

國立政治大學英國語文學系碩士班碩士論文

指導教授：張郇慧 博士

Advisor: Dr. Claire H. H. Chang

普通單字在特殊文類的呈現：

針對商業年報語言詞塊之語料庫研究

**General Lexis in Specialized Genre:
A Corpus Study on Formulaic Language in Business Reports**



研究生：陳俊宏 撰

Name : Chen, Chun-hung

中華民國 102 年 1 月

January 2013

General Lexis in Specialized Genre:
A Corpus Study on Formulaic Language in Business Reports

A Master Thesis
Presented to
Department of English,

National Chengchi University



In Partial Fulfillment
of the Requirements for the Degree of
Master of Arts

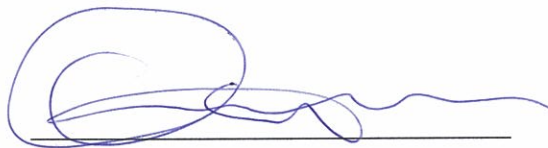
by
Chen, Chun-hung
January 2013

The members of the Committee approve the thesis of Chun-hung Chen
defended on January 10th, 2013.



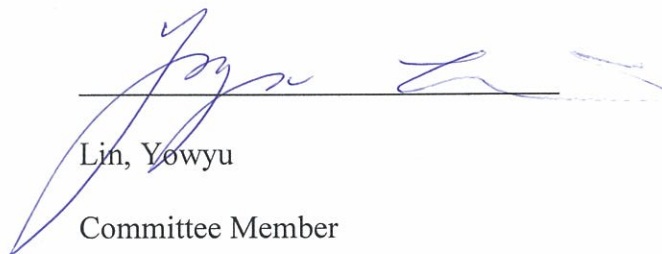
Chang, Claire H. H.

Professor Directing Thesis



Her, One-Soon

Committee Member



Lin, Yowyu

Committee Member

Approved:



Chih-hsin Lin, Chair, Department of English

Acknowledgement

This thesis is in fact a composition by many intellectual minds. This study was first motivated by the late Professor Cynthia Hsin-feng Wu [吳信鳳教授], who had suggested me to invest my study into the ESP field at the entrance interview of TESOL Program of Department of English of Chengchi University. My competence of knowledge in TESOL and ESP was further broadened and deepened by Professor Leah C.Y. Yeh [葉潔宇教授] and Professor Yi-Ping Huang [黃怡萍教授]; my performance in academic writing was enhanced by Professor Ming-chung Yu [余明忠教授] and Professor Judy H. Y. Yu [尤雪瑛教授] with basics of thesis writing and advanced course on discourse organization, respectively. Many thanks to Professor Siaw-Fong Chung [鍾曉芳教授], who has filled me in with essence of corpus linguistics and statistics and equipped me with studied attitude and administrative skills on doing research. This piece of academic work is accomplished with instruction of Professor Claire H. H. Chang's [張郇慧教授], who provides me with the largest degree of freedom to have my enthusiasm and creativity fulfilled. Professor Chang always gives me advice on the level of logic of argument, without which this thesis would fail in maintaining coherence and consistency.

Thanks to professors who had commented on my thesis at proposal and final oral test. Comments from Professor Cheung, Hintat [張顯達教授], Professor Yi-Ping Huang [黃怡萍教授], Professor Her, One-Soon [何萬順教授], and Professor Lin, Yowyu [林祐瑜教授] really help optimize this study from aspects of logic of argument, framing of research questions and hypotheses, implications on TESOL, and other technical issues.

Thanks to all faculty members of the Department of English, who have built a warm environment to make academic excellence realizable. And I would like to express my sincere gratitude to my classmates: Jonathan Wang, Rachel Tseng, Emily Hung, and Mandy Huang; they have sharpened my thoughts with their keen observation and insightful opinions shared in class.

And lastly, I would like to dedicate this work to Lord Jesus Christ and my dear wife, who found my life with miracles and unfailing love.



Table of Contents

CHAPTER 1 INTRODUCTION	1
CHAPTER 2 LITERATURE REVIEW	5
2.1 COMMON WORDS IN ENGLISH FOR SPECIFIC PURPOSES.....	5
2.1.1 <i>Common core hypothesis</i>	6
2.1.2 <i>Specifiable words in specific genres</i>	8
2.1.3 <i>Lexical patterns in different contexts</i>	9
2.1.4 <i>Tentative research question</i>	12
2.2 EXTRACTION OF WORDS.....	12
2.2.1 <i>Common words</i>	13
2.2.2 <i>Tentative research question</i>	14
2.2.3 <i>Specialized words</i>	15
2.2.4 <i>Summary on methods for extracting words</i>	18
2.2.5 <i>Operational definition of common words</i>	19
2.3 IDENTIFICATION OF FORMULAIC LANGUAGE.....	20
2.3.1 <i>Collocation</i>	21
2.3.2 <i>Lexical bundles (N-Grams)</i>	24
2.3.3 <i>Cumulative frequency</i>	26
2.4 INTEGRATING CORPUS RESEARCH WITH GENRE-BASED METHOD.....	28
2.5 SUMMARY OF CHAPTER AND RESEARCH QUESTIONS.....	31
CHAPTER 3 RESEARCH DESIGN	35
3.1 CREATING BUSINESS REPORTS CORPUS.....	35
3.1.1 <i>Source of data</i>	35
3.1.2 <i>Corpus size</i>	36
3.1.3 <i>Representativeness and balance</i>	37
3.1.4 <i>Dealing with technical compounds</i>	37
3.1.5 <i>Procedure for creating BRC</i>	39

3.2 EXTRACTING COMMON WORDS.....	46
3.2.1 <i>Excluding technical words and compounds</i>	46
3.2.2 <i>Keyword, words with unusually frequency</i>	46
3.2.3 <i>Matching words in Senior High English Wordlist for Reference</i>	47
3.2.4 <i>Key-keyword, keywords sorted according to text coverage</i>	48
3.2.5 <i>Procedure for extracting common words with WordSmith Tools</i>	48
3.3 LOCATING FORMULAIC LANGUAGE FROM DIFFERENT GENRES.....	57
3.3.1 <i>Selecting genres for comparing formulaic language</i>	57
3.3.2 <i>Syntagmatic variation and paradigmatic variation of formulaic language</i>	59
3.3.3 <i>Procedure for identifying formulaic language with AntConc</i>	59
3.4 FLOWCHART OF RESEARCH DESIGN.....	66
CHAPTER 4 RESULTS AND DISCUSSION.....	67
4.1 EXTRACTED COMMON WORDS.....	67
4.2 LENGTH OF FORMULAIC LANGUAGE ACROSS GENRES.....	74
4.3 COMPOSITION OF FORMULAIC LANGUAGE ACROSS GENRES.....	77
4.4 SUMMARY OF CHAPTER.....	89
CHAPTER 5 CONCLUSION.....	91
5.1 OVERVIEW OF RESEARCH QUESTIONS AND RESEARCH RESULTS.....	91
5.2 SIGNIFICANCE AND IMPLICATION.....	93
5.3 RESEARCH IN PROSPECT.....	94
REFERENCES.....	97
APPENDICES.....	103

List of Figures

Figure 3.1.1 Downloading 20-F documents of queried company on SEC website.....	40
Figure 3.1.2 “Error” codes in html 20-F document	42
Figure 3.1.3 Querying for special characters with original codes in Dreamweaver.....	43
Figure 3.2.1 Adjusting setting for html brackets and special characters	49
Figure 3.2.2 Uploading accounting compounds as stop list	51
Figure 3.2.3 Adjusting setting for computing keywords	52
Figure 3.2.4 Key-keywords assorted based on degree of text coverage	53
Figure 3.2.5 Adjusting setting for Match List with the Senior High English Wordlist for Reference	54
Figure 3.2.6 Immediate results of the matching process	55
Figure 3.2.7 Final results of the matching process	56
Figure 3.3.1 Querying for collocates of applicable in Business Reports Corpus with AntConc	61
Figure 3.3.2 Querying applicable together with a context word the.....	62
Figure 3.3.3 Results of querying for collocates of applicable with its context word the	63
Figure 3.3.4 Final results of finding collocates of applicable with the method of cumulative frequency.....	64
Figure 3.3.5 Concordance lines of applicable together with its extracted collocates	65
Figure 3.4.1 Flowchart of research design.....	66
Figure 4.1.1 Results of extracted common words assorted based on text coverage.....	69

List of Tables

Table 3.1.1 Basic information of the Business Reports Corpus and Brown Corpus	45
Table 3.3.1 Basic information of Business Reports Corpus and the two subdivisions of Brown Corpus for investigation	58
Table 4.1.1 Distribution of extracted key-keywords in BRC	70
Table 4.1.2 Distribution of extracted key-keywords in SHEWR [高中英文參考詞彙 表]	71
Table 4.2.1 Length of formulaic sequences composed by <i>annual</i>	74
Table 4.2.2 Length of formulaic sequences composed by <i>applicable</i>	75
Table 4.2.3 Length of formulaic sequences composed by <i>financial</i>	75
Table 4.2.4 Length of formulaic sequences composed by <i>significant</i>	75
Table 4.3.1 Composition of formulaic language composed by <i>annual</i>	77
Table 4.3.2 Composition of formulaic language composed by <i>applicable</i>	82
Table 4.3.3 Composition of formulaic language composed by <i>financial</i>	84
Table 4.3.4 Composition of formulaic language composed by <i>significant</i>	86





摘要

對專業英語教學而言，如何擇取語言教學內容，是一項相當重要的工作。學界一般建議專業英語教學教師，以與專業科目教師協同合作的方式，來共同協商出語言教學內容。但此協同方式的運作，往往受限於教學資源及教師之間對何謂知識的認知差距，常常窒礙難行。本研究即著力於探討語料庫方法運用在擇取專業英語教學內容的實際操作，期盼能提供一套有效的解決方案。

本研究將研究內容鎖定在普通單字及普通單字所組成的語言詞塊，並特別將研究範圍著重於商業文類。為了使本研究順利進行，作者應用語料庫相關知識，建立了一個商業年報語料庫。在擷取普通單字及語言詞塊時，也嚴謹運用了相關的語料庫技術，如單字顯著程度、單字涵蓋範圍、及累積詞頻法等等。此研究找出了數量適中的普通單字，能利於在授課時間有限的專業英語教學，進行課程規劃。本研究也證實了普通單字，會因為所處的文類不同，而組成不同的語言詞塊。

本論文的研究結果及研究流程，展示了以語料庫方法擷取專業英語教學的語學內容，效率良好，結果明確。同時本研究的研究結果，也將語言的慣性組成模式，顯明出來。最後，本研究的研究結果所顯示的普通單字，乃是專業英文與一般英文共享重疊的單字，可視之為一般英語教學銜接至專業英語教學的重要教學內容。由於研究過程採用了高中英語參考詞彙表，本研究之結果特別適用於台灣的英語教學情境，尤其對大學商學院之英語教學，更是貼切實用。

關鍵字：專業英語教學、單字表、高中英文參考詞彙表、商業英文、語料庫、文類分析、詞塊



Abstract

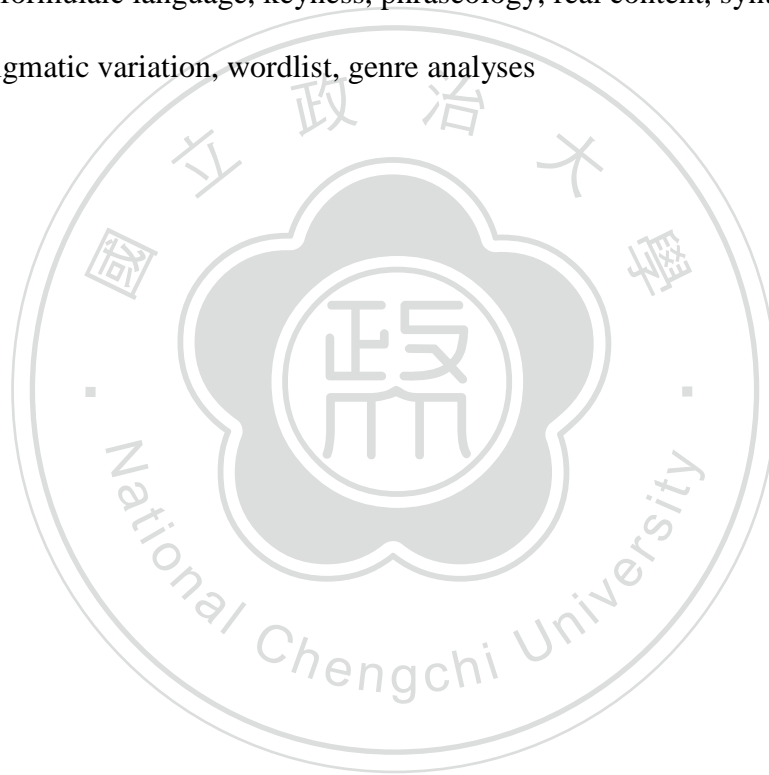
This study aims to apply corpus method on filtering out language content for ESP course. Following Dudley-Evans and St John's (1998) idea of real content, this thesis focuses on locating general lexis from different genres.

For this study, general lexis is viewed as common words and formulaic language composed with common words. In this project, common words are extracted from a homemade Business Reports Corpus (BRC) with reference to Senior High English Wordlist for Reference (SHEWR) [高中英文參考詞彙表] and Chinese-English Translation of Important Accounting Terms [重要會計用語中英對照] in Taiwan, also with notions of corpus linguistics including keyness and text coverage. Then with number of extracted common words being compared to the total amount of SHEWR, practicality of SHEWR applied in business genre is measured and a subset word bank of SHEWR is identified to link instruction of English of General Purpose to that of English of Specific Purposes in the context of Taiwan.

Secondly, formulaic language is regarded as multi-word units composed with previously extracted common words. Instead of Mutual Information, t-score, or lexical bundles, Danielsson's (2007) method of cumulative frequency is adopted to locate formulaic language. Cross-genre analyses on formulaic language identified from BRC, Subdivision A of Brown Corpus (texts of reportage) and Subdivision H of Brown Corpus (official documents) are conducted to verify whether composition of formulaic language is correlated with types of genre, which may affect description of phraseology as well as practice of the English teaching method Lexical Approach (Lewis, 2000).

This thesis has successfully demonstrated that the corpus method is a handy and efficient approach for ESP practitioners to sort out language content for instruction. One by-product of this study is that the difficulty level (as designated in SHEWR) of extracted common words is attached to allow further investigation together with data of text coverage and overall frequency.

Keywords: Corpus, English for Specific Purposes (ESP), English for Business Purpose (EBP), formulaic language, keyness, phraseology, real content, syntagmatic variation, paradigmatic variation, wordlist, genre analyses



Chapter 1 Introduction

The teaching of English for Specific Purposes (ESP) has been generally regarded as an identifiable activity in English Language Teaching (ELT). The distinctness of ESP could be attributed to the consensus that ESP teachers are required to bring about practical outcomes for learners with specific professional backgrounds in limited course time (Dudley-Evans & St John, 1998, p. 1). Most ESP teachers, assumedly equipped with training in teaching General Purpose English (GPE), are thus positioned in difficult situations, where they do not possess the subject knowledge in fields which their students specialize in (Wu & Badger, 2009). This knowledge gap not only intrinsically changes the status of ESP practitioners from “primary knower” to “language consultant” in the perspective of language teaching methodology, but also implies that practitioners need assistance in designing courses under such circumstance (Dudley-Evans & St John, 1998, pp. 13-16). Arranging language content, a central process in course design and material management, is of primary concern for ESP practitioners (Hyland, 2002, pp. 393-394).

To deal with actual content with limited subject knowledge, we can consult Dudley-Evans and St John’s (1998) classification of carrier content and real content (p. 11). *Carrier content* refers to language items that carry context related to a specific subject; for example, the word *germinate* in a course on life cycle of a plant. In comparison, *real content* denotes language that performs certain communicative function in the context. For example, the bolded preposition *as* and morphemes *s* in the sentence *As the plant germinates, the seed swells* are language of describing process, which is the real focus of this biology course. It is also noted that real content does not entail technical knowledge as carrier content does (Dudley-Evans & St John, 1998, p. 11).

Furthermore, we can observe that the real content comes as multiple lexical units formed with relatively general lexeme¹ like the preposition *as* and the morphemes *s*. For ESP practitioners, real content is the real focus for instruction (Dudley-Evans & St John, 1998, pp. 11, 16) and it is an issue of practicality as for how to sort out real content from ESP materials with efficiency. To resolve this issue, collaboration² between ESP practitioners and subject teachers has been proposed as an ideal and feasible solution for content management as well as course development (Dudley-Evans & St John, 1998, pp. 15-16).

Although studies on collaborative teaching report successful outcomes (Dudley-Evans, 2001), some researchers have expressed qualified agreement on the efficiency of collaboration. Belcher (2006) has indicated that modest resources provided by institutions and reluctance of teachers to cooperate with each other can obstruct collaboration (p. 140). The reluctance, according to Barron (2003), can be attributed to the inherent and incompatible diversity of collaborators' respective

¹ According to Crystal's (2003) definition, lexemes refer to minimal distinctive units for discerning meanings. Either occurring as grammatical variants such as *-s*, *-ing*, and *-ed* in *walks*, *walking*, and *walked*, or appearing as multiple lexical units to express complete meanings, like *kick the bucket* (= "die"), lexemes differ from words in that the term *word* treats language items based on orthography, not semantic unit (pp. 265-266).

² Collaboration, when viewed in the aspect of the degree of relative involvement between language teachers and subject teachers, is further categorized into co-operation, collaboration, and team-teaching (Dudley-Evans, 2001, pp. 226-228; Dudley-Evans & St John, 1998, pp. 42-47). The process of collaboration, according to Barron (2003), is suggested to be administered under the methodology of constructivism. Classification and methodology of collaboration, though a vital issue in the ESP field, is not the major concern for the present thesis.

philosophy of knowledge from different fields (p. 297). Motivated by these difficulties in collaborative teaching between ESP practitioners and subject teachers, this thesis considers the feasibility of another solution that can manage ESP content with minimum human interference. Corpus research, famed for its objective and automatic approach to text analysis, is thus put under contemplation to locate real content from ESP texts. The present thesis aims to prove that application of corpus method, in addition to collaboration, is an effective and practical approach to identify content that doesn't involve subject knowledge for ESP practitioners.

Having observed that real content involves multiple lexical units composed by general lexis³, the next chapter will narrow down the scope of research focus to the level of word to review how words, specifically common ones⁴, are conceptualized in the ESP field and from the perspective of corpus linguistics. Based on understanding of nature of words, the rest of Chapter 2 continues by introducing corpus techniques to extract words of interest as well as multiple lexical units. Chapter 2 ends with proposing research questions which are framed for achieving the objective of this project. Chapter 3 gives report on methodology along with its induced research design that encompasses the process of creating a corpus for exploration, extracting common words of interest, and identifying multiple lexical units. Chapter 4 demonstrates results on both extracted words and identified multiple lexical units, which in turn signify the efficacy of corpus approach to locate real content for ESP instruction. Finally, Chapter 5 concludes the

³ The term lexis is used as an umbrella term to refer to the vocabulary of a language (Crystal, 2003, p. 268).

In the present thesis, lexis includes morphemes, words, and multiple lexical units. Lexis differs from lexeme in that the former does not consider the semantic aspect but the latter does.

⁴ *Common words* and *general words* are used interchangeably in this thesis.

present study with implication of research results on ESP pedagogy; other issues for prospect research will also be addressed.



Chapter 2 Literature Review

In Introduction, it is mentioned that real content, the real focus of ESP courses, can be regarded as multiple lexical units composed with general lexemes that do not possess technical meanings of a specific subject. To make identifying real content a feasible task, this chapter discusses the issue of general lexis in ESP and the phenomenon of multiple lexical units by reference to related researches in the ESP field and those in corpus linguistics, respectively.

2.1 Common words in English for Specific Purposes

One principal issue in ESP is to account the specificity of any given types of ESP⁵ to distinguish themselves from General Purpose English (GPE). Account of specificity has been given from a number of different perspectives, including language content, communication skills, and needs analysis (Dudley-Evans & St John, 1998), of which language content concerns most for this thesis. The following sections introduce specificity of language content from three different but interwoven prospects towards conceptualizing ESP: (a) Corder's (1973, cited in Bloor & Bloor, 1986, pp. 16-21) common core hypothesis, which assumes the existence of a basic core in ESP content; (b) Basturkmen's (2006) interpretation of the notion of specificity, in which specifiable

⁵ Traditionally, ESP has been segmented into two main areas: English for Academic Purposes (EAP) and English for Occupational Purposes (EOP) (Dudley-Evans & St John, 1998, pp. 5-6). Subordinate categorization goes on to generate seemingly endless acronyms, e.g., English for Academic Legal Purposes (EALP), English for Academic Business Purposes (EABP), English for Academic Medical Purposes (EAMP), and so forth (Belcher, 2006, p. 134). For the following literature review, this thesis will touch on English for Academic Purposes (EAP) and English for Business Purposes (EBP) by referring them with the same acronym ESP, since classification of types of ESP is not the major concern for this study.

language items occur in specific genres (pp. 26-28); and (c) Sinclair's idiom principle (1991, pp. 110-112), which promotes a phraseological view in describing language (Hunston, 2002, pp. 137-157).

2.1.1 Common core hypothesis. *Common core hypothesis* (Corder 1973, cited in Bloor & Bloor, 1986, pp. 16-21), a widely-accepted perspective on specificity of ESP language, posits that there is a basic set of vocabulary and structures shared among language varieties, and this basic core is to be prioritized in syllabus. Common core hypothesis is composed of two constructs: common core and common core plus. *Common core* represents high-frequency words and sentence structures that are all-purpose in any situation, whereas language items fall outside the common core are regarded as *common core plus*, which gives specific-purpose meanings in target disciplines or occupations. For applied linguists who adopt this hypothesis, the common core is a construct that can be realized as a specific set of lexis and grammatical constructions; this set is supposed to be basics for language learning, upon with purpose-specific language can be added in syllabus in sequence. This belief is shown in Coxhead and Nation's (2001) argument advocating application of their categorization of academic vocabulary: "When learners have mastered control of the 2,000 words of general usefulness in English, it is wise to direct vocabulary learning to more specialized areas depending on the aims of the learners" (pp. 252-253).

Common core hypothesis can be observed in a number of lexical researches. For example, Coxhead's (2000) Academic Word List (AWL) was compiled by excluding words in West's (1953, cited in Coxhead, 2000) General Service List (GSL) from a collection of written academic texts. In Coxhead's study, GSL was viewed as

representative of the common core for exclusion, and the remaining words, filtered with corpus linguistic procedure to be identified as academic vocabulary, were implicitly regarded as the common core plus. In addition to academic purposes (Coxhead, 2000; Coxhead & Nation, 2001), this thread of research can also be seen in many studies for compiling wordlists in other specialized domains, such as business (Chujo & Genung, 2004), engineering (Ward, 2009a), medical academics (Baker, 1988; Wang, Liang, & Ge, 2008), and so forth.

However, common core hypothesis receives some criticism regarding validity of the common core. Bloor and Bloor (1986) doubts that the common core pre-exists independently from any language varieties⁶ (pp. 18-20). They argue that, in language teaching, any language items must be presented “within the context of some variety” and it is inevitable to “present the basic elements in some kind of linguistic context” (Bloor & Bloor, 1986, p. 18). In other words, the common core does not exist in vacuum a priori, but has existence a posteriori since the common core is derived from linguistic contexts. Bloor and Bloor (1986) emphasize the importance of context, as they affirm:

A language learner is as likely to acquire “the language” from one variety as from another, but the use of language, being geared to situation and participants, is learned in appropriate contexts. This view supports a theory of language use and the basis of language acquisition theory. (p. 28)

Bloor and Bloor’s criticism on common core hypothesis is accepted by Flowerdew and Peacock (2001), who add that common core hypothesis neglects the

⁶ Bloor and Bloor (1986) use the terms varieties and registers nearly synonymously, both of which refer to language in use (pp. 20-21).

semantic aspect of language. They state the fact that meanings of one identical item may vary according to contexts and thus meanings cannot be divorced from context⁷. As Flowerdew and Peacock (2001) affirm: “[M]eaning is determined by context, if meaning is to be incorporated into the common core hypothesis, it is not possible to escape from the notion of specific varieties” (p. 17).

On common core hypothesis, the above criticisms can be synthesized as that the common core is an a posteriori construct that can only be derived from specific language uses. In my view, although the common core seems a theoretical construct, the hypothesis is necessary for syllabus design in that management of materials requires systematically arranging language content, and using form of language (lexical units and grammatical structures) as criteria to grade contents is the most objective and replicable method for pedagogical practice. Thus the present thesis supports an adapted version of common core hypothesis, in which the existence of common core is not an assumption but a set of language items derived empirically from specific contexts.

At this point, we are drawn to a further discussion about forms of language and context where language items occur, which will be taken up by the next section.

2.1.2 Specifiable words in specific genres. To conceptualize forms of language and context of language in a framework for English for Specific Purposes (ESP), Basturkmen (2006) presents the concept of specificity on two different layers of notions (pp. 12-13, pp. 26-28). In Basturkmen’s framework, specificity of ESP language is

⁷ A prominent example would be the word *memory*: Generally it refers to events or experience remembered by human, but it will specifically mean capacity of computer disk once it occurs in a technical genre.

dealt with by categorizing ESP content as language uses and language systems.

Language uses refer to units of language use, such as speech acts and genres, which is *specific* in target disciplines and workplace. On the other hand, *language systems* refer to sentence-level features such as certain forms or patterns that occur in specific situation. Language features that occur in one specific environment are not exclusive to those in other contexts, and are therefore considered as *specifiable* elements in ESP language. In other words, one kind of language system appears in multiple language uses.

With Basturkmen's ESP framework, since genres are relatively easy to be recognized or predetermined, the focus of this literature review is to be confined to sentence-level language features to proceed to our study on common words in ESP.

2.1.3 Lexical patterns in different contexts. When language is put under scrutiny at the sentence level, conventional grammar theorists mostly adhere to dichotomy of vocabulary and syntax, which treats words and grammar structures separately. This traditional approach to meaning interpretation assumes that lexis and syntax operate at different level of language; that is, a sequence of linguistic elements are segmented and can be presented in a tree structure and nodes in the tree are points open to lexical choices, which is termed by Sinclair (1991) as *open-choice principle* for describing organization of language.

Nevertheless, Sinclair (1991) doubts whether this "slot-and filler" model is the primary one for language organization and description (p. 114). According to Sinclair (1991), the open-choice principle loses validity in describing frequently co-occurring multi-words (e.g., *of course*, *hard evidence*, *set eyes on*, etc.) and lexical choices along language sequences are in fact restrained by semantic environment and discourse

organization in speech acts or genres with specific social functions (pp. 110-112). This feature of unrandomness is termed as *idiom principle* by Sinclair (1991), who maintains that it should be the primary model that dominates language phenomenon (pp. 110, 114); where one sequence opens to variation, the open-choice principle comes into play as the complement rule to govern lexical choices. These two principles together adequately describe the phenomenon of multi-words that vary with different degree of fixedness (Sinclair, 1991, p. 114).

Sinclair's argument and other studies in the field of lexical research have challenged the traditional vocabulary-grammar dichotomy and lead to a phraseological view toward language (Schulze & Römer, 2008). Phraseology refers to that vocabulary behaves in preferred sequences that may assign specific meanings (Hunston, 2002, p. 137). Research on phraseology can be seen from lots of lexical studies focusing on collocation (Durrant, 2009), lexical bundles (N-Grams) (Hyland, 2008; Jablonkai, 2010; Römer, 2008), pattern grammar (Charles, 2006; Groom, 2005; Hunston & Francis, 1998), semantic sequences (Hunston, 2008), and so forth, all of which can be covered under the over-arching term formulaic language (Schmitt, 2010). The view of phraseology has laid a firm base for the lexical approach to English Language Teaching (ELT) (Lewis, 2000, pp. 147-148).

The phenomenon of phraseology can be demonstrated by the word *maintain*. According to Hunston (2002), *maintain* patterns in three different ways to express dissimilar meanings—"maintain something," "maintain that something is true," and "maintain something at a level" (p. 139); these phraseologies unambiguously express multifaceted meanings of the identical word *maintain*. This phenomenon is especially

essential for frequent words, which behave in rather fixed phrases (Hunston, 2002, p. 102); hence phraseologies of frequent words are significant teaching points in language instruction (Hunston, 2002, p. 139).

Researchers have indicated that phraseologies of words are closely related to disciplines and genres. Durrant (2009), in his study on compiling a collocation list for students in academic context, has indicated that lexical knowledge required for learners in Arts and Humanities are strikingly different from that for those in the domains of Natural Science and Social Science. Charles (2006), with her comparison of reporting clauses (that-clause complement) in two corpora of Politics/International Relations and Material Science, has found that language of Materials Science features in the use of the pattern *find/show/observe that*. In addition to cross-disciplinary observation, Groom (2005) has added another aspect on genre comparison (research articles and book reviews) on exploring the two patterns *it v-link ADJ that-* and *it v-link ADJ to-inf*, and has reported quantitative difference in distribution across the involved two genres and two disciplines (History and Literary Criticism). Disciplinary variation of lexical bundles has also been observed by Hyland (2008), who has analyzed frequencies of 4-word bundles in four disciplines: Biology, Electrical Engineering, Applied Linguistics, and Business Studies.

The above studies signify the role of contexts for variance of language patterns. They also echo Basturkmen's ESP framework (Section 2.1.2), in which language items are specifiable in specific genres. Moreover, taking common core hypothesis, genres, and phraseology all together on board, it invites an interesting inquiry for variation of formulaic language composed by common words across genres.

2.1.4 Tentative research question. To conclude, conceptualization of common words in ESP must be based on specific language use. And since that words pattern in relation to context, it awaits our further exploration on phraseology of common words across genres. Here one research question is tentatively proposed:

How formulaic language, composed with identical common words, vary across genres?

This question indirectly relates to the purpose of the present thesis. This study aims to prove that corpus method can provide ESP practitioners assistance to identify real content for language instruction; if the proposed research question can be responded with definite result, this study in itself would serve as a demonstration for the application of corpus in arranging ESP materials. Moreover, the proposed question directly relates to real content of ESP, which refers to multiple lexical units composed with general vocabulary (see Introduction). Therefore, research results obtained for the research question would simply be real content of ESP.

The following sections continue the discussion about common words and phraseology from one another perspective: methodology for extracting common words and for identifying formulaic language.

2.2 Extraction of words

Before touching on methods for extracting common words from texts, we have to discuss how to define common words. The idea of common words, or general vocabulary, is necessitated by pedagogical need to creating reading materials or setting vocabulary goal for language learners. Common words are generally accounted for in a rather quantitative manner; Schmitt (2010) has viewed common words as “the

higher-frequency vocabulary necessary to achieve a basic functionality with a language” (p. 75). The concepts of frequency and functionality of words will be surveyed by reviewing three studies on creating word lists.

2.2.1 Common words. One well-known venture for compiling common words is West’s (1953, cited in Schmitt, 2010, p. 75) General Service List (GSL), which contains about 2,000 words that can function as basic words for illustrating more advanced vocabulary for language learners. With this pedagogical purpose, words in GSL were selected by considering word frequency, word-building capacity, and types of genre, and so forth. Colloquial, slang words, and technical words for specialized fields were excluded. GSL has been a prestigious groundwork for subsequent lexical researches. For example, GSL has been employed to compile the Academic Word List (AWL) (Coxhead, 2000) and has also been applied as yardstick to measure learners’ vocabulary knowledge (Ward, 2009b).

Functionality of GSL can be evaluated by measuring the proportion of GSL words in texts of different genres or disciplines. According to Coxhead’s review (Coxhead, 2000, pp. 213-214), words in GSL cover up to more than 75% of fiction, non-fiction, and academic texts; the high percentage demonstrates the practicality of GSL.

