.TXT
exclusionList-K.TXT
uwords-DTSpeech-2005-6.TXT
uwords-DTSpeech-2006-7.TXT

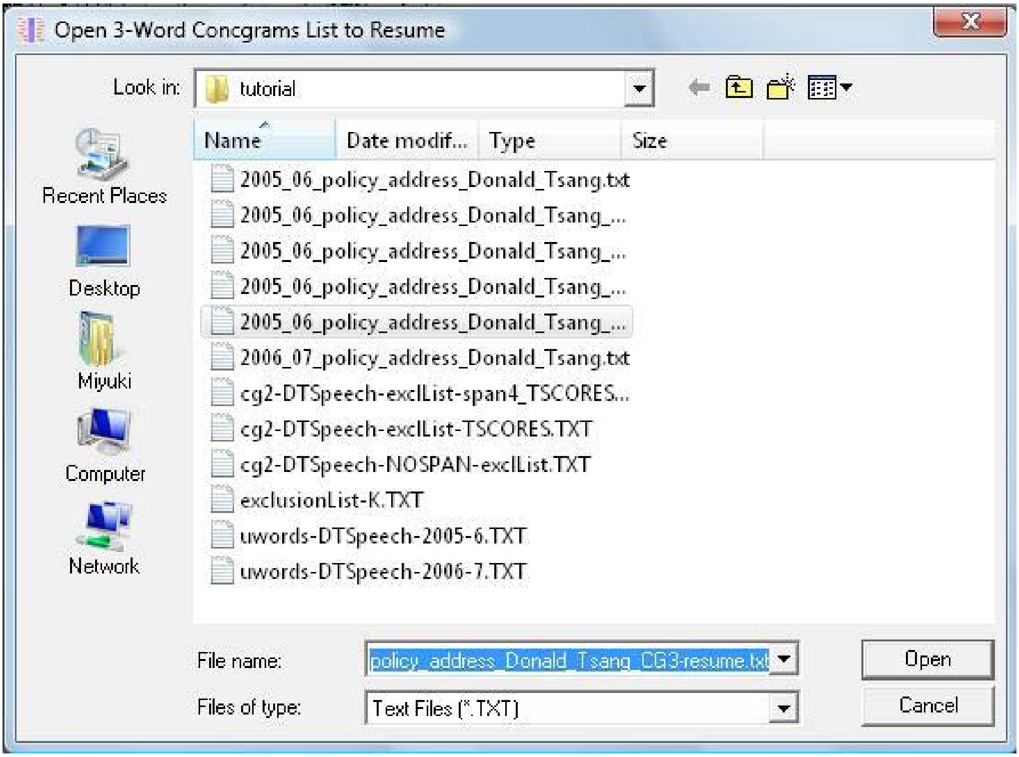
File name: 05_06_policy_address_Donald_Tsang_CG2.txt

Files of type: Text Files (*.TXT)

Resuming a 3-word concgram list requires that you first load the 2-word concgram list created from the original corpus file.

Then open the 3-word concgram list to resume:

Figure 31. Open the 3-word concgram list to resume



Look in: tutorial

2005_06_policy_address_Donald_Tsang.txt
2005_06_policy_address_Donald_Tsang_...
2005_06_policy_address_Donald_Tsang_...
2005_06_policy_address_Donald_Tsang_...
2005_06_policy_address_Donald_Tsang_...
2006_07_policy_address_Donald_Tsang.txt
cg2-DTSpeech-exclList-span4_TSCORES...
cg2-DTSpeech-exclList-TSCORES.TXT
cg2-DTSpeech-NOSPAN-exclList.TXT
exclusionList-K.TXT
uwords-DTSpeech-2005-6.TXT
uwords-DTSpeech-2006-7.TXT

File name: policy_address_Donald_Tsang_CG3-resume.txt

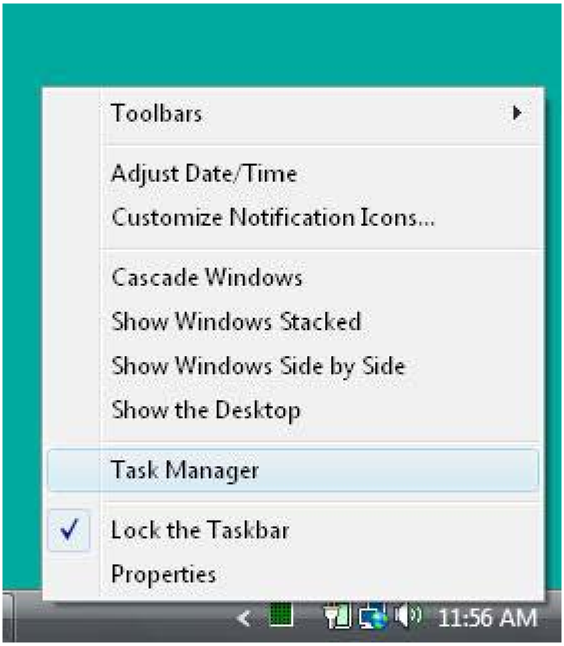
Files of type: Text Files (*.TXT)

Similarly, to resume a 4-word concgram list needs the original 3-word list to be loaded first.

4.3 Monitoring virtual memory

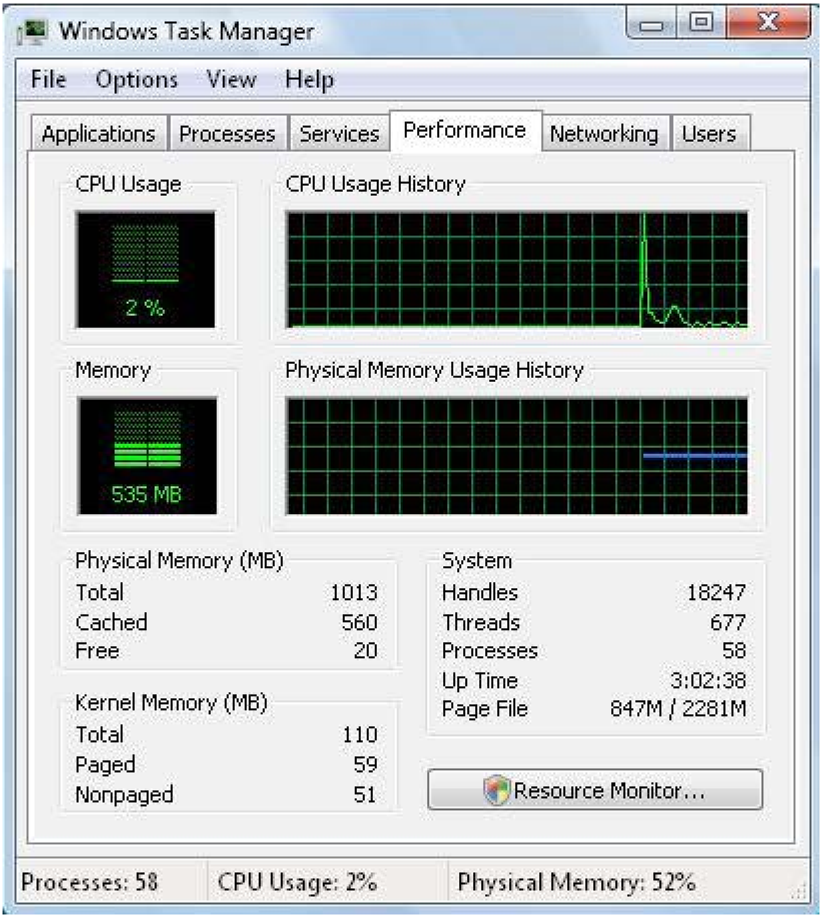
Creating lists using the automated searches can be very demanding on virtual memory use, and to avoid getting an "Out of virtual memory" error you may need to monitor the use of virtual memory when creating the lists. To do so, use the 'Task Manager' in Windows by right clicking the mouse in the bottom right hand corner (next to the clock time) and selecting 'Task Manager' from the popup menu that appears.

Figure 32. Selecting the Windows Task Manager



The Task Manager appears, and clicking the 'Performance' tab produces a display similar to the following:

Figure 33. The Windows Task Manager display after selecting the 'Performance' tab



Physical Memory (MB)
Total 1013
Cached 560
Free 20

Kernel Memory (MB)
Total 110
Paged 59
Nonpaged 51

System
Handles 18247
Threads 677
Processes 58
Up Time 3:02:38
Page File 847M / 2281M

Processes: 58 CPU Usage: 2% Physical Memory: 52%

This computer has 1014 MB of RAM, and 'virtual memory' can be monitored under the "Physical Memory Usage History". If the blue line reaches the top line (about 80% of the usage) the search results should be stopped and saved before resuming later.

Simply closing the search file is not enough to release all the memory allocated to the process, but you can free the memory by closing the program and restarting it. The performance monitor will then show that the virtual memory has been released and you can resume the list and the memory allocated will be less.

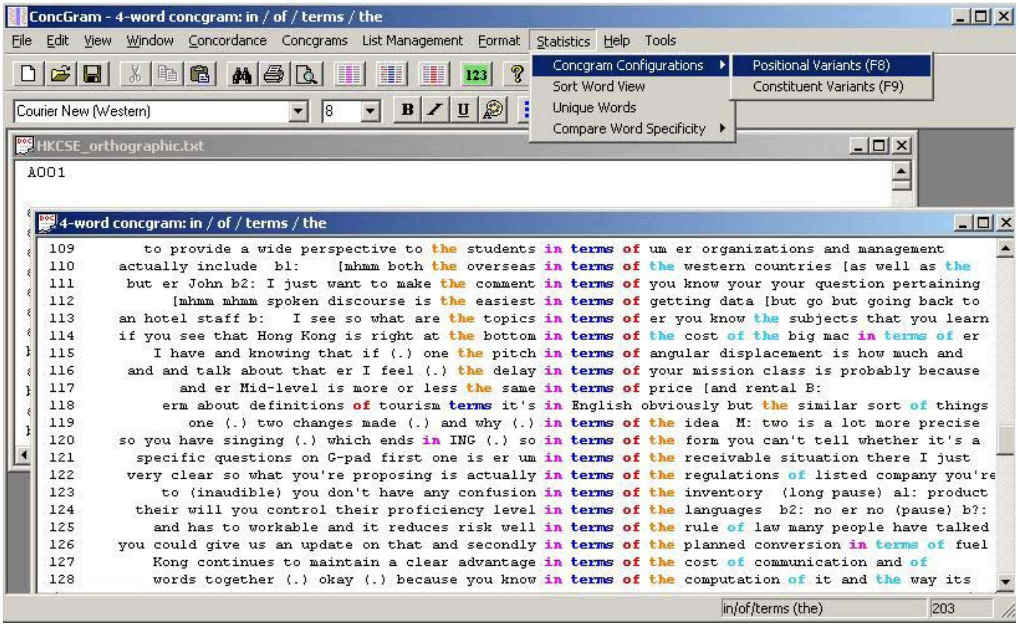
Chapter Five:

Using the Configuration List Boxes

5.1 Concgram configurations

The positional and constituent configurations of a concgram can be listed in the 'Concgram Configurations' List Boxes using 'Statistics >> Concgram Configurations' from the 'Statistics Menu'. You must first create a 'Concgram Search' window (Figure 34):

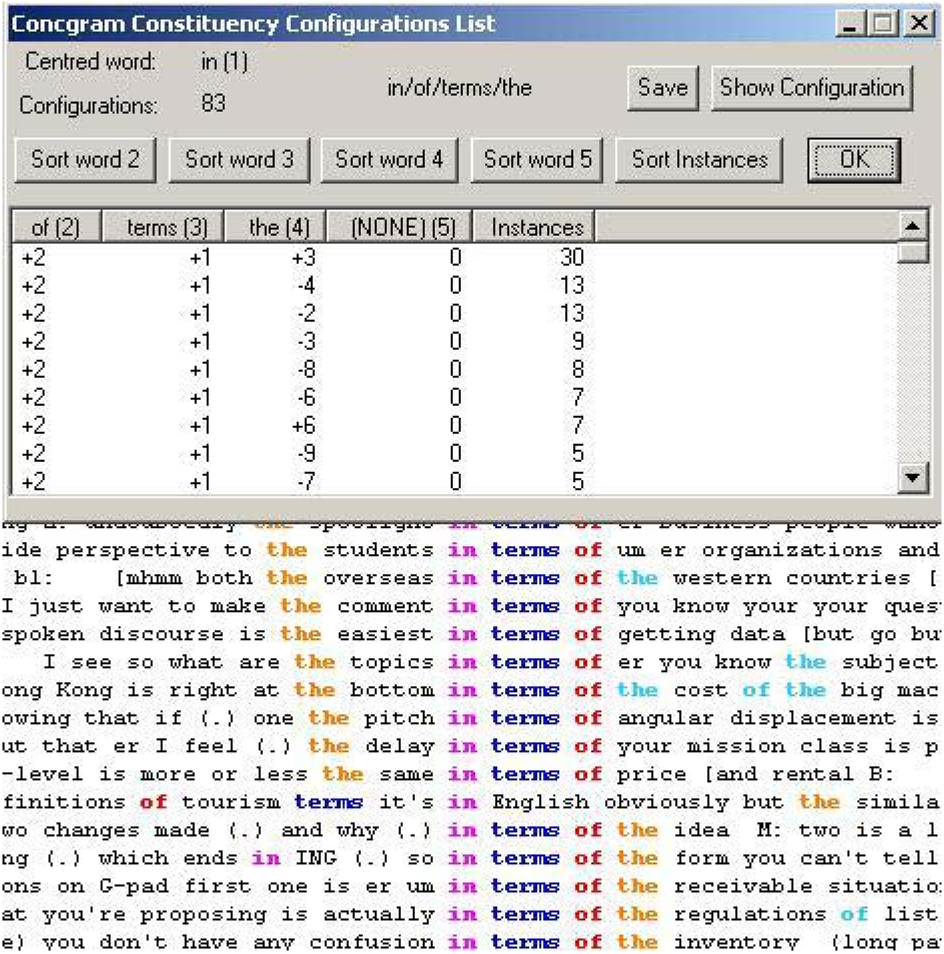
Figure 34. A Concgram Search window with the Concgram Configurations menu item selected


ConcGram - 4-word concgram: in / of / terms / the
File Edit View Window Concordance Concgrams List Management Format Statistics Help Tools
Concgram Configurations
Sort Word View
Unique Words
Compare Word Specificity
Positional Variants (F8)
Constituent Variants (F9)
Courier New (Western)
8
HKCSE_orthographic.txt
A001
4-word concgram: in / of / terms / the

```
109        to provide a wide perspective to the students in terms of um er organizations and management
110     actually include  b1:    [mhmm both the overseas in terms of the western countries [as well as the
111      but er John b2: I just want to make the comment in terms of you know your your question pertaining
112            [mhmm mhmm spoken discourse is the easiest in terms of getting data [but go but going back to
113     an hotel staff b:   I see so what are the topics in terms of er you know the subjects that you learn
114     if you see that Hong Kong is right at the bottom in terms of the cost of the big mac in terms of er
115         I have and knowing that if (.) one the pitch in terms of angular displacement is how much and
116      and and talk about that er I feel (.) the delay in terms of your mission class is probably because
117            and er Mid-level is more or less the same in terms of price [and rental B:
118          erm about definitions of tourism terms it's in English obviously but the similar sort of things
119             one (.) two changes made (.) and why (.) in terms of the idea  M: two is a lot more precise
120     so you have singing (.) which ends in ING (.) so in terms of the form you can't tell whether it's a
121       specific questions on G-pad first one is er um in terms of the receivable situation there I just
122      very clear so what you're proposing is actually in terms of the regulations of listed company you're
123           to (inaudible) you don't have any confusion in terms of the inventory  (long pause) al: product
124       their will you control their proficiency level in terms of the languages  b2: no er no (pause) b?:
125          and has to workable and it reduces risk well in terms of the rule of law many people have talked
126     you could give us an update on that and secondly in terms of the planned conversion in terms of fuel
127          Kong continues to maintain a clear advantage in terms of the cost of communication and of
128          words together (.) okay (.) because you know in terms of the computation of it and the way its
                                                                          in/of/terms (the)     203
```

Figure 35 shows the 'Concgram Constituency Configurations':

Figure 35. The Concgram Constituency Configurations List



Centred word: in (1)

Configurations: 83

in/of/terms/the

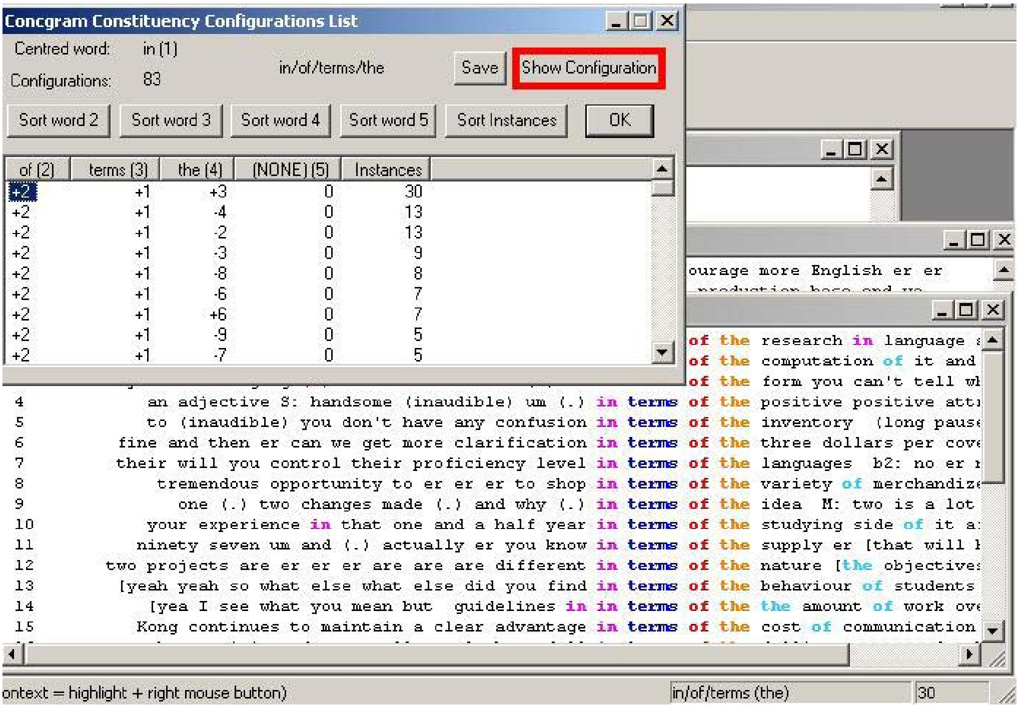
of (2)	terms (3)	the (4)	(NONE) (5)	Instances
+2	+1	+3	0	30
+2	+1	-4	0	13
+2	+1	-2	0	13
+2	+1	-3	0	9
+2	+1	-8	0	8
+2	+1	-6	0	7
+2	+1	+6	0	7
+2	+1	-9	0	5
+2	+1	-7	0	5

```
ng B: undoubtedly the spotlight in terms of er business people who
ide perspective to the students in terms of um er organizations and
 bl:     [mhmm both the overseas in terms of the western countries [
I just want to make the comment in terms of you know your your ques
spoken discourse is the easiest in terms of getting data [but go bu
   I see so what are the topics in terms of er you know the subject
ong Kong is right at the bottom in terms of the cost of the big mac
owing that if (.) one the pitch in terms of angular displacement is
ut that er I feel (.) the delay in terms of your mission class is p
-level is more or less the same in terms of price [and rental B:
finitions of tourism terms it's in English obviously but the simila
wo changes made (.) and why (.) in terms of the idea  M: two is a l
ng (.) which ends in ING (.) so in terms of the form you can't tell
ons on G-pad first one is er um in terms of the receivable situatio
at you're proposing is actually in terms of the regulations of list
e) you don't have any confusion in terms of the inventory  (long pa
```

Each configuration is listed separately, and sorted initially by the number of instances. The most common arrangement for this concgram is +1 +2 +3 from the centred word, accounting for 30 occurrences (+1 means 1 word to the right of the centred word, -1 means 1 word to the left).

