Cheng, Winnie (2007) Concgramming: A Corpus-Driven Approach to Learning the Phraseology of Discipline-Specific Texts. *CORELL: Computer Resources for Language Learning* 1, 22-35.

Concgramming: A Corpus-Driven Approach to Learning the Phraseology of Discipline-Specific Texts

Winnie Cheng
The Hong Kong Polytechnic University, China

Abstract. The paper introduces an innovative computer-based methodology, 'concgramming', to automatically identify the phraseological profile, and hence the 'aboutness', of a text or a corpus. This methodology can be employed by students and teachers of any disciplines, so that their awareness of the importance of the phraseological tendency in language will be raised, and their knowledge and skills regarding phraseology enhanced. The paper outlines both specific components and pedagogical implications of the methodology, by describing and exemplifying a set of engaging, interesting and collaborative activities that involve students in a data-driven learning mode.

Key words. aboutness, concgramming, concgrams, data-driven language learning, phraseological profile, phraseology

1. Introduction

When search engines were first designed to identify texts, single words were the typical means, i.e. Keyword In Context (KWIC) (Tribble 1997; Scott 1997, 2000, 2001). Even when more than one word is entered, the search does not treat the words as a linguistic structure, but rather as a set of co-ordinates that define a virtual space within which lie the texts required if the user is successful (Sinclair 2005). The over-reliance on keywords results in the loss of most of the important information regarding the content of individual texts. Sinclair (2005) argues that 'a word on its own is usually not distinctive enough to deliver a stable and precise meaning (outside the protected words which are recognised as technical terms – and even they are always at risk)'.

This paper focuses on examining textual meanings through studying the phraseology of the text. Following Clear (1993), the paper defines 'phraseology' as 'the recurrent co-occurrence of words' (ibid.: 277). Corpus linguists examining word co-occurrences in a range of text corpora have largely contributed to our understanding of, for example, pattern grammar (e.g. Hunston and Francis 2000), phraseology (e.g. Sinclair 1987; Sinclair 1996; Sinclair 2004a; Cowie 1998; Stubbs 2001; Tognini-Bonelli 2001, 2002; Halliday, Teubert and Yallop 2002; Teubert 2005), lexical 'clusters' (e.g. Biber *et al.* 1999; Biber *et al*. 2004; Simpson and Swales 2001), and semantic prosody (e.g. Louw 1993; Sinclair 1991). Biber *et al*. (1999), for instance, discuss lexical 'bundles' in terms of register variation across speech and writing, and classify them according to the structural patterns that they encompass and the grammatical category of the end word of a lexical bundle (ibid: 996-997). Carter and McCarthy (2006: 504-505) also analyse the structure of 'clusters' along with their functions across different genres. The phraseological tendency of natural language, whereby words are co-selected, rather than being selected separately constrained only by grammar, underlies Sinclair's (2004a: 29) 'idiom principle'. Underlying his notion of idiomaticity of natural language are the five categories of co-selection in his description of a lexical item (Sinclair, 1996, 2004a), namely the two obligatory categories of semantic prosody and the invariable core, and the three optional categories of semantic preference, collocation and colligation.

A few studies that examine the phraseological patterns in a range of public texts in Hong Kong have been reported (Cheng 2004, 2006). Cheng (2004) describes the analysis of twelve public speeches made by The Honourable Tung Chee-hwa, Chief Executive of Hong Kong Special Administrative Region (HKSAR), between October and December 2001. The speeches were examined in terms of the ways in which a public speaker constructs a relationship with the audience and conveys particular meanings and ideological positions by means of making phraseological and intonational choices, both directly and indirectly. Another study (Cheng 2006) examines a selection of spoken discourse events collected in Hong Kong during and in the immediate aftermath of the SARS crisis in 2003. The findings show that once the overlapping patterns of co-selection of the most frequently occurring lexical words in the SARS corpus have been determined, it is possible to describe the cumulative effects of the habitual co-selection in the lexical items (Sinclair 1996, 2004a) that contribute to textual meanings and coherence within and across the public texts. Cheng (2006) argues that, compared to lexical cohesion (i.e. lexical reiteration and collocation) (Halliday and Hasan 1976), patterns of co-selection provide a fuller picture of textual and intertextual coherence. The relation of phraseology and the communicative role of discourse intonation is also examined by Cheng and Warren (in press). Analyzing the one-million-word Hong Kong Corpus of Spoken English (HKCSE) (prosodic), Cheng and Warren describe the extent to which the lexical patterning in the form of word co-occurrences also reveal predictable patterns of discourse intonation (Brazil 1997). They find that for most of the word co-occurrences examined, the patterns of tone unit boundaries and prominence selection are closely related to the notion of phraseology, and so builds on the work of other studies in the field (e.g. Biber *et al.* 1998; Biber *et al*. 1999; Simpson and Swales 2001).