An equivalent work of General Service List (GSL) in East Asia is Jeng and his colleagues’ Senior High English Wordlist for Reference⁸ [高中英文參考詞彙表] (Jeng, Chang, Cheng, & Gu [鄭恆雄、張郁慧、程玉秀與顧英秀], 2002). The creation of this

⁸ The compilation of Senior High English Wordlist for Reference is partly activated by its 1998 version, created by Chang and his colleagues [張武昌等] (1998, cited in Jeng, Chang, Cheng, & Gu [鄭恆雄、張郁慧、程玉秀與顧英秀] 2002).

word list is motivated by need for reference word list that helps delimit the amount of vocabulary required for senior high school students in Taiwan to take the College Entrance Examination. Senior High English Wordlist for Reference is graded into six levels, each with 1,080 words to aggregate 6,480 words in total, which were extracted from a collection of texts from 21 kinds of source covering from textbooks from Taiwan and the U.S. and ready-made wordlists from Mainland China, Japan, and English-speaking countries like Canada and the British. Its major principle to enlist words is based on a specified threshold of word frequency. For example, words enlisted from Level 1 to Level 4 were determined by a threshold of 5 occurrences in the collection of texts. Socio-cultural factors, grammatical categories, word families, and so forth, were considered in creating and displaying Senior High English Wordlist for Reference. This word list is anticipated to serve as a primary benchmark for vocabulary learning in the pedagogical context of Taiwan.

As for functionality of Senior High English Wordlist for Reference (hereafter SHEWR), there seem scant studies on measuring proportion of SHEWR words in specific genres and disciplines. The present thesis thus sets one of its research aims as to evaluate the functionality of SHEWR in specific context.

2.2.2 Tentative research question. Here one tentative research question is proposed:

To what extent do SHEWR [高中英文參考詞彙表] words occur in texts of specific genre?

By proposing and responding to this question, we are able to evaluate the functionality of SHEWR in specific language use. The same research result would also corroborates

Dudley-Evans and St John's (1998, pp. 11, 16) argument that it is the real content, language that doesn't involve technical knowledge, should be the focus of ESP practitioners' instruction (see Introduction of this thesis). And more importantly, by investigating functionality of SHEWR against an ESP background, we have a chance to survey whether the common-word list brings benefits to English teaching of one specific occupation or context and to signify the usefulness of SHEWR for ESP pedagogy in Taiwan. Section 2.4 will address which kind of genre is to be put under exploration; the following section continues the discussion on word extraction with wordlist made for specific purpose.

2.2.3 Specialized words. The idea of functionality in making common word lists can also be observed in the process of creating other word lists for specific purpose. The well-known Academic Word List (AWL) (Coxhead, 2000), possessing some important notions on functionality underlying its delicate compiling procedure, will be attended below as an exemplar for specialized wordlists.

The AWL contains 570 word families (all derivatives closely related to one stem is counted as one word family) selected from a 3.5-million-word collection of written academic texts covering 28 subjects from 4 disciplines, and this word list was compiled by

- (a) excluding word families listed in West's (1953) General Service List (GSL), and
- (b) including any word family members with a minimum occurrence of 15 times in the 28 subjects and 10 times in each of the 4 disciplines, and
- (c) including any word family members with a minimum occurrence of 100 times in the established Academic Corpus.

There are four fundamental notions underlying the creation process of AWL: i) seeing members of the same word family as one word, ii) text coverage, iii) frequency, and iv) excluding GSL to identify AWL. Each of these four notions is worthy of further discussion.

The first notion underlying AWL is that, words in AWL are presented in the form of word family. For instance, the word *indicate* is treated as one single entry, leaving aside its derivatives (*indication, indicative, indicator*) and all inflections (*indicated, indicates, indicating, indications, indicators*) (Coxhead, 2000, p. 218). The reason for Coxhead to count closely morphologically related words as one single unit for wordlist inclusion is based on psycholinguistic researches, which maintain that derivatives and inflections of one stem are easily accessed and hence might be stored as one component in the mental lexicon.

However, from the perspective of genre analyses, different members of the same word family may possess distinctive connotation in specific genre. For example, in exploration on a collection of cancer research articles, Gledhill (1995, as cited in Hunston, 2002, p. 201) has pointed out that choices between *is* and *was* reflect specific philosophy of the discipline under study. Therefore, it is a subjective decision as whether to present morphologically related words in one single entry or to display them as separate entries. For the present thesis with its attempt to describe common words in patterns (Section 2.1.4), it is more appropriate to present words in their different forms separately to allow subsequent identification of formulaic language composed by common words.

The second notion applied in AWL we can take notice of is word coverage among text, which can be seen from Coxhead's threshold of word frequency in terms of all

considered subjects (minimum set as 15 occurrences) and disciplines (minimum set as 10 occurrences). Generally the composition of a corpus is designed according to predetermined purpose by deliberately balancing proportion of texts of various types (Hunston, 2002, pp. 28-30). By observing word coverage among all texts within a corpus, one can ensure that selected words occur across contexts and do not bias for certain types of texts. This notion of word coverage has been applied in corpus software. For example, Wordsmith Tools (Scott, 1997, 2011b) possesses the “Key-Keywords” function to compute words with information of their degree of coverage among texts. The notion of word coverage will be applied in this thesis to extract common words of interest.

The third notion involved in creating AWL is frequency of word occurrences. Similar to GSL and Senior High English Wordlist for Reference mentioned above, AWL has utilized frequency as threshold for word selection (threshold set as a minimum of 100 occurrences). The idea of frequency presumes that one word with higher frequency in a corpus signifies its greater probability of occurrence in real language use, which therefore implies its pedagogical importance (Hunston, 2002, p. 194). Nevertheless, discerning significance of words based solely on raw frequency often results in grammatical words such as prepositions *of*, *to*, *in*, and so forth, because they come with relatively higher occurrence compared to content words (Hunston, 2002, pp. 3-5).

The problem of high-frequency of grammatical words can be resolved by utilizing the notion of *unusual frequency*. Unusual frequency refers to outstandingness of words, or *keyness* of words in the software Wordsmith Tools (Scott, 1997, 2011b), which is measured by comparing word frequencies between two corpora. With corpus software,

degree of outstandingness of word is computed by comparing their frequencies in one corpus to those in another reference corpus with statistic measures such as chi-squared or loglikelihood. The same notion of unusual frequency will be employed by the present thesis to locate common words in order to avoid the problem of high-frequency words.

Lastly, in the process of creating AWL, words in General Service List (GSL) were assumed to represent general vocabulary and were excluded to narrow down the scope of candidate words pertinent to academic usage. This procedure indicates that in lexical research, general words and specialized ones are regarded to be mutually defined; in any collection of texts, specialized words can be specified by excluding general ones, and vice versa. For the present study, common words that do not involve technical meanings are to be extracted by excluding specialized words from target corpus.

In addition to corpus techniques for extracting specialized words, as those in creating AWL mentioned above, using a technical dictionary has been proved a reliable method for identifying specialized words. Chung and Nation (2004), in their study on comparing efficacy of four different approaches to identify specialized words in anatomy text, indicate that using a technical dictionary can have an accurate rate about 80%, which is comparable to that with corpus method. For the present study, which has previously acknowledged that general words and technical ones are mutually defined, it will adopt a specialized wordlist made by professionals in the field to identify common words.

2.2.4 Summary on methods for extracting words. Section 2.2.1 and Section 2.2.3 have studied on methods for extracting words of interest to synthesize working definition for common words. With our review on West's GSL (1953, cited in Schmitt, 2010, p. 75) and Jeng et al.'s [鄭恆雄等] SHEWR [高中英文參考詞彙表] (2002), we

have concluded that the degree of functionality of wordlist can be evaluated by measuring the portion those words bear in any discipline or genre; and SHEWR, a common word list geared to needs of English learners in Taiwan, requires to be evaluated with a similar method. Also, the detailed survey on Coxhead's (2000) AWL has amalgamated some practical notions and corpus techniques for word extraction. These notions and corpus techniques will contribute to give operational definition of common words in the present thesis, as will be illustrated below.

2.2.5 Operational definition of common words. For the present thesis, the abovementioned notions involved in compiling AWL on extracting words of interest will be applied to yield common words of pedagogical value for English learners in Taiwan. To put it specifically, common words in this study will be extracted by

- (a) excluding words that possess specialized meanings in the target corpus, and
- (b) selecting words of significant outstandingness by comparing them to their equivalents in a reference corpus, and
- (c) matching previously resulted words to those in the Senior High English Wordlist for Reference (SHEWR) [高中英文參考詞彙表] (Jeng, et al. [鄭恆雄等], 2002), and
- (d) sorting previously resulted words according to degree of coverage in the target corpus.

Among the above steps, it is anticipated that Step (c) will create pedagogical value for English learners in Taiwan by bridging the gap between ESP and public education, since the resulted words are going to be evaluated with their proportion in specific genre and used for locating formulaic language across genres. For this thesis,

the series of procedure serve as the operational definition of common words of interest.

Application of this definition will be reported in Chapter 3.

To recap, Section 2.2 has reviewed West's General Service List (GSL), Jeng and his colleagues' [鄭恆雄等] Senior High English Wordlist for Reference (SHEWR) [高中英文參考詞彙表] and Coxhead's Academic Word List (AWL) to gain insights regarding compiling word lists for either general or specific purposes. Pedagogical purpose of expected wordlist affects subjective decisions on which kind of corpus and ready-made wordlist to be adopted. With corpus methods, words can be extracted by objective computation considering word coverage, simple/unusual frequency, and techniques of matching or exclusion. For the present thesis, with one of its aims to identify formulaic language composed by common words, the issue of utilizing which kind of corpus and wordlist will be discussed in Section 2.4. Prior to Section 2.4, the following section will investigate a rather difficult topic in corpus linguistics: methods for identifying formulaic language.

2.3 Identification of formulaic language

As been mentioned in section 2.1.3, corpus linguists have noted the phenomena that language items tend to occur in patterns that cannot be explained with vocabulary-grammar dichotomy, which leads to the phraseological view toward language organization and description. Under the overarching term formulaic language, phraseology has been approached from various perspectives including collocation, lexical bundles (N-grams), pattern grammar, and so forth. These various descriptions signify the complexity of formulaic language, on which formulaic language is defined by Wray (2002, as cited in Schmitt, 2010, p. 120) as:

[A] sequence, continuous or discontinuous, of words or other elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar.

The latter half of Wray's definition focuses on the view of psycholinguistics.

Here we center on the first line of Wray's definition: formulaic language can be continuous or discontinuous sequence of words.

Take the sequence *a(n) _____ ago* for illustration (Schmitt, 2010, p. 132). This string of formulaic language is composed with continuous lexical items, along which the underlined slot is open to temporal lexemes, such as *hour*, *year* or *very long time*, making this string discontinuous at the slot. It is this feature of internal lexical variation that imposes difficulty on identification of formulaic language (Schmitt, 2010, p. 120). Difficulties about identifying formulaic language will be explored further by examining collocation and lexical bundles, with which corpus method for locating formulaic language with lexical variation is to be discussed.

2.3.1 Collocation. Collocation refers to the language phenomenon that two words co-occur with biased tendency, which can be measured with corpus statistics (Hunston, 2002, p. 68). Tendency of any two words to co-occur can be gauged via two conceptually-different approaches: One measures the strength of association of the two words with mutual information (MI) and the other considers the probability of co-occurrence of them with asymptotic hypothesis test (Schmitt, 2010, p. 124; Clear, 1993, pp. 279-282), both of which are discussed below.

Mutual information (MI) denotes the strength of co-occurrence of two words by comparing “the actual co-occurrence of the two items with their expected co-occurrence” (Hunston, 2002, p. 71). The actual co-occurrence of the two words is the observed occurrences appeared in a pre-determined span within which one of the two items is set as the node word; the expected co-occurrence refers to expected instances of the two words in the target corpus as a whole within the same width of span. MI is computed by dividing the Observed (O) with the Expected (E), and converting the resulted ratio to a base-2 logarithm, as demonstrated by the formula below:

$$MI = \log_2 \frac{O}{E} \quad (\text{Schmitt, 2010, p. 130})$$

Conventionally, a MI-score above 3 can be regarded as a significant value for identifying collocation (Hunston, 2002, p. 71). For example, to find co-occurring words of the word *gaze* in the Bank of English (Hunston, 2002, pp. 69-71), *gaze* is queried as the node word with span of node set as 4:4; that is, four words to the left of the node word and four words to the right. Results show that the collocate (co-occurring word of the queried word) of *gaze* with highest MI-score (12.1) is the word *Angelopoulos*, which is the name of the director of the film *Ulysses Gaze*. The high score of MI of *gaze* and *Angelopoulos* is motivated by the strong relation between the film and the director.

Another approach to identify collocation considers probability of co-occurrence of any two words by testing hypothesis. Statistic measures (t-score, z-score, chi-squared, and log-likelihood) are applicable to hypothesis tests, in which the null hypothesis is designated as that “words appear together no more frequently than we would expect by chance alone” (Schmitt, 2010, p. 124). Take t-score for illustration, the formula of t-score is that the observed occurrence minus the expected occurrence of the two words,

and the result is divided by the standard deviation (Hunston, 2002, p. 70), as demonstrated below:

$$t - score = \frac{O - E}{\sqrt{O}} \quad (\text{Schmitt, 2010, p. 126})$$

Normally, a t-score above 2 is taken to be significant (Hunston, 2002, p. 72).

Take the same example of *gaze* in the Bank of English, the collocate with the highest t-score (25.2) is *his*, with which we can claim with certainty that *gaze* lexically prefers *his*.

Both abovementioned approaches to identify collocation are objective and reliable because conceptions of statistics are applied. However, employment of statistics might make identifying collocation an unnecessary heavy task. Danielsson (2003), critiquing on Church and Hanks' (1989, as cited in Danielsson, 2003, p. 112) advocating the use of t-score to discern *strong support* as a more common collocation than *powerful support*, indicates that the same conclusion would be reached by simply observing raw frequency of collocation, since that *strong support* occurred 175 times and *powerful support* occurred only twice in their collection of texts. Applying raw frequency to identify collocation would turn our attention to lexical bundles (N-Grams), which will be introduced in the next section.

To consider identification of collocation together with one of the objectives of the present thesis, locating formulaic language formed with common words, those methods for identifying collocation are obviously incapable of picking out multi-word expression such as *a(n) ____ ago* mentioned previously. Moreover, variation within formulaic language is beyond the scope of a collocation list, unless collocates are to be observed from concordance lines (Hunston, 2002, p. 77).

In summary, the identification of collocation, when viewed under corpus linguistic methodology to identify formulaic language, is restricted on locating multi-word expressions as well as lexical variables. Over-reliance of statistics on collocation identification might also be a potential defect in locating formulaic language.

The problem of identifying multi-word units and avoidance of use of statistics can be dealt with if formulaic language is regarded as lexical bundles, which will be discussed in the next section.

2.3.2 Lexical bundles (N-Grams). In the previous section, we have discussed corpus statistics for identifying collocation. In addition to collocation, there is another type of formulaic language called lexical bundles, which refers to words that co-occur continuously within a predetermined length of a word string with high raw frequency (Biber et al. 1999, as cited in Biber, 2009, p. 282). The conventional method to locate lexical bundles with corpus software is to set length of word sequence as 3-word, 4-word, or 5-word, and word bundles can be sorted out based on simple frequency. Lexical bundles are also called N-Grams because they come with fixed string of N length (Schmitt, 2010, p. 123).

Research on lexical bundles brings benefits to ESP course. For example, Jablonkai (2010) has conducted a study on English documents of the European Union (EU) and collected 4-word lexical bundles of different discourse functions, which can serve as language focus for ESP writing instruction. For instance, *it is necessary to* is used to claim importance of a following proposition; *as set out in* clarifies topic of discussion; *as referred to in* is used for text-deixis, and so forth (Jablonkai, 2010, p. 262).

According to Jablonkai, a large portion of EU texts consist of lexical bundles, which therefore should be explicitly taught in class on English for EU purposes.

Although the process of identifying lexical bundles seems a reliable procedure, in which all words in an identified bundle are sorted out based on simple frequency, this approach is not without problem. As previously shown, lexical bundles tend to be composed with prepositions (e.g., *as*, *to*, *in*), and prepositions often position at the boundary of identified bundles. Sometimes there comes with multi-words that are difficult to interpret, such as *in so for as* or *to in Article #⁹*, which are also listed in Jablonkia's (2010) results. This phenomenon is largely due to the fact that in English, prepositions are words that have relatively higher frequency than other words do (see Hunston, 2002, pp. 69-70), and prepositions' high frequency make themselves the regular members of lexical bundles. High frequency of prepositions also causes identified lexical bundles to have prepositions as boundary words, as simple frequency is used in the identification process.

One more drawback of lexical bundles is that they do not allow variable items to be observed along the located word string. Identified lexical bundles are composed of continuous sequence of words, such as *it is necessary to* (Jablonkai, 2010, p. 262), with which we are not able to inquiry whether the sequence opens to variation, such as *it is vital to*.

The abovementioned demerits of lexical bundles, as well as those in collocation, are sought to be avoided by continuing discussion on other available method for locating formulaic language in the next section.

⁹ The symbol # represents numbers or figures in text.

2.3.3 Cumulative frequency¹⁰. To better the process and results of locating multi-word sequence, Danielsson (2007) proposes an alternative¹¹ that can solve the above problems in the identification of collocation and lexical bundles.

Instead of presetting length of sequences as in identifying lexical bundles, Danielsson (2007) sets off locating multi-words by seeking the most frequent co-occurring word (F1) of any target node item (N) within a span of 9 words (i.e., four words in two respective sides of the queried word). Once F1 is recognized, F1 and N are then used together in the subsequent step with the same 9-word-span to proceed the search for the next most frequent word (F2). The same procedure is repeated until the frequency of newly found co-occurring word fall to below 5. For example, Danielsson (2007) launched a query with the word *jam*, and ended it up in a small number of concordance embedded with a definite multi-word sequence *stuck in a traffic jam*.

There is one central point that is worth noting in Danielsson's method. In identifying collocates (high-correlated words) of the target item, Danielsson (2007) has left aside conventional statistics computation such as Mutual Information or t-score based on the reason that these statistics are largely derived from the assumption of random distribution; the assumption doesn't hold true in language since language is an obviously non-random system (p. 18). Therefore, the notion of simple frequency is adopted in

¹⁰ The term *cumulative frequency* comes from Hunston's (2008) discussion on the role of frequency in locating semantic sequences.

¹¹ The software ConcGram (Greaves, 2005) has been developed for locating multi-words with variation of constituents identified without querying a target word in advance (Cheng, Greaves, & Warren, 2006). Since the present study is devoted to locate formulaic language with pre-assigned common words, the method of ConcGram is not dealt with in this thesis.

Danielsson's method, in which the most frequent co-occurring words observed cumulatively are taken as collocates of the queried word.

Another advantage of Danielsson's method is that syntagmatic and paradigmatic variations of multi-word sequence are allowed to be recognized or testified.

Syntagmatic variation refers to that each word in a sequence is open to other modifying items inserted in between. For example, an intuitive guess at syntagmatic variation of Danielsson's result *stuck in a traffic jam* had been *stuck in a hellish traffic jam*; however, the guess is not verified with observation on all extracted concordance lines. The other kind of variation, paradigmatic variation, refers to that one word in a sequence may be substituted with other words at the same slot. This variation can be checked by testing each word in a string to see if there are slots open to variations. For instance, the position of *stuck* in the sequence *stuck in a traffic jam* was checked by observing items ahead of the sequence *in a traffic jam*, and results showed that *sitting, waiting, caught*, were interchangeable with *stuck*. Based on the above findings, Danielsson (2007) affirms that this test for paradigmatic variation is valuable for discovering a set of lexis whose meanings denote certain associations that is only interpretable along with the specified word string.

One other vantage of Danielsson's (2007) method is that, in my view, the length of any identified formulaic language can be measured with number of extracted collocates. This measure would gauge formulaic language with quantitative data and give information that cannot be provided with those methods for identifying collocation and lexical bundles (Section 2.3.1 & 2.3.2).

Because of the abovementioned advantages of Danielsson's (2007) method, the approach of cumulative frequency will be adopted by this thesis to identify common-word patterns. As variation of formulaic language across genres is one of this study's concerns, identified sequences will be investigated in terms of lengths of formulaic language as well as syntagmatic and paradigmatic variation of formulaic language among different genres.

2.4 Integrating corpus research with genre-based method

To investigate the research question proposed in Section 2.1.4 that inquires how formulaic language composed with identical common words varies across genres, the previous sections have addressed on the operational definition of common words in Section 2.2.5 and the chosen method for identifying formulaic language in Section 2.3.3. This section focuses on the issue of genre, especially when genre is taken into account with corpus research.

The research of genre is closely connected to English for Specific Purposes (ESP) (Dudley-Evans & St John, 1998, p. 87). For Swales (1990, as cited in Kay & Dudley-Evans, 1998, p. 309), a genre "comprises a class of communicative events, the member of which share some set of communicative purposes." With his research on academic texts, Swales sees communicative purposes as "moves," which are comprised of "steps" that are presented by textual elements. For example, the Research Article Introduction is regarded as a specific genre with structure as such:

Move 1 Establishing a territory

Move 2 Establishing a niche

Move 3 Occupying the niche

To achieve Move 1, textual elements below are to be applied:

Step 1 Claiming centrality

Step 2 Making topic generalizations

Step 3 Reviewing items of previous research

(Swales, 1990, cited in Kay & Dudley-Evans, 1998, p. 309)

This genre approach for describing specificity of ESP can also be employed to account for Business English. Bhatia (1993, as cited in Dudley-Evans & St John, 1998, p. 91), in his analysis on sales promotion letter and job application letter, concludes that these two types of letter belong to the same genre as they share nearly identical pattern of moves:

	<i>Sales Promotion Letter</i>	<i>Job Application Letter</i>
Move 1	Establishing credentials	Establishing credentials
Move 2	Introducing the offer	Introducing the candidature
Move 3	Offering incentives	Offering incentives
Move 4	Enclosing documents	Enclosing documents
Move 5	Soliciting response	Using pressure tactics
Move 6	Using pressure tactics	Soliciting response
Move 7	Ending politely	Ending politely

Other researches on business genre explore on the level of textual features to gain more concrete evidence for specificity. Yeung (2007), by analyzing 22 authentic business reports, indicates that the genre of business reports is one sub-genre of the prototypical scientific reports with their shared linguistic features of impersonality and use of statistics to build credibility of documents themselves, as persuasion and

recommendation for decision-making is the ultimate purpose of business reports.

Flowerdew and Wan (2010), with the similar integration of genre-based and corpus-based approach applied on researching audit reports, have observed how moves (i.e., communicative function), such as *opinion*, *qualified opinion*, *disclaimer of opinion* and so forth, are realized with linguistic patterns.

As can be seen, the integration of genre-based and corpus-based approach to text analyses yields valuable results that are practical to ESP courses. This integration has been advocated by Flowerdew (2005), and it is suggested that a specialized corpus composed of texts of a specific genre is required for text analyses (L. Flowerdew, 1998; Schulze & Römer, 2008).

As for the present thesis, which aims to explore how formulaic language, composed with identical common words, vary across genres, the researcher would like to follow the research on business genre in the accounting field with a homemade corpus consisted of annual reports for text analysis.

Here the term “corpus-based” needs further discussion regarding corpus linguistic methodology. In corpus linguistics, a *corpus-based* method refers to that pre-assumed linguistic categorization from theory (e.g., word class) is applied in analyses; while a *corpus-driven* method does not hold a priori assumptions about language items (Biber, 2009, p. 276). The present thesis, which aims to identify formulaic language with extracted common words, may be regarded as a hybrid of corpus-driven and corpus-based research. It can be seen as corpus-driven because of that all closely morphologically related words will be treated as separate words, in which grammatical categorization is not considered (Section 2.2.3). Meanwhile, this study can also be seen

as corpus-based since formulaic language will be identified with predetermined common words, which involves assigning certain attributes a priori to words under exploration. Therefore, the researcher would simply regard this study as a corpus research integrated with consideration of genres without explicitly labeling it as a corpus-driven or corpus-based study.

2.5 Summary of chapter and research questions

In this chapter, literature has been reviewed to attend to the issue of real content of ESP introduced in the beginning of this thesis with the attempt to prove that corpus method is a feasible approach, in addition to collaboration, to assist ESP practitioners in arranging materials. Considering that real content comes with multi-words composed with common lexis, we have discussed how common words are conceptualized with common core hypothesis and ESP framework, and concluded that analysis of common words shall not be divorced from genres and phraseology.

The idea of common words is explored further by examining previous wordlists in the perspective of corpus linguistic methodology, and a set of operational definition of common words has been constructed for the present study. During our discussion on common words, our attention has been directed to the functionality of Jeng et al.'s [鄭恆雄等] SHEWR [高中英文參考詞彙表] (2002), a common-word list made for English learners in Taiwan. This thesis will evaluate the functionality of SHEWR by measuring its proportion in one specific genre.

Afterwards, the discussion on extraction of words is expanded to the sentence level in which methods for identifying formulaic language are scrutinized with collocation and lexical bundles. Drawbacks in identification of collocation and lexical

bundles are to be avoided by employing the method of cumulative frequency to locate formulaic language.

This thesis is a genre-based corpus study, in which a specialized corpus of company annual reports will be created.

To advocate that the application of corpus method is a feasible approach, besides collaboration, in sorting out real content for ESP instruction, two research questions have been tentatively proposed:

- (a) To what extent do SHEWR [高中英文參考詞彙表] words occur in texts of specific genre? (Section 2.2.2)
- (b) How formulaic language composed with identical common words vary across genres? (Section 2.1.4)

Considering that business texts have been chosen as the target genre and the operational definition of common words have been constructed for this study, the original questions are to be rephrased with logic of research procedure:

Research Question (1)

To what extent do common words occur in company annual reports?

Research Question (2)

How formulaic language composed with identical common words vary across genres?

And Research Question (2) can be rephrased from two different perspectives:

Research Question (2-1)

Do the lengths of identified formulaic language vary across genres?

Research Question (2-2)

Does composition of identified formulaic language allow syntagmatic variation or paradigmatic variation?

Implications of the proposed research questions are worth reiterating. To evaluate the functionality of Jeng et al.'s [鄭恆雄等] SHEWR [高中英文參考詞彙表] in specific genre, the expected research results responding to Research Question 1 are expected to display proportion of SHEWR words occurred in business reports, in which the usefulness of common words are to be evaluated with quantitative analysis. The same results are also expected to signify the role of common words in specific genre, specifically usefulness of SHEWR in the business genre.

The pedagogical value of results for Research Question 2 is that real content (see Introduction) of ESP will be identified as multiple lexical units composed with common words, which requires little technical knowledge for ESP practitioners to give language instruction. Also, the same results will be presented with lengths of formulaic language as well as syntagmatic and paradigmatic variation, unveiling the nature of formulaic-language variation across genres. For Research Question 2, it is hypothesized that lengths and composition of formulaic language are related to type of genre.

Results for both Research Question 1 and Research Question 2 together are anticipated to prove the efficacy of corpus method for identifying real content for ESP instruction, in that real content is regarded as multiple lexical units composed of common lexis. Moreover, with the adoption of Jeng et al.'s [鄭恆雄等] SHEWR [高中英文參考

詞彙表] to identify formulaic language, this thesis serves to build the connection between General Purpose English (GPE) instructed in public schools in Taiwan and English for Specific Purposes (ESP) in the international business field.

Now that the research objectives and questions of this thesis have been clearly addressed, we now move to discuss research design in the next chapter.



Chapter 3 Research Design

Under corpus methodology integrated with genre-based method, this thesis attempts to filter out real content that doesn't involve subject knowledge of a certain profession for ESP instruction. With reviewed literature, real content of ESP has been regarded as formulaic language composed with common words. This chapter proceeds the study by reporting research design relevant to the purpose of the present project, including creation of a specialized corpus, procedure of extracting common words of interest, and proceedings of locating formulaic language constituted with extracted common words.

3.1 Creating Business Reports Corpus

3.1.1 Source of data. As been mentioned in Section 2.4, this study will continue Yeung (2007) and Flowerdew and Wan's (2010) research on texts in the accounting field, specifically annual business reports of corporations. 20-F, one kind of business reports, is chosen to create a Business Reports Corpus (BRC) for this research. 20-F is one specific type of document designated by U.S. Securities and Exchange Commission (SEC) as prerequisite for foreign corporations who are involved in issuing financial securities to the U.S. capital market (U.S. Securities and Exchange Commission, 2010, December 13). 20-F contains information about performance of a company presented in figures and texts, stored in electronic form and free for download on SEC official website¹²; and since documents of 20-F are regulated as documents that have to be issued to the investment public, consideration regarding copyright should be ignored.

¹² The official website of SEC for downloading 20-Fs:

<http://www.sec.gov/edgar/searchedgar/companysearch.html>

In addition to the homemade specialized corpus BRC, another corpus is needed to serve as a reference corpus for computing keyness of words (Section 2.2.3). According to Baron, Rayson, and Archer (2009), keyness is open to bias due to spelling difference; since BRC is composed with American English, bias of keyness could be avoided if the reference corpus is also presented by a corpus consisting of American English as well. Brown Corpus (Francis & Kucera, 1979), a one-million-word general corpus compiled with texts from the United States, is hence adopted as the reference corpus in the present thesis.

3.1.2 Corpus size. The mainstream of corpus linguistics on size of corpus has adhered to the belief that “a corpus should be as large as possible” (Sinclair, 1991, p. 18). This position is mainly grounded on lexicographic reason, in that only large amount of instances can ensure thorough description of language particularities (Sinclair, 1991, p. 98). The Bank of English, evolved from the Birmingham collection of English texts in Sinclair’s time, is one exemplar of large corpora; it contains about 400 million words, and a query for the word *point* will yield about 143,000 hits (Hunston, 2002, p. 25). In contrast to lexicographic need for large corpora, ESP studies with pedagogical concern calls for small corpora composed of texts relevant to particular research interest. For instance, to investigate lexis in Business English, Nelson (2000) created a one-million-word Business English Corpus by collecting business-specific texts such as emails, meetings, newspapers, annual reports, and so forth. Moreover, with objectives of study clearly defined, a small corpus is particularly valuable in researching multi-word sequences that exhibit special function in target text (Hunston, 2008, p. 293).

For the present thesis, the adopted keyness computation also affects the consideration of corpus size. Theoretically, computation of keyness requires a relatively small corpus to be compared to a larger reference corpus (Scott, 2011b). At this point, it is expected that the size of the anticipated BRC should be less than the one-hundred-million word of Brown Corpus.

3.1.3 Representativeness and balance. One another issue regarding corpus creation is to collect language samples representative of language in real use. The issue of representativeness also involves how to achieve balance among different types of texts with appropriate proportion. For example, the one-million-word Brown Corpus (Francis & Kucera, 1979), designed for representing contemporary American English, contains 500 pieces of prose printed in the United States. Balance of the Brown Corpus is achieved by collecting samples of 2000+ words each, with genres covering from reportage to humor essays. An ESP-oriented example, Nelson's (2000) Business English Corpus, compiled to reflect language used in business context, obtains its balance by bearing due proportion on texts labeled as written/spoken, doing/talking business, types of industry, and so forth.

As for the present study on business reports, the balance of the Business Reports Corpus (BRC) is to be achieved by sampling annual reports of corporations from various types of industry. Moreover, reports from each company will be collected for three consecutive years to eliminate possible effects from market fluctuation on language in texts.

3.1.4 Dealing with technical compounds. Research on words in ESP has seen the emergence of some flaws. Lexical research which gives weight to observing

co-occurrence of words may run the risk of neglecting the role of compound words in subject-specific text. For instance, in analyzing how words co-occur in science textbooks, Menon and Mukundan (2010) have reported that there are some immediate 2-word collocations such as *chain reaction* (p. 247), *guard cell* (p. 250), *atomic mass* (p. 253), and so forth; Menon and Mukundan concluded that these combinations were important teaching points not only because of their relatively high frequency of co-occurrences but also that their extended meanings could not be literally understood. Meanwhile, these 2-word combinations were discussed by the researchers with reference to the *Oxford Dictionary of Science*, indicating the fact that specialized compounds are treated as independent entries in their own right with definite meaning. Menon and Mukundan's study is a practical research with pedagogical value, but its view on subject compounds seems self-contradictory. Since these 2-word combinations are not learned word by word, it is unnecessary to see these compounds as separate words with significant correlation. For ESP learners, a compound is actually regarded as a complete meaningful unit in any disciplines. Therefore, additional caution should be exercised in handling ESP compounds by assuring that correct length of multi-words is assigned to each distinct compound. In my view, any specialized compound, no matter how many words it contains, should be regarded as a single word in corpus analyses.

