Online corpora for specific purposes

Martin Warren, Research Centre for Professional Communication in English English Department, The Hong Kong Polytechnic University

1 Introduction

A team based at the Research Centre for Professional Communication in Hong Kong has compiled two specialised corpora which are publicly available online. One aims to represent the English used by engineering professionals in Hong Kong and the other the English used by professionals in Hong Kong's financial services sector. Each corpus is the largest of its kind compiled to date, but they are also of significance for two other reasons. The corpora were compiled with the usual users in mind (i.e. researchers plus learners and teachers of English for specific purposes), but the main target groups are engineers and financial services professionals in Hong Kong and beyond who, it is hoped, will find the corpora useful resources to enhance their professional communication competencies. Some of the ways that these professionals have been introduced to the corpora are described. The other significant innovation is the software, ConcGramOnline[1], which allows users to go beyond concordances of individual words and clusters. It does this by automatically identifying instances of phraseological variation when two or three words or phrases co-occur, irrespective of either constituency or positional variation.

2 Background

This paper describes two specialised online corpora compiled to represent the English language use of two key sectors of the Hong Kong economy, engineering and financial services, and which are intended to be used by professionals working in those sectors, along with students and teachers of English for specific purposes, as a language support, or language learning, resource. The two corpora are the Hong Kong Engineering Corpus (HKEC) and the Hong Kong Financial Services Corpus (HKFSC).

Sinclair describes the difference between a 'general corpus', which can inform the researcher about the "language as a whole", and a 'specialised corpus', which can provide information about the "characteristics of the genre" (2001: xi). The design of any corpus worthy of study, whether it is a general or specialised corpus, requires the application of corpus design principles. Sinclair (2005a) details what he considers these to be and they are summarized below.

a. Corpus contents are selected based on their communicative purpose without regard for the language that they contain.
b. The corpus is as representative as possible.
c. Only components in the corpus that are designed to be independently contrasted are contrasted
d. Criteria determining the structure of the corpus are small in number, separate from each other, and efficient at delineating a corpus that is representative.
e. Any information about a text is stored separately from the plain text and only merged when needed.
f. Samples of language for the corpus, whenever possible, consist of entire texts.
g. The design and composition of the corpus are fully documented with full justifications.
h. The corpus design includes, as target notions, representativeness, and balance.
i. The control of subject matter in the corpus is imposed by the use of external, and not internal, criteria.
j. The corpus aims for homogeneity in its components while maintaining adequate coverage, and rogue texts should be avoided.

(Sinclair 2005a: 2–17)

The research team involved in the compilation of the two corpora described in this paper adhered to the above principles as far as possible, but, as Sinclair (2005b: 99) points out, "some kinds of data are inherently difficult or even impossible to obtain, and a measure of compromise is often necessary". Below, examples of how the above principles were observed, and how, on occasion, compromises were made, when building the corpora are described and discussed.

3 Specialised corpora

There has been an explosion of studies of specialised corpora as the electronic versions of texts have become the new norm, often replacing hard copies completely, enabling corpus linguists to build corpora, especially written corpora, to meet specific needs more easily. Specialised corpora range in both their size and in the degree of their specificity. So, for example, the Longman/Lancaster Spoken English Corpus is relatively small at 53,000 words while the Cambridge Corpus of Financial English has 55m words. In terms of specificity, there are highly specialised corpora which focus on one particular text type, move, or function, for example, personal advertisements (Ooi 2001), introductions to speakers (Henry and Roseberry 2001), dictionary definitions (Barnbrook and Sinclair 2001), grant proposals (Connor and Upton 2004) and direct mail letters (Connor and Gladkov 2004). Other corpora, while specialised, cover a wider range of text types or language use. Examples of such specialised corpora are the Michigan Corpus of Academic Spoken English (MICASE), which contains a range of spoken academic text types (see, for example, Powel and Simpson 2001), and the International Corpus of English, each component of which contains the same range and quantity of spoken and written text samples drawn from a particular variety of English.

A number of general corpora are publicly accessible online, at least in part, and this is also the case for some specialised corpora. However, to access the full versions of general corpora such as the Bank of English or the British National corpus, the user has to either purchase the corpus or pay some form of subscriber fee. The same is also true for some specialised corpora such as the International Corpus of Learner English, the latest version of which contains 3.7m words of EFL writing from learners representing 16 mother tongue backgrounds, and which has to be purchased (Granger *et al*. 2009). Likewise, the prosodically transcribed 1m-word Hong Kong Corpus of Spoken English is on a CD that comes with the book detailing the corpus and some of the main research studies conducted using the corpus (Cheng, Greaves and Warren 2008). Mentioning the fact that some corpora have to be purchased, or subscribed to, is not intended as a criticism of the compilers, but is done merely to make the point about the availability of corpora; that is whether or not they are freely available which in turn can be an inherent design feature of a corpus. Indeed, many specialised corpora are not made available at all to others in the field which is regrettable. Those that are publicly available are in the minority and these include some well-known specialised corpora such as MICASE, mentioned earlier, and the Corpus of London Teenager Language (COLT) (see Stenström *et al*. 2002), which contains 500,000 words of naturally occurring talk between London teenagers aged 13 to 17 years of age, both of which are fully available online.

The HKEC and the HKFSC were designed to be relatively large, 9.2m and 7.3m words respectively, to cover a wide range of text types and to be publicly available online. The rationale for these design features is detailed in the next section.