The notion of phraseology has presented exciting challenges for both researchers in applied English language studies and students and teachers of the English language. Recent textbooks on phraseology and collocation (e.g. McCarthy 2005; Sinclair 2003; Sinclair and Renouf 1991; Stubbs 2002) represent attempts at bridging research and pedagogy in the field of phraseology. This paper describes a set of instructional activities which involve university students at BA and MA levels in applying a new computer-mediated research methodology in the learning of phraseology in subjects of lexical studies, corpus linguistics, and discourse analysis The intended learning outcomes are that upon successful completion of the activities, students will be able to complete assignments that both show their awareness of patterns of phraseology and their ability in identifying and describing patterns of phraseology in the corpora that they examine. This paper, therefore, builds on the work of researchers and academics who have advocated the use of corpora and corpus linguistics in language learning in general (e.g. Aston 1997; Bernardini 2000, 2002; Braun 2005; Kennedy and Miceli 2002; Sinclair 2004b), and the use of concordancing in particular (e.g. Bernardini 2000, 2002; Cobb 1997; Gaskell and Cobb 2004; Johns 1991; Sinclair 2003; Stevens 1991).

2. ConcGram and concgramming

Cheng, *et al.* (2006) compare n-grams (in the form of bi-grams, tri-grams, and so on), 'skipgrams' (Wilks 2005), and 'phrase frames' (Fletcher 2006) in terms of their search functions and potential in revealing phraseological patterns in natural language. They go on and describe a search engine, ConcGram© (Greaves 2005), which is able to extract recurrent concgrams (i.e. sets of between 2 and 5 co-occurring words) fully automatically, within a wide span (up to 12

words on either side of the origin[1]), and which include all of a concgram's configurations irrespective of any constituent variation (e.g. AB and A*B) and positional variation (e.g. AB and BA) present. Cheng *et al*. (2006) argue that the identification of concgrams facilitates a fuller appreciation and understanding of Sinclair's (2004a) idiom principle, by revealing the co-selections made by the speakers and writers represented in a text or a corpus. Concgrams are, therefore, a useful starting point for quantifying the extent of phraseology in a text or a corpus, and thus determining the phraseological profile of the language contained within it. By 'phraseological profile', Cheng *et al.* (2006) mean the identification of the meaningful word associations in a text or a corpus, which is linked to what Phillips (1983, 1989) refers to as the 'aboutness' of a text. Phillips' notion of 'aboutness' is a product of the global patternings in the text, i.e. 'macrostructure'. Phillips argues that the macrostructure of texts should be determined by computational means in order to ensure that the results are derived from the texts itself and not from external features. The basic assumption of this position is that meanings in language are ultimately constructed by lexical items, or the associations of lexical items (Sinclair 1996, 2004a). This basic assumption also underpins 'concgramming' (Greaves and Warren 2007) and the activities described later in this paper. Three significant contributions that ConcGram, together with concgramming, can make to CALL are discussed by Greaves and Warren (2007). ConcGram serves as a tool for textual analysis, and can be used to help to raise students' awareness of the idiom principle, in that it helps students to find co-occurring words and 'chunks' (Sinclair and Mauranen 2006) in general, and as a result, enables students to master the discourses and genres of their specific disciplines (e.g. Bhatia 2004; Swales 2004).

3. Using concgramming in learning and teaching

This paper does not only describe the pedagogical procedures of the activities and the phraselogical profiles as a result of concgramming texts, but also offers a sample of output that is expected of the students, as a result of them following the classroom implementation procedures. More specifically, it describes a set of activities that aims to both raise students' awareness of the importance of phraseology in English language, and provide them with the necessary knowledge and computational and analytical skills to be able to conduct a study in identifying and describing patterns of phraseology in naturally occurring texts.