These can be listed separately in a new window by selecting this item in the list and selecting the 'Show Configuration' button (Figure 36):

Figure 36. Results of clicking the 'Show Configuration' button (most frequent pattern)



Centred word: in (1)

Configurations: 83

in/of/terms/the

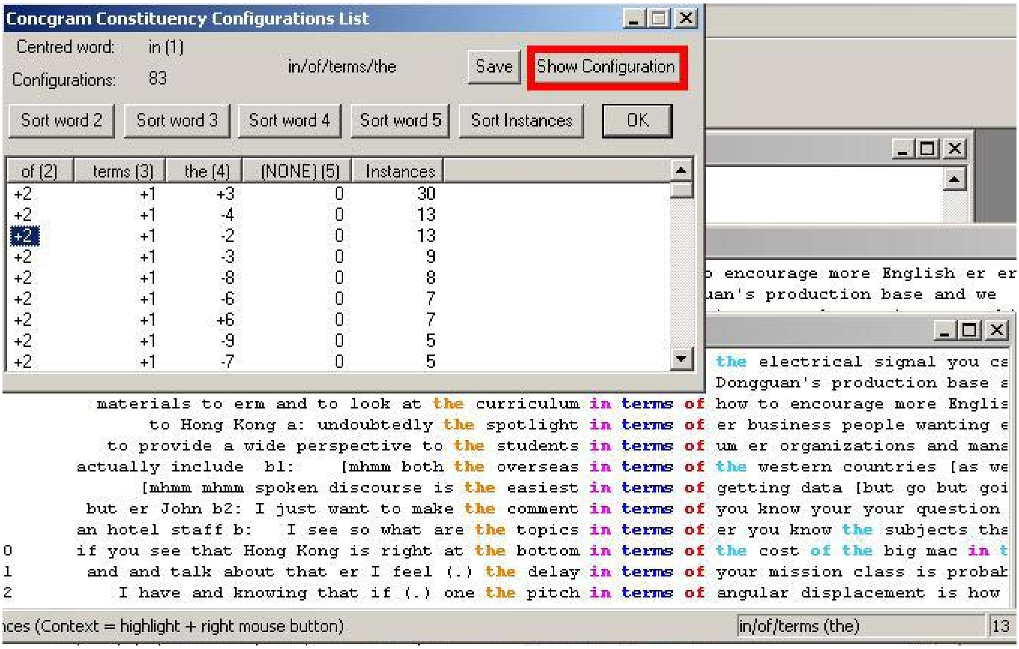
of (2)	terms (3)	the (4)	(NONE) (5)	Instances
+2	+1	+3	0	30
+2	+1	-4	0	13
+2	+1	-2	0	13
+2	+1	-3	0	9
+2	+1	-8	0	8
+2	+1	-6	0	7
+2	+1	+6	0	7
+2	+1	-9	0	5
+2	+1	-7	0	5

```
ourage more English er er
                                                                          of the research in language
                                                                          of the computation of it and
                                                                          of the form you can't tell wh
4        an adjective S: handsome (inaudible) um (.) in terms of the positive positive att
5        to (inaudible) you don't have any confusion in terms of the inventory  (long paus
6     fine and then er can we get more clarification in terms of the three dollars per cov
7    their will you control their proficiency level in terms of the languages  b2: no er
8         tremendous opportunity to er er er to shop in terms of the variety of merchandiz
9            one (.) two changes made (.) and why (.) in terms of the idea  M: two is a lot
10       your experience in that one and a half year in terms of the studying side of it a
11      ninety seven um and (.) actually er you know in terms of the supply er [that will
12    two projects are er er er are are are different in terms of the nature [the objective
13     [yeah yeah so what else what else did you find in terms of the behaviour of students
14         [yea I see what you mean but  guidelines in in terms of the the amount of work ov
15      Kong continues to maintain a clear advantage in terms of the cost of communication
```

ontext = highlight + right mouse button) in/of/terms (the) 30

The next most common pattern is +1+2–2 from the centred word (Figure 37):

Figure 37. Results of clicking the 'Show Configuration' button (next most frequent pattern)



Concgram Constituency Configurations List

Centred word: in (1)

Configurations: 83

in/of/terms/the

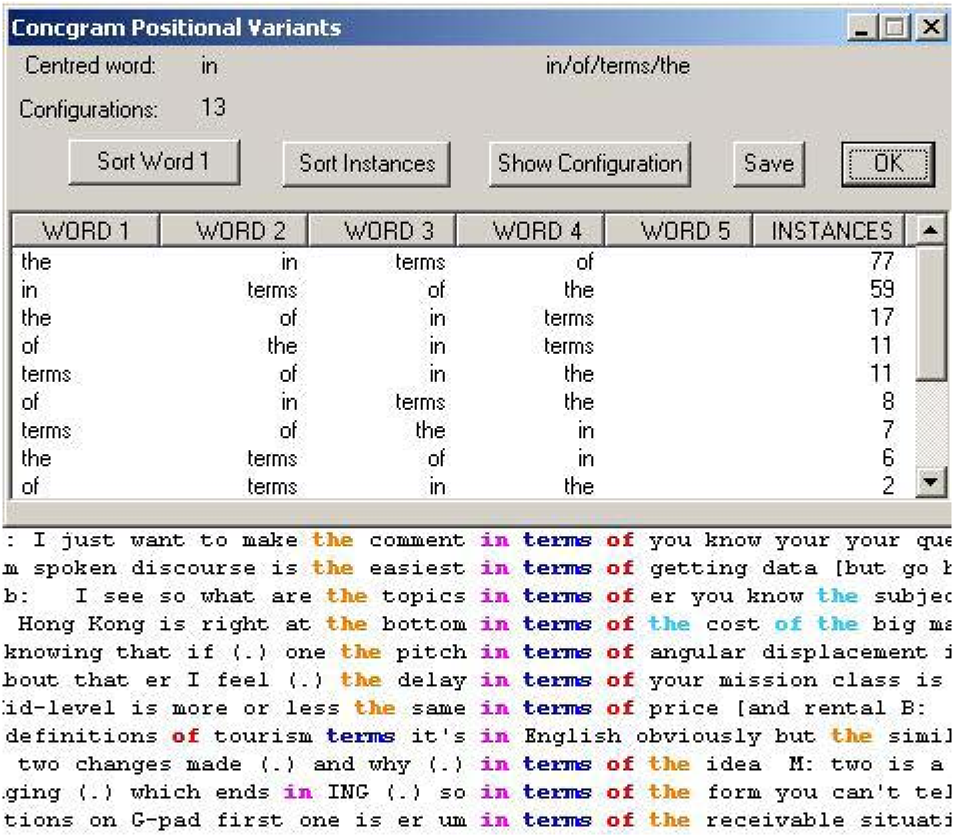
of (2)	terms (3)	the (4)	(NONE) (5)	Instances
+2	+1	+3	0	30
+2	+1	-4	0	13
+2	+1	-2	0	13
+2	+1	-3	0	9
+2	+1	-8	0	8
+2	+1	-6	0	7
+2	+1	+6	0	7
+2	+1	-9	0	5
+2	+1	-7	0	5

```
                                                                            o encourage more English er er
                                                                            uan's production base and we
                                                                              the electrical signal you ca
                                                                              Dongguan's production base a
         materials to erm and to look at the curriculum in terms of how to encourage more Englis
              to Hong Kong a: undoubtedly the spotlight in terms of er business people wanting e
          to provide a wide perspective to the students in terms of um er organizations and mana
        actually include  b1:    [mhmm both the overseas in terms of the western countries [as we
              [mhmm mhmm spoken discourse is the easiest in terms of getting data [but go but goi
         but er John b2: I just want to make the comment in terms of you know your your question
        an hotel staff b:   I see so what are the topics in terms of er you know the subjects tha
0       if you see that Hong Kong is right at the bottom in terms of the cost of the big mac in t
1        and and talk about that er I feel (.) the delay in terms of your mission class is probab
2           I have and knowing that if (.) one the pitch in terms of angular displacement is how
```

ces (Context = highlight + right mouse button) in/of/terms (the) 13

Lastly Figure 38 shows the List Box for configurations based on positional variation when this is selected. We can see from this that there are 13 variants, with 'the/in/terms/of' being the most frequent positional variant at 77 of the total occurrences, and 'in/terms/of/the' is the second most frequent positional variant with 59.

Figure 38. Concgram Positional Configurations List



Centred word: in in/of/terms/the

Configurations: 13

WORD 1	WORD 2	WORD 3	WORD 4	WORD 5	INSTANCES
the	in	terms	of		77
in	terms	of	the		59
the	of	in	terms		17
of	the	in	terms		11
terms	of	in	the		11
of	in	terms	the		8
terms	of	the	in		7
the	terms	of	in		6
of	terms	in	the		2

```
: I just want to make the comment in terms of you know your your que
m spoken discourse is the easiest in terms of getting data [but go b
b:   I see so what are the topics in terms of er you know the subjec
 Hong Kong is right at the bottom in terms of the cost of the big ma
knowing that if (.) one the pitch in terms of angular displacement i
bout that er I feel (.) the delay in terms of your mission class is
id-level is more or less the same in terms of price [and rental B:
definitions of tourism terms it's in English obviously but the simil
 two changes made (.) and why (.) in terms of the idea  M: two is a
ging (.) which ends in ING (.) so in terms of the form you can't tel
tions on G-pad first one is er um in terms of the receivable situati
```

Whether these word co-occurrences are significant is for the user to determine. By using the 'Concgram Configurations' List Boxes, we can find the most frequently occurring configurations of concgrams. Frequency of occurrence is one way of determining significance. Other techniques for determining significance using statistical tests, such as t-score, are discussed later in Chapter 7.

Chapter Six

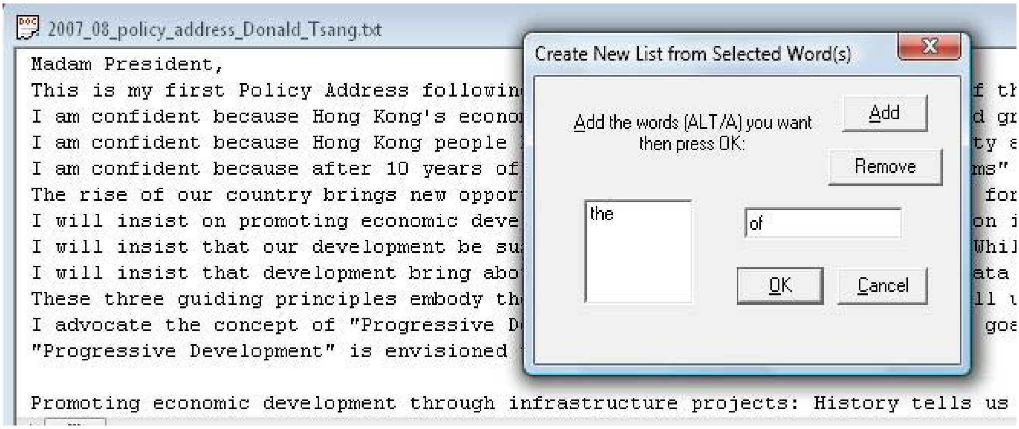
Automatic Searches from Specified Words

6.1 Specifying search words

When you make the lists using the automatic concgram search functions, you can create lists for words which you specify rather than using all the words in a text. Only the word or words you specify will be used for Origin Word searches, and only the words which are found to co-occur with these origins are stored in the concgrams list. Search words can be specified in two ways by using the 'Concgram Selected Words' Dialog, or an 'Inclusion List'.

Figure 39 below shows the 'Concgram Selected Words' Dialog, which appears when you select 'Concgrams >> Create New Concgram List (Automatic) >> Using Specified Words Only' from the Concgrams Menu. Type the word or words you want to use as Single Origins, and then select 'Add'. When you have finished, select 'OK'. Make sure you add a word at least once, or the Dialog will return an empty list.

Figure 39. The Concgram Selected Words Dialog


2007_08_policy_address_Donald_Tsang.txt
Create New List from Selected Word(s)
Add the words (ALT/A) you want then press OK:
Add
Remove
OK
Cancel

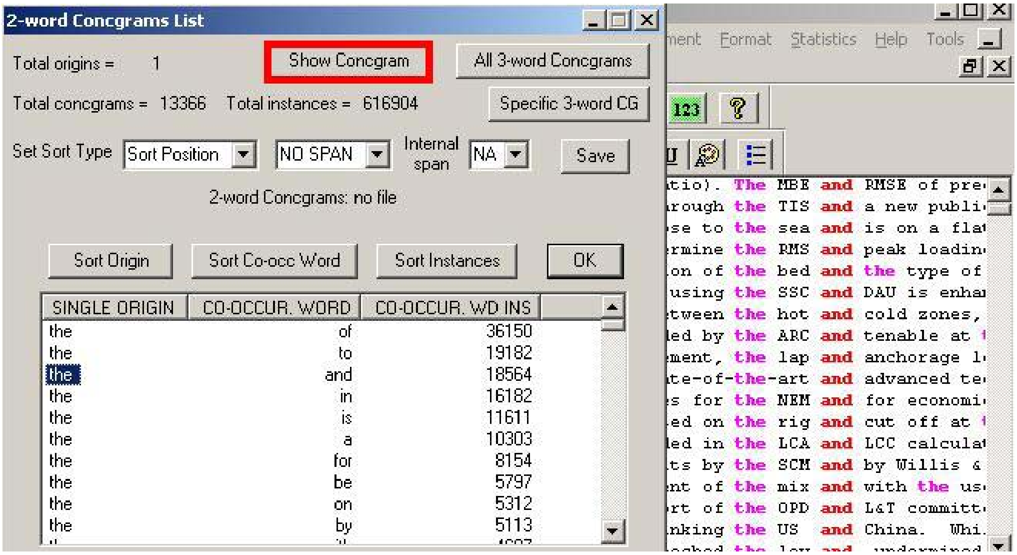
```
Madam President,
This is my first Policy Address followin                      f th
I am confident because Hong Kong's econo                      d gr
I am confident because Hong Kong people                       ty a
I am confident because after 10 years of                      ms"
The rise of our country brings new oppor                      for
I will insist on promoting economic deve                      on i
I will insist that our development be su                      Whil
I will insist that development bring abo                      ata
These three guiding principles embody th                      ll u
I advocate the concept of "Progressive D                      goa
"Progressive Development" is envisioned

Promoting economic development through infrastructure projects: History tells us
```

the

of

Figure 40 shows the resulting Concgrams List for 'the' and 'and' after the search words 'the/and' have been specified by using the 'Concgram Selected Words' Dialog.

Figure 40. Resulting Concgrams List with 'the/and' selected



Total origins = 1

Total concgrams = 13366 Total instances = 616904

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

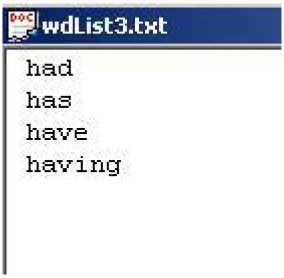
2-word Concgrams: no file

SINGLE ORIGIN	CO-OCCUR. WORD	CO-OCCUR. WD INS
the	of	36150
the	to	19182
the	and	18564
the	in	16182
the	is	11611
the	a	10303
the	for	8154
the	be	5797
the	on	5312
the	by	5113

```
tio). The MBE and RMSE of pre
rough the TIS and a new publi
se to the sea and is on a fla
rmine the RMS and peak loadin
on of the bed and the type of
using the SSC and DAU is enha
tween the hot and cold zones,
led by the ARC and tenable at
ment, the lap and anchorage l
te-of-the-art and advanced te
s for the NEM and for economi
ed on the rig and cut off at
led in the LCA and LCC calcula
ts by the SCM and by Willis &
nt of the mix and with the us
rt of the OPD and L&T committ
nking the US and China. Whi.
```

Search words can also be specified by adding them to an 'Inclusion List', by selecting 'Concgrams >> Create New Concgram List (Automatic) >> Using Only Words Specified in an Inclusion List' from the Concgrams Menu. The text is then searched only for words in the list, an example is shown below:

Figure 41. An Inclusion List



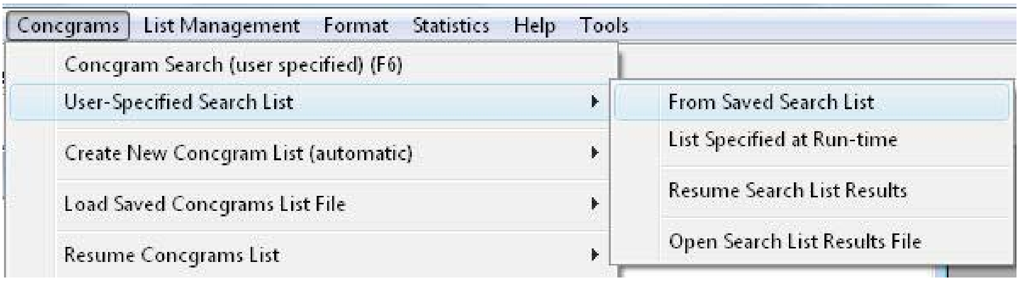
```
had
has
have
having
```

Inclusion Lists are useful if you have a number of words you want to investigate. You can use any text editor to create the lists, or use the 'List Management >> Make Inclusion List' function from the 'List Management' Menu. As Figure 41 shows, Inclusion Lists are also useful for lemmatizing irregular verbs, and a list of all irregular verbs and their lemmas would allow an automatic search to be performed with these words.

6.2 User Specified Search Lists

These lists are for creating 2-word concgrams from a specified list, and work by searching for each word in the list against every other word in the list.

Figure 42. Choosing the User Specified Search List

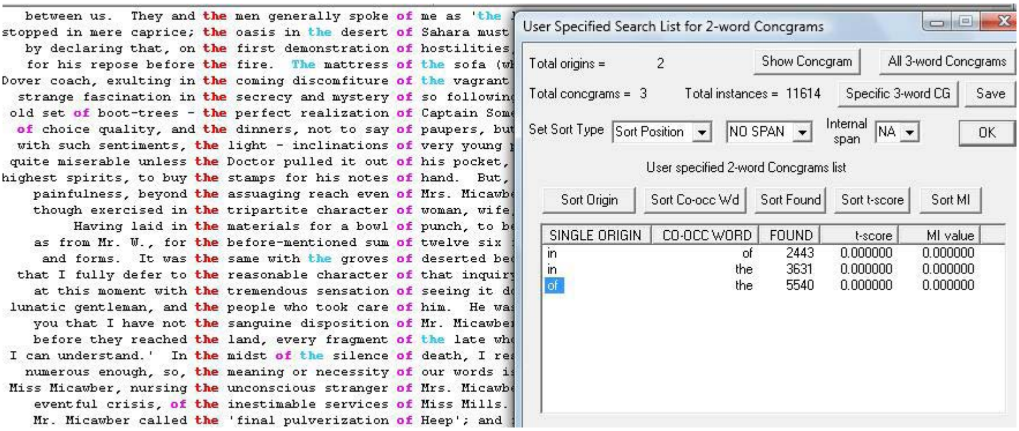

Concgrams
List Management
Format
Statistics
Help
Tools
Concgram Search (user specified) (F6)
User-Specified Search List
Create New Concgram List (automatic)
Load Saved Concgrams List File
Resume Concgrams List
From Saved Search List
List Specified at Run-time
Resume Search List Results
Open Search List Results File

The difference between the inclusion list described in **6.1** and that in **6.2** is that **6.1** is the same as performing an automated search with all the words which occur in the vicinity of the search word being listed, but only for the words which have been specified in the inclusion list, whereas in **6.2** the inclusion list is like performing a user nominated search for 2-word concgrams where each word in the list is one word in the concgram and every other word in the list becomes the other word in the 2-word concgram. This function is therefore suitable for searching for collocational and organizational frameworks, for example

For example, if the list contains the words "to, the, in, of" as a User Specified Search List, then the search will be for six 2-word concgrams "in/of, in/to, in/the", "of/the, of/to", and "the/to" (using no duplicates). Similarly, if the list contains the words "either, or, and, both, consequently" then the search will be for ten concgrams "and/either, and/or, and/both, and/consequently", "both/either, both/or, both/consequently", "consequently/either, consequently/or", and "either/or".

