Verbs in the Written English of Chinese Learners:

A Corpus-based Comparison

between Non-native Speakers and Native Speakers

by

Xiaotian Guo

A thesis submitted to the University of Birmingham

for the degree of DOCTOR of PHILOSOPHY

Supervisor: Professor Susan Hunston

The Department of English

The University of Birmingham

The School of Humanities

October 2006

Abstract

This thesis consists of ten chapters and its research methodology is a combination of quantitative and qualitative. Chapter One introduces the theme of the thesis, a demonstration of a corpus-based comparative approach in detecting the needs of the learners by looking for the similarities and disparities between the learner English (the COLEC corpus) and the NS English (the LOCNESS corpus). Chapter Two reviews the literature in relevant learner language studies and indicates the tasks of the research. The data and technology are introduced in Chapter Three. Chapter Four shows how two verb lemma lists can be made by using the Wordsmith Tools supported by other corpus and IT tools. How to make sense of the verb lemma lists is the focus of the second part of this chapter. Chapter Five deals with the individual forms of verbs and the findings suggest that there is less homogeneity in the learner English than the NS English. Chapter Six extends the research to verb–noun relationships in the learner English and the NS English and the result shows that the learners prioritise verbs over nouns. Chapter Seven studies the learners' preferences in using the patterns of *KEEP* compared with those of the NSs, and finds that the learners have various problems in using this simple verb. In this chapter, too, my reservations about the traditional use of 'overuse' and 'underuse' are expressed and a finer classification system is suggested. Chapter Eight compares another frequently-occurring verb, *TAKE,* in the aspect of collocates and yields similar findings that the learners have problems even with such simple vocabulary. In Chapter Nine, the research findings from Chapter Four to Chapter Eight are revisited and discussed in relation to the theme of the thesis. The concluding chapter, Chapter Ten, summarises the previous chapters and envisages how learner language studies will develop in the coming few years.

Acknowledgements

First and foremost, I would like to thank my supervisor Professor Susan Hunston. She spent a large amount of time on my thesis and guided me from the design of the research to the last version of each chapter. As an experienced supervisor and teacher, she knows very well when to leave me free exploring for something useful and when to bring my attention back to things with value. She hardly tells me what to do, but offers suggestions, comments, and clues for further development, leaving me enough time to reflect and digest. Undoubtedly, the knowledge I obtained from her supervision will be the most valuable assets for my academic career.

Secondly, my thanks should go to my beloved wife, Xiaorong (Wang). Actually, she sacrificed so much for my PhD study that I can hardly find appropriate words to express my gratitude. Different from many students who were funded by one means or another, my PhD was self-sponsored. Therefore, my finance became the dominating difficulty of my PhD study. In order to overcome this obstacle, she worked extremely hard and underwent great hardship and suffering. Even though she deserves a long break after the submission of my thesis, the unfortunate damage caused to her health may take the rest of her life to mend. In this sense, any words of thanks are incredibly weak and inadequate.

Thirdly, my sincere thanks go to my colleagues and friends who have supported me in many different aspects. Without their help my thesis could not have been accomplished by now. The names to follow are only some of them (with all the given names first and surnames last to be consistent): Richard (Zhonghua) Xiao, Scott (Songlin) Piao, Wenzhong Li, Pernilla Danielsson, Seo-In Shin, and Frank (Maocheng) Liang for their help in IT and corpus technologies; Geoff Barnbrook, Antoinette Renouf, Wenzhong Li and Jinbang Du for their valuable comments and suggestions; Sylviane Granger, John Milton, Angela Hasselgren, Shichun Gui, Jianzhong Pu, and Michael Rundell for their articles, PhD theses or other information sent to me when I was in desperate need of them; Wenjin Zhao, Zequan Liu, Laiqi Zhang, Junhua Zhang and Yaodong Wang for their encouragement and support as friends. There are others who helped me in one way or anther, but I am afraid I cannot list them all here.

Fourthly, I am grateful to my external examiner Mike Scott and internal examiner Martin Hewings for their valuable comments and advice and the chair to my viva Murray Knowles for his valuable time.

In addition, I am deeply indebted to my sister who looked after my parents together with my brother while I could not fulfil my part of duty as a son. I also thank my wife's family, Shulin and his family for their encouragement and support. My special thanks go to my daughter who accompanied me through the ups and downs of the years, especially when my wife had to work in another place. She also helped me with the proofreading of the Chinese pin-yin (the remaining errors still belong to me, of course).

Furthermore, thanks are overdue to the Great Britain-China Education Trust and Sino-British Fellowship Trust for the £1000 fellowship which was sent to me on the very day of the Chinese Spring Festival of 2003. It was the only funding I gained throughout my PhD study. Even though such an amount was far from liberating me from the financial strains, the very act of providing such a grant justified my study and greatly encouraged me to go through the rest of the difficulties. It meant a lot to me.

Last but not least, I must thank the University of Birmingham, especially the staff members of the Department of English, the School of Humanities, the Information Service, the Academic Office and the International Office for their unfailing and patient support.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

BoE	The Bank of English
BNC	The British National Corpus
CA	Contrastive Analysis
CCED	Collins Cobuild English Dictionary
CIA	Contrastive Interlanguage Analysis
CLC	Computer Learner Corpus
CLEC	The Chinese Learner English Corpus
COLEC	The Chinese College Learner English Corpus
DDL	Data-Driven Learning
EA	Error Analysis
EL	English language
ELT	English language teaching
GSL	A General Service List of English Words
ICLE	The International Corpus of Learner English
IL	interlanguage
KWIC	key word in context
L1	first language
L2	second language
LEA	The Longman Essential Activator
LLC	The Longman Learners' Corpus
LOCNESS	Louvain Corpus of Native English Essays
NL	native language
NNS	non-native speaker
NS	native speaker
POS	part of speech
SL	second language
SLA	Second Language Acquisition
TL	target language

Chapter One

Introduction

1.1 The theme and aim of the research

This thesis reports on a study of verb-related features of Chinese learner English. The aim of the research is to demonstrate how a corpus linguistic approach to learner English studies can help us to find out the similarities and disparities between the written English of a group of non-native speakers (NNSs) and that of a group of native speakers (NSs). It is hoped that the identification of similarity and difference between the learner English and the NS English will help us to identify the needs of the learners in essay writing.

1.2 Introducing computer learner corpus research

In the late 1980s and early 1990s, learner language research saw the birth of computer learner corpora (CLC), which are defined as follows by Granger (2002: 7):

> Computer learner corpora are electronic collections of authentic EL/SL textual data assembled according to explicit design criteria for a particular SLA/ELT purpose. They are encoded in a standardised and homogeneous way and documented as to their origin and provenance.

On the use of computer learner corpora, she comments thus (Granger 2002: 4):

> Using the main principles, tools and methods from corpus linguistics, it aims to provide improved descriptions of learner language which can be used for a wide range of purposes in foreign/second language acquisition research and also to improve foreign language teaching.

The core of learner corpus research lies in "contrastive interlanguage analysis" (CIA) as she maintains (Granger 1998b; 2002) though it is possible to carry out non-contrastive analysis (for example, Li 2003).

Unlike the previous learner language studies such as contrastive analysis (CA) and error analysis (EA) which will be reported in Section 1.3 of this chapter, this new approach to learner language study treats learner language as an entity in its own right. As Leech (1998:

xvii) insightfully summarises:

> "It enables us to investigate the non-native speaking learners' language (in relation to the native speakers') not only from a negative point of view (what did the learner get wrong?) but from a positive one (what did the learner get right?). For the first time it also allows a systematic and detailed study of the learners' linguistic behaviour from the point of view of 'overuse' (what linguistic features does the learner use more than a native speaker?) and 'underuse' (what features does the learner use less than a native speaker?)".

Apart from this, the new approach allows us to see the similarity and disparity between learner English and NS English when the learner English data and the NS English data are compared. On the whole, similarity points to, though it does not necessarily lead to, a degree of mastery by the learners, while disparity points to, but does not necessarily lead to, a kind of non-mastery by them. The features which are used by the NSs, but not by the learners, would be necessary for the learners to acquire if they wish to achieve the naturalness and 'nativeness' of the NS English (if the influence of the difference in topics between the two corpora is ignored for the moment).

1.3 The background to this research

A detailed review of the earlier studies concerning learner language will be found in Chapter Two. This section briefly relates the current research to the background from which CLC has emerged.

Earlier research in learner language may be traced to EA. It was generally maintained before the EA era, for instance in CA, that the learner's errors are undesirable because they are a sign of non-acquisition. Since the CA researchers found a relationship between the learner's errors and the difference between the learner's mother tongue (L1) and their second language (L2), they tried to pinpoint the source of errors by contrasting the two languages. In a comment to language teachers on the use of CA, Corder (1967, reprinted in Richards 1974: 19) remarks:

> Teachers have not always been very impressed by [the contribution from CA researchers] for the reason that their practical experience has usually already shown them where these difficulties lie and they have not felt that the contribution of [the researchers] has provided them with any significantly new information.

It was a significant advance when EA researchers to have placed the learner language (rather

than L1 and L2) under examination. A central consensus among EA researchers was that the learner's errors, instead of being seen as negative, should be treated as positive. The learner's language was treated as "interlanguage" (Selinker 1972) or as an "approximative system" (Nemser 1971). This is invaluable indeed for a better understanding of how second language acquisition takes place. However, there are some serious limitations with EA, one of which is that errors have been studied in isolation (see 2.1.1 for more details). Apart from this, the correct use of learner language was not as fully attended to as it deserves. EA prevailed in the 1960s and 1970s but was gradually submerged in a more general study in the field of L2 acquisition which is known as second language acquisition (SLA) today.

The major concern of SLA has been the nature of language acquisition *process* and the *factors* which affect language learners (Larsen-Freeman 1991). When the learner's output is considered, the focus of the research is rather more on the output of individual learners than on the output of a group of learners with the same background. Actually, the collective aspect of learner English should be a facet of SLA research and should not be neglected, according to Leech (1998: xix).

1.4 The impetus of this research

As mentioned above, even though there have been some advances in our understanding of how L2 acquisition takes place, obviously some important problems remain unsolved. EA was over-dependent on the error aspect of learner language, and therefore it is impossible for EA researchers to draw up a more complete profile of learner language as it is. As far as SLA is concerned, it is hard to find answers to questions concerning the nature of the language produced by a group of learners since its research focus is on the individual mind rather than on the output of the group. I would argue that in a world where English is mostly taught and learned in classes and groups, it is the information on group learner English that requires most of the attention of language researchers and teachers. If we wish to probe into the needs of learners, it is imperative that we examine the English produced by a group of learners rather than by individuals. If we suppose teachers wish to tailor their teaching to the needs of their students and help them to achieve a target level which is similar to the norm they have selected, there are some questions that must be solved first before any remedial work is carried out. What does it mean for learners to extend their vocabulary? What is the overall

size of the learners' vocabulary? Learners very often express their intention to expand their vocabulary and teachers strive hard to help their students to attain this end, but before students try to expand their vocabulary, the question arises: have they reached the full degree of vocabulary use for each word they think they know, especially the commonly used simple words? Among the different senses of polysemous and multiple part-of-speech (POS) words, to what level of complexity can the students operate? In a new approach to learner language studies, all these questions are likely to have an answer.

1.5 The focus and research questions of the research

In looking at the behaviour of the learner English this research focuses on the aspect of verbs. For one thing, it is not possible to concentrate on every POS. However, one important reason for having selected verbs rather than other parts of speech is that "nouns are more topic-related than other parts of speech" (Leech 2001: 332) and "Verbs are less topic-sensitive than nouns, and the most frequently used verbs may thus provide a good starting point for an assessment of linguistic features characteristic of one group of learners" (Ringbom 1998a: 192). Another reason is that "The choice of the verb system as the focus of study in second language acquisition (SLA) is based on the assumption that this is a centrally important area for the structure of any language which is moreover likely to pose major learning problems of any age (Harley 1986; Palmer 1975)", according to Housen (2002: 78). Given that the focus of the thesis is on verbs, the following are the overall research questions:

1) What are the salient similarities and disparities between the learner English and the NS English in the aspect of the width and depth of verbs? (By the width of verbs, I mean the size of vocabulary in verbs. By the depth of verbs, I mean the range of senses of verbs and the many words which, while being other POS, have a verbal function.)
2) What kinds of techniques could be used to answer the previous research question?
3) What are the pedagogical implications of this research?

1.6 The methodology of the research

This research uses a corpus-based approach to study group learner written English, i.e. the COLEC learner English. To highlight the features of the learner English, a reference corpus LOCNESS is used for comparison (for details of the two corpora including their contents,

sizes, and comparability, see Chapter Three). The standard text retrieval software used is mostly the WordSmith Tools (3.0) (Scott 1999) plus some use of a newer version of the WordSmith Tools (4.0) (Scott 2004) where necessary. In cases where the reference corpus is found insufficient for some enquiries, a larger and general NS corpus, the Bank of English (BoE) is used. In addition, the Google search engine (henceforward Google) is occasionally used to back up some intuitions about a particular usage.

In the cline of quantitative research and qualitative research in CLC, critical remarks by Nesselhauf (2004: 136) are worth noting:

> Many studies are exclusively or primarily quantitative. … While such studies can be interesting starting points for further quantitative analyses, they do not usually in themselves contribute much to language learner analysis, let alone to language teaching. If progress is to be made, it is imperative that this current stage is left behind and that more qualitative analyses are carried out.

Bearing this in mind, my research employs a method which is a combination of both the quantitative and the qualitative approaches. It is my belief that only by taking both approaches can we take full advantage of the current computer technology as well as the insightful practice and theories in corpus linguistics and other relevant areas such as English language teaching (ELT) (see 9.2.1 for more discussion of the quantitative *versus* the qualitative approach in corpus linguistics).

1.7 Two assumptions behind this research

In this thesis it is assumed, as is usual in this newly-born field of learner language study, that the NS English in the reference corpus can be regarded as a norm for the learners and the state of NS English is regarded as the ideal or target state for the learners to arrive at. Another assumption I need to make is that learners of English from the same background (L1, culture, age, education system, etc.) share similarities in their production of L2. This is also implied in the practice of learner corpora researchers. In other words, what appears to be frequent in the group is considered to be a commonly held characteristic of the majority of the group. To look at the question of similarity among learners with a similar background, refer to Raupach (1984) (cited in Hasselgren 2002: 154-55).

1.8 The structure of the thesis

As reviewed by Lenko-Szymanska (2002: 218), the majority of CIA studies focus either on the breadth or the depth of learners' vocabulary knowledge, whereas actually both of the aspects "constitute equally important and vital components of the overall lexical ability". Bearing this in mind, this thesis explores both the breadth and the depth of the learners' lexicon in the aspect of verbs. In Chapters Four and Five, the research focuses on the breadth of the learners' lexicon in verbs. Chapters Seven and Eight then switch to analysis-in-depth of the use of two frequently occurring verbs. The contents of each chapter are described below.

Chapter One mainly introduces the theme and the aim of this research, the background to it and the impetus behind it. This chapter also introduces the birth of the learner corpora studies to which this research methodologically belongs. It then sets out the agenda for the whole dissertation. Chapter Two reviews the literature of corpus linguistics focusing on its application in language pedagogy and education. Chapter Three introduces the data to be used in the research and the methodologies adopted in the investigation. From Chapter Four to Chapter Eight, I will report on my research which aims at a presentation of the advantages of a corpus-based method in the exploration of learner English. To be specific, Chapter Four first illustrates the creation of two verb lemma lists (one from the learner corpus and the other from the NS corpus) based upon annotated COLEC and LOCNESS and other modern technologies and then continues to explore how to make sense of the verb lemma lists by categorising individual verb lists semantically into groups. Chapter Five looks at the disparity in verb form distribution between the two corpora. Chapter Six deals with the disparity between the two corpora in terms of the distribution of verbal function and nominal function in some multiple POS vocabulary. In Chapter Seven I will choose a commonly used verb, *TAKE,* to look at all its collocates in the two corpora and see how well the learners' performance approximates the NSs' performance. In Chapter Eight, I will choose another commonly used verb, *KEEP,* to investigate how the learners' performance approximates that of the NS in terms of patterns (in line with Hunston and Francis 1999). Chapter Nine summarises the findings of the research chapters and discusses the advances this research has made in learner corpora studies. The pedagogical implications of this research will be addressed in this chapter and some possible studies in the area of learner corpora study will also be identified. Chapter Ten summarises the research and points out the limitations of the

research. It also envisages the near future of learner language studies.

Chapter Two

A Literature Review of Learner Language Studies

Computer learner corpus research is a very young branch of study of learner language (Granger 1998a, Leech 1998 & 2001, Nesselhauf 2004 and many others). "With roots both in corpus linguistics and second language acquisition (SLA) studies, it uses the methods and tools of corpus linguistics to gain better insights into authentic leaner language", as Granger summarises (1998a: xxi). Since EA is considered to be an earlier period of SLA (Ellis 1994: 68), this chapter starts from a review of EA and then revisits the territory of SLA. This review questions the relationship between synchronic CLC and SLA. After a brief recall of the birth of CLC, a few prominent learner corpora and the major learner corpus typology will be introduced. Some important issues relating to CLC will be discussed in some detail. Some striking features of learner English as found by many researchers so far will be presented and illustrated in detail. In the end, some inadequacies of and reservations about the current CLC studies will be addressed in relation to the topics of this thesis.

2.1 Earlier learner language studies

Since CLC originates to some extent from EA, a much earlier approach to learner language studies which also aims to focus on the product rather than the process of learner language, this section recalls the practice and decline of EA. The relationship between CLC and SLA will be revisited because it is my view that the widely-held view that SLA is the root of CLC (Leech 1998; Granger 1998a; Granger 2002) might be amended as CLC studies continue.

2.1.1 Error analysis recalled

Before EA, errors were treated as negative signs of acquisition or in the words of George (1972) "unwanted forms" (cited in Ellis 1994: 47). Errors 'should' not occur if native-likeness is targeted. This faulty view was challenged by many EA scholars including Corder (1967, reprinted in Richards 1974: 25) who brought to light the significance of learners' errors:

> A learner's errors, then, provide evidence of the system of the language that he is using (i.e. has learned) at a particular point in the course (and it must be repeated that he is using some system, although it is not yet the right system). They are significant in three ways. First to the teacher, in that they tell him, if he undertakes a systematic analysis, how far towards the goal the learner has progressed and consequently, what remains for him to learn. Second, they provide to the researcher evidence of how language is learned or acquired, what strategies or procedures the learner is employing in his discovery of the language. Thirdly (and in a sense this is their most important aspect) they are indispensable to the learner himself, because we can regard the making of errors as a device the learner uses in order to learn. It is a way the learner has of testing his hypothesis about the nature of the language he is learning.

In explaining the process of how EA scholars conduct error analysis, Ellis (1994: 68-69) has summarised it in four stages, i.e. the collection of errors, the identification of errors, the description of errors and the explanation of errors. The following is his illustration of the four stages:

> The first step in carrying out an EA was to collect a massive, specific, or incidental sample of learner language. The sample could consist of natural language use or be elicited either clinically or experimentally. It could also be collected cross-sectionally or longitudinally. The second stage involved identifying the errors in the sample. Corder distinguished errors of competence from mistakes in performance and argued that EA should investigate only errors. …The third stage consisted of description. Two types of descriptive taxonomies have been used: linguistic and surface strategy. The former provides an indication of the number and proportion of errors in either different levels of language (i.e. lexis, morphology, and syntax) or in specific grammatical categories (for example, articles, prepositions, or word order). The latter classifies errors according to whether they involve omission, additions, misinformations, or misordering. The fourth stage involves an attempt to explain the errors psycholinguistically.

EA prevailed in the 1960s and 1970s. In an article by Schachter and Celce-Murcia (1977: 442), a vivid depiction of the prevalence of EA is presented thus:

> A cursory glance at the titles and abstracts in recent issues of journals such as this one [*TESOL Quarterly*] (and others such as Language Learning and IRAL) would indicate that the advocates of EA have prevailed and that EA currently appears to be the "darling" of the 70's.

However, EA was not without problems. It was virtually in the heyday of EA when Schachter and Celce-Murcia (1977: 441) courageously and insightfully voiced their reservations concerning EA. There are six areas in error analysis which exhibit potential weakness: "(1) the analysis of errors in isolation; (2) the classification of identified errors; (3) statements of error frequency; (4) the identification of points of difficulty; (5) the ascription of causes to systematic errors; (6) the biased nature of sampling procedures. These altogether limit the usefulness of error analysis in describing the acquisition process of the second language

learner." Among the six areas, at least three deserve some more elaboration here, i.e. (1), (2) and (5). According to Schachter and Celce-Murcia, the first weakness comes from the limited perspective on understanding learner English, i.e. *the analysis of errors in isolation*. EA researchers took the trouble to extract learner errors from the data available. However, after the errors were analysed the data would be discarded from consideration. Schachter and Celce-Murcia (1977: 445) used examples to illustrate their point that it is inadequate and therefore harmful to investigate errors as if they could exist in isolation. The second weakness of EA lies in the difficulty of a proper classification of identified errors. As Schachter and Celce-Murcia noted, it is not always easy to decide whether an error is a deviation from the target language. Even though it is possible to make such a decision, it would be more difficult to locate what structure this error is in. The authors also used examples to show that there is always more than one decision to make in judging what structure or category an error belongs to. This point (together with the following one) is important for this thesis in that it justifies my decision not to take the stance of concentrating on errors in my research. The fifth weakness arises from "the ascription of causes to systematic errors". There might be multiple causes for this ascription; for example, interlingual (those due to the disparity between languages) and intralingual (those due to overgeneralisation within a language). It is a common practice for EA investigators to do some analysis of some isolated errors within a limited scope and then label them with interlingual or intralingual causes. Schachter and Celce-Murcia (1997: 44) comment that "It would be wise, then, for investigators to suggest causes of error only very cautiously. What we see happening, however, is just the reverse". What is paramount in the weaknesses that Schachter and Celce-Murcia listed is the isolated treatment of errors by EA investigators and the difficult situation which arises with the classification and ascription of errors. It is evident that looking at errors only will not lead to a comprehensive idea of how a language is produced by learners. As stated by Ellis (1994: 67):

> A frequently mentioned limitation is that EA fails to provide a complete picture of learner language. We need to know what learners do correctly as well as what they do wrongly.

Due to the faulty perspective adopted in methodology, EA went out of fashion and was largely submerged by a more general area of learner language study: SLA.

2.1.2 Second language acquisition reviewed

"There is no simple answer to the question 'What is second language acquisition?' … Second language acquisition is a complex, multifaceted phenomenon and it is not surprising that it has come to mean different things to different people", according to Ellis (1994:15). After a few decades of development from the end of the 1960s, "SLA research has become a rather *amorphous* field of study with *elastic* boundaries" (Ellis 1994: 2, italics added). Among the few researchers who attempt to define the borders of SLA are Larsen-Freeman and Long (1991, cited in Ellis 1994: 3), who believe that the territory of SLA is primarily the nature of the language acquisition *process* and the *factors* which affect language learners. Even though analysis has been made from groups of learners in SLA, it still remains a peripheral interest of SLA and most of the attention has been given to the individual learner's acquisition process and the factors that influence the process of acquisition.

Apart from the fact that collective learner English is not a major concern of current SLA research, there are also some limitations that current SLA research suffers in terms of data collection. This was pointed out explicitly by Granger (2002: 5-6) as follows:

> SLA research has traditionally drawn on a variety of data types, among which Ellis (1994: 670) distinguishes three major categories: language use data, metalingual judgements and self-report data …. Much current SLA research favours experimental and introspective data and tends to be dismissive of natural language use data. There are several reasons for this, prime among which is the difficulty controlling the variables that affect learner output in a non-experimental context. As it is difficult to subject a large number of informants to experimentation, SLA research tends to be based on a relatively narrow empirical base, focusing on the language of a very limited number of subjects, which consequently raises questions about the generalizability of the results.

In agreement with Granger (1998b: 5), I also firmly believe that "There is clearly a need for more, and better quality, data and this is particularly acute in the case of natural language data" and "learner corpora are a valuable addition to current SLA data sources" .

2.1.3 Conclusion

On one hand, EA failed to provide a complete picture of learner English though it attempted to depict a picture of learners' errors for clear pedagogical purposes. On the other hand, a very important area, i.e. the collective aspect of learner English, has received relatively little

attention in the current SLA research. As my research will gradually show, this is an area where CLC can play a better part by investigating the features of group learner English, which has been "unduly neglected", according to Leech (1998: xix).

2.2 Computer learner corpora: a new era

As discussed above, though EA was used to analyse learners' errors, it was on a much smaller scale, in no way comparable to the present CLC. CLC did not come into being until the late 1980s when NS corpora technology and analysis became fairly mature. As Aston (2000: 11) points out, the study of and research into NS corpora contribute to the description of the native language alone and provide "no information as to the relative difficulty and learnability of particular features to be taught" and studies "based on the analysis of native-speakers behaviour fail to consider the productivity of particular features from the learner's perspective". In the words of Granger (1998b: 7), "native corpora cannot ensure fully effective EFL learning and teaching, mainly because they contain no indication of the degree of difficulty of words and structures for learners" and for her it is doubtful that ELT materials should be designed "with a very fuzzy, intuitive, non-corpus-based view of the needs of an archetypal learner" (*ibid*.). As a result, NS corpora will not be able to shed any light upon how a language is acquired by NNSs. In emphasising the role of CLC, Leech (2001: 339) states that "corpus-based interlanguage analysis enables us to identify areas of difficulty which are not derivable from NS corpora alone, and which can often be attributed to particular causes, especially L1 transfer." Biber and Reppen (1998: 157) also maintain that "it is only by investigating actual language use in natural discourse that we can begin to understand how best to help students develop competence in the kinds of language they will encounter on a regular basis." More recently, Nesselhauf (2004: 125-126) also adopts the same tone, as follows:

> Hardly anyone will doubt any longer that native speaker corpora are indeed useful for the improvement of language teaching. They are useful mainly because they can reveal – better than native speaker intuition – what native speakers of the language in question typically write or say (either in general or in a situation / in a certain text type). For language teaching, however, it is not only essential to know what native speakers typically say, but also what the typical difficulties of the learners of a certain language, or rather of certain groups of learners of this language, are.

As seen above, there is a wide consensus over the limit of NS corpora and the necessity to look at learner corpora when a clear aim is to be achieved regarding the difficulties of a certain group of learners and the features of this group's learner English. The following part of this section introduces some of the prominent learner corpora and the corpus which is associated with this thesis, CLEC.

2.2.1 The International Corpus of Learner English

The International Corpus of Learner English (ICLE) is an international computer corpus of advanced learner English. This project was launched in the early 1990s by Sylviane Granger, of the Catholic University of Louvain, Belgium, with a world-wide collaboration of several universities. The corpus contains argumentative essays written by university students of English from different mother tongue backgrounds. By 2003, ICLE was composed of 15 subcorpora and each subcorpus represents the written English of a national variety with a size controlled at a level of 200,000 words. (The number of the subcorpora is increasing. See the website of ICLE for more information.[1]) The major scripts of the corpus are student essays of approximately 500 words. The variety and the size of the corpus keep expanding regularly. The corpus is well documented in the sense that it contains information about the individual writers' attributes such as age, sex, mother tongue, region, other foreign languages, and English proficiency level. The corpus is both POS-tagged and error-tagged. Information can be retrieved by computer automated software. ICLE was made available to public research in 2002 and researchers are now able to "enjoy the first harvest in the form of an ICLE CD-ROM" (Tono 2003: 800). The significance of the construction of this corpus cannot be overstated, because it has opened up a new avenue to exploring and interpreting learner language from a fresh perspective. As reported in Granger's edited work in CLC (Granger 1998), most initial studies in learner English analysis are based on this very corpus: ICLE.

2.2.2 The Longman Learners' Corpus

Another prominent learner corpus is the Longman Learners' Corpus (LLC), which aims to assist the compilation of English language teaching dictionaries and other ELT resources,

1 http://www.fltr.ucl.ac.be/fltr/germ/etan/cecl/Cecl-Projects/Icle/icle.htm, accessed on September 22, 2005.

according to Gillard and Gadsby (1998). The collection of the samples of learners writing was started in 1987 by Longman. In 1998 this corpus was reported to contain 10 million words in 27,000 individual scripts written by students of 117 nationalities at different levels of proficiency. This corpus is POS-tagged and has records of the writers' nationality, level of English, text type, target variety and country of residence. The earliest application of the corpus was in writing the Longman Language Activator which was published in 1993 (Gillard and Gadsby 1998: 160). The LLC played an important role in the compilation of the third edition of the *Longman Dictionary of Contemporary English* in 1995 and later the *Longman Essential Activator* (LEA) in 1997 (*ibid.*). The detailed application of LLC in the compilation of CIA will be discussed in section 2.8.3. This corpus is now available commercially to public research. Compared with ICLE, LLC has yielded a much smaller number of investigations (cf. Biber and Reppen 1998; Rundell and Ham 1994). However, this corpus is still significant in that it is one of the earliest learner corpora and also the one with the greatest number of nationalities among its contributors.

2.2.3 The Hong Kong University of Science and Technology Learner Corpus

The Hong Kong University of Science and Technology Learner Corpus has been collected and maintained by John Milton at the Hong Kong University of Science and Technology since 1992 (Milton 1998). It is composed of the writings of Hong Kong students submitted in electronic form. The "monitor archive", as Milton calls it, is ever-increasing, at a rate of about 3 million words (or about 6,000 scripts) a year. In January 2001, the size reached 25 million running words (or about 40,000 scripts). As the size grows, the topics expand too. The corpus is tagged for POS with CLAWS7 tagset. Errors are tagged manually and then the tagged texts are checked by a NS to ascertain the precision of the tagging. Since texts are collected automatically into the corpus by a central server when students submit their writing, it is becoming one of the largest learner corpora in the world.

2.2.4 The Chinese Learner English Corpus

The Chinese Learner English Corpus (CLEC) project was launched in 1997 in mainland China, with S. Gui and H. Yang as its leaders (Yang, 2001, Gui and Yang 2002). The corpus

contains student compositions of different levels of the English writing of learners ranging from middle school students to English-major university students taking degrees in English. The CLEC corpus has been used heavily especially by teachers of English in China since it was made available to researchers. As a component of CLEC, the College Learner English Corpus (COLEC, as I will call it henceforward), mainly made up of examination essays by university students not taking English as their main subject, will be explored in detail in this thesis. The whole corpus of CLEC is error-tagged but not POS-tagged and keeps the raw text version for possible individual research purposes.

This CLEC was made available for public research by the Shanghai Foreign Languages Education Press in the form of the book *Chinese Learner English Corpus*. This book is written in Chinese and introduces the construction of the corpus, the design of the error tags and some statistical analysis in the interpretation and description of CLEC writers. Attached to the book is a CD which contains the corpus CLEC and some concordancing tools: TACT, the WordSmith Tools[2], LEXA, and Corpus Concordancer (in Chinese interface). Some tables made in MS Excel are also provided on the CD. This saves researchers from repeating many laborious jobs if they retrieve the same thing. What is more, it has transferred the relevant data directly to the MS Excel environment and this makes further analysis much easier (for more details, see Gui and Yang 2002). It was planned that CLEC would be transferred onto the internet so that online retrieval could be undertaken, according to Yang (2001).[3] Even though an attempt has been made to list all the learner corpus projects around the world (Tono, 2003), it seems almost impossible to draw up an exhaustive list of all of them because of the fast development of the establishment of CLC studies world-wide.

2.2.5 Computer learner English studies as a 'newborn baby' of applied linguistics

Currently CLC studies appear to be mainly in Europe and Asia (Pravec 2002: 81); they are at the moment rare and sporadic in North America. But this has already been observed by North American researchers such as Cobb (2003). It can be envisaged that before long CLC will

2 WordSmith Tools, provided on the CD, is limited in function. For full function, registration is required.

3 Online concordancing is available at http://www.clal.org.cn/corpus/ChiSearchEngine.aspx, accessed on June 13, 2006.

spread more widely not only geographically but academically. Among the major journals studying English language learning and teaching, *TESOL Quarterly* has arranged a special-topic issue (Volume 37, Number 3, 2003) attempting to show "the multifaceted connections between corpus linguistics and TESOL" (editor's note). Another important journal in SLA, *Studies in Second Language Acquisition*, published a couple of book reviews introducing corpus linguistics as well. What is more exciting is the appearance of some corpus-based studies of learner language in *Second Language Research*, another key journal of SLA. These studies include Myles (2005) and Oshita (2000). However, another journal covering the same broad field, *Language Learning*, according to my recent survey of their volumes,[4] has had no publications on corpus linguistics at all, let alone CLC studies. This might be caused by a mistrust of the new methodology by researchers in the neighbouring disciplines. On the other hand, it seems that CLC researchers have not made the new methodology appealing enough to researchers in the neighbouring disciplines.