This growth in corpus linguistics has resulted in the development of new language learning and teaching methodologies (e.g. Burnard and McEnery 2000; Ghadessy *et al.* 2001; Hunston 1995, 2002; Granger 1998), particularly the use of corpora in the learning and teaching of English as a second language (e.g. Johns 1989, 1991; Thurstun and Candlin 1998; Flowerdew 1998; Hunston 2002; Ghadessy *et al.* 2001; Cheng *et al*. 2005). In the last ten years or so, a growing number of research studies have been reported in support of data-driven learning (DDL), showing how data from corpora can be used by students to further their language learning (Tribble 1999, 2000; Kettemann 1995; Johns 1997; Tribble and Jones 1990; Thurston and Candlin 1997; Wichmann *et al.* 1997). The originator of DDL is Tim Johns, formerly based at Birmingham University, UK, who believes that the language learner is at the same time a language researcher, and that in order to more effectively learn the target language, the learner needs to be able to have available authentic linguistic data (Johns 1991, 2002). He coins the term DDL to describe this approach to language learning. Using corpora as the source of spoken and written texts, DDL

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

brings to the class abundant examples of authentic language samples that can be studied and exploited in many ways. Such an approach usurps the traditional roles of the teacher, researcher and student because, as Johns (1991: 2) claims, 'research is too serious to be left to the researchers'. The teacher becomes a facilitator of language study instead of being seen as the language expert responsible for both teaching and research, and the students acquire a new role as language investigators in addition to that of language learners.

In the English Department of the university in Hong Kong in which the author works, corpus linguistics has been taught to both BA and MA students for six years. The intended learning outcomes are that after successfully completing this subject, students will be able to:

a. apply corpus methods to different types of corpora to investigate a wide range of linguistic features;
b. develop strategies to learn how language works, through carrying out corpus investigations, both individually and collaboratively;
c. report, in the form of an oral presentation and a written report, on a corpus-driven language study that they have conducted; and
d. develop the ability to critically reflect on their learning experience in the subject.

In the first lectures, students learn the basic concepts, theories and analytical techniques of corpus linguistics, including corpus building and concordancing. They are then introduced to the broad notion of phraseology, with examples to illustrate its forms and functions at the lexical-grammatical, discoursal and pragmatic levels. They are then shown the phraseologies found in different specialized corpora, e.g. the Hong Kong Financial Services Corpus and the Hong Kong Engineering Corpus (both of which are compiled by the author and colleagues), and how these phraseologies compare to those in general English, e.g. British National Corpus, Bank of English, the British English component of the International Corpus of English (ICE-GB), etc. Students will also be introduced to the Concgram© search engine, and trained how to use its 'concgramming' functions.

The methodology, 'concgramming', employed in the subject is described by Greaves and Warren (2007). It is a new computer-based methodology used to automatically identify the phraseological profile of any text or corpus, and hence the 'aboutness' of the text or corpus. Greaves and Warren (2007) outline the methodology and describe, with examples and classroom activities, its potential for use by language students in a data-driven learning mode, and discuss the wider implications of concgramming, and the concgrams so generated, with regard to CALL. In addition to DDL, there are major educational and learning theories that underpin the design and implementation of these activities, namely constructivist learning theory and situational cognition. Constructivist learning theory (Dewey 1916; Jonassen 1991) maintains that knowledge should be actively constructed by cognition. The teacher plays two major roles: first a facilitator and an adviser of instruction to help learners to create a knowledge construction environment, and second somebody to give guidance and support to help learners become actively involved in the learning process and construct their own knowledge. The theory of situational cognition states that learning should be applied to real-life situations and should emphasize students' involvement and understanding in the learning process (Bandura 1977; Lave and Wenger 1991).