The anticipated problem on technical compounds will be tackled by applying subject-specific wordlist. In the present research on Business Reports, the *Chinese-English Translation of Important Accounting Terms* [重要會計用語中英對照]¹³ (*Accounting Research and Development Foundation of the Republic of China, 2011*) [中

¹³ This specialized wordlist can be downloaded from <http://www.ardf.org.tw/html/tifrs1001115.pdf>

華民國會計研究發展基金會] will be adopted to locate technical terms since in business practice, external Business Reports are compiled or audited by certificated accountants. Accounting compounds such as *account receivable* can be replaced by adding a dash in between as *account-receivable* to be seen as a single word by corpus software. For the present thesis, this work of replacement will be processed with a text processing software, Useful File Utilities (ReplSoft.com, 2010), with which blanks between compound words can be converted into dashes in batch processing. After the conversion, each entry of accounting compounds will be read by corpus software as one single word.

3.1.5 Procedure for creating BRC. The abovementioned concerns in the anticipated Business Reports Corpus can be seen from the creation procedure reported below.

First of all, to survey which and how many Chinese corporations are engaged in issuing securities in the U.S. stock market, the financial website, cnYes.com [鉅亨網]¹⁴ was logged on to. Then all the Chinese corporations were recorded onto an Excel spreadsheet.

Later on, English names of all recorded Chinese companies were queried into the official website of the Securities and Exchange Commission (SEC)¹⁵ under the “Company Search” interface to find more information of the queried company.

¹⁴ Website of cnYes.com [鉅亨網]: <http://www.cnyes.com/usastock/adrprice.aspx>

¹⁵ Website of SEC for company search: <http://www.sec.gov/edgar/searchedgar/companysearch.html>

Figure 3.1.1 displays result of a sample query for Taiwan Semiconductor Manufacturing Corporation (TSMC) [台灣積體電路有限公司]: the SIC (Standard Industrial Classification) code is shown and all available 20-F documents are chronically listed. SIC code is for classifying industry type of the company of interest¹⁶. For instance, the queried Taiwan Semiconductor Manufacturing Corporation is labeled as 3674, which refers to “semiconductors & related devices.” The 20-F documents can be downloaded by clicking on the “Documents” bottom on the web page and saved as a single htm file on to disk.

The screenshot shows the EDGAR Search Results page for TAIWAN SEMICONDUCTOR MANUFACTURING CO LTD. The SIC code is 3674 - SEMICONDUCTORS & RELATED DEVICES. The page lists several 20-F filings with the following details:

Filings	Format	Description	Filing Date	File/Film Number
20-F	(Documents)	Annual and transition report of foreign private issuers [Sections 13 or 15(d)] Acc-no: 0001193125-12-161528 (34 Act) Size: 3 MB	2012-04-13	001-14700 12757455
20-F	(Documents)	Annual and transition report of foreign private issuers [Sections 13 or 15(d)] Acc-no: 0000950123-11-035858 (34 Act) Size: 2 MB	2011-04-15	001-14700 11760934
20-F	(Documents)	Annual and transition report of foreign private issuers [Sections 13 or 15(d)] Acc-no: 0000950123-10-034985 (34 Act) Size: 2 MB	2010-04-15	001-14700 10750669
20-F	(Documents)	Annual and transition report of foreign private issuers [Sections 13 or 15(d)] Acc-no: 0001145549-09-000622 (34 Act) Size: 2 MB	2009-04-17	001-14700 09755004
20-F	(Documents)	Annual and transition report of foreign private issuers [Sections 13 or 15(d)] Acc-no: 0000950144-08-002830 (34 Act) Size: 2 MB	2008-04-15	001-14700 08755835

Figure 3.1.1 Downloading 20-F documents of queried company on SEC website

Downloaded 20-Fs were named systematically. Each htm file was named with this order: SIC code_Company Name_Date of File. For example, the first 20-F document of TSMC in Figure 3.1.1 was named as 3674_Taiwan_Semiconductor_Manufacturing_2012-04-13.

¹⁶ SIC code of industries can be looked up in this page: <http://www.sec.gov/info/edgar/siccodes.htm>

At the same time, data relevant to all queried Chinese corporations were input in the previous created Excel spreadsheet. The spreadsheet keeps record of company name, SIC code, and the amount of 20-F documents available of each corporation. This spreadsheet was then used to sample companies for creating the Business Reports Corpus (BRC).

The step for finding out Chinese corporations issuing securities in the U.S. was completed on November 18, 2010, and 89 companies from 35 different types of industry were recorded. To achieve representativeness and balance in creating the Business Reports Corpus (BRC), 35 corporations from 35 industries were sampled for collection according to alphabetical order of company names¹⁷. Those chosen corporations all maintain an accumulation of 20-F documents for at least 3 years. Location of companies (such as Taiwan, Hong Kong, or Mainland China) can be one of the considerations for the sampling process, which is neglected in the present thesis for the sake of simplicity in execution of research. For creating the Business Reports Corpus, 20-F documents for 3 consecutive years of the 35 chosen corporations were collected and a total of 105 texts were compiled. The complete list of the 35 companies is attached in Appendix 3.1.1.

¹⁷ Due to this adopted sampling method, the previously mentioned TSMC [台灣積體電路有限公司] was not collected into BRC. Under the category of “semiconductors & related devices”, annual reports of Advanced Semiconductor Engineering Incorporated [日月光半導體] were collected to BRC.

Unfortunately, these saved 20-F documents contain some flaws when surveyed with corpus software. As Figure 3.1.2 shows, some “error” codes such as `“` and `”` occur when downloaded htm files were under process of the software WordSmith Tools. One solution to this problem is to find out what entities, that is, certain special characters, are represented by the shown original codes. This solution is demonstrated on Figure 3.1.3.

Concord	
File Edit View Compute Settings Windows Help	
N	Concordance
1	confirmed to us that Circular 157 is not applicable to entities that qualify for <code>&#147;3-year</code> exemption plus 3-year half rate <code>&#148;</code> tax holiday as
2	for <code>&#147;3-year</code> exemption plus 3-year half rate <code>&#148;</code> tax holiday as <code>&#147;high</code> and new technology enterprises <code>&#148;</code> and are registered in
3	local tax authority, and given the fact that our subsidiaries that are enjoying <code>&#147;3-year</code> exemption plus 3-year half rate <code>&#148;</code> as <code>&#147;high</code> and
4	technology enterprise <code>&#148;</code> and is also in a tax holiday period, including <code>&#147;2-year</code> exemption plus 3-year half rate, <code>&#148;</code> <code>&#147;5-year</code>
5	tax holiday period, including <code>&#147;2-year</code> exemption plus 3-year half rate, <code>&#148;</code> <code>&#147;5-year</code> exemption plus 5-year half rate <code>&#148;</code> and other tax
6	period, including <code>&#147;2-year</code> exemption plus 3-year half rate, <code>&#148;</code> <code>&#147;5-year</code> exemption plus 5-year half rate <code>&#148;</code> and other tax
7	that are enjoying <code>&#147;3-year</code> exemption plus 3-year half rate <code>&#148;</code> as <code>&#147;high</code> and new technology enterprises <code>&#148;</code> are registered in the
8	Although the term <code>&#147;de facto</code> management bodies <code>&#148;</code> is defined as <code>&#147;management</code> bodies which has substantial and overall management
9	the operation, human resources, accounting and assets of the enterprise, <code>&#148;</code> the circumstances under which an enterprise <code>&#146;</code> s <code>&#147;de</code>

Figure 3.1.2 “Error” codes in html 20-F document

Figure 3.1.3 shows that, with the web-design software Dreamweaver, corresponding entities of `“` can be identified as the opening quotation “ by querying `“` in the dialogue window of searching function in the software. Remaining “error” codes were found out by querying ampersand `&`¹⁸ in the concordance function in the WordSmith Tools, and corresponding entities of those found codes were checked up with Dreamweaver by applying the same method repeatedly.

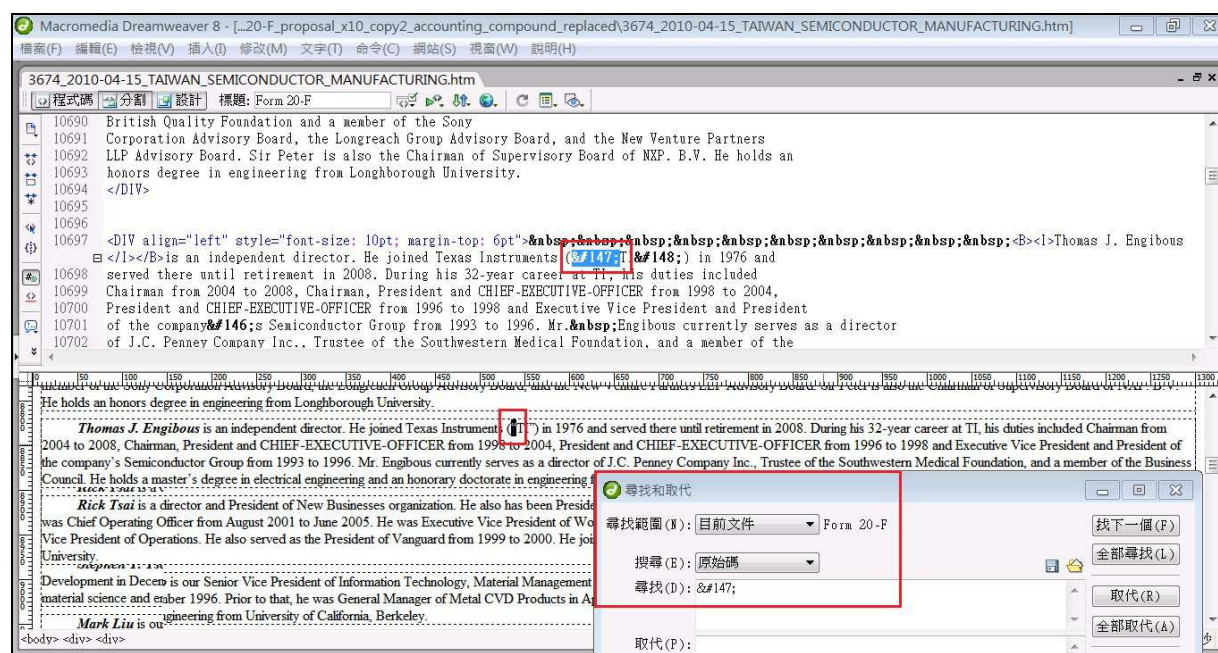


Figure 3.1.3 Querying for special characters with original codes in Dreamweaver

¹⁸¹⁸ For treatment of html codes, Professor Mike Scott recommends applying the ASCII (American Standard Code for Information Interchange) table (<http://www.asciitable.com>) (Scott, 2011b). Since original codes of all entities start with ampersand &, & was used for the querying process in the present thesis to locate special characters. More discussions about dealing with html entities can be seen on the on-line forum: <https://groups.google.com/group/wordsmithtools>

Those attained special characters as well as their original codes were together compiled in a txt file, as shown in Appendix 3.1.2. This txt file will help translating every “error” codes into their corresponding characters once this txt file is uploaded in WordSmith Tool.

Another necessary preparation in creating the Business Reports Corpus (BRC) is to identify technical compounds (Section 3.1.4). To do this, the website of Accounting Research and Development Foundation of the Republic of China [中華民國會計研究發展基金會] was logged on to download a piece of document called “Chinese-English translation of important Accounting Terms” [重要會計用語中英對照]¹⁹, which is compiled by professors and practitioners in the accounting field. This document contains 1,558 entries of important terms presented in bilingual format, as can be seen from one sample page presented in Appendix 3.1.3. As already mentioned in Introduction, the present project aims to locate ESP real content that does not involve subject knowledge; those entries of accounting compounds, each of which possesses complete technical meaning, are to be processed as separate single items and excluded with corpus software.

As introduced in Section 3.1.4, to make compounds as single items, all entries in the Chinese-English Translation of Important Accounting Terms [重要會計用語中英對照] and their hyphenated equivalents were all input into the batch replacer tool in the text processing software Useful File Utilities (ReplSoft.com, 2010). With target texts uploaded, all multiple lexical units such as *Amount Recoverable* will be replaced with its

¹⁹ The specialized wordlist can be seen in the web page www.ardf.org.tw/html/tifrs2.html

hyphenated form *Amount-Recoverable*, regardless of their cases. Snapshots of this hyphenation process are attached in Appendix 3.1.4.

After the hyphenation process, the creation of the Business Reports Corpus (BRC) is now successfully completed. The compiled BRC maintains 9,642,956 word tokens distributed across 105 texts. If all re-occurrences of the same item are counted only once, the 9,642,956 word tokens are in fact composed with 28,727 different word types. Basic statistic information of BRC, together with that of Brown Corpus, is presented in Table 3.1.1 below.

Table 3.1.1 Basic information of the Business Reports Corpus and Brown Corpus

	Total No. of Texts	Total No. of Word Types	Total No. of Word Tokens
Business Reports Corpus	105	28,727	9,642,956
Brown Corpus	500	48,846	1,054,599

3.2 Extracting common words

Having created the Business Reports Corpus (BRC) and assigned Brown Corpus as the reference corpus, this study proceeds to extract common words of interest. In section 2.2.5, common words in this project have been operationally defined as words which

- (a) involve no technical words with specialized meanings, and
- (b) occur with unusual frequency, and
- (c) match words in Jeng et al.'s Senior High English Wordlist for Reference (SHEWR)

[高中英文參考詞彙表]

The resulted words will be assorted according to degree of coverage in the target corpus. These operational definitions of common words will be applied in the procedure reported below.

3.2.1 Excluding technical words and compounds. The exclusion procedure requires in advance that the hyphenated accounting compounds to be uploaded in the stoplist function in the setting of WordSmith Tools. Part of this stoplist is displayed in Appendix 3.2.1. Accounting-specific terms in this stoplist, either single words or compounds, will be ignored in the following computation. Snapshots of this procedure will be reported in Section 3.2.5.

3.2.2 Keyword, words with unusually frequency. The identification of common words begins with computing keyness of words in BRC by comparing BRC to Brown Corpus via statistic tool. As been mentioned previously (Section 2.2.3), keyness demonstrates degree of unusual frequency of words in quantitative manner, and computation of keyness requires a smaller corpus to be compared to a larger one (Section

3.1.2). As the size of BRC (9,642,956 running words) is overwhelmingly bigger than Brown Corpus (1,054,599 running words), we cannot compare BRC to Brown Corpus. The alternative procedure for computing keyness would be comparing each of the 105 texts of BRC to the whole Brown Corpus respectively. With this procedure, 105 keyword lists belonging to the 35 corporations can be obtained with information of keyness of compared words. In WordSmith Tool, computation of keyness can be processed with either chi-squared or log likelihood²⁰ is adopted in the present thesis since it excels in processing small-size collection of texts (Dunning, 1993). Results of keyness computation will be reported in Section 3.2.5.

3.2.3 Matching words in Senior High English Wordlist for Reference. Now that we are expected to have obtained words with the highest keyness, we proceed to identify words in the results that are also members of the Senior High English Wordlist

²⁰ According to Scott (2011b), the formula of Log Likelihood is

$$2 \text{ times } (a \ln a + b \ln b + c \ln c + d \ln d$$

$$- (a+b) \ln (a+b)$$

$$- (a+c) \ln (a+c)$$

$$- (b+d) \ln (b+d)$$

$$- (c+d) \ln (c+d)$$

$$+ (a+b+c+d) \ln (a+b+c+d))$$

where

a = joint frequency

b = frequency of word 1

c = frequency of word 2

d := frequency of pairs involving neither w1 nor w2

and "Ln" means Natural Logarithm.

for Reference (SHEWR) [高中英文參考詞彙表]. The step reported in this section reflects pedagogical concern of the present thesis, which aims to bridge the gap between ESP and English Language Teaching (ELT) in public school of Taiwan (Section 2.2.5). Prior to identifying SHEWR words from the previously resulted words, all 6,480 words across the 6 levels in SHEWR have to be enlisted in a text file to be loaded as a match list in WordSmith Tools. Part of this text file is attached in Appendix 3.2.2 for reference. Detailed procedure and results of matching SHEWR words to the previously resulted keywords will be demonstrated in the Section 3.2.5.

3.2.4 Key-keyword, keywords sorted according to text coverage. Now words are going to be assorted in terms of their coverage among texts. Degree of coverage of a word is investigated by measuring how many texts the word occurs. Since BRC contains 105 files in total, words with the widest coverage are anticipated to occur across all those 105 texts. Computation of word coverage is designed as the “Key-Keywords” function in WordSmith Tools (Section 2.2.3). Procedure and results of key-keywords sorting will be presented in Section 3.2.5.

3.2.5 Procedure for extracting common words with WordSmith Tools. This section gives detailed reports on procedure for identifying common words of interest.

Before launching on any computation, we first adjust settings behind WordSmith Tools for later process. Figure 3.2.1 displays the window for adjusting settings in WordSmith Tools, in which “Mark-up to ignore” designates that word strings less than 200 words within the bracket for hyper-text-markup-language (HTML) are to be ignored in later computation, and “Custom settings” assigns html files for the type of text for processing. Previously made txt file of html entities (Appendix 3.1.2) is uploaded in the interface of “Entity File”; as the figure shows, there are 66 entries of special characters to be translated with WordSmith Tools.

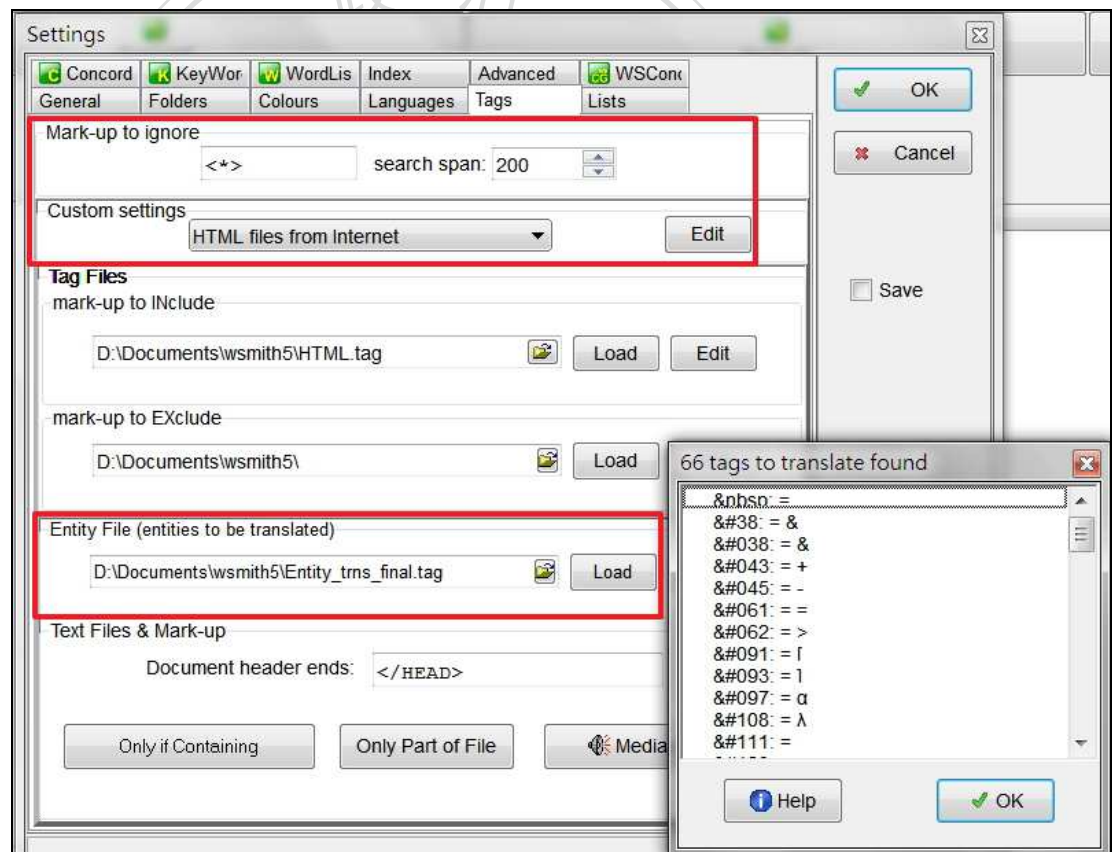


Figure 3.2.1 Adjusting setting for html brackets and special characters

As our anticipated common words do not involve technical terms, we upload technical words and compounds in the Chinese-English Translation of Important Accounting Terms [重要會計用語中英對照] (Accounting Research and Development Foundation of the Republic of China, 2011) (Section 3.1.4) as stop list in WordSmith Tools. As Figure 3.2.2 shows, 1,620²¹ entries of accounting terms were uploaded in the interface of “stop list”; by so doing, words with specialized meanings will be ignored in subsequent computation.



²¹ There are 1,558 entries of accounting terms in the Chinese-English Translation of Important Accounting Terms [重要會計用語中英對照], but the stoplist ends with 1,620 entries due to that some single terms could be presented in multiple forms; for example, “accumulated (amortization, interest, profit or loss)” was keyed in the stoplist as three separate entries.

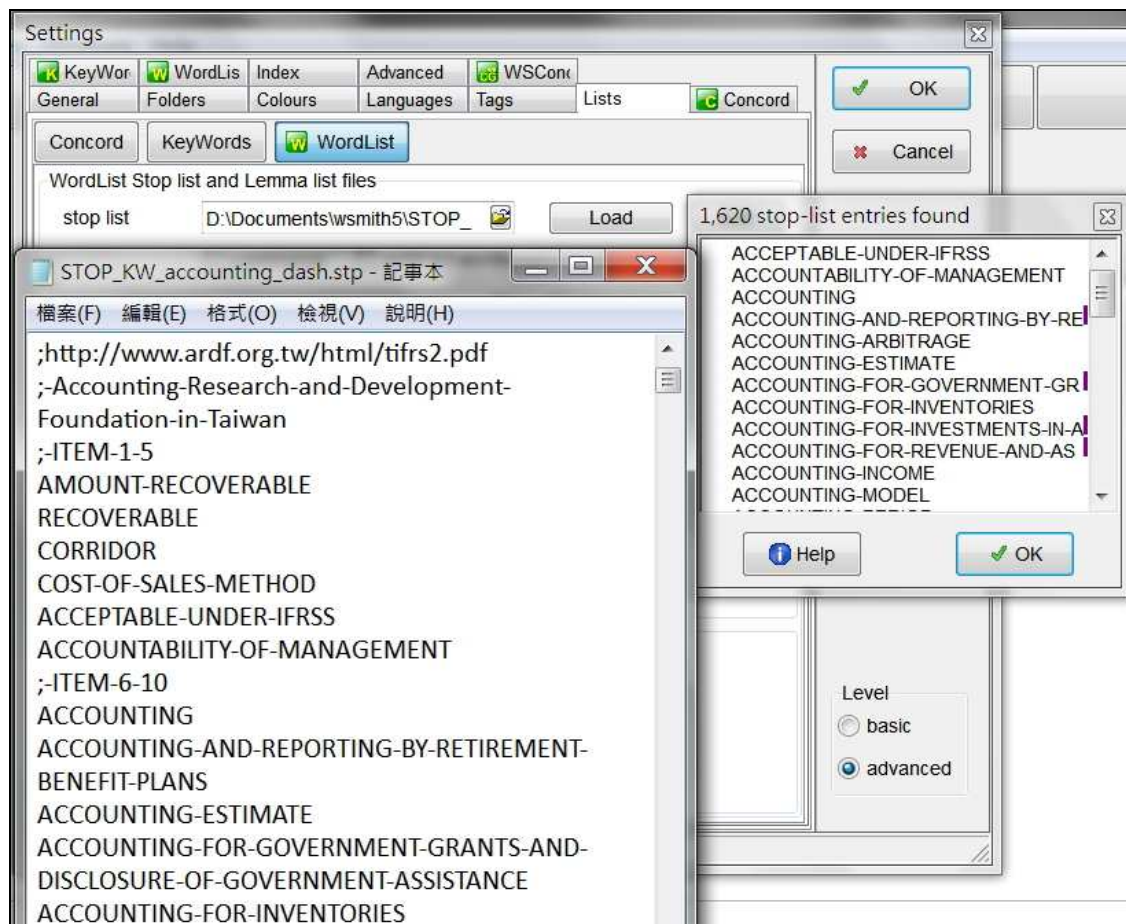


Figure 3.2.2 Uploading accounting compounds as stop list

One other setting to be made is about computing keyness of words. Figure 3.2.3 demonstrates that in the setting for KeyWords, the wordlist of Brown Corpus²² was designated in the “reference corpus” interface, and log likelihood was adopted as the statistic measure, with p value set as 0.000001²³. Among resulted keywords, only words that occur with more than 3 occurrences in each text of BRC will be displayed.

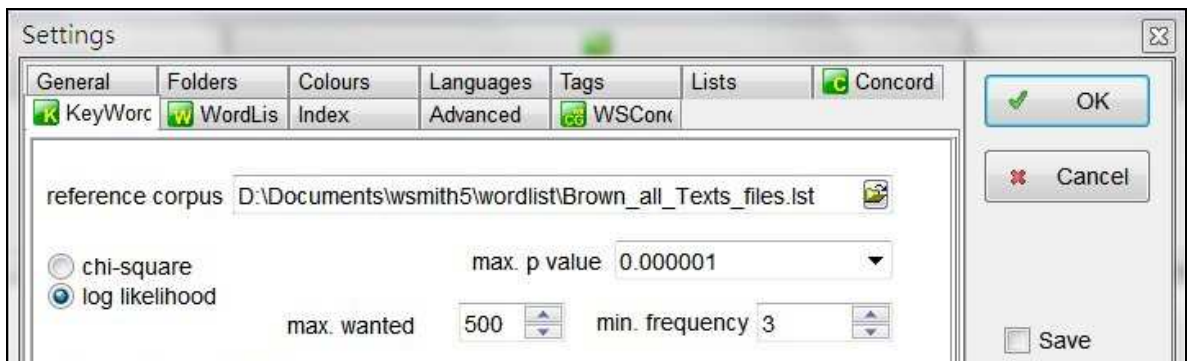


Figure 3.2.3 Adjusting setting for computing keywords

After the above necessary adjustments in WordSmith Settings, the process of extracting words of interest was launched by uploading all 105 texts of 35 companies and creating 105 wordlists for each text. Then the 105 wordlists were compared to the wordlist of Brown Corpus respectively with log likelihood to establish 105 keyword lists.

²² The wordlist of Brown Corpus was created in advance via the WordList function of WordSmith Tools.

²³ This means that result of this computation could be wrong with a risk of one in a million (Scott, 2011b).

With the 105 keyword lists, all keywords (words with unusual frequency computed with loglikelihood) are to be assorted in terms of coverage of texts. Figure 3.2.4 shows part of the key-keyword list, in which keywords were descended according to number of texts those keywords occur in. The complete list of key-keywords is presented in Appendix 4.1.1 and Appendix 4.1.2

N	KW	Texts	%	Overall Freq.
1		105	100.00	1,087,423
2	ANNUAL	105	100.00	9,168
3	APPLICABLE	105	100.00	5,651
4	CHINA	105	100.00	26,427
5	DECEMBER	105	100.00	30,234
6	EXCHANGE	105	100.00	11,302
7	FINANCIAL	105	100.00	14,944
8	INCLUDING	105	100.00	11,142
9	OPERATING	105	100.00	10,953
10	OPERATIONS	105	100.00	12,592
11	OTHER	105	100.00	33,559
12	RISKS	105	100.00	3,085
13	SHARES	105	100.00	19,630
14	SIGNIFICANT	105	100.00	7,013
15	ASSETS	104	99.00	12,932
16	BASED	104	99.00	8,726
17	COMPANIES	104	99.00	8,379
18	CONSOLIDATED	104	99.00	5,934
19	EMPLOYEES	104	99.00	6,085
20	HOLDERS	104	99.00	4,406

Figure 3.2.4 Key-keywords assorted based on degree of text coverage

Based on this key-keyword list, we proceed to identify words that match words in the Senior High English Wordlist for Reference (SHEWR) [高中英文參考詞彙表] (Jeng, et al. [鄭恆雄等] 2002). Prior to this matching work, SHEWR words must be uploaded as match list in WordSmith Tools. As shown in Figure 3.2.5, all 6,480²⁴ words in SHEWR were uploaded in the interface “KeyWords Match List.” Then words in our previous key-keywords list that match SHEWR can be identified.

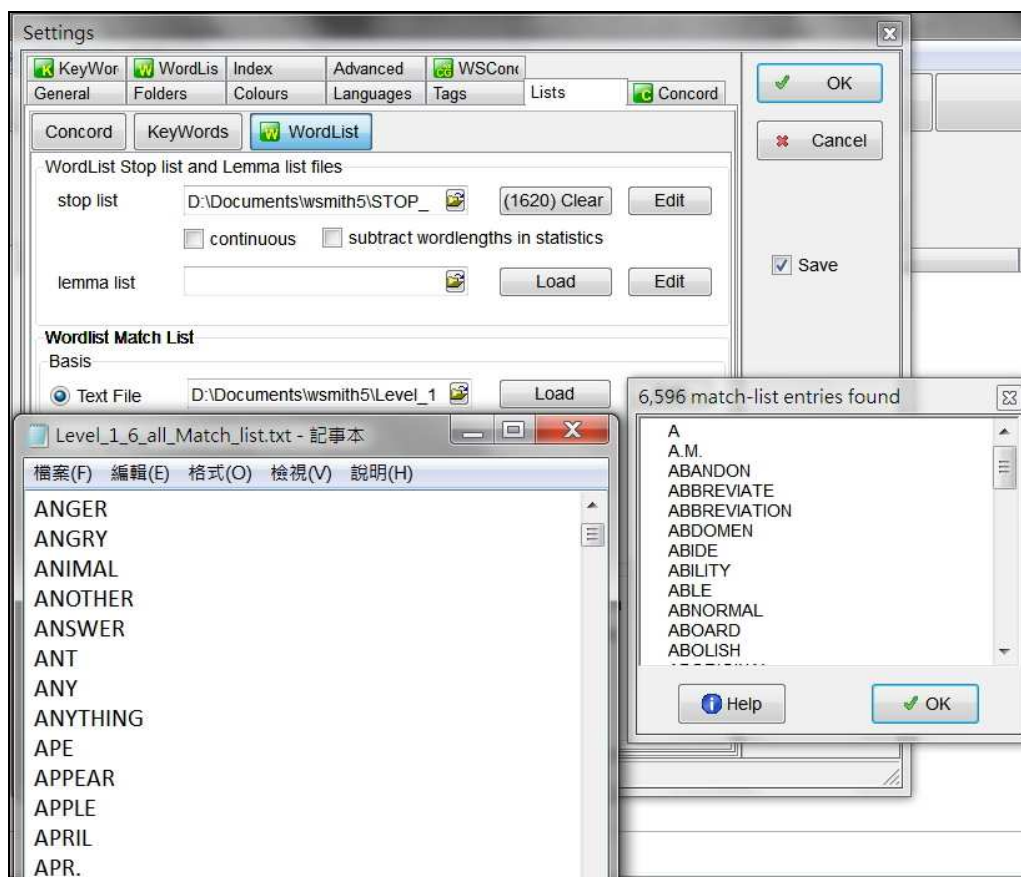


Figure 3.2.5 Adjusting setting for Match List with the Senior High English Wordlist for Reference

²⁴ It can be seen from Figure 3.2.5 that the match list possesses more words (6,596) than 6,480, which is resulted from some duplicate forms of words appeared in SHEWR, such as *Apr.* and *April.*

Figure 3.2.6 demonstrates result of the matching action. There were 733 words in the key-keywords list that match SHEWR [高中英文參考詞彙表]. This result will be discussed in detail in Chapter 4.

N	KW	Texts	%	Overall Freq.
1		105	100.00	4,087,423
2	ANNUAL	105	100.00	9,168
3	APPLICABLE	105	100.00	5,651
4	CHINA	105	100.00	26,427
5	DECEMBER	105	100.00	30,234
6	EXCHANGE	105	100.00	11,302
7	FINANCIAL	105	100.00	14,944
8	INCLUDING	105	100.00	11,142
9	OPERATING	105	100.00	10,953
10	OPERATIONS	105	100.00	12,592
11	OTHER	105	100.00	33,559
12	RISKS	105	100.00	3,085
13	SHARES	105	100.00	19,630
14	SIGNIFICANT	105	100.00	7,013
15	ASSETS	104	99.00	12,932
16	BASED	104	99.00	8,726
17	COMPANIES	104	99.00	8,379
18	CONSOLIDATED	104	99.00	5,934
19	EMPLOYEES	104	99.00	6,085
20	HOLDERS	104	99.00	4,406

Figure 3.2.6 Immediate results of the matching process

Figure 3.2.7 displays key-keywords that match SHEWR [高中英文參考詞彙表] words. As can be seen, there are 9 words among the extracted common words that have the widest text coverage (105 texts). Due to space constraints, only the four adjectives *annual*, *applicable*, *financial*, and *significant* will be employed for identifying formulaic language.