4 Compiling the HKEC and the HKFSC

As mentioned earlier, in compiling the HKEC and the HKFSC, Sinclair's design principles were followed as closely as possible by the project team and this necessitated the involvement of professionals, professional associations, and organisations in the respective fields. The project team is comprised of neither engineers nor financial services professionals and so we relied heavily on the advice of these insider sources in order to ensure the corpus contents were both representative and balanced. Professional bodies, such as the Hong Kong Institution of Engineers, the Hong Kong Institute of Certified Public Accountants, the Hong Kong Securities Institute, government departments, and private organisations, as well as individual professionals from the engineering and financial services sectors, all provided invaluable advice in terms of both the texts to include and the proportions to be assigned to certain text types. Their advice was essential to determine the range of text types that professionals read, write, speak, or listen to in English and what weighting specific text types should be accorded in the corpora. While for certain text types, such as 'code of practice' or 'ordinance', there is a finite number of relevant texts in existence and so the upper limit as to the number of such texts to include is pre-determined; for other text types, for example media releases and annual reports, expert advice was needed to ensure they are not over- or under-represented in the two corpora. However, inevitably, there were potential providers of desirable texts for the two corpora who were approached because their texts would have either contributed to the representativeness of the corpora, or their balance, or both, and who simply refused to give their permission for us to include the texts. This has implications for the representativeness and balance of the corpora, but we simply had to accept that we could not include certain texts and did our best to find equivalent sources, where possible.

An overarching aim was for these two specialised corpora to be representative of the texts that engineering and financial services professionals read, write, speak or listen to in English. The reason for this is that the corpora are primarily intended as a resource for these professionals themselves, and for those researching, learning, or teaching the language of these two professions. This aim makes the two corpora different from, for example, the 55m word Cambridge Corpus of Financial English[2] which is made up of books and articles relating to economics and finance from the UK and USA and is part of the Cambridge International Corpus, but which does claim to be made up of the kinds of texts that financial services professionals actually engage with on a regular basis in their workplaces.

Since another aim was to make the corpora publicly available, and to comply with our own code of ethics, it was important to obtain the consent of the copyright holders, but this also meant that, although many of the texts are publicly accessible on the internet, we could not just download them. As a result, the contents are at times a compromise resulting from certain texts being explicitly withheld. Also, a concern expressed by a number of those organisations which did give their consent was that the texts should not be downloadable. As a result, it was agreed that the amount of additional co-text available to users who wished to view more than the concordance line should be restricted to 30 words on either side of the centred word.

Seeking advice from experts in the field is necessary when building a specialist corpus, but, in terms of the HKEC and HKFSC, these experts are also viewed as stakeholders in another sense. These two corpora are aimed at researchers in corpus linguistics, learners and teachers of English for Specific Purposes, and, most importantly, engineers and financial services professionals in Hong Kong and beyond. In order to build a resource for the latter group, it was considered essential to involve them from the outset, not only to advise on the contents of the two corpora, but also to give the corpora credibility in the eyes of these two communities. This approach has then made it easier when we work to promote the corpora and encourage these professionals to use them as a language resource.

5 Contents of the corpora

The contents of the two corpora are detailed in Tables 1 and 2 below:

Table 1: Contents of the HKEC

Text type	Words	Text type	Words
About Us	647,013	Plans	4,173
Abstracts	94,671	Position Documents	75,660
Agreements	127,895	Publicity Material	599,407
Circulars	143,313	Product Descriptions	611,549

Codes of Practice	997,228	Project Summaries	115,829
Conference Proceedings	196,498	Q & A	27,703
Consultation Papers	111,494	Reports	979,170
Fact Sheets	26,059	Review papers	106,506
Frequently Asked Questions	55,726	Speeches	2,822
Guides	783,805	Standards	136,024
Handbooks	67,284	Technical papers	65,731
Letters to the Editor	3,492	Tender Notices	4,242
Manuals	296,299	Transaction Discussions (HKIE)	7,149
Media releases	1,566,742	Transaction Notes (HKIE)	79,058
Notes	156,255	Transaction Proceedings (HKIE)	1,055,248
Ordinances	139,176	TOTAL	9,224,384

Table 2: Contents of the HKFSC

Text type	Words	Text type	Words
Annual Reports	1,274,618	Insurance Product Descriptions	103,439
Brochures	12,770	Investment Product Descriptions	339,853
Bank Service Charges	20,354	Model Agreements	6,913
Codes of Practice	24,151	Media Releases	886,516
Corporate Announcements	88,246	Ordinances	384,910
Circulars	401,038	Procedures	1,198
Fund Descriptions	18,109	Principles	1,467
Fund Reports	70,842	Prospectuses	1,959,111
Factsheets	14,893	Rules	5,711
Guidelines	131,452	Results Announcements	321,426
General Meetings	28,226	Standards	12,766
Insurance Policies	16,407	Speeches	609,821
Interim Reports	603,862	TOTAL	7,334,908

It can be seen that almost all of the text types in both corpora are written. This reflects the reality of the language use of most professionals in Hong Kong (Evans and Green 2001) where Cantonese, and increasingly Putonghua (Mandarin), is used for spoken communication and English is used for almost all written communication. As Evans and Green's study makes clear (2001:

247), since the handover of Hong Kong to China in 1997, English continues to be “the unmarked language of internal and external written communication in both the public and private sectors”. For example, in a meeting involving Hong Kong Chinese professionals in the workplace, Cantonese is used and then the minutes are written up in English. Similarly, while the ordinances in Hong Kong are drafted in both Chinese and English, it is the English version which takes precedence in the courts. These manifestations of Hong Kong’s official language policy and professional communication practices, locally termed ‘biliteracy and trilingualism’, result in the two corpora being comprised mostly of written texts. The HKFSC contains more spoken texts due to the larger number of public speeches in this field.