A variety of activities and tasks are implemented to help students to achieve the learning outcomes. Students will be asked to build three small corpora for comparison purposes. The data are three political speeches. They are the three Policy Addresses given by the Chief Executive of

Hong Kong (the holder of this post is the head of the Government of the Hong Kong Special Administrative Region) in October 2006, October 2005 and October 2007 (Office of the Chief Executive, Hong Kong SAR Government), which are of inherent interest to students in Hong Kong. In the Policy Addresses, the Chief Executive outlines the political agenda of the government for the coming twelve months. These speeches are much anticipated in Hong Kong, and they are the subject of considerable speculation before they are given and much analysis by the media, political parties, and academics afterwards. It is, therefore, decided that these texts would be of interest to the students in Hong Kong to analyse, in terms of their respective phraseological profiles and, from these, their respective aboutness.

Another reason for using these texts is that they are long enough (the 2005 Policy Address is 12,811 words, the 2006 Policy Address is 8,251 words, and the 2007 Policy Address is 14,132 words) to produce sufficient instances of patterning. Students will be able to generate up to three-word concgrams in less than one hour on a regular desktop computer, and even faster if an exclusion list (Cheng *et al.* 2006) is used. Students can also concgram the three texts outside of class so that in class they can concentrate on working in small groups analysing and discussing their findings. As observed by Greaves and Warren (2007), lists of two-word and three-word concgrams are usually sufficient to yield an initial phraseological profile of a text, and the amount of data for the students to analyse in one to two hours is also manageable. The following, which builds on Greaves and Warren (2007), outline the steps and specifics of the activities for students to determine the aboutness of different texts.

1. To compile a list of the ten most frequent words in each of the three Policy Addresses, and combine inflected forms, where appropriate.
2. To compile a list of the twenty most frequent phrases or word co-occurrences in each of the three Policy Addresses, and combine inflected forms where appropriate.
3. To monitor and record the frequencies with which the most frequent words, phrases and word co-occurrences found in the 2005 Policy Address occur in the 2006 Policy Address, and vice versa. The step is repeated so that comparisons are made of all of the three texts.
4. In groups, to discuss the findings derived from the three Policy Addresses.
5. Throughout the analysis of the three Policy Addresses, to make note that the word lengths of the three Policy Addresses are different, and so normalized frequencies need to be worked out.

Regarding inflected forms of words, studies (e.g. Mindt 1991; Tognini-Bonelli 2001) have shown that inflected forms tend to be associated with different meanings and functions, and so careful analysis of the concordance lines needs to be carried out. This can generate a lot of discussion and promote language awareness, and this will be explicitly explained to the students.

Below are the lists of the ten most frequent lexical words in the three Policy Addresses, including inflected forms when considered appropriate.

Top 10 most frequent words

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

Figure 1. The 2005 Policy Address of the Hong Kong SAR Government (Total words spoken: 12,811)

Top 10 most frequent words

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

Figure 2. The 2006 Policy Address of the Hong Kong SAR Government (Total words spoken: 8,251, i.e. 36% shorter than the 2005 Policy Address)

Top 10 most frequent words

Ranking	Lexical word	Frequency	(Frequency in 2006)
1	Hong Kong	149	(68)
2	development/develop[2]	140	(76)
3	government	104	(71)
4	year(s)[3]	73	(47)
5	people	55	(17)
6	promote	54	(26)
7	new	52	(25)
8	public	49	(33)

[2] The most frequent words are arrived at by combining all the forms of a word where appropriate, e.g. *development* and *develop*.

[3] The most frequent words are arrived at by combining plural and singular forms where appropriate, e.g. *years* and *year*.

9	services	48	(31)
10	community(ies)	47	(38)

Figure 3. The 2006 Policy Address of the Hong Kong SAR Government
(Total words spoken: 14,132, i.e. 71% longer than the 2006 Policy Address)

In groups, students will discuss and compare the three lists of single lexical words in terms of relative frequencies and rankings, with a view to describing the aboutness of the three policy addresses, which reflect the varying focuses and priorities pertaining to the Hong Kong SAR Government in the three consecutive years, 2005-2007.

Students will also generate lists of two-word phrases or word co-occurrences. As described in Cheng *et al.* (2006), the computer program ConcGram© developed by Greaves (2005) is designed with the goal of identifying all the potential configurations of between 2 and 5 words in any corpus, based on a window of any size, to include the co-occurring words, even if they occur in different positions relative to one another (i.e. positional variation) and even when one or more words occur in between the associated words (i.e. constituency variation). Most important of all, this search engine can conduct fully automated searches throughout the data with no prior nomination of any parameters from the researcher; in other words, it will nominate the groupings itself.