N	KW	Texts	%	Overall Freq.
1	ANNUAL	105	100.00	9,168
2	APPLICABLE	105	100.00	5,651
3	CHINA	105	100.00	26,427
4	DECEMBER	105	100.00	30,234
5	EXCHANGE	105	100.00	11,302
6	FINANCIAL	105	100.00	14,944
7	INCLUDING	105	100.00	11,142
8	OTHER	105	100.00	33,559
9	SIGNIFICANT	105	100.00	7,013
10	SUBJECT	104	99.00	10,560
11	FOREIGN	103	98.00	11,109
12	INTERNAL	103	98.00	5,003
13	MARKET	103	98.00	9,126
14	RELEVANT	103	98.00	5,000
15	INFORMATION	102	97.00	12,089
16	OR	102	97.00	76,832
17	OUR	102	97.00	127,492
18	ACCORDANCE	101	96.00	4,531
19	COMPANY	101	96.00	35,198
20	CORPORATE	101	96.00	4,387
21	NET	101	96.00	14,272
22	REPORT	101	96.00	7,680
23	DUE	100	95.00	8,366
24	JANUARY	100	95.00	7,435
25	TOTAL	100	95.00	14,764
26	DATE	99	94.00	7,116

KW database associates filenames notes

Figure 3.2.7 Final results of the matching process

3.3 Locating formulaic language from different genres

The previous sections have reported on preparation of a homemade Business Reports corpus (Section 3.1) and process of extracting common words of interest (Section 3.2). Based on those extracted words, this section continues to identify formulaic language composed by common words.

Because the proposed Research Questions 2 (Section 2.5) asks how formulaic language varies across genres, we now turn to the issue of selecting genres, in which formulaic language is going to be identified.

3.3.1 Selecting genres for comparing formulaic language. To proceed further comparison among genres, two subdivisions of the Brown Corpus, Subdivision A and Subdivision H, are chosen by this study. Subdivision A of Brown Corpus (BC-A) is composed by the genre of reportage. It contains 44 pieces news writing, each with 2000+ words, with topics covering from politics, sports, culture, and so forth (complete list of contents is attached in Appendix 3.3.1). Subdivision A of Brown Corpus contains 92,022 running words in total. Another adopted subcorpus of Brown Corpus, the Subdivision H (BC-H), labeled as miscellaneous, is composed of 30 pieces of articles (each with 2000+ words), most of which comes with the form of official document such as government document, foundation reports and industry reports, and so forth (complete list is presented in Appendix 3.3.2). There are 63,870 running words in Subdivision H of Brown Corpus. Table 3.3.1 below displays basic statistics of the previously created Business Reports Corpus (BRC), Subdivision A and Subdivision H of Brown Corpus.

Table 3.3.1 Basic information of Business Reports Corpus and the two subdivisions of Brown Corpus for investigation

	Total No. of Articles	Total No. of Word Types	Total No. of Word Tokens
Business Reports Corpus	105	28,727	9,642,956
Subdivision A of Brown Corpus (BC-A, texts of reportage)	44	13,141	92,022
Subdivision H of Brown Corpus (BC-H, official documents)	30	7,554	63,870

The reason for choosing Subdivision A and Subdivision H of Brown Corpus is that, these two sub-corpora share some similarities with Business Reports Corpus (BRC) respectively. BRC, composed of annual business reports, tell facts as Subdivision A do, but contrasts with Subdivision A in target audience; business reports in BRC are prepared specifically for the investment public, while Subdivision A of Brown Corpus, reportage articles in press, are for the general public without specific purposes for information. In other words, BRC and Subdivision A differ in types of audience for communication. If formulaic language identified with identical common words in BRC compared to those in Subdivision A of Brown Corpus, the similarities or diversities of formulaic language could suggest that composition of formulaic language correlates with types of genre, where genre is defined from the aspect of expected audience of communication in written mode.

On the other hand, BRC and Subdivision H of Brown Corpus (mainly government documents) share similarities in format, in that they both come as official documents with clear structure and headed paragraphs. However, BRC differs from Subdivision H in type of audience; texts in BRC are prepared specifically for the investment public, while

those in Subdivision H are mostly articles for parties or groups engaged in government affairs. For example, texts in Subdivision H cover an array of governmental functions such as federal aids (Sample H01), tax law (Sample H05), and medicine policy (Sample H10), and so forth. If comparison of formulaic language in BRC and those in Subdivision H shows variance, this could be an indication of how formulaic language, composed with identical common words, patterns in relation to different types of genre that are defined from the perspective of anticipated audience.

Since genres for comparison have been decided, we now continue to discuss in what aspects identified formulaic language are to be examined.

3.3.2 Syntagmatic variation and paradigmatic variation of formulaic language.

Variations of formulaic language across genres are going to be investigated in the perspective of syntagmatic and paradigmatic variation. As been demonstrated in Section 2.3.3, syntagmatic variation refers to that a sequence of words is open to modifying items inserted in the original string (*stuck in a traffic jam & stuck in a **hellish** traffic jam*), while paradigmatic variation refers to that one word in a string could be alternated with other words at the same slot (***stuck** in a traffic jam & **caught** in a traffic jam*). These variations can only be observed from strings of language, hence concordance lines of formulaic language need to be identified in advance for this observation. The following section demonstrates application of the corpus software AntConc (Anthony, 2010) to cumulatively identify collocates of target common words and concordance lines composed with those collocates as well.

3.3.3 Procedure for identifying formulaic language with AntConc. In the end of Section 3.2.5, it is mentioned that in the present study, only the four common words

annual, *applicable*, *financial*, and *significant* are kept for subsequent investigation. This section takes the word *applicable* as example to display the procedure of identifying sequence of formulaic language composed with *applicable*.

The main purpose of the procedure demonstrated here is to extract collocates cumulatively of the query item with a 4:4 span (a 9-word-span). With AntConc, the first step is to input the word *applicable* in the query box with the span set as 4:4 under the function of Collocates. Figure 3.3.1 shows results of this query in BRC: *applicable* has 5,377 hits; *nbsb* has 2,307 hits; *the* has 2,045 hits. These results mean that within the span of 4 word on each side of *applicable*, *nbsb* (the html codes for non-breaking space)²⁵ is the most frequent collocate of *applicable* and *the* is the second frequent one of *applicable*. Since *nbsb* is not a valid word, *the* was adopted for the subsequent work.

²⁵ To the researcher's knowledge, AntConc doesn't have the function for translating html entities as WordSmith Tools does.

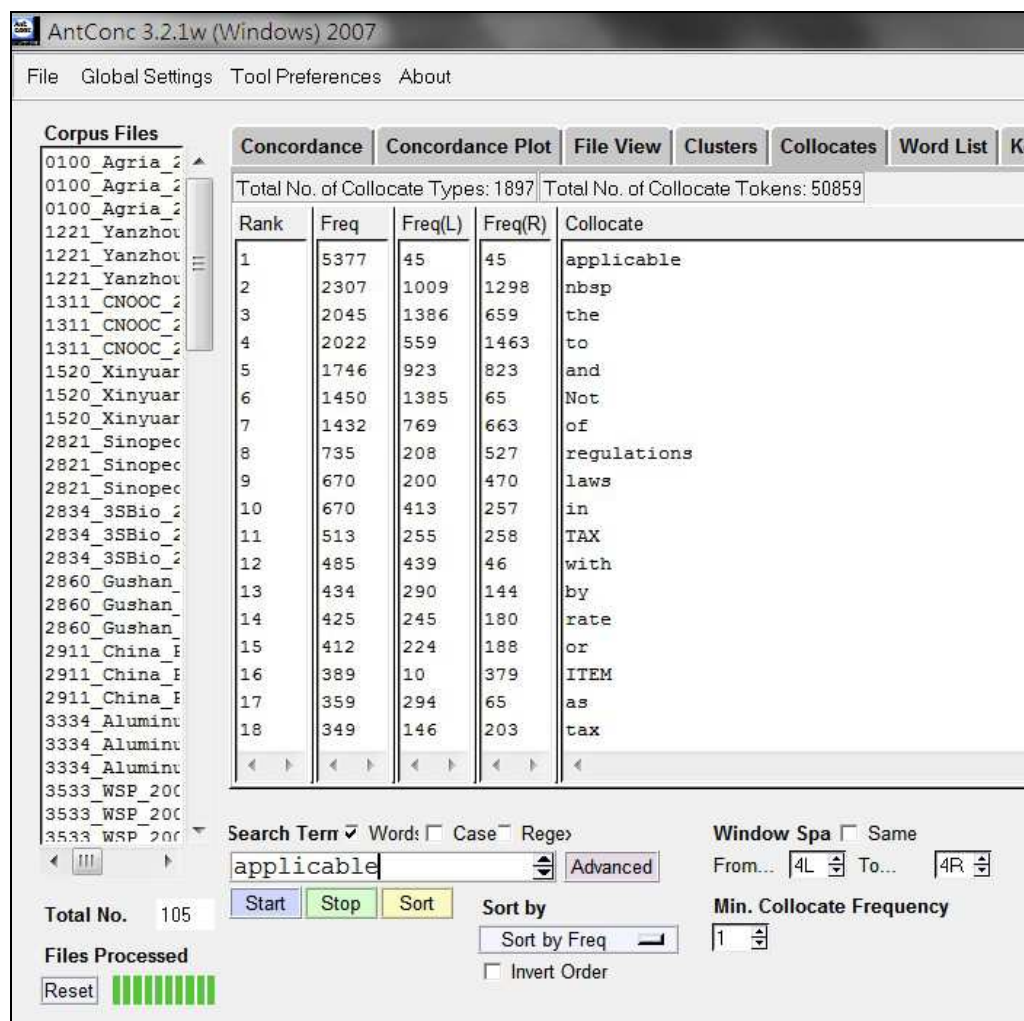


Figure 3.3.1 Querying for collocates of *applicable* in Business Reports Corpus with AntConc

Figure 3.3.2 displays that the Advance Search function of AntConc was applied, in which the main query remains *applicable* with the context word input as *the* within the Context Horizon set as 4:4.

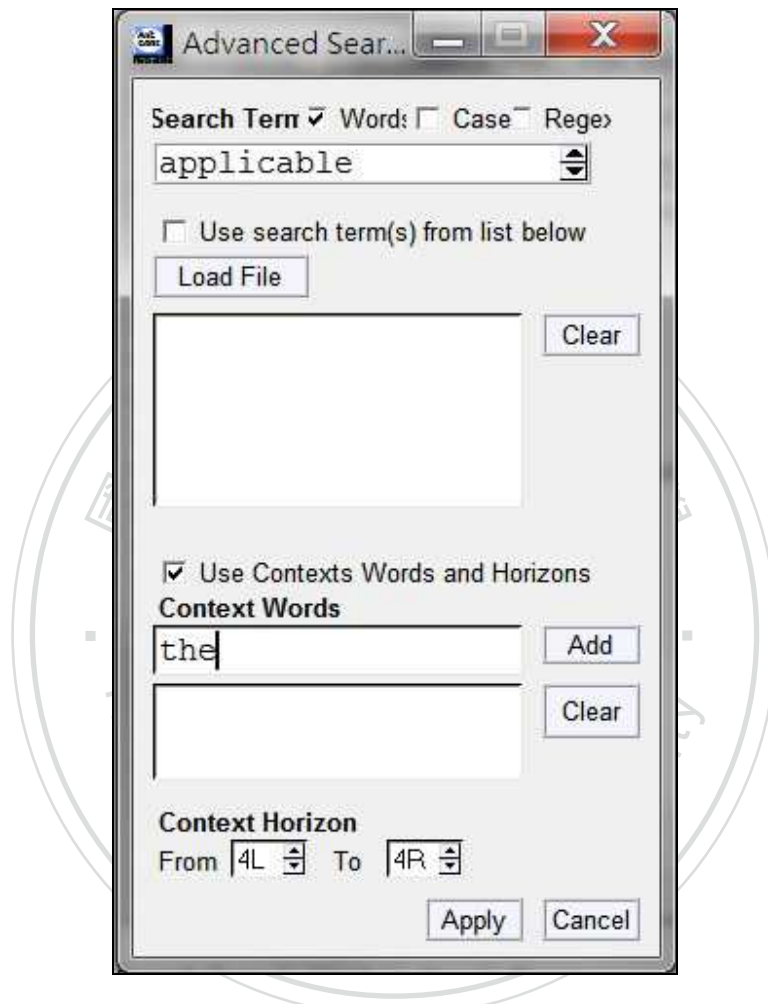


Figure 3.3.2 Querying *applicable* together with a context word *the*

Result of the previous step is displayed in Figure 3.3.3, in which the most frequent word is *the* (2,045 hits), the second frequent one is *applicable* (1,807 hits), and *to* is the third frequent word (695 hits). The word *to* is adopted in the subsequent work with the same procedure, which was repeated until the frequency of extracted collocates is below 5.²⁶

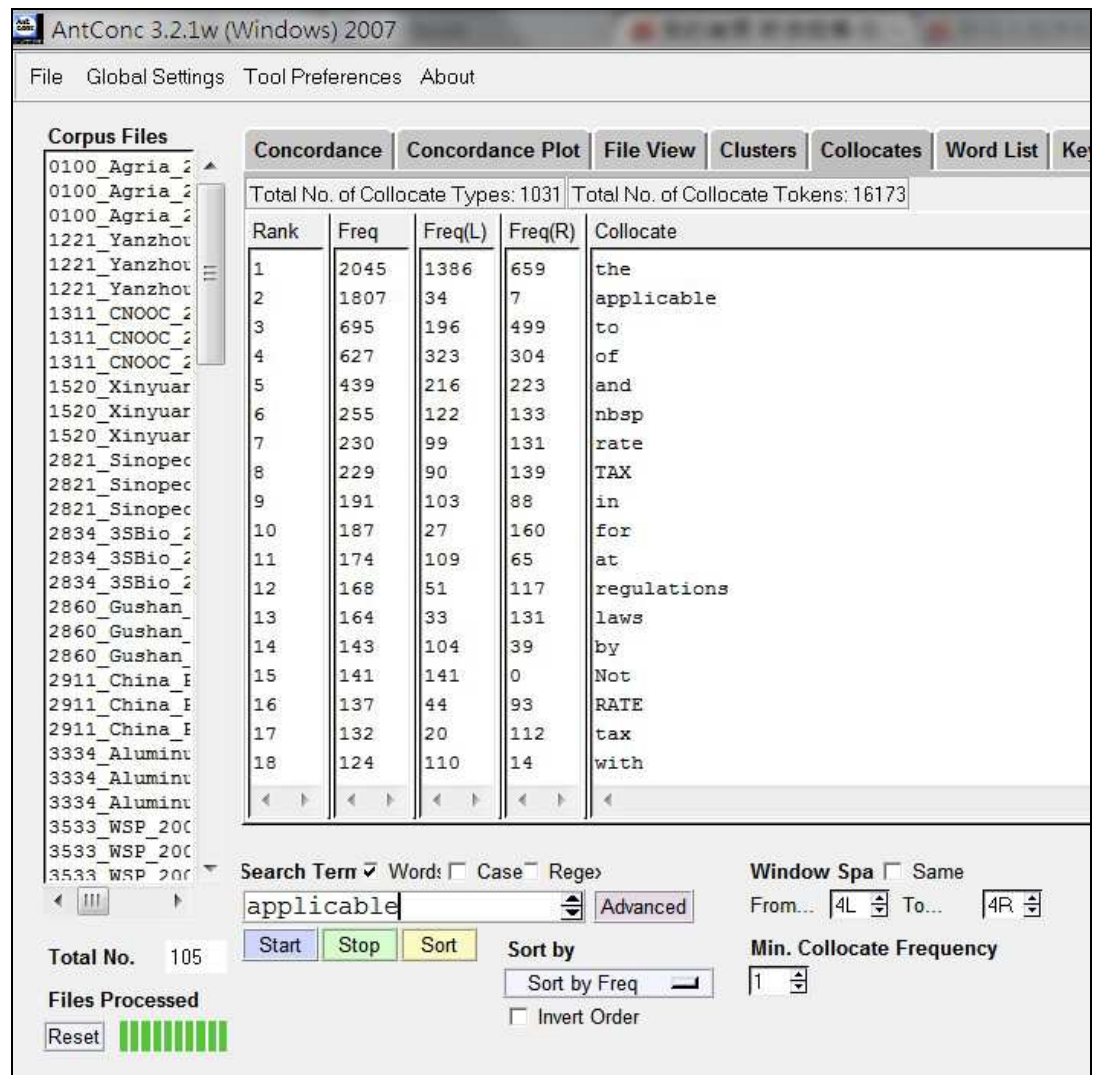


Figure 3.3.3 Results of querying for collocates of *applicable* with its context word *the*

²⁶ The criterion of frequency of collocates follows Danielsson's (2007) original method, as also been introduced in Section 2.3.3.

Figure 3.3.4 demonstrates the final result of the query of *applicable*, in which the collocates of *applicable* with frequency above 5 are *to*, *the*, *texts*, *regulations*, *of*, *laws*, *for* and *and*.

AntConc 3.2.1w (Windows) 2007

File Global Settings Tool Preferences About

Corpus Files

0100_Agria_2
0100_Agria_2
0100_Agria_2
1221_Yanzhou
1221_Yanzhou
1311_CNOOC_2
1311_CNOOC_2
1311_CNOOC_2
1520_Xinyuar
1520_Xinyuar
1520_Xinyuar
2821_Sinopec
2821_Sinopec
2821_Sinopec
2834_3SBio_2
2834_3SBio_2
2834_3SBio_2
2860_Gushan
2860_Gushan
2860_Gushan
2911_China_E
2911_China_E
2911_China_E
3334_Aluminu
3334_Aluminu
3334_Aluminu
3533_WSP_200
3533_WSP_200
3533_WSP_200

Concordance Concordance Plot File View Clusters Collocates Word List Ke

Total No. of Collocate Types: 9 Total No. of Collocate Tokens: 72

Rank	Freq	Freq(L)	Freq(R)	Collocate
1	8	8	0	to
2	8	8	0	the
3	8	8	0	texts
4	8	0	8	regulations
5	8	8	0	of
6	8	0	8	laws
7	8	0	8	for
8	8	0	0	applicable
9	8	0	8	and

Search Term Word: Case: Regexp:

applicable Advanced

Window Spacing Same

From... 4L To... 4R

Total No. 105

Files Processed

Reset

Start Stop Sort

Sort by

Sort by Freq

Invert Order

Min. Collocate Frequency 1

Figure 3.3.4 Final results of finding collocates of *applicable* with the method of cumulative frequency

Under the function of Concordance, concordance lines of *applicable* and all of its collocates can be identified, as Figure 3.3.5 shows. These concordance lines can be extracted in complete sentence respectively for further investigation.

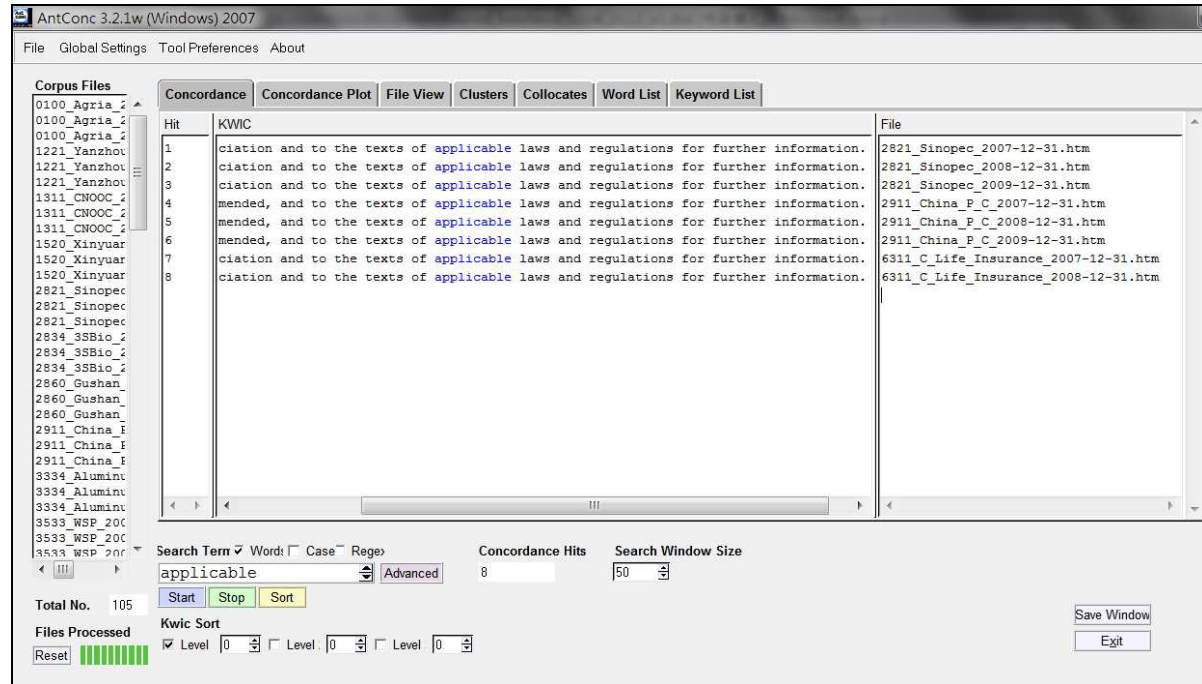


Figure 3.3.5 Concordance lines of *applicable* together with its extracted collocates

It should be noted that when the method of cumulative frequency as applied in Subdivision A and Subdivision H of Brown Corpus, the criterion of frequency hits of collocates was adjusted from 5 to 2 due to comparatively small size of the two sub-corpora.

3.4 Flowchart of research design

The following figure summaries complete procedure of the research design in the present study.

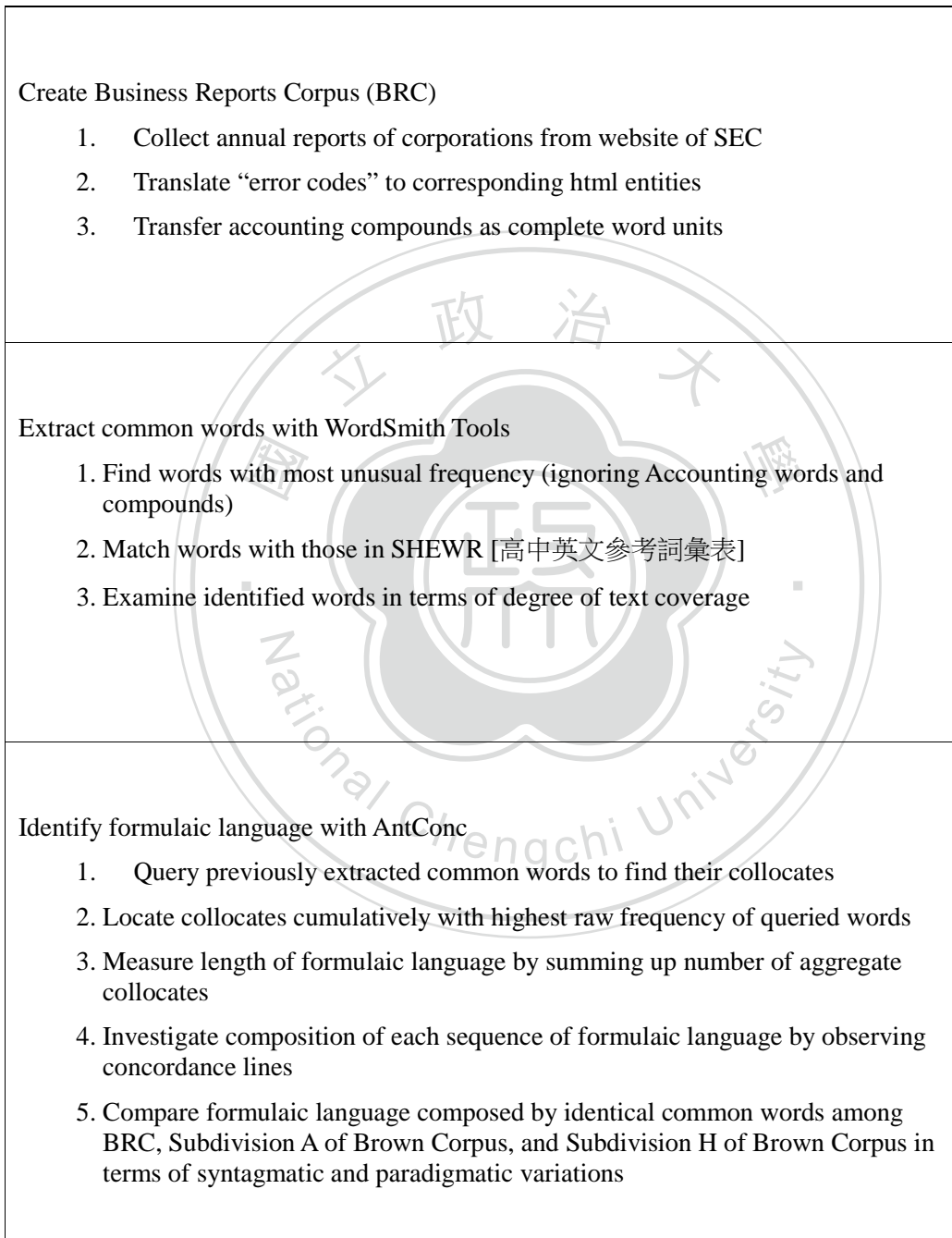


Figure 3.4.1 Flowchart of research design

Chapter 4 Results and Discussion

The previous chapter has reported the research design for compiling a Business Reports Corpus (BRC), extracting common words, and identifying formulaic language composed with extracted words. This procedure was designed for answering the research questions proposed in Section 2.5:

Research Question (1)

To what extent do common words occur in company annual reports?

Research Question (2)

How formulaic language composed with identical common words vary across genres?

Research Question (2) has been refined to explore formulaic language from two different perspectives:

Research Question (2-1)

Do the lengths of identified formulaic language vary across genres?

Research Question (2-2)

Does composition of the identified formulaic language allow syntagmatic variation or paradigmatic variation?

This chapter demonstrates results of research in an attempt to answer the above research questions.

4.1 Extracted Common words

This section gives results to answer Research Question (1) that asks for how common words are distributed in company annual reports. Figure 4.1 below displays

results of extracted common words. There are 733 words²⁷ in Business Reports Corpus (BRC) that have unusual frequency when compared to their equivalents in Brown Corpus, match words in the Senior High English Wordlist for Reference [高中英文參考詞彙表] (Jeng, et al. [鄭恆雄等] 2002), and does not involve technical words or compounds in the accounting field. Words in this list of key-keywords are displayed in descending order of word coverage among texts; the first nine words from *annual* to *significant* have the widest coverage among all 105 texts.



²⁷ Complete lists of extracted words are attached as Appendix 4.1.1 and 4.1.2. These two lists may be helpful for readers who are interested in whether difficulty level of word is connected with keyness or text coverage of word.

KeyWords					
File Edit View Compute Settings Windows Help					
N	KW	Texts	%	Overall Freq.	
1	ANNUAL	105	100.00	9,168	
2	APPLICABLE	105	100.00	5,651	
3	CHINA	105	100.00	26,427	
4	DECEMBER	105	100.00	30,234	
5	EXCHANGE	105	100.00	11,302	
6	FINANCIAL	105	100.00	14,944	
7	INCLUDING	105	100.00	11,142	
8	OTHER	105	100.00	33,559	
9	SIGNIFICANT	105	100.00	7,013	
10	SUBJECT	104	99.00	10,560	
11	FOREIGN	103	98.00	11,109	
12	INTERNAL	103	98.00	5,003	
13	MARKET	103	98.00	9,126	
14	RELEVANT	103	98.00	5,000	
15	INFORMATION	102	97.00	12,089	
16	OR	102	97.00	76,832	
17	OUR	102	97.00	127,492	
18	ACCORDANCE	101	96.00	4,531	
19	COMPANY	101	96.00	35,198	
20	CORPORATE	101	96.00	4,387	
21	NET	101	96.00	14,272	
22	REPORT	101	96.00	7,680	
23	DUE	100	95.00	8,366	
24	JANUARY	100	95.00	7,435	
25	TOTAL	100	95.00	14,764	
26	DATE	99	94.00	7,116	

KW database associates filenames notes

Figure 4.1.1 Results of extracted common words assorted based on text coverage

Table 4.1.1 Distribution of extracted key-keywords in BRC

Key-keywords in BRC (105 texts)	No. of words	Percentage (%)
105 texts (100% of 105 texts)	9	1.2
85-104 texts (81-99% of 105 texts)	50	6.8
65-84 texts (61-80% of 105 texts)	27	3.7
44-64 texts (41-60% of 105 texts)	44	6.0
23-43 texts (21-40% of 105 texts)	89	12.1
1-22 texts (1-20% of 105 texts)	514	70.1
Total No. of key-keywords in Business Reports Corpus	733	100.0

Table 4.1.1 presents how the 733 common words are distributed in Business Reports Corpus (BRC). Distribution of occurrence of the 733 words in BRC can be observed in the first and second column of this table, in which the 733 words in BRC are arranged based on their occurrence with different proportion in the total 105 texts. For instance, there are 9 of the extracted common words (1.2% of 733 words) occurring in all the 105 texts.

At this point, the above research results respond to Research Question 1, which enquires about how common words are distributed in company annual reports. The extracted 733 words, which comply with the operational definition of common words of this study (Section 2.2.5), have occurrence in the homemade Business Reports Corpus.

About 70% of the 733 words (514 words) has occurrence in less than 22 texts of the 105-text collection and only 1.2% of the 733 words (9 words) has the widest coverage across the collection. Since BRC is compiled with annual reports from 35 different industries evenly (Section 3.1.3 & 3.1.5), those figures representing distribution of the 733 words indicate the importance of the extracted common words regarding their degree of practicality in the business genre across various industries. Table 4.1.1 guides ESP practitioners looking for useful vocabulary for learners with business purpose.

As for further scrutinizing the overall practicality of SHEWR [高中英文參考詞彙表], which is the present generally accepted common-word list in Taiwan, Table 4.1.2 is to be consulted.

Table 4.1.2 Distribution of extracted key-keywords in SHEWR [高中英文參考詞彙表]

Key-keywords in BRC (105 texts)	No. of words	Percentage (%)
Total No. of key-keywords in BRC	733	11.3
SHEWR words that neither bear outstandingness nor occur in BRC	5,747	88.7
Total of SHEWR words	6,480	100.0

Table 4.1.2 tabulates proportion of extracted common words as well as SHEWR [高中英文參考詞彙表] words that neither bear significance in BRC nor have occurrence in BRC. The very bottom of second column of this table presents the total number of words in SHEWR (6,480); the difference between 6,480 and 733 (5,747) is the amount of

SHEWR words that neither bear outstandingness nor occur in BRC. As can be seen, the 733 words bear a proportion of 11.3% in SHEWR, indicating that there are about one-tenth of SHEWR words consistently occur in Business Reports Corpus. Our attention will be focused on the figure 11.3% to discuss practicality of SHEWR. The figure 11.3% cannot be used to claim high practicality of SHEWR words in business genre if viewed with the conventional computation in which practicality of wordlist is evaluated by comparing raw frequency of occurring words to the whole word collection²⁸. However, a strict interpretation needs to be placed in the figure 11.3% (733 words) since the concept of unusual frequency was applied in the extraction process.

With the research design of this thesis, the 733 words were identified with the notion of unusual frequency. Hence, the 733 words escape the interference of high-frequency words (such as grammatical words) by comparing each 105 business texts to the Brown Corpus; these extracted words are words with unusual high frequency in BRC (presenting business English) than in Brown Corpus (presenting general English), which means that readers of these business texts are exposed to the 733 words with higher probability than in common situation. With this interpretation, the figure 11.3% makes a conservative claim that the 733 words, though bearing low proportion in the complete SHEWR [高中英文參考詞彙表], compose a vocabulary bank compulsory for English learners who attempt to advance to English for Business Purpose (EBP) on the basis of English for General Purpose (EGP). The figure of 11.3% (733 words) should not be used to conclude a low practicality of SHEWR words in business context; on the

²⁸ In Coxhead's review (2000, pp. 213-214) on General Service List (GSL), practicality of GSL was evaluated with percentage of GSL words in fiction, non-fiction, and academic texts.

contrary, this figure provides strong assistance for ESP practitioners because it indicates that ESP courses could be arranged with a manageable amount of vocabulary.