The list is created in the same way, one word per line, it is the program which treats them differently.

Figure 43. A User Specified Search List after clicking the ShowConcgram button



```
   between us.  They and the men generally spoke of me as 'the
stopped in mere caprice; the oasis in the desert of Sahara must
   by declaring that, on the first demonstration of hostilities
    for his repose before the fire.  The mattress of the sofa (w
Dover coach, exulting in the coming discomfiture of the vagrant
  strange fascination in the secrecy and mystery of so followin
 old set of boot-trees - the perfect realization of Captain Som
  of choice quality, and the dinners, not to say of paupers, bu
  with such sentiments, the light - inclinations of very young
 quite miserable unless the Doctor pulled it out of his pocket,
highest spirits, to buy the stamps for his notes of hand.  But,
    painfulness, beyond the assuaging reach even of Mrs. Micawb
    though exercised in the tripartite character of woman, wife
         Having laid in the materials for a bowl of punch, to b
    as from Mr. W., for the before-mentioned sum of twelve six
     and forms.  It was the same with the groves of deserted be
  that I fully defer to the reasonable character of that inquir
    at this moment with the tremendous sensation of seeing it d
 lunatic gentleman, and the people who took care of him.  He wa
    you that I have not the sanguine disposition of Mr. Micawbe
    before they reached the land, every fragment of the late wh
 I can understand.'  In the midst of the silence of death, I re
   numerous enough, so, the meaning or necessity of our words i
 Miss Micawber, nursing the unconscious stranger of Mrs. Micawb
    eventful crisis, of the inestimable services of Miss Mills.
    Mr. Micawber called the 'final pulverization of Heep'; and
```

User Specified Search List for 2-word Concgrams

Total origins = 2

Total concgrams = 3   Total instances = 11614

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

User specified 2-word Concgrams list

SINGLE ORIGIN	CO-OCC WORD	FOUND	t-score	MI value
in	of	2443	0.000000	0.000000
in	the	3631	0.000000	0.000000
of	the	5540	0.000000	0.000000

Chapter Seven:

Statistical tests

7.1 The t-score and MI tests

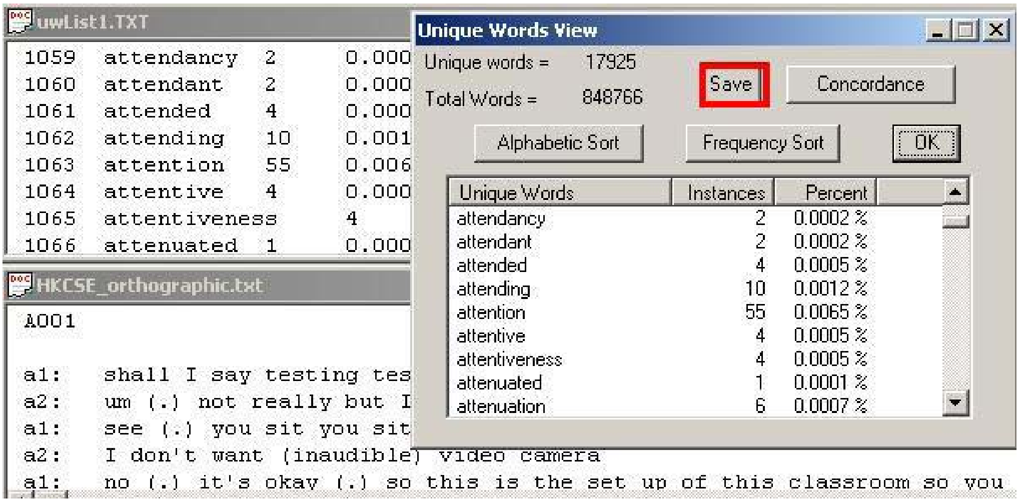
The reason for administering statistical tests using ConcGram is to attempt to calculate the significance of co-occurring words in context. While the fully automatic concgram search will find all of the contiguous and non-contiguous co-occurring words, including both constituent (AB, ACB) and positional (AB, BA) variations, that constitute 2-word, 3-word, 4-word and 5-word concgrams, there are co-occurrences that do not prove to be meaningfully associated when examined in context. For these reasons, ConcGram can run statistical tests to generate t-score and MI (Mutual Information) value to help to decide the statistically significant cut-offs for concgram lists, and to provide the user with indications as to which word co-occurrences are more likely to prove to be meaningful, and which ones the user can reasonably afford to ignore. Readers are advised to read a study by Stubbs (1995) which reviews the usefulness of such tests in language studies.

These tests are only available for 2-word concgrams as the formulas for calculating both t-score and MI value only provide values for the co-occurrence of 2 words. The formulas used for calculating both t-score and MI value and the cut-offs suggested are those given by Barnbrook (1996). Two steps are necessary before these tests can be listed:

1. You must first create an ordinary 2-word concgram list and save it.
2. A list of all the unique words for the corpus you are using must be created and saved.

To create the list of unique words for the corpus, select the 'Unique Words View' Dialog (Statistics >> Unique Words). The list gives the total number of instances for each word as well as the total number of words in the file, both of which are needed for the calculations (Figure 44):

Figure 44. Saved list of Unique Words



uwList1.TXT

```
1059   attendancy  2      0.000
1060   attendant   2      0.000
1061   attended    4      0.000
1062   attending   10     0.001
1063   attention   55     0.006
1064   attentive   4      0.000
1065   attentiveness      4
1066   attenuated  1      0.000
```

HKCSE_orthographic.txt

```
A001

a1:    shall I say testing tes
a2:    um (.) not really but I
a1:    see (.) you sit you sit
a2:    I don't want (inaudible) video camera
a1:    no (.) it's okay (.) so this is the set up of this classroom so you
```

Unique Words View

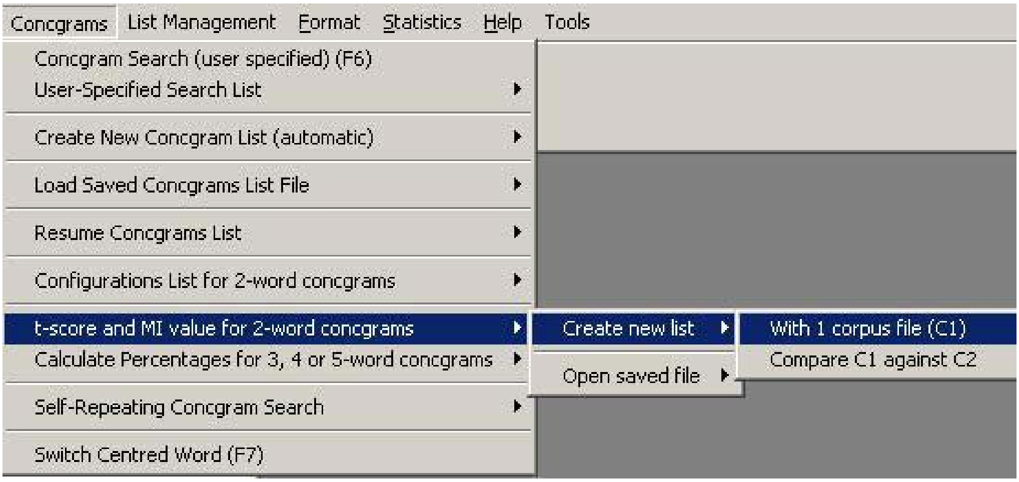
Unique words = 17925

Total Words = 848766

Unique Words	Instances	Percent
attendancy	2	0.0002 %
attendant	2	0.0002 %
attended	4	0.0005 %
attending	10	0.0012 %
attention	55	0.0065 %
attentive	4	0.0005 %
attentiveness	4	0.0005 %
attenuated	1	0.0001 %
attenuation	6	0.0007 %

The list can then be created by selecting from the 'Concgrams Menu', and the figures for both t-score and MI value listed next to each concgram (Figure 45).

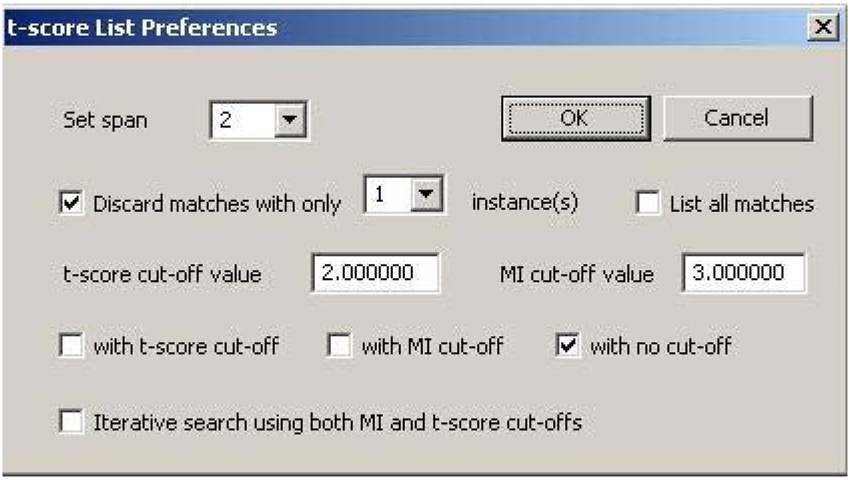
Figure 45. Concgrams Menu options to create a list with t-score and MI value for each concgram



Concgrams
List Management
Format
Statistics
Help
Tools
Concgram Search (user specified) (F6)
User-Specified Search List
Create New Concgram List (automatic)
Load Saved Concgrams List File
Resume Concgrams List
Configurations List for 2-word concgrams
t-score and MI value for 2-word concgrams
Calculate Percentages for 3, 4 or 5-word concgrams
Self-Repeating Concgram Search
Switch Centred Word (F7)
Create new list
Open saved file
With 1 corpus file (C1)
Compare C1 against C2

After selecting this menu item and selecting the saved 2-word concgram list to operate on, the user will be prompted to select from the options in the 't-score List Preferences' Dialog (Figure 46):

Figure 46. The t-score List Preferences Dialog



t-score List Preferences

Set span: 2

OK | Cancel

☑ Discard matches with only 1 instance(s)   ☐ List all matches

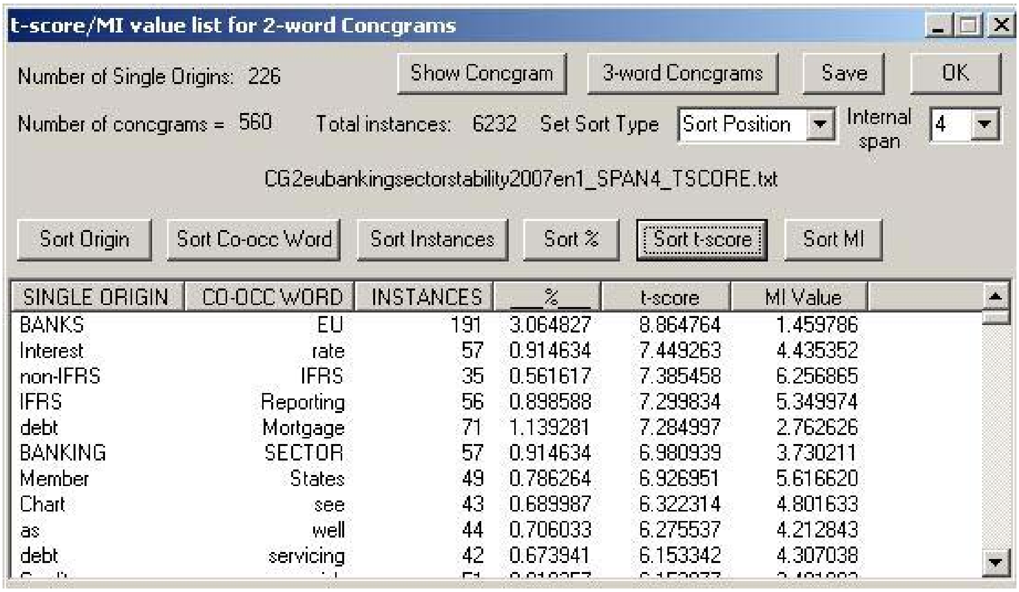
t-score cut-off value: 2.000000   MI cut-off value: 3.000000

☐ with t-score cut-off   ☐ with MI cut-off   ☑ with no cut-off

☐ Iterative search using both MI and t-score cut-offs

You can choose either t-score or MI cut-offs, use both, or have no cut-off at all. Using cut-offs greatly reduces the length of the resulting lists. After selecting any of these options, and the list for unique words to use for the figures, a search for all the 2-word concgrams listed is performed and the t-score and MI values for each will be calculated. Finally the 't-score List' Dialog will be displayed as in Figure 47 below:

Figure 47. The t-score List Dialog showing t-score and MI values listed for each 2-word concgram



t-score/MI value list for 2-word Concgrams

Number of Single Origins: 226

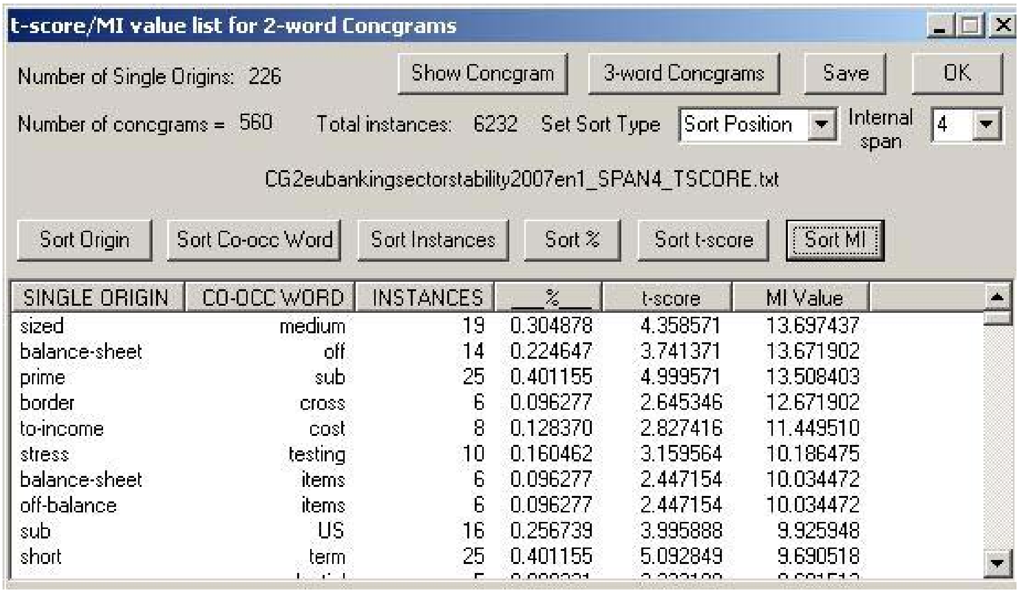
Number of concgrams = 560   Total instances: 6232   Set Sort Type: Sort Position   Internal span: 4

CG2eubankingsectorstability2007en1_SPAN4_TSCORE.txt

SINGLE ORIGIN	CO-OCC WORD	INSTANCES	%	t-score	MI Value
BANKS	EU	191	3.064827	8.864764	1.459786
Interest	rate	57	0.914634	7.449263	4.435352
non-IFRS	IFRS	35	0.561617	7.385458	6.256865
IFRS	Reporting	56	0.898588	7.299834	5.349974
debt	Mortgage	71	1.139281	7.284997	2.762626
BANKING	SECTOR	57	0.914634	6.980939	3.730211
Member	States	49	0.786264	6.926951	5.616620
Chart	see	43	0.689987	6.322314	4.801633
as	well	44	0.706033	6.275537	4.212843
debt	servicing	42	0.673941	6.153342	4.307038

Figure 47 shows a display which has been sorted by t-scores. Figure 48 shows that a sort by MI values would produce a different ranking. No grammatical words are included, only lexical words. If you want to exclude such words from the list, using an MI cut-off might be useful. But if you want to study grammatical words, using a t-score cut-off is more useful.

Figure 48. The t-score List Dialog sorted by MI values



Number of Single Origins: 226

Number of concgrams = 560 Total instances: 6232 Set Sort Type: Sort Position Internal span: 4

CG2eubankingsectorstability2007en1_SPAN4_TSCORE.txt

SINGLE ORIGIN	CO-OCC WORD	INSTANCES	%	t-score	MI Value
sized	medium	19	0.304878	4.358571	13.697437
balance-sheet	off	14	0.224647	3.741371	13.671902
prime	sub	25	0.401155	4.999571	13.508403
border	cross	6	0.096277	2.645346	12.671902
to-income	cost	8	0.128370	2.827416	11.449510
stress	testing	10	0.160462	3.159564	10.186475
balance-sheet	items	6	0.096277	2.447154	10.034472
off-balance	items	6	0.096277	2.447154	10.034472
sub	US	16	0.256739	3.995888	9.925948
short	term	25	0.401155	5.092849	9.690518

Chapter Eight:

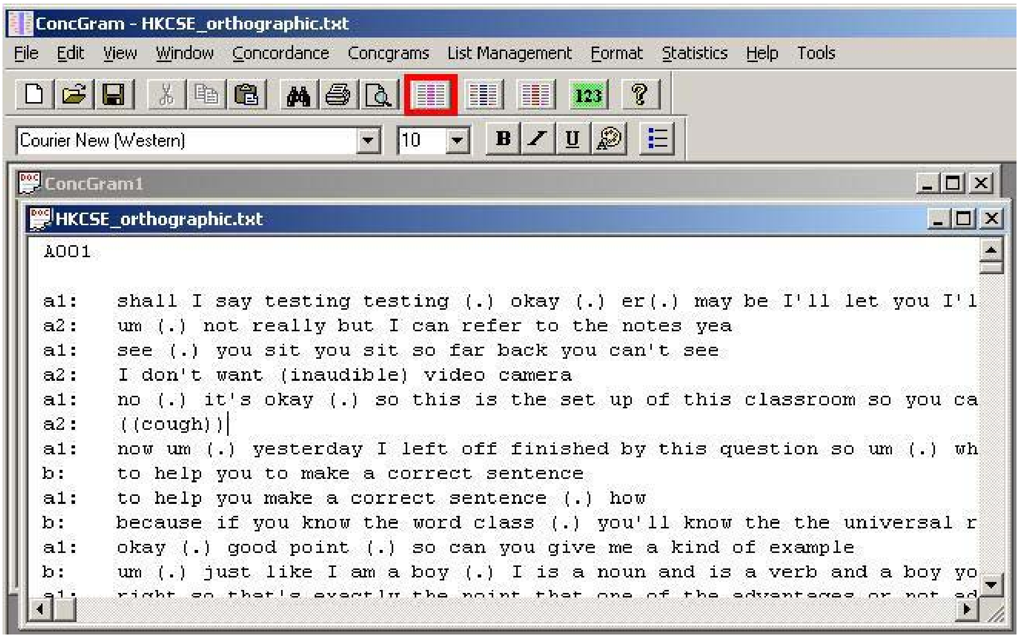
User nominated searches with ConcGram

8.1 Single word searches

We will illustrate user nominated searches with the use of a corpus file which has first been opened using the 'File >> Open' function, described in Chapter 3. With the user nominated search, ConcGram will search all the files which have been opened, so several files can be opened.