Below are the lists of the twenty most frequent phrases or word co-occurrences in the three Policy Addresses, including inflected forms when considered appropriate. A forward slash indicates that there is variation, either constituency, positional or both, in the phrase or word co-occurrence.

20 most frequent phrases or co-occurring words

Ranking	Phrase or word co-occurrence	Frequency	(Frequency in 2006)
1	Chief Executive	18	(8)
2	the SAR Government	17	(14)
2	Legislative Council	17	(8)
3	the Central Authorities	15	(2)
4	Hong Kong/people	14	(2)
5	social harmony/harmonious society	13	(4)
6	Hong Kong/development/develop	11	(10)
6	food safety	11	(0)
7	the Basic Law	10	(2)
8	community/support	9	(8)
8	powers and functions	9	(0)
8	the/government/continue	9	(0)
9	economic/economy/development	8	(7)
9	government/support	8	(4)
9	air quality	8	(8)
10	world city	7	(1)
10	one country two systems	7	(0)
10	principal officials	7	(0)
10	emissions reduction	7	(3)
11	Commission on Strategic Development	6	(1)

Figure 4. The 2005 Policy Address of the Hong Kong SAR Government
(Total words spoken: 12,811)

20 most frequent phrases or co-occurring words

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

Figure 5. The 2006 Policy Address of the Hong Kong SAR Government (Total words spoken: 8,251, i.e. 36% shorter than the 2005 Policy Address)

Top 20 most frequent phrases or co-occurring words

Ranking	Phrase or word co-occurrence	Frequency	(Frequency in 06/07)
1	health care	23	(0)
2	next/year	20	(2)
3	our country	20	(6)
4	economic development	13	(7)
4	Hong Kong/development[4]	13	(10)
4	Hong Kong financial	13	(0)
4	promote/development	13	(3)
8	HKSAR/government	12	(0)
9	heritage conservation	10	(0)
9	historic/buildings	10	(0)
9	Hong Kong people	10	(1)
9	infrastructure projects	10	(0)
9	our people	10	(0)
9	Basic Law	10	(2)
9	environment/protect	10	(6)
16	development/country/our	9	(3)
16	social enterprises	9	(0)
16	young people	9	(0)

[4] The associated words 'Hong Kong/development' include instances such as *Hong Kong's development* and *development of Hong Kong*.

19	Hong Kong/Mainland	8	(2)
20	Financial Secretary	7	(1)

Figure 6. The 2007 Policy Address of the Hong Kong SAR Government (Total words spoken: 14,132, i.e. 71% longer than the 2006 Policy Address)

Once the students have compiled the frequency lists of 2-word co-occurrences, they will be encouraged to compare the contents of the lists and compare the frequencies of phrases or word co-occurrences in one text against the number of instances, if any, in the other text. Below are the kinds and levels of analysis that are expected of the students for them to demonstrate the expected outcome of the comparative study, based on phrase or word co-occurrence frequency lists. For example, the main points of the Chief Executive's Policy Address 2007, as compared to Policy Address 2006, include:

1. There are new phrases such as 'health care' (23 instances), 'heritage conservation' (10), 'historic buildings' (10), infrastructure projects' (10), 'social enterprises' (10), 'young people' (10). None of these are found in the 2006 Policy Address.
2. There is also a preponderance of certain phrases compared to the 2006 Policy Address, for example, 'next/year' (20 versus 2), 'our country' (20 versus 6), 'promote/development' (13 versus 3), 'Hong Kong people' (10 versus 1), 'Basic Law' (10 versus 2), 'development/country/our' (9 versus 3), 'Hong Kong/Mainland' (8 versus 2), and 'Financial Secretary' (7 versus 1).
3. Some of the phrases in the 2006 Policy Address 2006 have completely disappeared in the 2007 Policy Address, namely 'development of/political system' (6), 'strong governance' (6), 'foster/family' (5), and 'provide/parents' (5).
4. Phrases frequent in the 2006 Policy Address drop in frequency in the 2007 Policy Address, e.g. 'air quality' (8 versus 5), 'Chief Executive' (8 versus 4), 'family members' (8 versus 3), 'Legislative Council' (8 versus 3), 'last year' (7 versus 2), 'mutual support' (6 versus 1), and 'film industry' (5 versus 1).
5. On a more technical note, 'SAR Government' (14) in the 2006 Policy Address is replaced by the more precise 'HKSAR Government' (12) in the 2007 Policy Address.
6. In terms of single word frequencies, seven of the top ten remain the same. Three words drop out of the top ten: 'support', 'family(ies)', and 'provide/provision', and these are replaced by 'people', 'promote', and 'new'.
7. The phrases or word co-occurrences used by the Chief Executive are a better indicator of his Policy Address contents from year to year than single word frequencies.