Moreover, the 733 words are highly valuable for Taiwan English learners for ESP courses due to the fact that these words are assumed to be acquired in the period of senior high school, which in a way bridges the gap between EGP and ESP.

The above interpretation about the figure 11.3% (733 words of SHEWR) limits itself in making reference if we are to theorize the link between English for General Purpose (EGP) and English for Specific Purposes (ESP) because of the wordlists and text employed. Wordlists applied in this study include SHEWR [高中英文參考詞彙表] and Chinese-English Translation of Important Accounting Terms [重要會計用語中英對照], both of which are compiled with Taiwan perspectives; texts collected as Business Reports Corpus (BRC) are limited in Chinese corporations involved in issuing securities and these texts cannot be regarded as the only representative of business genre. Hence the above interpretation on 11.3% and inference about link between EGP and ESP holds validity only in the pedagogical context of Taiwan, especially under discussion on the gap between English teaching in the senior high level and College English instruction for learners of the commerce school.

As for identifying formulaic language composed with common words, only the four adjectives *annual*, *applicable*, *financial*, and *significant* will be discussed due to constraints of space.

4.2 Length of formulaic language across genres

This section answers Research Question 2-1, which inquires about length of formulaic language composed by identical common words across genres. Research results are presented in the order of four extracted common word *annual*, *applicable*, *financial*, and *significant*.

Table 4.2.1 gives data about length of formulaic language composed by *annual* across the three target corpora. The length of formulaic language was evaluated with number of collocates of the queried word with Danielsson's method of cumulative frequency (Section 2.3.3). As Table 4.2.1 shows, *annual* possesses 7 collocates in BRC and 3 collocates in both BC-A (texts of reportage) and BC-H (official documents). There are three groups of collocates in BC-A divided by semicolon.

Table 4.2.1 Length of formulaic sequences composed by *annual*

Queried word: <i>annual</i>	Collocates identified with Cumulative Frequency	Amounts of collocates
Business Reports Corpus (BRC)	report, this, in, included, elsewhere, F, on	7
Subdivision A of Brown Corpus (BC-A, texts of reportage)	the, at, s; the, at, of; the, at, meeting	3
Subdivision H of Brown Corpus (BC-H, official documents)	the, of, at	3

In Table 4.2.2 below, we can see that *applicable* have 8 collocates in BRC, 1 collocate in BC-H, but it has no occurrences in BC-A.

Table 4.2.2 Length of formulaic sequences composed by *applicable*

Queried word: <i>applicable</i>	Collocates identified with Cumulative Frequency	Amounts of collocates
Business Reports Corpus (BRC)	to, the, texts, regulations, of, laws, for, and	8
Subdivision A of Brown Corpus (BC-A, texts of reportage)	No hits of <i>applicable</i> in this subcorpus.	
Subdivision H of Brown Corpus (BC-H, official documents)	to	1

The other two common words *financial* and *significant* still have the most collocates in BRC, as can be seen in Table 4.2.3 and Table 4.2.4; these two words both have more collocates in Brown Corpus-H than in Brown Corpus-A.

Table 4.2.3 Length of formulaic sequences composed by *financial*

Queried word: <i>financial</i>	Collocates identified with Cumulative Frequency	Amounts of collocates
Business Reports Corpus (BRC)	the, statements, of, consolidated, and	5
Subdivision A of Brown Corpus (BC-A, texts of reportage)	the, of, support	3
Subdivision H of Brown Corpus (BC-H, official documents)	the, of, condition, upon	4

Table 4.2.4 Length of formulaic sequences composed by *significant*

Queried word: <i>significant</i>	Collocates identified with Cumulative Frequency	Amounts of collocates
Business Reports Corpus (BRC)	the, of, a, portion, and	5
Subdivision A of Brown Corpus (BC-A, texts of reportage)	a	1
Subdivision H of Brown Corpus (BC-H, official documents)	in, the	2

According to the results presented above, formulaic language composed with identical common words are generally longer in BRC than those in BC-A (texts of reportage) and BC-H (official documents), which suggests that types of genre play a role in length of formulaic language. As introduced in Section 3.3.1, BRC is the genre specifically for the investment public, BC-H (official documents) is for the interested parties engaged in government affairs, and BC-A (texts of reportage) is for the general public without specific purpose for information. In this respect, texts in BRC (20-F designated by SEC) is relatively more specific than those in BC-A (texts of reportage) and in BC-H (official documents), which should be the reason that formulaic language is generally longer in BRC. With the same reason, when formulaic language in BC-A (texts of reportage) and BC-H (official documents) are compared to each other, there is a tendency that formulaic language in official documents is longer than those in texts of reportage; this can be interpreted that texts in BC-H (official documents) have higher degree of specificity than those in BC-A (texts of reportage). This inference may seem insufficient with *annual* in Table 4.2.1, but it is believed that the same inference would be arrived with larger amount of texts under investigation.

To answer Research Question 2-1, we can conclude with certainty that types of genre affect length of formulaic language, which is the same as our earlier hypothesis (Section 2.5). If type of readers of texts is restricted to certain kind of audience with specific purpose for information, a common word will form relatively longer formulaic language in the texts.

4.3 Composition of formulaic language across genres

In the previous section, we have discussed on length of formulaic language across genres; in this section, we continue to explore composition of formulaic language in order to answer Research Question 2-2, which asks about whether those sequences of formulaic language composed with extracted collocates allow syntagmatic or paradigmatic variation. The following four tables demonstrate composition of formulaic language composed by *annual*, *applicable*, *financial*, and *significant* respectively, which were identified with application of the corpus software AntConc (Section 3.3.3).

Table 4.3.1 Composition of formulaic language composed by *annual*

Queried word: <i>annual</i>	Units of formulaic language	Order and variation of units within identified formulaic language
Business Reports Corpus (BRC)	report, this, in, included, elsewhere, F, on	included-elsewhere-in-this-annual-report-on-Form- 20-F (66 concordance hits)
Subdivision A of Brown Corpus (BC-A, texts of reportage)	the, at, s; the, at, of; the, at, meeting	at-the-X's (PV)*-annual (2 concordance hits) at-the- X(SV)*-annual-Y(PV)-of-Z(PV) (2 concordance hits) at-the- X(SV)-annual-meeting (2 concordance hits)
Subdivision H of Brown Corpus (BC-H, official documents)	the, of, at	at-the-annual-X(PV)-of (1 concordance hit)

*PV is an abbreviation for paradigmatic variation and SV is for syntagmatic variation.

Table 4.3.1 presents order and variation of units in identified formulaic language composed by *annual*; content in the column of “units of formulaic language” is the same as those in the column of “collocates identified with cumulative frequency” in Table

4.2.1. Based on the same data, the order and variation of those units were further checked up by observing concordance lines²⁹.

For instance, the 7 collocates of *annual* in BRC is verified to be aligned in the order *included-elsewhere-in-this-**annual**-report-on-Form-20-F*, in which dashes are used to denote that no items occur between any two of those collocates. This sequence of formulaic language was found to have 66 concordance hits in BRC, among which two sentences are presented below:

- (1) You should read the following discussion and analysis of our financial condition and results of operations in conjunction with our CONSOLIDATED-FINANCIAL-STATEMENTS and the related notes *included elsewhere in this **annual** report on Form 20-F.*
(0100_Agria_2007-12-31.htm)
- (2) The following selected consolidated financial information for the periods and as of the dates indicated should be read in conjunction with our CONSOLIDATED-FINANCIAL-STATEMENTS and accompanying notes *included elsewhere in this **annual** report on Form 20-F.*
(7371_Longtop_2010-3-31.htm)

In (1) and (2), the capitalized CONSOLIDATED-FINANCIAL-STATEMENTS are accounting compounds which were referred from Chinese-English Translation of Important Accounting Terms [重要會計用語中英對照] and had been hyphenated with the software Useful File Utilities (Section 3.1.4). As can be seen, formulaic language

²⁹ Paradigmatic variation can also be verified by applying wild-card query onto each slot of identified formulaic language. However, this procedure was omitted to give concise report on research results.

composed by *annual* is a fixed sequence where no syntagmatic or paradigmatic variation occur.

In BC-A, *annual* have three possible groups of collocates segregated with semicolon: the first group contains *the, at, s*; the second group contains *the, at, of*; and the last group contains *the, at, meeting*.

The group of *annual, the, at, and s* has 2 concordance hits in BC-A (texts of reportage), as (3) and (4) shows:

(3) Martin , who has been in office in Washington , D. C. , for 13 months spoke

at the council's annual

meeting at the Multnomah Hotel. (a10.xml)³⁰

(4) The poll was taken

at the Center's annual

builders' intentions conference. (a27.xml)

From (3) and (4), the formulaic language of *annual* in BC-A are induced as at-the-X's (PV)-*annual*, where PV is an abbreviation for paradigmatic variation and X denotes a set of words that alternate paradigmatically in the slot of X. Here the paradigmatic variation is solely for presenting the possessive form.

(5) and (6) are the two sentences in BC-A with the group of *annual, the, at, and of*:

³⁰ Here the notation for source of text follows those in Brown Corpus, where *a* represents Subdivision A of Brown Corpus and *h* for Subdivision H of Brown Corpus; the number following the English alphabet denotes the order of text sequence in the sub-collection.

(5) The crowd

*at the twenty-first **annual** K. of C. Games,*

final indoor meet of the season, got a thrill a few minutes earlier when a slender, bespectacled woman broke the one-week-old world record in the half-mile run. (a11.xml)

(6) The presentation was made before several hundred persons

*at the **annual** meeting of the League*

at Olney Hall, College of Marin , Kentfield. (a25.xml)

From (5) and (6), the formulaic language *at-the-X(SV)-annual-Y(PV)-of-Z(PV)* can be generalized, where SV denotes syntagmatic variation; X, Y, and Z refers to sets of words that can be alternated or inserted in the slot. The syntagmatic variation occurs to present ordinal (*twenty-first*) followed by *annual*, and the other two slots (underlined *K. and C. Games* in (5) and *the League* in (6)) are paradigmatic variation open for nouns.

The group of *annual, the, at, and meeting* compose the sequence of formulaic language *at-the- X(SV)-annual-meeting*, where X is a set open to syntagmatic variation, as shown in (7) and (8) below:

(7) Martin , who has been in office in Washington , D. C. , for 13 months spoke

*at the council's **annual** *meeting**

at the Multnomah Hotel . (a10.xml)

(8) The presentation was made before several hundred persons

*at the **annual** *meeting**

of the League at Olney Hall, College of Marin, Kentfield. (a25.xml)

Based on (7) and (8), only syntagmatic variation, occurring as possessive form, is allowed within the string of formulaic language *at-the- X(SV)-annual-meeting*.

As for BC-H (official document), there were only 1 concordance line found with *annual, the, of, and at* to compose *at-the-annual-X (PV)-of*, as shown in (9):

(9) An exhibit, `` Macropathology -- An Ancient Art, A New Science ", was presented *at the **annual meeting** of* the American Medical Association.

(h10.xml)

Due to the limited size of BC-H (only 30 pieces of texts), we are unable to see other items allowable in the slot for paradigmatic variation in the genre of official documents.

From Table 4.3.1 and the demonstration above, it can be observed that *annual* composes the most strict formulaic language in business annual reports (BRC), in that the identified sequence does not maintain any syntagmatic or paradigmatic variation. It seems that the formulaic language composed by *annual* allows for variation in a larger degree in texts of reportage (BC-A) than official documents in (BC-H).

Table 4.3.2 reports formulaic language composed by *applicable* across the three corpora, in which *applicable* compose relatively strict formulaic language in BRC. *Applicable* but does not have any occurrence in BC-A (texts of reportage). Meanwhile, *applicable* composes the phrase *applicable-to* in BC-H (official documents); variation of *applicable-to* was not checked since it is too short a sequence for further investigation.

Table 4.3.2 Composition of formulaic language composed by *applicable*

Queried word: <i>applicable</i>	units of formulaic language	order and variation of units in formulaic language
Business Reports Corpus (BRC)	to, the, texts, regulations, of, laws, for, and	and-to-the-texts-of- <i>applicable</i> -laws-and-regulations -for-further-information (8 concordance hits)
Subdivision A of Brown Corpus (BC-A, texts of reportage)	No hits of <i>applicable</i> in this subcorpus.	
Subdivision H of Brown Corpus (BC-H, official documents)	to	<i>applicable-to</i> (2 concordance hits)

8 concordance hits of formulaic language were found in BRC, among which two sentences are presented below:

(10) You should refer to the text of the Articles of Association

*and to the texts of **applicable** laws and regulations for further information.*

(2821_Sinopec_2008-12-31.htm)

(11) You and your advisors should refer to the text of our articles of association,

as amended,

*and to the texts of **applicable** laws and regulations for further information.*

(2911_China_P_C_2008-12-31.htm)

Applicable forms the phrase *applicable-to* in BC-H (official documents), as shown in (12) and (13):

(12) ... the scope and adequacy of State mine-safety laws

applicable to

such mines and the enforcement of such laws . (h09.xml)

(13) The deposit of rupees to the account of the Government of the United States of America in payment for the commodities and for ocean transportation costs financed by the Government of the United States of America (except excess costs resulting from the requirement that United States flag vessels be used) shall be made at the rate of exchange for United States dollars generally

applicable to

import transactions... (h09.xml)

Applicable has no hits in texts of reportage (BC-A), but forms the phrase *applicable-to* in official documents (BC-H) and the string of formulaic language *and-to-the-texts-of-applicable-laws-and-regulations-for-further-information* in business annual reports (BRC), which may indicate that the word *applicable* is especially pertinent to texts of legal connotation and deserve some attention in ESP courses.

Table 4.3.3 displays composition of formulaic language composed by the common word *financial*.

Table 4.3.3 Composition of formulaic language composed by *financial*

Queried word: <i>financial</i>	units of FS	order and variation of units in FS
Business Reports Corpus (BRC)	the, statements, of, consolidated, and	and-the-consolidated-financial-statements-of (2 concordance hits)
Subdivision A of Brown Corpus (BC-A, texts of reportage)	the, of, support	the-financial-support-of (2 concordance hits)
Subdivision A of Brown Corpus (BC-H, official documents)	the, of, condition, upon	upon-the-financial-condition-of (2 concordance hits)

In BRC, *financial* composed the sequence and-the-consolidated-financial-statements-of, as shown in (14) and (15):

(14) ...for at HISTORICAL-COST

and the consolidated financial statements of

the Company prior to ... (2911_China_P_C_2008-12-31.htm)

(15) ... for at HISTORICAL-COST

and the consolidated financial statements of

the Company prior to... (_China_P_C_2007-12-31.htm)

(16) and (17) demonstrate the formulaic language *the-financial-support-of* in

BC-A (texts of reportage):

(16) Family Service could not open its doors to a single family without

the financial support of

the United Givers Fund. (a33.xml)

(17) The local community maintains responsibility for

the financial support of

its own library program , facilities , and services , but wider resources and additional services become available through membership in a system.

(a44.xml)

In BC-H (official documents), *financial* forms the sequence *upon-the-financial-condition-of*, as shown in (18) and (19):

(18) It should be kept in mind that the ease or difficulty with which a town or city

can convert to the proposed plan is directly dependent

upon the financial condition of

that town or city. (h07.xml)

(19) However, it must be stressed that much depends

upon the financial condition of

the individual cities and towns involved. (h07.xml)

Results show that there are no syntagmatic or paradigmatic variation within formulaic language composed by *financial*.

Table 4.3.4 gives results about composition of formulaic language with the word *significant*.

Table 4.3.4 Composition of formulaic language composed by *significant*

Queried word: <i>significant</i>	Units of formulaic language	Order and variation of units in formulaic language
Business Reports Corpus (BRC)	the, of, a, portion, and	and-a-significant-portion-of-the (9 concordance hits) and-X(PV)-a-significant-portion-of-the (6 concordance hits)
Subdivision A of Brown Corpus (BC-A, texts of reportage)	a	a-significant (2 concordance hits)
Subdivision H of Brown Corpus (BC-H, official documents)	in, the	significant-X(PV)-in-Y(PV)-the (1 concordance hit)

In BRC, *significant* forms two sequences of formulaic language. One is *and-a-significant-portion-of-the*, as shown in (20), and the other is *and-X(PV)-a-significant-portion-of-the*, as shown in (21) and (22).

(20) Because the Group is limited in the types of investments as permitted by China Insurance Regulatory Commission *and a significant portion of the* portfolio is in government bonds...(6311_C_Life_Insurance_2007-12-31.htm)

(21) However, since we regard subscription-based services as our current core business *and allocate a significant portion of the* advertising inventories of our websites...(7389_C_Finance_Online_2007-12-31.htm)

(22) Historically, our pre-sales activities have allowed us to generate CASH-FLOWs relatively early in the development cycle *and to fund a significant portion of the* capital required for existing projects, reducing financing needs and associated costs. (1520_Xinyuan_2007-12-31.htm)

From (21) and (22), it can be seen that X is a slot for paradigmatic variation open to infinitive without the participle *to* (*allocate* in (21)) or infinitive with the participle *to* (*to fund* in (22)).

In BC-A (texts of reportage), *significant* follows the indefinite article *a*, as shown in (23) and (24).

(23) This is *a significant* advance but its import should not be exaggerated. (a35.xml)

(24) In 1957 Nixon delivered *a significant* opinion that a majority of Senators had the power to adopt new rules at the beginning of each new Congress...(a37.xml)

In BC-H (official documents), there was only one concordance line found, as (25) shows.

(25) In still others which are barely on the threshold of the transition into modernity, the decade can bring *significant* progress in launching the slow process of developing their human resources...(h02.xml)

Due to limited amount of texts in BC-H (official documents), the formulaic language composed with significant was tentatively induced as *significant-X(PV)-in-Y(PV)-the*, where X is a slot for paradigmatic variation open to noun, and Y is another slot for paradigmatic variation open to gerund.

From Table 4.3.1 to Table 4.3.4, we can observe that in general, formulaic language composed by common words allow least variation in Business Reports Corpus, relatively more variation in BC-H (official documents), and most variation in BC-A (texts of reportage). This phenomenon of different degree of variation across genres indicates that type of genre is a crucial factor for composition of formulaic language, which verifies our earlier hypothesis (Section 2.5). The most fixed composition of formulaic language occurs in BRC, which contains texts with information prepared specifically for the investment public. When the same common word occurs in BC-H (official documents), which possesses texts prepared for concerned parties engaged in government services, the formulaic language seems still fixed with shorter composition but allows some variation. Generally formulaic language with the highest degree of variation occurs in BC-A (texts of reportage), by which we could conclude that reportage texts are relatively flexible for variation of formulaic language.

As for Research Question 2-2, which asks whether variation within formulaic language occur across genres, this research provides positive results as has been shown above. Furthermore, it is evidenced that type of expected readers of texts plays a significant role in composition of formulaic language.

4.4 Summary of chapter

In this chapter, research results of extracted common words and identified formulaic language have been presented. Following the proposed operational definition of common word (Section 2.2.5), words were extracted with corpus methodology such as keyness and text coverage. Then the extracted words were investigated to see how they are distributed in the homemade Business Reports Corpus (BRC) as well as in the pedagogical wordlist SHEWR [高中英文參考詞彙表]. Figures about proportion of extracted words in BRC directly answer the enquiry of Research Question 1 and gives guidance for ESP practitioners to design vocabulary course in business genre. The figure of percentage of extracted words in SHEWR [高中英文參考詞彙表] shows a low percentage (11.3%), which seemingly indicates low practicality of SHEWR in business genre. Nevertheless, due to the prudent application of the notion of unusual frequency with corresponding corpus technique, the low figure in fact makes a conservative claim on the usefulness of those extracted words, which come with a manageable amount and occur consistently in context of business. With this evidence, common words identified in this thesis can provide reliable assistance for ESP practitioners in arranging courses.

The extracted common words were later used to identify formulaic language across genres of business, news, and official document. By counting collocates of queried words and observing composition of formulaic language, it was found that length and composition of formulaic language composed with identical common word vary across different genres, as been hypothesized in Section 2.5. With seeing genre according to types of anticipated audience, research results show that a common word tends to compose lengthier formulaic language with restricted variation in texts prepared

for readers with specific need for information. On the contrary, common words were found to form shorter formulaic language that allows more variation in texts for general purpose. In other words, the more specific the genre is, the less variation the structure seems to be. These result answers Research Question 2 and signifies the importance of genre in lexical research.



Chapter 5 Conclusion

This chapter gives overview of research questions and obtained results, discussions on significance and implications of the present study, and prospect on future work.

5.1 Overview of research questions and research results

From Chapter 3 and Chapter 4, this thesis has presented research design and research results geared to answer research questions proposed in Chapter 2:

Research Question (1)

How common words are distributed in company annual reports?

Research Question (2)

How formulaic language composed with identical common words vary across genres?

These two questions were enlightened by the common core hypothesis from the ESP field (Corder, 1973, cited in Bloor & Bloor, 1986, pp. 16-21), phraseological view toward language from corpus linguistics (Hunston, 2002, pp. 137-157; Sinclair, 1991, pp. 110-112), and Basturkmen's (2006) framework on conceptualizing ESP (pp. 26-28), with all of which it is assumed that language can be described as dissimilar formations of formulaic language composed with identical common words. This assumption follows Dudley-Evans and St John's (1998) support on real content of ESP, who maintain that real content involves non-technical language items with certain communicative function and does not require specialized knowledge for language teachers (p. 11). The non-technical language items, paraphrased as common words in this study, have been operationally defined with assistance of corpus linguistic techniques as well as the

common word list Senior High English Wordlist for Reference [高中英文參考詞彙表] made for pedagogical reason in Taiwan (Jeng, Chang, Cheng, & Gu [鄭恆雄、張郁慧、程玉秀與顧英秀] 2002). By answering Research Question 1, the functionality of Senior High English Wordlist for Reference (SHEWR) [高中英文參考詞彙表] has also been assessed.

In an attempt to respond the proposed research questions, corpus methodology was adopted. The notions of unusual frequency (keyness) and text coverage (key-keyword) were employed to extract common words. The function of stop-list and match-list in the corpus software WordSmith Tools (Scott, 2011a) was also utilized to avoid technical compounds and to match words in SHEWR [高中英文參考詞彙表]. Then the extracted common words were investigated to see how these words are distributed in the homemade Business Reports Corpus (BRC) as well as their proportion in SHEWR. To survey composition of formulaic language (composed with identical common words extracted previously) in the aspect of syntagmatic and paradigmatic variation, the notion of cumulative frequency (Danielsson, 2007) was applied with assistance of another corpus software AntConc (Anthony, 2010) to identify collocates of the queried common words as well as their concordance lines.

Research results for Research Question 1 on distribution of extracted common words in the homemade Business Reports Corpus (BRC) have shown that about one-tenth (11.3%) of SHEWR [高中英文參考詞彙表] words consistently occur in annual reports of business corporations, which suggests that these words are relevant to English learners who need to acquaint with business genre. As for Research Question 2, research results successfully demonstrate variations of formulaic language across

different genres; identified concordance lines show that length and syntagmatic/paradigmatic variation of formulaic language bear certain relation with types of genre. Texts created for specific kind of audience tend to contain longer sequences of formulaic language which allow relatively constrained syntagmatic/paradigmatic variation, while texts for the general public tend to maintain shorter strings that allow more flexibility for variation. Results for Research Question 2 indicate that the instruction of language content for English for Specific Purposes (as well as English for Business Purposes, or EBP) relies heavily on the awareness of genre since types of genre fundamentally affects language content to be taught.

5.2 Significance and implication

Obtained with careful research design, research results of this thesis carry important implication in instruction of ESP in Taiwan, specifically in vocabulary teaching. Also, this study in itself realizes the significance of the application of corpus methodology in English teaching.

The result that 11.3% (733 words) of SHEWR [高中英文參考詞彙表] (a common-word list followed by public schools in Taiwan) consistently occur in the document of 20-F (a business genre with international currency) indicates that this common-word list, though originally designed for English for General Purposes (EGP), provides reliable assistance for English learners who attempt to specialize in the business world. In other words, this 11.3% (733 words) of SHEWR plays a role as a bridge between EGP and ESP for English teaching in the context of Taiwan. Furthermore, the subsequent employment of the extracted common words in this thesis on identifying formulaic language in various genres has realized Dudley-Evans and St John's (1998)

notion of real content, which does not involve terms with technical knowledge but sequences of common lexis with certain communicative function. Hence this study contributes in objectively identifying real content for ESP practitioners with corpus methodology.

The identification of formulaic language in the present thesis also evidences the existence of phraseology with consideration of genres. By comparing formulaic language composed with identical common words across genres, this study asserts that observation of variation of formulaic language cannot be divorced from specific genres, and indicates the importance of context in vocabulary teaching. This thesis not only provides concrete evidence for Sinclair's (1991) idea of idiom principle with consideration of genres, but also indicates that Lexical Approach (Lewis, 2000), a teaching method that incorporates the notion of idiom principle, shall lay more emphasis on the factor of genre.

On the perspective of methodology, this thesis presents a successful demonstration of corpus method on extracting language content for ESP instruction. With objectives of curriculum clearly defined (such as an ESP program focusing on non-technical elements), corpus tools are handy equipment for ESP practitioners to locate teaching points. The researcher hence advocates employment of corpus tools in the process of curriculum design for ESP courses, along with collaboration between subject teachers and language teachers.

5.3 Research in prospect

Based on this thesis, future research can be continued in lexical study, genre analyses, and experimentation of another corpus software ConcGram (Greaves, 2005).

This study has investigated functionality of SHEWR [高中英文參考詞彙表], in which vocabulary is categorized in six levels. Words belonging to the six levels could be further analyzed with the same Business Reports Corpus and Brown Corpus to scrutinize the degree of difficulty of the six-level word bank. Moreover, wordlist for English teaching in junior high school could be covered to expand this thread of research on common-word lists in pedagogical context of Taiwan to reveal the connection between EGP and ESP.

As for genre analyses, further investigation of formulaic language could be executed in different sections of 20-F. Viewed as one definite genre, the document of 20-F is in fact composed with different sections with particular aims for information disclosure, such as information of the company (Item 4 in 20-F), unresolved staff comments (Item 4A in 20-F), corporate governance (Item 16G in 20-F), and so forth. Deeper observation can be made among these various sections within the same genre. Also, cross-genre analyses can be conducted with larger amount of texts in reportage, official documents or other types of texts. This thesis is limited with the size of Brown Corpus and it is believed that a larger corpus will help enhance future research. More knowledge on analyzing genre from genre theories also needs to be consulted to study communicative function of formulaic language.

Lastly, in the perspective of corpus linguistics, this thesis could be executed with one another corpus software ConcGram (Greaves, 2005). ConcGram is designed to scrutinize lexical permutations without querying the node item in advance. It is believed that the application of ConcGram will throw light on the same issue of variation of formulaic language across genres.



References

- Accounting Research and Development Foundation of the Republic of China. (2011). 重要會計用語中英對照 [Chinese-English translation of important Accounting terms]. Retrieved from <http://www.ardf.org.tw/html/tifrs1000128.pdf>
- Anthony, L. (2010). AntConc (Version 3.2.1) [Computer Software]. Tokyo, Japan: Waseda University. Retrieved from <http://www.antlab.sci.waseda.ac.jp/>
- Baker, M. (1988). Sub-technical vocabulary and the ESP teacher: An analysis of some rhetorical items in medical journal articles. *Reading in a Foreign Language*, 4(2), 91-105.
- Baron, A., Rayson, P., & Archer, D. (2009). Word frequency and key word statistics in corpus linguistics. *Anglistik: International Journal of English Studies*, 20(1), 41-67.
- Barron, C. (2003). Problem-solving and EAP: Themes and issues in a collaborative teaching venture. *English for Specific Purposes*, 22(3), 297-314.
- Basturkmen, H. (2006). *Ideas and options in English for specific purposes*. Mahwah, NJ: Lawrence Erlbaum.
- Belcher, D. D. (2006). English for Specific Purposes: Teaching to perceived needs and imagined futures in worlds of work, study, and everyday life. *TESOL Quarterly*, 40(1), 133-156.
- Biber, D. (2009). A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics*, 14(3), 275-311. doi: 10.1075/ijcl.14.3.08bib

- Bloor, M., & Bloor, T. (1986). *Language for specific purposes: Practice and theory*. (CLCS occasional paper no. 19). Dublin: Trinity College.
- Charles, M. (2006). Phraseological patterns in reporting clauses used in citation: A corpus-based study of theses in two disciplines. *English for Specific Purposes*, 25(3), 310-331. doi: DOI: 10.1016/j.esp.2005.05.003
- Cheng, W., Greaves, C., & Warren, M. (2006). From n-gram to skipgram to conogram. [Article]. *International Journal of Corpus Linguistics*, 11(4), 411-433.
- Chujo, K., & Genung, M. (2004). Comparing the three specialized vocabularies used in 'Business English,' TOEIC, and British National Corpus Spoken Business Communications. *Practical English Studies*, 11, 1-15.
- Chung, T. M., & Nation, P. (2004). Identifying technical vocabulary. *System*, 32(2), 251-263.
- Clear, J. (1993). From Firth Principles: Computational tools for the study of collocation. In M. Baker, G. Francis & E. Tognini-Bonelli (Eds.), *Text and technology: In honour of John Sinclair* (pp. 271-292). Philadelphia: John Benjamins.
- Coxhead, A. (2000). A New Academic Word List. *TESOL Quarterly*, 34(2), 213-238.
- Coxhead, A., & Nation, P. (2001). The specialised vocabulary of English for academic purposes. In J. Flowerdew & M. Peacock (Eds.), *Research perspectives on English for Academic Purposes* (pp. 252-267). Cambridge: Cambridge University Press.
- Crystal, D. (2003). *A dictionary of linguistics and phonetics* (5 ed.). Oxford: Blackwell Publishing.

- Danielsson, P. (2003). Automatic extraction of meaningful units from corpora: A corpus-driven approach using the word stroke. *International Journal of Corpus Linguistics*, 8(1), 109-127.
- Danielsson, P. (2007). What constitutes a unit of analysis in language? *Linguistik online*, 31. Retrieved from http://www.linguistik-online.de/31_07/danielsson.pdf
- Dudley-Evans, T. (2001). Team-teaching in EAP: Changes and adaptations in the Birmingham approach. In J. Flowerdew & M. Peacock (Eds.), *Research perspectives on English for Academic Purposes* (pp. 225-238). Cambridge: Cambridge University Press.
- Dudley-Evans, T., & St John, M. J. (1998). *Developments in ESP: A multi-disciplinary approach*. Cambridge: Cambridge University Press.
- Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1), 61-74.
- Durrant, P. (2009). Investigating the viability of a collocation list for students of English for academic purposes. *English for Specific Purposes*, 28(3), 157-169.
- Flowerdew, J., & Peacock, M. (2001). Issues in EAP: A preliminary perspective. In J. Flowerdew & M. Peacock (Eds.), *Research Perspectives on English for Academic Purposes* (pp. 8-24). Cambridge: Cambridge University Press.
- Flowerdew, J., & Wan, A. (2010). The linguistic and the contextual in applied genre analysis: The case of the company audit report. *English for Specific Purposes*, 29(2), 78-93.
- Flowerdew, L. (1998). Corpus linguistic techniques applied to textlinguistics. *System*, 26(4), 541-552.