2.3 Typology of CLC data

To describe learner corpus typology, Granger (2002: 11) deploys four dichotomies, namely, monolingual vs. bilingual, general vs. technical, synchronic vs. diachronic and written vs. spoken. In fact, there are other perspectives to classifying corpus data types. For example, in terms of notation, the CLC can be kept clean and called "raw corpus" or "un-annotated corpus" or "plain text", or it can be added with special values such as POS or learner errors in which case it is labelled as an "annotated corpus". In this section, I will focus only on the following dichotomies: *synchronic* vs. *diachronic* and *written* vs. *spoken*, and *un-annotated* vs. *annotated*. (See McEnery and Wilson (1996), Kennedy (1998) and Horvath (1999) for a further classification of corpus typology).

2.3.1 Synchronic vs. diachronic

A *synchronic* corpus is a collection of texts written at a particular time and is used to reveal and "describe learner use at a particular point in time" (Granger 2002: 11). Contrary to a

4 This survey was conducted in December, 2005.

synchronic corpus, a *diachronic* corpus is used "to trace the development of aspects of a language over time" (Hunston, 2002: 20). This second type of corpus could also be called "longitudinal corpus" (Granger 2002: 11). Unfortunately, due to the difficulty of collection there are very few of this type so far, especially learner English corpora (*ibid*.). Since great interest exists in the development of group interlanguage (IL), researchers are trying to use a kind of corpus of learner English from different ages (from young to old) or levels of proficiency (from novice level to advanced level) so that the corpus resembles the structure of a longitudinal one. This type of corpus is termed "quasi-longitudinal" by Granger (*ibid*.). So far, most studies in CLC are based on synchronic learner corpora even though some research is also carried out in a quasi-longitudinal way (see Housen (2002) for an example). Diachronic CLC has a closer relationship with SLA than synchronic CLC because SLA has more concerns with the longitudinal development of learner language as discussed previously (see 2.1.2).

2.3.2 Written vs. spoken

Most current learner corpora fall into the 'written' category. As Leech (1998: xviii) says: "Writing is an exceedingly important skill for most foreign language learners, and well deserves the expenditure of effort to collect corpora of written learner language." Like the development of NS corpora, the compilation of NNS corpus has followed the pattern of written corpus first and spoken corpus second. "This tendency has dogged corpus linguistics from the start: the truth is that whereas humans are built primarily to process speech, computers are built primarily for the written word" (*ibid*.). Spoken data have to be transcribed into computer-readable codes. The advantage of a *written corpus* is the accurate rendering of the form of the language without distraction from spoken language features such as interruptions and repetitions. However, it does not expose the process of thinking and word-seeking information as a spoken corpus may. A *spoken corpus* contains the spontaneous utterance of language, which is more naturally produced. Compared with a written learner corpus, a spoken learner corpus may contain more errors because transcribers themselves make mistakes. Even though 'errors' of written learner English also exist the accuracy will increase when students submit their essays through digital form on computers and the raw data are automatically transferred into the corpus.

2.3.3 Un-annotated vs. annotated

An *un-annotated corpus*, as we noted above, is a body of clean text without externally added information such as POS or learner errors. It is generally known as "plain text" or "raw corpus". An *annotated corpus* is one with specifically designed "interpretative" and "linguistic information" encoded in a body of clean text (Leech 1997: 2). Since corpus annotation is becoming widely practised and acknowledged "as a crucial contribution to the benefit a corpus brings", it has become "an important and fascinating area" of linguistic enquiries as Leech observes (*ibid.*). There are competing ideas about the use of annotated corpora, which will be discussed in the following section.

2.4 Clean-text policy and annotation

There are two strikingly different views as to whether corpora should be kept clean as raw texts or annotated with more information such as POS or error-tagged information. Sinclair proposes a "clean-text policy" (Sinclair 1991: 21-22). The two strong reasons he holds are as follows:

> Firstly, each particular investigation is likely to view the language according to different priorities. Its analytic apparatus may well be valuable and interesting to the next investigator, and even adaptable to the new needs; but not so standardized that it can become an integral part of the corpus.
>
> Secondly, although linguists leap effortlessly to abstractions like 'word' (meaning lemma) and beyond, they do not all leap in the same way, and they do not devise precise rules for the abstracting. Hence, even the bedrock of assumptions of linguistics, like the identification of words, assignment of morphological division, and primary word class, are not at all standardized. Each study helps the others, but does not provide a platform on which the others can directly build.

Contrary to Sinclair's "clean-text" policy, Leech (1997: 2) views annotation as an added value to a raw corpus because "it enriches the corpus as a source of linguistic information for future research and development". Leech (1997: 4-6) provides three advantages of corpus annotation: "extracting information", "re-usability" and "multi-functionality". Leech argues that corpora become useful only when knowledge or information can be extracted from them. To realise this extraction, researchers would normally have to insert information into a corpus, which is

adding annotations. Leech does not believe that in its orthographical form a raw corpus can provide any direct information. One of the examples Leech (1997: 4) raises is the word *left*:

> Consider the word spelt *left.* As a word meaning the opposite of *right*, it can be an adjective ('my *left* hand'), and adverb (turn *left*) or a noun ('on my *left*'). As a past tense or past participle of *leave*, it is a verb ('I *left* early'). *Left* is therefore a very versatile piece of language – but its various meanings and uses cannot be detected from its orthographic form.

Accordingly, Leech points out that a grammatically-tagged corpus (POS-tagged) will make this distinction possible. With regard to "re-usability", Leech claims that "once the annotation has been added to the corpus, the resulting annotated corpus is a more valuable resource than the original corpus, and can now be handed on to other users" (Leech 1997: 5). He attaches a heavy weighting to this point since he views the feature of "re-usability" as a powerful one. Considering the fact that corpus annotation is a business entailing considerable expense and time, Leech emphasises, "We do not want to waste resources by 're-inventing the wheel' time and time again – i.e. by re-analysing or re-annotating the same corpus material" (Leech 1997: 5). As far as the third advantage, "multi-functionality", is concerned, Leech points out a multitude of applications of annotated corpora in practice. Among those mentioned are lexicography (as in his example of *left*), speech synthesis, machine-aided translation and information retrieval. Apart from the multiple applications of annotated corpora, annotation facilitates investigations with added value to a corpus in the general sense, making the use of the corpus open to multiple purposes. In connecting "multi-functionality" with the "re-usability" point, Leech continues to argue that "The re-usability of annotated corpus is enhanced by the fact that there are many different purposes for which others may wish to make use of the annotations: purposes which the original annotations of the corpus may not even have thought of" (Leech 1997: 6).

Even though strong opposition exists in the theories as to whether to keep texts clean or the other way around, this difference is not absolute. Actually, what Sinclair (1991: 29) advocates is not the total prohibition but the minimum use of annotations ("abstractions" in his own term) as shown in the following quotation:

> Hence, it is good policy to defer the use of them [abstract categories or abstractions] for as long as possible, to refrain from imposing analytical categories from the outside until we have had a chance to look very closely at the physical evidence.

On the other hand, Leech acknowledges that "we should not see annotations as having the claim to reality and authenticity which belongs to the corpus itself. For a written corpus, the text itself is the data …, and the annotations are superimposed on it" (Leech 1997: 4). This is perhaps the closest convergence point between the two lines of theories.

To adopt the practice of annotation or the "clean policy" may be dependent on the varying purposes and tasks of individual researchers. Hunston (2002) divides corpus analysis methodologies into two kinds: the "word-based" method and "category-based" method. According to her observations (Hunston 2002: 92), researchers prioritising individual words tend to go along with a plain text corpus, namely, one with a minimal annotation (for example, a corpus which is POS-tagged but not parsed). Yet, those who prioritise categories often have a preference for an annotated corpus, although with exceptions. In discussing whether to opt for a word-based or a category-based method of corpus analysis, Hunston (2002: 94) suggests "a synergy" between the two in which they can inform each other, "much as qualitative and quantitative methods of research complement each other". In the examples she raises (Hunston 2002: 94) Biber and his colleagues move between the two categories as needed in much of their corpus analysis; Thomas and Wilson move between frequency and interpretation in terms of phraseology when they work on semantic annotation. Hunston agrees with Conrad in that future investigations need to go beyond individual words but draws attention to the fact that "the interpretation of information found by looking beyond the concordance line frequently involves returning to those same concordance lines" (ibid). This is in agreement with Sinclair's emphasis on the use of plain text: "even in the time when annotated texts are becoming available and more choices are open to researchers, adequate attention should be drawn to the strength of patterning emerging from the rawest un-annotated data" (Sinclair 1991: 117). Since there needs to be constant movement between using sophisticated search techniques in an annotated corpus and looking at the raw data of language, Hunston (2002: 94) proposes "a mixture of plain text and annotation".

In line with Hunston's view, my thesis uses annotation technology to deal with verb lemmas (as in Chapter Four) and verb forms (as in Chapter Five) and raw data to study the syntactic patterns of the verb *KEEP* (as in Chapter Seven) and collocates of the verb *TAKE* (as in Chapter Eight). In cases where both the annotated version and the raw version can do the job

(as in Chapter Six), I prefer the raw version because "text becomes grossly overstuffed with tags" (Sinclair 2004: 191). To conclude, a selective and cautious use of annotation is my policy in this research.

2.5 Learner corpus annotation

Following the practice of corpus annotation in NS corpora, learner corpora are also widely annotated (see Pravec 2002), but mainly with POS and "errors". In terms of the POS-tagging to learner corpora, Aarts and Granger (1998: 140) claim that their study in tag sequence (based on traditional POS classification) in learner corpora "highlights the benefits of tagged corpora over raw corpora for the analysis of grammar and discourse features". Researchers can hope to gain totally new insights into learner grammar and discourse by adding the technique of tag sequence extraction to their supply of heuristic devices. Granger also favours annotation for particular research purposes. She encourages the use of annotated and in particular POS-tagged learner corpora because they facilitate "refined linguistic analysis" (Granger 2002: 18). As one of the leading pioneers in corpus annotation, Leech (1997: 15) also advocates error-tagging to learner corpora:

> The function of such corpora [learner corpora] is to advance our knowledge of how languages are learned as a second language: for example, to what extent does the English of non-native speakers reflect the influence of their native tongue? For this kind of investigation, it is very useful to annotate the corpus with classes of errors, or features of non-native language behaviour. Such 'error tags' make use of grammatical and lexical classifications, for example, but also take into account the relation between the non-native and corresponding native phenomena.

In explaining how to attach error-tagging to a learner corpus and how to benefit from it, Granger (1998b: 15) says: "Once an error taxonomy has been drawn up and error tags inserted into the text files, the learner corpus can be queried automatically and comprehensive lists of specific error types can be produced". To automate error-tagging, special software (Error Editor) is used in the Centre for English Corpus Linguistics in the Catholic University of Louvain.

However, the theory and practice have potential problems which need to be solved before error-tagging becomes widely accepted, and thus deserve more discussion (see 2.9.7 for more

details).

2.6 Contrastive Interlanguage Analysis and its data processing approaches

2.6.1 The notion of Contrastive Interlanguage Analysis (CIA)

As rightly pointed out by Hunston (2002: 206), "the essence of work on learner corpora is comparison: between corpora produced by different sets of learners, and between corpora produced by learners and those produced by native or expert speakers" (see also Tono 2003: 803-4 for the same view). The characteristics of learner IL will become obvious only when learner output is put into a comparison with some kind of norm (even though it is impossible to establish a norm acceptable to all the researchers in this field). Researchers also compare IL1 with IL2 for a specific purpose such as to clarify whether a certain kind of overuse by learners is caused by mother-tongue influence. Considering the contrastive approach of traditional Contrastive Analysis (CA), Selinker calls this new approach to comparison 'a new type of CA' and Granger refers to it as Contrastive Interlanguage Analysis (CIA) (Granger 1998b: 12). Literally, CIA seems to refer to the analysis between ILs, but actually Granger means not only the comparison between IL1 and IL2 but also between a particular IL and the target language.

2.6.2 Quantitative plus qualitative: approaching CLC data

Using computer software to retrieve information is the most salient feature of CLC, arising from the fact that CLC are originally made in such a way that the data can be stored in large quantity in computers and, what is more important, they can be easily retrieved by software. There are several kinds of retrieval software in use for different research purposes. Some researchers develop their own software for special purposes. MicroConcord, the WordSmith Tools are among the most often-used ready-made retrieval tools in CIA. MicroConcord is a DOS-based concordancer with the function of KWIC (key word in context). The number of concordance lines is limited to around 1500 and a concordance can only be saved as a text file. The WordSmith Tools (3.0) is Windows-interfaced and accepts different text formats such as DOS, Text only, ASCII and ASNI (Scott 1999: 10). The WordSmith Tools (3.0) can compute as many as 16368 lines of concordance using Concord each time (for details of the software,

see Chapter Three, 3.3.1).

However, no matter how helpful computer software has been in retrieving the information researchers need "a computerized approach has linguistic limitations," as Granger (1998b: 16) acknowledges. She suggests that researchers should not limit their investigations to what the computer can do. She insists that a computer approach is ideally suitable for the analysis of lexis and to some extent grammar but it is much less useful for discourse studies, and stresses the necessity of manual analysis where existing software is inadequate. Apart from the applicability of computer approach in different aspects of linguistics, there is a problem of superficiality of computer retrieved data. "Surface differences – or similarities- between aspects of native and non-native language always require further qualitative investigation" (Meunier 1998: 36). Computer automation is a vital assistant to any corpus analysis. But without intelligent human scrutiny, the computer-retrieved data are nothing more than a list of figures and codes. Computer-retrieved output is clearly preliminary in nature and only serves as a starting point for further analysis. Filtering the computer-retrieved data for meaningful information should be the core of CIA and it takes strategies to transform the raw computer data into a refined piece of work potentially useful for the investigation (see De Cock *et al*. 1998 for the three steps they take in order to get a list of potential formulae for vagueness tags). In a review of learner corpus studies, Hunston points out (2002: 207): "The studies in Granger's collection [Learner English on Computer] are quantitative rather than qualitative in nature, but there are interesting qualitative generalisations to be made."

2.7 Learner English features

CLC is a fairly young field of study but is growing at a fantastic speed as Leech has acknowledged in several places (for example, Leech 1998 & 2001). The investigations are beginning to yield enlightening results. This section will review some of the striking features of learner English as reported in the literature (for a review of learner English features, see Hunston (2002: 206-212)).

2.7.1 The informal and speechlike features of written learner English

Essays are normally expected to be formal and academic. However, a striking feature of learner English found by many researchers is the informality and speech-like nature of learner English writing. Typically, this involves features of speech such as a large amount of use of first and second person pronouns (Granger and Rayson 1998), high writer/reader (W/R) visibility according to Petch-Tyson (1998), more use of verbs over nouns (see Chapter Six, also Guo 2003), less use of prepositions (Aarts and Granger 1998). The following are some examples of studies that report the evidence of the informal style of learner English writings.

Granger and Rayson (1998) compare French-speaking learners' argumentative essays from ICLE with LOCNESS[5] in an attempt to identify the salient features of learner English writings. In their study, "the learner data is shown to display many of the stylistic features of spoken, rather than written, English" (*ibid.*: 119). For example, learners dramatically overuse[6] the first person and second person pronouns. A number of scholars[7] "associate the feature with the involved nature of speech and point to the low frequency of indices of personal reference in academic writing" (*ibid.*: 126). In the detailed study of verbs, the overuse of auxiliaries is the first striking feature, "a characteristic of conversational English" (*ibid.*: 128-129). The second striking feature concerns the underuse of the finite form of lexical verbs and participles (both present and past) and the overuse of infinitives. This is not what one would expect from an academic text. According to Chafe and Danielewicz (1987: 101) (cited in Granger and Rayson 1998: 129), "language other than academic writing makes considerably less use of participles". Also, in O'Donnell's view (*ibid.*), "a high frequency of infinitives, which goes together with a high frequency of auxiliaries, is indicative of speech". With regard to the use of nouns, Johansson (1985: 30) and Svartvik and Ekedahl (1995: 27) (cited in Granger and Rayson 1998: 128) link the underuse of nouns to the category of imaginative texts and conversations. Biber *et al.*'s study (1999: 65) reaches the conclusion that "Nouns (excluding pronouns) are more frequent in news and in academic prose than in other registers, and least frequent in conversation" (cited in Hunston 2002: 162). In studying the underused

---

5 The Louvain Corpus of Native English Essays (LOCNESS); for details see 3.2.2, Chapter Three.

6 I am using 'overuse' and 'underuse' in order to follow the currently popular terms in learner corpora research. For my reservations with the use of these terms, see 2.9.6, 7.1, 7.2 and 7.4 of the thesis.

7 Poole and Field 1976, Chafe 1982, Chafe and Danielewicz 1987, Biber 1988, Petch-Tyson 1998, etc., cited in Granger and Rayson (1998: 126).

nouns, Granger and Rayson (1998: 128) find that learner English is short of a set of items which are normally considered to be the vocabulary of argumentative writing such as: *argument, issue, belief, reasoning, claim, debate, controversy, dispute, support, advocate, supporter, proponent, denial.* They emphasise that "The overall underuse of nouns that characterizes French learner argumentative writing is thus clearly a further sign of a tendency towards oral style". They call upon further research into nominalisations which have been shown to be of great importance in academic writing by Chafe and Danielewicz (1987: 99, cited in Granger and Rayson 1998: 128). In my earlier research (Guo 2003) I found that learners use more verbs whereas NSs (from LOCNESS) prefer nouns (also see Chapter Six). In almost all the 25 verbal concepts I choose, learners have a much stronger tendency to use the verb form than the noun form. For example, the learners in my NNS corpus use the verb *accept* 41 times but do not use the noun *acceptance* at all whereas NSs use the verb form 182 times and the noun form 33 times. The learners use the verb *introduce* 12 times but the noun *introduction* only 2 times while NSs use the verb form 61 times and the noun form 44 times. On average, the writers use verbs two and a half times as often as the native writers in comparison with nouns. This further supports Granger and Rayson's findings (1998) that underuse of nouns is a characteristic of learner English, which contributes to the overall feature of orality and informality of learner English writings.

Biber (1988: 102) and Biber *et al*. (1998: 148) (cited in Hunston 2002: 164-65) find a co-relation between the use of nouns and prepositions. The co-occurrence of one linguistic feature with another is regarded as an example of 'association patterns' by Biber (1996: 173, cited in Hunston 2002: 164). In the words of Biber (cited in Hunston *ibid.*), these are "the systematic ways in which linguistic features are used in association with other linguistic and non-linguistic features". According to the research findings of Biber (1988: 102) and Biber *et al.* (1998: 148) (cited in Hunston 2002: 165), nouns not only co-occur with prepositions but also with other formal register linguistic features such as long-length words, a large number of types relative to the number of tokens, agentless passives and reduced relative clauses beginning with past participles. The underuse of nouns and prepositions is also discovered in Granger and Rayson's study (1998: 127).

Altenberg and Tapper (1998) analyse the Swedish subcorpus of ICLE and find that the

Swedes in their argumentative essays produce a language similar to fiction and informal talk. For example, learners tend to overuse the contrastive connector *but.* They (*ibid.*: 87-88) claim: "This is a clear indication that the Swedish learners tend to avoid formal contrastive conjuncts like *however*, and *yet*, replacing them with more informal equivalents."

In their comparative study of complement clauses (*that*-clause, *to*-clause, *ing*-clause, and *WH*-clause), Biber and Reppen (1998: 157) find that "the patterns of use in the learner essays are very similar to those found in native conversation and fiction, but strikingly different from those found in native academic prose". While an obvious difference exists between conversation and academic prose in NS (in the case of complement clauses with *think, say, know, show,* and *hope*, almost no difference is shown in the usage of NNSs (French, Spanish, Chinese and Japanese) (Biber and Reppen 1998: 152-153). Biber and Reppen (1998: 154) report that "all four languages are additionally similar to conversation in that they use the verb *want* very frequently controlling *to*-clauses".

Aarts and Granger (1998: 137) find learners' underuse of sentence-initial nouns, in parallel with an overuse of sentence-initial pronouns, which is "undoubtedly at least partly related to the higher degree of involvement that characterizes learner writing". They also find an underuse by learners of the structure: the sentence-initial preposition-headed "-ing" clause, such as: "By arguing that …", and "By using this example", which plays an important frame-setting or linking role in academic writing.

Unlike the perspectives of the learner English studies above, Petch-Tyson (1998) compares learner English with NS English from the view of W/R visibility, which means writers interact directly with their readers. The features of W/R visibility are mainly marked by "high use of, among others, first person reference, pragmatic markers (such as *I mean*, *you know*), fuzzy reference and direct quotes" (*ibid.*: 109). All the learner writers were found to use to some extent almost all of the features of W/R visibility much more often than the control NS writers, and can thus be said to focus more on interpersonal involvement. The features under investigation include first person singular pronouns *(I, me, my, mine*), first person plural pronouns (*we, us, our, ours*), second person pronouns (*you, your, yours*), fuzziness words (*kind/sort of*, *and so on*, *etc*.), emphatic particles (*just, really*), and reference to situations of

writing/reading (*here, now, this essay*). As a result, Petch-Tyson (1998: 116) believes, their writing may be felt to deviate from "the conventions of the particular genre."

In a replication of Petch-Tyson's work of 1998 in W/R visibility, Cobb (2003) carried out an investigation into all the first and second person pronouns in a corpus of advanced learner English built in Quebec. The first and second person pronominal amount reaches "a total of 6.47% of the words in the advanced learner corpus", signalling strong interpersonal involvement, as opposed to message content, for these Quebec learners (*ibid.*: 418). He vividly describes the oral nature of learner English as "talk written down" (*ibid*.: 415).
In accordance with Petch-Tyson (1998) and Cobb (2003), Wen *et al*. (2003) probed into the features of advanced learners of English in China and found that there is an obvious employment of a spoken type of discourse in learner English writing. Even though disparity exists in different learner groups of different mother tongue backgrounds, their study also shows the obvious universality of high W/R visibility in all the learner groups under investigation compared with the usage of NSs. In the continuum of W/R visibility they make, the order of sequence from high to low is: Swedish, Finnish, Chinese, Dutch, and French. On average these learners overuse the high W/R parameters by about three times according to the continuum (Wen *et al*. 2003: 271).

Ringbom (1998b: 48) also found the overuse of some auxiliaries and personal pronouns, and the underuse of prepositions. Furthermore, there are other important findings of learner English writing that add up to a generally raised degree of orality as against literacy and informality as against formality. For example, learners underuse the passive voice compared with the active voice (Granger 1997). There is a strong tendency for learners to overuse the base form of a verb among all its other forms (Guo 2003), and to overuse direct questions (Virtanen 1998). Due to the large amount of studies and fast development in this aspect of learner English study, it is difficult to be exhaustive here in talking of the outstandingly informal and oral style of written learner English.

After abundant support is provided to show that learner English writing style is strongly characterised by oral and informal English, it is natural to develop an idea that learner English speech will resemble the style of NS speech since it has been discovered by a large number of

studies that learners are familiar with the oral and informal style of the language as evidenced in their writings. Surprisingly, however, just as learner English writings are not like the style of written English by NSs, learner English speech is unlike the style of spoken English of NSs. In the study of learner English "phrasicon" by De Cock *et al*. (1998), they found learners underuse expressions for vagueness markers such as *sort of* and *kind of*. NSs normally use verbs (30-35%) to follow these two markers for this purpose whereas learners follow them with nouns almost without exception. To be vague is one of the most essential features of informal conversation according to Crystal and Davy (1975) (cited in De Cock *et al*.: 98). The vagueness, however, is absent in the learner English De Cock *et al*. studied.

Reports on the resemblance between learners' spoken English and NSs' written English are rare at this moment because learner English study is young and the most recent investigations are mainly committed to written English, which is rather unbalanced considering the proportions of speech in language use. However, it is envisaged that before long similar findings will appear in learner English studies.

2.7.2 Small vocabulary range, overuse of general vocabulary and the 'teddy bear principle'

Apart from the oral and informal style of learner written English, there is another prominent feature: small vocabulary range and overuse of general vocabulary in learner English. According to the studies by Gillard and Gadsby (1998: 161):

> One of the first things that is easily noticeable about learners' vocabulary is the way they use the most common words in the language, particularly the common adjectives. These words are much more common in learners' English than in native speakers' English, and they are more common in lower-level learners' English than in higher-level learners' English.

They compared some learners's use of two commonly used adjectives: *nice* and *happy* against the British National Corpus (BNC). The result shows that the average use of *nice* by learners is about ten times more and the average use of *happy* is six times more than that of the NSs (*ibid*.). They ascribe the overuse of the commonly used vocabulary to the lack of alternatives in the mental lexicons of learners. Their study also shows that learners do not usually have access to a wide range of synonyms for particular meanings. They tend to show a particular preference towards a particular concept. For example, instead of using *big, enormous, massive*

and *huge* alternatively, NNSs are more likely to use *big* as a default term for the other alternatives (*ibid.*).

Ringbom (1998b), in his cross-linguistic research into learner English, found that the learners overuse high-frequency verbs. For example, the NSs use *think* only 6 times per 10,000 words whereas the NNSs (of French, Spanish, Finnish, Finland-Swedish, Swedish, Dutch and German) on average use it 23 times per 10,000, which means the NNSs use this verb nearly 4 times as often as the NSs (*ibid.*: 44). Other conspicuously overused high frequency words include *get, make, become, want, take, find, know, use, go* and *live* (*ibid.*: 44).

In terms of the comparison of nouns, Granger and Rayson (1998: 128) mention the "overuse of general and/ or vague nouns such as *people, thing, phenomenon, problem, difficulty, reality, humanity*". Kaszubski's comparison between the Polish and NS corpora (Kaszubski 1998b: 181) indicates that "Poles overuse hypernyms as a whole set, and also in a number of individual cases – five lemmas: *case, factor, kind, situation, thing*; and two word-forms *conditions* and *time*." This is supported by Cobb's replicating work (Cobb 2003). His comparison evidences learners' overuse in "general, unnuanced lexical items" such as *things, problem, position, change, strong* and *everyone* (Cobb 2003: 402).

The existence of small vocabulary range or overuse of general vocabulary in learner English can be interpreted as the 'teddy bear principle', which is explicitly illustrated in Hasselgren's study into the English of some Norwegian learners (Hasselgren 1994). By proposing the 'teddy bear principle', Hasselgren compares learners who are over-dependent on the easy set of vocabulary items they are familiar with and stick to it constantly to children who hold their teddy bears before going to sleep. According to her study, 'core items' [general vocabulary] such as *very (much), a lot (of),* and *extreme(ly)* as intensifiers are much more likely to occur in learner English than in NS English.

In their creation of a "new conceptual map of English", Rundell and Ham (1994: 178) make use of the multi-nationed learner corpus (LLC) to display the feature of generality in learner English vocabulary:

> [W]hen students want to convey a message which they lack the lexical resources to express

> precisely, they tend to start from the basic-level terms they already know. This resort to high-frequency default terms is a classic 'communication strategy' of the type described by Pit Corder and others (see e.g. Pit Corder 1983). And the use of a 'superordinate-plus-paraphrase' strategy (for example, 'steal from a shop' for *shoplift*, or 'listen in secret' for *eavesdrop*) is a pervasive feature of learners' text particularly at intermediate level and above.

They exemplify the use of default terms with a set of words such as *interesting, fascinating, intriguing,* and *riveting.* Rundell and Ham (*ibid.*) report that "the first item is easily the most frequent of the four in all types of text" and that hundreds of similar sets of patterns can be found in the corpus. The finding of the default term use of learner English helps them in the process of concept creation and concept naming.

2.7.3 More open-choice-principled than idiom-principled

Sinclair (1991) put forward his influential proposal of the 'open-choice' principle and 'idiom principle'. This theoretical construction influences the corpus study of learner English. There are a number of reports that suggest that learner English is more controlled by the 'open-choice' principle than by the 'idiom principle'. The following is one of the examples.

Ting and Wen (2003), in studying the relationship between the command of formulaic sequences and oral English performance, find that learners lack knowledge of formulaic sequences. This is especially true when some sequences have no similar counterparts in the native language (NL): Chinese. For example, there is no evidence to show mastery of sequences such as 'no sooner had he … than …', 'it looked as though,' and 'forced its way'. They also detected that where there are alternative sequences realising a particular meaning learners tend to choose the one closest to the NL. For example, between 'went immediately' and 'immediately went', most students choose the latter whose sequence order in Chinese is the same as the NL. In the conclusion, they recommend study by memorizing formulaic sequences. This point agrees with what has been put forward by Kjellmer (1991: 125) in terms of learning collocations as follows:

> Pupils and students who have acquired 'collocational learning habits' at an early stage can be expected with some confidence to pursue their further studies of lexis in a more fruitful way than would otherwise have been the case. It is only when the student has acquired a good command of a very considerable number of collocations that the creative element can be relied on to produce phrases that are acceptable and natural to the native speaker.

In comparing how NSs and learners manage to speak the English language, Kjellmer (1991: 124) maintains that NSs have acquired a large portion of "prefabs" which learners can only hope to use whereas the learner's building material is "individual bricks rather than prefabricated sections". Since formulaic sequence plays an essential role in the acquisition of a language, the characteristics of the learners in using it will be addressed in Chapter Seven and Chapter Eight.

Cobb (2003: 411-412) extended De Cock *et al.*'s examination by looking at the 'phrasicons' in the pattern "verb + out" and the findings suggest that "As with phrases in general, these advanced learners clearly do use out-phrases, but fewer of them and with more repetition." Cobb also examined some other verbs followed by "out" and similar results are obtained. He even carried out the examination with other phrase types and the result yields similar and complementary findings. As a result of the replication, Cobb's work reinforced the impression that learners do indeed use the 'idiom principle' but not as thoroughly and appropriately as NSs do. After the analysis of phrases, Cobb concludes (2003: 412, italics added) that "the pattern is the same for phrases as it was for basic vocabulary in the replication of Ringbom: *fewer items repeated more*".

2.7.4 Proficiency level and fossilised errors

There is very little doubt about the everlasting nature of development in adult SLA. It should follow that if the development is adequate, learners' errors would disappear completely once learners reach a certain high proficiency level. However, the current studies in CLC do not support this hypothesis.

In a study of four groups of English learners (university English majors from Year One to Year Four), Wen *et al.* (2003: 272) examined the written English of these students from the perspective of W/R visibility. The data reveals an apparent tendency to decrease from Year One to Year Four. For example, the occurrences (per million) of the plural form of the first person pronoun (*we, us, ours*) change from Year One through Year Four as follows: 326, 280, 255 and 80. However, no matter how much the number drops, and how obvious the decreasing tendency is, there is no sign for the overuse of pronouns to disappear from the

W/R variables with the development of the student's English level as a whole.

Chen (2002) compared the percentage of the misuse of the passive voice by several groups of different levels and found that as a whole the higher the level, the lower the misuse percentage. She claims that this proves the continuous process of interlanguage improvement. Although the amount of misuse of the passive voice decreases with the improvement of the English level of the group learners, the difficulties shared by all the groups remain the same across all the sub-categories of passive voice misuse: for example spelling mistakes in the verb, and the underuse of the passive voice.

Cobb's replication work (2003: 404) also supports this point that the overuse seems to decline with time and greater proficiency, although slowly, even though he has no comments on the universality of difficult points for learners.

In addition, the problem of overuse of existential *there* in the community of Chinese students is raised and studied by Lei (2003). She compared three groups of English learners in CLEC. The result shows that with the increase of the learners' English proficiency the existential *there* tends to drop in frequency. But its use even by the highest-level students is far above the average use in NS writings, resembling that of NSs in conversation. This means that the overuse of the existential *there* is less problematic when the learner English improves in groups. It is still out of the question for these learners to reach the stage of native use. The difference between the different levels of learners is only a matter of quantity rather than quality.

However, there is an interesting counter-example in the work of Cui and Huang (2003). Instead of showing a drop of a certain item across groups of learners with the increase in English proficiency, their data shows that the number of difficult points increases among groups of learners with the development of proficiency in English. The difficult point under investigation is the use of affixes. Unlike many other linguistic items which begin to emerge at quite an early stage in the process of language acquisition, affixes start to be used by learners rather late. This creates the pseudo-message as stated above. One of the possible reasons Cui and Huang (2003) provide (citing Hatch and Brown 2001) is that affixes are

avoided by learners at early stages of acquisition. Only when learners reach a certain degree of proficiency and realise the importance of affixes in the new language will they start practising them, leading to errors now and then. It can be predicted that if there are sufficient groups of high-level learners to be observed, the general tendency for occurrences of difficult points to decrease with the increase in proficiency will gradually appear. Even in this seemingly counter-trend example, there is one problem found to be predominant in all the categories of error throughout the examined groups: the spelling of the affixes. They conclude that learners with different proficiencies are faced with similar learning difficulties.