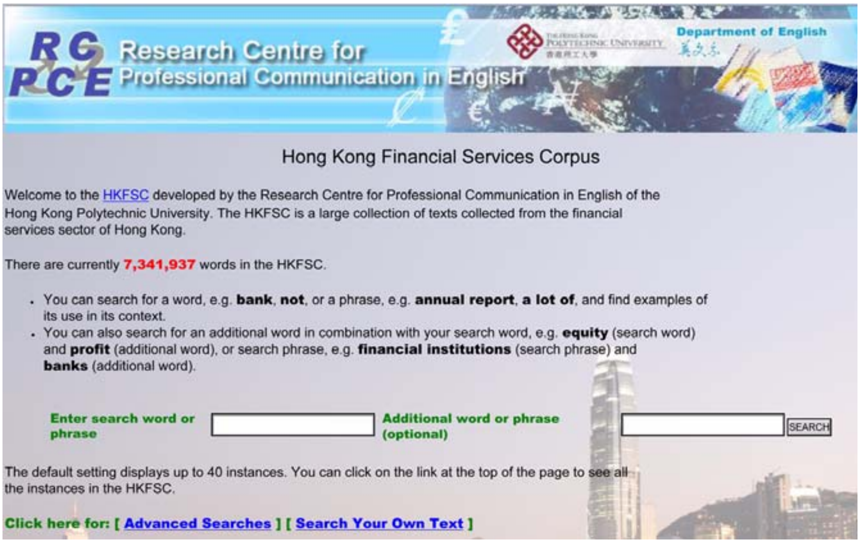
Most of the text types are self-explanatory, but one or two require brief explanations. The text type ‘About Us’ is often found on the websites of organisations where they describe the nature of their work, and the goods or services they provide, and has now come to replace traditional hard copy material such as brochures. In the HKEC, there are three kinds of ‘Transaction’ – ‘Discussions’, ‘Notes’ and ‘Proceedings’ – almost all of which were provided by the Hong Kong Institution of Engineers and denote three text types in the journal of engineering research and development published each year by this professional body. ‘Discussions’ are very short pieces on a point of current interest, ‘Notes’ are short pieces covering work in progress, and ‘Proceedings are full-blown papers detailing research and development across all of the main fields of engineering. In the HKFSC, ‘Prospectuses’ are the documents which must be made publicly available by companies planning an initial public offering (IPO) on the stock exchange. Also in the HKFSC, ‘Model Agreements’ are agreement templates provided by professional bodies for organisations to adapt to suit their needs.

There are certain text types in common across the two corpora such as ‘Media releases’ related to their respective fields and, interestingly, these are relatively more common in the engineering corpus. Both corpora also have ‘Codes of Practice’ and, again, these texts, which are widely referred to by both sets of professionals, constitute a far bigger slice of the HKEC (10.8 per cent) than the HKFSC (0.3 per cent). This reflects the higher number of different codes of practice across the various fields of engineering compared to the financial services sector. Conversely, ‘ordinances’ are a more significant source of reference material for professionals working in financial services (5.25 per cent) than in engineering (1.5 per cent). ‘Speeches’, including ‘Conference Proceedings’ in the HKEC, which were referred to indirectly earlier, are more likely to be made, or listened to, by those in the financial services sector than by engineers. They represent 8.3 per cent of the HKFSC and 2.16 per cent of the HKEC. ‘Product Descriptions’ of various kinds are found to be proportionally very similar at 6.63 per cent in the HKEC and 6.2 per cent in the HKFSC.

When searching these corpora online, the default setting is to search the full corpus, but the user can select any of the text types detailed in Tables 1 and 2. The text type selected is then treated as a separate sub-corpus for the purposes of conducting searches.

6 Online interface

As has been mentioned before, these two corpora are primarily aimed at engineering and financial services professionals who have no knowledge of corpus linguistics and no training in how to interrogate a corpus. With this in mind, the interface is designed to be as simple and as user-friendly as possible. The corpora are accessed by means of a link on the home page of our research centre which then takes the user to a list of the various corpora available. The user clicks on the corpus of choice and, by default, is taken to the page containing very basic search functions as shown in Figure 1:


RC PCE
Research Centre for Professional Communication in English
The Hong Kong Polytechnic University
Department of English

Hong Kong Financial Services Corpus

Welcome to the HKFSC developed by the Research Centre for Professional Communication in English of the Hong Kong Polytechnic University. The HKFSC is a large collection of texts collected from the financial services sector of Hong Kong.

There are currently **7,341,937** words in the HKFSC.

- You can search for a word, e.g. **bank**, **not**, or a phrase, e.g. **annual report**, **a lot of**, and find examples of its use in its context.
- You can also search for an additional word in combination with your search word, e.g. **equity** (search word) and **profit** (additional word), or search phrase, e.g. **financial institutions** (search phrase) and **banks** (additional word).

Enter search word or phrase
Additional word or phrase (optional)
SEARCH

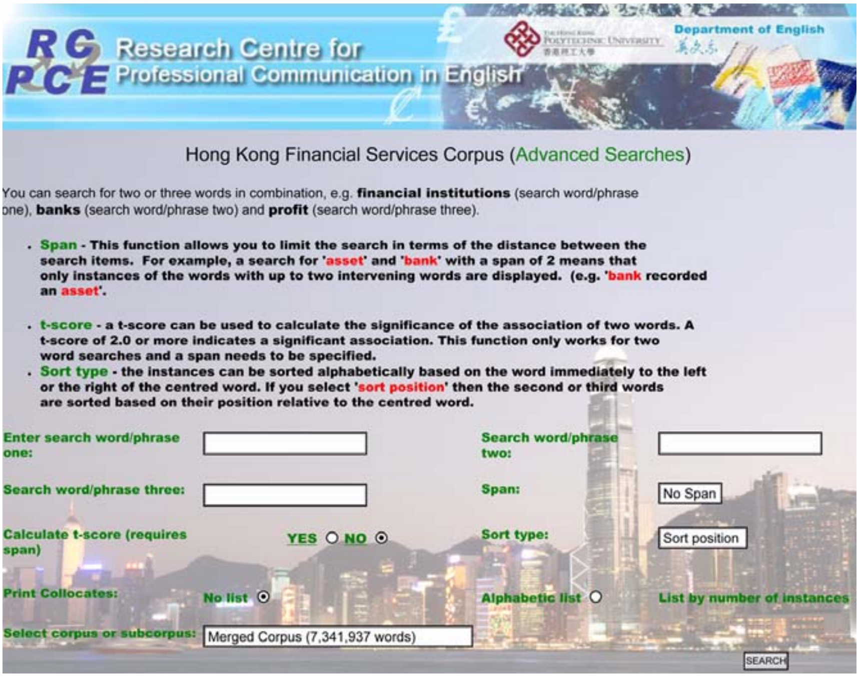
The default setting displays up to 40 instances. You can click on the link at the top of the page to see all the instances in the HKFSC.