- Flowerdew, L. (2005). An integration of corpus-based and genre-based approaches to text analysis in EAP/ESP: Countering criticisms against corpus-based methodologies. *English for Specific Purposes*, 24(3), 321-332.
- Francis, W. N., & Kucera, H. (1979). Brown Corpus manual: Manual of information to accompany a standard corpus of present-day edited American English, for use with digital computers. Retrieved from <http://icame.uib.no/brown/bcm.html>
- Greaves, C. (2005). Introduction to ConcGram©. *Tuscan Word Centre International Workshop*.
- Groom, N. (2005). Pattern and meaning across genres and disciplines: An exploratory study. *Journal of English for Academic Purposes*, 4(3), 257-277.
- Hunston, S. (2002). *Corpora in Applied Linguistics*. New York: Cambridge University Press.
- Hunston, S. (2008). Starting with the small words: Patterns, lexis and semantic sequences. *International Journal of Corpus Linguistics*, 13(3), 271-295.
- Hunston, S., & Francis, G. (1998). Verbs observed: A corpus-driven pedagogic grammar. *Applied Linguistics*, 19(1), 45-72.
- Hyland, K. (2002). Specificity revisited: how far should we go now? *English for Specific Purposes*, 21(4), 385-395. doi: 10.1016/s0889-4906(01)00028-x
- Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*, 27(1), 4-21. doi: 10.1016/j.esp.2007.06.001
- Jablonkai, R. (2010). English in the context of European integration: A corpus-driven analysis of lexical bundles in English EU documents. *English for Specific Purposes*, 29(4), 253-267. doi: 10.1016/j.esp.2010.04.006

- Kay, H., & Dudley-Evans, T. (1998). Genre: what teachers think. *ELT Journal*, 52(4), 308-314.
- Lewis, M. (2000). Language in the lexical approach. In M. Lewis (Ed.), *Teaching collocation* (pp. 126-154). Hove, England: Language Teaching Publications.
- Menon, S., & Mukundan, J. (2010). Analysing collocational patterns of semi-technical words in science textbooks. *Pertanika Journal of Social Science & Humanities*, 18(2), 241-258.
- Nelson, M. (2000). *A corpus-based study of the lexis of Business English and Business English teaching materials*. Unpublished thesis. University of Manchester. Manchester. Retrieved from <http://users.utu.fi/micnel/thesis.html>
- Römer, U. (2008). Identification impossible?: A corpus approach to realisations of evaluative meaning in academic writing. *Functions of Language*, 15(1), 115-130.
- ReplSoft.com. (2010). Useful File Utilities (Version 3.6) [Computer software].
- Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Houndmills, Basingstoke, Hampshire; New York: Palgrave Macmillan.
- Schulze, R., & Römer, U. (2008). Introduction: Patterns, meaningful units and specialized discourses. *International Journal of Corpus Linguistics*, 13(3), 265-270.
- Scott, M. (1997). PC analysis of key words -- And key key words. *System*, 25(2), 233-245.
- Scott, M. (2011a). WordSmith Tools (Version 5.0) [Computer Software]. Liverpool: Lexical Analysis Software.

- Scott, M. (2011b). WordSmith Tools Help [Software Manual]. Liverpool: Lexical Analysis Software.
- Sinclair, J. (1991). *Corpus, concordance, collocation*. New York: Oxford University Press.
- U.S. Securities and Exchange Commission. (2010, December 13). EDGAR Filer Manual (Volume II) EDGAR Filing (Version 16).
- Wang, J., Liang, S.-l., & Ge, G.-c. (2008). Establishment of a Medical Academic Word List. *English for Specific Purposes*, 27(4), 442-458.
- Ward, J. (2009a). A basic engineering English word list for less proficient foundation engineering undergraduates. *English for Specific Purposes*, 28(3), 170-182.
- Ward, J. (2009b). EAP reading and lexis for Thai engineering undergraduates. *Journal of English for Academic Purposes*, 8(4), 294-301.
- Wu, H., & Badger, R. G. (2009). In a strange and uncharted land: ESP teachers' strategies for dealing with unpredicted problems in subject knowledge during class. *English for Specific Purposes*, 28(1), 19-32.
- Yeung, L. (2007). In search of commonalities: Some linguistic and rhetorical features of business reports as a genre. *English for Specific Purposes*, 26(2), 156-179. doi: DOI: 10.1016/j.esp.2006.06.004
- 鄭恆雄, 張郁慧, 程玉秀, & 顧英秀. (2002). 高中英文參考詞彙表 [Senior High English Wordlist for Reference]. Retrieved from http://www.ceec.edu.tw/Research2/doc_980828/ce37/5.pdf

Appendices

Appendix 3.1.1 Complete List of 35 Companies

No.	公司名稱	Company Name	SIC (Standard Industrial Classification)	Industry Title	Periods of Collected Files
1	華奧物種	Agria Corp	0100	AGRICULTURAL PRODUCTION-CROPS	2007; 2008; 2009
2	兗州煤業	Yanzhou Coal Mining Co Ltd	1221	BITUMINOUS COAL & LIGNITE SURFACE MINING	2007; 2008; 2009
3	中國海洋石油	CNOOC Ltd	1311	CRUDE PETROLEUM & NATURAL GAS	2007; 2008; 2009
4	鑫苑置業	Xinyuan Real Estate Co Ltd	1520	GENERAL BLDG CONTRACTORS - RESIDENTIAL BLDGS	2007; 2008; 2009
5	上海石油化工	Sinopec Shanghai Petrochemical Co Ltd	2821	PLASTIC MATERIALS, SYNTH RESINS & NONVULCAN ELASTOMERS	2007; 2008; 2009
6	三生制藥	3SBio Inc	2824	PHARMACEUTICAL PREPARATIONS	2007; 2008;

No.	公司名稱	Company Name	SIC (Standard Industrial Classification)	Industry Title	Periods of Collected Files
					2009
7	古杉環境	Gushan Environmental Energy Ltd	2860	INDUSTRIAL ORGANIC CHEMICALS	2007; 2008; 2009
8	中國石化	China Petroleum Chemical Corp	2911	PETROLEUM REFINING	2007; 2008; 2009
9	中國鋁業	Aluminum Corp of China Ltd	3334	PRIMARY PRODUCTION OF ALUMINUM	2007; 2008; 2009
10	WSP 石油管製造	Wsp Holdings Ltd	3533	OIL & GAS FIELD MACHINERY & EQUIPMENT	2007; 2008; 2009
11	日月光半導體	Advanced Semiconductor Engineering Inc.	3674	SEMICONDUCTORS & RELATED DEVICES	2007; 2008; 2009
12	中國醫療技術	China Medical Technologies	3841	SURGICAL & MEDICAL INSTRUMENTS & APPARATUS	2008; 2009; 2010

No.	公司名稱	Company Name	SIC (Standard Industrial Classification)	Industry Title	Periods of Collected Files
13	廣深鐵路	Guangshen Railway Co Ltd	4011	RAILROADS, LINE-HAUL OPERATING	2007; 2008; 2009
14	中國東方航空	China Eastern Airlines Corp Ltd	4512	AIR TRANSPORTATION, SCHEDULED	2007; 2008; 2009
15	藝龍網	eLong Inc	4700	TRANSPORTATION SERVICES	2007; 2008; 2009
16	中華電信	Chunghwa Teletcom Co Ltd	4812	RADIOTELEPHONE COMMUNICATIONS	2007; 2008; 2009
17	中國移動	China Mobile Ltd	4813	TELEPHONE COMMUNICATIONS (NO RADIOTELEPHONE)	2007; 2008; 2009
18	掌上靈通	Linktone Ltd	4822	TELEGRAPH & OTHER MESSAGE COMMUNICATIONS	2007; 2008; 2009
19	華友世紀	Hurray Holding Co Ltd	4899	COMMUNICATIONS SERVICES, NEC	2007; 2008; 2009
20	華能國	Huaneng Power	4911	ELECTRIC SERVICES	2007;

No.	公司名稱	Company Name	SIC (Standard Industrial Classification)	Industry Title	Periods of Collected Files
	際電力	International Inc			2008; 2009
21	歐陸科儀	Euro Tech Holdings Co Ltd	5040	WHOLESALE-PROFESSIONAL & COMMERCIAL EQUIPMENT & SUPPLIES	2007; 2008; 2009
22	橡果國際	Acorn International Inc	5900	RETAIL-MISCELLANEOUS RETAIL	2007; 2008; 2009
23	海王星 晨醫藥	China Nepstar Chain Drugstore Ltd	5912	RETAIL-DRUG STORES AND PROPRIETARY STORES	2007; 2008; 2009
24	匯豐控股	HSBC Holdings Plc	6035	SAVINGS INSTITUTION, FEDERALLY CHARTERED	2007; 2008; 2009
25	中國人壽	China Life Insurance Co Ltd	6311	LIFE INSURANCE	2007; 2008; 2009
26	泛華保險	Cninsure Inc	6411	INSURANCE AGENTS, BROKERS & SERVICE	2007; 2008; 2009
27	易居中國	E House China Holdings Ltd	6531	REAL ESTATE AGENTS & MANAGERS (FOR OTHERS)	2007; 2008; 2009

No.	公司名稱	Company Name	SIC (Standard Industrial Classification)	Industry Title	Periods of Collected Files
28	如家快捷酒店	Home Inns Hotels Management Inc	7011	HOTELS & MOTELS	2007; 2008; 2009
29	航美傳媒	AirMedia Group Inc.	7311	SERVICES-ADVERTISING AGENCIES	2007; 2008; 2009
30	前程無憂	51Job Inc	7361	SERVICES-EMPLOYMENT AGENCIES	2007; 2008; 2009
31	百度	Baidu com Inc	7370	SERVICES-COMPUTER PROGRAMMING, DATA PROCESSING, ETC.	2007; 2008; 2009
32	東南融通	Longtop Financial Technologies Ltd	7371	SERVICES-COMPUTER PROGRAMMING SERVICES	2008; 2009; 2010
33	九城關貿	NINETOWNS INTERNET TECHNOLOGY GROUP CO LTD	7372	SERVICES-PREPACKAGED SOFTWARE	2007; 2008; 2009
34	中國金融在線	China Finance Online Co LTD	7389	SERVICES-BUSINESS SERVICES, NEC	2007; 2008; 2009

No.	公司名稱	Company Name	SIC (Standard Industrial Classification)	Industry Title	Periods of Collected Files
35	全美測評軟體	ATA Inc	8200	SERVICES-EDUCATIONAL SERVICES	2008; 2009; 2010



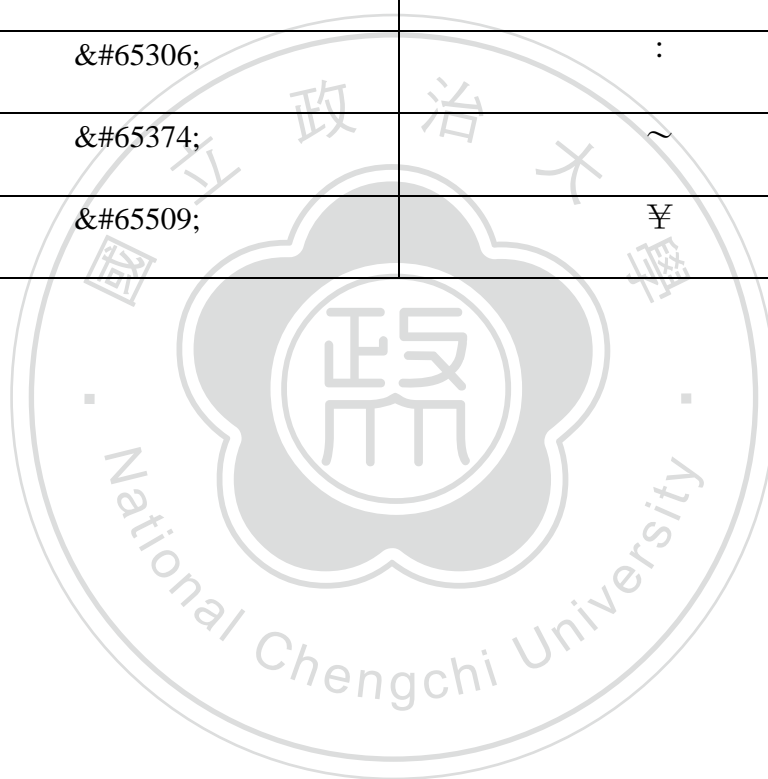
Appendix 3.1.2 List of Entity File

No.	HTML Codes	Corresponding Numeric Codes or Characters
1	 sp;	 (space)
2	&	& (ampersand)
3	&	&
4	+	+
5	-	-
6	=	=
7	>	>
8	[[
9]]
10	a	α
11	l	λ
12	o	 (space)
13	x	 (space)
14	~	~
15	†	†
16	‘	‘
17	’	’
18	“	“
19	”	”

No.	HTML Codes	Corresponding Numeric Codes or Characters
20	•	•
21	–	–
22	—	—
23	™	™
24	Ÿ	ÿ
25	 	 (space)
26	¡	ı
27	£	≤
28	¥	¥
29	§	§
30	¨	 (space)
31	­	
32	®	®
33	±	±
34	³	≥
35	·	·
36	¾	¾
37	×	×
38	à	à
39	ý	ÿ

No.	HTML Codes	Corresponding Numeric Codes or Characters
40	þ	 (space)
41	Ÿ	ÿ
42	ˇ	˘
43	ο	o
44	 	 (space)
45	–	—
46	—	—
47	―	—
48	‘	‘
49	’	’
50	“	“
51	”	”
52	†	†
53	•	•
54	…	...
55	─	 (space)
56	■	■
57	●	●
58	　	 (space)
59		 (space)

No.	HTML Codes	Corresponding Numeric Codes or Characters
60	％	%
61	，	,
62	－	—
63	／	/
64	：	:
65	～	~
66	￥	¥



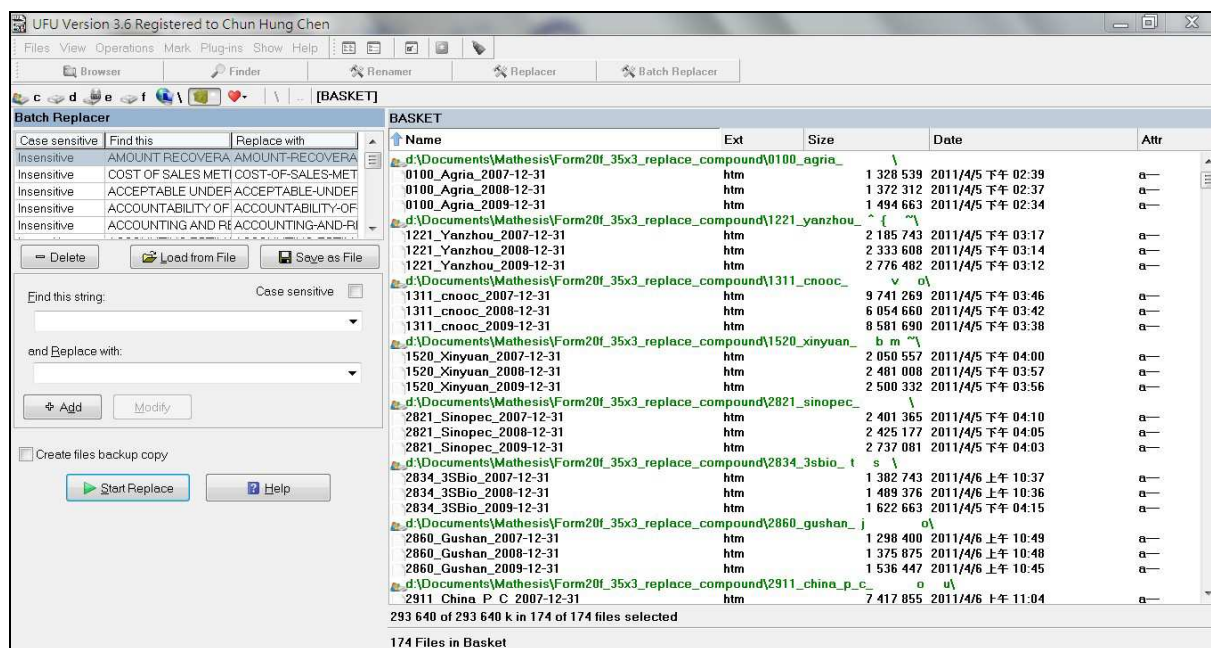
Appendix 3.1.3 One Sample Page of Chinese-English Translation of Important

Accounting Terms [重要會計用語中英對照]

「重要會計用語中英對照」

Item	Term in English	Term in Chinese
1	[Amount] recoverable	可回收(金額)
2	'Corridor'	「緩衝區」
3	'Cost of sales' method	「銷貨成本」法
4	Acceptable under IFRSs	國際財務報導準則可接受
5	Accountability of management	管理階層之課責性
6	Accounting	會計
7	Accounting and Reporting by Retirement Benefit Plans	退休福利計畫之會計與報導
8	Accounting estimate	會計估計
9	Accounting for Government Grants and Disclosure of Government Assistance	政府補助之會計及政府補助之揭露
10	accounting for inventories	存貨會計
11	Accounting for Investments in Associates	投資關聯企業之會計
12	Accounting income	會計收益
13	Accounting model	會計模式
14	Accounting period	會計期間
15	Accounting Policies, Changes in Accounting Estimates and Errors	會計政策、會計估計變動及錯誤
16	Accounting policy	會計政策
17	Accounting principle	會計原則
18	Accounting profit	會計利潤
19	Accounting record	會計紀錄
20	Accounting treatment	會計處理
21	Accounts receivable	應收帳款
22	Accrual basis	應計基礎
23	Accrual basis of accounting	應計基礎會計
24	Accrued liabilities	應計負債
25	Accumulated (amortisation, interest, profit or loss)	累計(攤銷、利息、損益)
26	Accumulated profit or loss	累計損益
27	Accumulating compensated absences	累積帶薪假
28	Achieve comparability	達成可比性/達成...可比性
29	acquired entity	(被)收購(之)個體
30	acquired goodwill	收購(之)商譽

Appendix 3.1.4 Hyphenating Process with Useful File Utilities



The left part of the above figure shows the window of Batch Replacer in Useful File Utilities, while the right part shows texts in the BRC that will undergo the hyphenating process. The window of Batch Replacer shows that accounting compounds will be replaced with their hyphenated equivalents.

Find and Replace string

Name	Size before	Size after	Changed size	Amount	Status
d:\Documents(Mathesis\Form20F_35x3_replace_compound\0100_Agria_華興物種\0100_Agria_2007-12-31.htm	1 328 548	1 328 539	-9	1 374	successful
d:\Documents(Mathesis\Form20F_35x3_replace_compound\0100_Agria_華興物種\0100_Agria_2008-12-31.htm	1 372 332	1 372 312	-20	1 539	successful
d:\Documents(Mathesis\Form20F_35x3_replace_compound\0100_Agria_華興物種\0100_Agria_2009-12-31.htm	1 494 685	1 494 663	-22	1 587	successful
d:\Documents(Mathesis\Form20F_35x3_replace_compound\1221_Yanzhou_兗州煤業\1221_Yanzhou_2007-12-31.htm	2 185 789	2 185 743	-46	1 514	successful
d:\Documents(Mathesis\Form20F_35x3_replace_compound\1221_Yanzhou_兗州煤業\1221_Yanzhou_2008-12-31.htm	2 333 654	2 333 608	-46	1 659	successful
d:\Documents(Mathesis\Form20F_35x3_replace_compound\1221_Yanzhou_兗州煤業\1221_Yanzhou_2009-12-31.htm	2 776 532	2 776 482	-50	2 052	successful
d:\Documents(Mathesis\Form20F_35x3_replace_compound\1311_CNOOC_中國海洋石油\1311_CNOOC_2007-12-31.htm	9 741 305	9 741 269	-36	1 870	successful
d:\Documents(Mathesis\Form20F_35x3_replace_compound\1311_CNOOC_中國海洋石油\1311_CNOOC_2008-12-31.htm	6 054 678	6 054 660	-18	1 447	successful
d:\Documents(Mathesis\Form20F_35x3_replace_compound\1311_CNOOC_中國海洋石油\1311_CNOOC_2009-12-31.htm	8 581 705	8 581 690	-15	1 360	successful
d:\Documents(Mathesis\Form20F_35x3_replace_compound\1520_Xinyuan_鑫苑置業\1520_Xinyuan_2007-12-31.htm	2 050 557	2 050 557	0	1 644	successful
d:\Documents(Mathesis\Form20F_35x3_replace_compound\1520_Xinyuan_鑫苑置業\1520_Xinyuan_2008-12-31.htm	2 481 008	2 481 008	0	2 187	successful
d:\Documents(Mathesis\Form20F_35x3_replace_compound\1520_Xinyuan_鑫苑置業\1520_Xinyuan_2009-12-31.htm	2 500 336	2 500 332	-4	1 839	successful
d:\Documents(Mathesis\Form20F_35x3_replace_compound\2821_Sinopec_中國石化\2821_Sinopec_2007-12-31.htm	2 401 401	2 401 365	-36	1 960	successful
d:\Documents(Mathesis\Form20F_35x3_replace_compound\2821_Sinopec_中國石化\2821_Sinopec_2008-12-31.htm	2 425 212	2 425 177	-35	1 941	successful
d:\Documents(Mathesis\Form20F_35x3_replace_compound\2821_Sinopec_中國石化\2821_Sinopec_2009-12-31.htm	2 737 120	2 737 081	-39	2 156	successful
d:\Documents(Mathesis\Form20F_35x3_replace_compound\2834_3SBio_三生製藥\2834_3SBio_2007-12-31.htm	1 382 769	1 382 743	-26	1 216	successful
d:\Documents(Mathesis\Form20F_35x3_replace_compound\2834_3SBio_三生製藥\2834_3SBio_2008-12-31.htm	1 489 403	1 489 376	-27	1 344	successful
d:\Documents(Mathesis\Form20F_35x3_replace_compound\2834_3SBio_三生製藥\2834_3SBio_2009-12-31.htm	1 622 690	1 622 663	-27	1 467	successful
d:\Documents(Mathesis\Form20F_35x3_replace_compound\2860_Gushan_古杉生物藥油\2860_Gushan_2007-12-31.htm	1 298 431	1 298 400	-31	1 500	successful
d:\Documents(Mathesis\Form20F_35x3_replace_compound\2860_Gushan_古杉生物藥油\2860_Gushan_2008-12-31.htm	1 375 905	1 375 875	-30	1 638	successful
d:\Documents(Mathesis\Form20F_35x3_replace_compound\2860_Gushan_古杉生物藥油\2860_Gushan_2009-12-31.htm	1 536 481	1 536 447	-34	1 672	successful
d:\Documents(Mathesis\Form20F_35x3_replace_compound\2911_China_P_C_中國石油化工\2911_China_P_C_2007-12-31.htm	7 417 871	7 417 855	-16	1 180	successful
d:\Documents(Mathesis\Form20F_35x3_replace_compound\2911_China_P_C_中國石油化工\2911_China_P_C_2008-12-31.htm	8 312 844	8 312 824	-20	1 303	successful
d:\Documents(Mathesis\Form20F_35x3_replace_compound\2911_China_P_C_中國石油化工\2911_China_P_C_2009-12-31.htm	7 236 154	7 236 117	-37	1 874	successful

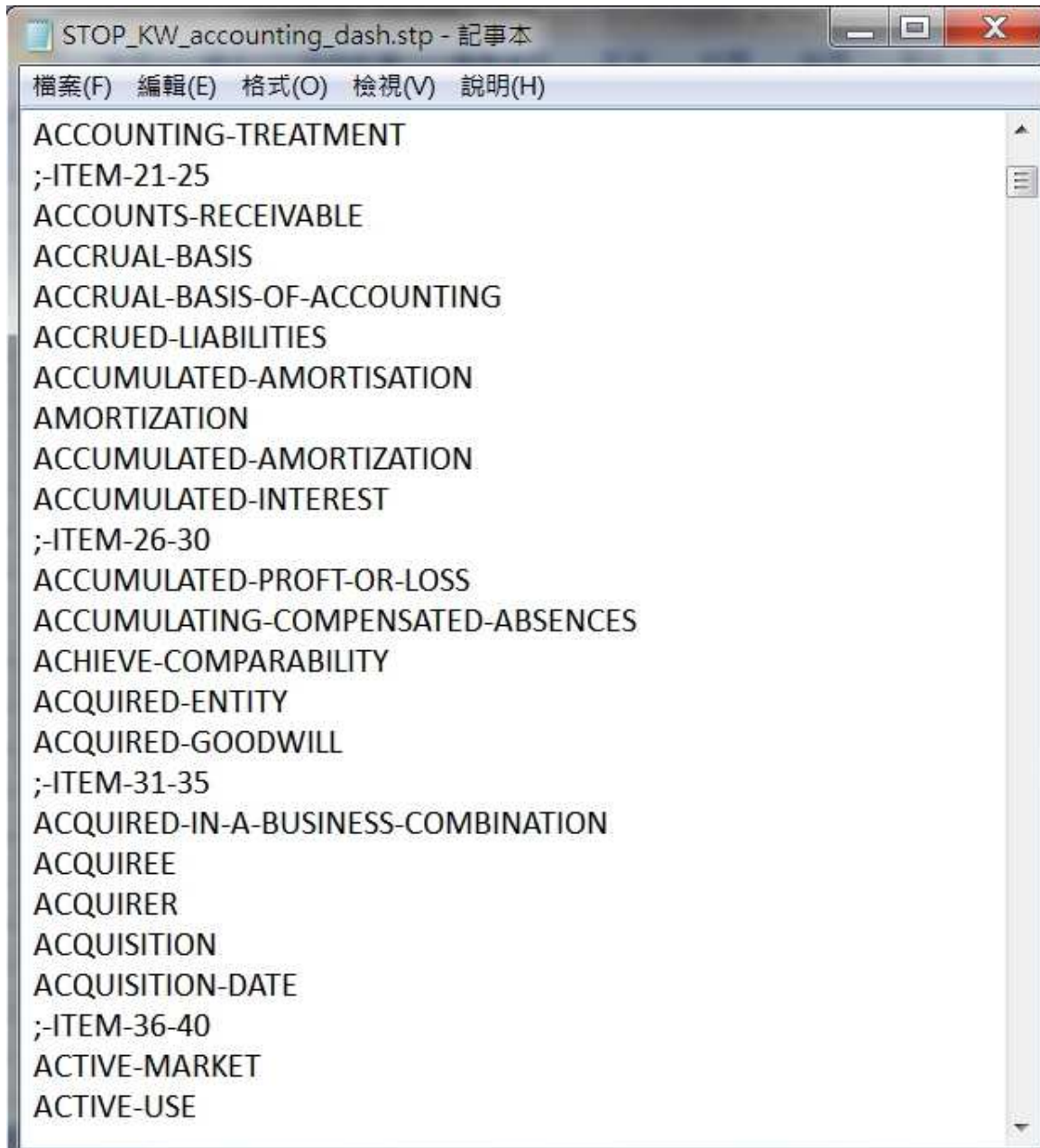
	Number of files	Size before	Size after	Changed size
All files	174	300 687 367	300 685 744	-1 623
All unchanged	69	880 909	880 909	0
All successful	105	299 806 458	299 804 835	-1 623
All failed	0	—	—	—

Files to show: All Successful Unchanged Failed

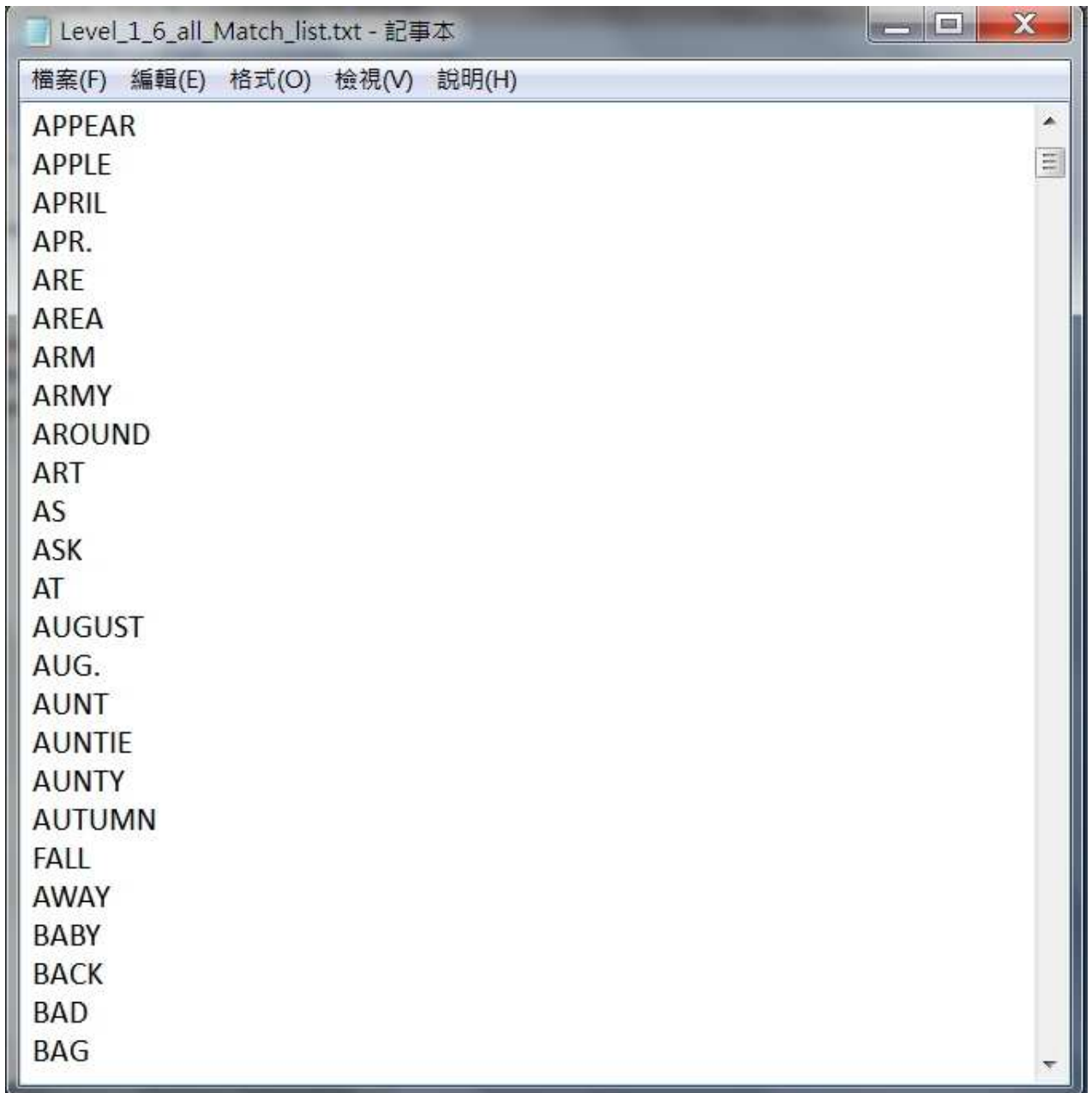
Operation completed!

This figure displays results of the hyphenating process, in which 105 files in BRC were all processed, with size downsized for 1,623 bytes. After this procedure, all accounting compounds were added with dash in between, which will be seen as single words by corpus software.