2.7.5 The essential role of L1 in L2 production

As one of the most often discussed issues, the essential role that L1 plays in L2 production is ascertained by authentic learner English data. In analysing the underuse of prepositions of Finns, especially multifunctional prepositions such as *with, by* and *at,* Ringbom (1998b: 48) states: "This must be seen against the background of the Finnish language: in Finnish the relationships expressed by prepositions in the Germanic languages are normally indicated by case endings, which, however, have several other functions as well."

In his survey of Chinese learners' use of English verbs in grammatical and lexical patterns, Pu (2000a: 37-41) noticed the existence of one-to-one semantic mapping from the learners' native language, Chinese, to English. For example, in the first place, learners map the sense of *serve* to *fu wu* (in Chinese pinyin[8]), i.e. *serve = fu wu*. Since *fu wu* is more often intransitive, linked to a noun by the preposition *wei,* which means *for*, learners will tend to apply the idiomatic structure '*wei … fu wu*' to the English language situation after shifting the position of the Chinese preposition and placing it after *serve*, attempting to meet the requirements of the English system. Predictably, phrases such as '*serve for the people*' and '*serve for the society*' will appear in the interlanguage of English learners whose L1 is Chinese. An advanced search[9] in Google yields 92 hits of '*serve for the people*', most of which have a link to the Chinese community.

---

8 Chinese pinyin resembles English phonetic symbols in that both of them have the function of marking the pronunciation of the written form of the language.

9 Conducted on May 14, 2004.

In the same vein, Lu (2002) found that learners of English in China tend to overuse some expressions that have direct translatable equivalents such as *we/us college students*, *with the development of* and *if you want to do something*, which indicates L1 transfer in learner English.

In a study similar in nature, but from a perspective of case grammar, Yang and Ning (2002) compared learners' interlanguage with L1 (Chinese) and L2 (English) and concluded that it is the negative transfer of cases in L1 that accounts for the difficulties of English learning that cause the deviances of learner English.

In the four complementary clauses investigation (*that*-clauses, *to*-clauses, *ing*-clauses, and *WH*-clauses) by Biber and Reppen (1998: 150-151), the learners (of French, Spanish, Chinese and Japanese) are found to underuse '*ing*-clauses' and '*WH*-clauses'. Since none of these L1s (in their judgment) allows participial clauses or '*WH*-clauses' serving as complement clauses, Biber and Reppen concluded that the differences between NSs and NNSs seem to reflect L1 transfer to the target language. Biber and Reppen (1998: 151) refer to the transference of preferred use in patterns from a first language to a second language as the 'use of transfer' (citing Wu, 1995).

2.7.6 A narrower range of senses in the use of vocabulary

Some CLC studies have also found that learners use a narrower range of senses of multiple-sensed vocabulary. This feature of learner English is not as much reported as other features above. Nevertheless, it is too important a point to miss out in the construction of a linguistic map of learner English.

Ringbom's analysis (1998: 44-45) in his cross-linguistic learner English data detects an unbalanced sense spread. In the four main uses (as he summarises), learners overuse the structure of 'get + objective' in which the meaning of the verb is 'obtain'. NSs use this word 2 times per 10,000 in the structure whereas NNSs use it as often as 8 times on average.

Pu (2000b) describes in his survey of learner English the behaviour of English verbs in grammatical and lexical patterns as follows: "The meanings that the learners intend to convey by the use of a certain verb tend to be uniform and unvaried, while the native speakers often use the same verb to convey varied meanings." For example, in the three verbs examined, Pu noticed the overwhelmingly dominant use of *serve* in the pattern of 'V+N' (63%) as in *serve the society* and *serve the people*. But no cases were found in the patterns such as 'be V-ed' and 'V as N' in learner English (34% for these two patterns in the Brown Corpus).

The features as evidenced in the literature are only the most outstanding ones from the perspective of this thesis. There are other findings and classifications to the findings in this field that may appeal to other investigations (see Tono 2003: 804-806).

2.8. Applications of research results

Applications of CIA are mainly evidenced in language learning and teaching. Within this broad area, great efforts have been made to probe into the possibilities and approaches to utilising the research products both in the context of classroom and electronic background. Textbooks are being written to enhance the writing competence of English learners. Dictionary compilation is another area where the features of IL are considered as a priority compared with traditional dictionaries. The following are some examples to show how the research results in learner corpora could be applied to the above-mentioned areas.

2.8.1 TeleNex

Introduced by Allan (2002), TeleNex is an internet network designed for teacher training of second level English teachers in Hong Kong. This network is based on the TELEC [10] Secondary Learner Corpus (TSLC) which contains over two million words. The TeleNex network comprises two hyper-linked databases called TeleGram and TeleTeach and a series of theme-based conference corners. [11] While TeleGram serves as a resource of grammar

10 TELEC refers to Teachers of English Language Education Center, University of Hong Kong.

11 Even though full access is restricted to registered English teachers in Hong Kong, a sampler of files can be viewed at http://www.TeleNex.hku.hk, accessed on February 17, 2004.

instructions customised to Hong Kong teachers, TeleTeach offers materials supplementary to course books, which can be printed and used in classes. Raw data from TSLC is used to produce teaching files for TeleTeach. What is recommendable with TeleNex is the referential use of modern English corpora such as the BoE when investigating the learner corpus. These investigations in turn reveal significant information which is later drawn on to answer teachers' questions through TeleNex conference corners. For example, among the interesting findings from the learner corpus, *besides* is found to be apparently overused, especially at the sentence initial position (90%). However, the data in real modern English shows that its syntactic function is both intra-sentential and inter-sentential. Such exploratory work not only helps teachers to check and correct conventional reference grammar books and dictionaries but also helps them to explain and illustrate points of grammar and usage. By means of systematic linguistic analyses of the difficulties Hong Kong secondary students experience, the problems of students are classified into twelve function areas and made into files under 'Students' problems' in TeleGram. Primarily, TeleGram is designed for teachers, for pedagogical purposes, containing five core files: Overview, Teachers' quiz, Misconceptions, Students' problems and Teaching implications. Through these files teachers' interest in and awareness of key points are aroused and teachers' attention is drawn to the areas of misunderstandings shared by students. Afterwards, specific problems of Hong Kong secondary students are focused on and what is most important, at the end, methodologies are shown of how the grammatical information with regard to a particular area can be dealt with. TeleNex has not only contributed to English language teaching and learning in Hong Kong, but also has yielded quite a number of academic articles.[12]

2.8.2 CALL Tools

Milton (1998) conducted a study in a learner corpus based on POS- and error-tagged data. In the first stage of his study, he made an analysis of the learner corpus from a lexical-grammatical view in which it was made possible to find the most common and serious errors of the learner group. In the second stage he carried out a word-sequence analysis which resulted in significant findings. The essence of these findings is that "the NNSs make use of a

12 For detailed research output based on TeleNex, see the following website (accessed on February 17, 2004): http://www.telenex.hku.hk/telec/smain/sintro/intro.htm.

much smaller amount of word sequences than the NSs, but the degree of NNSs' using high frequency expressions in their capacity such as *First of all* and *On the other hand* is 'startling'" (Milton 1998: 191). In contrast to this, the NSs use the most common expressions infrequently but appropriately because NSs have a much wider repertoire of lexis and syntax and are not limited to any one string all the time. The data in the previous analyses was exploited to develop tutorial exercises and CALL tools to assist these learners to be sensitised to and to correct the most frequently occurring errors of their own learning community and to reduce learners' liability to stick to a small subset of expressions. The brief outline of the components of the electronic tool is quoted below (Milton 1998: 192):

- an error recognition (i.e. 'proofreading' or 'editing') exercise intended to sensitize learners to the most common or most 'serious' errors exposed by the first analysis;

- a hypertext online grammar designed to give context-sensitive feedback, based on these errors;

- databases of the 'underused' lexical and grammatical phrases exposed by the second analysis; and made interactively available to learners from their word processor; and

- a list-driven concordancer which interacts with text in these programs and databases.

To use CLC study findings in CALL is a fascinating area of language learning and teaching and helps teachers to make classroom tasks easier than ever before.

2.8.3 Dictionary compilation

English dictionaries for advanced learners have been compiled with frequency in consideration; examples include the *Oxford Advanced Learners' Dictionary*, the *Collins Cobuild Dictionary* and the *Longman Dictionary of Contemporary English* (Leech 2001: 329). But they are not based on the evidence of learner English. Gillard and Gadsby (1998) report their exploration in making extensive use of a learner corpus LLC in compiling the LEA. The first step towards compiling a dictionary is to decide what to include in this dictionary in order to maximise its usefulness for target users. To make this decision, they generated frequency listings from the LLC so see which vocabulary was being used by learners at a

particular level. In the process of examining learner English, they found the dominating feature to be the overuse of common adjectives such as *nice* and *happy*. This information was decisive in helping them to make a general blueprint for the LEA. Unlike most other dictionaries, LEA was made into a "production dictionary", which means it was designed for producing English rather than consulting to find the meaning of a new word or phrase encountered. The words in this dictionary are shown with near-synonyms under approximately 1000 "concepts" such as *WALK* to go together with *stroll, stride, amble,* and *jog*. In order for students to distinguish these words, definitions and examples are given in detail. It is also shown exactly how and when a particular word should be chosen over others. Gillard and Gadsby examined each name used for the "concept" in LLC to confirm that the vocabulary they selected for the name of each "concept" posed no problems for students. They claim that: "The skill of lexicography for ELT dictionaries lies in being able to write definitions which are clear and which accurately pinpoint the key aspects of the word or phrase being defined" (Gillard and Gadsby 1998: 163). To meet the level of target users, the intermediate level for LEA, words and phrases are defined within the basic 2000 words, in which LLC was consulted and checked, and afterwards testing was done to scale the knowledge of students about the words in context. By finding out the most often-occurring errors common to all learners and "correctable enough" to them, they drew up a number of *help boxes* to warn users not to make such errors in their production of English. The ultimate aim of the dictionary is to help learners of English to withdraw from the over-reliance on a small number of common words in their early acquisition, and to accurately and naturally make use of a much wider range of words and phrases. Gillard and Gadsby (1998: 170) believe that it would be "a very odd idea" if an ELT dictionary were compiled without access to the information from a learner corpus. In contrast to this, "By having constant access to a very large body of students' writing, lexicographers are sensitised to and reminded of the needs of their audience far more thoroughly than they could achieve through their previous teaching experience" (Gillard and Gadsby 1998: 163). This obviously has set a new trend for lexicographers to take the features of learner English into consideration for dictionary compilation.

2.8.4 Textbook enhancement

Attempts to enhance teaching materials by using the findings from CLC studies have been rare but valuable. Kaszubski (1998b) for example, based on his investigation into ICLE, points out that text books designed for international learners around the world may not suit perfectly a particular learner community. He suggests innovations to the traditional writing textbooks in Poland by adding some specific information as listed below (*ibid*.: 183):

- longer lists of synonymous items, accompanied with frequency band information, register/style description, and (gradable) overuse/underuse/misuse warnings (if applicable). In cases of misuse, Polish and NS contrasting samples could be given;

- […] lists of common collocations, with additional information on contrasts between Polish and NS use;

- listings of commonly misused words and phrases as well as examples of serious over- and underuse.

These suggestions may not only apply to the Polish textbook writers, but also to textbook writers of other nationalities. Kaszubski advocates a strong collaboration between CLC research teams and ELT publishing houses so that the right type of learning aid can be developed to meet the needs of target users. Since CLC is a new phenomenon and analyses based on CLC studies are far from comprehensive, "it is not surprising that learner corpus studies have not yet had any remarkable impact on pedagogic material", according to Nesselhauf (2004: 137). As a result, "there has been no influence so far on printed teaching material such as textbooks or workbooks" (*ibid*.). The topic of how to use learner English study findings in textbook design will be discussed in more detail in Chapter Nine.

2.8.5. Data-driven learning

Data-driven learning (DDL), developed by Tim Johns, is another area where learner corpora can be used for language pedagogy. (For an introduction to DDL, see Hunston 2002: 170ff and for a discussion of more details see Sripicharn 2002.) The essence of DDL, according to Hunston (2002: 170), is that "students act as 'language detectives' (Johns 1997: 101), discovering facts about the language they are learning for themselves, from authentic examples". A considerable amount of investigations have been made by using NS corpora, for

example, Gan *et al*. (1996), Blappert (1998), Kennedy and Miceli (2001) (cited in Sripicharn 2002), Cobb and Horst (2001), Dodd (1997), (cited in Hunston 2002: 171-172), Hunston (2002), Sripicharn (2002) and Hadley (2002). However, DDL with learner corpora is much less reported due to the nature of learner corpora and the newness of CLC studies (Nesselhauf 2004). The first DDL study based on both NS English data and NNS English data was most probably made by Granger and Tribble (1998). Among the few reported DDL studies with learner language data are Flowerdew 2001; Horvath 2001, Milton and Hyland 1999; Ragan 2001 (cited in Nesselhauf 2004: 139) and Sripicharn 2002.

The shift from concordancing NS data alone in DDL to comparing NNS data with that of NS data is a great leap forward for CLC studies. As observed by Nesselhauf (2004: 140), there are advantages of using learner data compared with using NS language data:

> One of these advantages is that asking learners to look for mistakes, or rather for differences in learner and native speaker language, can increase learner autonomy and train the learners' general ability to notice such differences. In addition, such a procedure might also lead to a more positive attitude towards mistakes, because mistakes are then no longer merely a feature that has to be corrected, but also a feature that can be discovered. … Data-driven learning with learner data is probably particularly useful for points which have already been covered in the classroom, possibly even repeatedly, but which the learners nevertheless still get wrong, learners have the opportunity to get something right, namely to identify and explain the mistake in question.

Since DDL by learner data is a brand new area of language learning and teaching, Nesselhauf calls for more empirical studies to solve the problems such as "for which areas, for which learners and with what procedures data-driven learning with learner corpora is most efficient". Considering the importance and potential applications of DDL in CLC, this topic will be picked up in Chapter Nine when pedagogical implications based on this research are considered at length.

2.9 Some limitations of previous CLC researches

Though tremendous achievements have been made in the new, exciting and fast-developing field of research, there are some problems worthy of further discussion and investigation. This section discusses some limitations of previous CLC studies and relates them to the topics of

this thesis.

2.9.1 Lack of systematic study of lexis

Nesselhauf (2004: 135) maintains that "at the moment there is a wide variety of disconnected studies, usually concentrating on a few words or uses of words, and there are hardly any studies that look at a phenomenon in more depth"; and if CLC results are to be translated into pedagogical implications and applications directly, sporadic and non-systematic studies are inadequate. She stresses that more effort must be made to carry out systematic research in a particular area. It goes without saying that when there are enough areas which have been studied in detail the CLC investigation would yield meaningful results in pedagogical application. This thesis puts verbs in the centre of the theme and tries to examine their width and the depth. This is a first step to interpreting learner English if systematisation is to be achieved.

2.9.2 Lack of POS segmentation for multiple-POS words

Though POS annotation has been practised in corpus analysis as a whole, as far as I know, it is very rare for this to be carried out in learner corpora study. Since the English language contains many words which can be more than one part of speech, it is necessary to separate the verb use from the noun use, and to separate the verb use from other uses such as adjective and occasionally adverb. In a preliminary study (Guo 2003), I tried to look at the discrepancies between NSs and NNSs in their preferences with regard to verb use and noun use among 25 sets of verb and noun pairs such as *include* and *inclusion*. Even though it reveals the NSs' preference for nouns and the NNSs' preference to verbs, it is not known whether this trend exists in the vocabulary with more than one POS but the same morphology such as *charge*, *control* and *desire*. This thesis continues to explore in this direction, but with more detail.

2.9.3 Lack of semantic segmentisation for multiple-sensed words

It has been extensively reported that learner English is largely characterised by a smaller range of vocabulary as detailed in the previous part of this chapter. There have been very few

reports, however, on the use and distribution of different senses for multiple-sensed lexis (exceptions are Pu 2000a and Ringbom 1998a). It is harmful to limit investigations of learner vocabulary to the quantity or size of vocabulary acquired by learners, which hides the problem of whether a particular lexis has been mastered properly or not, especially the multi-sensed vocabulary. This kind of investigation may lead to pedagogical suggestions that learners should enlarge their vocabulary size if they wish to increase their English competence, which is indubitably correct. However, it ignores a major issue: whether learners should make full use of the word forms they seem to know already. Sinclair and Renouf (1988: 155), in an appraisal of lexical approach to language teaching, suggest the following:

> [T]he lexical syllabus does not encourage the piecemeal acquisition of a large vocabulary, especially initially. Instead, it concentrates on making full use of the words that the learner already has, at any particular stage. It teaches that there is far more general utility in the recombination of known elements than in the addition of less easily usable items.

It is essential for researchers to know how much of a word has become a part of the learners' English and how much is yet to be learned and used by the learners. This will have both theoretical significance (e.g. in understanding the process of vocabulary acquisition) and pedagogical implications (e.g. in curriculum design). This issue will be addressed in Chapter Seven and Chapter Eight when patterns (of *KEEP*) and collocates (of *TAKE*) are investigated in detail.

2.9.4 Lack of in-depth exploration in learner language feature identification

Due to the newness of CLC study, there is tremendous room for improvement. Apart from her appeal for more systematic studies, Nesselhauf (2004: 135) also stresses the importance of making more in-depth explorations as follows:

> [M]any, if not the majority, of learner corpus studies so far have concentrated on phenomena that can easily be studied automatically. Almost all studies look either at certain individual words, at continuous word sequences, or at other features that can be easily extracted from the corpus.

This thesis, along with trying to show how automatic functions of KWIC software (as in the identification of collocations) could be employed to a fuller extent, also attempts to explore how discontinuous word sequences could be worked out effectively and insightfully with the

aid of automation of the software (such as verb-related patterns in Chapter Seven). (For the facilities of the software that enable discontinuous word search, see Chapter Three, 3.3.1).

2.9.5 No linguistic standards to scale the level of learner English

In most of the studies, the criteria are not specified for what counts as "advanced level" or "intermediate level" or "novice level". Among the few studies in which the criteria are mentioned, the parameters are all external and bear no relation whatsoever to the internal linguistic proficiency. For example, in a learner corpus study, Cobb (2003) makes his judgement for what constitutes "advanced learners" by the fact that the NSs have passed the admission criteria of a TESL training programme at a university. Similarly, he attaches the label "intermediate level" to the test essays written by those applicants for ESL courses at the same institution. Actually, the lack of standards in this aspect has long pervaded the history of SLA, as these cases exist throughout the literature of SLA studies. Biskup (1992: 88) labelled some Polish and German students of English as "very advanced learners" because they "had received an average of ten years' instruction in English". The time spent on English study is certainly important in grading the current status of learner English, but it does not guarantee any improvement of the competence of learner English. Being instructed 20 hours a week is surely different from two hours a week. External parameters do not automatically validate the subjectivity in grading the levels of learner English. White (2002) (cited in Long 2003: 507) describes her subject, a Turkish woman as follows: "She is a fluent, 'advanced' speaker, as judged by her score of 93 percent on a University ELI placement test…" It can be easily observed from Table 2.1 that Hasselgren has very different standards from Granger and De Cock *et al*. A notable difference should exist between 'Norwegian sixth form students and first year university students of English' and third or fourth year university students of English.

My reservation on measuring learner language by external factors is also very well echoed by Tono (2003: 801):

> Selection based upon external criteria such as school year or age does not necessarily ensure that the subjects selected are comparable in terms of language proficiency. This happens to be the case for the Japanese-speaking EFL learner group. Although their learner profile fulfilled all the criteria, their proficiency levels are so markedly lower than those from other European countries that the inclusion of the Japanese data seems to skew the overall results.

Lorenz (1999: 10) also complains that the current learner language studies suffer from a lack of principles in giving criteria to 'advanced' learners. He believes that it is difficult to classify the levels of learner English based on linguistic grounds. Therefore, he has to accommodate the problem as follows while he addresses the issue of 'advanced learners' (*ibid*.):

> The present definition [of advanced learners] is therefore based on external factors and inductive reasoning: advanced learners are learners who have to meet advanced foreign language requirement, i.e. learners who *are generally* expected to have mastered the basic rules and regularities of the language they are learning.

No matter whether theoretically or empirically, there is a need to establish a relative norm so that when someone mentions "advanced learners", it will be explicitly understood as the same (or approximately the same) thing with the same parameters in the measurement of the learners' English. Table 2.1 shows the incompatibility in labelling the degree of of learners' attainment:

Table 2. 1 A sample of some studies which have no comparability between each other

Author	Level	Mother Tongue	Criteria
Granger (1998b)	Advanced	Various	University undergraduates in English Language and Literature in their third or fourth year
De Cock *et al.* (1998)	Advanced	French	Third and fourth year university students
Hasselgren (1994)	Advanced	Norwegian	Norwegian sixth form students and first year university students of English
Cobb (2003)	Advanced	Unspecified	Successful candidates to a TESL training programme at UQAM[13] with a writing task
Cobb (2003)	Intermediate	Unspecified	Successful applicants for ESL courses at UQAM
White (2002)	Advanced	Turkish	High score of 93 percent on a university placement text
Biber *et al*. (1998)	Intermediate or advanced	French, Spanish Chinese and Japanese	Non-specified

The criteria in Table 2.1 are various, ranging from the period of time spent on study to successful entry to a particular course. What they lack is uniformity. Without exception, they are all external criteria which provide no information of any linguistic parameters. This thesis does not try to seek a solution because it is not the purpose of this thesis to have this

13 The Université du Québec à Montréal.

complicated problem solved in passing. It would need a whole thesis to try to establish a prototype in the area of testing. My thesis only raises the awareness of the problem and tries to propose a perspective to a possible solution (see Chapter Nine).

2.9.6 Some reservations about the use of 'overuse' and 'underuse'

Though it is possible to examine a learner corpus without comparison (for example Li 2003), "the essence of work on learner corpora is comparison" (Hunston 2002: 206), as mentioned above. When an item is compared in two or more corpora, it is natural for there to be a discrepancy in frequency. It would be very unusual if the item under study were exactly the same. Inevitably, comparative corpus analysis will involve different frequencies between the corpora under study. If the item is used more by the learners, that is *overuse* by the learners. On the contrary, if this item is used less by the learners, that is *underuse* by the learners. As far as I can see, however, there exists a loophole in these two terms. Since CLC is a brand-new branch of study and there are so many appealing areas to investigate, it seems that attention has not been given to this issue. This thesis reveals this problem and proposes a more refined distinction between different types of overuse and underuse (see Chapter Seven for the detail).

2.9.7 Some reservations with error-tagging

One general impression from a review of CLC is that too much attention has been given to error-tagging, given that errors in learners' IL are only a small portion of the entire IL system. Even though CLC is widely annotated with learners' misuse of the TL, there are at least two questions to raise regarding the practice of error-tagging to a learner corpus. "What is an 'error' of learner English?" And "Can 'errors' be annotated properly?"

Essentially, it is almost impossible to answer the question: What is an 'error' of learner English? More often than not, what is judged as an 'error' by one person may sound acceptable to another. It is becoming extremely difficult to find a standard for so-called 'correct English' when English is becoming a lingua franca of the world, resulting in great difficulties in labelling what is correct use and what is incorrect use. NSs very often rely on

their intuition for the linguistic resources stored in their minds, but actually, human beings' intuition does not always work accurately and properly. Furthermore, NSs do not always share the same intuition between themselves and sometimes their intuition can be inaccurate and unreliable. When people resort to their intuitions they draw mainly on the aspects of language they have encountered throughout their lives. In fact the almost unlimited language resources beyond their vision may invalidate people's intuitions. Looking at a bit of language is different from looking at a lot of language (Hunston, 2002). (Also see 3.3.1 in Chapter Three for a more detailed discussion of this issue). When one is looking at a lot of language, people's intuitions toward the language get tested, clarified, improved, and sometimes even corrected. Until one's intuition is proved well grounded, one can never be certain about the validity of one's own intuitions. Let me illustrate this point with a few examples. The following two sentences produced by English learners are judged to be 'misuse' from the view of Gillard and Gadsby (1998:167).

*They live in a very lovely house near the sea.
*The cake was very delicious.

According to them, *lovely* and *delicious* are non-gradable adjectives and thus should not be modified by grading lexis such as *very*. However, if we open the BoE and type the cluster of "very+lovely" and "very+delicious", we obtain 116 and 16 cases for them respectively in the whole corpus. Two examples are:

1. He even admitted: "The house is very lovely. In this garden, one can … (Corpus usbooks/US)
2. But served with boiled potatoes and hollandaise, it's very delicious. (Corpus times/UK)

These two examples reflect the possible controversy as to what is acceptable and what is not even in the eyes of NSs. In fact, many adjectives which seem to be traditionally non-gradable can be found to be modified by the intensifier *very* in modern English. For example, *available* (9), *definite* (207), *right* (75), *wrong* (332), *true* (278). This shows a change in conventional standards and the emergence of new practices.

English changes over time like other languages. What was not accepted a couple of years ago may come into daily and active use of the language. A few years ago, it could be considered irrational if "email" were used in its plural form. But who would bother today if someone said "More than 1000 emails flooded into his internet website"[14]. To conclude, it is not at all easy to set a standard for 'correct English' and apply it in error-tagging. Without having a uniform standard towards what is right and what is wrong, it is questionable as to whether any claims on a study based on an 'error-tagged' learner corpus are dependable.

If the first question ("What is an 'error' of learner English?") is more theoretical, then the second question ("Can 'errors' be annotated properly?") is more practical. In an error-tagged learner corpus, researchers are to a great extent liable to be restricted by the error taxonomy. What can be observed and found out will be mostly (if not all) based on this taxonomy. Can this taxonomy be sorted out properly? Tono has the following summary (2003: 801):

> As shown in the history of error analysis, categorizing learner errors is a laborious and oftentimes fruitless job, for there are various ways of classifying errors, depending on research interest and theories involved and it is often the case that the classification is only as valid as the theory it is based on. Also, most people have different perspectives on error types, thus leading to very low inter-rater (or classifier) reliability.

To avoid falling again into the pitfalls of the "thorny issue" which arose in the 1970s, of attempting to look for an error taxonomy but without success, Tono warns researchers not to attempt to create a generic error taxonomy for all purposes. He maintains that research goals must be the first consideration in the assessment of the validity of error-tagging. This implies that to annotate a learner corpus without knowing the research goals of potential researchers is something like putting the cart before the horse.

Error-tagging can be very demanding due to the possibility that learners may produce any deviant form from the norm of NS English. Tono (2003: 804) observes: "There are often cases where there is insufficient evidence to assign one unambiguous interpretation of an error." "It should be noted," in Milton's words (1998: 188), "that the determination of error is not possible when the semantic or pragmatic intention of the writer is not clear or the syntax or lexis is so entangled that the most heroic measures cannot disambiguate meaning - especially

14 From BoE, Corpus/sunnow/uk, accessed on February 11, 2004.

common among the weakest writers." If the content to be tagged is not understandable to the human taggers, how can we expect computer taggers to make sound judgments?

Furthermore, there is something else that makes error-tagging unreliable. In the production of learner English, there exist some "mistakes" and "errors", according to Corder (1967, reprinted in Richards 1974). While the former refers to the performance inadequacy caused by "slips of the tongue (or pen)" or other chance circumstances (*ibid*., 24), the latter is reserved for the deficiency of the learner's "underlying knowledge of the language to date" (*ibid*., 25). The problem with error-tagging is that human taggers can hardly work out which misuse is a 'mistake' and which misuse is an 'error'. If this distinction cannot be made, how can we know exactly what is the portion already acquired and what is the portion still to be acquired? Embarrassingly, this is a problem error-tagging not only cannot resolve but may actually hide from researchers (cf. Gui and Yang 2002: 2).

Another pair of concepts in learner English deserving fine distinction according to Corder (1971) is 'overtly idiosyncratic' and 'covertly idiosyncratic'. The former refers to the feature of ill-formedness in terms of the rules of the target language and the latter stands for the sentences that are perfect in form but erroneous in context. Most probably, error-tagging will be too much attracted to the 'overtly idiosyncratic' type and meanwhile the other is entirely ignored.

As a conclusion, there is no widely accepted standard in treating so-called 'errors' and it is hardly possible to annotate the deviant features of learner English properly. Considering these problems, it is my belief that the practice of error-tagging should be re-evaluated. But it must be pointed out here that it is not the case that I resist the practice of error-tagging but propose to exert extra caution while this is done. It can be envisaged that error-tagging will be improved and used widely in the long run. This thesis, however, due to the unsolved problems and potential loopholes with error-tagging, will not use the error-tagged version of COLEC (see Chapter Three for details of the data). Instead, the raw version will be used as the basis for investigation and in some of the chapters (mainly Chapter Four), the data will be POS-tagged for specific purposes.

2.10 Conclusion

Great achievements have been made since the start of the era of CLC study in the late 1980s and early 1990s. The compilation and analysis of CLC experienced a pioneer stage when researchers sought better ways of dealing with the data in the new enterprise. Even though there are 'competing methods' (Hunston 2002: 92) in the field of corpus linguistics as to how to access the data, their research outcome is equally persuasive and promising. They have observed some of the most salient features of learner English. They have attempted to annotate learner corpora either manually or with the assistance of annotation tools. However, it is still too early to be positive about the way we are currently dealing with the learner English data and with the benefit SLA researchers wish to obtain from CLC study. For example, how reasonable are the theory and practice of error-tagging in relation to a learner corpus when the definition of 'error' can hardly be made? How much insight can group interlanguage study shed on the assessment and evaluation of individual learner's interlanguage? Why do learners stick to a limited range of options when there are plenty of alternatives for them to choose? For this young academic domain of CLC, it seems that there are more questions yet to answer than questions already answered. There are more myths to explore than any feat to be proud of. As a whole, learner English study is a fast developing domain of study double-edged with promise and challenge.

While CIA researchers may be excited by the idea that "we are on the verge of a learner corpus boom" (Granger 1998a: xxii), it seems that it is not an easy task to paint a picture of learner corpus rosy enough to attract the eyes of our neighbouring researchers, at present filled with doubt and mistrust (Leech 1998). This thesis attempts to add my colour to the large picture.

Chapter Three

The Data and the Tools

3.1 Introduction

This chapter will introduce the learner corpus and the reference NS corpus which are to be compared and analysed in this thesis. The rationale for using a reference corpus in learner English study will be discussed and it will be explained why this particular reference corpus has been chosen. The issue of comparability between the two corpora will be addressed. Finally, the tools to be used for this research will be introduced.

3.2 The data

3.2.1 The Learner Corpus – COLEC

The COLEC corpus contains about half a million words and the greater part of this corpus was selected from university students' compositions in the nation-wide English examinations called College English Test (Band 4) and College English Test (Band 6) (shortened to Band 4 and Band 6 henceforward). Students first attend Band 4 before they proceed to Band 6 some time later (normally one year later). These two tests have been conducted regularly at certain times every year in China for some years. The titles of the compositions which were collected for COLEC involve social issues such as "The Shortage of Fresh Water", "The Harmfulness of Fake Commodities", "Health Gains in the Developing Countries", campus-related issues such as "Getting to Know the World outside the Campus", and job-related issues such as "My View on Job-hopping", "My Ideal Job", and daily life topics such as "Practice Makes Perfect" and "Haste Makes Waste". There are 1500 essays chosen from both Band 4 and Band 6. According to the marking scheme, 15 is the full score for both Band 4 and Band 6 compositions.[15] To add a supplementary source to the corpus, 1000 essays of free writing

15 At the pilot study stage, it was found that the texts below the score of 6 were of little value for inclusion to the corpus because they were fragmentary sentences and were too short for the minimum requirement of words.

were collected from several universities. The free writing essays were graded by scores using College English Test marking criteria. My research will treat the two bands of essays as a homogeneous unit of learner language. There are several reasons why I should have done so. The first reason is that these two bands of essays are homogeneous in the aspects of cultural background, education level, and learning objectives, as Li (1999) acknowledges. The second reason is that although there is some improvement in the degree of English from Band 4 to Band 6 such as total lemmatised types, standard type–token ratio, and a lower error percentage, a considerable homogeneity exists between these two groups according to the analysis by Gui and Yang (2002: 52). What is more important is that difficult points in Band 4 will basically remain the same in Band 6, as is demonstrated in Wen *et al*.'s quasi-longitudinal research (Wen *et al*. 2003).