Click here for: [Advanced Searches] [Search Your Own Text]

Figure 1: Basic search functions

The user is able to search for a single word or phrase, or the user can include one additional word or phrase in a search. The resulting concordance is then displayed. If there are more than forty instances in the corpus, only forty randomly chosen concordance lines are displayed. There is, however, the option to see all of the instances, and it is possible to search for more specific forms of the centred word by using a search function at the bottom of the page displaying the concordance.

As can be seen in Figure 1, at the bottom of the page there are links to ‘Advanced Searches and ‘Search your own text’. The latter option allows users to upload their own text, or corpus, and then search it using ConcGramOnline. This facility allows users to take advantage of the innovative software to identify the phraseology, especially phraseological variation, in their own texts and corpora. Selecting ‘Advanced Searches’ takes the user to the page shown in Figure 2:



Hong Kong Financial Services Corpus (Advanced Searches)

You can search for two or three words in combination, e.g. **financial institutions** (search word/phrase one), **banks** (search word/phrase two) and **profit** (search word/phrase three).

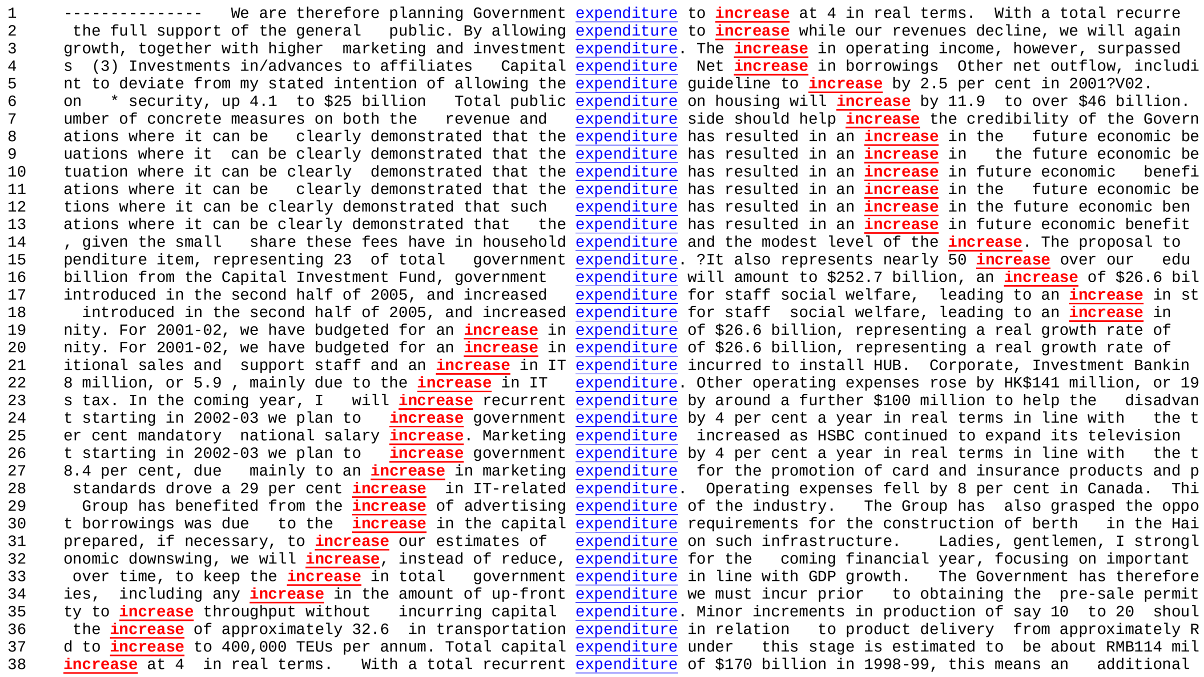
- **Span - This function allows you to limit the search in terms of the distance between the search items. For example, a search for 'asset' and 'bank' with a span of 2 means that only instances of the words with up to two intervening words are displayed. (e.g. 'bank recorded an asset'.**

- **t-score - a t-score can be used to calculate the significance of the association of two words. A t-score of 2.0 or more indicates a significant association. This function only works for two word searches and a span needs to be specified.**
- **Sort type - the instances can be sorted alphabetically based on the word immediately to the left or the right of the centred word. If you select 'sort position' then the second or third words are sorted based on their position relative to the centred word.**

Enter search word/phrase one:
Search word/phrase two:
Search word/phrase three:
Span: No Span
Calculate t-score (requires span) YES ○ NO ●
Sort type: Sort position
Print Collocates: No list ● Alphabetic list ○ List by number of instances
Select corpus or subcorpus: Merged Corpus (7,341,937 words)