Appendix 3.2.1 Part of Stoplist (Hyphenated Accounting Compounds)



Appendix 3.2.2 Part of Match-list (Level 1 of SHEWR)



Appendix 3.3.1 Contents of Subdivision A (Reportage) in Brown Corpus

Topics	Number of Samples		
	Daily	Weekly	Total
Political	10	4	14
Sports	5	2	7
Society	3	0	3
Spot News	7	2	9
Financial	3	1	4
Cultural	5	2	7
Total			44

Appendix 3.3.2 Contents of Subdivision H (Miscellaneous) in Brown Corpus

Topics	Number of Samples
Government Documents	24
Foundation Reports	2
Industry Reports	2
College Catalog	1
Industry House organ	1
Total	30

Appendix 4.1.1 Complete Key-Word List Assorted According to Text Coverage

N	KW	Texts	%	Overall Freq.	Level in SHEWR
1	ANNUAL	105	100	9,168	4
2	APPLICABLE	105	100	5,651	6
3	CHINA	105	100	26,427	3
4	DECEMBER	105	100	30,234	1
5	EXCHANGE	105	100	11,302	3
6	FINANCIAL	105	100	14,944	4
7	INCLUDING	105	100	11,142	4
8	OTHER	105	100	33,559	1
9	SIGNIFICANT	105	100	7,013	3
10	SUBJECT	104	99	10,560	2
11	FOREIGN	103	98	11,109	1
12	INTERNAL	103	98	5,003	3
13	MARKET	103	98	9,126	1
14	RELEVANT	103	98	5,000	6
15	INFORMATION	102	97	12,089	4
16	OR	102	97	76,832	1
17	OUR	102	97	127,492	1
18	ACCORDANCE	101	96	4,531	6
19	COMPANY	101	96	35,198	2

N	KW	Texts	%	Overall Freq.	Level in SHEWR
20	CORPORATE	101	96	4,387	6
21	NET	101	96	14,272	2
22	REPORT	101	96	7,680	1
23	DUE	100	95	8,366	3
24	JANUARY	100	95	7,435	1
25	TOTAL	100	95	14,764	1
26	DATE	99	94	7,116	1
27	PRIOR	99	94	4,185	5
28	UNDER	99	94	19,897	1
29	US	99	94	37,490	1
30	MILLION	98	93	28,419	2
31	CURRENT	97	92	5,179	3
32	DECREASE	96	91	3,578	4
33	LOSS	96	91	7,119	2
34	MAY	96	91	28,456	1
35	RATE	96	91	8,460	3
36	AFFECT	95	90	3,822	3
37	PURCHASE	95	90	4,174	5
38	AMOUNT	94	89	7,795	2
39	CERTAIN	94	89	8,701	1

N	KW	Texts	%	Overall Freq.	Level in SHEWR
40	INCREASE	94	89	9,775	2
41	WE	93	88	58,840	1
42	ADDITION	92	87	6,179	2
43	ARE	92	87	48,535	1
44	YEAR	92	87	18,640	1
45	EFFECTIVE	91	86	4,941	2
46	FUTURE	91	86	7,551	2
47	STOCK	91	86	7,068	5&6
48	TABLE	90	85	15,295	1&2
49	FOLLOWING	89	84	6,390	2
50	OUTSTANDING	89	84	3,759	4
51	OWNERSHIP	89	84	2,639	3
52	AGREEMENT	88	83	8,012	1
53	ANY	88	83	21,469	1
54	CONTENTS	88	83	13,870	4
55	INDEPENDENT	88	83	4,034	2
56	BASIS	87	82	5,469	2
57	PERIOD	87	82	6,771	2
58	VALUE	87	82	6,150	2
59	AND	85	80	216,535	1

N	KW	Texts	%	Overall Freq.	Level in SHEWR
60	RESPECT	84	80	4,866	2
61	PRINCIPAL	83	79	3,716	2
62	SENIOR	83	79	3,175	4
63	APPROVAL	82	78	3,160	4
64	DIRECTOR	82	78	5,694	2
65	EXECUTIVE	82	78	3,616	5
66	PRICE	81	77	4,755	1
67	DOLLAR	80	76	2,832	1
68	EXPENSE	80	76	4,118	3
69	REGARDING	80	76	2,344	4
70	COMMITTEE	79	75	5,490	3
71	INCLUDE	78	74	3,473	2
72	OF	77	73	261,173	1
73	TECHNOLOGY	77	73	5,624	3
74	EFFECT	76	72	4,905	2
75	ADDITIONAL	75	71	4,008	3
76	LAW	75	71	6,795	1
77	PAYMENT	74	70	2,629	1
78	SUCH	74	70	14,225	1
79	DISTRIBUTION	73	69	4,180	4

N	KW	Texts	%	Overall Freq.	Level in SHEWR
80	GLOBAL	73	69	2,569	3
81	CUSTOMER	71	67	4,198	2
82	OBTAIN	71	67	2,041	4
83	REGISTRATION	71	67	3,575	4
84	ABILITY	70	66	3,439	2
85	LEGAL	68	64	2,424	2
86	HOLDER	66	62	1,950	2
87	PER	63	60	10,202	2&4
88	CURRENCY	62	59	1,231	5
89	EMPLOYEE	62	59	1,947	3
90	IMPACT	62	59	2,260	4
91	PROVIDE	60	57	3,828	2
92	EXERCISE	59	56	2,182	2
93	PLAN	58	55	4,384	1&5
94	FINANCE	57	54	3,038	4
95	INDUSTRY	57	54	3,085	2
96	REGULATION	57	54	1,345	4
97	APRIL	56	53	2,840	1
98	DEPOSIT	56	53	1,337	3
99	EXHIBIT	56	53	2,356	4

N	KW	Texts	%	Overall Freq.	Level in SHEWR
100	PORTION	56	53	1,905	3
101	INSURANCE	55	52	9,105	4
102	JUNE	55	52	3,051	1
103	PROPERTY	55	52	3,988	3
104	RESIDENT	54	51	1,144	5
105	SOFTWARE	54	51	4,714	4
106	DOMESTIC	53	50	2,601	3
107	OVERSEAS	53	50	1,436	2
108	QUARTER	53	50	1,576	2
109	SERVICE	53	50	6,552	1
110	DATA	52	49	3,876	2
111	FORTH	51	48	1,675	3
112	PASSIVE	51	48	800	4
113	REFERENCE	51	48	2,208	4
114	FUND	50	47	2,101	3
115	PERCENTAGE	50	47	1,716	4
116	RESPECTIVE	50	47	970	6
117	FEE	48	45	1,824	2
118	LICENSE	48	45	1,909	4
119	OVERALL	48	45	1,004	5

N	KW	Texts	%	Overall Freq.	Level in SHEWR
120	STATEMENT	48	45	2,904	1
121	CONTINUE	47	44	1,687	1
122	INTERNATIONAL	47	44	4,584	2
123	MARCH	47	44	4,992	1&3
124	AVERAGE	46	43	2,919	3
125	BENEFICIAL	45	42	791	5
126	COMPREHENSIVE	45	42	958	6
127	NETWORK	45	42	3,325	3
128	GRANT	44	41	1,521	5
129	INCENTIVE	44	41	870	6
130	SUBSTANTIAL	44	41	1,246	5
131	CORPORATION	43	40	3,008	5
132	NUMBER	43	40	4,384	1
133	SAFE	43	40	1,422	1
134	CONNECTION	42	40	1,399	3
135	EXPIRE	42	40	461	6
136	PRODUCT	42	40	3,395	3
137	RECEIVE	42	40	1,265	1
138	NOTICE	41	39	1,393	1
139	AUGUST	40	38	1,768	1

N	KW	Texts	%	Overall Freq.	Level in SHEWR
140	OTHERWISE	40	38	2,191	4
141	SUMMARY	40	38	918	3
142	TRANSLATION	40	38	1,380	4
143	AS	39	37	29,709	1
144	CONDITION	39	37	1,341	3
145	FILE	39	37	1,759	3
146	INTERNET	39	37	3,465	4
147	ASSURANCE	38	36	730	4
148	GROWTH	38	36	2,971	2
149	MAINTAIN	38	36	998	2
150	MEASURES	38	36	1,238	4
151	CONDUCT	37	35	922	5
152	ENVIRONMENTAL	37	35	1,658	3
153	MINISTRY	37	35	1,036	4
154	ADMINISTRATIVE	36	34	1,032	6
155	CONSIST	36	34	610	4
156	JULY	36	34	1,631	1
157	OFFER	36	34	1,177	2
158	PRODUCTION	36	34	6,016	4
159	RELY	36	34	529	3

N	KW	Texts	%	Overall Freq.	Level in SHEWR
160	CHAIRMAN	35	33	1,512	5
161	EXCEPT	34	32	2,480	1
162	MEETING	34	32	2,086	2
163	OCTOBER	34	32	1,636	1
164	OFFICER	34	32	1,161	1
165	BRAND	33	31	1,028	2
166	EXPAND	33	31	554	4
167	WEBSITE	33	31	722	4
168	COMPUTER	32	30	741	2
169	EQUIPMENT	32	30	1,998	4
170	IN	32	30	72,723	1
171	NOVEMBER	32	30	1,723	1
172	RETAIN	32	30	465	4
173	TERM	32	30	1,894	2
174	TRADEMARK	32	30	673	5
175	CREDIT	31	29	2,539	3
176	GAIN	31	29	1,165	2
177	PLEDGE	31	29	650	5
178	DISPOSAL	30	28	647	6
179	REQUIRE	30	28	979	2

N	KW	Texts	%	Overall Freq.	Level in SHEWR
180	SELL	30	28	905	1
181	SUPERVISOR	30	28	747	5
182	ACCORDINGLY	29	27	639	6
183	ADMINISTRATION	29	27	2,007	6
184	BALANCE	29	27	1,324	3
185	DEPUTY	29	27	1,349	6
186	FIN	29	27	433	5
187	KEY	29	27	1,064	1
188	SEPTEMBER	29	27	1,257	1
189	OFFERING	28	26	725	6
190	ASSURE	27	25	599	4
191	CONSTRUCTION	27	25	2,283	4
192	FEBRUARY	27	25	1,184	1
193	NOTE	27	25	2,559	1
194	VICE	27	25	1,159	6
195	ACQUIRE	26	24	490	4
196	CAPACITY	26	24	1,713	4
197	COMPETITIVE	26	24	592	4
198	CONTENT	26	24	2,083	4
199	GENERATE	26	24	382	6

N	KW	Texts	%	Overall Freq.	Level in SHEWR
200	ON	26	24	21,126	1
201	PROVINCE	26	24	1,157	5
202	ASSOCIATION	25	23	1,318	4
203	EMPLOYMENT	25	23	713	3
204	EXCLUSIVE	25	23	739	6
205	FROM	25	23	13,754	1
206	MANAGER	25	23	1,149	3
207	PROVINCIAL	25	23	460	6
208	SUBSEQUENT	25	23	508	6
209	ACCOUNTANT	24	22	336	4
210	ORDINARY	24	22	662	2
211	BELIEVE	23	21	1,285	1
212	CERTIFICATE	23	21	479	5
213	DECLINE	23	21	609	6
214	DISTRIBUTE	23	21	377	4
215	ENSURE	23	21	443	5
216	EXPIRATION	23	21	256	6
217	MEDIA	23	21	1,357	3
218	RAW	23	21	1,149	3
219	TRANSPORTATION	23	21	1,814	4

N	KW	Texts	%	Overall Freq.	Level in SHEWR
220	COMPETITION	22	20	597	4
221	DEMAND	22	20	920	4
222	ELIGIBLE	22	20	492	6
223	PAY	22	20	1,050	1&3
224	UNLESS	22	20	1,547	3
225	CIRCULAR	21	20	467	4
226	INTELLECTUAL	21	20	634	4
227	PROJECT	21	20	1,668	2
228	SUBSCRIPTION	21	20	560	6
229	ENHANCE	20	19	208	6
230	EXCESS	20	19	478	5
231	GUIDANCE	20	19	663	3
232	LOCAL	20	19	1,661	2
233	REPUBLIC	20	19	869	3
234	SCHEME	20	19	1,001	5
235	BILLION	19	18	4,157	3
236	BONUS	19	18	345	5
237	COMMERCIAL	19	18	1,668	3
238	DELIVERY	19	18	653	3
239	EXPECT	19	18	678	2

N	KW	Texts	%	Overall Freq.	Level in SHEWR
240	FORM	19	18	1,637	2
241	GENERAL	19	18	2,306	1&2
242	PROCESS	19	18	929	3
243	REDUCTION	19	18	661	4
244	RETIREMENT	19	18	471	4
245	AWARD	18	17	667	3
246	DIGITAL	18	17	1,241	4&6
247	REASONABLE	18	17	501	3
248	RENTAL	18	17	316	6
249	TERMINATE	18	17	255	6
250	ACT	17	16	928	1
251	ARRANGEMENT	17	16	367	2
252	BOARD	17	16	1,550	2
253	EXPANSION	17	16	581	4
254	FACILITY	17	16	372	4
255	OPERATION	17	16	889	4
256	PLANT	17	16	2,068	1
257	STRATEGIC	17	16	439	6
258	VARIABLE	17	16	501	6
259	APPLICATION	16	15	469	4

N	KW	Texts	%	Overall Freq.	Level in SHEWR
260	BRANCH	16	15	798	2
261	GROSS	16	15	1,147	5
262	HEDGE	16	15	312	5
263	MANAGE	16	15	350	3
264	OPERATE	16	15	522	2
265	PLATFORM	16	15	586	2
266	QUALIFICATION	16	15	410	6
267	QUALITY	16	15	532	2
268	SUPPLY	16	15	788	2
269	ACCESS	15	14	553	4
270	BACHELOR	15	14	163	5
271	CODE	15	14	382	4&5
272	COMMON	15	14	3,159	1
273	ENGINEERING	15	14	638	4
274	ESTATE	15	14	3,423	5
275	EXPORT	15	14	744	3
276	EXTENT	15	14	645	4
277	LAND	15	14	1,970	1
278	MAINTENANCE	15	14	988	5
279	MONTHLY	15	14	432	4

N	KW	Texts	%	Overall Freq.	Level in SHEWR
280	OIL	15	14	4,416	1
281	PETROLEUM	15	14	1,691	6
282	RATIO	15	14	777	5
283	RETAIL	15	14	851	6
284	STRATEGY	15	14	482	3
285	THESE	15	14	2,651	1
286	TRADE	15	14	865	2
287	TRANSMISSION	15	14	488	6
288	BEHALF	14	13	289	5
289	FAIL	14	13	282	2
290	FUEL	14	13	758	4
291	INTEND	14	13	215	4
292	MOBILE	14	13	3,394	3&5
293	PATENT	14	13	490	5
294	PROTECTION	14	13	518	3
295	RECEIPT	14	13	200	3
296	REDUCE	14	13	307	3
297	RULE	14	13	398	1
298	SERIES	14	13	1,009	5
299	UNIT	14	13	952	1

N	KW	Texts	%	Overall Freq.	Level in SHEWR
300	BELOW	13	12	695	1
301	CONSUMER	13	12	759	4
302	CRUDE	13	12	1,474	6
303	EACH	13	12	1,579	1
304	POTENTIAL	13	12	321	5
305	SURPLUS	13	12	318	6
306	WITHIN	13	12	1,708	2
307	ASSESSMENT	12	11	322	6
308	CHEMICAL	12	11	948	2
309	CHIEF	12	11	605	1
310	DURING	12	11	1,969	1
311	ECONOMIC	12	11	1,019	4
312	EXTERNAL	12	11	543	5
313	GAS	12	11	2,463	1&3
314	BUREAU	11	10	348	5
315	BY	11	10	10,214	1
316	CHANGE	11	10	1,162	2
317	DEVELOP	11	10	291	2
318	DOUBTFUL	11	10	184	3
319	EASTERN	11	10	827	2

N	KW	Texts	%	Overall Freq.	Level in SHEWR
320	FOR	11	10	10,945	1
321	INVEST	11	10	247	4
322	ITS	11	10	3,813	1
323	JOINT	11	10	430	2
324	MANUFACTURE	11	10	258	4
325	MINORITY	11	10	202	3
326	SAFETY	11	10	464	2
327	TARIFF	11	10	433	6
328	TRAVEL	11	10	1,054	2
329	ADVERTISEMENT	10	9	170	3
330	ATTRACT	10	9	138	3
331	COMPETE	10	9	143	3
332	CONSENT	10	9	146	5
333	COVERAGE	10	9	177	6
334	ENGAGE	10	9	148	3
335	EXPERTISE	10	9	135	6
336	EXPLORATION	10	9	1,141	6
337	FURTHERMORE	10	9	181	4
338	INVENTORY	10	9	325	6
339	PARTY	10	9	837	1

N	KW	Texts	%	Overall Freq.	Level in SHEWR
340	PHONE	10	9	449	2&5
341	SHALL	10	9	933	1
342	STAMP	10	9	135	2
343	VIRGIN	10	9	225	4
344	AIR	9	8	1,370	1
345	AIRLINE	9	8	326	2
346	COAL	9	8	4,592	2
347	CONSUMPTION	9	8	346	6
348	DRUG	9	8	350	2
349	ELECTRICITY	9	8	720	3
350	EQUIVALENT	9	8	238	6
351	EXCEED	9	8	273	5
352	GOODS	9	8	235	4
353	IMPLEMENT	9	8	97	6
354	INSTITUTE	9	8	267	5
355	INTERPRETATION	9	8	206	5
356	LINE	9	8	972	1
357	MAINLAND	9	8	784	5
358	MEDICAL	9	8	859	3
359	REGIONAL	9	8	564	3

N	KW	Texts	%	Overall Freq.	Level in SHEWR
360	RESERVATION	9	8	200	4
361	RESIDENTIAL	9	8	706	6
362	RESTRICT	9	8	119	3
363	SECURITY	9	8	256	3
364	SUPERVISION	9	8	155	6
365	SYSTEM	9	8	1,161	3
366	THIRD	9	8	568	1
367	TON	9	8	248	3
368	TRAFFIC	9	8	616	2
369	update	9	8	121	5
370	USAGE	9	8	639	4
371	AGENT	8	7	254	4
372	APPLY	8	7	227	2
373	AUTHORITY	8	7	354	4
374	CLIENT	8	7	319	3
375	COUNSEL	8	7	131	5
376	ELECTRONIC	8	7	620	3
377	ENDING	8	7	175	2
378	ENERGY	8	7	757	2
379	ENFORCE	8	7	106	4

N	KW	Texts	%	Overall Freq.	Level in SHEWR
380	ENFORCEMENT	8	7	136	4
381	EXPOSURE	8	7	671	4
382	FURTHER	8	7	395	2
383	INDIVIDUAL	8	7	819	3
384	POWER	8	7	4,554	1
385	PRODUCE	8	7	256	2
386	REFORM	8	7	215	4
387	REGION	8	7	548	2
388	SEARCH	8	7	888	2
389	SITE	8	7	318	4
390	TREASURY	8	7	373	5
391	VOLUME	8	7	394	3
392	WHOLESALE	8	7	182	5
393	YIELD	8	7	328	5
394	ACCOUNT	7	6	610	3
395	CALCULATION	7	6	127	4
396	CARGO	7	6	587	4
397	CUMULATIVE	7	6	237	6
398	DEVICE	7	6	239	4
399	DIRECT	7	6	1,433	1

N	KW	Texts	%	Overall Freq.	Level in SHEWR
400	ECONOMICS	7	6	103	4
401	EVALUATION	7	6	153	4
402	FAIR	7	6	318	2
403	MAJORITY	7	6	194	3
404	MIX	7	6	143	2
405	NEWS	7	6	306	1
406	OPERATIONAL	7	6	288	6
407	PACKAGE	7	6	277	2
408	PRIMARY	7	6	1,312	3
409	REGISTER	7	6	133	4
410	RELATE	7	6	69	3
411	REPRESENT	7	6	142	3
412	RESOURCE	7	6	593	3
413	STAFF	7	6	330	3
414	SUBSCRIBE	7	6	61	6
415	TECHNICAL	7	6	288	3
416	TELEVISION	7	6	353	2&4
417	treatment	7	6	377	2
418	USE	7	6	940	1
419	VIOLATION	7	6	167	4

N	KW	Texts	%	Overall Freq.	Level in SHEWR
420	waste	7	6	142	1
421	WEB	7	6	293	3
422	AIRPORT	6	5	542	1
423	APPRECIATION	6	5	199	4
424	BASE	6	5	347	1
425	BETWEEN	6	5	738	1
426	CALENDAR	6	5	118	2
427	CANCER	6	5	215	2
428	CASUALTY	6	5	330	6
429	CAUSE	6	5	146	1
430	CHAIN	6	5	464	3
431	CLINICAL	6	5	369	6
432	COMMERCE	6	5	162	4
433	CONSULT	6	5	72	4
434	DISCOUNT	6	5	102	3
435	EAST	6	5	378	1
436	FIBER	6	5	159	5
437	FREIGHT	6	5	641	5
438	GENERATION	6	5	423	4
439	HOTEL	6	5	1,235	2

N	KW	Texts	%	Overall Freq.	Level in SHEWR
440	IMPORT	6	5	339	3
441	LIFE	6	5	2,437	1
442	MEDICINE	6	5	208	2
443	MINE	6	5	978	2
444	NATURAL	6	5	900	2
445	PAGE	6	5	904	1
446	PASSENGER	6	5	976	2
447	PERMIT	6	5	229	3
448	POLLUTION	6	5	90	4
449	PORT	6	5	121	2
450	PREDECESSOR	6	5	116	6
451	PREDICT	6	5	69	4
452	PREPARATION	6	5	416	3
453	PRIVATE	6	5	881	2
454	PROMOTION	6	5	164	4
455	REAL	6	5	3,058	1
456	RECOGNIZE	6	5	156	3
457	RENEW	6	5	66	3
458	SECRETARY	6	5	300	2
459	SPECTRUM	6	5	184	6

N	KW	Texts	%	Overall Freq.	Level in SHEWR
460	SYNTHETIC	6	5	382	6
461	TEST	6	5	1,857	2
462	TICKET	6	5	134	1
463	TV	6	5	1,665	2
464	UPON	6	5	443	2
465	USEFUL	6	5	152	1
466	VIDEO	6	5	238	2&4
467	ACCEPTANCE	5	4	134	4
468	CATALOGUE	5	4	89	4
469	CENT	5	4	3,405	1&4
470	CHARGE	5	4	223	2
471	DEPARTMENT	5	4	299	2
472	ELECTRIC	5	4	277	3
473	FAILURE	5	4	123	2
474	GASOLINE	5	4	158	3
475	HARDWARE	5	4	136	4
476	HOUSING	5	4	163	5
477	IRON	5	4	177	1
478	LEARNING	5	4	401	4
479	METHOD	5	4	275	2

N	KW	Texts	%	Overall Freq.	Level in SHEWR
480	POLICY	5	4	690	2
481	SATELLITE	5	4	185	4
482	SEED	5	4	691	1
483	STRUCTURE	5	4	148	3
484	SUPPORT	5	4	291	2
485	THROUGH	5	4	976	2
486	TOURISM	5	4	39	3
487	UNDERGROUND	5	4	147	2
488	UPGRADE	5	4	51	6
489	VALID	5	4	86	6
490	ACCIDENT	4	3	398	3
491	ADVISOR	4	3	32	3
492	AI	4	3	125	5
493	APPROPRIATE	4	3	320	4
494	ARTICLE	4	3	217	2&4
495	AVAILABLE	4	3	298	3
496	BARREL	4	3	70	3
497	BLUE	4	3	204	1
498	CABLE	4	3	84	2
499	CELLULAR	4	3	683	5

N	KW	Texts	%	Overall Freq.	Level in SHEWR
500	CHANNEL	4	3	94	3
501	CHIP	4	3	76	3
502	CONCESSION	4	3	541	6
503	DEGREE	4	3	198	2
504	DISCHARGE	4	3	70	6
505	DISPOSE	4	3	48	5
506	DISTANCE	4	3	446	2
507	DIVISION	4	3	163	2
508	ELECTION	4	3	99	3
509	ENGINEER	4	3	89	3
510	ENTERTAINMENT	4	3	83	4
511	ETHICS	4	3	64	5
512	EVALUATE	4	3	60	4
513	HEALTH	4	3	474	1
514	ISSUE	4	3	309	5
515	LIMIT	4	3	60	2
516	LOYALTY	4	3	76	4
517	MINERAL	4	3	94	4
518	PHASE	4	3	347	6
519	PRESCRIPTION	4	3	116	6

N	KW	Texts	%	Overall Freq.	Level in SHEWR
520	RECOVERY	4	3	129	4
521	REFLECT	4	3	205	4
522	REPUTATION	4	3	47	4
523	REQUIREMENT	4	3	84	2
524	RESPONSIBLE	4	3	128	2
525	REWARD	4	3	76	4
526	SCHEDULE	4	3	80	3
527	SECTOR	4	3	213	6
528	SKY	4	3	106	1
529	STATION	4	3	175	1
530	SULFUR	4	3	25	5
531	TECHNOLOGICAL	4	3	46	4
532	THOUSAND	4	3	771	1
533	TO	4	3	4,419	1
534	TOPIC	4	3	86	2
535	TRANSPORT	4	3	70	3
536	TREATY	4	3	59	5
537	VOTE	4	3	99	2
538	WHARF	4	3	40	5
539	WHICH	4	3	1,608	1

N	KW	Texts	%	Overall Freq.	Level in SHEWR
540	ACID	3	2	115	4
541	ACTUAL	3	2	139	3
542	ADDRESS	3	2	79	1
543	ADVANCED	3	2	270	3
544	AGRICULTURAL	3	2	263	5
545	AIRCRAFT	3	2	905	2
546	ALUMINUM	3	2	3,506	4
547	AM	3	2	741	1&4
548	ANALYTICAL	3	2	65	6
549	APPOINT	3	2	38	4
550	ARTIST	3	2	148	2
551	ATM	3	2	98	4
552	AUCTION	3	2	89	6
553	AUDIO	3	2	55	4
554	AUTOMOBILE	3	2	116	3
555	AVIATION	3	2	460	6
556	BACKBONE	3	2	36	5
557	BAY	3	2	111	3
558	BEVERAGE	3	2	37	6
559	BIRD	3	2	126	1

N	KW	Texts	%	Overall Freq.	Level in SHEWR
560	BROADCAST	3	2	84	2
561	CALL	3	2	242	1
562	CARBON	3	2	111	5
563	CARD	3	2	352	1
564	CASUAL	3	2	90	3
565	CD	3	2	89	4
566	CELL	3	2	283	2
567	CHAIRWOMAN	3	2	18	5
568	CIRCUIT	3	2	67	5
569	CITY	3	2	358	1
570	CIVIL	3	2	223	3
571	CLASS	3	2	338	1
572	COMMENCE	3	2	51	6
573	COMMUNIST	3	2	127	5
574	COMPLEX	3	2	132	3
575	COMPRISE	3	2	39	6
576	CONVENIENCE	3	2	63	4
577	CORN	3	2	596	1
578	COURSE	3	2	302	1
579	CROP	3	2	62	2

N	KW	Texts	%	Overall Freq.	Level in SHEWR
580	CULTURE	3	2	114	2
581	CUSTOMS	3	2	71	5
582	DAILY	3	2	218	2
583	DECLARATION	3	2	123	5
584	DELIVER	3	2	48	2
585	DESIGN	3	2	121	2
586	DIAGNOSIS	3	2	62	6
587	DISPATCH	3	2	118	6
588	DISPLAY	3	2	106	2&6
589	DIVERSIFY	3	2	24	6
590	DIVERT	3	2	25	6
591	DOSAGE	3	2	33	6
592	DRILL	3	2	146	4
593	DRUGSTORE	3	2	475	2
594	DURATION	3	2	85	5
595	ECONOMY	3	2	118	4
596	EDUCATIONAL	3	2	242	3
597	EFFICIENCY	3	2	229	4
598	ELECTRICAL	3	2	89	3
599	ELECTRONICS	3	2	178	4

N	KW	Texts	%	Overall Freq.	Level in SHEWR
600	EMPLOYER	3	2	63	3
601	ENVIRONMENT	3	2	125	2
602	ESTABLISHMENT	3	2	75	4
603	ESTIMATE	3	2	59	4
604	EXTRACT	3	2	33	6
605	FERTILIZER	3	2	30	5
606	FISH	3	2	737	1
607	FLEET	3	2	71	6
608	FLIGHT	3	2	261	2
609	FORK	3	2	42	1
610	FORTUNE	3	2	153	3
611	FORWARD	3	2	70	2
612	FREQUENCY	3	2	90	4
613	GAME	3	2	212	1
614	GARDEN	3	2	309	1
615	GENETIC	3	2	36	6
616	GENIUS	3	2	139	4
617	GEOGRAPHICAL	3	2	336	5
618	GLORY	3	2	81	3
619	GRADUATE	3	2	80	3

N	KW	Texts	%	Overall Freq.	Level in SHEWR
620	GREEN	3	2	143	1
621	HANG	3	2	123	2
622	HISTORICAL	3	2	116	3
623	HOME	3	2	703	1
624	HUMAN	3	2	539	1
625	IDENTICAL	3	2	48	4
626	IMMUNE	3	2	27	4&6
627	INDICATION	3	2	54	4
628	INN	3	2	36	3
629	INSPECTION	3	2	118	4
630	INTEGRATION	3	2	93	6
631	INTELLIGENCE	3	2	62	4
632	INTERMEDIATE	3	2	224	4
633	JOB	3	2	179	1
634	KIN	3	2	26	5
635	KIT	3	2	34	3
636	LABEL	3	2	218	3
637	LAUNCH	3	2	78	4
638	LAYER	3	2	54	5
639	LEISURE	3	2	45	3

N	KW	Texts	%	Overall Freq.	Level in SHEWR
640	LINK	3	2	76	2
641	LIVESTOCK	3	2	69	5
642	LOCOMOTIVE	3	2	33	5
643	LOGO	3	2	27	5
644	LOWER	3	2	539	2
645	MACHINERY	3	2	105	4
646	MAIL	3	2	256	1
647	MAJOR	3	2	131	3
648	MAR	3	2	102	1&6
649	MATCH	3	2	77	1&2
650	MECHANISM	3	2	52	6
651	MEMBERSHIP	3	2	84	3
652	MICROSCOPE	3	2	39	4
653	MIGRATION	3	2	60	6
654	MINIMUM	3	2	139	4
655	MUSIC	3	2	1,305	1&4
656	NATIONAL	3	2	403	2
657	NEWSPAPER	3	2	97	1
658	NO	3	2	573	1
659	NORTH	3	2	614	1

N	KW	Texts	%	Overall Freq.	Level in SHEWR
660	ORIGIN	3	2	58	3
661	OUTDOOR	3	2	149	3
662	OUTGOING	3	2	59	5
663	OUTPUT	3	2	102	5
664	PACIFIC	3	2	295	5
665	PACT	3	2	124	6
666	PARTICIPANT	3	2	48	5
667	PEARL	3	2	67	3
668	PERSONAL	3	2	897	2
669	PIPE	3	2	351	2
670	POSTURE	3	2	99	6
671	PRESIDENT	3	2	239	2
672	PREVIOUS	3	2	117	3
673	PRINT	3	2	318	1
674	PROMOTE	3	2	50	3
675	QUALIFICATIONS	3	2	68	6
676	RAIL	3	2	169	5
677	RAILROAD	3	2	226	1
678	RECORD	3	2	136	2
679	RECRUIT	3	2	150	6

N	KW	Texts	%	Overall Freq.	Level in SHEWR
680	REFER	3	2	53	4
681	REMAINDER	3	2	81	6
682	REPAIR	3	2	63	3
683	RESOLUTION	3	2	77	4
684	REST	3	2	440	1
685	RESTORATION	3	2	85	6
686	RIDGE	3	2	181	5
687	ROSE	3	2	730	1
688	ROUND	3	2	118	1
689	ROUTE	3	2	153	4
690	RUBBER	3	2	55	1
691	RUN	3	2	261	1
692	SAVINGS	3	2	175	3
693	SEA	3	2	279	1
694	SECONDARY	3	2	184	3
695	SENSITIVITY	3	2	175	5
696	SHEEP	3	2	606	1
697	SILICON	3	2	45	6
698	SIMILAR	3	2	108	2
699	SLOT	3	2	44	6

N	KW	Texts	%	Overall Freq.	Level in SHEWR
700	SOLE	3	2	59	5
701	SOLUTION	3	2	148	2
702	SOUTH	3	2	220	1
703	SPARE	3	2	74	4
704	SPLENDID	3	2	181	4
705	SQUARE	3	2	495	2
706	STAPLE	3	2	38	6
707	STAR	3	2	199	1
708	STEEL	3	2	436	2
709	STORE	3	2	306	1
710	SUCCESS	3	2	130	2
711	SUCCESSFUL	3	2	97	2
712	SUPPLEMENT	3	2	68	6
713	TELEPHONE	3	2	442	2
714	TIGER	3	2	70	1
715	TIME	3	2	623	1
716	TRADITIONAL	3	2	219	2
717	TRAIN	3	2	237	1
718	TRIAL	3	2	122	2
719	TYPE	3	2	348	2

N	KW	Texts	%	Overall Freq.	Level in SHEWR
720	UNDERTAKE	3	2	68	6
721	UNIQUE	3	2	66	4
722	UNIVERSAL	3	2	134	4
723	URBAN	3	2	93	4
724	UTILIZE	3	2	34	6
725	VARIOUS	3	2	150	3
726	VEGETABLE	3	2	168	1
727	VENTURE	3	2	79	5
728	VILLAGE	3	2	109	2
729	VOICE	3	2	159	1
730	WATER	3	2	251	1
731	WEEKLY	3	2	199	4
732	WIRE	3	2	102	2
733	WISDOM	3	2	75	3

Appendix 4.1.2 Complete Key-Word List Assorted According to Overall Frequency

N	KW	Texts	%	Overall Freq.	Level in SHEWR
72	OF	77	73	261,173	1
59	AND	85	80	216,535	1
17	OUR	102	97	127,492	1
16	OR	102	97	76,832	1
170	IN	32	30	72,723	1
41	WE	93	88	58,840	1
43	ARE	92	87	48,535	1
29	US	99	94	37,490	1
19	COMPANY	101	96	35,198	2
8	OTHER	105	100	33,559	1
4	DECEMBER	105	100	30,234	1
143	AS	39	37	29,709	1
34	MAY	96	91	28,456	1
30	MILLION	98	93	28,419	2
3	CHINA	105	100	26,427	3
53	ANY	88	83	21,469	1
200	ON	26	24	21,126	1
28	UNDER	99	94	19,897	1
44	YEAR	92	87	18,640	1
48	TABLE	90	85	15,295	1&2