Since all of the learner texts were handwritten they had to be retyped, which means typing errors would be almost unavoidable, and a lot of proofreading was required all the way from the beginning of the keyboarding to the end of error-tagging. In the future students can be asked to submit their writings in digital form to avoid such hard labour.

The corpus was error-tagged and each essay is marked with non-linguistic information such as the score of the essay, the gender and the university code of the writer and the test band of the essay, 4 or 6. For the convenience of research, this corpus was made in two versions, the error-tagged version and the raw version. As discussed in Chapter two, due to my reservations on the practice of error-tagging, only the raw version is used in this thesis.

The COLEC corpus was made available on CD for corpus study as early as 1998. A considerable amount of comparative studies based on COLEC have been conducted ever since. Lu (2002) conducted her PhD research on learner English in Singapore using this corpus. Several papers presented to the International Conference on Corpus Linguistics in Shanghai also used COLEC.[16]

Therefore, a decision was made that only papers with a score of 6 and above would be selected and all those below would be discarded.

16 Researchers who wish to conduct learner language studies may choose to use the updated and expanded version of learner English, which is called CLEC (see Gui and Yang 2002 for details).

3.2.2 The Native Speaker Corpus - LOCNESS

NNS corpora alone will not suffice if we wish to trace features of learner language which deviate from those of NSs. They will show up only in contrast with a reference corpus (also called control corpus). When there are plenty of NS corpora available today, which one to opt for becomes the next question. In the centre of this question lies a thorny issue, i.e. comparability. Whose language production should be considered the norm, the experts' performances or the performances of native learners of a comparable age? Even though there are comparative studies between learner corpora and adult expert corpora (for example Yang 2001, Gui and Yang 2002), "Optimally, we also need *targeted* corpora – corpora targeted to represent as closely as possible the learner's future communicative needs," as Leech (2001: 333) insightfully suggests. Obviously we would not expect learners to learn a very general use of English, disregarding the special needs of learners in written argumentative English production. Apart from this, there are other reasons for us to take on board while we consider the need for a comparable reference corpus (or 'targeted corpus' by Leech). Kaszubski (1998a) argues strongly for such a need in CIA:

> Corpus-based error analysis is ideally based on maximum comparability of corpora: the more variables can be controlled, the more dependable results are supported. … mere text type congruity does not always warrant a sufficient degree of comparability. Since learner language remains greatly influenced by extralinguistic developmental factors, the age and experience of contributors whose output is analysed are also very significant variables. … it is psycholinguistically more appropriate to compare EFL learner corpora not with ideal "expert performances" in the target language but with attainable performance of native learners of a comparable age.[17]

Based upon such a necessity, I have chosen LOCNESS as the reference corpus. There are two reasons for having made this selection. One is the considerable comparability between COLEC and LOCNESS which will be detailed below. The other is that LOCNESS is the NS corpus most commonly used for comparison so far. For some examples, Ringbom (1998b) compared seven western European learner corpora with LOCNESS. Lorenz (1998) used

17 This paragraph was taken from the TaLC 1998 website http://users.ox.ac.uk/~talc98/kaszubski.htm but unfortunately, it was removed later. In order to track it back, the archive website of the internet can be used by opening the website at http://web.archive.org and enter the old URL at the prompt.

LOCNESS in a comparison between the NNS English of German learners and that of NSs. Virtanen (1988) studied direct questions in argumentative student writing by comparing several NNS subcorpora of ICLE and LOCNESS. It was in Granger and Rayson's comparative research between ICLE and LOCNESS that the stylistic features of spoken rather than written English were displayed. Aarts and Granger (1998) discovered some distinctive interlanguage patterns of tag sequences of some NNSs by comparing the argumentative essays written by Dutch, Finnish and French-speaking advanced learners and the essays extracted from LOCNESS. Aijmer (2002) examined modality by comparing advanced Swedish learners' written interlanguage and the English in LOCNESS, although together with the English from other sources. Lin (2002) discussed the overuse and underuse of *it* in the writing of Chinese learners of English against the reference corpus LOCNESS. Undoubtedly, LOCNESS has gained a solid reputation in serving as a reference corpus in the domain of learner English study.

LOCNESS was built by the Centre for English Corpus Linguistics at the Catholic University of Louvain, Belgium and made available for public use in 1998. The texts of the corpus are essays produced by British and American native speakers from 1991 to 1995. The corpus is composed of four components, i.e., essays of British A-Level students, essays of British university students, argumentative essays of American students and literary-mixed essays of American students. The texts of the corpus include examination papers, timed essays and free essays. No reference tools were used in examination papers whereas in some timed essays and free essays reference tools were used. The length of essays is around 500 words. The age of students is mostly between 17 and 23 although there are a very small number of students who are much older. Although the NS profile for the essays of British A-Level students and of British university students is not available, it can be assumed that most of the students are NSs of English. The texts cover a very wide range of topics from social problems such as water pollution, nuclear power, sex, violence, and gender roles to campus-related issues such as cheating in college, controversy in the classroom, and prayer in schools. Both parts of the corpus, the British essays and the American essays, have country-specific topics. For example, in British essays, a large amount of texts talk about the parliamentary system, foxhunting, the

national lottery, and BSE[18] and British beef. In contrast to this, the major interest in American essays is found in quite different areas such as the Confederate Flag, the US government, book banning in America, gun control and the legalisation of marijuana. From the topics covered by the texts, the overall feature of the writing style can be interpreted as argumentative. The whole corpus is not tagged but is coded with information on the essay titles. Based upon the features of COLEC and LOCNESS, a table has been drawn up to give a description of the comparability between the two corpora (see Table 3.1).

Table 3. 1 Comparison of some parameters of COLEC and LOCNESS (Comp = Comparability)

Parameter	COLEC	LOCNESS	Comp
Essay type	Exam papers and non-exam papers	Exams, timed essays and free essays	HIGH
Size[19]	480063	322464	AVERAGE
Use of reference tools	Some yes	Some yes	AVERAGE
Length of each essay (tokens)	200	500	LOW
Age of students	16-24	Mostly 17-23	HIGH
Topics	Shortage of fresh water, fake commodities, job-hunting, views on how to get to know the world, etc.	Water pollution, nuclear power, gender roles, violence, sex, drugs, parliament, freedom and religion, etc.	LOW
Genre	Mainly expository and descriptive	Mainly argumentative	LOW
Authoritativeness of the compilers	Professionals in Linguistics, testing and TEFL	Professional in computer learner corpus	HIGH
Time of completion	1998	1998	HIGH

Both similarities and differences exist in the two corpora, implying comparability *and* incomparability. On the one hand, for example, the two corpora were completed in the same year, 1998, representing comparability in the time of production of the data. Both of the two

18 BSE is the acronym for bovine spongiform encephalopathy. A more commonly used name is mad cow disease.

19 COLEC contains a large quantity of codings such as the writer's ID, gender, age, and the essay's title. This will cause a problem of inaccurate word count. What is more serious, it will be difficult to see the true presentation of the text written by students due to the existence of a considerable amount of lexical words including verbs in essay titles. To avoid these problems, all the words in the diamond brackets including the brackets themselves in COLEC have been deleted (by typing \<*\> in the Find and Substitute order of MS Word). Although LOCNESS contains much less coding of diamond brackets, to make the two corpora more comparable, the same kind of deletion was done to LOCNESS.

corpora were directed by professionals in teaching and language corpora. Both groups of students are from almost exactly the same age-group (between 16 and 24), which again increases the comparability. On the other hand, however, individual essays of LOCNESS are much longer than those of COLEC. There are about 500 words in an essay in LOCNESS whereas there are only 200 words in an essay in COLEC. This may affect the comparison to some extent. Most COLEC essays are expository or descriptive whereas LOCNESS writings are mainly argumentative. This will also make a difference in the vocabulary used because different genres or text types will involve a different lexis. While LOCNESS students are writing about their western way of life (such as water pollution, nuclear power, gender roles, violence, sex, drugs, government, parliament, freedom and religion), COLEC students are talking about something different which plays a role in their life (such as global shortage of fresh water, the harmfulness of fake commodities, the ways to get to know society, and one's view on job-hunting). A disparity will definitely emerge here due to the difference in the topics (for a detailed discussion of the influence of a topic on the mode of language, cf. Tarone and Yule 1989). It will be impossible to find a perfect control corpus which is similar in every aspect. The existence of different cultures alone will make it hard to achieve such a goal. It should be borne in mind that a reference corpus should serve only as a tool of reference for comparison in general. Pu (2000b), holds the belief that ideally a reference corpus should not only have the same size but also the same topics, and even have respondents from a similar background (for example, university students). Sometimes, in the real world, however, we have to reach a compromise between what is desirable and what is available. Granger (1998b: 13) acknowledged this important problem thus:

> Criticisms can be levelled against most control corpora. Each has its limitations and the important thing is to be aware of them and make an informed choice based on the type of investigation to be carried out.

Considering that a fairly large degree of similarity exists in the two corpora, it is feasible to carry out a comparison between them, especially when the reference corpus LOCNESS is treated as a presumed norm.

3.2.3 The back-up resources

3.2.3.1 The Bank of English

Normally, commonly used words or structures can be found in the control corpus, LOCNESS. If a popular word or structure is found in the learners' corpus, it would also be found in the control corpus. If something exists in the learners' corpus but not in the NS corpus, there might be two explanations. One possibility is that the use in the learners' corpus is correct but that it is not found in the reference corpus because it is too small or because the topics of the corpus would not allow such a use to happen. The other possibility is that the use in the learners' corpus is incorrect and therefore there is no match in the NS corpus. Intuition has a role to play in making a judgement as to whether a situation belongs to the first possibility or the second, but it will not work all the time. For the sake of safety, I choose to use a much larger NS corpus, the Bank of English (BoE) as a backup corpus. A detailed introduction to the BoE will not be attempted here because online information with regard to this well-known corpus is abundant. [20] What follows is only a brief introduction to some issues relevant to the current research.

The BoE is a collection of samples of modern English language after 1990 and is currently maintained by the University of Birmingham. It contains both written and spoken English from hundreds of different sources. Written texts mainly include newspapers, magazines, fiction and non-fiction books, brochures, leaflets, reports and letters whereas the spoken part comprises transcriptions of everyday casual conversation, radio broadcasts, meetings, interviews and discussions. The material is up-to-date, as the majority of texts have originated since 1990. It can be used for the analysis of words, meanings, grammar and usage.[21] Since it is a monitor corpus which tracks language change, it is increasingly expanding in size (see Sinclair 1991: 9, 24-26 and Hunston 2002: 16, 30-31 for a detailed discussion of monitor corpora). By January, 2001, it reached the figure of 450 million words. It has benefited a number of areas of academic research, for example, Sinclair, Jones and Daley in the 1970s (see Krishnamurthy 2004); Renouf and Sinclair (1991); Sinclair (1991); Hunston and Francis (1998); Moon (1998); Hunston and Francis (1999) and also of language pedagogy, for

20 See the homepage of the BoE at http://www.titania.bham.ac.uk/docs/about.htm.

21 http://www.titania.bham.ac.uk/docs/about.htm, accessed on January 1, 2006.

example, Sinclair and Renouf (1988); Willis (1990); Johns (1991).

Undoubtedly, the BoE is a useful tool for language description. Its potential as a back-up tool in CIA has not, however, been properly explored. This thesis, while undertaking a comparative analysis, shows in passing how the BoE could be used wisely to assist with this seemingly irrelevant work, CIA.

3.2.3.2 The Google search engine

The BoE as a backup tool in this research meets most confirmatory requirements. However, there are cases where even more data is desirable. Occasionally, I will turn to Google which takes advantage of the Web as a reservoir of English data. In comparing corpora in the traditional sense and the web as corpus, Fletcher (2004: 275) outlines the advantages of the latter thus:

> A static corpus represents a snapshot of issues and language use known when it was compiled. The great expense of setting up a large corpus precludes frequent supplementation or replacement, and contemporary content can grow stale quickly. On contrast, new documents appear on the Web daily, so up-to-date content and usage tend to be well represented online. In addition, even a very large corpus might include few examples of infrequent expressions or constructions that can be found in abundance on the web. Moreover, certain content domains or text genres may be underrepresented in an existing corpus or even missing entirely. With the Web as a source one usually can locate documents from which to compile an ad-hoc corpus to meet the specific needs of groups of investigators, translators or learners. Finally, while existing corpora may entail significant fees and require specialized hardware and software to consult, Web access is generally inexpensive, and desktop computers to perform the necessary processing are now within the reach of students as well as researchers.

Of course, we need to be fully aware of the drawbacks of using the web as corpus. In the words of Fletcher (2004: 275), "The quantity of information online greatly surpasses its overall quality." In this research I intend to use the web only as another back-up (in cases where the BoE is not big enough to provide a support or where a special need comes up). In other words, I only need to see whether a particular expression is used online and if it is used at all, how often it is used, rather than to see how it behaves as related to the co-text.

3.3 The WordSmithTools

This section introduces in some detail the software used in this research, i.e. the WordSmith Tools (henceforward, WordSmith) (Scott 1999), which is probably the most widely used software for a general purpose of KWIC retrieval and corpus linguistic analysis.[22] There are three major functions in WordSmith (Version 3.0), i.e. Concord, the concordancer, WordList, which produces word lists in a number of ways, and Keyword which yields key words in a file and key key words in a number of files (see Scott 1997 for a detailed discussion of key words and key key words). Since this research will use only Concord and WordList, the following sections will briefly introduce their functions and some important concepts involved in this research.

3.3.1. Concord

In order to introduce Concord properly, it is helpful to introduce some of the most often used terms in corpus linguistics in general, *concordance, the node word* and *concordancer*. "A concordance is a collection of the occurrences of a word-form, each in its own textual environment" (Sinclair 1991: 32). A concordancer is a program used to search the specified data for all the instances of a word or phrase selected by the user and then present them in the middle of the computer screen (Hunston 2002: 39). The selected word is known as *the node word*. Concordances help us to arrive at several aims which are not easy to achieve intuitively. According to Hunston (2002: 42ff), concordances help us to observe the "central and typical" behaviour of a language, meaning distinctions and details of language use. Scott (1999: 13) believes that: "It is through changing the shape of data, reducing it and then re-casting it in a different format, that the human capacity for noticing patterns comes to the fore… Human beings are good at noticing, and particularly good at noticing visual patterns". In a corpus, the language is displayed vertically whereas in conventional texts it is read horizontally (Bonelli 2001 & 2004). Looking at language by concordances is different from looking at language by texts (Sinclair 1991; Hunston 2002 and others). How to benefit most from examining concordance lines is one of the major topics of this thesis, especially Chapter Nine; thus it is

22 See Mike Scott's homepage for the use of the software in research publications (accessed on January 5, 2006): http://www.lexically.net/wordsmith/corpus_linguistics_links/papers_using_wordsmith.htm.

adequate here to quote Scott's insight into the pedagogical use of reading concordances.

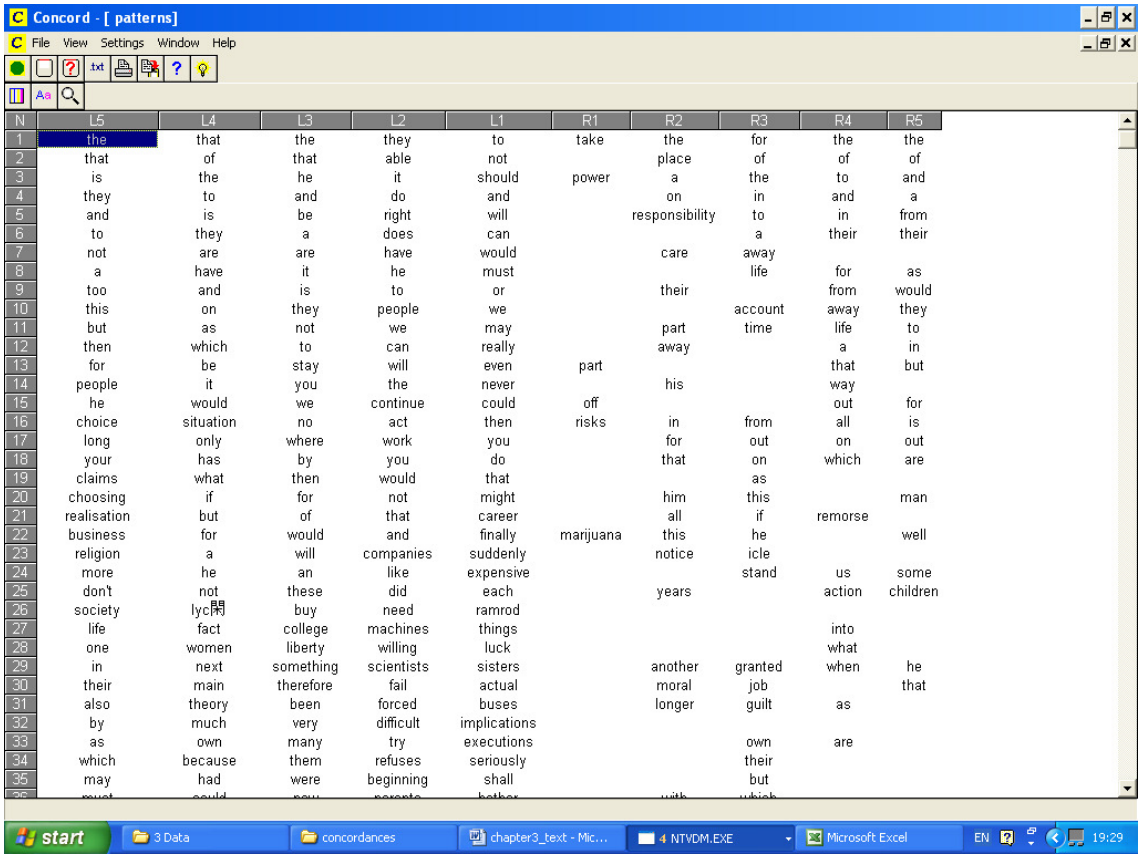
> Language students can use a concordancer to find out how to use a word or phrase, or to find out which other words belong with a word they want to use. For example, it's through using a concordancer that you could find out that in academic writing, a *paper* can *describe*, *claim*, or *show*, though it doesn't *believe* or *want* (**this paper wants to prove that*…).

> Language teachers can use the concordancer to find similar patterns so as to help their students. They can also use Concord to help produce vocabulary exercises, by choosing two or three search-words, blanking them out, then printing. (Scott 1999: 55).

To read concordance lines as they appear in the default setting does not always meet the user's research purpose. Concord provides the possibility of re-sorting to the concordance lines. Switching from one re-sorting to another enables the program to reveal the frequently occurring features more thoroughly. Researchers build up their skills while they conduct their searches. Experience has a role to play in the selection of the right re-sorting type. This issue will be discussed in the research chapters in more detail. Concord not only yields single word searches but also *clusters*, and *patterns*. Scott (1999: 81) uses *clusters* to mean "the words which are found repeatedly in each other's company". It seems to me that the term *cluster* resembles fixed phrases, or multiple units physically, but they are not necessarily identical. For example, *take care of* is both a cluster and a phrase but *care of the* is only a cluster. Since computers do not distinguish clusters from phrases, it is the task of the human beings to identify meaningful information from the clusters. Another important issue that needs to be pointed out here is that there is a distinction between the *pattern* in technical sense, i.e. in Scott's term (Scott 1999), and the *pattern* in syntactic sense, i.e. in Hunston and Francis's term (1999). The pattern used by Scott is the general sense which refers to the frequency relationship between the node word and its environment, either on the left or on the right, either immediate to or a few positions away from the node word. It allows the most frequent items in the specified neighbourhood of the search word to "float up" to the top (Scott 1999: 68) (see Figure 3.1 for an example).

Since a whole chapter will address the issue of patterns in Hunston and Francis's sense (see Chapter Seven for details), to avoid confusion, this thesis will stick to their concept of patterns.

Figure 3. 1 A screenshot of the pattern of *take* (from LOCNESS) by WordSmith

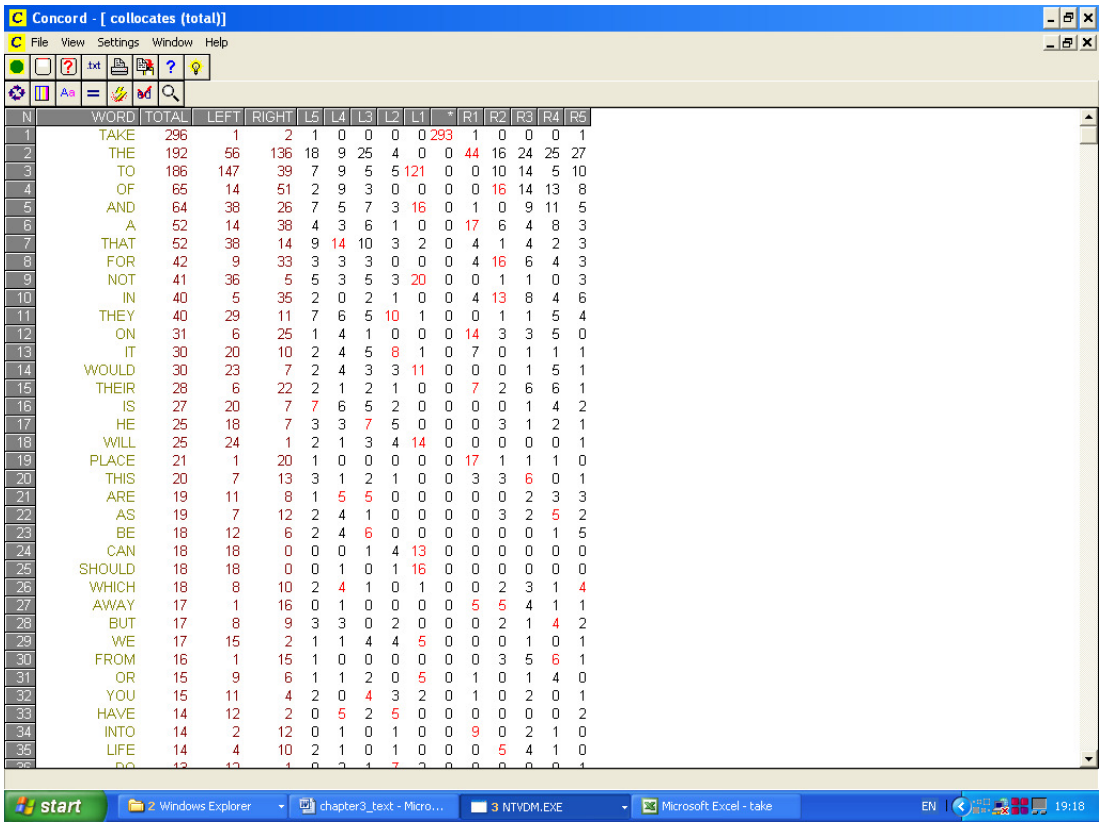


N	L5	L4	L3	L2	L1	R1	R2	R3	R4	R5
1	the	that	the	they	to	take	the	for	the	the
2	that	of	that	able	not		place	of	of	of
3	is	the	he	it	should	power	a	the	to	and
4	they	to	and	do	and		on	in	and	a
5	and	is	be	right	will		responsibility	to	in	from
6	to	they	a	does	can			a	their	their
7	not	are	are	have	would		care	away		
8	a	have	it	he	must			life	for	as
9	too	and	is	to	or		their		from	would
10	this	on	they	people	we			account	away	they
11	but	as	not	we	may		part	time	life	to
12	then	which	to	can	really		away		a	in
13	for	be	stay	will	even	part			that	but
14	people	it	you	the	never		his		way	
15	he	would	we	continue	could	off			out	for
16	choice	situation	no	act	then	risks	in	from	all	is
17	long	only	where	work	you		for	out	on	out
18	your	has	by	you	do		that	on	which	are
19	claims	what	then	would	that			as		
20	choosing	if	for	not	might		him	this		man
21	realisation	but	of	that	career		all	if	remorse	
22	business	for	would	and	finally	marijuana	this	he		well
23	religion	a	will	companies	suddenly		notice	icle		
24	more	he	an	like	expensive			stand	us	some
25	don't	not	these	did	each		years		action	children
26	society	lych	buy	need	ramrod					
27	life	fact	college	machines	things				into	
28	one	women	liberty	willing	luck				what	
29	in	next	something	scientists	sisters		another	granted	when	he
30	their	main	therefore	fail	actual		moral	job		that
31	also	theory	been	forced	buses		longer	guilt	as	
32	by	much	very	difficult	implications					
33	as	own	many	try	executions			own	are	
34	which	because	them	refuses	seriously			their		
35	may	had	were	beginning	shall			but		

Apart from *clusters* and *patterns*, another type of search in Concord is the search for *collocates*. "A word which occurs in close proximity to a word under investigation is called a *collocate* of it" and "Collocation is the occurrence of two or more words within a short space of each other in a text", according to Sinclair (1991: 170). One central function of Concord is to find collocates of a given word (or phrase). By examining its collocates, "You shall know a word by the company it keeps" (Firth 1957). Scott (1999: 57) exemplifies the notion of collocates by saying that collocates of *letter* might include *post*, *stamp*, and *envelope*. "Collocates can be counted and this measurement is called the *span*" and "A span of -4, +4 means that four words on either side of the node word will be taken to its relevant verbal environment", in the words of Sinclair (1991: 175). The notions of *the node word*, *span*, *collocate* should always be interpreted in relation to each other. Take the word form *take* for example; when it is searched in Concord, it becomes *the node word*, when the collocate search is carried out, a certain number of *collocates* are produced including both grammatical words such as *the*, *to* and *of* and content words such as *place* which predominantly occur immediately to the right of the node word. This position could be expressed in a *span* of "+1"

or in Scott's term "R1". The most frequent collocate will be signalled in red (the less invisible ones in this black only printing) (see Figure 3.2). The amount of collocates Concord can produce is dependent on the value settings, such as span and minimum frequency of the word, in the specified neighbourhood of the search word. The default value of the span (*horizon* in Scott's term) for collocates is set at five on both sides, i.e. L5, and R5. This basically confirms the view held by Sinclair, Jones, and Daley 1970, cited in Sinclair (1991: 106) that "beyond four words from the node there were no statistical indications of the attractive power of the node". Very recently, after a re-calculation of a much larger corpus, Sinclair has added that "five words to the left and four words to the right might result in a slightly stronger improvement of semantic relevance" (Krishnamurthy 2004: xix). His explanation is that "it seems that the patterning in general is a little stronger on the left than on the right" (*ibid.*). Even though the default value of collocates is set at L5 and R5, users may adjust the horizon to as far apart as L25 and R25, which is able to meet the requirement of exceptionally specific queries. (See Figure 3.2 for a brief idea of how the collocates of *take* look like as produced by WordSmith).

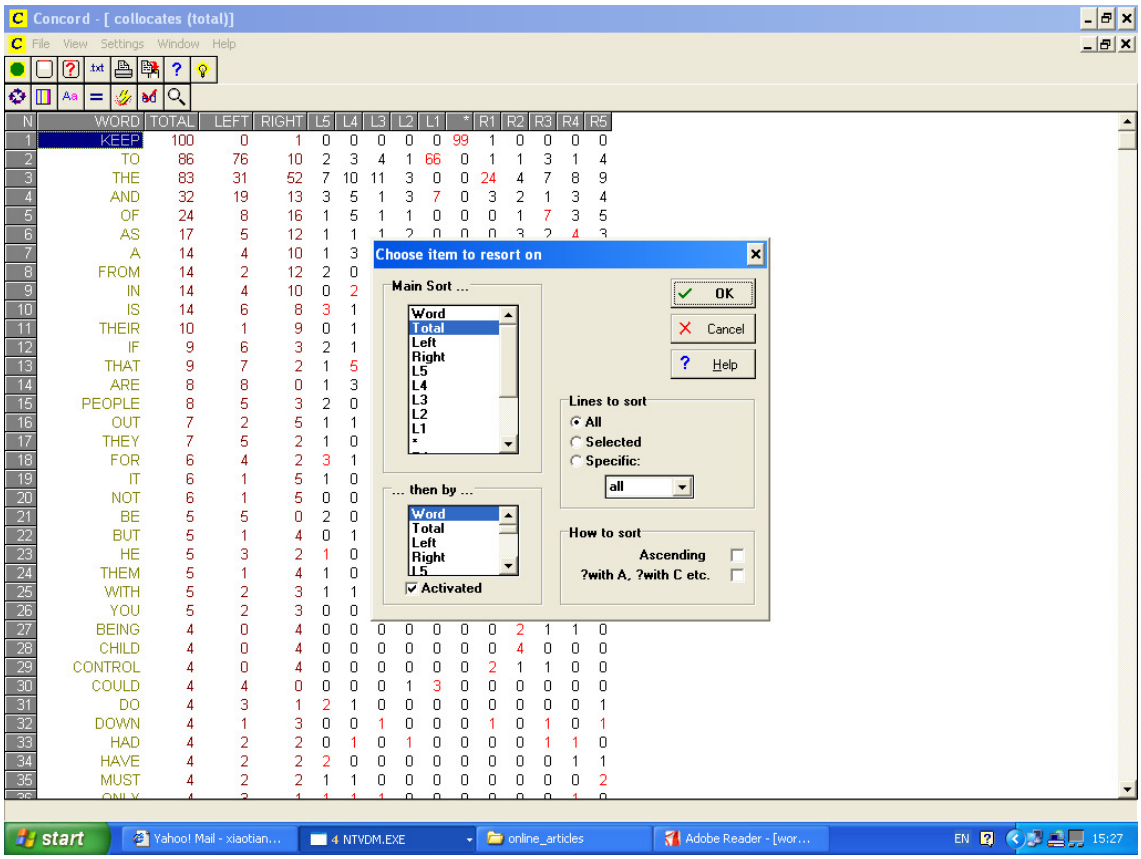
Figure 3. 2 A screenshot of the collocates of *take* (from LOCNESS) by WordSmith



N	WORD	TOTAL	LEFT	RIGHT	L5	L4	L3	L2	L1	*	R1	R2	R3	R4	R5
1	TAKE	296	1	2	1	0	0	0	0	293	1	0	0	0	1
2	THE	192	56	136	18	9	25	4	0	0	44	16	24	25	27
3	TO	186	147	39	7	9	5	5	121	0	0	10	14	5	10
4	OF	65	14	51	2	9	3	0	0	0	0	16	14	13	8
5	AND	64	38	26	7	5	7	3	16	0	1	0	9	11	5
6	A	52	14	38	4	3	6	1	0	0	17	6	4	8	3
7	THAT	52	38	14	9	14	10	3	2	0	4	1	4	2	3
8	FOR	42	9	33	3	3	3	0	0	0	4	16	6	4	3
9	NOT	41	36	5	5	3	5	3	20	0	0	1	1	0	3
10	IN	40	5	35	2	0	2	1	0	0	4	13	8	4	6
11	THEY	40	29	11	7	6	5	10	1	0	0	1	1	5	4
12	ON	31	6	25	1	4	1	0	0	0	14	3	3	5	0
13	IT	30	20	10	2	4	5	8	1	0	7	0	1	1	1
14	WOULD	30	23	7	2	4	3	3	11	0	0	0	0	1	5	1
15	THEIR	28	6	22	2	1	2	1	0	0	7	2	6	6	1
16	IS	27	20	7	7	6	5	2	0	0	0	0	1	4	2
17	HE	25	18	7	3	3	7	5	0	0	0	3	1	2	1
18	WILL	25	24	1	2	1	3	4	14	0	0	0	0	0	1
19	PLACE	21	1	20	1	0	0	0	0	0	17	1	1	1	0
20	THIS	20	7	13	3	1	2	1	0	0	3	3	6	0	1
21	ARE	19	11	8	1	5	5	0	0	0	0	0	2	3	3
22	AS	19	7	12	2	4	1	0	0	0	0	3	2	5	2
23	BE	18	12	6	2	4	6	0	0	0	0	0	0	1	5
24	CAN	18	18	0	0	0	1	4	13	0	0	0	0	0	0
25	SHOULD	18	18	0	0	1	0	1	16	0	0	0	0	0	0
26	WHICH	18	8	10	2	4	1	0	1	0	0	2	3	1	4
27	AWAY	17	1	16	0	1	0	0	0	0	5	5	4	1	1
28	BUT	17	8	9	3	3	0	2	0	0	0	2	1	4	2
29	WE	17	15	2	1	1	4	4	5	0	0	0	1	0	1
30	FROM	16	1	15	1	0	0	0	0	0	0	3	5	6	1
31	OR	15	9	6	1	1	2	0	5	0	1	0	1	4	0
32	YOU	15	11	4	2	0	4	3	2	0	1	0	2	0	1
33	HAVE	14	12	2	0	5	2	5	0	0	0	0	0	0	2
34	INTO	14	2	12	0	1	0	1	0	0	9	0	2	1	0
35	LIFE	14	4	10	2	1	0	1	0	0	0	5	4	1	0

Another important value the user can customise is the minimum frequency for a collocate to appear. Since this thesis aims at examining the most salient features of learner English as contrasted against NS English, it does not make sense to focus on those infrequent cases which can hardly represent the English of the whole group. Thus a proper value should be set which considers both representativeness and the size of the corpora. If this value is set too low, the program produces too many collocates, causing too much noise; and if the value is set too high, the corpora may not be large enough for the program to produce enough significant collocates. Sorting collocates, as Figure 3.2 shows, is only one way (the default setting) of doing this. There are actually several different ways of re-sorting the collocates. It helps the user to see clearly the outstanding collocates in different positions. See Figure 3.3 for a screenshot of the task box for setting the values for re-sorting collocates.