To start the search select 'Search' from the 'Concordance' Menu, or use the left search toolbar button as shown in Figure 49:

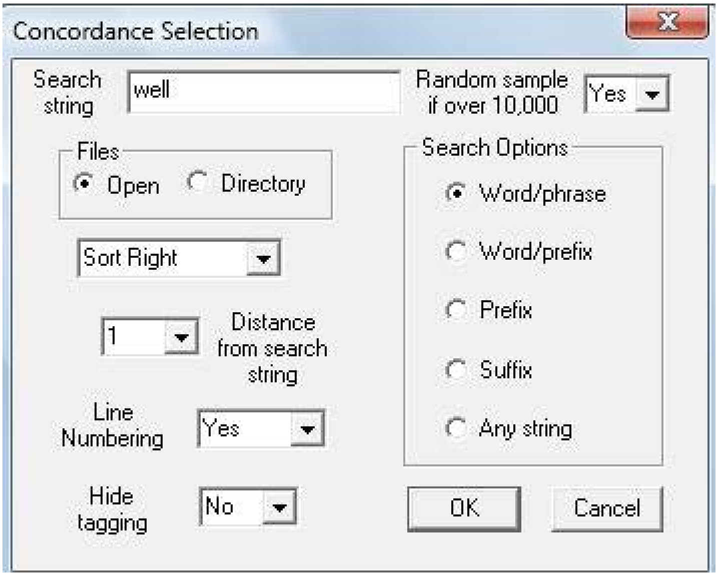
Figure 49. The User Nominated Single Word Concordance Search button



```
A001

a1:   shall I say testing testing (.) okay (.) er(.) may be I'll let you I'l
a2:   um (.) not really but I can refer to the notes yea
a1:   see (.) you sit you sit so far back you can't see
a2:   I don't want (inaudible) video camera
a1:   no (.) it's okay (.) so this is the set up of this classroom so you ca
a2:   ((cough))
a1:   now um (.) yesterday I left off finished by this question so um (.) wh
b:    to help you to make a correct sentence
a1:   to help you make a correct sentence (.) how
b:    because if you know the word class (.) you'll know the the universal r
a1:   okay (.) good point (.) so can you give me a kind of example
b:    um (.) just like I am a boy (.) I is a noun and is a verb and a boy yo
a1:   right so that's exactly the point that one of the advantages or not ad
```

The 'Concordance Selection' Dialog Box appears, as in Figure 50:

Figure 50. The Concordance Selection Dialog Box



Search string: well

Random sample if over 10,000: Yes

Files: ● Open ○ Directory

Sort Right

1 Distance from search string

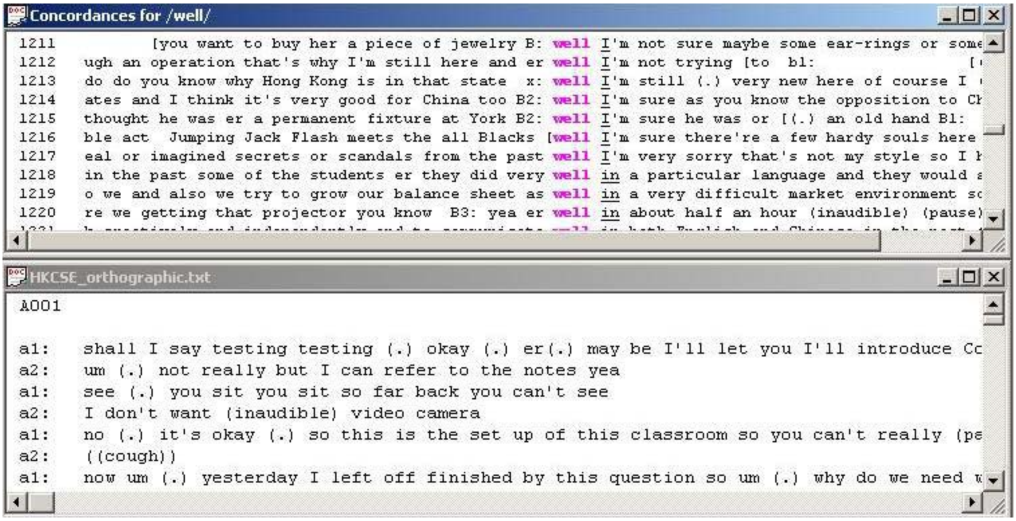
Line Numbering: Yes

Hide tagging: No

Search Options: ● Word/phrase ○ Word/prefix ○ Prefix ○ Suffix ○ Any string

Type in the search word (e.g. 'well') and leave the default settings to do a simple search using the open text file. Click 'OK' to start the search, and the concordance lines appear listed in a new window:

Figure 51. The Concordance Window

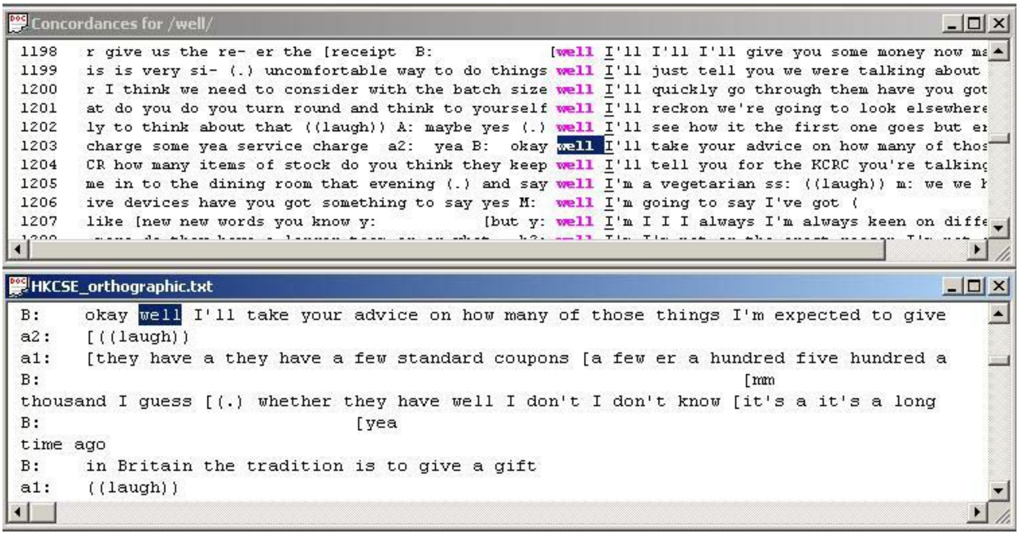


```
1211        [you want to buy her a piece of jewelry B: well I'm not sure maybe some ear-rings or some
1212   ugh an operation that's why I'm still here and er well I'm not trying [to  b1:              [
1213   do do you know why Hong Kong is in that state  x: well I'm still (.) very new here of course I
1214   ates and I think it's very good for China too B2: well I'm sure as you know the opposition to Ch
1215   thought he was er a permanent fixture at York B2: well I'm sure he was or [(.) an old hand B1:
1216   ble act  Jumping Jack Flash meets the all Blacks [well I'm sure there're a few hardy souls here
1217   eal or imagined secrets or scandals from the past well I'm very sorry that's not my style so I k
1218   in the past some of the students er they did very well in a particular language and they would a
1219   o we and also we try to grow our balance sheet as well in a very difficult market environment sc
1220   re we getting that projector you know  B3: yea er well in about half an hour (inaudible) (pause)
```

```
A001

a1:   shall I say testing testing (.) okay (.) er(.) may be I'll let you I'll introduce Cc
a2:   um (.) not really but I can refer to the notes yea
a1:   see (.) you sit you sit so far back you can't see
a2:   I don't want (inaudible) video camera
a1:   no (.) it's okay (.) so this is the set up of this classroom so you can't really (pe
a2:   ((cough))
a1:   now um (.) yesterday I left off finished by this question so um (.) why do we need w
```

In the above illustration, the windows have been tiled for easy viewing. Right click the mouse on any line in the concordance window to see the word in context in the text file which is displayed in the context window, as in Figure 52.

Figure 52. The Context Window



Concordances for /well/

```
1198   r give us the re- er the [receipt  B:           [well I'll I'll I'll give you some money now ma
1199   is is very si- (.) uncomfortable way to do things well I'll just tell you we were talking about
1200   r I think we need to consider with the batch size well I'll quickly go through them have you got
1201   at do you do you turn round and think to yourself well I'll reckon we're going to look elsewhere
1202   ly to think about that ((laugh)) A: maybe yes (.) well I'll see how it the first one goes but er
1203   charge some yea service charge  a2:  yea B:  okay well I'll take your advice on how many of thos
1204   CR how many items of stock do you think they keep well I'll tell you for the KCRC you're talking
1205   me in to the dining room that evening (.) and say well I'm a vegetarian ss: ((laugh)) m: we we h
1206   ive devices have you got something to say yes M:  well I'm going to say I've got (
1207   like [new new words you know y:          [but y: well I'm I I I always I'm always keen on diffe
```

HKCSE_orthographic.txt

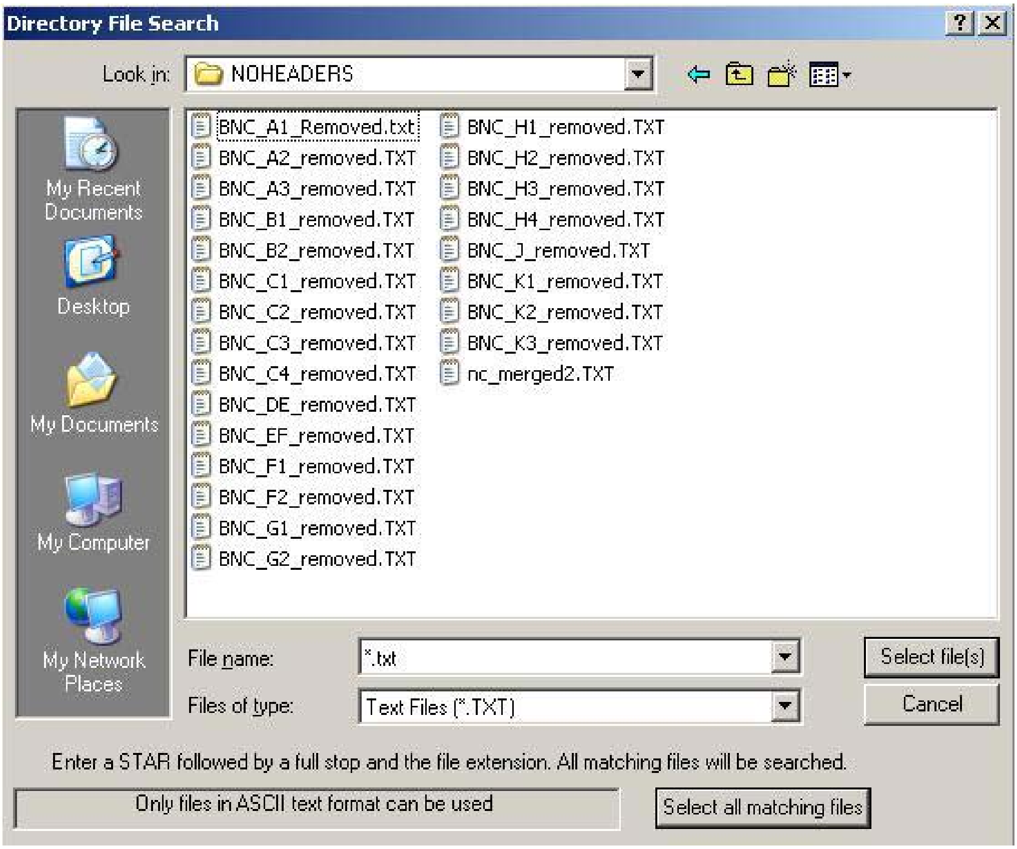
```
B:    okay well I'll take your advice on how many of those things I'm expected to give
a2:   [((laugh))
a1:   [they have a they have a few standard coupons [a few er a hundred five hundred a
B:                                                   [mm
thousand I guess [(.) whether they have well I don't I don't know [it's a it's a long
B:                             [yea
time ago
B:    in Britain the tradition is to give a gift
a1:   ((laugh))
```

8.2 Setting the concordance search options

This section describes the steps for setting the concordance search options:

1. With automatic concgram searches described in Chapter 6, only one file which has first been opened can be searched. With user nominated searches, however, files can be searched either as open files (more that one file can be opened and searched) or as directory files. The 'Directory File Search' (Figure 53) allows the user to select one or more files which have not been opened. The advantages are that the files can be larger than those which can be opened simultaneously, and the files are not left open after searching. You can select only the files you want to search or select all the files which have the same extension. Both options require that files must be in the same directory.

Figure 53. The Directory File Search Dialog



Look in: NOHEADERS

BNC_A1_Removed.txt
BNC_A2_removed.TXT
BNC_A3_removed.TXT
BNC_B1_removed.TXT
BNC_B2_removed.TXT
BNC_C1_removed.TXT
BNC_C2_removed.TXT
BNC_C3_removed.TXT
BNC_C4_removed.TXT
BNC_DE_removed.TXT
BNC_EF_removed.TXT
BNC_F1_removed.TXT
BNC_F2_removed.TXT
BNC_G1_removed.TXT
BNC_G2_removed.TXT
BNC_H1_removed.TXT
BNC_H2_removed.TXT
BNC_H3_removed.TXT
BNC_H4_removed.TXT
BNC_J_removed.TXT
BNC_K1_removed.TXT
BNC_K2_removed.TXT
BNC_K3_removed.TXT
nc_merged2.TXT

File name: *.txt

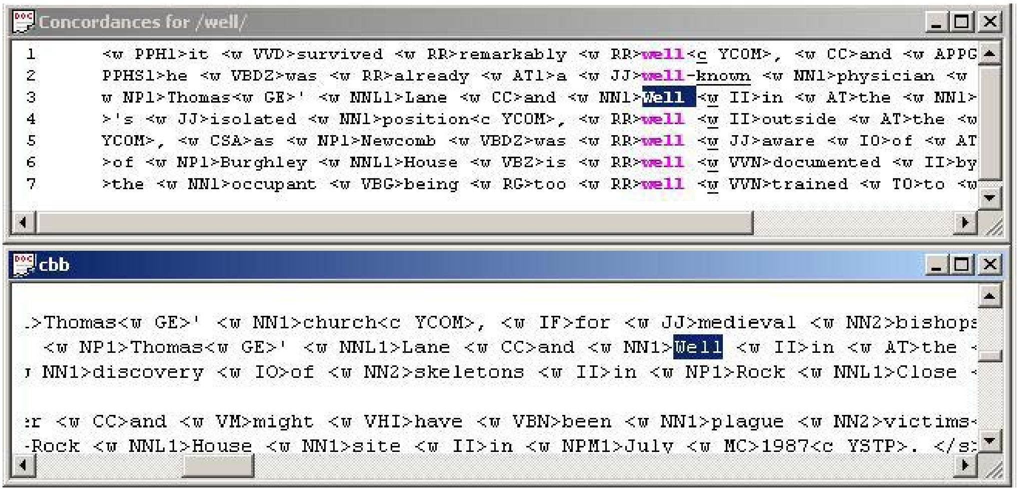
Files of type: Text Files (*.TXT)

Enter a STAR followed by a full stop and the file extension. All matching files will be searched.

Only files in ASCII text format can be used

2. The next choice is to select 'Sort Right' (default), 'Sort Left', or 'Unsorted' in the 'Sort' option. The function sorts alphabetically on the co-occurring word to the right or left of the search word. You can set the distance from the search word at 1, 2 or 3 words away.
3. Choose whether or not you want to have the concordance lines numbered (default is 'Yes').
4. 'Hide tagging' refers to if you are using a corpus which has been tagged. If you are not, this setting does not apply. However, if you are using a tagged corpus, you can change this option to 'Yes' if you want to view the concordance lines without the tags. You can see the results of searching with and without tags in Figures 54 and 55 (the 'Hide Tagging' function only applies to the concordance lines, not the context view).

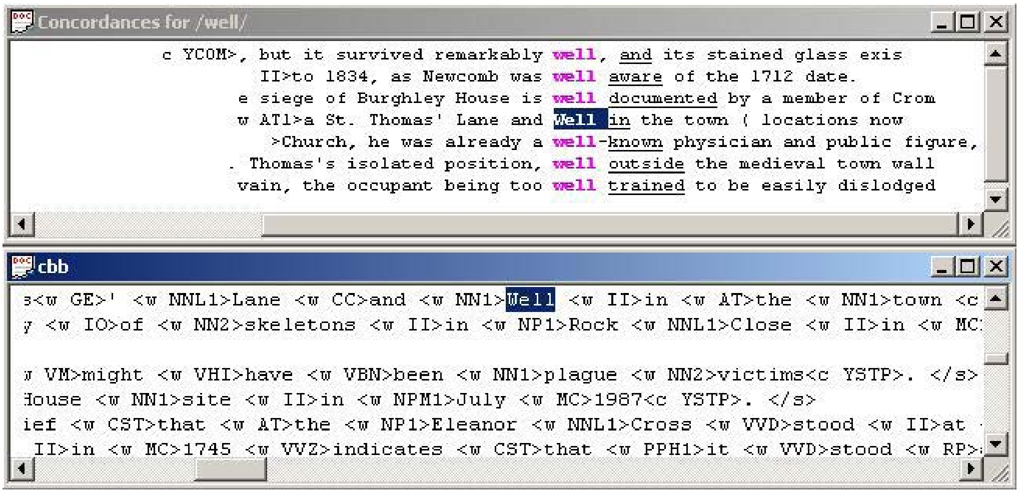
Figure 54. A concordance search with tagging shown in the Concordance lines



```
1       <w PPH1>it <w VVD>survived <w RR>remarkably <w RR>well<c YCOM>, <w CC>and <w APPG
2       PPHS1>he <w VBDZ>was <w RR>already <w AT1>a <w JJ>well-known <w NN1>physician <w
3       w NP1>Thomas<w GE>' <w NNL1>Lane <w CC>and <w NN1>Well <w II>in <w AT>the <w NN1>
4       >'s <w JJ>isolated <w NN1>position<c YCOM>, <w RR>well <w II>outside <w AT>the <w
5       YCOM>, <w CSA>as <w NP1>Newcomb <w VBDZ>was <w RR>well <w JJ>aware <w IO>of <w AT
6       >of <w NP1>Burghley <w NNL1>House <w VBZ>is <w RR>well <w VVN>documented <w II>by
7       >the <w NN1>occupant <w VBG>being <w RG>too <w RR>well <w VVN>trained <w TO>to <w
```

```
.>Thomas<w GE>' <w NN1>church<c YCOM>, <w IF>for <w JJ>medieval <w NN2>bishops
 <w NP1>Thomas<w GE>' <w NNL1>Lane <w CC>and <w NN1>Well <w II>in <w AT>the <
ı NN1>discovery <w IO>of <w NN2>skeletons <w II>in <w NP1>Rock <w NNL1>Close <

:r <w CC>and <w VM>might <w VHI>have <w VBN>been <w NN1>plague <w NN2>victims<
-Rock <w NNL1>House <w NN1>site <w II>in <w NPM1>July <w MC>1987<c YSTP>. </s>
```

Figure 55. A concordance search with tagging hidden in the Concordance lines



```
        c YCOM>, but it survived remarkably well, and its stained glass exis
                 II>to 1834, as Newcomb was well aware of the 1712 date.
               e siege of Burghley House is well documented by a member of Crom
               w AT1>a St. Thomas' Lane and Well in the town ( locations now
                  >Church, he was already a well-known physician and public figure,
              . Thomas's isolated position, well outside the medieval town wall
               vain, the occupant being too well trained to be easily dislodged
```

```
s<w GE>' <w NNL1>Lane <w CC>and <w NN1>Well <w II>in <w AT>the <w NN1>town <c
y <w IO>of <w NN2>skeletons <w II>in <w NP1>Rock <w NNL1>Close <w II>in <w MC:

w VM>might <w VHI>have <w VBN>been <w NN1>plague <w NN2>victims<c YSTP>. </s>
House <w NN1>site <w II>in <w NPM1>July <w MC>1987<c YSTP>. </s>
ief <w CST>that <w AT>the <w NP1>Eleanor <w NNL1>Cross <w VVD>stood <w II>at
 II>in <w MC>1745 <w VVZ>indicates <w CST>that <w PPH1>it <w VVD>stood <w RP>:
```

5. The remaining 'Search Options' refer to the type of search you wish to make. Only one of these buttons can be selected for each search. They refer to search strings, as follows:

   a. **'Word/phrase'.** This searches for any word or phrase (e.g. 'going to') which is bounded by a non- alphabetic character to the left and right.
   b. **'Word/prefix'.** This looks for a word or prefix which includes the word, for example, entering 'go' will find all instances of 'go', 'goes', 'going', 'got', etc.