N	KW	Texts	%	Overall Freq.	Level in SHEWR
6	FINANCIAL	105	100	14,944	4
25	TOTAL	100	95	14,764	1
21	NET	101	96	14,272	2
78	SUCH	74	70	14,225	1
54	CONTENTS	88	83	13,870	4
205	FROM	25	23	13,754	1
15	INFORMATION	102	97	12,089	4
5	EXCHANGE	105	100	11,302	3
7	INCLUDING	105	100	11,142	4
11	FOREIGN	103	98	11,109	1
320	FOR	11	10	10,945	1
10	SUBJECT	104	99	10,560	2
315	BY	11	10	10,214	1
87	PER	63	60	10,202	2&4
40	INCREASE	94	89	9,775	2
1	ANNUAL	105	100	9,168	4
13	MARKET	103	98	9,126	1
101	INSURANCE	55	52	9,105	4
39	CERTAIN	94	89	8,701	1
35	RATE	96	91	8,460	3
23	DUE	100	95	8,366	3

N	KW	Texts	%	Overall Freq.	Level in SHEWR
52	AGREEMENT	88	83	8,012	1
38	AMOUNT	94	89	7,795	2
22	REPORT	101	96	7,680	1
46	FUTURE	91	86	7,551	2
24	JANUARY	100	95	7,435	1
33	LOSS	96	91	7,119	2
26	DATE	99	94	7,116	1
47	STOCK	91	86	7,068	5&6
9	SIGNIFICANT	105	100	7,013	3
76	LAW	75	71	6,795	1
57	PERIOD	87	82	6,771	2
109	SERVICE	53	50	6,552	1
49	FOLLOWING	89	84	6,390	2
42	ADDITION	92	87	6,179	2
58	VALUE	87	82	6,150	2
158	PRODUCTION	36	34	6,016	4
64	DIRECTOR	82	78	5,694	2
2	APPLICABLE	105	100	5,651	6
73	TECHNOLOGY	77	73	5,624	3
70	COMMITTEE	79	75	5,490	3
56	BASIS	87	82	5,469	2

N	KW	Texts	%	Overall Freq.	Level in SHEWR
31	CURRENT	97	92	5,179	3
12	INTERNAL	103	98	5,003	3
14	RELEVANT	103	98	5,000	6
123	MARCH	47	44	4,992	1&3
45	EFFECTIVE	91	86	4,941	2
74	EFFECT	76	72	4,905	2
60	RESPECT	84	80	4,866	2
66	PRICE	81	77	4,755	1
105	SOFTWARE	54	51	4,714	4
346	COAL	9	8	4,592	2
122	INTERNATIONAL	47	44	4,584	2
384	POWER	8	7	4,554	1
18	ACCORDANCE	101	96	4,531	6
533	TO	4	3	4,419	1
280	OIL	15	14	4,416	1
20	CORPORATE	101	96	4,387	6
93	NUMBER	43	40	4,384	1&5
132	PLAN	58	55	4,384	1
81	CUSTOMER	71	67	4,198	2
27	PRIOR	99	94	4,185	5
79	DISTRIBUTION	73	69	4,180	4

N	KW	Texts	%	Overall Freq.	Level in SHEWR
37	PURCHASE	95	90	4,174	5
235	BILLION	19	18	4,157	3
68	EXPENSE	80	76	4,118	3
55	INDEPENDENT	88	83	4,034	2
75	ADDITIONAL	75	71	4,008	3
103	PROPERTY	55	52	3,988	3
110	DATA	52	49	3,876	2
91	PROVIDE	60	57	3,828	2
36	AFFECT	95	90	3,822	3
322	ITS	11	10	3,813	1
50	OUTSTANDING	89	84	3,759	4
61	PRINCIPAL	83	79	3,716	2
65	EXECUTIVE	82	78	3,616	5
32	DECREASE	96	91	3,578	4
83	REGISTRATION	71	67	3,575	4
546	ALUMINUM	3	2	3,506	4
71	INCLUDE	78	74	3,473	2
146	INTERNET	39	37	3,465	4
84	ABILITY	70	66	3,439	2
274	ESTATE	15	14	3,423	5
469	CENT	5	4	3,405	1&4

N	KW	Texts	%	Overall Freq.	Level in SHEWR
136	PRODUCT	42	40	3,395	3
292	MOBILE	14	13	3,394	3&5
127	NETWORK	45	42	3,325	3
62	SENIOR	83	79	3,175	4
63	APPROVAL	82	78	3,160	4
272	COMMON	15	14	3,159	1
95	INDUSTRY	57	54	3,085	2
455	REAL	6	5	3,058	1
102	JUNE	55	52	3,051	1
94	FINANCE	57	54	3,038	4
131	CORPORATION	43	40	3,008	5
148	GROWTH	38	36	2,971	2
124	AVERAGE	46	43	2,919	3
120	STATEMENT	48	45	2,904	1
97	APRIL	56	53	2,840	1
67	DOLLAR	80	76	2,832	1
285	THESE	15	14	2,651	1
51	OWNERSHIP	89	84	2,639	3
77	PAYMENT	74	70	2,629	1
106	DOMESTIC	53	50	2,601	3
80	GLOBAL	73	69	2,569	3

N	KW	Texts	%	Overall Freq.	Level in SHEWR
193	NOTE	27	25	2,559	1
175	CREDIT	31	29	2,539	3
161	EXCEPT	34	32	2,480	1
313	GAS	12	11	2,463	1&3
441	LIFE	6	5	2,437	1
85	LEGAL	68	64	2,424	2
99	EXHIBIT	56	53	2,356	4
69	REGARDING	80	76	2,344	4
241	GENERAL	19	18	2,306	1&2
191	CONSTRUCTION	27	25	2,283	4
90	IMPACT	62	59	2,260	4
113	REFERENCE	51	48	2,208	4
140	OTHERWISE	40	38	2,191	4
92	EXERCISE	59	56	2,182	2
114	FUND	50	47	2,101	3
162	MEETING	34	32	2,086	2
198	CONTENT	26	24	2,083	4
256	PLANT	17	16	2,068	1
82	OBTAIN	71	67	2,041	4
183	ADMINISTRATION	29	27	2,007	6
169	EQUIPMENT	32	30	1,998	4

N	KW	Texts	%	Overall Freq.	Level in SHEWR
277	LAND	15	14	1,970	1
310	DURING	12	11	1,969	1
86	HOLDER	66	62	1,950	2
89	EMPLOYEE	62	59	1,947	3
118	LICENSE	48	45	1,909	4
100	PORTION	56	53	1,905	3
173	TERM	32	30	1,894	2
461	TEST	6	5	1,857	2
117	FEE	48	45	1,824	2
219	TRANSPORTATION	23	21	1,814	4
139	AUGUST	40	38	1,768	1
145	FILE	39	37	1,759	3
171	NOVEMBER	32	30	1,723	1
115	PERCENTAGE	50	47	1,716	4
196	CAPACITY	26	24	1,713	4
306	WITHIN	13	12	1,708	2
281	PETROLEUM	15	14	1,691	6
121	CONTINUE	47	44	1,687	1
111	FORTH	51	48	1,675	3
227	COMMERCIAL	19	18	1,668	2
237	PROJECT	21	20	1,668	3

N	KW	Texts	%	Overall Freq.	Level in SHEWR
463	TV	6	5	1,665	2
232	LOCAL	20	19	1,661	2
152	ENVIRONMENTAL	37	35	1,658	3
240	FORM	19	18	1,637	2
163	OCTOBER	34	32	1,636	1
156	JULY	36	34	1,631	1
539	WHICH	4	3	1,608	1
303	EACH	13	12	1,579	1
108	QUARTER	53	50	1,576	2
252	BOARD	17	16	1,550	2
224	UNLESS	22	20	1,547	3
128	GRANT	44	41	1,521	5
160	CHAIRMAN	35	33	1,512	5
302	CRUDE	13	12	1,474	6
107	OVERSEAS	53	50	1,436	2
399	DIRECT	7	6	1,433	1
133	SAFE	43	40	1,422	1
134	CONNECTION	42	40	1,399	3
138	NOTICE	41	39	1,393	1
142	TRANSLATION	40	38	1,380	4
344	AIR	9	8	1,370	1

N	KW	Texts	%	Overall Freq.	Level in SHEWR
217	MEDIA	23	21	1,357	3
185	DEPUTY	29	27	1,349	6
96	REGULATION	57	54	1,345	4
144	CONDITION	39	37	1,341	3
98	DEPOSIT	56	53	1,337	3
184	BALANCE	29	27	1,324	3
202	ASSOCIATION	25	23	1,318	4
408	PRIMARY	7	6	1,312	3
655	MUSIC	3	2	1,305	1&4
211	BELIEVE	23	21	1,285	1
137	RECEIVE	42	40	1,265	1
188	SEPTEMBER	29	27	1,257	1
130	SUBSTANTIAL	44	41	1,246	5
246	DIGITAL	18	17	1,241	4&6
150	MEASURES	38	36	1,238	4
439	HOTEL	6	5	1,235	2
88	CURRENCY	62	59	1,231	5
192	FEBRUARY	27	25	1,184	1
157	OFFER	36	34	1,177	2
176	GAIN	31	29	1,165	2
316	CHANGE	11	10	1,162	2

N	KW	Texts	%	Overall Freq.	Level in SHEWR
164	OFFICER	34	32	1,161	1
365	SYSTEM	9	8	1,161	3
194	VICE	27	25	1,159	6
201	PROVINCE	26	24	1,157	5
206	MANAGER	25	23	1,149	3
218	RAW	23	21	1,149	3
261	GROSS	16	15	1,147	5
104	RESIDENT	54	51	1,144	5
336	EXPLORATION	10	9	1,141	6
187	KEY	29	27	1,064	1
328	TRAVEL	11	10	1,054	2
223	PAY	22	20	1,050	1&3
153	MINISTRY	37	35	1,036	4
154	ADMINISTRATIVE	36	34	1,032	6
165	BRAND	33	31	1,028	2
311	ECONOMIC	12	11	1,019	4
298	SERIES	14	13	1,009	5
119	OVERALL	48	45	1,004	5
234	SCHEME	20	19	1,001	5
149	MAINTAIN	38	36	998	2
278	MAINTENANCE	15	14	988	5

N	KW	Texts	%	Overall Freq.	Level in SHEWR
179	REQUIRE	30	28	979	2
443	MINE	6	5	978	2
446	PASSENGER	6	5	976	2
485	THROUGH	5	4	976	2
356	LINE	9	8	972	1
116	RESPECTIVE	50	47	970	6
126	COMPREHENSIVE	45	42	958	6
299	UNIT	14	13	952	1
308	CHEMICAL	12	11	948	2
418	USE	7	6	940	1
341	SHALL	10	9	933	1
242	PROCESS	19	18	929	3
250	ACT	17	16	928	1
151	CONDUCT	37	35	922	5
221	DEMAND	22	20	920	4
141	SUMMARY	40	38	918	3
180	AIRCRAFT	3	2	905	1
545	SELL	30	28	905	2
445	PAGE	6	5	904	1
444	NATURAL	6	5	900	2
668	PERSONAL	3	2	897	2

N	KW	Texts	%	Overall Freq.	Level in SHEWR
255	OPERATION	17	16	889	4
388	SEARCH	8	7	888	2
453	PRIVATE	6	5	881	2
129	INCENTIVE	44	41	870	6
233	REPUBLIC	20	19	869	3
286	TRADE	15	14	865	2
358	MEDICAL	9	8	859	3
283	RETAIL	15	14	851	6
339	PARTY	10	9	837	1
319	EASTERN	11	10	827	2
383	INDIVIDUAL	8	7	819	3
112	PASSIVE	51	48	800	4
260	BRANCH	16	15	798	2
125	BENEFICIAL	45	42	791	5
268	SUPPLY	16	15	788	2
357	MAINLAND	9	8	784	5
282	RATIO	15	14	777	5
532	THOUSAND	4	3	771	1
301	CONSUMER	13	12	759	4
290	FUEL	14	13	758	4
378	ENERGY	8	7	757	2

N	KW	Texts	%	Overall Freq.	Level in SHEWR
181	SUPERVISOR	30	28	747	5
275	EXPORT	15	14	744	3
168	AM	3	2	741	2
547	COMPUTER	32	30	741	1&4
204	EXCLUSIVE	25	23	739	6
425	BETWEEN	6	5	738	1
606	FISH	3	2	737	1
147	ASSURANCE	38	36	730	4
687	ROSE	3	2	730	1
189	OFFERING	28	26	725	6
167	WEBSITE	33	31	722	4
349	ELECTRICITY	9	8	720	3
203	EMPLOYMENT	25	23	713	3
361	RESIDENTIAL	9	8	706	6
623	HOME	3	2	703	1
300	BELOW	13	12	695	1
482	SEED	5	4	691	1
480	POLICY	5	4	690	2
499	CELLULAR	4	3	683	5
239	EXPECT	19	18	678	2
174	TRADEMARK	32	30	673	5

N	KW	Texts	%	Overall Freq.	Level in SHEWR
381	EXPOSURE	8	7	671	4
245	AWARD	18	17	667	3
231	GUIDANCE	20	19	663	3
210	ORDINARY	24	22	662	2
243	REDUCTION	19	18	661	4
238	DELIVERY	19	18	653	3
177	PLEDGE	31	29	650	5
178	DISPOSAL	30	28	647	6
276	EXTENT	15	14	645	4
437	FREIGHT	6	5	641	5
182	ACCORDINGLY	29	27	639	6
370	USAGE	9	8	639	4
273	ENGINEERING	15	14	638	4
226	INTELLECTUAL	21	20	634	4
715	TIME	3	2	623	1
376	ELECTRONIC	8	7	620	3
368	TRAFFIC	9	8	616	2
659	NORTH	3	2	614	1
155	ACCOUNT	7	6	610	4
394	CONSIST	36	34	610	3
213	DECLINE	23	21	609	6

N	KW	Texts	%	Overall Freq.	Level in SHEWR
696	SHEEP	3	2	606	1
309	CHIEF	12	11	605	1
190	ASSURE	27	25	599	4
220	COMPETITION	22	20	597	4
577	CORN	3	2	596	1
412	RESOURCE	7	6	593	3
197	COMPETITIVE	26	24	592	4
396	CARGO	7	6	587	4
265	PLATFORM	16	15	586	2
253	EXPANSION	17	16	581	4
658	NO	3	2	573	1
366	THIRD	9	8	568	1
359	REGIONAL	9	8	564	3
228	SUBSCRIPTION	21	20	560	6
166	EXPAND	33	31	554	4
269	ACCESS	15	14	553	4
387	REGION	8	7	548	2
312	EXTERNAL	12	11	543	5
422	AIRPORT	6	5	542	1
502	CONCESSION	4	3	541	6
624	HUMAN	3	2	539	1

N	KW	Texts	%	Overall Freq.	Level in SHEWR
644	LOWER	3	2	539	2
267	QUALITY	16	15	532	2
159	RELY	36	34	529	3
264	OPERATE	16	15	522	2
294	PROTECTION	14	13	518	3
208	SUBSEQUENT	25	23	508	6
247	REASONABLE	18	17	501	3
258	VARIABLE	17	16	501	6
705	SQUARE	3	2	495	2
222	ELIGIBLE	22	20	492	6
195	ACQUIRE	26	24	490	4
293	PATENT	14	13	490	5
287	TRANSMISSION	15	14	488	6
284	STRATEGY	15	14	482	3
212	CERTIFICATE	23	21	479	5
230	EXCESS	20	19	478	5
593	DRUGSTORE	3	2	475	2
513	HEALTH	4	3	474	1
244	RETIREMENT	19	18	471	4
259	APPLICATION	16	15	469	4
225	CIRCULAR	21	20	467	4

N	KW	Texts	%	Overall Freq.	Level in SHEWR
172	RETAIN	32	30	465	4
326	CHAIN	6	5	464	2
430	SAFETY	11	10	464	3
135	EXPIRE	42	40	461	6
207	AVIATION	3	2	460	6
555	PROVINCIAL	25	23	460	6
340	PHONE	10	9	449	2&5
506	DISTANCE	4	3	446	2
215	ENSURE	23	21	443	5
464	UPON	6	5	443	2
713	TELEPHONE	3	2	442	2
684	REST	3	2	440	1
257	STRATEGIC	17	16	439	6
708	STEEL	3	2	436	2
186	FIN	29	27	433	5
327	TARIFF	11	10	433	6
279	MONTHLY	15	14	432	4
323	JOINT	11	10	430	2
438	GENERATION	6	5	423	4
452	PREPARATION	6	5	416	3
266	QUALIFICATION	16	15	410	6

N	KW	Texts	%	Overall Freq.	Level in SHEWR
656	NATIONAL	3	2	403	2
478	LEARNING	5	4	401	4
297	ACCIDENT	4	3	398	1
490	RULE	14	13	398	3
382	FURTHER	8	7	395	2
391	VOLUME	8	7	394	3
199	CODE	15	14	382	6
271	GENERATE	26	24	382	4&5
460	SYNTHETIC	6	5	382	6
435	EAST	6	5	378	1
214	DISTRIBUTE	23	21	377	4
417	treatment	7	6	377	2
390	TREASURY	8	7	373	5
254	FACILITY	17	16	372	4
431	CLINICAL	6	5	369	6
251	ARRANGEMENT	17	16	367	2
569	CITY	3	2	358	1
373	AUTHORITY	8	7	354	4
416	TELEVISION	7	6	353	2&4
563	CARD	3	2	352	1
669	PIPE	3	2	351	2

N	KW	Texts	%	Overall Freq.	Level in SHEWR
263	DRUG	9	8	350	3
348	MANAGE	16	15	350	2
314	BUREAU	11	10	348	5
719	TYPE	3	2	348	2
424	BASE	6	5	347	1
518	PHASE	4	3	347	6
347	CONSUMPTION	9	8	346	6
236	BONUS	19	18	345	5
440	IMPORT	6	5	339	3
571	CLASS	3	2	338	1
209	ACCOUNTANT	24	22	336	4
617	GEOGRAPHICAL	3	2	336	5
413	CASUALTY	6	5	330	3
428	STAFF	7	6	330	6
393	YIELD	8	7	328	5
345	AIRLINE	9	8	326	2
338	INVENTORY	10	9	325	6
307	ASSESSMENT	12	11	322	6
304	POTENTIAL	13	12	321	5
493	APPROPRIATE	4	3	320	4
374	CLIENT	8	7	319	3

N	KW	Texts	%	Overall Freq.	Level in SHEWR
305	FAIR	7	6	318	6
389	PRINT	3	2	318	4
402	SITE	8	7	318	2
673	SURPLUS	13	12	318	1
248	RENTAL	18	17	316	6
262	HEDGE	16	15	312	5
514	GARDEN	3	2	309	5
614	ISSUE	4	3	309	1
296	REDUCE	14	13	307	3
405	NEWS	7	6	306	1
709	STORE	3	2	306	1
578	COURSE	3	2	302	1
458	SECRETARY	6	5	300	2
471	DEPARTMENT	5	4	299	2
495	AVAILABLE	4	3	298	3
664	PACIFIC	3	2	295	5
421	WEB	7	6	293	3
317	DEVELOP	11	10	291	2
484	SUPPORT	5	4	291	2
288	BEHALF	14	13	289	5
406	OPERATIONAL	7	6	288	6

N	KW	Texts	%	Overall Freq.	Level in SHEWR
415	TECHNICAL	7	6	288	3
566	CELL	3	2	283	2
289	FAIL	14	13	282	2
693	SEA	3	2	279	1
407	ELECTRIC	5	4	277	2
472	PACKAGE	7	6	277	3
479	METHOD	5	4	275	2
351	EXCEED	9	8	273	5
543	ADVANCED	3	2	270	3
354	INSTITUTE	9	8	267	5
544	AGRICULTURAL	3	2	263	5
608	FLIGHT	3	2	261	2
691	RUN	3	2	261	1
324	MANUFACTURE	11	10	258	4
216	EXPIRATION	23	21	256	6
363	MAIL	3	2	256	3
385	PRODUCE	8	7	256	2
646	SECURITY	9	8	256	1
249	TERMINATE	18	17	255	6
371	AGENT	8	7	254	4
730	WATER	3	2	251	1

N	KW	Texts	%	Overall Freq.	Level in SHEWR
367	TON	9	8	248	3
321	INVEST	11	10	247	4
561	CALL	3	2	242	1
596	EDUCATIONAL	3	2	242	3
398	DEVICE	7	6	239	4
671	PRESIDENT	3	2	239	2
350	EQUIVALENT	9	8	238	6
466	VIDEO	6	5	238	2&4
397	CUMULATIVE	7	6	237	6
717	TRAIN	3	2	237	1
352	GOODS	9	8	235	4
447	EFFICIENCY	3	2	229	3
597	PERMIT	6	5	229	4
372	APPLY	8	7	227	2
677	RAILROAD	3	2	226	1
343	VIRGIN	10	9	225	4
632	INTERMEDIATE	3	2	224	4
470	CHARGE	5	4	223	2
570	CIVIL	3	2	223	3
702	SOUTH	3	2	220	1
716	TRADITIONAL	3	2	219	2

N	KW	Texts	%	Overall Freq.	Level in SHEWR
582	DAILY	3	2	218	2
636	LABEL	3	2	218	3
494	ARTICLE	4	3	217	2&4
291	CANCER	6	5	215	4
386	INTEND	14	13	215	4
427	REFORM	8	7	215	2
527	SECTOR	4	3	213	6
613	GAME	3	2	212	1
229	ENHANCE	20	19	208	6
442	MEDICINE	6	5	208	2
355	INTERPRETATION	9	8	206	5
521	REFLECT	4	3	205	4
497	BLUE	4	3	204	1
325	MINORITY	11	10	202	3
295	RECEIPT	14	13	200	3
360	RESERVATION	9	8	200	4
423	APPRECIATION	6	5	199	4
707	STAR	3	2	199	1
731	WEEKLY	3	2	199	4
503	DEGREE	4	3	198	2
403	MAJORITY	7	6	194	3

N	KW	Texts	%	Overall Freq.	Level in SHEWR
481	SATELLITE	5	4	185	4
318	DOUBTFUL	11	10	184	3
459	SECONDARY	3	2	184	6
694	SPECTRUM	6	5	184	3
392	WHOLESALE	8	7	182	5
337	FURTHERMORE	10	9	181	4
686	RIDGE	3	2	181	5
704	SPLENDID	3	2	181	4
633	JOB	3	2	179	1
599	ELECTRONICS	3	2	178	4
333	COVERAGE	10	9	177	6
477	IRON	5	4	177	1
377	ENDING	8	7	175	2
529	SAVINGS	3	2	175	1
692	SENSITIVITY	3	2	175	3
695	STATION	4	3	175	5
329	ADVERTISEMENT	10	9	170	3
676	RAIL	3	2	169	5
726	VEGETABLE	3	2	168	1
419	VIOLATION	7	6	167	4
454	PROMOTION	6	5	164	4

N	KW	Texts	%	Overall Freq.	Level in SHEWR
270	BACHELOR	15	14	163	5
476	DIVISION	4	3	163	5
507	HOUSING	5	4	163	2
432	COMMERCE	6	5	162	4
436	FIBER	6	5	159	5
729	VOICE	3	2	159	1
474	GASOLINE	5	4	158	3
456	RECOGNIZE	6	5	156	3
364	SUPERVISION	9	8	155	6
401	EVALUATION	7	6	153	4
610	FORTUNE	3	2	153	3
689	ROUTE	3	2	153	4
465	USEFUL	6	5	152	1
679	RECRUIT	3	2	150	6
725	VARIOUS	3	2	150	3
661	OUTDOOR	3	2	149	3
334	ARTIST	3	2	148	3
483	ENGAGE	10	9	148	3
550	SOLUTION	3	2	148	2
701	STRUCTURE	5	4	148	2
487	UNDERGROUND	5	4	147	2

N	KW	Texts	%	Overall Freq.	Level in SHEWR
332	CAUSE	6	5	146	5
429	CONSENT	10	9	146	1
592	DRILL	3	2	146	4
331	COMPETE	10	9	143	3
404	GREEN	3	2	143	2
620	MIX	7	6	143	1
411	REPRESENT	7	6	142	3
420	waste	7	6	142	1
541	ACTUAL	3	2	139	3
616	GENIUS	3	2	139	4
654	MINIMUM	3	2	139	4
330	ATTRACT	10	9	138	3
380	ENFORCEMENT	8	7	136	4
475	HARDWARE	5	4	136	4
678	RECORD	3	2	136	2
335	EXPERTISE	10	9	135	6
342	STAMP	10	9	135	2
462	ACCEPTANCE	5	4	134	1
467	TICKET	6	5	134	4
722	UNIVERSAL	3	2	134	4
409	REGISTER	7	6	133	4

N	KW	Texts	%	Overall Freq.	Level in SHEWR
574	COMPLEX	3	2	132	3
375	COUNSEL	8	7	131	5
647	MAJOR	3	2	131	3
710	SUCCESS	3	2	130	2
520	RECOVERY	4	3	129	4
524	RESPONSIBLE	4	3	128	2
395	CALCULATION	7	6	127	4
573	COMMUNIST	3	2	127	5
559	BIRD	3	2	126	1
492	AI	4	3	125	5
601	ENVIRONMENT	3	2	125	2
665	PACT	3	2	124	6
473	DECLARATION	3	2	123	2
583	FAILURE	5	4	123	5
621	HANG	3	2	123	2
718	TRIAL	3	2	122	2
369	DESIGN	3	2	121	5
449	PORT	6	5	121	2
585	update	9	8	121	2
362	RESTRICT	9	8	119	3
426	CALENDAR	6	5	118	2

N	KW	Texts	%	Overall Freq.	Level in SHEWR
587	DISPATCH	3	2	118	6
595	ECONOMY	3	2	118	4
629	INSPECTION	3	2	118	4
688	ROUND	3	2	118	1
672	PREVIOUS	3	2	117	3
450	AUTOMOBILE	3	2	116	6
519	HISTORICAL	3	2	116	6
554	PREDECESSOR	6	5	116	3
622	PRESCRIPTION	4	3	116	3
540	ACID	3	2	115	4
580	CULTURE	3	2	114	2
557	BAY	3	2	111	3
562	CARBON	3	2	111	5
728	VILLAGE	3	2	109	2
698	SIMILAR	3	2	108	2
379	DISPLAY	3	2	106	4
528	ENFORCE	8	7	106	1
588	SKY	4	3	106	2&6
645	MACHINERY	3	2	105	4
400	ECONOMICS	7	6	103	4
434	DISCOUNT	6	5	102	3

N	KW	Texts	%	Overall Freq.	Level in SHEWR
648	MAR	3	2	102	1&6
663	OUTPUT	3	2	102	5
732	WIRE	3	2	102	2
508	ELECTION	4	3	99	3
537	POSTURE	3	2	99	2
670	VOTE	4	3	99	6
551	ATM	3	2	98	4
353	IMPLEMENT	9	8	97	6
657	NEWSPAPER	3	2	97	1
711	SUCCESSFUL	3	2	97	2
500	CHANNEL	4	3	94	3
517	MINERAL	4	3	94	4
630	INTEGRATION	3	2	93	6
723	URBAN	3	2	93	4
448	CASUAL	3	2	90	4
564	FREQUENCY	3	2	90	3
612	POLLUTION	6	5	90	4
468	AUCTION	3	2	89	4
509	CATALOGUE	5	4	89	3
552	CD	3	2	89	6
565	ELECTRICAL	3	2	89	4

N	KW	Texts	%	Overall Freq.	Level in SHEWR
598	ENGINEER	4	3	89	3
489	TOPIC	4	3	86	6
534	VALID	5	4	86	2
594	DURATION	3	2	85	5
685	RESTORATION	3	2	85	6
498	BROADCAST	3	2	84	2
523	CABLE	4	3	84	2
560	MEMBERSHIP	3	2	84	2
651	REQUIREMENT	4	3	84	3
510	ENTERTAINMENT	4	3	83	4
618	GLORY	3	2	81	3
681	REMAINDER	3	2	81	6
526	GRADUATE	3	2	80	3
619	SCHEDULE	4	3	80	3
542	ADDRESS	3	2	79	1
727	VENTURE	3	2	79	5
637	LAUNCH	3	2	78	4
649	MATCH	3	2	77	1&2
683	RESOLUTION	3	2	77	4
501	CHIP	4	3	76	3
516	LINK	3	2	76	4

N	KW	Texts	%	Overall Freq.	Level in SHEWR
525	LOYALTY	4	3	76	4
640	REWARD	4	3	76	2
602	ESTABLISHMENT	3	2	75	4
733	WISDOM	3	2	75	3
703	SPARE	3	2	74	4
433	CONSULT	6	5	72	4
581	CUSTOMS	3	2	71	5
607	FLEET	3	2	71	6
496	BARREL	4	3	70	3
504	DISCHARGE	4	3	70	6
535	FORWARD	3	2	70	3
611	TIGER	3	2	70	2
714	TRANSPORT	4	3	70	1
410	LIVESTOCK	3	2	69	3
451	PREDICT	6	5	69	4
641	RELATE	7	6	69	5
675	QUALIFICATIONS	3	2	68	6
712	SUPPLEMENT	3	2	68	6
720	UNDERTAKE	3	2	68	6
568	CIRCUIT	3	2	67	5
667	PEARL	3	2	67	3

N	KW	Texts	%	Overall Freq.	Level in SHEWR
457	RENEW	6	5	66	3
721	UNIQUE	3	2	66	4
548	ANALYTICAL	3	2	65	6
511	ETHICS	4	3	64	5
576	CONVENIENCE	3	2	63	4
600	EMPLOYER	3	2	63	3
682	REPAIR	3	2	63	3
579	CROP	3	2	62	2
586	DIAGNOSIS	3	2	62	6
631	INTELLIGENCE	3	2	62	4
414	SUBSCRIBE	7	6	61	6
512	EVALUATE	4	3	60	4
515	LIMIT	4	3	60	2
653	MIGRATION	3	2	60	6
536	ESTIMATE	3	2	59	5
603	OUTGOING	3	2	59	4
662	SOLE	3	2	59	5
700	TREATY	4	3	59	5
660	ORIGIN	3	2	58	3
553	AUDIO	3	2	55	4
690	RUBBER	3	2	55	1

N	KW	Texts	%	Overall Freq.	Level in SHEWR
627	INDICATION	3	2	54	4
638	LAYER	3	2	54	5
680	REFER	3	2	53	4
650	MECHANISM	3	2	52	6
488	COMMENCE	3	2	51	6
572	UPGRADE	5	4	51	6
674	PROMOTE	3	2	50	3
505	DELIVER	3	2	48	5
584	DISPOSE	4	3	48	2
625	IDENTICAL	3	2	48	4
666	PARTICIPANT	3	2	48	5
522	REPUTATION	4	3	47	4
531	TECHNOLOGICAL	4	3	46	4
639	LEISURE	3	2	45	3
697	SILICON	3	2	45	6
699	SLOT	3	2	44	6
609	FORK	3	2	42	1
538	WHARF	4	3	40	5
486	COMPRISE	3	2	39	3
575	MICROSCOPE	3	2	39	6
652	TOURISM	5	4	39	4

N	KW	Texts	%	Overall Freq.	Level in SHEWR
549	APPOINT	3	2	38	4
706	STAPLE	3	2	38	6
558	BEVERAGE	3	2	37	6
556	BACKBONE	3	2	36	5
615	GENETIC	3	2	36	6
628	INN	3	2	36	3
635	KIT	3	2	34	3
724	UTILIZE	3	2	34	6
591	DOSAGE	3	2	33	6
604	EXTRACT	3	2	33	6
642	LOCOMOTIVE	3	2	33	5
491	ADVISOR	4	3	32	3
605	FERTILIZER	3	2	30	5
626	IMMUNE	3	2	27	4&6
643	LOGO	3	2	27	5
634	KIN	3	2	26	5
530	DIVERT	3	2	25	5
590	SULFUR	4	3	25	6
589	DIVERSIFY	3	2	24	6
567	CHAIRWOMAN	3	2	18	5