Figure 3. 3 A screenshot of value setting for collocate re-sorting



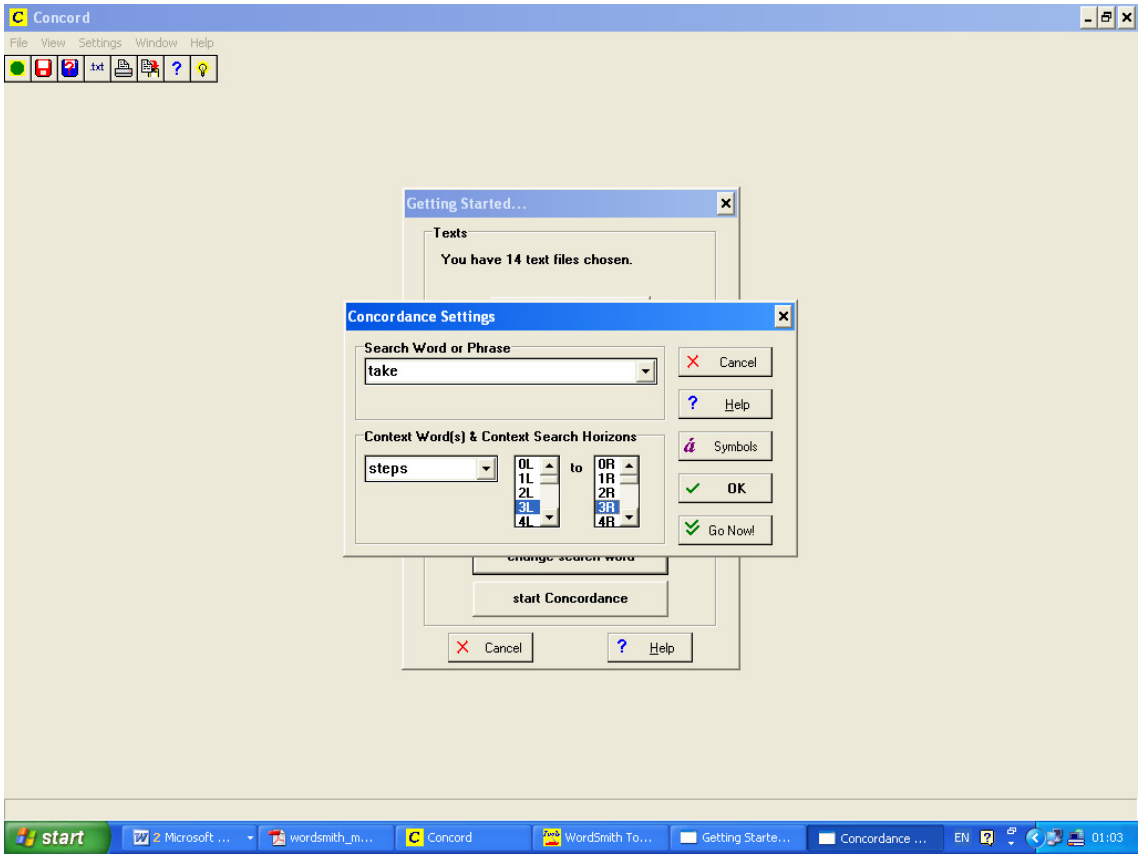
In the same way as *pattern* search, *collocate* search helps the user to see the most salient lexical attractions to the node. But unlike *pattern* search, *collocate* search provides in detail the number of occurrences of the collocates which appear in different positions whereas *pattern* search puts the most often occurring word in a certain position (compare Figure 3.1

and Figure 3.2 to see the disparity). Switching between *pattern* search and *collocate* search helps Concord to display which words appear, how often and in what positions.

Words do not co-occur randomly. NSs can resort to their intuition to judge how likely it is that one word will co-occur with another. But very often it is far from adequate (Sinclair 1991; Hunston 2002 and many others).

Every KWIC software has its own unique search queries. The BoE, for example, has its own Lookup (Sinclair 1987) while the British National Corpus has its Sara (Aston and Burnard 1998) and then Xara (Burnard and Dodd 2003). Concord has its own query language which Scott calls Search Word Syntax (Scott 1999: 60-61). This search word syntax allows accurate and case-sensitive search as well as wild-card advanced searches. This is important for my research because it helps to identify how complicated or how advanced NS English can be. For example, if I need to examine the verb *SURPRISE*, the wild card will enable complicated searches such as for all the lemma forms including *surprise*, *surprises*, *surprised* and *surprising*.

Figure 3. 4 A screenshot of the Concordance Settings box of WordSmith



In the Concordance Setting box, as shown in Figure 3.4, where the node word or phrase is entered, an advance search facility is provided, i.e. the specification of context words and context search horizons. This is extremely important for my research because it makes many complicated searches possible. In Chapter Eight, for example, while the node word *take* is examined, it is possible to see how many times *steps* appears on the right side of the node word and how many times it occurs on the left side of the node word. It helps me to reach a conclusion whether learners tend to use *steps* on the left or on the right of the node word, which has great potential pedagogical significance.

Before I move on to WordSmith's WordList, there is one more important thing to introduce, i.e. the user-defined categories in concordance line examination. This function is useful if the researcher needs to categorise the concordance lines according to a certain need, for example, the POS. Raw corpora do not have POS information; if the researcher wishes to distinguish the verb use from the noun use of multiple POS words, this is the right thing to do. As many as 52 categories can be given with English letters (both lower case and upper case).[23] Self-defining categorisation will be used frequently in this research.

3.3.2 WordList

WordList is another important component of WordSmith for producing word lists in both alphabetical and frequency order.[24] It can be used for the following purposes (Scott 1999: 84):

1. simply in order to study the types of vocabulary used;

2. to identify common word clusters;

3. to compare the frequency of a word in different text files or across genres;

4. to compare the frequencies of cognate words or translation equivalents between different languages.

---

23 Version (4.0) has more possibilities.

24 WordList indexes can also be made, to examine for stylistic or comparative purposes.

Closely associated to the generation of word list are the concepts of *type* and *token*, and then *the type/token ratio*. A *token* is a running word whereas a *type* is any distinct word in the text (Krishnamurthy 2004: 34). A text of 100 words long contains 100 *tokens* but much fewer *types* because some words are repeated, such as articles and prepositions.

To produce a lemma list by WordList is another important function of WordSmith. This research will use WordList (together with programming other software) to produce a special lemma list: a verb lemma list for both COLEC and LOCNESS. Since it requires a full-length description to illustrate the whole process, I will leave this issue to be discussed in Chapter Four.

3.4 Conclusion

This chapter has introduced the corpora to be examined in this research, the learner corpus COLEC and the NS corpus LOCNESS. The issue of comparability is addressed in detail. This comparative analysis presupposes that the standard reached by the LOCNESS writers is treated as the norm of the COLEC writers (as also mentioned in Chapter One), and aims at pinning down the distance between the COLEC writers (as a group) and the LOCNESS writers (also as a group) mainly in the area of verbs. This chapter has also introduced the tools to be utilised in this research, WordSmith, the BoE and Google. I will treat the BoE and Google only as back-up tools rather than as reference corpora.

Chapter Four

Making and Making Sense of Two Verb Lemma Lists

4.1 Introduction

Making or using a word list is not a new concept, either in native English corpora studies (for example, Sinclair 1987, Scott 1999) or in learner corpora studies (for example, Gui and Yang 2002; van Rooy and Schafer 2003). But making or using a *verb* lemma list in the studies of NS corpora is much less reported (two exceptions are Kilgarriff 1997 and Leech *et al*. 2001), not to mention the studies of learner corpora. As a matter of fact, to gain access to the knowledge of learners' general lexicon in verbs is not only of great value for the learner language researcher but also for the English language teacher, and learners themselves. Knowing how many verbs are used in a learner corpus and how many verbs are used in a NS corpus means knowing the distance between the learner English and the NS English. And knowing the distance of the learners from the target means knowing the learners' needs in vocabulary growth which is essential for interlanguage research into group learners and indeed for ELT. Attracted by the research and pedagogical significance, I have drawn up two verb lemma lists from COLEC and LOCNESS. Since two randomly arranged verb lemma lists have limited value for the researcher and the teacher, I am going to arrange the verb lemmas effectively so that they may be more illuminating. This chapter reports on the progress of making the two verb lemma lists and proposes a way of grouping the verbs which makes sense not only to learner language researchers but also ELT practitioners and learners. The research questions of this chapter are set out as follows:

1) What is the range of verbs used in COLEC and what is the range of verbs used in LOCNESS?
2) What is the similarity and disparity between the COLEC writers and the LOCNESS writers as far as verbs are concerned?
3) How many verbs are used only in LOCNESS and what are they?
4) How could the research findings based on the previous three questions be used for the improvement of ELT?

4.2 Some issues in making a verb lemma list

4.2.1 The significance of making a verb lemma list

Learner language researchers have long been talking about improving the vocabulary size of learners. However, crucial questions about the actual vocabulary size of a group of learners seem not have been worked out. How similar is the learner language to the target norm? And how deviant is the learner English from the target norm? As I am going to show in this chapter these questions may now have answers. On the pedagogical side of the picture, it is possible for ELT practitioners to see very accurately how many verbs are produced by the learners as a group, and what these verbs are. When a verb lemma list for NSs is produced and the learner lemma list is compared with it, it is possible for teachers to detect which verbs are used only by the NSs, which verbs are used only by the learners, and also which verbs are shared by both groups in terms of verb types. All in all, when these questions have proper answers, learners' needs in vocabulary expansion become accessible to the teacher and other ELT practitioners including the writers of teaching materials, the syllabus writers, the evaluators and others. Fuzzy speculation can now give way to the accurate identification of the features of learner English under investigation when a learner corpus is compared with a NS corpus.

4.2.2 Some notions

Some key concepts will be involved in the process of making verb lemma lists. The first key notion is called *lemma* and some of its associated terms are *word-form* and *lemmatisation*. The word *lemma* is not easy to define because it can be used either narrowly as "a set of lexical forms having the same stem and belonging to the same major word class, differing only in inflection and/or spelling" (as defined by Francis and Kucera (2004), cited in Knowles and Zuraidah 2004: 70), or broadly as a covering term for all the lexical forms under a given dictionary entry. For example, as Sinclair exemplified (Sinclair 1991: 173), the word-forms *give*, *gives*, *giving*, *gave* and *given* will be lemmatised into the lemma *GIVE*. For another example, the word forms *eat*, *eats*, *eating*, *ate* and *eaten* belong to the lemma *EAT* (Hunston 2002: 17-18). This thesis uses the narrow sense of the notion to refer to the inflectional forms of the same POS only. So instead of referring to all the forms that cut across POS boundaries

(such as *believe, believes, believing, believed, belief, beliefs, believable, believably*), when I mention the lemma *BELIEVE*, I only mean the inflectional forms within the same POS category verb, i.e. *believe, believes, believing,* and *believed*. For the sake of convenience, different spellings (British and American) of the same word, such as *ORGANISE* (including *organise*, *organises*, *organising* and *organised*) and *ORGANIZE* (including *organize*, *organizes*, *organizing*, and *organized*) will be treated as two separate lemmas.

4.2.3 The difficulties in making a verb lemma list

The first problem in making a verb lemma list lies in the unavailability of a verb lemma base or template for a corpus to match. One of the few available lemma lists (not verb lemma list) was produced by Yasumasa Someya (details will be introduced in 4.3.1). A problem with this list, however, is that verb function and noun function are not separated within multiple-POS words. If this lemma list is to be used, nouns have to be removed from the list first. This may sound easy but it is not as straightforward as expected, since the lemma list does not contain the crucial information needed such as POS -tagging or demarcation marks between verbs and nouns. To cross nouns out manually is obviously one option, but there seems to be a better solution. Section 4.3.1 below will describe how the nouns are removed from the whole lemma list. A second problem is that Someya's lemma list most of the time provides only American spellings. Since LOCNESS contains a large number of essays by British students (see 3.2.2 in Chapter 3 for the composition of LOCNESS), and also, the COLEC writers use both American and English spellings, the words with corresponding British spellings must be added to the list if it is to be used as the base of a reference lemma list (see 4.3.1 below for details). A third problem is that both COLEC and LOCNESS were originally not POS-tagged. Without annotation, it is impossible to tell whether a multiple-POS word such as *change* is being used as a verb or a noun. By the same token, without annotation it is not possible to identify whether a word ending with "s" is the third person singular form of a verb or the plural form of a noun (as in *supports*). Therefore, to make such distinctions, the two corpora need to be POS-tagged. Details of how the corpora are tagged are provided in 4.3.2 below.

4.2.4 Two approaches to making a verb list

There are two possible approaches that have presented themselves in the process of making my verb lemma lists from COLEC and LOCNESS. One is to work out all the word forms that are used as verbs from among the corpora. This objective is achievable now because a corpus can be annotated by POS as introduced in 2.3.3 in Chapter Two. A KWIC search into a POS-tagged corpus (either by WordSmith or a home-made program) will be able to produce all the word forms used as verbs. When all these word forms used as verbs are produced as a complete list, it is possible to make a verb lemma template that includes all the forms of verbs. However, the problem that immediately arises from this perspective is that it is hard to make such a verb lemma list template that contains all the forms used by learners because there are unexpected and incorrect forms. For example, some COLEC writers use *solute* as a verb (the result of misuse for *solve*). This is hardly predictable for the most knowledgeable designers of a verb lemma list unless they are informed by the real learners' production data. Therefore, it is difficult, if not impossible, to work out a lemma list that will cover both the correct and incorrect forms of all the verbs used by learners.

The second approach is to make a basic verb list first and then match the POS-tagged learner corpus against the verb list, ignoring the incorrect uses of verbs and assuming that the correctly used forms form the majority of the learner English. This approach is not without problems because there are situations where the writer uses word-formation rules such as affixes (such as *re-*, *de-*, *co-*, *un-* etc) to meet the special needs in the context. Another disadvantage with the second approach is that the category of verbs of English is not always unchangeably set. For example, conversions from noun to verb are abundant in NS language use (see Davies 2004 for a detailed discussion about noun to verb conversion in English). To recognise such conversions and tag them correctly is a challenge to POS-taggers. Any lemma list which aims at a high accuracy will have to be exhaustive enough to predict all these complex situations and include everything that appears in actual language use.

It seems that no matter which approach we take it is impossible to achieve perfection. Since the aim of this thesis is to measure the distance in language use between a group of learners and a group of NSs, and it is less important to know how the learners use verbs incorrectly

than to know how far from the target their use of verbs (the types of verb lemmas) is, the second approach seems to be less affected by the limitations of the current POS-tagging technologies. Therefore, it is the second approach that I took in producing a verb lemma list out of COLEC and LOCNESS respectively.

4.3 Making two verb lemma lists

This section describes the process of making two verb lemma lists. The following section explains the process of making two verb lemma lists, a list from COLEC and a list from LOCNESS, which mainly concerns three stages, i.e. making a reference lemma list, tagging the raw corpora, and lemmatising the word forms.[25]

4.3.1 The lemma list archetype

The e-lemma list compiled by Yasumasa Someya in 1998 is probably the most exhaustive lemma list at the time of writing[26]. It contains 40,569 words (tokens) and 14,762 lemma groups. The following is a sample of the list:

accept -> accepts, accepting, accepted
acceptance -> acceptances
acknowledge -> acknowledges, acknowledging, acknowledged
acknowledgement -> acknowledgements
know -> knows, knowing, knew, known
organize -> organizes, organizing, organized
organization -> organizations

Someya's lemma list contains not only verbs like *acknowledge* and *know* but also nouns with singular and plural forms like *acceptance* and *acceptances*. There is no doubt that this lemma list is a useful tool for anybody who wishes to conduct lemmatisation. Though not perfect, researchers may find it useful in using it as a base for their own lemmatisation. Of course,

25 Some of the points in this section may not be exactly the same as in my approach because this description is based on recollection.

26 Yasumasa Someya's lemma list is available on the homepage of Dr. Mike Scott:
http://www.lexically.net/downloads/version4/downloading%20BNC.htm.

they may find it necessary to make corresponding modifications according to their special needs. For example, the LOCNESS corpus contains writings by both British students and American students; therefore, there exist two systems of spelling such as *organise* vs. *organize*, *realise* vs. *realize*, *favour* vs. *favor*. Someya's list, however, in most cases, provides only the American spelling and does not include the British spelling variants such as *organise*, *realise*. In other cases, it provides only the British spelling such as *favour* but misses out the American spelling *favor*. For the lemma list to cover the data as extensively as possible, appropriate manual modifications would need to be carried out. Apart from the above problems, some other minor problems should be solved before the lemma list is put into use. In dealing with the lemma *MEET*, for instance, Someya lists it as two lemmas, as follows:

meet -> meets,met
meeting -> meetings

This is a controversial arrangement because if this word list is used as a base for lemmatisation the word-form 'meeting' will not be included in the verb lemma *MEET* along with 'meets' and 'met'. In order to avoid the problem the lemma *MEET* will here be rearranged as follows:

meet -> meets, meeting, met

Since POS-tagging is expected to solve the distinction between noun use and verb use (see 4.3.2 for details), there is no need to worry about the possibility that the use of 'meeting' as a noun will also be calculated in the verb lemma *MEET*.

Another problem that prevents Someya's lemma list from direct use is that it contains lemmas of nouns and other parts of speech such as indefinite articles (*a* and *an*), and adjectives (such as *big*, *bigger* and *biggest*). Since the research aim is to make verb lemma lists, and other POS words may become a noise in the process of lemmatisation, a decision was made to detect all non-verb lemmas. It is certainly possible to conduct a complete manual deletion, but it would be very time-consuming because lemmas of different POSes are mixed alphabetically. To save time, I have chosen to use MS Excel to replace the greater part of the hard manual labour. After a series of edition to Someya's lemma list, it is ready for verb lemma processing (for a

description of the process of how the non-verb lemmas were deleted and a verb lemma base worked out, see Appendix 1).

4.3.2 Tagging the corpora

POS-tagging the corpora is the second important part of making the verb lists. The POS-tagging of the two corpora was conducted by using CLAWS7, which allows verbs to be differentiated by different forms, i.e the base form (including the finite form and the infinitive form), the third person singular form, the V-ing form, the past form, and the past participle form and also allows multiple POS words to be differentiated by different tags. The following are some examples to show how different forms are tagged differently:

Figure 4. 1 Different forms of *TAKE* tagged by CLAWS7

```
1       _RR over_II the_AT world_NN1 finally_RR take_VV0 a_AT1 stand_NN1 for_IF what_DDQ
2       _NN1 ;_; something_PN1 that_CST I_PPIS1 take_VV0 for_IF granted_VVN ,_, yet_RR o

3       VVG societies_NN2 one_PN1 had_VHD to_TO take_VVI account_NN1 of_IO the_AT hidden
4       .  </s> <s> They_PPHS2 decide_VV0 to_TO take_VVI action_NN1 to_TO improve_VVI th

5       om_II Michigan_NP1 to_II California_NP1 takes_VVZ about_II 30-35_MCMC hours_NNT2
6       n_II a_AT1 fellow_JJ human_JJ being_NN1 takes_VVZ a_AT1 lot_NN1 of_IO courage_NN

7       ofessors_NN2 when_CS they_PPHS2 are_VBR taking_VVG advanced_JJ courses_NN2 that_
8       CSN it_PPH1 is_VBZ an_AT1 act_NN1 of_IO taking_VVG advantage_NN1 of_IO an_AT1 op

9       </s> <s> Many_DA2 black_JJ students_NN2 took_VVD advantage_NN1 of_IO the_AT whit
10      t_NN1 ,_, but_CCB whereas_CS Christ_NP1 took_VVD away_RL men_NN2 's_GE sins_NN2

11      is_NP1 10_MC )_) In_II a_AT1 survey_NN1 taken_VVN across_II the_AT USA_NP1 ,_, d
12      S21 if_CS22 people_NN are_VBR being_VBG taken_VVN advantage_NN1 of_IO by_II Kevo
```

As shown in the previous concordances (Figure 4.1), the different forms of the verb *TAKE* are tagged differently (VV0 for the finite form, VVI for the infinitive form, VVZ for the third person singular form, VVG for the -ing form, VVD for the past form and VVN for the past participle form).

Not only are the different forms of a verb distinguishable, but the POS distinction is also realised in the POS-tagging. As expected, the 'hunt' in the following sentence is tagged as a noun (NN1):

<s>The_AT fox_NN1 **hunt_NN1** is_VBZ a_AT1 lengthy_JJ and_CC extreemly_RR cruel_JJ process_NN1 ._. </s>

and tagged as verb in the following sentence:

<s>In_II the_AT modern_JJ world_NN1 there_EX is_VBZ no_AT need_NN1 to_TO **hunt_VVI** in_BCL21 order_BCL22 to_TO obtain_VVI our_APPGE food_NN1 ._. </s>

For another example, 'fixed' is tagged as adjective (JJ) in the following use

they_PPHS2 can_VM sell_VVI them_PPHO2 to_II the_AT EU_NP1 at_II a_AT1 **fixed_JJ** price_NN1 ._. </s>

and tagged as verb in its past participle form (VVN) in the following use:

there_EX is_VBZ a_AT1 problem_NN1 that_CST must_VM be_VBI **fixed_VVN** ._.

Table 4. 1 A sample of the verb list from LOCNESS

	Lemma	V-e	V-s	V-ing	Ved	V-n	Total
1	make	426	113	129	88	231	987
2	take	289	76	111	59	132	667
3	see	306	48	27	35	219	635
4	use	198	52	96	27	190	563
5	become	209	69	75	60	86	499
6	say	178	110	68	76	61	493
7	give	164	51	61	40	137	453
8	go	201	91	79	34	37	442
9	feel	280	70	13	57	13	433
10	want	215	105	16	71	19	426

The third stage involves lemmatisation. After POS annotation, it is possible to separate nouns and other POS words from verbs. This makes it possible to focus on verbs and to calculate the frequencies of all the forms of a verb. In this way, the important information with regard to the frequency of a lemma (rather than of its individual forms), and then the frequencies of all the verb forms is available to the researcher. Table 4.1 is a sample of some verb lemmas (the first 10 most often-used lemmas in LOCNESS) with their individual forms separated.[27] The frequencies of the individual forms and the lemmas are provided in the table as well. The

27 I am grateful to Richard Xiao for having helped me with the POS-tagging of the corpora and Scott Piao for having written a program to arrange the verb lemma lists with all the forms of a verb in one row.

word forms are expressed by V-e for the base form, V-s for the third person singular form, V-ing for the ing form, V-ed for the past form and V-n for the past participle.

At the end of this section, it should be noted that there exists a problem of accuracy rate of the POS tagging. According to the report of the CLAWS7, the accuracy rate of the word-class tagger is between 95% and 98% depending on types of text. [28] I have to admit that such a rate of accuracy has not been checked in this research. What is more, the rate of accuracy in the learner corpus should be much lower than that of the NS corpus simply because the learner corpus has unexpected uses and the tagging system (presumably designed for NS English data) is not expected to work as well on learner corpora. It is certainly true that the more accurate the POS tagging is, the more confident we are with the research result. However, CLC researchers would have to reach a compromise between what is desirable and what is available.

4.3.3 Editing the raw verb lemma lists

Even though my intention is to make verb lemma lists, some verbs are not taken into consideration because they are problematic in one way or another for the production of lemma lists. What is more important is that they do not contribute significantly to my research question: how many verbs are used and what are they in each of the corpora? These lemmas include auxiliaries such as *DO*, *HAVE*, *CAN*, *MAY*, *WILL, DARE,* etc. Even though the lemma *GO* is in the lists, the catenative use of *going* as in *be going to* has been counted separately and is not included in the lists. The use of *going* could be viewed as the non-catenative use as in "There is wide debate *going* on about […]". The base form and the infinitive form are not distinguished in making the verb lemma lists because the purpose of making the verb lists is to see the range of the verb lemmas which are used by the NSs and NNS and how they compare. Researchers who have such an interest could separate them because CLAWS7 is able to produce different tags for them by labelling the base form (as in *he works hard*) as VV0 and the infinitive (as in *to give* and it *will work*) as VVI.

In the raw verb lemma lists, there are altogether 758 lemmas identified in COLEC and 1238

28 http://www.comp.lancs.ac.uk/computing/users/eiamjw/claws/claws7.html, accessed October 8, 2006.

lemmas identified in LOCNESS. These two figures are too pointlessly exact and can hardly be taken seriously for immediate application, because some lemmas are very infrequent and others are mostly used as nouns rather than verbs. The lemma lists would stand a better chance of pedagogical application if these problems could be sorted out first. It has therefore been decided that the verb lemma lists should be trimmed.

4.3.3.1 Dealing with small-frequency lemmas

Verb lemma frequencies vary from verb to verb. Some verb lemmas have high frequencies such as *TAKE* and *KEEP* while some verbs have very few occurrences in the whole corpus. There are many hapax legomena: items that occur only once. The fewer occurrences there are for a verb lemma, the less confident we are in making a judgement about whether this verb as a lemma has become a part of the vocabulary of a group of writers as a whole. For the infrequent words such as *indulge* (2), *dwell* (2), *spray* (1), and *thrill* (1) in COLEC, my teaching experience and intuition about Chinese university students does not support the speculation that these words are a fairly representative performance of the whole group. Therefore, it would be irrational to see these infrequent words as part of the learned and mastered vocabulary of the group. By the same token, the low-frequency lemmas in LOCNESS could also be a result of occasional need and are not a feature shared by a large number of the group. Based upon this understanding, all the lemmas with total frequencies below 3 times (inclusive) have been deleted from the lists.

4.3.3.2 Detecting wrongly used lemmas

My intuition as a NNS from a similar background to the learners and as a teacher of English for many years helps me to identify some verb lemmas misused by the learners. For example, there are some cases of the lemma *CREASE*; in this case, my intuition as a learner of English myself suggests to me that this this is probably a misuse, and a check of the concordances shows that all the cases are in fact misuses for *INCREASE*. In another example, the English verb *SERVE* (7) is misused for two variants, one being *SERVICE* and the other being *SEVER*.[29] Another lemma misused in the same manner is *LEAN* for *LEARN*[30]. These lemmas

29 Such information could be used for exercises such as multiple choices.

have been removed from the verb lemma list of COLEC.

After the processing as explained above, the verb lemmas in the two lists are reduced to 569 in COLEC and 893 in LOCNESS. The detailed verb lemma lists are presented in Appendix 2 (COLEC) and Appendix 3 (LOCNESS). Such trimmed verb lemma lists give me extra confidence in conducting the following interpretation and analysis.

4.4 Making sense of the two verb lemma lists

A random verb lemma list without a certain level of categorisation can be said to have very little value for research and pedagogy. This section discusses how grouping and aligning verb lemma lists from certain aspects could help us discover similarity and disparity in the two groups of learners. Apart from revealing the similarity and disparity, I will also try to explore how such a grouping and aligning of verb lemmas could benefit language pedagogy.

4.4.1 A rational study

4.4.1.1 Some explorations in semantic theory applications in vocabulary teaching

There have been sizeable studies exploring how a word can be best displayed in relation to its semantically related associates since the 1980s (Channell 1981 & 1988, Godman 1982, Harvey 1983 and Stieglitz 1983, to name only a few). Joanna Channell and Arthur Godman are two researchers who have explored extensively in this field.

As a result of the traditional grammar teaching in the English classroom, learners become grammatically strong but lexically weak. In a valuable study about how to apply semantic theory to vocabulary teaching, Channell (1981: 115-116) states that learners make errors because they do not possess a native-like selection in using the right vocabulary. On the disparity between a native speaker and a learner in judging the acceptability of a sentence, she comments as follows:

---

30 Information gained from making verb lemma lists could well be used for language learning and teaching. (This issue will be fully discussed in Chapter Nine.)

> The native speaker is in possession of [all the information needed to speak his native language correctly] and he can use it to judge the acceptability of any sentence. For him, the subtle distinctions between *an attractive girl/a pretty girl/a beautiful girl/a good-looking girl/a nice girl* are things he makes use of in his everyday conversation without giving them a second thought. These distinctions are, however, the despair of any foreign learner unless there exists a systematic way of representing them, and therefore of being able to teach them.

In order to find this 'systematic way', Channell (1981: 116) proposes two aspects of semantic theory: 'semantic field theory' and 'componential analysis'. Semantic field refers to the many interrelating networks of relations between words (*ibid*.). One word in a semantic field such as 'stroll' in the semantic field *walk, run, stroll, amble, trot, job* may also be grouped into another semantic field consisting of *wander, stroll, roam, ramble* (*ibid*.: 117). Channell stresses that: "It is in this sense that *the vocabulary of a language should be seen as a set of interrelating networks*." (*ibid*., italics added). With regard to *componential analysis*, Channell (*ibid*.: 117-118) expounds as follows:

> Words can be said to belong to the same semantic field when they share some aspects of meaning. At the same time they hardly ever share all aspects. For example *walk* and *run* are similar in both being verbs describing ways in which animate beings with legs move, yet they differ in that *run* implies a different, usually faster, movement of the legs than *walk*. Componential analysis offers a systematic way of describing such similarities and differences. It consists, simply, of breaking down the meaning of a word or words into different pieces known as semantic components.

By using the semantic theory, Channell was able to describe meaning systematically in the materials for vocabulary teaching. The following paragraphs are her descriptions in detail (ibid.: 118):

> Imagine a text describing very unexpected events, in which the learner meets *astound* and *flabbergast*. He may look them up in his dictionary, and find definitions using a word he does know – *surprise*. However, as van Buren has pointed out, "the possibilities of misuse and misunderstanding" arising from definitions "seem endless". Neither definitions nor citations will give the learner much help with the two questions he needs to answer if these words are to enter his active vocabulary – 1 how do they relate to other words with similar meaning? and 2 which other words can they be used with, and in which contexts? This is where diagrammatic representation using [semantic] field theory and componential analysis can help.

In order to show how these two words could be better understood, Channell uses componential analysis to break down the meaning of each word into different grids (see Figure 4.2).

Channell (*ibid*.: 119) claims that by making explicit the differences and disparities between *astound* and *flabbergast*, and between these two words and the others (*surprise, astonish, amaze*), the learner is provided with the exact information he needs to know for correct interpretation and production.