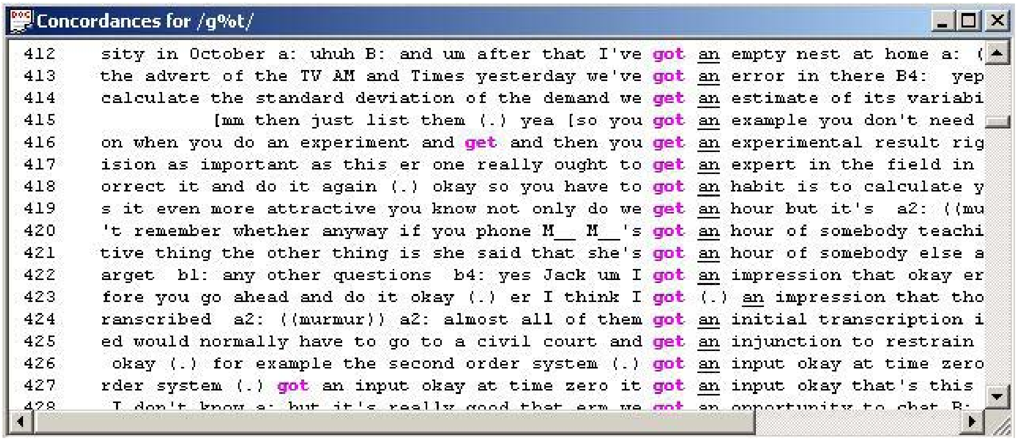
c. **'Prefix'.** Only where the search string is a prefix will the instance be stored, so 'go', if entered as a prefix, will find 'goes', 'going', 'got', etc., but not the word 'go' on its own.

d. '**Suffix'.** Only where the search string is a SUFFIX will the instance be stored, so 'go', if entered as a suffix, will find 'ago', 'undergo', 'Chicago', etc., but not the word 'go' on its own.

e. **'Any string'**. This will find all instances of the search string, whether as a word or anywhere as part of another string. So if 'go' is entered, all instances where 'go' occurs will be stored, for example, 'go', 'undergo', 'going' and 'negotiate' will all be listed.

6. A wildcard '%' can be used to represent any single letter, for example, 'g%t' will find both 'get' and 'got' (Figure 56):

Figure 56. Results of using the wildcard '%' for 'g%t'



Concordances for /g%t/

```
412    sity in October a: uhuh B: and um after that I've got an empty nest at home a: (
413    the advert of the TV AM and Times yesterday we've got an error in there B4:  yep
414    calculate the standard deviation of the demand we get an estimate of its variabi
415              [mm then just list them (.) yea [so you got an example you don't need
416    on when you do an experiment and get and then you get an experimental result rig
417    ision as important as this er one really ought to get an expert in the field in
418    orrect it and do it again (.) okay so you have to got an habit is to calculate y
419    s it even more attractive you know not only do we get an hour but it's  a2: ((mu
420    't remember whether anyway if you phone M__ M__'s got an hour of somebody teachi
421    tive thing the other thing is she said that she's got an hour of somebody else a
422    arget  b1: any other questions  b4: yes Jack um I got an impression that okay er
423    fore you go ahead and do it okay (.) er I think I got (.) an impression that tho
424    ranscribed  a2: ((murmur)) a2: almost all of them got an initial transcription i
425    ed would normally have to go to a civil court and get an injunction to restrain
426     okay (.) for example the second order system (.) got an input okay at time zero
427    rder system (.) got an input okay at time zero it got an input okay that's this
428     I don't know a: but it's really good that erm we got an opportunity to chat B:
```

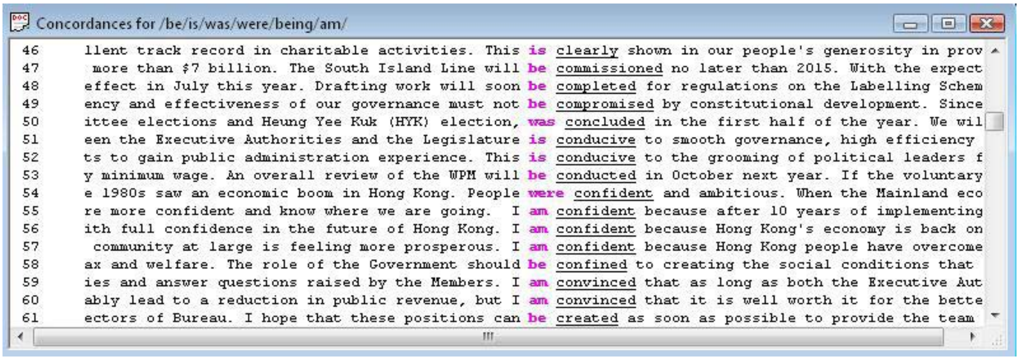
In the 'Concgram Search' Dialog Box, there are a number of options which can be selected, many of which are the same search options as for a single word concordance. First enter the word which is to be centred, and set the search options as given. Two of these options are also available for the first co-occurring word in the concgram (a concgram search must have at least 'one' co-occurring word). The co-occurring word or words may be anywhere in the string, or only to the left or right of the centred word.

7. There are two ways you can **lemmatize** a search — either by using the Concordance Search Dialog, or by entering the words separately using the Concgram >> Create New Concgram List >> Using Specified Words ONLY (or use an Inclusion Word

List). The outputs are different, and if you use the Concordance Search Dialog then the words will all be mixed together, and the totals accordingly, whereas with the specified words option all the words are treated individually and their totals are measured separately. Words must be separated by a forward slash — for example, by entering "a/an" in the search edit box will search for "a" and "an", or entering "be/is/was/were/being/am" will search for all 6 forms of the verb "BE", as shown below.

8. The dropdown box next to the Search String edit box applies to words such as "THE" which may have more than 10,000 occurrences in a corpus, and the concordance lines will therefore be greater accordingly. By default, the program limits the display to no more than 9,999 which are taken as a random selection from the total number. If you want all concordances to be displayed the dropdown box must be set to "NO". All the concordances will then be displayed, but the colouring of the centred and co-occuring words will take longer to format, and can be cancelled if this is not required. The actual number of occurrences of the Search String is printed in the bottom right-hand corner of the display.

Figure 57. Results of searching for the lemma "be/is/was/were/being/am"



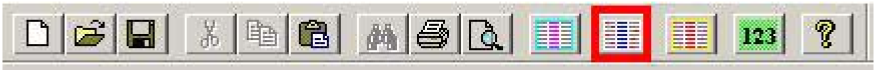
Concordances for /be/is/was/were/being/am/

```
46     llent track record in charitable activities. This is clearly shown in our people's generosity in prov
47      more than $7 billion. The South Island Line will be commissioned no later than 2015. With the expect
48     effect in July this year. Drafting work will soon be completed for regulations on the Labelling Schem
49     ency and effectiveness of our governance must not be compromised by constitutional development. Since
50     ittee elections and Heung Yee Kuk (HYK) election, was concluded in the first half of the year. We wil
51     een the Executive Authorities and the Legislature is conducive to smooth governance, high efficiency
52     ts to gain public administration experience. This is conducive to the grooming of political leaders f
53     y minimum wage. An overall review of the WPM will be conducted in October next year. If the voluntary
54     e 1980s saw an economic boom in Hong Kong. People were confident and ambitious. When the Mainland eco
55     re more confident and know where we are going.  I am confident because after 10 years of implementing
56     ith full confidence in the future of Hong Kong. I am confident because Hong Kong's economy is back on
57      community at large is feeling more prosperous. I am confident because Hong Kong people have overcome
58     ax and welfare. The role of the Government should be confined to creating the social conditions that
59     ies and answer questions raised by the Members. I am convinced that as long as both the Executive Aut
60     ably lead to a reduction in public revenue, but I am convinced that it is well worth it for the bette
61     ectors of Bureau. I hope that these positions can be created as soon as possible to provide the team
```

8.3 The user nominated concgram search

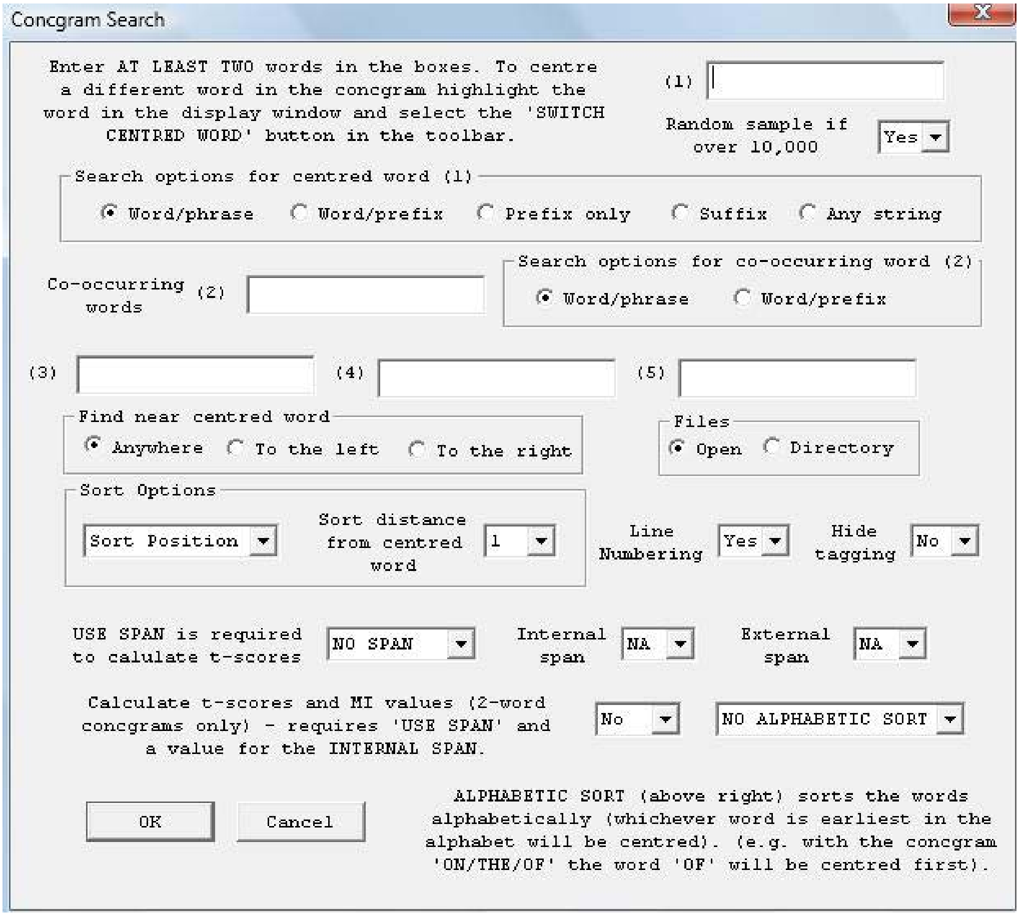
This function is available from the 'Concgrams >> Concgram Search' menu option, or by clicking the centre search button in the toolbar (Figure 58):

Figure 58. The Concgram Search button in the toolbar



Selecting this option initially presents the user with the 'Concgram Search' Dialog Box, shown in Figure 59.

Figure 59. The Concgram Search Dialog Box



Concgram Search

Enter AT LEAST TWO words in the boxes. To centre a different word in the concgram highlight the word in the display window and select the 'SWITCH CENTRED WORD' button in the toolbar.

(1)

Random sample if over 10,000 Yes

Search options for centred word (1)
● Word/phrase ○ Word/prefix ○ Prefix only ○ Suffix ○ Any string

Co-occurring words (2)

Search options for co-occurring word (2)
● Word/phrase ○ Word/prefix

(3) (4) (5)

Find near centred word
● Anywhere ○ To the left ○ To the right

Files
● Open ○ Directory

Sort Options
Sort Position
Sort distance from centred word 1

Line Numbering Yes

Hide tagging No

USE SPAN is required to calulate t-scores NO SPAN

Internal span NA

External span NA

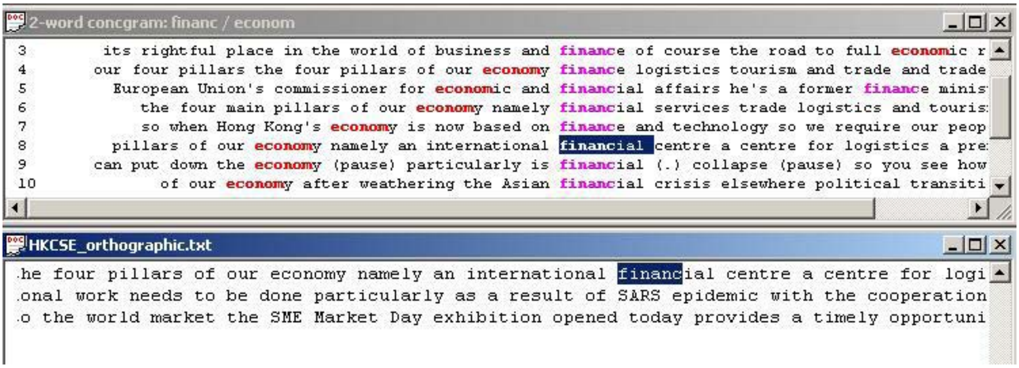
Calculate t-scores and MI values (2-word concgrams only) - requires 'USE SPAN' and a value for the INTERNAL SPAN. No

NO ALPHABETIC SORT

ALPHABETIC SORT (above right) sorts the words alphabetically (whichever word is earliest in the alphabet will be centred). (e.g. with the concgram 'ON/THE/OF' the word 'OF' will be centred first).

As well as the same search options available for the single word search, the 'Concgram Search' Dialog has a number of additional options for the user to select, as required. Both the first and second words in the concgram can be selected for word prefixes, and Figure 60 shows the display resulting for a search for 'financ' and 'econom' with the 'Word/Prefix' option selected for both.

Figure 60. Results of selecting 'financ' and 'econom' both with the 'Word/prefix' option

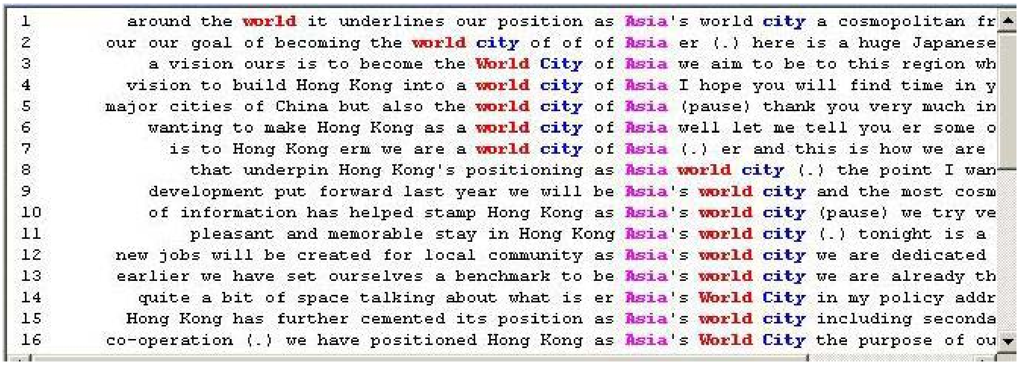


```
3        its rightful place in the world of business and finance of course the road to full economic r
4       our four pillars the four pillars of our economy finance logistics tourism and trade and trade
5        European Union's commissioner for economic and financial affairs he's a former finance minis
6            the four main pillars of our economy namely financial services trade logistics and touris
7             so when Hong Kong's economy is now based on finance and technology so we require our peop
8        pillars of our economy namely an international financial centre a centre for logistics a pre
9      can put down the economy (pause) particularly is financial (.) collapse (pause) so you see how
10             of our economy after weathering the Asian financial crisis elsewhere political transiti
```

```
.he four pillars of our economy namely an international financial centre a centre for logi
.onal work needs to be done particularly as a result of SARS epidemic with the cooperation
.o the world market the SME Market Day exhibition opened today provides a timely opportuni
```

The 'Sort' options are the same as for the single word concordancer, but by default an additional option is selected for concgrams, 'Sort Position'. This sorts according to the character positions of the co-occurring words relative to the centred word, as illustrated in Figure 61:

Figure 61. The 'Sort Position' option

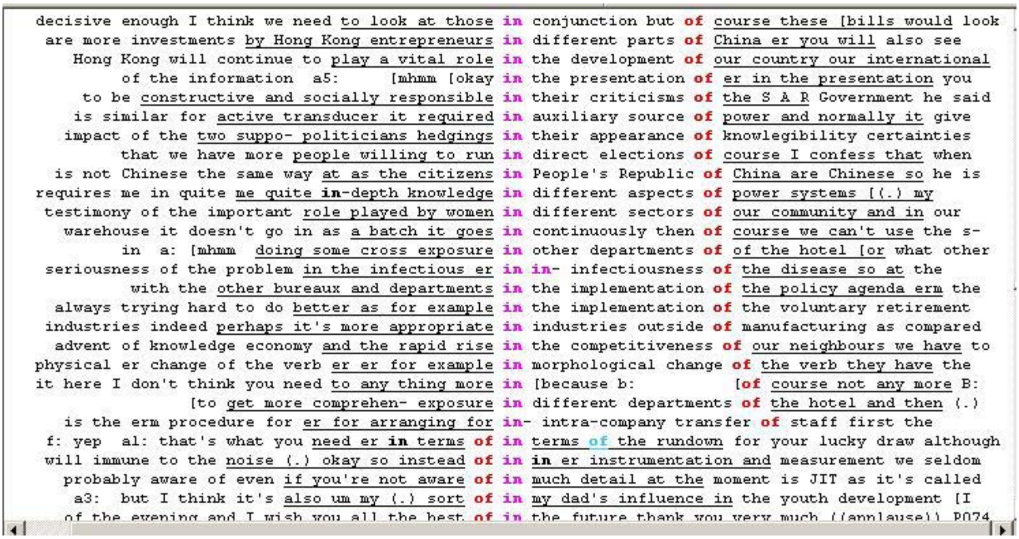


```
1         around the world it underlines our position as Asia's world city a cosmopolitan fr
2       our our goal of becoming the world city of of of Asia er (.) here is a huge Japanese
3           a vision ours is to become the World City of Asia we aim to be to this region wh
4         vision to build Hong Kong into a world city of Asia I hope you will find time in y
5       major cities of China but also the world city of Asia (pause) thank you very much in
6           wanting to make Hong Kong as a world city of Asia well let me tell you er some o
7             is to Hong Kong erm we are a world city of Asia (.) er and this is how we are
8               that underpin Hong Kong's positioning as Asia world city (.) the point I wan
9           development put forward last year we will be Asia's world city and the most cosm
10          of information has helped stamp Hong Kong as Asia's world city (pause) we try ve
11              pleasant and memorable stay in Hong Kong Asia's world city (.) tonight is a
12       new jobs will be created for local community as Asia's world city we are dedicated
13       earlier we have set ourselves a benchmark to be Asia's world city we are already th
14         quite a bit of space talking about what is er Asia's World City in my policy addr
15        Hong Kong has further cemented its position as Asia's world city including seconda
16      co-operation (.) we have positioned Hong Kong as Asia's World City the purpose of ou
```

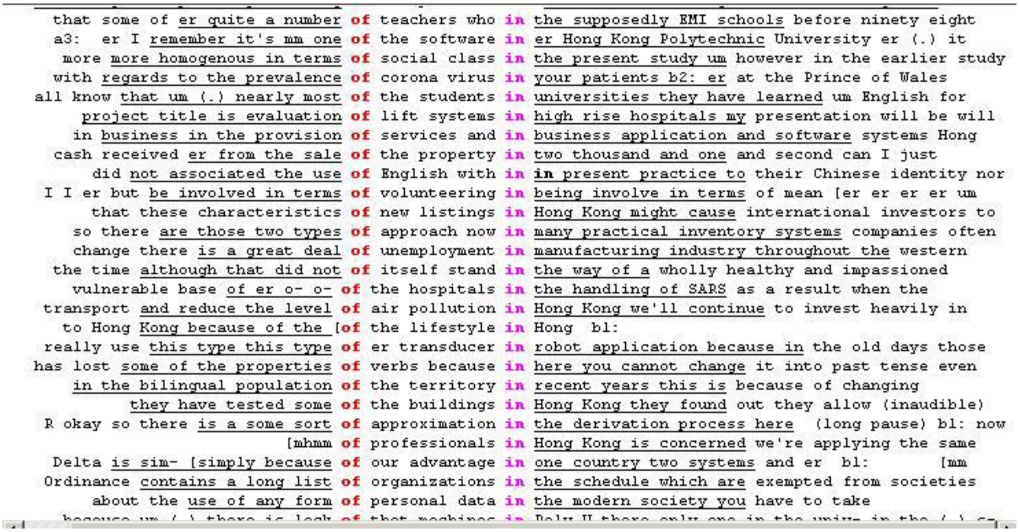
For the 'Sort Right' or 'Sort Left' options, the user can set the co-occurring word that is to be alphabetically sorted either 1, 2 or 3 words to the right or left of the centred word. This parameter is ignored for the 'Sort Position' option. You can choose whether or not you want to have the concordance lines numbered (default is 'Yes'), and whether to hide tagging.