Figure 2: Advanced search functions

The advanced searches allow for up to three words or phrases to be searched at a time with further options such as setting a span, sorting left or right, calculating t-scores, displaying collocates, and selecting a sub-corpus based on a text type contained in the full corpus. These functions are fairly standard for any concordancing software, but it is ConcGramOnline's ability to automatically find all of the co-occurrences of up to three words/phrases nominated by the user which is innovative. It needs to be pointed out that the software is limited in functionality by the fact that it is being used online. This means that it is unable to fully automatically concgram a whole corpus, or sub-corpus, because of online logistical constraints. As a result, users have to nominate the word(s) or phrase(s) for each search. The co-occurrences found in such searches are called 'concgrams' (Cheng, Greaves and Warren 2006). A concordance based on a search for all of the co-occurrences of *expenditure* and *increase* is shown in Figure 3 and illustrates the extent of the phraseological variation uncovered by ConcgramOnline. The conventions for writing concgrams is to write the words in a concgram in alphabetical order separated by a forward slash, in this case *expenditure/increase*.



```
1    ---------------   We are therefore planning Government expenditure to increase at 4 in real terms.  With a total recurre
2     the full support of the general   public. By allowing expenditure to increase while our revenues decline, we will again
3    growth, together with higher  marketing and investment expenditure. The increase in operating income, however, surpassed
4    s  (3) Investments in/advances to affiliates   Capital expenditure  Net increase in borrowings  Other net outflow, includi
5    nt to deviate from my stated intention of allowing the expenditure guideline to increase by 2.5 per cent in 2001?V02.
6    on   * security, up 4.1  to $25 billion   Total public expenditure on housing will increase by 11.9  to over $46 billion.
7    umber of concrete measures on both the   revenue and   expenditure side should help increase the credibility of the Govern
8    ations where it can be   clearly demonstrated that the expenditure has resulted in an increase in the   future economic be
9    uations where it  can be clearly demonstrated that the expenditure has resulted in an increase in   the future economic be
10   tuation where it can be clearly  demonstrated that the expenditure has resulted in an increase in future economic   benefi
11   ations where it can be   clearly demonstrated that the expenditure has resulted in an increase in the   future economic be
12   tions where it can be clearly demonstrated that such   expenditure has resulted in an increase in the future economic ben
13   ations where it can be clearly demonstrated that   the expenditure has resulted in an increase in future economic benefit
14   , given the small   share these fees have in household expenditure and the modest level of the increase. The proposal to
15   penditure item, representing 23  of total   government expenditure. ?It also represents nearly 50 increase over our   edu
16   billion from the Capital Investment Fund, government   expenditure will amount to $252.7 billion, an increase of $26.6 bil
17   introduced in the second half of 2005, and increased   expenditure for staff social welfare,  leading to an increase in st
18     introduced in the second half of 2005, and increased expenditure for staff  social welfare, leading to an increase  in
19   nity. For 2001-02, we have budgeted for an increase in expenditure of $26.6 billion, representing a real growth rate of
20   nity. For 2001-02, we have budgeted for an increase in expenditure of $26.6 billion, representing a real growth rate of
21   itional sales and  support staff and an increase in IT expenditure incurred to install HUB.  Corporate, Investment Bankin
22   8 million, or 5.9 , mainly due to the increase in IT   expenditure. Other operating expenses rose by HK$141 million, or 19
23   s tax. In the coming year, I   will increase recurrent expenditure by around a further $100 million to help the   disadvan
24   t starting in 2002-03 we plan to   increase government expenditure by 4 per cent a year in real terms in line with   the t
25   er cent mandatory  national salary increase. Marketing expenditure  increased as HSBC continued to expand its television
26   t starting in 2002-03 we plan to   increase government expenditure by 4 per cent a year in real terms in line with  the t
27   8.4 per cent, due   mainly to an increase in marketing expenditure  for the promotion of card and insurance products and p
28    standards drove a 29 per cent increase  in IT-related expenditure.  Operating expenses fell by 8 per cent in Canada.  Thi
29     Group has benefited from the increase of advertising expenditure of the industry.   The Group has  also grasped the oppo
30   t borrowings was due   to the  increase in the capital expenditure requirements for the construction of berth   in the Hai
31   prepared, if necessary, to increase our estimates of   expenditure on such infrastructure.    Ladies, gentlemen, I strongl
32   onomic downswing, we will increase, instead of reduce, expenditure for the   coming financial year, focusing on important
33    over time, to keep the increase in total   government expenditure in line with GDP growth.   The Government has therefore
34   ies,  including any increase in the amount of up-front expenditure we must incur prior   to obtaining the  pre-sale permit
35   ty to increase throughput without   incurring capital  expenditure. Minor increments in production of say 10  to 20  shoul
36    the increase of approximately 32.6  in transportation expenditure in relation   to product delivery  from approximately R
37   d to increase to 400,000 TEUs per annum. Total capital expenditure under   this stage is estimated to  be about RMB114 mil
38   increase at 4  in real terms.   With a total recurrent expenditure of $170 billion in 1998-99, this means an   additional
```

Figure 3: The two-word concgrams expenditure/increase in the HKFSC

In the above concordance, it can be seen that the display of the concordance lines aims to be reader-friendly. The various concgram configurations for *expenditure/increase*, when *expenditure* is the centred word, start with those instances where *increase* is to the right and closest to *expenditure*, and subsequent lines show instances ordered according to the distance between *expenditure* and *increase*. Once all of the instances to the right of the centred word have been displayed, those to the left are shown based on the same principle. Creating a reader-friendly display is a key consideration because of the way in which the software finds all of the co-occurrences of up to three words regardless of constituency (ranging from one intervening word, "increase in expenditure", to five, "the increase of approximately 32.6 in transportation expenditure") or positional variation (for example, "increase in total government expenditure" versus "Total public expenditure on housing will increase by 11.9 to over $46 billion"). When such a range of variation is presented, the concordance needs to be arranged to make it as orderly as possible for the user to then analyse the patterning. The use of colour to represent each of the words in a concgram also makes it much easier to identify the members of a concgram immediately.

An important feature of ConcGramOnline, which is derived from the more powerful ConcGram (Greaves 2009), becomes clear as soon as a concgram concordance is displayed because the highlighted concgrams present a very different view of word co-occurrences compared with the traditional KWIC display. KWIC displays highlight the node, termed 'centred word' in Concgram (Cheng *et al*. 2006), and this has tended to make the node the main focus of any analysis. As a result, any associated words risk being viewed as in some way subordinate relative to it. ConcGramOnline, following the design principles of ConcGram, highlights all of a concgram's co-occurring words in a concordance. This then has the beneficial effect of making word co-occurrences very much the focus of study rather than the node.