Figure 4. 2 Channell's componential analysis of *SURPRISE*, *ASTONISH*, *AMAZE*, *ASTOUND*, and *FLABBERGAST* (Channel 1981: 119)

	affect with wonder	because unexpected	because difficult to believe	so as to cause confusion	so as to leave one helpless to act or think
surprise	+	+			
astonish	+		+		
amaze	+			+	
astound	+				+
flabbergast	+				+

Figure 4. 3 A table of three sense-related verbs based on Appendix 1, Godman (1982: 47)

Break Group	Divide Group	Cut Group
break	detach	amputate
burst	disconnect	carve
chip	disengage	chop
crack	disentangle	cleave
crumble	disperse	clip
disintegrate	dissipate	crop
fracture	dissociate	cut
rip	divorce	dissect
shatter	insulate	excise
smash	isolate	hack
snap	loose	hew
splinter	part	incise
split	scatter	lop
tear	segregate	mince
	separate	prune
	spread	sever
	uncouple	share
	unlock	shear
	unravel	shred
	untie	slash
		slice
		slit
		snip
		split
		trim

Largely in agreement with Channell, Godman (1982: 39) also maintains that vocabulary

should be presented in groups so that distinctions between each other appear clearly to students. In a study of verbs, he first challenges the traditional use of alphabetical dictionary entries as follows (Godman 1982: 39-40):

> A look at the usual entries for verbs in an alphabetical dictionary indicates that the entry is rarely full enough to give the L2 speaker confidence in his ability to use it correctly in all possible contexts. The exemplification of the term is usually more helpful in elucidating the correct definition, but that needs conscious effort on the part of the reader and it may not always be successful. A hierarchical system, using full definitions, can overcome the L2 speaker's difficulties […]

Figure 4. 4 A sense cluster map of the verb *BREAK* by Godman (1982: 47)

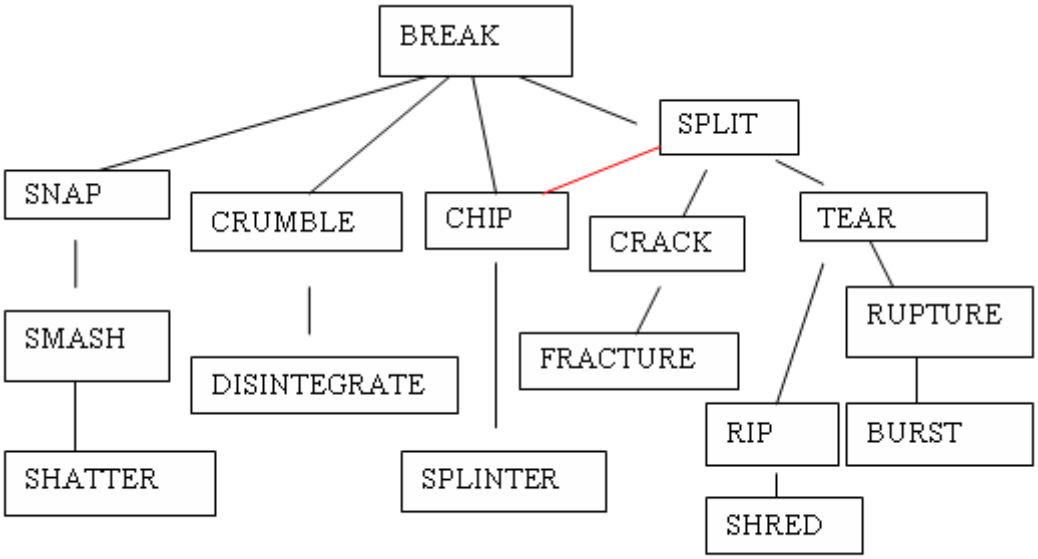


Figure 4. 5 A semantic field chart of the group headed by *BREAK* by Godman (1982: 49)

<table>
<tr><td colspan="2"></td><td colspan="3">Solids</td><td colspan="2">Surfaces</td><td>Coverings/ Thin Materials</td></tr>
<tr><td colspan="2" rowspan="2">Pieces Formed</td><td colspan="5">BREAK</td><td rowspan="2"></td></tr>
<tr><td></td><td colspan="2">SNAP</td><td colspan="2" rowspan="3"></td></tr>
<tr><td colspan="2" rowspan="2">Many pieces Formed</td><td rowspan="2">SHATTER
SMASH</td><td colspan="2">DISINTEGRATE</td><td rowspan="2">SHRED</td></tr>
<tr><td></td><td>CRUMBLE</td></tr>
<tr><td colspan="2">Small Pieces Separate</td><td></td><td>CHIP
SPLINTER</td><td></td><td colspan="2"></td><td></td></tr>
<tr><td rowspan="3">Thin Lines of Division Formed</td><td>Pieces Separate</td><td colspan="4">SPLIT</td><td rowspan="2">TEAR</td><td>RIP</td></tr>
<tr><td rowspan="2">No Separate Pieces</td><td rowspan="2">FRACTURE</td><td rowspan="2">CRACK</td><td rowspan="2"></td><td>RUPTURE</td></tr>
<tr><td>BURST</td><td></td><td>SPLIT</td><td>RIP</td></tr>
<tr><td colspan="2"></td><td>Hard Blow</td><td>Blow</td><td>Pressure</td><td>Blow</td><td>Tension</td><td></td></tr>
<tr><td colspan="2"></td><td colspan="6">Force</td></tr>
</table>

To construct such a hierarchy, Godman first collected a group of verbs sharing some element of meaning. The second task was to divide the verbs in this group into clusters which have a greater degree of elements in common (see Figure 4.3 for some examples).

Godman only studied the first 250 words in order of frequency of occurrence in the text he chose.[31] He arranged these words into 22 groups. There are two approaches that Godman employed in making the relations between group members explicit to learners. The first approach is a mono-dimensional one (see Figure 4.3) in which all the members of one group can be listed together alphabetically without showing the distinguishing features of each other. A more complex approach involves the use of two 'diagrammatic representations' (as Godman calls them 1982: 41). One is a common diagrammatic representation (see Figure 4.4) and the other is a multi-dimensional diagram (see Figure 4.5).

By analysing the elements of each verb as in these approaches, especially the multi-dimensional diagram in Figure 4.5, Godman clearly shows the semantic position of each verb and the relationship between one verb and another. The following is his own explanation:

> This term, BREAK, can be used in place of all the other verbs, with the exception of SHRED and RIP, by adding a suitable adverbial phrase of manner. The cluster can generally be represented by a phrasal verb which describes the common element of meaning. The clusters are hierarchically arranged in a group with the group displaying the common concept of a force changing a whole solid, surface, or thin material, into smaller pieces of forming divisions of such objects, which could eventually develop into complete separation.

Even though Channell and Godman use different ways to present the relationship between words in a semantic field, they are in agreement with each other in the belief that a word can be displayed more explicitly in the semantic field set to which it belongs. Channell (1981: 117) emphasises: "By analysing vocabulary into fields, we are no longer dealing with random lists, but with a systematic structure, and one which can be practically passed on to learners." She therefore advocates that "we should teach foreign-language vocabulary in semantic sets" (*ibid.*). Holding the same view, Godman states(1982: 46):

> The accurate meaning of a verb is fully comprehended only when it is placed in a set of verbs of

31 Unfortunately, he did not mention the size of the text.

> similar meaning, and the differences noted of its contextual features. Aids to achieving full comprehension include a hierarchical system of sets of verbs arranged in clusters, groups and families, and the arrangement of each member of the group.

Godman (*ibid*.: 45) firmly believes that treating verbs as demonstrated above should lead to a rapid vocabulary expansion. Channell has also applied the semantic theories to vocabulary teaching with her colleagues and met considerable success (see the two popular textbooks *The Words You Need* (Rudzka *et al.* 1981) and *More Words You Need* (Rudzka *et al*. 1985). (For more reports on the positive effect of the application of semantic field theories in ELT, see Channell 1981, Harvey 1983 and Stieglitz 1983.)

4.4.1.2 Some pioneering work concerning the presentation of vocabulary to learners

Apart from the modern researchers and ELT practitioners who hold the view that vocabulary should be presented in an arranged ways (rather than randomly), some ELT pioneers started this practice in their teachings long ago, even without the support of semantic theories. Far back in 1923, Horace Wyatt (1923, reprinted in Smith 2003), in trying to help his young Indian students to remember new vocabulary, proposed a practical way of grouping the vocabulary. Drawing upon his rich experience in teaching, Wyatt reinforced the notion that memory can be efficiently assisted "by treating the new vocabulary in groups which have a common topic or connecting bond" (*ibid*.: 35). Apart from grouping words (such as *walk* and *run*) that share a certain parameter (bodily movements), Wyatt (*ibid*.: 36) also stressed the importance of associating near-synonyms and contrasting antonyms:

> A further principle of association of use for introducing and practising new vocabulary is association by similarity and contrast. Words of the same, and words of opposite or rather of contrasted meaning, can be taught together, *black* at the same time as *white*, *high* and *low*, *big* and *small*, *short* and *tall*, *above* and *below*, *to cry* and *to laugh*, *to eat* and *to drink*, *to love* and *to hate*, *often* and *seldom*, *always* and *never*, and so on. Synonyms may also be taught together, or here again words of mainly similar meaning; for few words have exact synonyms (italics added to keep consistency with the format of the thesis).

Having reviewed the importance of applying semantic theories in vocabulary acquisition, it becomes apparent to me that the verb lemmas are best divided into a certain number of groups. While some groups can be created to contain near-synonyms, others can contain near-

antonyms. As long as the members in a group are related in some way semantically and systematically, it is expected that the association between the members would help learners with vocabulary acquisition.

4.4.1.3 Some explorations in verb classification based on syntactic constructions

The studies by the researchers mentioned above have shown that a verb's feature displays better within a certain semantically related construction, either in a hierarchical system as Godman (1982) suggests or in grids such as many researchers use (such as Channell 1981, Harvey 1983, and Stieglitz 1983). The next question that arises immediately is whether it is possible to discover and create such constructions for all the English verbs, because my research needs to cover all the verb lemmas used by both the learners and the NSs. Another question is: apart from the groupings of near-synonyms in the manner of Channell and Godman, are there other ways of classifying and grouping verbs?

One attempt to group English verbs was made by Levin (1993) who tried to classify 3000 English verbs based on diathesis alternations. Verbs that share the same alternation are grouped in the same class. The assumption behind Levin's verb classification is that "syntactic properties are semantically determined" (Levin 1993: 14). In other words, the meaning of a verb determines its syntactic properties such as diathesis alternation. One example from her classification is as follows (Levin 1993: 209):

> 37.7 Say Verbs
>
> Class members: announce, articulate, blab, blurt, claim, confess, confide, convey, declare, mention, note, observe, proclaim, propose, recount, reiterate, relate, remark, repeat, report, reveal, say, state, suggest

Levin's classification provides many sets of verb lists that are arranged in particular alternations as shown above. Even though she classified only 3000 English verbs, her classification outnumbers the total number of the verbs used by the COLEC writers and the LOCNESS writers. Therefore, Levin's classification remains a good source for consultation if the verbs in my two corpora are to be classified.

Figure 4. 6 The verbs and phrases that share the 'V that clause' structure by Francis *et al*. (1996: 98-99)

accept 2	confide	indicate 2	profess 1
acknowledge 1	conjecture 2	insinuate 1	promise 1
admit 1	contend 2	insist 1,2	pronounce 3
advise 1	crow 3	instruct 1	prophesy
advocate 1	declare 1,2	intimate 7	propose 1,3,4
affirm 1	decree 2	joke 2	protest 3
agree 1,3	demand 1	lament 1	quip 2
allege	deny 1	maintain 2	radio 6
allow 6	dictate 2	mandate 5	reason 4
announce 1,2,3	direct 12	marvel 1	recall 1
argue 1	disclose	mention 1	recollect
ask 2,3	divulge	moan 2	recommend 2
assert 1	emphasize	move 16	recount 1
attest	enthuse 1	muse 1	reflect 5
aver	estimate 1	note 10,11	regret 1
beg 1	explain 1,2	observe 3	remark 1
boast 1	forecast 2	opine	remonstrate
brag	foretell	ordain2	report 1
caution 2	grant 3	order 2.2	request 1
certify 1	groan 3	plead 1,4	reveal 1
claim 1	grouse 2	pledge 2	rule 7
command 1	grumble 1	posit	say 1,2
comment 1	guarantee 3	postulate 1	signal 2
complain 1	guess 1	pray 1	signify 2
concede 1	hazard 3	preach 2	sneer
conclude 1	hint 2	predict	specify 2
concur	hypothesize	pretend 3	speculate 1
confess 1	imply 1	proclaim 1,2	state 8
stipulate	swear 2,3	underscore 1	wager 2
stress 1	testify 1	urge 3	warn 1
submit 2	theorize	venture 3	warrant 4
suggest 1,3,4	threaten 1	volunteer 4	write 5
surmise 1	underline 1	vow 1	
(not) let on	point out 2	report back 1	
make out 3	put down 1		

One problem with Levin's exploration in verb classification, including some previous and preliminary attempts by Alexander and Kunz (1964) and Bridgeman *et al*. (1965) is that her research was seriously affected by the limited quantity of data. When, later, the computer technology allowed for a large store of texts, this problem was solved to a large extent. In an innovative program of COBUILD (see Sinclair 1987), Francis *et al*. (1996) tried to make an exhaustive classification of the English verbs according to the patterns verbs share (see 7.3 for a definition of pattern). The following is an extract of their inclusion of the verbs which share the structure of 'V that-clause'. The numbers in Figure 4.6 are the sense numbers that appear

in the *Collins Cobuild English Dictionary*. Francis *et al*.'s categorisation based on corpus data is impressive and should make it a useful source for my own verb lemma categorisation.

Since my ambition in this chapter is to list all the verb lemmas (that occur in the two corpora), and any one verb lemma will appear only once in the tables, some degree of reconciliation has to be made. Instead of using the classification by diathesis classification like Levin or the pattern classification like Francis *et al*., I have to use a very loose standard in my own classification.

4.4.1.4 Some explorations of the links between the known and unknown and between L1 and L2

It goes without saying that learners build up their knowledge of vocabulary gradually. New knowledge will be acquired more easily if it is based upon some existing knowledge in one's mind. A practical approach advocated by Stieglitz (1983: 71) for vocabulary reinforcement supports the notion that "[For students] new vocabulary items should be presented in known structures and, whenever possible, should be centered around one topic."

Apart from the need to link the known and the unknown in vocabulary expansion, some psycholinguistic studies have investigated how the learner's L1 lexicon relates to his L2 translation. As maintained by many psycholinguists such as Albert and Obler (1978), and Meara (1982) (cited in Channell 1988: 86), "there is interaction between the lexicons of the two languages [L1 and L2] in one user". Channell (1988: 86) quotes Albert and Obler as follows:

> It is clear that words in one language, and their translation equivalents in the other (when such exist), are related in the brain in a non-random way, much as a word and its synonym in the same language may be connected in an associated network. (Albert and Obler 1978: 246)

Based upon the notion and research findings above, it is envisaged that by positing the known verbs together with the unknown verbs (those that are used only by the NSs), there is a better chance for learners to become familiar with the new target vocabulary.

4.4.2 Working out a design for the grouping of the verb lemmas of COLEC and LOCNESS

Based on the rational study above, I am going to make a design for the layout of the verb lemmas that appear in COLEC and LOCNESS. Since a random list of verbs is of little value to the language learner and teacher, the verbs will be presented in groups. Obviously a table will be helpful for a tidy presentation.

A central aim of this research is to identify the range of the learners' lexicon in verbs, and the disparity and similarity between the learner English and the NS English, as shown in the first three research questions (see 4.1). A second aim is to see how a display of the similarity and disparity of the verb lemmas in a list could aid learners with large-scale vocabulary expansion. To realise the first aim the verb lemmas by the NSs and the learners are compared and contrasted. One column provides the lemmas of one corpus and another column provides the lemmas of the other (see Table 4.2 for a first impression). If a particular verb is used by both of the group writers, they are listed in both of the columns (with the frequency provided in brackets). This will show the similarity of the two groups of writers. If a particular verb is only used in one of the corpora (mostly in LOCNESS, but occasionally in COLEC), this verb is listed in the corresponding corpus. This will show the disparity between the two groups of writers. In order for people to see the similarity and disparity clearly, verbs that are used by both of the groups are aligned on the same line. If one verb occurs only in one corpus, it is listed only in this corpus (in bold) and a blank space will be used in the column of the other corpus to show the contrast. When all the verb lemmas of COLEC are added, the range of the learners' productive vocabulary in verbs will be known. In the same way, the range of the NSs productive vocabulary in verbs can be discovered.

These two columns would, however, answer only the first three research questions. To answer the fourth question (how could the research findings based on the similarity and disparity between the NS English and the NNS English be used by the language teacher, the learner and even the writer of teaching materials?), a decision was made to group the verb lemmas of the two corpora by semantic links. There are two reasons for doing this. One is that by grouping verb lemmas according to the semantic relationship (either synonymous[32] or antonymous, or

32 This word is not accurate because it is very doubtful that whether there exist true synonyms in a language. For

some members showing a scale of change from one pole to another), the psycholinguistic factors in vocabulary acquisition are taken into account, as Channell, Godman and Wyatt advocate. It is envisaged that when the compared and contrasted lemma list is used by the language teacher and the language learner it will help with vocabulary acquisition. While we are talking about an approach to facilitate easier vocabulary acquisition, it would certainly be helpful if the verb lemmas could be analysed in a semantic componential analysis approach to mark the semantic relation between one verb and another in a semantic field as Godman and Channell have done. However, given the space available in one chapter, I will not take this approach for my research. Instead, I am going to use a simple but easy-to-understand approach, Chinese pin-yin,[33] which in many cases has the advantage of showing the meaning components of each lemma, and the difference from other members of the same set when put in a semantic field set, but does not require too much detailed analysis in the style of Godman, Channell and others. Some detailed advantages of using Chinese pin-yin will be looked at later (4.4.2.1).

It must be pointed out that this design of the table columns and contents is not for a detailed clarification of the uses of the verb lemmas that appear in the two corpora. The tables are used primarily for seeking answers to the first, research-oriented, research questions and then seeking answers to the last, teaching-oriented, research question. When the verb list tables are passed to learners, they will have a chance to see what verbs are produced by learners who share the same background as themselves and what verbs are produced only by the NSs. For the new verbs, the learners would have a chance to associate them first with their L1 translations and then with their L2 clusters.

4.4.3 General principles of grouping the verb lemmas in COLEC and LOCNESS

Before large-scale groupings are undertaken, it is necessary to lay out some basic principles that are applicable to the categorisation of all the verb lemmas. The major sources to consult in making decisions as to whether a particular verb should be added to a group are Levin's

---

a discussion of this issue, see Ullman 1967 and Leech 1974.

33 In the future, if the use of the Chinese translation is taken seriously for pedagogical uses, Chinese characters will play a greater role than Chinese pin-yin because the characters are more recognisable and straightforward.

work on verb classifications (Levin 1993), the Grammar Patterns reference book by Francis *et al*. (1996). Occasionally, the *Merriam–Webster Online Thesaurus*[34] is also used. For the Chinese pin-yin translation, *A New English-Chinese Dictionary* (century edition) is the main source to be used. The frequencies of the verb lemmas in the tables are not normalised mainly because the intention of classifying the verb lemmas is to see which lemmas are used and which lemmas are not used in the two corpora, and there is no intention to compare the frequencies of each lemma in the two corpora. Another reason for not having normalised the frequencies is that it is easy to see which verbs are just above the cutting point (which is 3 and inclusive). If we look at Table 4.2, the numbers in the first left column (the **English** column) show the lemma number contrast between LOCNESS and COLEC. Take the 'house 2-1' for example; the '2' means that two verb lemmas are used in LOCNESS (*HOUSE* and *STORE*) and one verb lemma is used in COLEC (*STORE*) in this sense group.

In some cases the most commonly used verb lemma (maily in LOCNESS) is selected to be the cover verb to represent a group of verb such as the 'put' group (See Table 4.2). But in other cases, the cover verbs are not the most often used verb lemmas either in LOCNESS or in COLEC such as the 'relax' group (see Table 4.3). In the latter cases, a word which is more likely to represent the whole group than others will be chosen. My experience as a teacher of English will be used in making such decisions. As a general rule, the cover verbs are small and easy words. It is expected that when the lists are passed over to teachers and learners in the end, it would be easy for them to handle.

As mentioned above, the Chinese pin-yin will be used to link L1 and its L2 equivalents. As may be unknown to NNS of Chinese, in many cases where the English lemmas do not distinguish themselves clearly by form, the Chinese pin-yin has the advantage of being able to make some distinctions. In a sense group, the change in the pin-yin not only shows the subtle change from one lemma to another but also the connection between the two. The following four adjoining sense groups, entitled *put*, *house*, *fix* and *fill* will be used to validate my decision to make use of Chinese pin-yin to serve this purpose (see Table 4.2).

34 http://www.m-w.com/

Table 4. 2 A categorisation of the sense group of *PUT*, *HOUSE*, *FILL* and *FIX*

English	LOCNESS	COLEC	Chinese Pin-yin
Fill and Fix			
put 3-3	**file**(13)		gui-fang
		lay (20)	fang; ge-xia; pu-she
	place (60)	place (15)	fang-zhi
	put (182)	put (203)	an-fang
house 2-1	**house** (4)		shou-cang, cun-fang, gei…fang-zi-zhu
	store (8)	store (10)	chu-cang; chu-bei
fill 3-2		**crowd** (6)	ji-man; zhuang-man
	fill (27)	fill (33)	zhuang-man
	load (6)		zhuang-zai; zhuang-man
	pack (4)		bao-zhuang; zhuang-man
fix 2-2		**equip** (4)	zhuang-bei
	fix (10)	fix (7)	an-zhuang; shi…gu-ding
	install (6)		an-zhuang; zhuang-bei

In the sense group *put*, the verb *FILE* (gui-fang) and *PLACE* (fang-zhi) share the same element, i.e. 'fang', which can be glossed as *PUT* in English. But *FILE* has a special sense of 'gui' which means 'sorting out in a certain order' while *PLACE* does not. Instead, the component 'zhi' requires a specific place in *PLACE* but not in *FILE*. The verb *LAY* could also be represented by 'fang' in Chinese, but the other senses, i.e. 'ge-xia' (which means 'putting things on the ground') and 'pu-she' (which means 'setting or producing public facilities such as cables') are distinguishing. In other words, the similarity of 'fang' relates the verbs *FILE* and *PLACE* and *LAY* to each other but the other senses such as 'gui', 'zhi', 'ge-xia' and 'pu-she' separates them apart. Slightly away from the four verbs (*FILE*, *PLACE, LAY* and *PUT*) are two verbs (*HOUSE* and *STORE*) which are closely related to each other by the element 'cang', which implies 'reservation' or 'shelter'. A second meaning of *STORE* in Chinese is 'chu-bei' which is applicable only to goods. The following examples from LOCNESS show the uniqueness of the verb lemma *STORE*.

1 ey compete for the right to collect and **store** garbage from other states. It m
2 has change because of it. I'm able to **store** almost all the information I ne
3 erty. Money enables a human being to **store** what he owns, to pay his expenses,
4 will have. Aside from being able to **store** great amounts of info, I'm also ab
5 my computer. This means I'm able to **store** more information in an organized
6 t get stale as quickly when they are **stored** in these bags (without any air).
7 from the mother so that the eggs can be **stored** and used later if the pregnancy i
8 ast amount of knowledge and information **stored** in a computer is knowledge which
9 ast amount of knowledge and information **stored** in a computer is knowledge which

The verb *HOUSE*, however, is not limited to things only. This is reflected by the multiple translations in the Chinese pin-yin (shou-cang; cun-fang; gei … fang-zi-zhu). The following

two examples from LOCNESS can be translated into the third Chinese pin-yin, which is ‘gei … fang-zi-zhu’.

1) After the experiences of the American Revolution, the first Congress wanted to ensure that people could protect themselves by serving as armed citizens in the militia and that citizens could not be forced to *house* soldiers in peacetime.
2) This money may not come from the perpetrator of the crime but it will come from the individual's family, friends and acquantances in the form of tax dollars. As the number of criminals increases so to will the tax money needed to *house* these individuals

In the *fill* group, four very near synonyms are *CROWD*, *FILL*, *LOAD* and *PACK*. The similarity between the four words is echoed by the same sense ‘zhuang-man’ which means ‘put something into a container into an area to its full extent’. The senses that are not identical to each other are observable by the different translations. If we look at the column of Chinese pin-yin, we may have the following observations. In the verb *CROWD*, there is a unique sense which is not shared by others, i.e. ‘ji-man’ which means ‘to press, force, or thrust something (or some people) into a small space’. The sense ‘zhuang-zai’ (put things in a vehicle) is only unique to the verb *LOAD*. The peculiarity of *PACK* lies in the way how a container is filled, i.e. ‘bag’ or ‘package’ in English and ‘bao-zhuang’ in Chinese. If we look at the *fix* sense group, it can be found that the sense component ‘zhuang’ is shared by all the three verbs (*EQUIP*, *FIX* and *INSTALL*) as in the sense ‘an-zhuang’ which means ‘fit a piece of equipment so that it can work properly’. What contributes to the distinction between the three verbs is that the verb *FIX* usually requires a place for an equipment to be fit as shown by another sense of the verb ‘shi…gu-ding’ (attach something to a place firmly) while the other two verbs do not have this requirement. What relates the three verbs is that the verb *INSTALL* shares the sense ‘an-zhuang’ with the verb *FIX* but shares the sense ‘zhuang-bei’ with the verb *EQUIP* (provide tools or equipment to somebody).

The four groups are placed together under one sub-title “Fill and Fix” because there is a broad sense link between one and the others. This sense link can be seen by the existence of the Chinese components ‘an’, ‘zhuang’ and ‘cang’ which span two sense groups. For example, the Chinese pin-yin ‘an’ exists in the *put* group and the *fix* group (‘*an*-fang’ in the *put* group and

'*an*-zhuang' in the *fix* group, the element of 'zhuang' exists in the *fill* group and the *fix* group ('an-*zhuang'* in the *fix* group and *zhuang*-zai in the *fill* group). By seeing the same sense elements spanning two groups, I am assured that these verb lemmas can be classified under the same sub-title.

A drawback of using Chinese pin-yin rather than Chinese characters in the lemma lists is that some different senses are expressed by the same form of pin-yin, as in 'shi-yan' for both 'experiment' and 'test'. In such a case, the Chinese characters will have to be used to make a distinction in 'shi-yan (实验), shi-yan (试验)'. Due to the limited space in tables, some repeated expressions shared by two senses (or two versions of the same sense) are omitted by enclosing the senses or different versions of the same sense in square brackets as in the following table. The highlighted 'shi' is applicable to all the senses or different versions of the same sense (see Table 4.3).

Table 4. 3 A categorisation of the sense group of *RELAX* and its translations

relax 4-3	**ease** (13)		**shi**…[fang-song; shu-shi; an-xin]
	loose (11)		**shi**... song-chi
		relax (34)	**shi**…[song-chi; qing-song; xiu-xi]
	rest (7)	rest (10)	**shi**…[xiu-xi; qing-song; xie-xi]
	sit (32)	sit (59)	**shi**…jiu-zuo; zuo

To categorise the verbs one by one into a suitable sense group is not always easy because "Language, like the world of living organisms, does not yield up neat or exact taxonomies, but blurred interrelationships" (Godman 1982: 41). Sometimes, difficult decisions have to be made as to which sense group a verb belongs. Due to the fact that a verb may have multiple features shared by two or even three sense groups, the researcher has to choose painstakingly the sense group to which it best belongs. For example, *VARY* can go with *DIFFER* and *CHANGE*; *FAVO(U)R* with *LIKE* and *SELECT*; *IMPROVE* with *ADVANCE* and *CORRECT*; and *ENFORCE* has bearings not only on LEGAL VEBS, but also *CONDUCT* and *FORCE*. Manual checks in WordSmith Tools were frequently used to see which is the dominant sense used is in the concordances of the verb but it was unfortunately not possible to check every verb for the sense used because of time pressures. It must be admitted beforehand that even though the categorisation of the sense groups in this research is meant to put near synonymous verbs together, it should not be claimed that such a categorisation is watertight and has not problems. Researchers and teachers may find it necessary to make appropriate

changes to make themselves comfortable with the categorisation.

Having found a way to group all the verb lemmas according to a certain relationship between each other, it was decided that the two long original verb lists were to be chopped to several shorter lists so that they could be put under control and handled easily when and if they were passed over to language teachers, learners, course designers, and others who have an interest in interpreting this hitherto undiscovered feature of learner English. Originally, it was planned that two large groups would be used to cover all the verb lemmas, i.e. the neighbouring concept sense groups and the near antonymous sense groups. But it turned out to be difficult for the following reasons and some amendments had to be done.

1) Some concept groups are so special (such as SENSE VERBS, LEGAL VERBS) that they had better stand out from the rest of groups;
2) There are several sense groups (such as 'Say and Write', 'Know and Reason') that are outstanding in number; and therefore deserve to be listed separately;
3) There are some sense groups that became odds and ends after the majority of verbs were grouped to certain categories;
4) The neighbouring sense groups seemed too long and difficult to handle and therefore needed to be divided into two.

In the end, six groups are used and in such an order:

1) neighbouring concept groups (1),
2) neighbouring concept groups (2),
3) near antonymous groups,
4) five large family groups,
5) special concept groups and
6) miscellaneous groups.

Some words that are obviously wrongly annotated – verbs such as *TOUT* in LOCNESS and *FIRE* in COLEC – are not included in the final verb lemma lists.

4.4.3.1 Neighbouring concept groups (1)

By 'neighbouring concept groups', I am trying to be 'fuzzy' enough to be able to cover as many pairs or clusters as possible. Every pair has been given a name so that the sub-groups can be related to each other and become easy for teachers (and others) to manage (see Table 4.4). Some names are deliberately made funny so that the verbs in the subgroups can be easily remembered. There are altogether 27 groups in this section and all the groups of the neighbouring groups are named as binominals (such as 'join' and 'gather') for the sake of convenience. Other neighbouring groups exceeding two subgroups will be arranged in the next section.