The 'Span' options refer to the number of words between the centred word and the outer co-occurring word that is used for the search. You must first select 'Use Span', and then choose a number for the internal span, for example, if '2' is selected for the internal span value, only co-occurring words which are 1, 2 or 3 words from the centred word will be listed, as in the example below for 'in' with 'of' co-occurring (Figure 62):

Figure 62. Results of using the 'SPAN' option with internal span set to 2 and external span set to 4



```
 decisive enough I think we need to look at those in conjunction but of course these [bills would look
  are more investments by Hong Kong entrepreneurs in different parts of China er you will also see
     Hong Kong will continue to play a vital role in the development of our country our international
           of the information  a5:     [mhmm [okay in the presentation of er in the presentation you
      to be constructive and socially responsible in their criticisms of the S A R Government he said
     is similar for active transducer it required in auxiliary source of power and normally it give
    impact of the two suppo- politicians hedgings in their appearance of knowlegibility certainties
           that we have more people willing to run in direct elections of course I confess that when
   is not Chinese the same way at as the citizens in People's Republic of China are Chinese so he is
 requires me in quite me quite in-depth knowledge in different aspects of power systems [(.) my
  testimony of the important role played by women in different sectors of our community and in our
    warehouse it doesn't go in as a batch it goes in continuously then of course we can't use the s-
           in  a: [mhmm  doing some cross exposure in other departments of of the hotel [or what other
  seriousness of the problem in the infectious er in in- infectiousness of the disease so at the
            with the other bureaux and departments in the implementation of the policy agenda erm the
   always trying hard to do better as for example in the implementation of the voluntary retirement
  industries indeed perhaps it's more appropriate in industries outside of manufacturing as compared
   advent of knowledge economy and the rapid rise in the competitiveness of our neighbours we have to
 physical er change of the verb er er for example in morphological change of the verb they have the
 it here I don't think you need to any thing more in [because b:           [of course not any more B:
                 [to get more comprehen- exposure in different departments of the hotel and then (.)
    is the erm procedure for er for arranging for in- intra-company transfer of staff first the
  f: yep  a1: that's what you need er in terms of in terms of the rundown for your lucky draw although
  will immune to the noise (.) okay so instead of in in er instrumentation and measurement we seldom
    probably aware of even if you're not aware of in much detail at the moment is JIT as it's called
     a3:  but I think it's also um my (.) sort of in my dad's influence in the youth development [I
    of the evening and I wish you all the best of in the future thank you very much ((applause)) P074
```

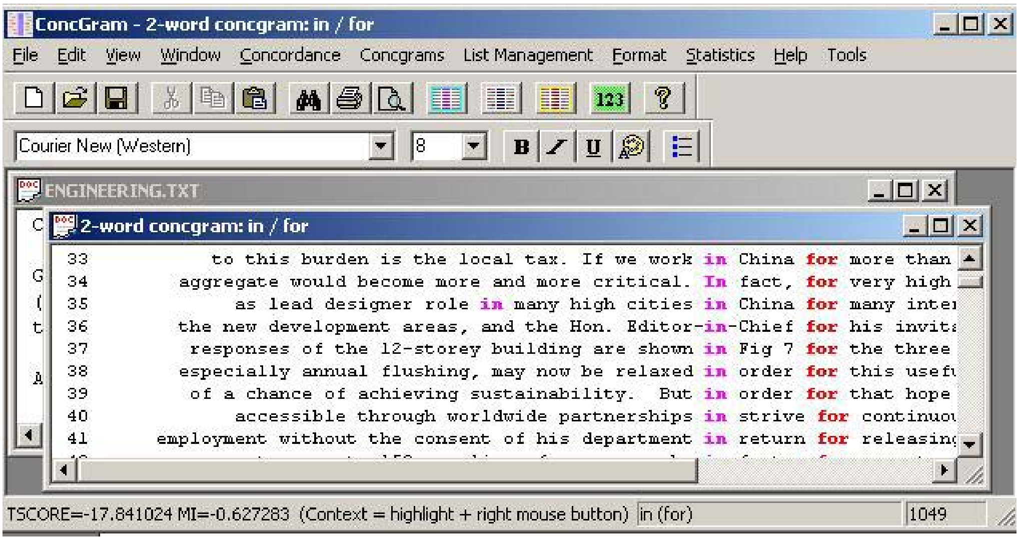


```
    that some of er quite a number of teachers who in the supposedly EMI schools before ninety eight
    a3:  er I remember it's mm one of the software in er Hong Kong Polytechnic University er (.) it
     more more homogenous in terms of social class in the present study um however in the earlier study
    with regards to the prevalence of corona virus in your patients b2: er at the Prince of Wales
  all know that um (.) nearly most of the students in universities they have learned um English for
       project title is evaluation of lift systems in high rise hospitals my presentation will be will
      in business in the provision of services and in business application and software systems Hong
    cash received er from the sale of the property in two thousand and one and second can I just
         did not associated the use of English with in in present practice to their Chinese identity nor
   I I er but be involved in terms of volunteering in being involve in terms of mean [er er er er um
        that these characteristics of new listings in Hong Kong might cause international investors to
       so there are those two types of approach now in many practical inventory systems companies often
      change there is a great deal of unemployment in manufacturing industry throughout the western
    the time although that did not of itself stand in the way of a wholly healthy and impassioned
      vulnerable base of er o- o- of the hospitals in the handling of SARS as a result when the
   transport and reduce the level of air pollution in Hong Kong we'll continue to invest heavily in
     to Hong Kong because of the [of the lifestyle in Hong  b1:
   really use this type this type of er transducer in robot application because in the old days those
  has lost some of the properties of verbs because in here you cannot change it into past tense even
      in the bilingual population of the territory in recent years this is because of changing
            they have tested some of the buildings in Hong Kong they found out they allow (inaudible)
   R okay so there is a some sort of approximation in the derivation process here  (long pause) b1: now
                             [mhmm of professionals in Hong Kong is concerned we're applying the same
    Delta is sim- [simply because of our advantage in one country two systems and er  b1:        [mm
   Ordinance contains a long list of organizations in the schedule which are exempted from societies
       about the use of any form of personal data in the modern society you have to take
      because um (.) there is lack of that machines in Poly U there only one in the univ- in the (.) c-
```

You have the option to search for the t-score or MI value of the concgram, but you need a 'Unique Word' list for the file that serves as your corpus (remember that user nominated searches can search multiple files). If this option is selected, you will first be prompted to specify a Unique Words list to use in calculating the t-score and MI value. These are displayed in the Status Bar at the bottom of the window.

Also, if this option is selected, the 'Use Span' option must also be selected and a value set for the internal span, as this is also required in the calculation of the statistical tests.

Figure 63. The t-score and MI value displayed in Status Bar below main window



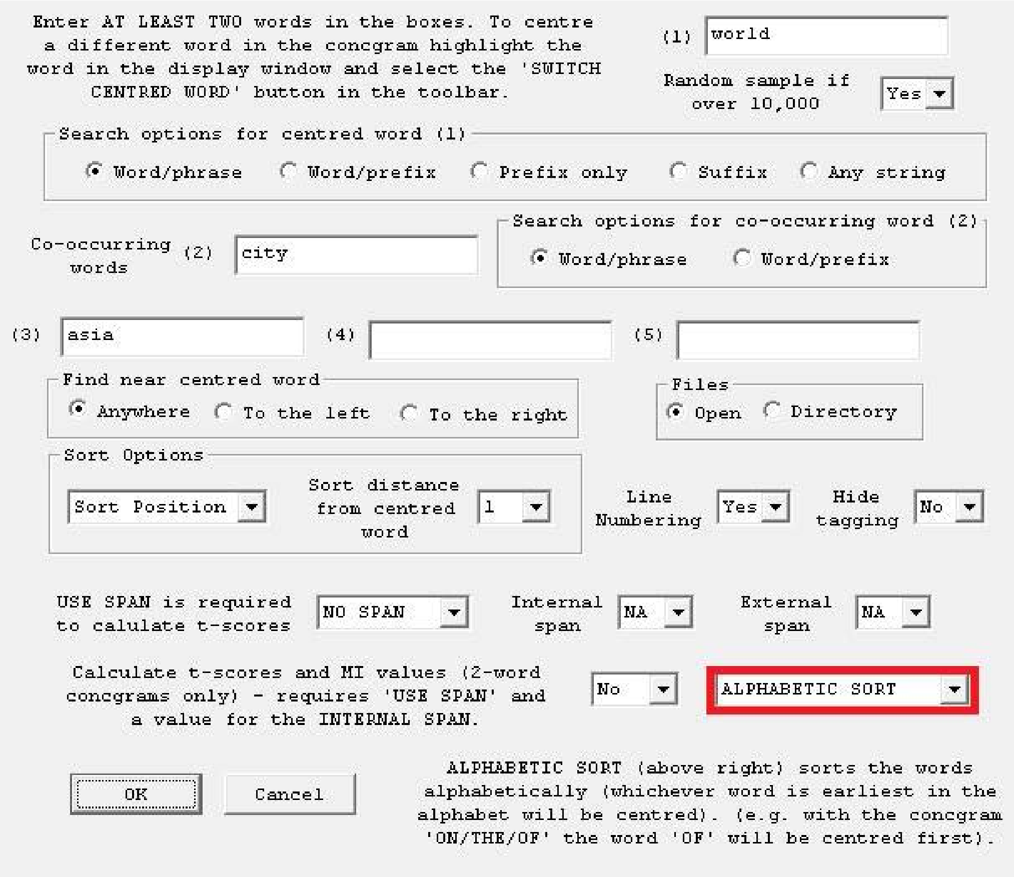
ConcGram - 2-word concgram: in / for
File Edit View Window Concordance Concgrams List Management Format Statistics Help Tools
Courier New (Western)
8
ENGINEERING.TXT
2-word concgram: in / for

```
33          to this burden is the local tax. If we work in China for more than
34        aggregate would become more and more critical. In fact, for very high
35            as lead designer role in many high cities in China for many inte
36        the new development areas, and the Hon. Editor-in-Chief for his invit
37         responses of the 12-storey building are shown in Fig 7 for the three
38        especially annual flushing, may now be relaxed in order for this usef
39         of a chance of achieving sustainability.  But in order for that hope
40            accessible through worldwide partnerships in strive for continuo
41      employment without the consent of his department in return for releasin
```

TSCORE=-17.841024 MI=-0.627283 (Context = highlight + right mouse button) in (for) 1049

The final option refers to the words which make up the concgram, and the words can be sorted alphabetically or not according to the user's preference. Figure 64 shows the effect of typing 'world/city/asia' into the 'Concgram Search' Dialog and selecting the 'Alphabetic Sort' option. The words in the concgram are simply sorted alphabetically, so that 'asia' is the first word and thus the centred word, and 'world' becomes word 3 in the concgram.

Figure 64. The alphabetically sorted concgram 'world/city/asia'



Enter AT LEAST TWO words in the boxes. To centre a different word in the concgram highlight the word in the display window and select the 'SWITCH CENTRED WORD' button in the toolbar.

(1) world

Random sample if over 10,000: Yes

Search options for centred word (1): ● Word/phrase ○ Word/prefix ○ Prefix only ○ Suffix ○ Any string

Co-occurring words (2) city

Search options for co-occurring word (2): ● Word/phrase ○ Word/prefix

(3) asia (4) (5)

Find near centred word: ● Anywhere ○ To the left ○ To the right

Files: ● Open ○ Directory

Sort Options: Sort Position; Sort distance from centred word: 1

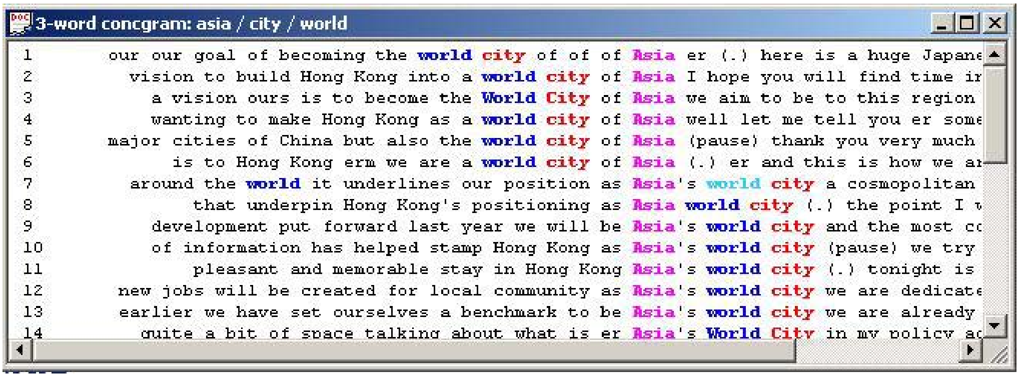
Line Numbering: Yes; Hide tagging: No

USE SPAN is required to calulate t-scores: NO SPAN; Internal span: NA; External span: NA

Calculate t-scores and MI values (2-word concgrams only) - requires 'USE SPAN' and a value for the INTERNAL SPAN. No

ALPHABETIC SORT

ALPHABETIC SORT (above right) sorts the words alphabetically (whichever word is earliest in the alphabet will be centred). (e.g. with the concgram 'ON/THE/OF' the word 'OF' will be centred first).



3-word concgram: asia / city / world

```
1        our our goal of becoming the world city of of of Asia er (.) here is a huge Japane
2          vision to build Hong Kong into a world city of Asia I hope you will find time ir
3            a vision ours is to become the World City of Asia we aim to be to this region
4            wanting to make Hong Kong as a world city of Asia well let me tell you er some
5        major cities of China but also the world city of Asia (pause) thank you very much
6               is to Hong Kong erm we are a world city of Asia (.) er and this is how we ar
7          around the world it underlines our position as Asia's world city a cosmopolitan
8                that underpin Hong Kong's positioning as Asia world city (.) the point I v
9            development put forward last year we will be Asia's world city and the most cc
10           of information has helped stamp Hong Kong as Asia's world city (pause) we try
11               pleasant and memorable stay in Hong Kong Asia's world city (.) tonight is
12        new jobs will be created for local community as Asia's world city we are dedicate
13        earlier we have set ourselves a benchmark to be Asia's world city we are already
14           quite a bit of space talking about what is er Asia's World City in my policy ac
```

You can use the lemmatization or the wildcard '%' sign in the Concgram Search Dialog, but only in the edit box for word (1), and with no alphabetic sort or word prefix for word (1).

Appendix

ConcGram Tutorial

Files for following the tutorial

There are several files to demonstrate the basic ConcGram functions. These are:

1. Two speeches by Donald Tsang, the Chief Executive of Hong Kong, one delivered in 2005 and one delivered in 2006:
   2005_06_policy_address_Donald_Tsang.txt
   2006_07_policy_address_Donald_Tsang.txt

2. The 2-word concgram list for the 2006–7 speech created without using any span (default) but using an Exclusion List based on Ahmad's most frequent words in the BNC:
   cg2-DTSpeech-NOSPAN-exclList.TXT

3. The Exclusion List used:
   exclusionList-K.TXT

4. And the Unique Words Lists for the two speeches:
   uwords-DTSpeech-2005-6.TXT
   uwords-DTSpeech-2006–7.TXT

All these files are included with the ConcGram Setup on the CD, and may be found in the folder marked 'Tutorial' in the folder where you installed ConcGram.

Appendix

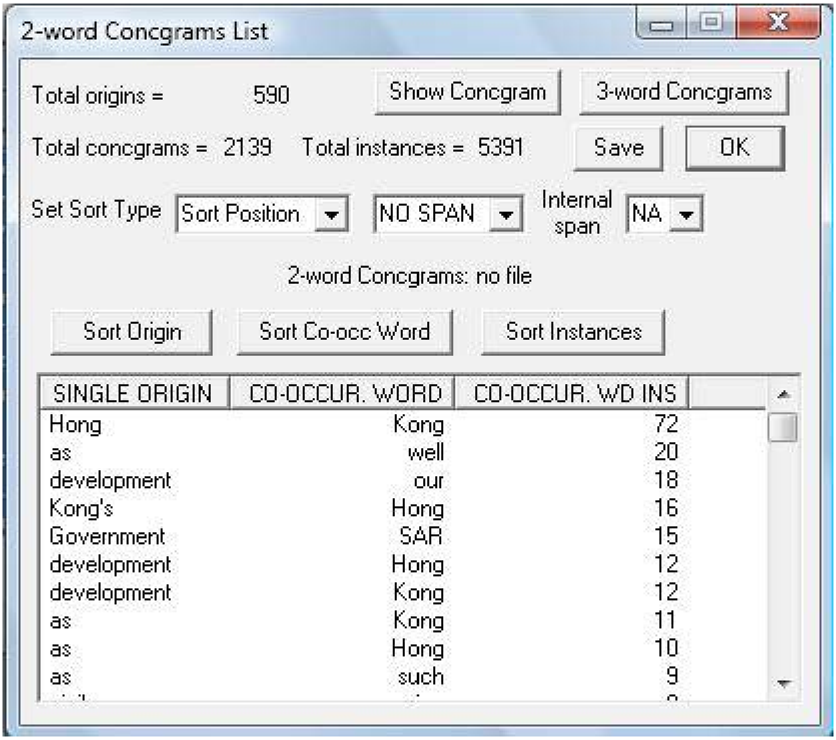
ConcGram Tutorial (1)

Creating the initial 2-word concgram list

1. Open the speech for 2006–7 (**FILE ➡ OPEN**) (“**2006_07_policy_address_Donald_Tsang.txt”).**
2. Create a new 2-word concgram list (**CONCGRAMS ➡ CREATE NEW CONCGRAM LIST (AUTOMATIC) ➡ USING ALL THE WORDS IN TEXT ➡ WITH NO INITIAL UW LIST.**
3. When the Concgram List Preferences Dialog Box appears leave the default settings EXCEPT you should check the “**Use Exclusion List**” box (by default this is unchecked).
4. Click “**OK**” and when prompted to choose the exclusion list select the above listed file (“**exclusionList-K.TXT**”)
5. As the file for the speech is small, and there are only 1968 searches to make (based on the count of unique words in the text), each search takes a short time
6. The resulting 2135 concgrams are displayed in the Concgrams List Box shown in Figure 1 below.
7. Click the “**Sort Instances**” button in the list box (by default entries are alphabetically sorted) and this will show the ‘aboutness’ of the text.

The 2-word concgram list appears as follows (only the top of the list is shown here)

Figure 1. The 2-word Concgrams List Box after the "Sort Instances" button has been clicked



2-word Concgrams List

Total origins = 590

Total concgrams = 2139 Total instances = 5391

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

2-word Concgrams: no file

SINGLE ORIGIN	CO-OCCUR. WORD	CO-OCCUR. WD INS
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

8. Lastly click on "**Save**" and click "**OK**" for "**no cut off**". When prompted, name the list "**cg2-DTSpeech-2006-7-NOSPAN-exclList.TXT**".