It has to be pointed out that, while the software is very effective when it comes to identifying up to three co-occurring words or phrases, the user needs to then study the concordance lines to determine whether they are meaningfully associated. This is a useful exercise in itself, especially for learners of English, and it can be seen in Figure 3 that the concgrams in a number of the lines, for example lines 4, 7–15, 17, 18, 25, 37 and 38, are arguably co-occurrences rather than meaningfully associated.

7 Phraseological profiles and aboutgrams

There have been a number of studies of keywords and the notion of keyness is well-known in corpus linguistics (see, for example, Scott and Tribble 2006). ConcGramOnline can be used to extend the notion of keyness beyond keywords to include a fuller range of phraseology. Concgrams provide useful raw data which can reveal the co-selections made by the speakers and writers in a corpus. They are the first stage towards quantifying the extent of phraseology in a corpus and, once their associatedness is confirmed, the phraseological profile of the language. There is evidence to suggest that n-grams (also known as ‘lexical bundles’ or ‘lexical clusters’) are genre-based (see, for example, Carter and McCarthy 2006: 828–837; Scott and Tribble 2006: 131–159). There is evidence that this is also the case for concgrams. Studies using concgrams to examine the aboutness of corpora (see Tognini Bonelli 2006; Greaves and Warren 2007; Cheng 2008; Milizia and Spinzi 2008; O’Donnell, Scott and Mahlberg 2008) show that this is indeed the case.

According to Phillips (1989), aboutness is a result of global patternings. He argues that it should be possible to identify them by computational means, so that they are derived from the corpus rather than external features. The phraseological profile is all of the word associations in a corpus, and the aboutness of the corpus can be derived from the word associations specific to that particular corpus. Phraseologies which are specific to a text or corpus, either because they only occur in a particular text or corpus, or because they occur more frequently, are termed ‘aboutgrams’ (Sinclair, personal communication). The notion of a phraseological profile of a corpus, and the identification of potential aboutgrams, are best illustrated with reference to Table 3 below:

Table 3: The twenty-five most frequent two-word concgrams in the HKEC and HKFSC

HKEC 2-word concgrams	frequency	HKFSC 2-word concgrams	frequency
Hong Kong	37,928	Hong Kong	34,942
less/than	4,705	HK/million(s)	15,043
energy/efficiency	4,559	exchange/stock	4,707
ebb/mid	4,507	per/share	4,165
flood/mid	4,497	interest/rate(s)	4,050
more/than	4,357	Hong (Kong)/limited	3,931
carried/out	3,854	offer/share(s)	3,806
air/quality	3,559	fair/value	3,568
HK/million(s)	3,128	out/set	3,373
environmental/protection	2,742	option(s)/share	3,206
department/services	2,680	HK/per	3,080
quality/water	2,417	more/than	3,054
consumption/energy	2,291	ended/month(s	2,872
co./ltd.	2,278	million/RMB	2,846
code/practice	2,267	ended/year	2,721
supply/water	2,171	financial/statement(s)	2,704
Hong Kong/university	2,051	income/net	2,617
electrical/mechanical	2,045	general/meeting	2,554
as/possible	1,982	section/under	2,493
monitoring/quality	1,978	ended/June	2,480
shall/provided	1,921	executive/non	2,415
measures/mitigation	1,902	exchange/fund	2,395
domestic/non	1,894	months/six	2,327
control/system	1,893	listing/rules	2,302
building/works	1,832	holdings/limited	2,056

Table 3 presents the twenty-five most frequent lexically-rich concgrams in the two corpora. A concgram is classified as lexically-rich if at least one of the words in the concgram is what is traditionally termed a 'lexical word'. The lists themselves form part of the phraseological profile of each corpus in that the contents are a partial description of the phraseologies to be found in each corpus. In terms of the aboutness of the two corpora, it can be seen that there are many obvious differences between the two lists with only three overlapping concgrams: *Hong Kong*, *HK/million(s)* and *more/than*. It can be argued that *Hong Kong* is an indicator of the aboutness of the two corpora, which specifically target language use in Hong Kong, and so, while not 'about' one specific corpus, it is an aboutgram one might predict to find in a Hong Kong-based corpus. *HK/million(s)* is frequent in both corpora, but it is almost five times more frequent in the HKFSC than the bigger HKEC and might be classified as a financial services-related aboutgram. *More/than* is interesting because it occurs with a similar frequency in both corpora and a search in a 5m-word sample of the British National Corpus (BNC5m) shows a similar proportion of this phraseology (2,692) in general language use. This suggests that *more/than* is not an aboutgram in the corpora.

There are a number of unsurprising aboutgrams in both lists in that they are clearly phraseologies one might expect to find in specialised corpora such as these. The HKEC has aboutgrams such as *energy/efficiency*, *electrical/mechanical*, and *control/system*, while the HKFSC has *exchange/stock*, *financial/statement(s)*, and *listing/rules*. More interesting, perhaps, are phraseologies which are less obviously engineering- or financial services-related, but which occur with a much higher frequency in one or both of the specialised corpora and so qualify as aboutgrams. An example of this type of aboutgram is *less/than* which occurs 4,705 times in the HKEC, 1,845 times in the HKFSC, and only 734 times in the BNC5m. In this case, an aboutgram has been identified for both specialised corpora with it being a stronger indicator of aboutness in the HKEC. The aboutgram *out/set* is another aboutgram common to both specialised corpora and is used by writers referring to other texts such as ordinances, guidelines and codes of practices: “the principles set out in this Code”, “the basis of allotment of new shares is set out in paragraphs 2 and 3 of the Appendix”, and “carrying out safety management functions as set out in section 5.1.4 of this Technical Memorandum”. It occurs 3,373 times in the HKFSC, 1,350 times in the HKEC, with only 251 instances in the BNC5m, and so it is a stronger indicator of aboutness in the HKFSC.