Table 4. 4 A categorisation of the verb lemma lists by neighbouring groups (1)

English	LOCNESS	COLEC	Chinese Pin-yin
Join and Gather			
join 5-4	attend (24)	attend (53)	chu-xi; can-jia; zhao-liao
	engage (13)	engage (56)	shi…can-jia; shi…juan-ru
	join (40)	join (101)	can-jia; jia-ru; shi-jie-he; lian-jie
	partake (7)		can-jia; can-yu; fen-xiang; fen-dan
	participate (19)	participate (13)	can-jia; can-yu; fen-xiang; fen-dan
gather 3-1	collect (7)	collect (10)	shou-ji; cai-ji; qu-zou
	gather (18)		shou-ji; sou-ji; ji-ju
	pool (4)		(zi-jin) ru-huo; gong-xiang; fen-xiang
Oppose and Contradict			
oppose 4-2	**defy** (6)		gong-ran-fan-kang; mie-shi; miao-shi
	object (6)		fan-dui; bu-zan-cheng
	oppose (37)	oppose (4)	fan-dui; fan-kang; di-kang
	resist (6)	resist (18)	di-kang; di-dang; di-zhi; kang-ju
contradict 2-0	**conflict** (4)		chong-tu; di-chu; dou-zheng; zheng-lun
	contradict (8)		mao-dun; di-chu
Move and Shake			
move 6-6		**climb** (14)	pan-deng; xiang-shang-pa; pa-dong
	flow (5)	flow (23)	liu-dong; liu-chu
	fly (15)	fly (19)	fei; fei-xing; fei-wu
	go (442)	go (962)	qu; li-qu
	leave (201)	leave (106)	li; li-qu; li-kai
	move (65)	move (37)	yi-zou; ban-zou; yi-dong
	roll (7)		gun-dong; fan-gun
shake 2-0	**shake** (13)		yao; dong-yao; yao-yun; dou-dong
	sway (4)		yao-dong, bai-dong; yao-bai; dong-yao
Retire and Relax			
retire 1-1	retire (8)	retire (9)	tui-xiu
relax 4-3	**ease** (13)		shi…[fang-song; shu-shi; an-xin]
	loose (11)		shi… song-chi
		relax (34)	shi…[song-chi; qing-song; xiu-xi]
	rest (7)	rest (10)	shi…[xiu-xi; qing-song; xie-xi]
	sit (32)	sit (59)	shi…jiu-zuo; zuo
Reach and Arrive			
reach 2-1	**culminate** (4)		da-dao-ding-dian
	reach (57)	reach (117)	di-da; da-dao; dao-da
arrive 3-3	approach (7)	approach (5)	xiang…jie-jin; kao-jin
	arrive (24)	arrive (34)	dao-da; lai-dao
	come (324)	come (331)	lai; lai-dao

Offset and Limit

offset 2-0	**counteract** (5)		zhong-he; zu-ai
	offset (6)		di-xiao
limit 3-1	**cap** (4)		xian-zhi (… shu-liang)
	limit (27)	limit (118)	xian-zhi, jian-shao
	restrict (21)		xian-zhi

Label and List

label 2-0	**label** (17)		biao-ji
	mark (9)		biao-ming; xian-shi; zheng-ming
list 4-3	count (7)	count (12)	shu
	list (6)	list (12)	lie-ju
	print (5)		da-yin
	register (5)		zhu-ce
		type (20)	dai-zi

Repair and Correct

repair 1-1		**repair** (18)	xiu-li, xiu-bu; bu-chang
	compensate (8)		bu-chang; pei-chang
correct 3-1	correct (7)	correct (7)	gai-zheng; jiao-zheng
	modify (5)		xiu-gai, geng-gai; huan-he
	renew (7)		geng-xin; bu-chong

Change and Differ

change 10-8	alter (31)	alter (15)	gai-bian, gai-dong
		alternate (4)	lun-liu, jiao-ti
	change (215)	change (1008)	gai-bian; bian-hua
	convert (8)	convert (9)	zhuan-bian, zhuan-hua
		exchange (23)	jiao-huan; geng-huan
	render (8)		shi…bian-wei
	replace (32)		geng-huan, ti-huan; jie-ti
	shift (6)		ti-huan; zhuan-huan
	switch (7)		zhuan-bian, zhuan-huan
	transfer (15)	transfer (6)	zhuan-ran, zhuan-yi; diao-dong
	transform (5)	transform (5)	zhuan-huan, gai-bian
	translate (4)	translate (11)	fan-yi
differ 3-2	differ (7)	differ (4)	yü…bu-yi-yang, xiang-yi
	diversify (4)		shi…duo-yang-hua
	vary (6)	vary (18)	gai-bian, shi…duo-yang-hua

Cope and Solve

cope 7-4	**address** (22)		dui-fu; chu-li
	cope (17)	cope (20)	dui-fu; (tuo-shan) chu-li
	deal (95)	deal (126)	dui-fu; ying-fu; chu-li; an-pai
	handle (15)	handle (6)	chu-li; guan-li; dui-dai (ren)
	organise (6)		an-pai; zu-zhi; shi…you-tiao-li
	organize (6)	organize (19)	an-pai; zu-zhi; shi…you-tiao-li
	tackle (23)		zhuo-shou-chu-li; dui-fu; jie-jue
solve 3-4	manage (27)	manage (28)	she-fa-jie-jue
	resolve (11)	resolve (20)	jie-jue; jie-da; jie-chu
		settle (17)	tiao-ting, jie-jue (zheng-duan)
	solve (51)	solve (175)	jie-jue; jie-da; jie-shi

Travel and Carry

travel 2-1	**sail** (5)		hang-xing
	travel (51)	travel (28)	lü-xing
carry 2-1	carry (117)	carry (62)	yun-song; yun-zai
	convey (8)		yun-song; shu-song; chuan-song

Adapt and Compromise

adapt 3-2	**accommodate** (6)		shi…[shi-ying; he-xie]; tiao-zheng
	adapt (12)	adapt (103)	shi-ying; shi-he; shi…[shi-ying; shi-he]
	adjust (5)	adjust (28)	gai-bian…yi-shi-ying; tiao-zheng

compromise 4-0	**align** (5)		dui-qi; shi…[cheng-yi-lie; cheng-yi-hang]
	balance (11)		bao-chi-ping-heng; xie-tiao; quan-heng
	compromise (14)		tuo-xie; rang-bu; he-jie
	reconcile (12)		tiao-jie; tiao-ting; tiao-he
Defeat and Compete			
defeat 2-2	defeat (6)	defeat (72)	zhan-sheng; ji-bai
	overcome (24)	overcome (38)	ke-fu
compete 8-2	**combat** (16)		zhan-dou; ge-dou
	compete (17)	compete (7)	bi-sai; jing-zheng; dui-kang
	contend (4)		jing-zheng; zheng-dou
	fight (91)	fight (38)	zhan-dou; bo-dou; fen-dou; dou-zheng
	protest (8)		fan-kang, kang-yi
	race (6)		bi-sai; jing-sai; jing-zheng
	rebel (8)		fan-pan
	revolt (6)		fan-pan
Lift and Grow			
lift 2-1	lift (5)	lift (4)	ju-qi, ti-qi
	uphold (4)		gao-ju
grow 8-5	**breed** (10)		yang-yu, wei-yang
	cultivate (5)	cultivate (5)	pei-yang
	feed (20)	feed (8)	wei-yang
	grow (84)	grow (105)	sheng-zhang
	nurture (8)		yang-yu, pei-yang
	plant (4)	plant (38)	zhong-zhi
	raise (73)	raise (65)	yang-yu, si-yang, zhong-zhi
	rear (4)		fu-yang, si-yang, zhong-zhi
Aim and Plan			
aim 5-4	aim (30)	aim (6)	mu-di-zai-yu; da-suan; qi-tu
	attempt (51)	attempt (4)	shi-tu; chang-shi
	strive (9)		nu-li, fen-dou, li-qiu; fan-kang
	struggle (19)	struggle (17)	dou-zheng, nu-li; zheng-zha
	try (266)	try (461)	chang-shi; shi-tu, nu-li
plan 6-3	arrange (5)	arrange (15)	an-pai
	design (19)		she-ji
	plan (22)	plan (68)	ji-hua
	prepare (46)	prepare (38)	zhun-bei
	program (7)		she-ji (jie-mu)
	project (7)		she-ji (xiang-mu)
Act and Conduct			
act 2-1	act (94)	act (55)	xing-dong
	behave (7)		xing-wei
conduct 10-6	**administer** (4)		shi-shi, shi-xing, zhi-xing
	commit (87)	commit (8)	fan…(zui-xing), gan
	conduct (25)	conduct (6)	shi-shi, chu-li, jin-xing
	enforce (17)		shi-shi, qiang-zhi-zhi-xing
	execute (12)		shi-shi, zhi-xing
	implement (21)		guan-che, zhi-xing
	perform (48)	perform (16)	lü-xing, zhi-xing
	play (147)	play (332)	ban-yan
	practice (13)	practice (43)	lian-xi
	practise (4)	practise (295)	lian-xi
Select and Focus			
select 7-6	choose (134)	choose (122)	tiao-xuan
	elect (27)	elect (5)	xuan-ju
	pick (17)	pick (10)	tiao-xuan
	prefer (19)	prefer (76)	geng-xi-huan, xuan-ze
	select (13)	select (45)	tiao-xuan
	tend (50)	tend (32)	qing-xian

	vote (16)		xuan-ju
focus 5-2	concentrate (18)	concentrate (27)	zhuan-xin
	center (4)		yi…wei-zhong-xing
	centre (5)		yi…wei-zhong-xing
	focus (38)	focus (12)	shi…ji-zhong
	highlight (7)		shi…xian-zhu, shi…tu-chu
Enable and Facilitate			
enable 1-1	enable (31)	enable (20)	shi…neng-gou
facilitate 1-0	**facilitate** (6)		shi…bian-li, shi…rong-yi
Point and Refer			
point 1-1	point (53)	point (27)	zhi-xiang
refer 1-1	refer (41)	refer (17)	zhi-xiang; can-kao
Satisfy and Amuse			
satisfy 1-1	satisfy (17)	satisfy (39)	man-zu; shi … man-yi
amuse 2-1	**entertain** (5)		shi…huan-le; gei… yu-le
	please (9)	please (5)	shi…gao-xing; shi…man-yi
Set and Decide			
set 1-1	set (74)	set (95)	gui-ding
decide 2-2	decide (153)	decide (92)	jue-ding
	determine (58)	determine (29)	jue-xin
Press and Impress			
press 1-0	**press** (6)		ya, ji-ya
impress 1-1	impress (5)	impress (8)	gei … ji-shen-de-yin-xiang; ming-ji
Confuse and Mistake			
confuse 1-1	confuse (7)	confuse (8)	mi-huo
mistake 0-1		**mistake** (4)	wu-jie
Occupy and Own			
occupy 1-1	occupy (5)	occupy (22)	zhan-you; qin-zhan
own 2-2	own (20)	own (30)	you, yong-you; zhi-pei
	possess (22)	possess (7)	zhan-you, yong-you; zhi-pei
Fall and Pour			
fall 2-1	fall (57)	fall (74)	die-luo; die-dao; zhui-luo; luo-xia
	stumble (4)		ban-die; ban-dao; jie-jie-ba-ba-de-shuo
pour 1-1	pour (10)	pour (17)	dao; qing-xie
Save and Spare			
save 2-1	**rescue** (4)		yuan-jiu; wan-jiu; ying-jiu
	save (78)	save (243)	jiu-zhu; wan-jiu; da-jiu; jie-sheng
spare 1-1	spare (4)	spare (17)	jian-sheng, jie-yue
Qualify and Deserve			
qualify 1-1	qualify (4)	qualify (6)	you…zi-ge
deserve 1-0	**deserve** (35)		zhi-de

There are 67 verb lemmas that occur only in the LOCNESS corpus and they are singled out as follows (see Figure 4.7):

Figure 4. 7 The verb lemmas that occur only in LOCNESS in Table 4.4.

accommodate address administer align balance behave breed cap center centre combat compensate compromise conflict contend contradict convey counteract culminate defy deserve design diversify ease enforce entertain execute facilitate gather highlight implement label mark modify nurture object offset organise partake pool press print program project protest race rear rebel reconcile register render renew replace rescue restrict revolt roll sail shake shift strive stumble sway switch tackle uphold vote

Broadly speaking, the COLEC writers are performing fairly close to the norm of the LOCNESS writers as far as the frequency is concerned. In the *join* sub-group of the 'Join and Gather' group, the performance of the learners approximates that of the NSs. Among the five verbs used in this group, the learners are found to be using four of them. However, it seems that there is not a single group in COLEC that matches completely the entire variety of verb lemmas in LOCNESS. In some sense groups, such as *contradict*, *offset* and *compromise*, no verb lemmas appear in COLEC. The 'teddy bear' phenomenon does exist in some of the sense groups such as *collect* and *correct* (see 2.7.2 in Chapter Two for an introduction to the 'teddy bear' phenomenon). Instead of using *collect* only, the NSs are also using *gather* and *pool*. Instead of using *correct* all the times (as the NNS do), the NSs are using *modify* and *renew* for the purposes of fine distinction.

4.4.3.2 Neighbouring concept groups (2)

There are 11 clusters of senses in the neighbouring concept groups (2) (see Table 4.5). As can be detected, there are identifiable changes of verb senses from one group to another. In the first sense group 'From the Beginning to the End', the verb sense starts from *BEGIN* and proceeds to *DEVELOP*, and shifts to *PROSPER* and then *COMPLETE* and finally ends with *STOP*. The 'teddy bear' problem is also identifiable in this section. A serous absence of varieties in the production of verbs in COLEC is manifest in many sense groups such as 'Destroy and Throw' and 'Rule and Control'.

Table 4. 5 A categorisation of the verb lemma lists by neighbouring groups (2)

English	LOCNESS	COLEC	Chinese Pin-yin
From the Beginning to the End			
begin 3-2	begin (178)	begin (180)	kai-shi
	embark (9)		kai-shi; zhuo-shou; cong-shi
	start (129)	start (53)	kai-shi
develop 9-9	advance (14)	advance (29)	jin-bu
		better (15)	jin-bu
	develop (90)	develop (347)	fa-zhan; jin-bu
	enhance (23)	enhance (12)	ti-gao
	improve (69)	improve (530)	ti-gao
	mature (6)		cheng-shu
	process (16)	process (4)	jia-gong
	progress (16)	progress (16)	jin-bu
	promote (39)	promote (18)	fa-yang; tui-xiao
	reform (5)	reform (13)	fa-zhan; gai-ge
prosper 2-0	**prosper** (5)		fan-rong, chang-sheng
	thrive (4)		xing-wang, fan-rong, chang-sheng
complete 5-5	accomplish (18)	accomplish (9)	jie-shu, wan-cheng
	complete (9)	complete (28)	jie-shu, wan-cheng

	finish (11)	finish (290)	jie-shu, wan-cheng
	fulfil(l) (18)	fulfil(l) (5)	shi-xian; wan-cheng
	graduate (5)	graduate (83)	bi-ye, jie-shu (xue-ye)
stop 6-3	**abort** (4)		liu-chan;zhong-zhi (中止)
	cease (16)		ting-zhi, jie-shu
	end (4)	end (24)	jie-shu; zhong-zhi (终止)
	quit (7)	quit (5)	ting-zhi; fang-qi
	resign (12)		ci-zhi
	stop (116)	stop (75)	ting-zhi
Fill and Fix			
put 3-3	**file**(13)		gui-fang
		lay (20)	fang; ge-xia; pu-she
	place (60)	place (15)	fang-zhi
	put (182)	put (203)	an-fang
house 2-1	**house** (4)		shou-cang, cun-fang, gei…fang-zi-zhu
	store (8)	store (10)	chu-cang; chu-bei
fill 3-2		**crowd** (6)	ji-man; zhuang-man
	fill (27)	fill (33)	zhuang-man
	load (6)		zhuang-zai; zhuang-man
	pack (4)		bao-zhuang; zhuang-man
fix 2-2		**equip** (4)	zhuang-bei
	fix (10)	fix (7)	an-zhuang; shi…gu-ding
	install (6)		an-zhuang; zhuang-bei
Seek and Find			
seek 4-3	hunt (7)	hunt (8)	zhui-bu; zhui-gan; da-lie
	resort (5)		su-zhu; qiu-zhu; ping-jie
	search (16)	search (19)	sou-xun; sou-cha
	seek (49)	seek (20)	zhui-qiu; tan-suo; xun-zhao
find 10-6	**detect** (5)		jue-cha; zheng-cha; fa-xian
	discover (62)	discover (13)	fa-xian, fa-jue; zhao-dao
	find (310)	find (1054)	zhao-dao; gan-dao; de-dao
	identify (19)	identify (15)	ren-chu; shi-bie; jian-ding
	note (9)		zhu-yi, liu-yi; ji-lu
	notice (19)	notice (27)	jue-cha-dao; zhu-yi-dao; ren-chu
	observe (11)	observe (10)	jue-cha-dao; zhu-yi-dao; guan-cha
	perceive (15)		yi-shi-dao; jue-cha; ba…kan-zuo
	recognise (30)		ren-chu; bian-ren
	recognize (45)	recognize (26)	ren-chu; bian-ren
explore 2-3		**explore** (4)	tan-suo; tan-ce
	investigate (5)	investigate (8)	diao-cha; shen-cha
	research (8)	research (12)	yan-jiu; tan-jiu
Destroy and Throw			
destroy 11-8	damage (17)	damage (23)	po-huai
	destroy (48)	destroy (31)	po-huai
		disable (4)	shi … can-fei
	disrupt (5)		po-huai; fen-lie, wa-jie
	distort (4)		wai-qu; qu-jie
	harm (13)	harm (133)	wei-hai
	infringe (7)		wei-fan (违反); wei-fan (违犯)
	injure (8)	injure (16)	shang-hai
		poison (5)	du-hai
	pollute (6)	pollute (203)	wu-ran
	ruin (6)		po-huai
		spoil (6)	guan-huai (chong-huai)
	undermine (9)		(an-zhong) po-huai
	violate (9)		wei-fan (违反); wei-fan (违犯)

break 7-5	break (69)	break (70)	da-po; da-duan; nong-huai
	crash (4)	crash (7)	za-sui; zhuang-ji
	crush (4)		ya-sui; ya-huai; zhen-ya; ya-kua
	cut (44)	cut (39)	jie; qie; kan; ge; shan; jian
	rip (8)		si; che; bo; hua-po; pi
	smash (6)	smash (57)	da-po, da-sui; fen-sui; wa-jie
	tear (9)	tear (4)	si-kai, si-po, si-lie, che-po
eliminate 13-3	**abandon** (11)		fang-qi, pao-qi
	abolish (27)		fei-chu; fei-zhi
	eliminate (32)	eliminate (5)	xiao-mie; xiao-chu
	eradicate (5)		gen-chu; xiao-mie
	erase (4)		qing-chu; xiao-mie
	hang (20)	hang (5)	jiao-si
	kill (196)	kill (57)	sha-si
	murder (26)		mou-sha, an-sha
	relinquish (4)		fang-qi; che-li
	remove (35)		xiao-chu; qü-diao
	repeal (9)		fei-chu; che-xiao
	revoke (4)		fei-chu; che-xiao
	slaughter (7)		tu-sha
throw 3-1	**cast** (4)		reng, tou, zhi
	dump (7)		qing-dao; pao-qi
	throw (46)	throw (29)	reng, tou, zhi
dismiss 3-0	**betray** (5)		bei-pan
	dismiss (12)		jie-gu
	sack (7)		jie-gu
rid 3-0	**deprive** (13)		bo-duo
	rid (4)		bo-duo
	strip (5)		bo-duo
Care and Worry			
care 4-3	care (30)	care (52)	guan-xin; dan-xin; hu-li; bao-yang
	comfort (8)		an-wei; kuan-wei
	concern (20)	concern (22)	guan-xin; gua-nian; dan-xin
	mind (14)	mind (11)	guan-xin; dang-xin; liu-xin; jie-yi
cure 2-2	cure (5)	cure (21)	zhi-yu; zhi-liao
	treat (47)	treat (37)	yi-zhi; zhi-liao
bother 3-2	**bother** (7)		ma-fan; fan-rao; fen-rao
		disturb (13)	da-rao; rao-luan
	plague (6)		zhe-mo; fan-rao; shi…ku-nao
	worry (24)	worry (70)	sao-rao; kun-rao; zhe-mo; shi…dan-xin
Contact from Mild to Wild			
kick 1-0	**kick** (6)		ti
touch 2-1	**tap** (6)		qing-pai; qing-qiao
	touch (6)	touch (36)	chu-mo; jie-chu
beat 7-6	beat (25)	beat (38)	qiao; da; chong-ji; da-bai
	bump (5)		zhuang-shang; shi…meng-ji
	hit (38)	hit (113)	da; da-ji; ji-zhong
		hurt (60)	shang-hai
	knock (6)	knock (62)	qiao; ji; da; qiao-diao
	shoot (23)	shoot (10)	fa-she; kai-qiang; fang-pao
	strike (13)	strike (9)	da; ji; zhuang-ji
	whip (10)		bian-chi, chou-da
From Expect to Suspect			
want 5-6	desire (16)	desire (5)	xiang-wang, ke-wang, xi-wang
	hope (25)	hope (117)	xi-wang
	intend (21)	intend (18)	da-suan, ji-hua, xiang-yao
		long (15)	ke-wang
	want (426)	want (1349)	xiang-yao, yao

	wish (56)	wish (51)	zhu-yuan; xiang-yao
expect 2-2	expect (57)	expect (54)	qi-dai; yu-liao
	predict (15)	predict (8)	yu-liao; yu-yan
imagine 8-4	**assume** (41)	*assume* (4)	yi-wei; jia-ding, she-xiang
	doubt (7)		cai-xiang; huai-yi
	envisage (4)		xiang-xiang, she-xiang
	figure (4)		gu-ji; pan-duan; ji-suan
	imagine (15)	imagine (25)	xiang-xiang; she-xiang; liao-xiang
	guess (10)	guess (13)	cai-ce; tui-ce
	suspect (4)		huai-yi; tui-ce; cai-xiang
	wonder (23)	wonder (13)	yi-huo, na-men
Rule and Control			
control 8-1	**conquer** (5)		zheng-fu; gong-ke
	control (55)	control (132)	kong-zhi; zhi-pei; guan-zhi
	dominate (24)		zhi-pei, tong-zhi, kong-zhi
	govern (12)		tong-zhi; zhi-pei; ying-xiang
	manipulate (17)		cao-zong; kong-zhi; cao-zuo
	regulate (17)		kong-zhi; tiao-zheng
	reign (4)		tong-zhi; jia-yü; kong-zhi
	rule (29)		tong-zhi; kong-zhi; gui-ding
overwhelm 2-0	**override** (4)		ya-dao; you-xian-yu
	prevail (6)		zhan-shang-feng; zhan-you-shi
monitor 1-0	**monitor** (6)		jian-shi; jian-kong
From Excited to Offended			
excite 0-1		**excite** (6)	shi.. xing-fen
surprise 1-1	**shock** (10)		shi…zhen-jing; shi-fen-nu
		surprise (7)	shi…chi-jing; shi…cha-yi
frighten 5-2	fear (29)	fear (11)	hai-pa; dan-xin; you-lü, kong-ju
	frighten (8)		shi…jing-kong; xia-hu
	haunt (4)		(gui-hun) chu-mo-yü; zhe-mo
	scare (6)		shi…kong-ju; shi…hai-pa
	threaten (17)	threaten (18)	wei-xie; kong-he; dong-he
offend 3-0	**anger** (10)		ji-nu, shi…fa-nu
	offend (6)		chu-nu; mao-fan
	upset (8)		shi…xin-fan-yi-luan
Walk, Jump and Flee			
walk 4-4	run (120)	run (183)	pao; ben-pao; ben-chi
	rush (9)	rush (95)	chong; ben; chuang
	step (10)	step (121)	xing-zou; bu-xing
	walk (30)	walk (135)	zou; bu-xing; san-bu
pass 3-2	cross (15)	cross (5)	tong-guo; yue-guo; du-guo
	pass (91)	pass (132)	jing-guo; chuan-guo; yue-guo; tong-guo
	slip (9)		liu; liu-zou; qiao-qiao-de-zou
jump 2-3	**bounce** (4)		tan-tiao
		dance (28)	tiao-wu; tiao-dong
		hop (19)	tan-tiao; dan-tui-tiao
	jump (11)	jump (67)	tiao; tiao-yue
escape 2-1	escape (28)	escape (5)	tao-tuo
	flee (4)		tao-pao
Happen and Exist			
cause 7-3	cause (195)	cause (269)	yin-qi; zao-cheng
	evoke (18)		yin-qi; huan-qi
	pose (16)		zao-cheng, xing-cheng
	provoke (9)		ji-qi; tiao-dou; you-dao
	result (58)	result (49)	dao-zhi; zao-cheng
	spark (5)		ji-fa
	stimulate (6)	stimulate (12)	ci-ji
happen 2-2	happen (145)	happen (175)	fa-sheng

	occur (93)	occur (25)	fa-sheng
exist 2-1	exist (92)	exist (40)	cun-zai
	perpetuate (4)		yong-cun

There are 79 verb lemmas that occur only in the LOCNESS corpus and they are singled out as follows (see Figure 4.8):

Figure 4. 8 The verb lemmas that occur only in LOCNESS in Table 4.5

abandon abolish abort anger assume betray bother bounce bump cast cease comfort conquer crush deprive detect dismiss disrupt distort dominate doubt dump embark envisage eradicate erase evoke figure file flee frighten govern haunt house infringe install kick load manipulate mature monitor murder note offend override pack perceive perpetuate plague pose prevail prosper provoke recognise regulate reign relinquish remove repeal resort revoke rid rip ruin rule sack scare shock slaughter slip spark strip suspect tap thrive undermine upset violate whip

4.4.3.3 Near antonymous groups

There are 19 pairs of near-antonyms arranged in this section (see Table 4.6) and 100 verb lemmas that are not present in COLEC. The 'teddy bear' problem is serious in some sense groups such as *get (again)*, *ignore*, *emphasise*, *ignore*, and *live*. Take the *get again* sub-sense of *GET* for example, only a general word *recover* is present and three synonyms for *recover* (*regain*, *reinstate*, and *restore*) are missing in COLEC. Take *ignore* for another example, it is found that COLEC writers are mainly using this general word while the NSs are using more synonyms, i.e. *disregard*, *neglect* and *overlook*. However, the 'teddy bear' problem is not always dominant in COLEC. For example, in the 'Open and Close' group, among the four verbs used by the NSs (*disclose*, *expose*, *open* and *reveal*), three verbs also appear in COLEC (*expose*, *open* and *reveal*); similarly, among the four verbs used by the NSs (*bury*, *close*, *cover* and *hide*), three verbs are found in COLEC (*close*, *cover* and *hide*). But as a rule, the verbs used by the learners are shorter, informal and more neutral in register. Formal verbs which could be used to replace the superordinates are mostly missing in COLEC. The use of more formal verbs for the sense of *GET* such as *attain*, *contract*, *earn* and *profit* has not been shown to be part of the production vocabulary of the learners. However, it must be admitted that there seems to be a gradation of proficiency in production.

Table 4. 6 A categorisation of the verb lemma lists by near antonymous groups

English	LOCNESS	COLEC	Chinese Pin-yin
Get and Give			

get 12-10	achieve (86)	achieve (58)	(jing-nu-li) da-dao; wan-cheng
	acquire (19)	acquire (26)	qu-de; huo-de; xue-dao
	attain (12)	attain (21)	da-dao; huo-de; dao-da
	benefit (35)	benefit (65)	shou-yi
	contract (31)		huan-shang…ji-bing
	earn (31)	earn (91)	ying-de; zhuan-de; bo-de
	gain (66)	gain (168)	huo-de; ying-de; zheng-de
	get (421)	get (2316)	huo-de; de-dao; ying-de
	obtain (46)	obtain (77)	huo-de; de-dao
	profit (5)		de-li; huo-yi
	receive (103)	receive (39)	shou-dao (收到); shou-dao (受到); jie-shou
	win (87)	win (47)	ying-de
get (again) 4-1	recover (11)	recover (9)	hui-fu; wan-hui
	regain (14)		hui-fu; shou-hui; fu-de; fan-hui
	reinstate (5)		shi…hui-fu; shi…zheng-chang
	restore (4)		hui-fu (恢复); hui-fu (回复); fu-bi
give 19-11	**allocate** (4)		fen-pei, fen-pai; hua-bo
	attribute (20)		gui-yin-yu
	award (4)		shou-yu; gei-yu
	confer (4)		shou-yu; fu-yu
	contribute (27)	contribute (61)	gong-xian; juan-xian
	dedicate (4)	dedicate (4)	feng-xian; xian-shen
	devote (5)	devote (97)	feng-xian; xian-shen
	distribute (5)		fen-fa; fen-pei; fen-song
	donate (9)		juan-xian; juan-zeng
	give (453)	give (299)	zeng-song; shou-yu; gei
	grant (21)	*grant* (69)	tong-yi; shou-yu
	hand (8)	hand (6)	jiao-chu, chuan-di, gei
	invest (12)		tou-zi
	offer (65)	offer (43)	ti-gong; ti-chu
	pay (145)	pay (224)	fu-kuan
	present (99)	present (5)	zeng-song, shou-yu; cheng-xian
	provide (132)	provide (91)	ti-gong
	sacrifice (20)		xian-chu; xi-sheng
	supply (9)	supply (19)	gong-ying, ti-gong
Remember and Forget			
remember 2-4	remember (42)	remember (327)	ji-de; xiang-qi, hui-yi-qi
		memorize (12)	ji-yi,
		recall (8)	hui-yi
	recite (8)	recite (32)	lang-song bei-song
forget 1-1	forget (31)	forget (133)	wang-ji
Include and Exclude			
include 4-3	contain (44)	contain (17)	bao-kuo
	entail (11)		qian-she
	include (73)	include (25)	bao-kuo
	involve (108)	involve (9)	qian-she
exclude 1-0	**exclude** (8)		pai-chu (bu-bao-kuo)
Emphasise and Ignore			
emphasize 4-1	**emphasise**(7)		qiang-diao
	emphasize (20)	emphasize (4)	qiang-diao
	reinforce (16)		qiang-diao
	stress (13)		qiang-diao
ignore 4-3	ignore (28)	ignore (38)	hu-shi, hu-lue
	disregard (6)		bu-li, hu-shi
	neglect (8)	neglect (21)	hu-shi, hu-lüe
		omit (6)	sheng-lue, shu-hu
	overlook (10)		hu-lue, kan-lou
Bring and Take			

bring 1-1	bring (211)	bring (364)	dai-lai
take 4-3	**pillage** (5)		lue-zou
	rob (8)	rob (10)	qiang-zou
	steal (18)	steal (12)	tou-zou
	take (987)	take (1231)	dai-zou
Honour and Dishonour			
honour 3-2	**honour** (4)		gei-yu…rong-yu; shi…zeng-guang
	praise (7)	praise (11)	zan-yang; zan-mei; cheng-zan
	reward(16)	reward (5)	bao-da; chou-lao; jiang-shang
scold 7-4	**accuse** (20)		ze-bei
	blame (33)	blame (6)	bao-yuan
	charge (7)	*charge* (15)	zhi-kong; kong-gao
	complain (7)	complain (16)	bao-yuan
	condemn (23)		qian-ze
	criticise (14)		pi-ping
	criticize (8)		pi-ping
		scold (7)	man-ma (ze-ma)
Borrow and Lend			
borrow 2-2	borrow (6)	borrow (6)	jie-(ru); zu-jie
	owe (4)	owe (11)	qian
lend 1-0	**lend** (4)		jie-(chu)
Teach and Learn			
teach 8-5	**direct** (18)		zhi-dao; zhi-yin
	educate (34)	educate (32)	jiao-yu; pei-yang; xun-lian
	guide (6)		zhi-dao; yin-dao; dai-ling
	head (14)	head (5)	shua-ling; zai…de-qian-tou
	instruct (7)		jiao; xun-lian; zhi-dao; zhi-shi
	lead (266)	lead (188)	ling-dao; shuai-ling; zhi-hui
	teach (102)	teach (124)	jiang; jiao-shou; jiao-yu; jiao-dao
	train (21)	train (39)	pei-yang; pei-xun; xun-lian
learn 2-3	learn (111)	learn (1623)	xue; xue-xi; xue-hui
		master (209)	jing-tong; zhang-wo; kong-zhi
	study (37)	study (860)	xue-xi; gong-du; yan-jiu; tan-tao
Pull and Push			
pull 3-2	draw (49)	draw (83)	lai; tuo; chou; yin-dao; ji-qu
	extract (4)		(yong-li) qu-chu; ti-lian; zhai-lu
	pull (12)	pull (19)	la; tuo; che; qian; ba; zhai; chou
push 1-1	push (32)	push (9)	tui; tui-dong; tui-jin
Protect and Attack			
protect 3-2	defend (20)	defend (6)	fang-wei; bao-wei; wei…bian-hu
	protect (44)	protect (157)	bao-hu; jing-jie
	safeguard (7)		bao-hu; han-wei; wei-hu
attack 1-0	**attack** (44)		xi-ji; gong-ji
Encourage and Discourage			
encourage 3-1	encourage (66)	encourage (47)	gu-li, ji-li
	inspire (5)		gu-wu; ji-qi
	motivate (4)		tui-dong, ji-fa
discourage 1-1	discourage (11)	discourage (11)	shi…xie-qi
Like and Hate			
like 10-8	admire (24)	admire (14)	qin-pei; zan-mei; xin-shang;
	appreciate (15)	appreciate (7)	xin-shang, shang-shi; gan-ji
		cherish (6)	zhen-xi, zhen-ai; ai-hu
	enjoy (53)	enjoy (101)	xin-shang, xi-ai; xiang-you
	favour (9)		xi-ai; zhi-chi; zan-cheng
	like (91)	like (920)	xi-huan
	love (35)	love (96)	xi-huan; re-ai; lian-ai
	miss (14)	miss (27)	dian-nian; huai-nian
	respect (23)	respect (94)	zun-zhong; zun-jin

	value (6)		zun-zhong; zhong-shi
	worship (4)		chong-bai; zun-jing; xin-feng
hate 5-3	dislike (9)	dislike (20)	bu-xi-huan, yan-wu
	hate (11)	hate (36)	zeng-hen; bu-xi-hua
	regret (5)	regret (4)	hui-hen; ao-hui
	repent (24)		hui-wu; hui-gai; hou-hui
	resent (6)		yuan-hen; fen-hen
Agree and Disagree			
agree 8-4	accept (168)	accept (41)	ren-ke; jie-shou
	acknowledge (14)		cheng-ren
	admit (39)	admit (6)	cheng-ren; gong-ren
	agree (80)	agree (44)	tong-yi; cheng-ren
		approve (4)	tong-guo; pi-zhun; zan-cheng
	condone (6)		kuan-shu
	confess (27)		gong-ren, cheng-ren, tan-pai
	consent (4)		zan-cheng; zhun-xu
	sign (20)		qian-zi (yi-shi-tong-yi)
disagree 6-4	deny (53)	deny (8)	ju-jue; xie-jue
	disagree (24)		bu-tong-yi, fou-ren
	forbid (5)	forbid (26)	bu-tong-yi; zheng-zhi
	refuse (63)	refuse (26)	ju-jue; ju-shou;ju-gei
	reject (85)	reject (13)	ju-jue; di-zhi; pai-chi
	veto (9)		fou-jue; jin-zhi; fan-dui
Increase and Decrease			
increase (1)		**accumulate** (20)	ji-lei, ji-ju
(desirably)	**amount** (4)		zeng zhang
8-9	broaden (7)	broaden (6)	tuo-kuan, kuo-da
		enlarge (28)	kuo-da
	expand (25)	expand (14)	kuo-da, kuo-zhang
		fasten (9)	jia-su
	increase (132)	increase (466)	zeng-zhang
	rise (19)	rise (98)	ti-gao, zeng-zhang
	speed (4)	speed (14)	jia-su
	spring (5)		(xun-su) zeng-zhang
	strengthen (29)	strengthen (21)	zeng-qiang
increase(2)		**accelerate** (5)	sheng-ji,
(undesirably)	**exacerbate** (4)		jia-ju, e-hua
2-2		**hurry** (57)	cui-cu; shi-jia-kuai
	worsen (5)		e-hua, bian-de-geng-huai
decrease	**alleviate** (8)		huan-ji, huan-he
13-7	**curtail** (5)		suo-duan; xue-jian
	decline (8)	decline (54)	shuai-tui; xia-jiang
	decrease (41)	decrease (197)	xia-jiang
	degrade (4)		bian-chu; shi…jiang-ji
	drop (29)	drop (31)	xia-jiang
	erode (4)		qin-shi; mo-sun
	lessen (8)		jian-shao, jiang-di
	lower (40)	lower (12)	jian-shao, jiang-di
	reduce (91)	reduce (147)	jian-shao, jiang-di
	relieve (12)	relieve (4)	jian-qing
	shorten (4)	shorten (4)	suo-duan
	weaken (12)		xue-ruo
Allow and Prevent			
allow 3-3	allow (270)	allow (11)	yun-xu, zhun-xu
	let (76)	let (156)	yun-xu, rang
	permit (11)	permit (10)	yun-xu, xu-ke; zhun-xu
prevent 9-4	avoid (37)	avoid (66)	bi-mian
	ban (98)	ban (8)	jin-zhi

	bar (4)		jin-zhi; fang-ai
	block (6)		zu-ai, fang-ai
	deter (4)		fang-zhi, zu-zhi
	inhibit (9)		jin-zhi, zu-zhi
	prevent (76)	prevent (166)	zu-zhi; yu-fang
	prohibit (13)	prohibit (30)	jin-zhi; zu-zhi
	shun (5)		bi-mian; hui-bi
Open and Close			
open 4-3	**disclose** (5)		jie-kai; jie-fa; tou-lou; xie-lou
	expose (29)	expose (5)	jie-lou; luo-lou; bao-lou
	open (42)	open (37)	kai; da-kai; zhang-kai; jie-kai
	reveal (51)	reveal (5)	zhan-xian; jie-shi; bao-lou
close 4-3	**bury** (4)		yan-cang; yan-mai
	close (15)	close (26)	guan-bi
	cover (18)	cover (40)	fu-gai; yan-gai; yan-shi
	hide (14)	hide (7)	yin-cang; yan-gai; yan-shi
Unite and Divide			
unite 16-7	**accompany** (6)		ban-sui; pei-ban
	associate (26)	associate (5)	lian-jie; jie-he
	bind (18)		jie-he; zhan-he; yue-shu
	bond (4)		jie-he; zhan-he; wei…zuo-bao
	combine (6)	combine (7)	lian-he; hun-he; zu-he
	connect (9)	connect (19)	lian-xi; lian-jie
	couple (4)		shi…hun-pei; lian-he; jie-he
	integrate (20)		hun-he; jie-he
	link (34)	link (6)	lian-jie; lian-xi
	marry (22)	marry (4)	jie-hun
	mix (11)		jie-he; he-bing; hun-he
	relate (44)	relate (22)	jiang…lian-xi-qi-lai
	reunite (6)		chong-ju; zai-lian-he; zai-ji-he
	tie (6)		lian-jie; lian-he; yue-shu
	unify (5)		tong-yi; shi…yi-yuan-hua
	unite (8)	unite (9)	lian-he; tuan-jie; jie-he
break 8-1	**discriminate** (11)		qi-shi
	divide (20)	divide (7)	fen-kai, ge-li
	divorce (13)		li-hun
	isolate (5)		ge-li; ge-jue
	prejudice (5)		qi-shi
	segregate (4)		shi-xing-zhong-zu-ge-li; fen-li; ge-li
	separate (34)		fen-li
	split (5)		pi-kai; si-lie
Enter and Emit			
enter 8-2	enter (53)	enter (84)	jin-ru; can-jia;
	import (5)		jin-kou; yin-jin; shu-ru
	inject (5)		zhu-she; zhu-ru
	insert (7)		cha-ru; qian-ru
	interfere (11)		gan-she; gan-yu; fang-ai
	intervene (9)		gan-she; gan-yu; jie-ru
	invade (4)	invade (6)	qin-ru; qin-lue; qin-fan
	tamper (5)		gan-she
emit 1-0	**emit** (4)		shi-fang (san-fang)
Keep and Lose			
keep 8-4	keep (164)	keep (390)	bao-cun; bao-liu; bao-shou
	maintain (40)		wei-chi; bao-chi; wei-hu
	preserve (13)	preserve (12)	bao-hu; wei-hu; wei-chi
	remain (87)	remain (10)	bao-chi-bu-bian; reng-shi
	retain (34)		bao-chi; bao-liu; bao-you
	reserve (4)		bao-cun; bao-liu; yu-ding

	stay (73)	stay (71)	bao-chi-xia-qu; ting-liu
	sustain (7)		wei-chi; gong-yang; zhi-cheng
lose 1-1	lose (183)	lose (211)	shi-qu diu-shi