4. Conclusions and implications

This paper has described some pedagogical implications of the new computer-mediated methodology, concgramming, which aim to facilitate the introduction of phraseology to language students at both undergraduate and postgraduate levels. Concgramming is a new way of identifying and categorising word co-occurrences. The concgrams of a corpus are preferably identified and generated without prior input from the user, other than to set the size of the span, as it is only a fully automated concgram search that can reveal all of the possible collocational patterns that exist in a corpus. Studying concgram search results, as in the case of those

generated from the three Policy Addresses of the Hong Kong SAR Government, can reveal the phraseological profile of the texts in a way that other searches do not. In the case of the latter, attention is primarily drawn to the user-nominated node word, a popular and traditional starting point for corpus queries which is replaced by the notion of 'origin' in concgram searches where the focus of attention is on word co-occurrences and their constituency and positional variations. As discussed in Cheng *et al.* (2006), concgram searches begin with an origin (single, double, triple or quadruple) and have the central aim of uncovering the phraseological patterns in the language.

This paper argues that phraseology, or the recurrent ways of expressing ideas, processes and propositions, is useful for understanding the meanings of texts. Language students need to be aware of and understand both the extent and the importance of phraseology in the English language, which is a major area of English language study that is currently given insufficient attention. The basic principles of the learning and teaching methodology described in this paper are aimed to ensure that the learning process is both interactive and collaborative in nature, from the perspectives of constructivist learning theory and situational cognition, and is derived from DDL (Johns 1991; Cheng *et al.* 2003) which casts the language learner in the role of language researcher (Johns 1991).

The concgramming learning and teaching activities described that highlight the main elements in the understanding and production of phraseology in English can be replicated in ESP and LSP subjects. Take the author's university as an example, specialised texts collected from such major academic disciplines as engineering, land surveying, business, financial studies, design, tourism and hospitality management, health sciences, and so on, could be concgrammed in order to determine the discipline-specific phraseological profiles. The distinctive usages in discipline-specific texts, for instance, business, will bear meanings that are specific to the field and different from general English usage. The lexical profiles identified contribute directly to the aboutness of the texts. It is important for both the writers and readers of business English to have a mastery of these lexical profiles. Those who fail to communicate using the conventional keywords and phraseology of business English might be misunderstood and, as readers, might misunderstand the subtle shifts in meanings that result in particular choices (Gledhill 2000a, 2000b; Sinclair 2004a; Kemppanen 2004). The phraseological profile established will hence enable the content, scope and argument of texts to be determined, and enable the extent to which texts deviate from the expected aboutness to be established.

Finally, in addition to enhancing teachers' and students' critical awareness of the nature and role of phraseology in the English language, the activities also enhance students' critical and creative thinking through the understanding, analysis, comparison and application of phraseology that is specific to individual text types or genre types.

Acknowledgements

The work described in this paper was substantially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region (Project No. PolyU 5480/06H).

References

Aston, G. 1997. "Small and large corpora in language learning". B. Lewandowska-Tomaszczyk and J. P. Melia, eds. *Practical Applications in Language Corpora*, Łodz: Łodz University Press. 51-62.

Bandura, A. 1977. *Social Learning Theory*. Englewood Cliffs (NJ): Prentice Hall.

Bernardini, S. 2000. "Systematising serendipity: Proposals for concordancing large corpora with language students". L. Burnard and T. McEnery, eds. *Rethinking Language Pedagogy from a Corpus Perspective*. Frankfurt: Peter Lang. 225-234.