An aboutgram found only in the HKEC is *carried/out* (3,854) which is far more frequent than in either the HKFSC (468) and the BNC5m (318). It seems that engineers often describe processes that have been or will be carried out, as in “This straightening should be carried out cold.”, and “The measurements were carried out in a scaled-down room …”. Two more aboutgrams which are specific to the HKEC are *as/possible* (1,982) and *shall/provided* (1,921). The former is much less frequent in the HKFSC (435), and the BNC5m (680), as is the latter with 201 instances in the HKFSC and five in the BNC5m. These aboutgrams are often used by engineers when they issue guidelines, “timber board support should be provided as far as possible to the exposed excavation face”, and invoke codes of practice, “Sufficient ties to the building structure shall be provided to ensure stability.” Two aboutgrams in the HKFSC are *ended/June* (2,480) and *months/six* (2,327) which outnumber the instances found in the HKEC (52 and 501, respectively) and the BNC5m, which has no instances of *ended/June* and 226 *months/six*. The presence of these aboutgrams can be explained by the fact that professionals in the financial services sector frequently prepare reports on a six monthly basis.

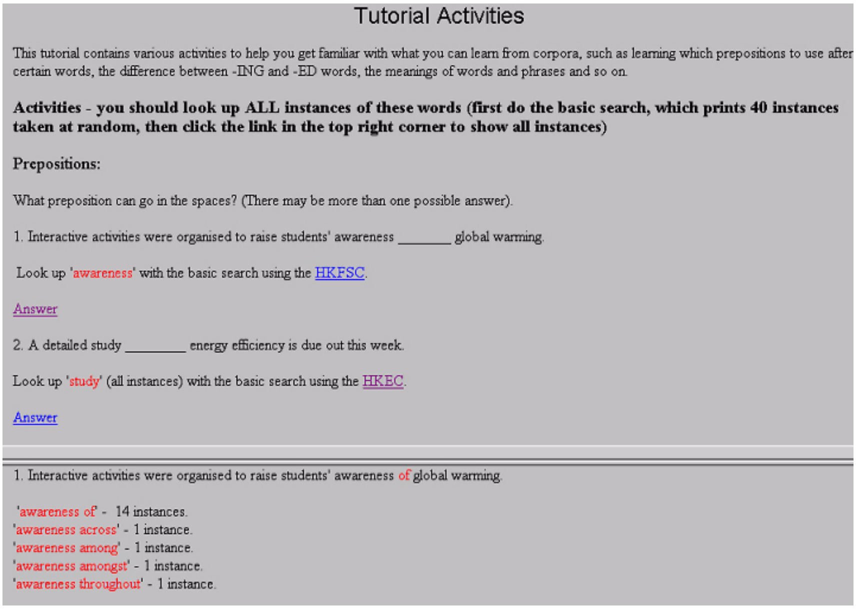
8 Promoting the use of the corpora to professionals

To be successful communicators in their workplaces, professionals need to have an awareness of language that enables them to analyse language use, make informed decisions about language choices, and be creative and in control of their use of language. Bhatia (2004: 146) identifies three elements which together constitute professional expertise: disciplinary knowledge, professional practice and discursive competence. It is the latter, discursive competence in professional contexts, which is being targeted in the development and promotion of the two specialised corpora. It operates at the levels of textual competence, generic competence, and social competence (Bhatia 2004: 144). Textual competence is the ability to master language and use textual, contextual, and pragmatic knowledge in the construction and interpretation of texts. Generic competence is the ability to respond to familiar and unfamiliar communicative situations by producing, interpreting and using generic conventions in one’s profession to achieve professional-specific goals. Bhatia (2004) describes the last of these competences, social competence, as the ability of professionals to use language beyond their professional community in a range of social and institutional contexts and so express their social identity.

The two profession-specific corpora described in this paper are seen as a significant language resource for professionals to use to further enhance their discourse competence. The only problems are: how do we inform professionals of their existence and how do we encourage them to make use of them?

As mentioned earlier in this paper, the team compiling the corpora worked closely with a number of professional bodies, organisations, and individual professionals to determine the contents. Upon completion of the corpora, all of these stakeholders have been informed and, with the help of professional bodies, a series of training workshops has been arranged to teach professionals how to use the corpora as a language resource. In Hong Kong, as elsewhere, professional bodies require that their members take a certain number of continuing professional development (CPD) credits in order to maintain their professional status. This requirement has meant that the training workshops have taken place under the auspices of CPD, hosted by professional bodies in engineering and financial services.

In addition, a number of tutorial activities have been developed to act as online support for users of the corpora, as illustrated in Figure 4 below:



Tutorial Activities

This tutorial contains various activities to help you get familiar with what you can learn from corpora, such as learning which prepositions to use after certain words, the difference between -ING and -ED words, the meanings of words and phrases and so on.

Activities - you should look up ALL instances of these words (first do the basic search, which prints 40 instances taken at random, then click the link in the top right corner to show all instances)

Prepositions:

What preposition can go in the spaces? (There may be more than one possible answer).

1. Interactive activities were organised to raise students' awareness _______ global warming.

Look up 'awareness' with the basic search using the HKFSC.

Answer

2. A detailed study ________ energy efficiency is due out this week.

Look up 'study' (all instances) with the basic search using the HKEC.

Answer

1. Interactive activities were organised to raise students' awareness of global warming.

'awareness of' - 14 instances.
'awareness across' - 1 instance.
'awareness among' - 1 instance.
'awareness amongst' - 1 instance.
'awareness throughout' - 1 instance.

Figure 4: Online tutorial support

In the activity depicted in Figure 4, users are asked a question and then instructed to search for a particular word in one of the specialised corpora to help them answer the question. Once users have an answer to the question, they can check their answer against a suggested answer by clicking on ‘Answer’. Currently, these activities include a range of potential applications and more will be added. The activities include choosing the best preposition, matters relating to co-selections such as collocation, colligation, semantic preference, semantic prosody, and the use of metaphors.