The saved concgram list appears in a new window:

```
 2-word concgrams
 2139 NOSPAN 590
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
civil          service       9
development    support       9
families       support       9
air            quality       8
as             centre        8
as              our             8
as              support         8
Chief           Executive       8
Council         Legislative     8
development     Government      8
Kong            our             8
need            our             8
set             up              8
as              civil           7
as              family          7
as              international   7
Centre          Kong            7
community       our             7
community       support         7
development     economic        7
development     future          7
environment     our             7
family          members         7
financial       Kong            7
Governance      Strong          7
Government      role            7
Government      support         7
Hong            our             7
important       our             7
international   Kong            7
Kong's          as              7
Kong's          development     7
last            year            7
our             support         7
```

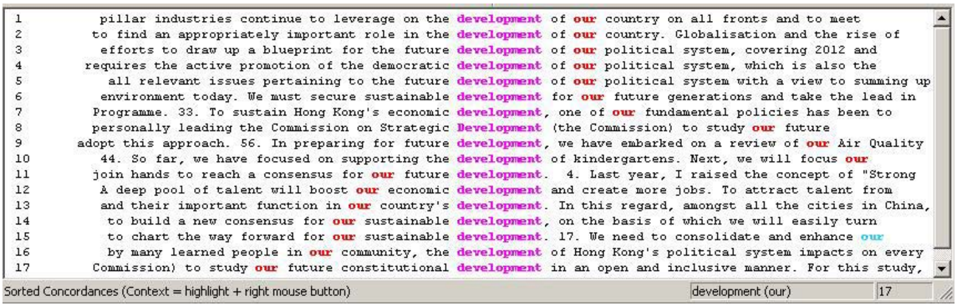
The starting number (2135) refers to the total number of concgrams in the list, and the end number (589) refers to the number of unique origins.

The 12 most frequent phrases which show the 'aboutness' of the speech are:

```
Hong            Kong          68
development     our           18
Government      SAR           15
development     support        9
support         families       9
air             quality        8
Chief           Executive      8
Council         Legislative    8
Executive       Chief          8
Government      Development    8
community       support        7
development     economic       7
```

To see a concgram concordance display for any of these, in the List Box select with the cursor one of these phrases and then click on the 'Show Concgram' button in the List Box. For example, if the 2-word concgram 'development/our' is selected and the 'Show Concgram' button is clicked, the following concgram will be displayed in a new window behind the List Box:

Figure 2. 2-word concgram for 'development / our'



```
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 up
6          environment today. We must secure sustainable development for our future generations and take the lead in
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 Quality
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 "Strong
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 China,
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 every
17        Commission) to study our future constitutional development in an open and inclusive manner. For this study,
```

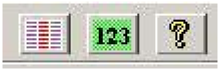
Sorted Concordances (Context = highlight + right mouse button)
development (our)
17

Appendix

ConcGram Tutorial (2)

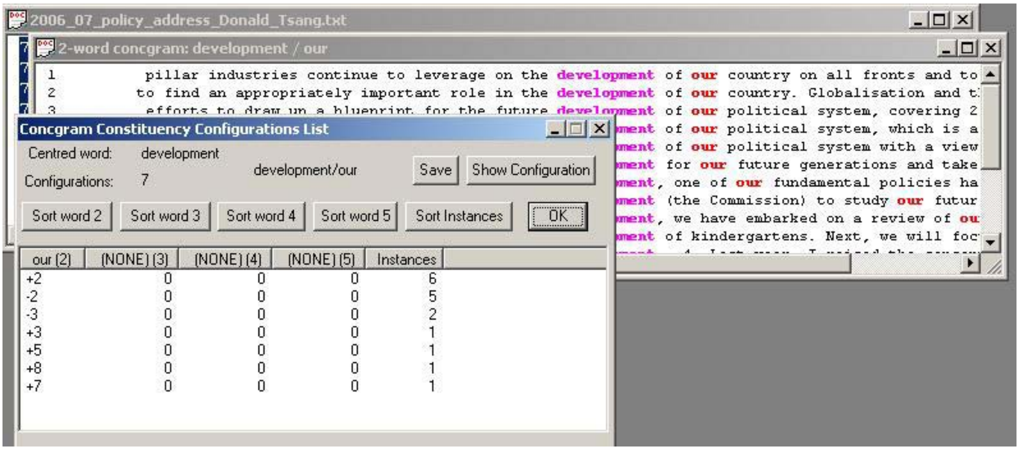
Concgram configurations

You can open the 'Concgram Constituency Configurations' List Dialog by clicking the '123' button in the toolbar (or 'Statistics ➡ Concgram Configurations ➡ Constituent Variants' from the menu).



For the 'development/our' concgram we can see that the highest incidence occurs at + 2, with 6 occurrences, followed by -2 with 5 occurrences. There are no contiguous examples.

Figure 3. The Constituency Configurations List


2006_07_policy_address_Donald_Tsang.txt
2-word concgram: development / our
Concgram Constituency Configurations List
Save
Show Configuration
Sort word 2
Sort word 3
Sort word 4
Sort word 5
Sort Instances
OK

```
1          pillar industries continue to leverage on the development of our country on all fronts and to
2         to find an appropriately important role in the development of our country. Globalisation and t
3          efforts to draw up a blueprint for the future development of our political system, covering 2
                                                            ment of our political system, which is a
                                                            ment of our political system with a view
                                                            ment for our future generations and take
                                                            ment, one of our fundamental policies ha
                                                            ment (the Commission) to study our futur
                                                            ment, we have embarked on a review of ou
                                                            ment of kindergartens. Next, we will foc
```

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

Before going on to Part 3, close all windows.

Appendix

ConcGram Tutorial (3)

Statistical tests: t-score and MI value

Another way to make the list shorter is by using the t-scores / MI Value function from the Concgram Menu (Concgram ➡ t-scores & MI Value For 2-Word Concgrams ➡ Create New List ➡ With One Corpus File). This function requires both the CG2 list (list of 2-word concgrams) and the UW list (Unique Word List) for the corpus file as well as the file itself ('2006_07_policy_address_Donald_Tsang.txt'). You also need to set the span as this value is required for the calculations.

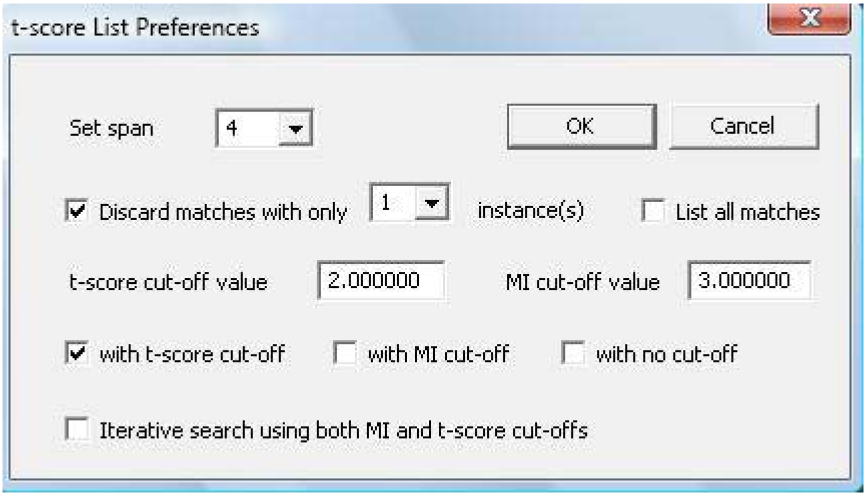
You have already created the CG2 list in Tutorial 1, but you now need to create the UW List for the same text. To do this, follow these steps:

1. Open the speech for 2006-7 (File ➡ Open), file '2006_07_policy_address_Donald_Tsang.txt'.
2. Create the Unique Word List from the Statistics Menu (Statistics ➡ Unique Words).
3. Select the 'Frequency Sort' button, then the 'Save' button.
4. Name the file 'uwords-DTSpeech-2006–7.TXT' and press 'OK'.
5. The saved file will be loaded in a new window.

Close all windows before proceeding. Then follow these steps to create the t-score and MI value list.

6. Open the speech for 2006-7 (File ➡ Open), file '2006_07_policy_address_Donald_Tsang.txt'.
7. Select the from the Concgram Menu (Concgram ➡ t-scores & MI Value For 2-Word Concgrams ➡ Create New List ➡ With One Corpus File).
8. At the prompt, open the CG2 list file you created before (point 8 in TUTORIAL 1) 'cg2-DTSpeech-2006-7-NOSPAN-exclList.TXT'.
9. At the next prompt for the UW List for the file, load the UW list you created in Steps 2–4 'uwords-DTSpeech-2006–7.TXT'.
10. The t-scores & MI value Preferences Dialog will then be displayed as in Figure 4.

Figure 4. The *t*-score/MI Values Preferences Dialog



t-score List Preferences

Set span 4

☑ Discard matches with only 1 instance(s) □ List all matches

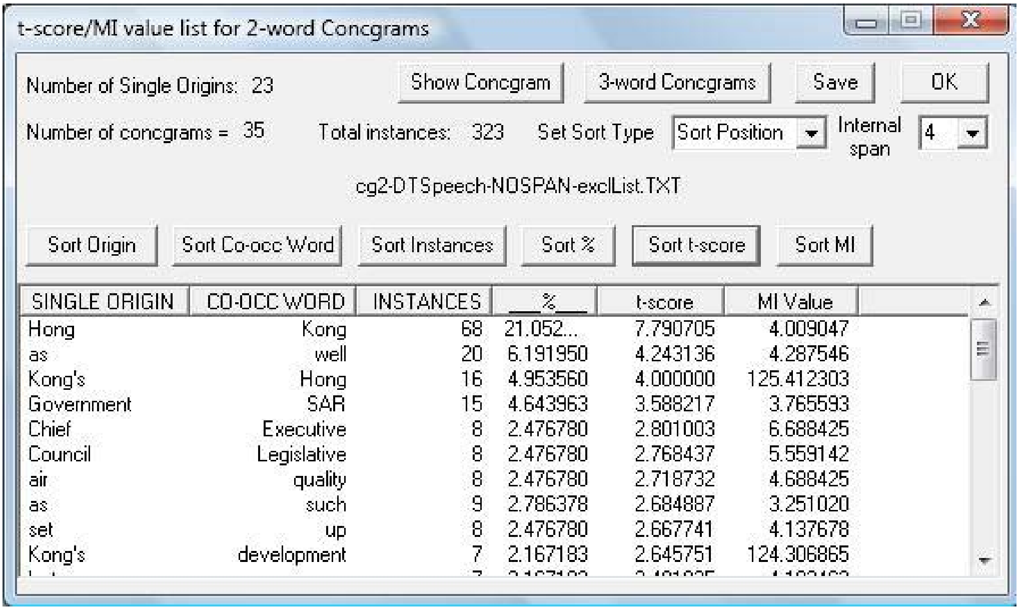
t-score cut-off value 2.000000 MI cut-off value 3.000000

☑ with t-score cut-off □ with MI cut-off □ with no cut-off

□ Iterative search using both MI and t-score cut-offs

10. Leave all the default values, except check the 'with t-score cut-off' box. Select your preferred *t*-score cut off value (here the default 2.000000 is used), and click 'OK'.
11. At the prompt to load the Unique Word List select the file you have created and saved as 'uwords-DTSpeech-2006–7.TXT'
12. The t-score and MI value list will be created as shown in Figure 5.
13. Save the file as 'cg2-DTSpeech-exclList-TSCORES.txt'. You will need to load this file in TUTORIAL 6.

Figure 5. The t-score/MI Value List for the 2006-7 speech after activation of 'Sort by t-score'



Number of Single Origins: 23

Number of concgrams = 35 Total instances: 323 Set Sort Type Sort Position Internal span 4

cg2-DTSpeech-NOSPAN-exclList.TXT

SINGLE ORIGIN	CO-OCC WORD	INSTANCES	%	t-score	MI Value
Hong	Kong	68	21.052...	7.790705	4.009047
as	well	20	6.191950	4.243136	4.287546
Kong's	Hong	16	4.953560	4.000000	125.412303
Government	SAR	15	4.643963	3.588217	3.765593
Chief	Executive	8	2.476780	2.801003	6.688425
Council	Legislative	8	2.476780	2.768437	5.559142
air	quality	8	2.476780	2.718732	4.688425
as	such	9	2.786378	2.684887	3.251020
set	up	8	2.476780	2.667741	4.137678
Kong's	development	7	2.167183	2.645751	124.306865

Using the t-score cutoff makes the list much shorter. It is also possible to compare the t-score created for a 2-word concgrams list created from one text or corpus with those from another text or corpus, and TUTORIAL 6 shows you how to do this.

Appendix

ConcGram Tutorial (4)

Comparing files for word specificity

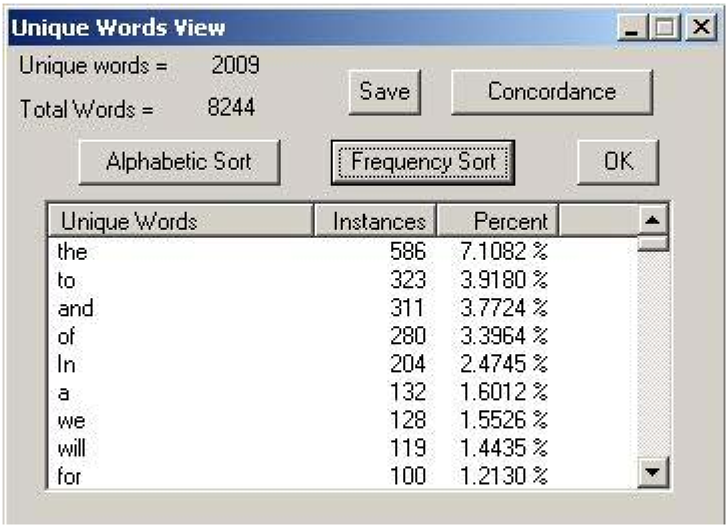
Another comparison uses Ahmad's (2005) 'weirdness' formula for the relative frequency of a single word in a text or corpus compared with a second or third text or corpus ('word specificity' is also known as 'keyness' (Scott and Tribble 2006). A score of 1 or less means the word is not specific to C1, whereas a score greater than 1 will indicate specificity.

We have listed this under the term 'Word Specificity' in preference to Ahmad's 'weirdness'. The 'Word Specificity' function is available by selecting from the Statistics Menu (Statistics ➡ Compare Word Specificity).

In Tutorial 4, we will compare 2006-7 and 2005-6 word specificity lists in order to find which items are more specific to the 2006-7 speech. In order to compare the two speeches, we must first create a Unique Word List for each of the speeches. To do this, follow these steps:

1. Open the speech for 2006–7 (FILE ➡ OPEN), choosing file '2006_07_policy_address_Donald_Tsang.txt'.
2. Create the Unique Word List from the Statistics Menu (Statistics ➡ Unique Words).
3. Select the 'Frequency Sort' button, then the 'Save' button.
4. Name the file 'uwords-DTSpeech-2006-7.TXT' and press 'OK'.
5. The saved file will be loaded in a new window.

Figure 6. Unique Word List for the 2006-7 speech after activation of 'Frequency Sort'



Unique words = 2009

Total Words = 8244

Unique Words	Instances	Percent
the	586	7.1082 %
to	323	3.9180 %
and	311	3.7724 %
of	280	3.3964 %
In	204	2.4745 %
a	132	1.6012 %
we	128	1.5526 %
will	119	1.4435 %
for	100	1.2130 %

Now close both the 2006–7 speech ('2006_07_policy_address_Donald_Tsang.txt') and the UW List for this text ('uwords-DTSpeech-2006-7.TXT') and repeat steps 1 to 5 for the 2005-6 speech ('2005_06_policy_address_Donald_Tsang.txt'). Name the new UW List file 'uwords-DTSpeech-2005-6.TXT'.

Once you have the UW List for each speech, you can compare them for specificity. To do this, follow the steps as follows:

1. Select from the Statistics Menu and choose the function to compare 2 texts (Statistics ➡ Compare Word Specificity ➡ Compare 2 Corpus Files).
2. At the prompt for Corpus 1, select the Unique Word List you created for the 2006-7 speech ('uwords-DTSpeech-2006-7.TXT').
3. At the next prompt for Corpus 2, select Unique Word List you created for the 2005-6 speech ('uwords-DTSpeech-2005-6.TXT').
4. The result initially lists the Word Specificities as follows:

```
Unique Words      Inst C1     Percent C1  Inst C2     Percent C2  Specificity C1/C2

1      and        311         3.7692 %    577         4.5039 %          0.8369

2      a          132         1.5998 %    205         1.6002 %          0.9998

3      as         65          0.7878 %    92          0.7181 %          1.0970

4      an         35          0.4242 %    42          0.3278 %          1.2939

5      are        34          0.4121 %    53          0.4137 %          0.9960

6      also       30          0.3636 %    40          0.3122 %          1.1645
```

```
7     At          28          0.3394 %    33          0.2576 %          1.3174
8     all         24          0.2909 %    38          0.2966 %          0.9806
9     air         16          0.1939 %    13          0.1015 %          1.9110
10    achieve     8           0.0970 %    12          0.0937 %          1.0351
11    among       8           0.0970 %    6           0.0468 %          2.0702
12    Address     7           0.0848 %    10          0.0781 %          1.0869
13    about       7           0.0848 %    11          0.0859 %          0.9881
14    arts        7           0.0848 %    7           0.0546 %          1.5527
15    areas       6           0.0727 %    11          0.0859 %          0.8469
16    adopt       6           0.0727 %    1           0.0078 %          9.3160
17    able        6           0.0727 %    4           0.0312 %          2.3290
18    am          5           0.0606 %    5           0.0390 %          1.5527
19    after       5           0.0606 %    3           0.0234 %          2.5878
20    actively    5           0.0606 %    13          0.1015 %          0.5972
21    application 5           0.0606 %    1           0.0078 %          7.7633
22    appropriate 5           0.0606 %    5           0.0390 %          1.5527
23    approach    5           0.0606 %    2           0.0156 %          3.8817
24    Authorities 4           0.0485 %    24          0.1873 %          0.2588
25    Asian       4           0.0485 %    0           0.0000 %          1.#INF
```

5. Now click the ‘Specificity (C1/C2) Sort’ in the List Box.
6. The resulting list looks like this:

```
Unique Words      Inst C1     Percent C1  Inst C2     Percent C2  Specificity C1/C2
1     Asian             4     0.0485 %    0           0.0000 %          1.#INF
2     aged              3     0.0364 %    0           0.0000 %          1.#INF
3     Answer            2     0.0242 %    0           0.0000 %          1.#INF
4     amongst           2     0.0242 %    0           0.0000 %          1.#INF
5     attain            2     0.0242 %    0           0.0000 %          1.#INF
6     allowing          2     0.0242 %    0           0.0000 %          1.#INF
7     achievers         2     0.0242 %    0           0.0000 %          1.#INF
8     addressing        2     0.0242 %    0           0.0000 %          1.#INF
9     anti-pollution    1     0.0121 %    0           0.0000 %          1.#INF
10    Anhui             1     0.0121 %    0           0.0000 %          1.#INF
11    accelerated       1     0.0121 %    0           0.0000 %          1.#INF
12    August            1     0.0121 %    0           0.0000 %          1.#INF
13    appropriately     1     0.0121 %    0           0.0000 %          1.#INF
14    adapting          1     0.0121 %    0           0.0000 %          1.#INF
```

```
15    aligned          1     0.0121 %    0           0.0000 %         1.#INF
16    affluence        1     0.0121 %    0           0.0000 %         1.#INF
17    accepts          1     0.0121 %    0           0.0000 %         1.#INF
18    applied          1     0.0121 %    0           0.0000 %         1.#INF
19    automotive       1     0.0121 %    0           0.0000 %         1.#INF
20    accessory        1     0.0121 %    0           0.0000 %         1.#INF
21    atmosphere       1     0.0121 %    0           0.0000 %         1.#INF
22    audience         1     0.0121 %    0           0.0000 %         1.#INF
23    Associations     1     0.0121 %    0           0.0000 %         1.#INF
24    applications     1     0.0121 %    0           0.0000 %         1.#INF
25    across           1     0.0121 %    0           0.0000 %         1.#INF
```