Bernardini, S. 2002. "Exploring new directions for discovery learning". B. Kettemann and G. Marko, Eds. *Teaching and Learning by Doing Corpus Analysis*. New York: The Edwin Mellen Press. 165-182.

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

Biber, D., S. Conrad and R. Reppen. 1998. *Corpus Linguistics: Investigating Language Structure and Use.* Cambridge (UK): Cambridge University Press.

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

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

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

Brazil, D. 1997. *The Communicative Role of Intonation in English*. Cambridge: Cambridge University Press.

Burnard, L. and T. McEnery, eds. 2000. *Rethinking Language Pedagogy from a Corpus Perspective*. Frankfurt: Peter Lang.

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

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

Cheng, W. 2006. "Describing the extended meanings of lexical cohesion in a corpus of SARS spoken discourse". J. Flowerdew and M. Mahlberg, eds. *Special Issue of International Journal of Corpus Linguistics: Corpus Linguistics and Lexical Cohesion*, 11(3): 325-344.

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

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

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

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

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

Cowie, A.P. 1998. *Phraseology: Theory, Analysis, and Applications*. Oxford: Clarendon Press.

Dewey, J. 1916. *Democracy and Education*. New York: Free Press.
Fletcher, W. H. 2006. "Phrases in English" Home. [http://pie.usna.edu]
Flowerdew, L. 1998. "CALL materials derived from integrating 'expert' and 'interlanguage' corpora findings on causality: discoveries from teachers and students". *English for Specific Purposes* 17: 329-346.
Gaskell, D. and T. Cobb. 2004. "Can students use concordance feedback for writing errors?" *System* 32 (3): 301-319.
Ghadessy, M., A. Henry and R.L. Roseberry, eds. 2001. *Small Corpus Studies and ELT:Theory and Practice*. Amsterdam/Philadelphia: John Benjamins.
Gledhill C. 2000a. "The discourse function of collocation in research article introductions".*Englishfor Specific Purposes* 19(2): 115-135.
Gledhill C. 2000b. "Collocations in science writing. *Language in Performance Series* 22. Tübingen: Gunter Narr Verlag.
Granger, S. ed. 1998. *Learner English on Computer.* London/New York: Longman.
Greaves, C. 2005. "Introduction to ConcGram©". *Tuscan Word Centre International Workshop*. Certosa di Pontignano, Tuscany, Italy, 25-29 June 2005.
Greaves, C. and M. Warren. 2007. "Concgramming: A computer-driven approach to learning the phraseology of English". *ReCALL Journal* 17(3): 287-306.
Halliday, M.A.K and R. Hasan. 1976. *Cohesion in English*. London/New York: Longman.
Halliday, M.A.K., W. Teubert and C. Yallop. 2002. *Perspectives in Lexicology and Corpus Linguistics.* London: Continuum.
Hunston, S. 1995. "A corpus study of some English verbs of attribution". *Functions of Language* 2(2): 133-158.
Hunston, S. 2002. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.
Hunston, S. and G. Francis. 2000. *Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English*. Amsterdam: John Benjamins.
Johns, T. 1989. "Whence and whither classroom concordancing?" T. Bongaerts, P. de Haan, S. Lobbe and H. Wekker, eds. *Computer Applications in Language Learning.*Dordrecht: Foris. 9-33.
Johns, T, 1991. "Should you be persuaded: Two samples of data-driven learning materials". T. Johns and P. King eds. *Classroom Concordancing.* English Language Research: Birmingham University. 1-16.
Johns, T. 2002. "Data-driven learning: The perpetual challenge". B. Kettemann and G. Marko, eds. *Teaching and Learning by Doing Corpus Analysis. Proceedings of the Fourth International Conference on Teaching and Language Corpora, Graz 19-24 July 2000*. Amsterdam: Rodopi. 107-117.
Jonassen, D.H. 1991. "Evaluating constructivist learning". *Educational Technology* 31 (9): 28-33.
Kemppanen, H. 2004. "Keywords and ideology in translated History texts: A corpus-based analysis". *Across Languages and Cultures* 5(1): 89-106.
Kennedy, C. and T. Miceli. 2002. "The *CWIC* project: Developing and using a corpus for intermediate Italian students". B. Kettemann and G. Marko, eds. *Teaching and Learning by Doing Corpus Analysis*. New York: The Edwin Mellen Press. 183-192.
Kettemann, B. 1995. "On the use of concordancing in ELT". *TELL & CALL* 4: 4-15.
Lave, J. and E. Wenger. 1991. *Situated Learning. Legitimate Peripheral Participation*. Cambridge: University of Cambridge Press.