9 Conclusions

This paper has described two new specialised online corpora: the HKEC and the HKFSC. These corpora have been designed with the help of professionals in the related fields and are primarily aimed at providing professional communication support to professionals working in the engineering and financial services sectors by empowering them to better understand the language of their professions. It has also introduced a new computer-based methodology, ‘concgramming’ (Cheng, Greaves and Warren 2006; Greaves and Warren 2007), and discusses how it can be used to facilitate the introduction of phraseology via the specially written software, ConcGramOnline. It is argued that the phraseology of the language contained in a text or a corpus that is specific to a profession constitutes the ‘aboutness’ (Phillips 1989) of the profession. The aim is to empower professionals, and learners of ESP, to learn the phraseology characteristic of their profession, and most importantly, to acquire the techniques needed to critically analyse the language, as a step towards achieving ‘textual competence’, and hence ‘discursive competence’ (Bhatia 2004: 146).

Close collaboration took place with Professional Associations, such as the Hong Kong Institute for Certified Public Accountants and the Hong Kong Institution of Engineers, to compile the specialised corpora. Also, a series of CPD workshops has been jointly hosted by our research centre and these professional bodies to promote the use of the corpora to their members as an invaluable resource capable of further enhancing their professional communication.

Acknowledgements

The work described in this paper was substantially supported by a grant from the Research Grants Council of the Hong Kong Special Administration Region (Project No. PolyU 5459/08H, B-Q11N). The author is grateful to all those who supported the compilation of the two corpora, and to Winnie Cheng and Chris Greaves who are fellow members of the project team.

Notes

1. ConcGramOnline is written by Chris Greaves, Senior Research fellow at The Hong Kong Polytechnic University, and is a simpler online version of ConcGram (Greaves 2009).
2. http://www.cambridge.org/elt/corpus/international_corpus.htm, accessed 22.11.2009.

References

Barnbrook, Geoff and John McHardy Sinclair. 2001. Specialised corpus, local and functional grammars. In M. Ghadessy, A. Henry and R. L. Roseberry (eds.). *Small corpus studies and ELT*, 237–276. Amsterdam: John Benjamins.

Bhatia, Vijay K. 2004. *Worlds of written discourse*. London and New York: Continuum.

Carter, Ron and Michael McCarthy. 2006. *Cambridge grammar of English*. Cambridge: Cambridge University Press.

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

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

Cheng, Winnie, Chris Greaves and Martin Warren. 2008. *A corpus-driven study of discourse intonation: The HKCSE (prosodic)*. Amsterdam: John Benjamins.

Connor, Ulla and Kostya Gladkov. 2004. Rhetorical appeals in fundraising letters. In U. Connor and T. Upton (eds.). *Discourse in the professions: Perspectives from corpus linguistics*, 257–286. Amsterdam: John Benjamins.

Connor, Ulla and Thomas Upton. 2004. The genre of grant proposals: A corpus linguistic analysis. In U. Connor and T. Upton (eds.). *Discourse in the professions: Perspectives from corpus linguistics*, 235–255. Amsterdam: John Benjamins.

Evans, Stephen and Christopher Green. 2001. Language in post-colonial Hong Kong: The roles of English and Chinese in the public and private sectors. *English World-Wide 22(2): 247–268.*

Granger, Sylviane, Estelle Dagneaux, Fanny Meunier and Magali Paquot. 2009. International Corpus of Learner English v2. Louvain-la-Neuve, Belgium: Presses universitaires de Louvain.

Greaves, Chris. 2009. *ConcGram 1.0: A phraseological search engine*. Amsterdam: John Benjamins.

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

Milizia, Denise and Cinzia Spinzi. 2008. The ‘terridiom’ principle between spoken and written discourse. *International Journal of Corpus Linguistics* 13(3) 322–350.

O’Donnell, Matthew Brook, Mike Scott and Michaela Mahlberg. 2008. Exploring text-initial concgrams in a newspaper corpus. Paper presented at the 7th International Conference of the American Association of Corpus Linguistics, Brigham Young University, Provo, Utah, USA, 12–15 March, 2008.

Ooi, Vincent B.Y. 2001. Investigating genres using the world wide web. In M. Ghadessy, A. Henry and R.L. Roseberry (eds.). *Small corpus studies and ELT,* 175–203. Amsterdam: John Benjamins.

Phillips, Martin. 1989. *Lexical structure of text. Discourse analysis monographs: 12*. English Language Research: University of Birmingham.

Powell, Christina and Rita Simpson. 2001. Collaboration between corpus linguists and digital librarians for the MICASE web search interface. In R. C. Simpson and J. M. Swales (eds.). *Corpus linguistics in North America: Selections from the 1999 symposium,* 32–47. Ann Arbor: University of Michigan Press.

Scott, Mike and Christopher Tribble. 2006. *Textual patterns: Key words and corpus analysis in language education*. Amsterdam: John Benjamins.

Sinclair, John McH. 2001. Preface. In M. Ghadessy, A. Henry and R. L. Roseberry (eds.). *Small corpus studies and ELT,* vii–xv. Amsterdam: John Benjamins.

Sinclair, John McH. 2005a. Corpus and text: Basic principles. In M. Wynne (ed.). *Developing linguistic corpora: A guide to good practice,* 1–21. Oxford: Arts and Humanities Data Service.

Sinclair, John McH. 2005b. Appendix: How to build a corpus. In M. Wynne (ed.). *Developing linguistic corpora: A guide to good practice,* 98–103. Oxford: Arts and Humanities Data Service.

Stenström, Anna-Brita, Gisle Andersen and Ingrid K. Hasund. 2002. *Trends in teenage talk: Corpus compilation, analysis and findings*. Amsterdam: John Benjamins.

Tognini-Bonelli, Elena. 2006. The corpus as an onion: The CÆT Corpus Siena (a corpus of academic economics texts). Paper presented at the International Seminar: Special and Varied Corpora. Tuscan Word Centre, Italy, 27–31 October, 2006.