7. The '1.#INF' means that the word does not appear at all in the second corpus (C2), and therefore has 'infinite' specificity. These words are listed first, ordered by instances in C1. Such instances need to be handled with caution by researchers and learners because many of them only occur once, and so they may not be specific to the text or corpus.
8. Scrolling down the list reaches the words which are found in both speeches, the highest specificity being for the word 'technology' which occurs 12 times in C1 but only once in C2:

```
Unique Words      Inst C1     Percent C1  Inst C2     Percent C2  Specificity C1/C2
820   technology       12    0.1454 %    1           0.0078 %         18.6319
821   vehicle          7     0.0848 %    1           0.0078 %         10.8686
822   adopt            6     0.0727 %    1           0.0078 %         9.3160
823   design           6     0.0727 %    1           0.0078 %         9.3160
824   expenditure      6     0.0727 %    1           0.0078 %         9.3160
825   research         6     0.0727 %    1           0.0078 %         9.3160
826   application      5     0.0606 %    1           0.0078 %         7.7633
827   emission         5     0.0606 %    1           0.0078 %         7.7633
828   Euro             5     0.0606 %    1           0.0078 %         7.7633
829   energy           5     0.0606 %    1           0.0078 %         7.7633
830   faced            5     0.0606 %    1           0.0078 %         7.7633
831   success          5     0.0606 %    1           0.0078 %         7.7633
832   supporting       5     0.0606 %    1           0.0078 %         7.7633
833   situation        5     0.0606 %    1           0.0078 %         7.7633
834   high             9     0.1091 %    2           0.0156 %         6.9870
835   per              9     0.1091 %    2           0.0156 %         6.9870
```

```
836   Academy          4    0.0485 %    1           0.0078 %         6.2106
837   Complex          4    0.0485 %    1           0.0078 %         6.2106
838   days             4    0.0485 %    1           0.0078 %         6.2106
839   election         4    0.0485 %    1           0.0078 %         6.2106
840   having           4    0.0485 %    1           0.0078 %         6.2106
841   his              4    0.0485 %    1           0.0078 %         6.2106
842   leading          4    0.0485 %    1           0.0078 %         6.2106
843   performing       4    0.0485 %    1           0.0078 %         6.2106
844   River            4    0.0485 %    1           0.0078 %         6.2106
845   reduced          4    0.0485 %    1           0.0078 %         6.2106
```

9. The word 'technology' has the highest specificity (18.6319), followed by 'vehicle' (10.8686) which occurs 7 times in C1 but only once in C2.
10. Finally if we scroll down to words which do not have specificity, we find, for example, words like 'the' (586 in C1 and 965 in C2) or 'SAR' (14 and 23) with a specificity of less than 1.

```
Unique Words      Inst C1     Percent C1  Inst C2     Percent C2  Specificity C1/C2
1563  from             22     0.2666 %    36          0.2810 %         0.9488
1564  two              11     0.1333 %    18          0.1405 %         0.9488
1565  SAR              14     0.1697 %    23          0.1795 %         0.9451
1566  the             586     7.1022 %    965         7.5326 %         0.9429
1567  Commission        9     0.1091 %    15          0.1171 %         0.9316
1568  allocate          3     0.0364 %    5           0.0390 %         0.9316
1569  city              6     0.0727 %    10          0.0781 %         0.9316
1570  Committee         6     0.0727 %    10          0.0781 %         0.9316
1571  capital           3     0.0364 %    5           0.0390 %         0.9316
```

Although Word Specificity is intended mainly to compare words from a more specialized text with a text or corpus which is more general in nature, Tutorial 4 shows that the function can also be used to compare texts of the same genre for their similarities and differences.

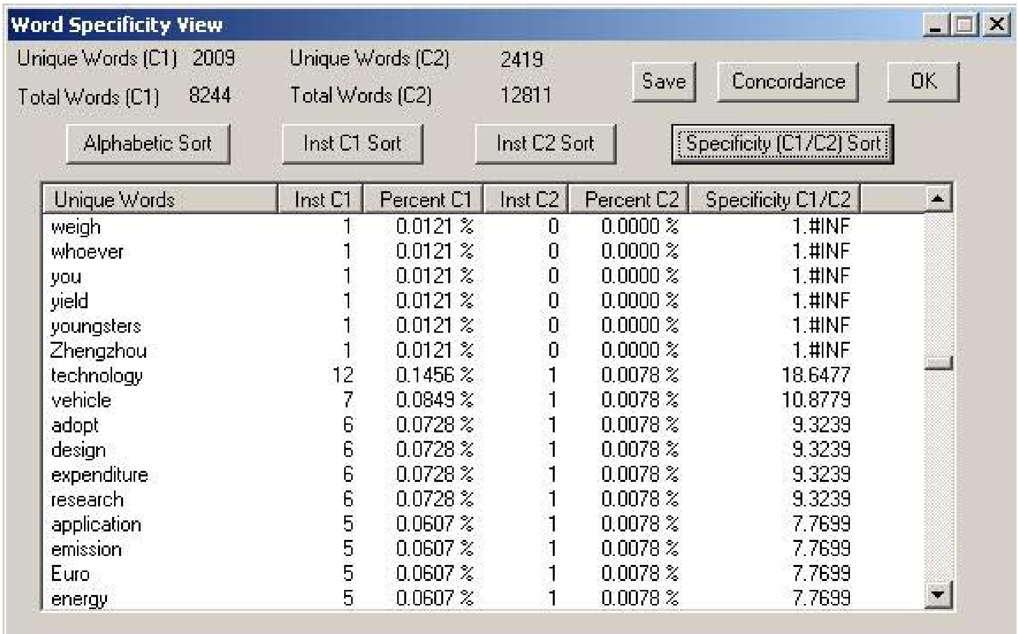
Appendix

ConcGram Tutorial (5)

The Configurations List for 2-word concgrams

Tutorial 5 illustrates the use of the Configurations List for 2-word Concgrams which is found under the Concgrams Menu. This displays the concgrams together with their positional and constituent variants, up to an internal span of two intervening words (i.e. AB, A * B and A * * B), so that a profile of each concgram in the list can be seen immediately. This function requires that a 2-word concgram list for the text has first been created and saved. This list is then loaded by the program and every concgram in the list is then used for a full search in the text, making the constituent and positional calculations at the same time. The result is then displayed in the List Box (shown in Figure 7).

Figure 7. The Word Specificity List Box



Unique Words (C1) 2009
Total Words (C1) 8244
Unique Words (C2) 2419
Total Words (C2) 12811

Unique Words	Inst C1	Percent C1	Inst C2	Percent C2	Specificity C1/C2
weigh	1	0.0121 %	0	0.0000 %	1.#INF
whoever	1	0.0121 %	0	0.0000 %	1.#INF
you	1	0.0121 %	0	0.0000 %	1.#INF
yield	1	0.0121 %	0	0.0000 %	1.#INF
youngsters	1	0.0121 %	0	0.0000 %	1.#INF
Zhengzhou	1	0.0121 %	0	0.0000 %	1.#INF
technology	12	0.1456 %	1	0.0078 %	18.6477
vehicle	7	0.0849 %	1	0.0078 %	10.8779
adopt	6	0.0728 %	1	0.0078 %	9.3239
design	6	0.0728 %	1	0.0078 %	9.3239
expenditure	6	0.0728 %	1	0.0078 %	9.3239
research	6	0.0728 %	1	0.0078 %	9.3239
application	5	0.0607 %	1	0.0078 %	7.7699
emission	5	0.0607 %	1	0.0078 %	7.7699
Euro	5	0.0607 %	1	0.0078 %	7.7699
energy	5	0.0607 %	1	0.0078 %	7.7699

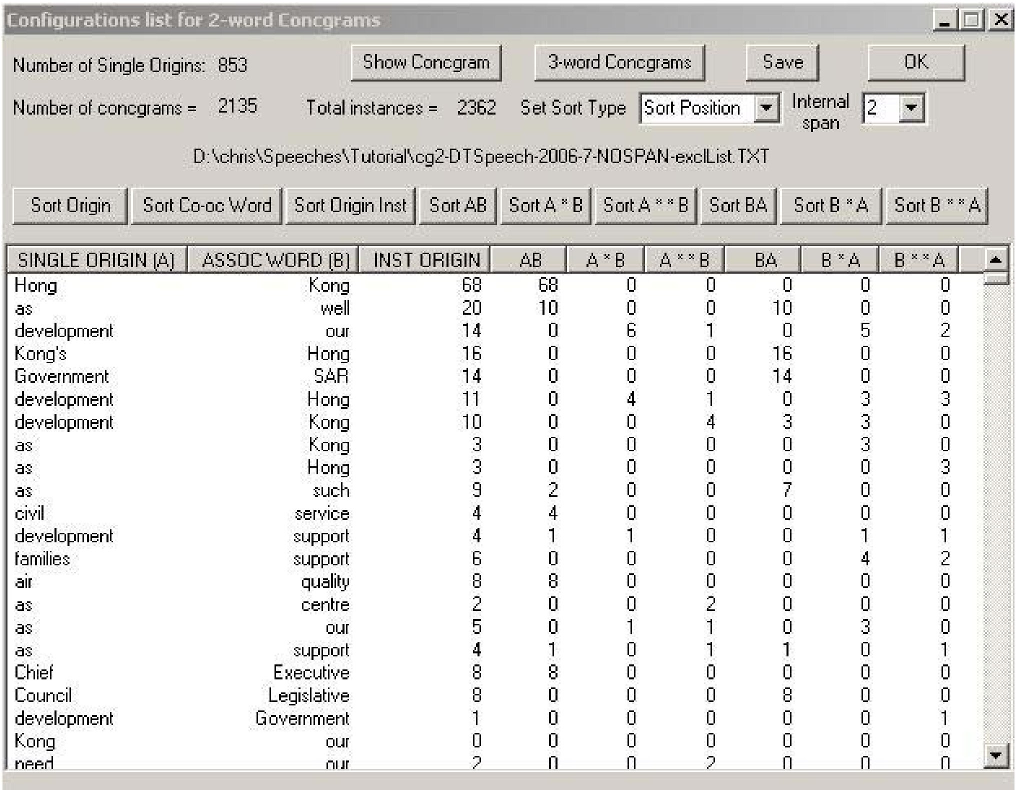
To create and save the initial list of 2-word concgrams, follow the same steps 1–6, as for Tutorial 1:

1. Open the speech for 2006–7 (File ➡ Open), choosing the '2006_07_policy_address_Donald_Tsang.txt' file.
2. Create a new 2-word concgram list (Concgrams ➡ Create New Concgram List (Automatic) ➡ Using All the Words in a Text.)
3. When the Concgram List Preferences Dialog Box appears, you should select the 'Use Exclusion List' Box (by default this is not selected). No span is used, use all words in the text and discard single instances.
4. Click 'OK'. When prompted, to choose the Exclusion List, select the file 'exclusion-List-K.TXT'.
5. Load the unique word list and then search. There are 894 searches if you search only for unique words which have 2 or more instances and 2009 searches if you search for all unique words.
6. Click 'Sort Instances' in the Concgram List Box that appears.
7. Lastly click on 'Save' and click 'OK' for 'no cut off'. When prompted, name the list 'cg2-DTSpeech-NOSPAN-exclList.TXT'.

Close all windows before starting the next stage.

8. First open the speech for 2006-7 (File ➡ Open) by choosing the file '2006_07_policy_address_Donald_Tsang.txt'.
9. Select from the Concgrams Menu options Concgrams ➡ Configurations List for 2-Word Concgrams ➡ Create New List.
10. When you are asked to choose the 2-word concgram list to be processed, choose the file you created in steps 1–7 ('cg2-DTSpeech-NOSPAN-exclList.TXT').
11. After the list has been selected, a 'Configurations List Preferences' Dialog Box gives you the opportunity to set the internal span. Set it to 2 (this is the default setting and so just click 'OK').
12. This time the program will make the configuration calculations for each item in the lists, and the results then displayed in the List Box (Figure 8).
13. After all the searches have been completed, click on the 'Sort Origin Inst' button (Figure 8).

Figure 8. The 2-Word Concgram Configurations List Box



Configurations list for 2-word Concgrams

Number of Single Origins: 853

Number of concgrams = 2135   Total instances = 2362   Set Sort Type: Sort Position   Internal span: 2

D:\chris\Speeches\Tutorial\cg2-DTSpeech-2006-7-NOSPAN-exclList.TXT

SINGLE ORIGIN (A)	ASSOC WORD (B)	INST ORIGIN	AB	A * B	A * * B	BA	B * A	B * * A
Hong	Kong	68	68	0	0	0	0	0
as	well	20	10	0	0	10	0	0
development	our	14	0	6	1	0	5	2
Kong's	Hong	16	0	0	0	16	0	0
Government	SAR	14	0	0	0	14	0	0
development	Hong	11	0	4	1	0	3	3
development	Kong	10	0	0	4	3	3	0
as	Kong	3	0	0	0	0	3	0
as	Hong	3	0	0	0	0	0	3
as	such	9	2	0	0	7	0	0
civil	service	4	4	0	0	0	0	0
development	support	4	1	1	0	0	1	1
families	support	6	0	0	0	0	4	2
air	quality	8	8	0	0	0	0	0
as	centre	2	0	0	2	0	0	0
as	our	5	0	1	1	0	3	0
as	support	4	1	0	1	1	0	1
Chief	Executive	8	8	0	0	0	0	0
Council	Legislative	8	0	0	0	8	0	0
development	Government	1	0	0	0	0	0	1
Kong	our	0	0	0	0	0	0	0
need	our	2	0	0	2	0	0	0

The List Box shows the positional and constituent variants for each concgram with an internal span of up to 2 intervening words. You can then select a specific single origin, and click to display all of the concordance lines for that particular concgram.

Appendix

ConcGram Tutorial (6)

Comparing the statistical values for one 2-word concgram list against two corpora

In TUTORIAL 3 you learned how to calculate t-score and MI value using a list of 2-word concgrams created from a single corpus file. To calculate these you needed to load three files — the corpus, which needed to be loaded as an open file so it could be searched, the list of 2-word concgrams created from the corpus, and the list of unique words created from the same corpus. The values were calculated and displayed in the t-score list box shown in Figure 5. You saved the file as 'cg2-DTSpeech-exclList-span4_TSCORES.txt'.

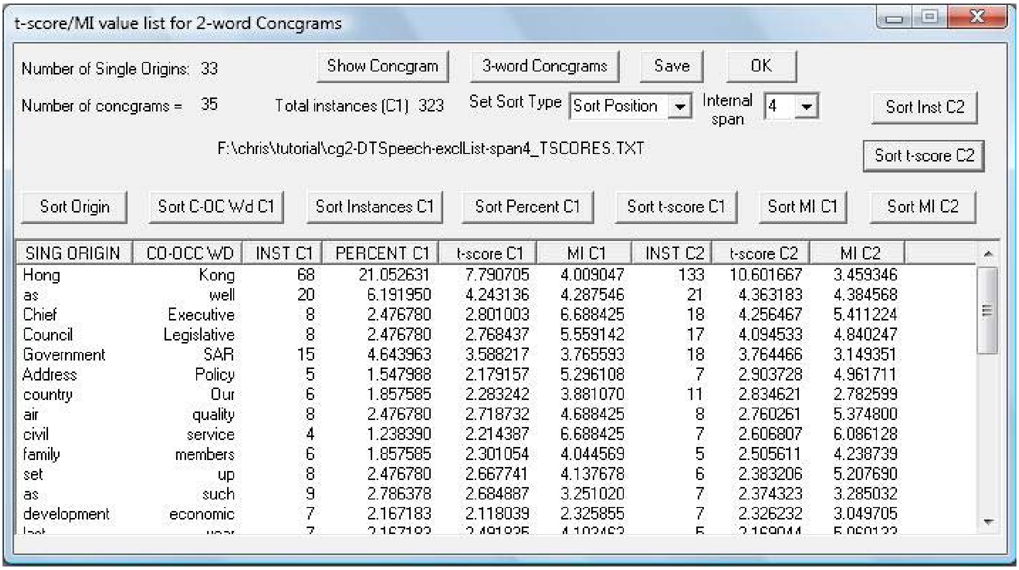
TUTORIAL 6 shows how to compare the t-score and MI value for the same 2-word concgram list with the values gained from a search of a second corpus (using the same concgrams list). Like the Word Specificity Dialog in TUTORIAL 4), this function is also designed to compare a short file in a particular genre with a larger corpus (which may also be in the same specialized genre). In TUTORIAL 6, however, we shall use the same files that we have used before to compare the speech from one year with that given the previous year.

1. First select the CONCGRAM ➡ t-score & MI VALUE FOR 2-WORD CONCGRAMS ➡ CREATE NEW LIST ➡ COMPARE C1 AGAINST C2.
2. The File Dialog prompt first asks you to select the 2-word concgram list containing the statistical values — open the file 'cg2-DTSpeech-exclList-span4_TSCORES.txt' that you created in TUTORIAL 3.
3. The t-score & MI value Preferences Dialog will then be displayed. Set the span to 4 as the 2-word concgrams list was created using span of 4. The other options will be ignored as **Corpus 2** is used for the concgram searches and all results, even if zero, will be stored.
4. The second File Dialog prompt asks you to select the Corpus 2 file to compare with — is this case select the file '2005_06_policy_address_Donald_Tsang.txt'.

5. The third File Dialog prompt asks you to select the Unique Word List file created from Corpus 2 — select the file 'uwords-DTSpeech-2005-6.TXT'.

The same 2-word concgrams list created for Corpus 1 is now used for Corpus 2, and the t-score generated are stored in the List Box for the 2 Corpus files to allow a comparison to be made, as in Figure 9.

Figure 9. The t-score/MI Value List for both speeches after activation of 'Sort t-score C2'



Number of Single Origins: 33

Number of concgrams = 35

Total instances (C1) 323

Set Sort Type: Sort Position

Internal span: 4

F:\chris\tutorial\cg2-DTSpeech-exclList-span4_TSCORES.TXT

SING ORIGIN	CO-OCC WD	INST C1	PERCENT C1	t-score C1	MI C1	INST C2	t-score C2	MI C2
Hong	Kong	68	21.052631	7.790705	4.009047	133	10.601667	3.459346
as	well	20	6.191950	4.243136	4.287546	21	4.363183	4.384568
Chief	Executive	8	2.476780	2.801003	6.688425	18	4.256467	5.411224
Council	Legislative	8	2.476780	2.768437	5.559142	17	4.094533	4.840247
Government	SAR	15	4.643963	3.588217	3.765593	18	3.764466	3.149351
Address	Policy	5	1.547988	2.179157	5.296108	7	2.903728	4.961711
country	Our	6	1.857585	2.283242	3.881070	11	2.834621	2.782599
air	quality	8	2.476780	2.718732	4.688425	8	2.760261	5.374800
civil	service	4	1.238390	2.214387	6.688425	7	2.606807	6.086128
family	members	6	1.857585	2.301054	4.044569	5	2.505611	4.238739
set	up	8	2.476780	2.667741	4.137678	6	2.383206	5.207690
as	such	9	2.786378	2.684887	3.251020	7	2.374323	3.285032
development	economic	7	2.167183	2.118039	2.325855	7	2.326232	3.049705

It is important to sort the origin instances first (for C1 or C2, whichever you wish to look at), before sorting the t-score or MI value.

References

1. Ahmad, K. (2005). Terminology in Text. Paper presented at the Tuscan Word Centre International Workshop: Dial a Corpus. Certosa di Pontignano, Italy.
2. Barnbrook, G. (1996). *Language and Computers: A Practical Introduction to the Computer Analysis of Language,* Edinburgh: Edinburgh University Press, pp 88–106.
3. Cheng, W., Greaves C. and Warren M. (2006) ‘From n-gram to skipgram to concgram’, *International Journal of Corpus Linguistics* 11/4: 411–433.
4. Fletcher, W. H. (2006) “Phrases in English” Home. Retrieved 15 February 2006, from http://pie.usna.edu/.
5. Renouf, A. J. and J. M. Sinclair (1991) ‘Collocational Frameworks in English’, in Ajimer and Altenberg (eds) *English Corpus Linguistics*, pp 128–43.
6. Scott, M. and Tribble, C. (2006). *Textual Patterns: Key words and corpus analysis in language education*. Amsterdam: John Benjamins.
7. Stubbs, M. (1995). Collocations and semantic profiles: on the cause of the trouble with quantitative methods . *Functions of Language*, 2/1: 23–55.
8. Wilks, Y. (2005). REVEAL: the notion of anomalous texts in a very large corpus. Tuscan Word Centre International Workshop. Certosa di Pontignano, Tuscany, Italy, 1–3 July 2005.