Louw, B. 1993. "Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies". M. Baker, G. Francis and E. Tognini- Bonelli, eds. *Text and Technology: In Honour of John Sinclair*. 157-176. Amsterdam/Philadelphia: John Benjamins.

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

Mindt, D. 1991. "Syntactic evidence for semantic distinctions in English". K. Aijmer and B. Altenberg, eds. *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. London/New York: Longman. 183-196

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

Phillips, M. 1989. *Lexical Structure of Text. Discourse Analysis Monographs: 12*. English Language Research: University of Birmingham.

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

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

Policy Address 2007/2008. [htttp://www.policyaddress.gov.hk/07-08/eng/pdf/speech.pdf]

Scott, M. 1997. "PC Analysis of key words - and key key words". *System* 25(1): 1-13.

Scott, M. 2000. "Focusing on the text and its key words". L. Burnard and T. McEnery, eds. *Rethinking Language Pedagogy from a Corpus Perspective* 2: 103-122. Frankfurt: Peter Lang.

Scott, M. 2001. "Comparing corpora and identifying key words, collocations, and frequencydistributions through the WordSmith Tools suite of computer programs". M. Ghadessy, A. Henry and R.L. Roseberry, eds. *Small Corpus Studies and ELT: Theory and Practice.* Amsterdam: Benjamins. 47- 67.

Simpson, R. and J. Swales, eds. 2001. *Corpus Linguistics in North America*. Ann Arbor (MI):University of Michigan Press.

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

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

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

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

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

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

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

Sinclair, J. McH. and A. Mauranen. 2006. *Linear Unit Grammar*. Amsterdam: John Benjamins.

Sinclair, J.M. and A. Renouf. 1991. "Collocational frameworks in English". J. McH. Sinclair, ed. *Lexis and Lexicography*. National University of Singapore: Unipress. 55-71.

Stevens, V. 1991. "Concordance-based vocabulary exercises: A viable alternative to gap- fillers". T. Johns and P. King, eds. *Classroom Concordancing: English Language Research Journal* 4: 47-63. Centre for English Language Studies: University of Birmingham.

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

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

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

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

Thurstun, J. and C. Candlin. 1998. "Concordancing and the teaching of the vocabulary of academic English". *English for Specific Purposes* 17: 267-280.

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

Tognini Bonelli, E. 2002. "Between phraseology and terminology in the language of Economics". S. Nuccorini, ed. *Phrases and Phraseology - Data and Descriptions*. Bern, Switzerland: Peter Lang. 65-83.

Tribble, C. 1997. "Improvising corpora for ELT: Quick-and-dirty ways of developing corpora for language teaching". B. Lewandowska-Tomaszczyk and P. J. Melia, eds. *Practical Applications in Language Corpora*. Lodz, Poland: Lodz University Press. 106-117.

Tribble, C. 1999. "Genres, keywords, teaching: Towards a pedagogic account of the language of project proposals". L. Burnard and A. McEnery, eds. *Rethinking Language Pedagogy from a Corpus Perspective: Papers from the Third International Conference on Teaching and Language Corpora*, (Lodz Studies in Language). Hamburg: Peter Lang.

Tribble, C. 2000. "Genres, keywords, teaching: towards a pedagogic account of the language of project proposals". Lou Burnard and Tony McEnery, eds. *Rethinking Language Pedagogy from a Corpus*. Frankfurt am Main: Peter Lang. 75-90.

Tribble C. and Jones, W. 1990. *Concordances in the Classroom: A Resource Book for Teachers*. Harlow: Longman.

Wichmann, A., S. Fligelstone, G. Knowles and A. McEnery, eds. 1997. *Teaching and Language Corpora.* London/New York: Longman.

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