The verbs that occur only in LOCNESS (93) are singled out as follows (see Figure 4.9):

Figure 4. 9 The verb lemmas that occur only in LOCNESS in Table 4.6

accompany accuse acknowledge alleviate allocate amount attack attribute award bar bind block bond bury charge condemn condone confer confess consent contract couple criticise criticize curtail degrade deter direct disagree disclose discriminate disregard distribute divorce donate emit emphasise entail erode exacerbate exclude extract favour grant guide honour import inhibit inject insert inspire instruct integrate interfere intervene invest isolate lend lessen maintain mix motivate overlook pillage prejudice profit regain reinforce reinstate repent resent reserve restore retain reunite sacrifice safeguard segregate separate shun sign split spring stress sustain tamper tie unify value veto weaken worsen worship

4.4.3.4 Six large family groups

Some groups contain so many components that it is worthwhile to single them out from other groups of categorisation display. These groups include six families, i.e. 'Say and Write', 'Know and Reason', 'Make and Work', 'Use', 'See', 'Show and Prove', and 29 subgroups (see Table 4.7). There are other big groups such as 'get' and 'give', 'unite' and 'break', but since they are roughly antonyms to each other they are grouped together under 4.4.3.3.

As mentioned in Chapter Two (and other chapters) of this thesis, learner written English is strongly featured by an oral style compared with the more formal English of NSs as a whole. This feature seems to be apparent everywhere in this list. In the 'argue' subgroup, formal and academic verbs such as *argue*, *debate*, *dispute* and *refute* are almost completely missing in COLEC. Instead of using these verbs, requisite for academic writing, the COLEC writers use only *quarrel* which is obviously a word from non-academic fields such as fiction.

Table 4. 7 A categorisation of the verb lemma lists by large family groups

English	LOCNESS	COLEC	Chinese Pin-yin
Say and Write			
argue 4-1	**argue** (162)		bian-lun; tao-lun
	debate (24)		bian-lun; tao-lun
	dispute (4)		zheng-lun; zheng-chao
		quarrel (5)	zheng-chao, chao-nao
	refute (19)		fan-bo, bo-chi
discuss 2-1	discuss (77)	discuss (10)	shang-tao, tao-lun
	consult (7)		zi-xun
say/write 18-15	answer (24)	answer (50)	hui-da
	assert (9)		xuan-cheng; duan-yan

	claim (81)	claim (10)	zi-cheng, zhu-zhang
	declare (8)	declare (4)	xuan-bu; xuan-cheng
	dictate (6)		kou-shu
	express (56)	express (23)	biao-da
	mention (36)	mention (22)	ti-qi, shuo-dao
	profess (7)		cheng-ren, biao-bai
		pronounce (6)	xuan-bu, fa-yin
	record (4)	record (5)	ji-lu
	repeat (6)	repeat (14)	chong-fu
	reply (10)		hui-da
	report (31)	report (29)	bao-dao
	say (493)	say (718)	shuo, jiang
		shout (10)	hu-han
	speak (65)	speak (287)	shuo, jiang
		spell (6)	pin-xie; pin-zi
	state (180)		chen-shu, chan-ming
	talk (70)	talk (117)	jiang-hua
	voice (9)		fa-yan, biao-da
	write (114)	write (292)	shu-xie
tell 11-7		**advertise** (6)	guang-gao
	announce (11)		xuan-gao
	broadcast (4)	broadcast (6)	guang-bo, chuan-bo
	inform (20)	inform (14)	gao-zhi
	issue (6)		ban-bu
	proclaim (9)		xuan-bu, sheng-ming
	publicise (6)		gong-bu, gong-gao
	publish (12)	publish (4)	fa-biao, chu-ban
	remind (12)	remind (4)	ti-xing
	tell (145)	tell (286)	gao-su
	televise (9)		dian-shi-bo-song
	warn (4)	warn (11)	jing-gao
spread 1-1	spread (14)	spread (14)	chuan-bo
ask 2-1	ask (113)	ask (169)	wen; xun-wen; zi-xun
	question (36)		fa-wen; xun-wen
demand 7-5		**command** (5)	ming-ling
	demand (23)	demand (12)	yao-qiu; qiang-qiu
	invite (14)	invite (8)	yao-qing
	invoke (5)		bao-you, qi-qiu
	order (10)	order (11)	ming-ling
	prescribe (11)		ming-ling, gui-ding
	pray (22)		qi-dao
	request (4)	request (6)	yao-qiu
quote 2-0	**cite** (14)		yin-yong, ju-li
	quote (7)		yin-yong, yin-zheng
describe 3-0	**describe** (53)		miao-shu; miao-xie
	depict (12)		miao-hui, miao-xie, miao-shu
	portray (19)		miao-hui, miao-xie, miao-shu
explain 5-3	account (6)	account (12)	jie-shi (… yuan-yin)
	clarify (6)		chan-ming, cheng-qing
	exemplify (5)	explain (34)	ju-li-shuo-ming
	explain (72)	illustrate (8)	jie-shi
	illustrate (41)		chan-shi
persuade 2-0	**convince** (12)		shuo-fu
	persuade (22)		shuo-fu
suggest 6-3	advise (9)	advise (11)	quan-gao; zhong-gao; jian-yi
	advocate (16)		ti-chang; zhu-zhang
	introduce (61)	introduce (11)	jie-shao; yin-jin
	preach (8)		bu-dao, jiang-dao; shuo-jiao

	propose (15)		jian-yi, ti-yi
	suggest (51)	suggest (32)	jian-yi, ti-yi; tui-jian
comment 1-0	**comment** (4)		ping-lun
mock 3-0	**mock** (11)		chao-nong, chao-xiao
	ridicule (15)		chao-nong, chao-xiao, xi-luo
	parody (5)		(hui-xie) mo-fang
cheat 3-1	cheat (24)	cheat (56)	qi-pian
	deceive (5)		qi-pian
	fool (5)		qi-pian, yu-nong
call 6-3	**appoint** (15)		ren-ming; wei-ren
	call (88)	call (91)	cheng-hu; ba…jiao-zuo
	define (71)	define (4)	gei…xia-ding-yi; que-ding…jie-xian
	name (5)	name (5)	gei…qu-ming; reng-ming; ti-ming
	nominate (5)		ti-ming; ren-ming; zhi-ding
	term (4)		ba…cheng-wei; ba…jiao-zuo
Know and Reason			
know 6-5		**acquaint** (4)	shi … liao-jie
	comprehend (8)		li-jie, ling-hui, dong
	interpret (17)		li-jie; jie-shi
	know (363)	know (2859)	zhi-dao; liao-jie; shu-xi
	realise (98)		ren-shi-dao; liao-jie
	realize (122)	realize (196)	ren-shi-dao; liao-jie
		specialize (5)	zhuan-gong, shan-chang
	understand (151)	understand (344)	li-jie, ming-bai; dong
think 8-6	consider (158)	consider (119)	ren-wei, ba…kan-zuo; kao-lü
	contemplate (8)		chen-si, si-cun
	deem (17)		ren-wei, xiang-xin
	feel (433)	feel (328)	fa-jue; gan-dao; ren-wei
	judge (74)	judge (13)	ren-wei; ping-pan; ping-jia
	regard (29)	regard (46)	ren-wei; ba…kan-zuo
	suppose (12)	suppose (15)	cai-xiang; xiang-xiang; jia-ding
	think (366)	think (2132)	xiang, ren-wei; si-suo
analyse 5-1	**analyse** (7)		fen-xi; jie-xi
	analyze (21)	analyze (13)	fen-xi; jie-xi
	diagnose (5)		zhen-duan; fen-xi (yuan-yin)
	induce (5)		gui-na
	reason (5)		tui-duan; bian-lun
distinguish 2-1	**classify** (8)		fen-lei, gui-lei
	distinguish (7)	distinguish (12)	qu-bie; bian-bie; shi-bie
compare 3-1	compare (49)	compare (52)	bi-jiao; dui-zhao
	contrast (10)		dui-zhao; dui-bi
	outweigh (20)		bi … zhong
Make and Work			
make 10-10	build (58)	build (123)	jian-she, jian-zao
	coin (4)		du-zhuan
		construct (5)	jian-zao,jian-she
	create (179)	create (18)	chuang-zao
		dig (6)	wa (jing)
	establish (53)	establish (21)	jian-li, she-li
	found (9)	found (9)	chuang-ban; chuang-jian
	generate (11)		sheng-zhi, chan-sheng, chuang-zao
	institute (10)		chuang-li; shi-xing
	invent (18)	invent (29)	fa-ming, chuang-zao
	make (987)	make (3856)	sheng-chan, zhi-zao
		manufacture (9)	sheng-chan, zhi-zao
	produce (81)	produce (221)	sheng-chan, zhi-zao
work 3-2	work (210)	work (819)	gong-zuo; chan-sheng
	function (11)		yun-zuo

	operate (10)	operate (16)	cao-zuo; gong-zuo
Use			
use	adopt (44)	adopt (27)	cai-yong, cai-na; cai-qu (tai-du)
(properly)	apply (56)	apply (65)	ying-yong, yun-yong, shi-yong
10-9	consume (20)	consume (18)	hua-fei; xiao-fei; hao-jin
	employ (16)		gu-yong; shi-yong
	exercise (13)	*exercise* (46)	yun-yong; xing-shi (zhi-quan)
		exert (9)	xing-shi; fa-hui (wei-li)
	hire (5)		gu-yong; zu-yong
	recycle (14)	recycle (14)	hui-shou-li-yong
	spend (99)	spend (237)	hua-fei; yong
	use (563)	use (1342)	yong; shi-yong; ying-yong
	utilize (12)	utilize (5)	li-yong
use	**abuse** (12)		lan-yong; nue-dai
(excessively) 5-2	exhaust (4)	exhaust (6)	yong-wan; hao-jin
	exploit (6)		kai-cai; li-yong; bo-xue
	misuse (13)		wu-yong; lan-yong; nue-dai
	waste (25)	waste (345)	lang-fei; lan-yong
See			
see 10-6	**encounter** (19)		yu-dao; ou-ran-peng-dao; zao-yu
	interview (4)		jie-jian, hui-jian; mian-shi
	look (205)	look (301)	kan, qiao, wang
	meet (64)	meet (191)	yu-dao, peng-jian; ju-hui
	see (635)	see (530)	kan-jian, kan-dao; li-jie
	stare (4)	stare (5)	ning-shi, ding
	view (72)		kan, guan-kan; cha-kan; kan-dai
	visit (11)	visit (33)	fang-wen; can-guan; bai-fang
	watch (91)	watch (213)	guan-kan; zhao-kan; kan-shou
	witness (25)		mu-du, mu-ji
Show and Prove			
show 11-6	**demonstrate** (33)		zheng-ming; (yi-shi-li) shuo-ming
	display (22)	display (7)	xian-shi; xian-lou; biao-xian
	epitomise (9)		biao-ming
	implicate (4)		an-han; yi-wei-zhe
	imply (22)	imply (8)	an-zhi; an-shi; yi-zhi
	indicate (10)	indicate (21)	zhi-shi; biao-ming; xiang-zheng; yu-shi
	manifest (5)		biao-ming; xian-shi; xian-lou
	reflect (37)	reflect (5)	fan-ying; biao-xian; chen-si; fan-xing
	show (350)	show (138)	biao-ming; shuo-ming; zheng-ming
	mean (224)	mean (351)	yi-zhi; yi-wei-zhe
	signify (4)		biao-shi; biao-ming; yi-wei-zhe
represent 3-2	**embody** (4)		dai-biao; shi…ju-ti-hua
	represent (61)	represent (4)	dai-biao
	symbolise (9)		xiang-zheng, biao-zhi
		symbolize (6)	xiang-zheng, biao-zhi
prove 10-4	**affirm** (4)		duan-yan, zheng-shi.; pi-zhuen
	assure (5)		shi…fang-xin; xiang…bao-zheng
	confirm (5)	confirm (5)	zheng-shi, ken-ding; que-ren
	ensure (33)	ensure (7)	bao-zheng-huo-de; dan-bao
	guarantee (16)		bao-zheng; dan-bao; que-bao
	insure (9)		tou-bao; que-bao
	justify (46)		zheng-ming…zheng-dang
	prove (118)	prove (83)	zheng-ming; zheng-shi
	promise (8)	promise (6)	bao-zheng; yun-nuo, xu-nuo
	secure (9)		bao-zheng; wei…dan-bao

The verbs that occur only in LOCNESS (80) are singled out as follows (see Figure 4.10):

Figure 4. 10 The verb lemmas that occur only in LOCNESS in Table 4.7

abuse advocate affirm analyse announce appoint argue assert assure cite clarify classify coin comment comprehend consult contemplate contrast convince debate deceive deem demonstrate depict describe diagnose dictate dispute embody employ encounter epitomise exemplify exercise exploit fool function generate guarantee hire implicate induce institute insure interpret interview invoke issue justify manifest misuse mock nominate outweigh parody persuade portray pray preach prescribe proclaim profess propose publicise question quote realise reason refute reply ridicule secure signify state symbolise televise term view voice witness

4.4.3.5 Special concept groups

Some groups in the language use seem to be too deviant from the verbs above. Therefore, there is a need to cover some special verbs such as link verbs and legal activity-related verbs (see Table 4.8).

The verbs that occur only in LOCNESS (11) are singled out as follows (see Figure 4.11):

Figure 4. 11 The verb lemmas that only occur in LOCNESS in Table 4.8

convict enact legalise legalize legislate sentence sue overhear kiss date time

Table 4. 8 A categorisation of the verb lemma lists by special concept groups

English	LOCNES	COLEC	Chinese Pin-yin
LINK VERBS 3-3	seem (294)	seem (88)	hao-xiang-shi
	sound (14)	sound (10)	ting-shang-qu
	become (499)	become (606)	bian-de
LEGAL VERBS 7-0	**convict** (10)		[zheng-ming; xuan-pan] you-zui
	enact (8)		[ban-bu; fa-bu] fa-ling;
	legalise (6)		shi-he-fa-hua
	legalize (28)		shi-he-fa-hua
	legislate (10)		li-fa; zhi-ding-fa-lü
	sentence (9)		pan-jue; xuan-pan
	sue (11)		xiang-fa-yuan-qi-su; ti-qi-su-song
LIGHT VERBS 1-3	burn (9)	burn (16)	ran-shao
		light (8)	zhao-liang
		shine (4)	shan-guang
SENSE VERBS 14-15	breathe (4)	breathe (7)	hu-xi
	hear (91)	hear (143)	ting-dao
	listen (31)	listen (260)	ting
	overhear (6)		tou-ting
	dream (9)	dream (18)	zuo-meng
	sleep (15)	sleep (18)	shui-jiao
	wake (6)	wake (11)	jue-xing
	eat (51)	eat (146)	chi
		smell (6)	wen

		taste (13)	chang
	drink (45)	drink (102)	he
	smoke (10)	smoke (18)	chou-yan
	kiss (8)		wen; jie-wen
	laugh (11)	laugh (32)	da-xiao
	smile (4)	smile (8)	xiao; wei-xiao
	cry (12)	cry (92)	ku-qi; jian-jiao
		sing (43)	chang-ge
WEATHER WORDS 2-6	blow (10)	blow (7)	gua-feng
		cool (4)	shi … leng-jing
	heat (4)	heat (4)	jia-re
		rain (97)	xia-yu
		warm (4)	wun-nuan
		water (9)	gei … jiao-shui
SPORTS VERBS 1-2		**skate** (23)	hua-bing
	swim (12)	swim (24)	you-yong
HOUSE WORK VERBS 1-2	cook (9)	cook (51)	peng-ren
TIME-RELATED 2-0	**date** (7)		que-ding … de ri-qi
	time (6)		an-pai … de shi-jian

Among the 11 verbs in Figure 4.10, seven are 'legal verbs'. This seems to suggest that this is a special topic in LOCNESS that is not shared by COLEC. The absence of *overhear* in COLEC is not surprising because its superordinate *hear* could have been used in its place. One point that needs some explanation is the verb *KISS*. This verb is used infrequently by the learners, presumably due to the cultural disparity which means that the act of kissing is not a public topic as it is in the western world. Arguably, this verb should not be listed for the learners to practise because my intuition is that the learners know how to use this word; it is only the cultural difference that prevents it from being used very often. This has shown a weak point of real data analysis because corpus-based studies deal only with what has been produced. For the unproduced part, it is hard to know whether it is due to avoidance (as probably in this case) or inability in production. Researchers using a corpus-based approach should be ready to consult their intuitions, and should not be totally dependent on the the corpus data.

4.4.3.6 The miscellaneous groups

Some verbs have become 'odds and ends' after a majority of verbs have been grouped according to my previous distinctions. I shall put these verbs into the 'miscellaneous' group (see Table 4.9). Unlike the other groups, this section has no group titles because it is hard to find proper names for its members.

Table 4. 9 A categorisation of the verb lemma lists: the miscellaneous groups

English	LOCNESS	COLEC	Chinese Pin-yin
add 3-3	add (45)	add (34)	jia; tian-jia
	adhere (12)	adhere (6)	fu-zhuo; yi-fu
	attach (11)	attach (27)	fu-jia; tie; fu; ji
affect 6-4	affect (81)	affect (42)	ying-xiang
	effect (25)	effect (19)	ying-xiang
	impact (6)		ying-xiang
	influence (28)	influence (26)	ying-xiang
	matter (12)	matter (5)	you-ying-xiang
	subject (7)		shou … ying-xiang
afford 1-1	afford (39)	afford (18)	mai-de-qi; jing-de-qi
attract 3-4		**absorb** (13)	xi-shou
	attract (20)	attract (11)	xi-yin
	appeal (15)	appeal (9)	xi-yin
	hook (4)		xi-yin, shang-yin
		interest (7)	xi-yin, shi…gan-xing-qu
celebrate 1-2	**celebrate** (6)		huan-qing
		greet (16)	wen-hou, huan-ying
		welcome (9)	huan-ying
challenge 1-1	challenge (9)	challenge (32)	tiao-zhan
check 5-3	**censor** (6)		shen-cha; shan-gai
	check (11)	check (16)	jian-cha; he-dui; kong-zhi
	examine (27)	examine (17)	jian-cha; diao-cha; shen-cha
	review (5)	review (18)	hui-gu; shen-shi; ping-lun
	screen (5)		shen-cha; jian-cha; zhen-bie
clean 3-4	clean (10)	clean (61)	qing-xi
	clear (8)	clear (8)	qing-li
		purify (10)	jing-hua
	wash (8)	wash (34)	chong-xi
compose 5-3		**compose** (6)	you … gou-cheng
	comprise (5)		you … gou-cheng, bao-han
	consist (12)	consist (13)	you … gou-cheng
	constitute (9)		gou-cheng, xing-cheng
	form (75)	form (26)	xing-cheng
	shape (7)		xing-cheng
conclude 1-1	conclude (29)	conclude (37)	jie-lun
conform 4-3	**coincide** (4)		qiao-he; chong-die
	conform (6)	conform (6)	zun-zhao; fu-he; yi-zhi
	comply (5)	comply (7)	zun-cong, shun-cong; zhao-ban
	obey (5)	obey (39)	zun-cong; zun-shou
contact 6-4	communicate (21)	communicate (22)	jiao-liu; jiao-ji; chuan-di
	contact (7)	contact (16)	lian-xi; jie-chu
	interact (10)		(hu-xiang) jiao-liu; ying-xiang
	negotiate (6)		xie-shang; tan-pan; yi-ding
	react (16)	react (6)	fan-ying; zuo-yong
	respond (18)		hui-da; xiang-ying; fan-ying
		telephone (9)	gei…da-dian-hua
continue 6-5	continue (163)	continue (48)	ji-xu; lian-xu; chi-xu
	extend (18)	extend (10)	yan-chang; yan-shen; kuo-zhan
	further (6)		cu-jin; tui-dong
	last (24)	last (21)	chi-xu; chi-jiu; jian-chi; zhi-cheng
	proceed (9)	proceed (9)	ji-xu-jin-xing; zhuo-shou
	prolong (4)	prolong (6)	yan-chang; la-chang; tuo-yan
cost 1-1	cost (36)	cost (26)	hua … qian/shi-jian
delay 3-1	delay (9)	delay (15)	dan-wu; yan-wu; tui-chi
	hinder (7)		zu-zhi, zu-ai; fang-ai

	postpone (4)		tui-chi, yan-qi, yan-huan
dress 2-2	dress (22)	dress (10)	gei…chuan-yi-fu, zhuang-shi
	wear (42)	wear (24)	chuan-zhuo; chuan-po
exceed 1-1	exceed (6)	exceed (6)	chao-chu
experience 8-4	bear (21)	bear (7)	ren-shou; cheng-shou
	endure (14)		ren-shou; ren-nai
	experience (62)	experience (40)	ti-yan; gan-shou; zao-shou; jingli
	inflict (11)		shi…zao-shou (tong-ku, shang-hai)
	stand (66)	stand (88)	ren-shou; jing-shou; ding-zhu
	suffer (82)	suffer (35)	ren-shou; zao-shou; meng-shou
	tolerate(6)		ren-shou; rong-ren; kuan-rong
	undergo (15)		jing-li, jing-shou; ren-shou
experiment 2-2	experiment (8)	experiment (5)	shi-yan (实验), shi-yan (试验)
	test (30)	test (18)	shi-yan (试验); ce-shi
face 2-2	face (98)	face (22)	mian-dui
	confront (17)	confront (10)	mian-dui
fit 4-3	belong (26)	belong (21)	he-shi; shi-yi; shi-he-yu
	fit (31)	fit (116)	shi-he; fu-he; shi-ying
	match (10)		he…pi-pei; he…xiang-chen
	suit (8)	suit (45)	shi-he; shi-yi; shi-ying
follow 5-5	chase (14)	chase (4)	zhui-zhu, zhui-gan; zhui-qiu
	copy (6)	copy (4)	mo-fang; chao-xi; fu-zhi
	follow (110)	follow (91)	zhui-sui; gen-sui; zhui-qiu
		imitate (4)	mo-fang, xiao-fang; wei-zao
	pursue (18)	pursue (30)	zhui-sui; zhui-qiu
	trace (5)		zhui-zong, gen-zong
force 6-2	**compel** (4)		qiang-po
	drive (40)	drive (72)	qu-shi
	force (82)	force (22)	qiang-po, bi-po
	impose (24)		bai…qiang-jia-yu; zheng-shui
	oblige (12)		gan-ji
	rape (16)		qiang-jian
forgive 1-0	**forgive** (8)		yuan-liang
fuse 2-0	**dissolve** (10)		rong-jie; rong-hua; jie-san
	fuse (4)		rong-he; rong-hua
help 8-2	**aid** (14)		yuan-zhu, bang-zhu; zi-zhu
	assist (16)		bang-zhu
	back (14)		zhi-chi
	endorse (8)		zhi-chi; zan-tong; ren-ke
	fund (5)		zi-zhu
	help (198)	help (343)	bang-zhu; yuan-zhu
	sponsor (4)		zi-zhu; zan-zhu
	support (127)	support (27)	zhi-chi; fu-chi; yuan-zhu
insist 3-3	insist (19)	insist (48)	jian-chi; jian-chi-zhu-zhang
	persist (6)	persist (6)	jian-chi; zhi-yi; chi-xu; cun-liu
	stick (13)	stick (37)	jian-chi
lack 1-1	lack (26)	lack (16)	que-shao; que-fa
locate 2-0	**locate** (7)		ding-wei, wei-yu
	map (4)		hui-zhi … de-di-tu
measure 5-1	**calculate** (8)		gu-ji; yu-ji; tui-ce; ji-suan
		estimate (4)	gu-ji; gu-liang; ping-jia
	measure (10)		ce-liang; ji-liang; heng-liang
	rate (4)		dui…gu-jia; dui…ping-jia
	score (9)		(ti-yu-bi-sai-zhong) [de-fen, ji-fen]
	weigh (13)		cheng…de-zhong-liang; quan-heng
numb 1-0	**desensitize** (5)		shi…ma-mu
originate 4-2	base (75)	base (112)	ba…jian-li-zai…de-ji-chu-shang
	derive (9)	derive (7)	qi-yuan; you-lai; pai-sheng-chu

	originate (4)		fa-yuan; chan-sheng; fa-qi
	stem (14)		qi-yuan; fa-sheng
park 0-1		**park** (4)	ting-che
pretend 1-0	**pretend** (4)		jia-zhuang
punish 1-1	punish (21)	punish (92)	cheng-fa
range 2-1	range (9)	range (6)	zai-yi-ding-fan-wei-nei-bian-hua
	rank (8)		gei…fen-deng; wei…pai-lie
read 1-2	read (79)	read (815)	yue-du
		skim (4)	cu-lue yue-du
rely 2-2	depend (24)	depend (73)	yi-kao, qu-jue
	rely (22)	rely (11)	yi-lai
require 2-2	need (285)	need (551)	xu-yao
	require (74)	require (74)	xu-yao, xu-qiu
return 3-1	return (38)	return (22)	hui-fu, gui-huan
	reverse (10)		dian-dao, fan-zhuan
	withdraw (9)		shou-hui, che-tui
revolve 2-3	**revolve** (8)		wei-rao
		ring (7)	huan-rao
		surround (7)	wei-rao
	turn (101)	turn (22)	wei-rao, xuan-zhuan
ride 1-1	ride (6)	ride (26)	qi; qi-ma; qi-che
risk 5-2	**bet** (5)		du-bo
	endanger (8)	endanger (7)	wei-ji shi…zao-shou-wei-xian
	gamble (5)		du-bo; tou-ji; mao-xian
	risk (12)	risk (9)	mao-xian; mao…de-wei-xian
	venture (6)		mao…de-wei-xian; na…zuo-du-zhu
send 5-2	deliver (8)	deliver (5)	tou-di; yun-zai; ti-gülong
	send (38)	send (53)	fa-song; ji
	submit (9)		cheng-di; ti-jiao
	transmit (14)		chuan-song; shu-song; chuan-di
	transport (15)		yun-shu, yun-song; shu-song
serve 1-1	serve (70)	serve (176)	fu-wu
share 1-1	share (35)	share (10)	fen-xiang
sympathise 2-0	**sympathise** (12)		tong-qing
	sympathize (7)		tong-qing
tax 1-0	**tax** (4)		dui…zheng-shui
thank 1-1	thank (10)	thank (5)	gan-ji, gan-xie
trust 2-2	believe (365)	believe (298)	xiang-xin; ren-wei
	trust (16)	trust (15)	xin-ren; xin-lai
wait 2-1	**await** (6)		deng-dai
	wait (24)	wait (42)	deng-dai

The verbs that occur only in LOCNESS (61) are singled out as follows (see Figure 4.12):

Figure 4. 12 The verb lemmas that occur only in LOCNESS in Table 4.9

aid assist await back bet calculate celebrate censor coincide compel comprise constitute desensitize dissolve endorse endure forgive fund further fuse gamble hinder hook impact impose inflict interact locate map match measure negotiate numb oblige originate postpone pretend rank rape rate respond reverse revolve score screen shape sponsor stem subject submit sympathise sympathize tax tolerate trace transmit transport undergo venture weigh withdraw

There are two groups in which disparity between the two corpora is large. One is the 'help'

group and the other is the 'measure' group. In the *HELP* group, the learners are using only *help* and *support* whereas the NSs are not only using these two general words, but also using some specific words such as *aid*, *assist*, *endorse*, *fund*, *sponsor*. While there are five verbs in LOCNESS (*calculate, measure, rate, score, weigh*), there is only one verb in COLEC (*estimate*). The 'teddy bear' principle is especially significant in the case of this group of verbs.

4.5 Research questions revisited and answered

After a long discussion about how these two verb lemma lists have been drawn up and how analytically the verb lemmas are grouped in the previous sections, it seems that there is a need to revisit the research questions and see how well they have been addressed.

Question One: What is the range of verbs used in COLEC and what is the range of verbs used in LOCNESS?

According to the verb lemma lists (see Appendix 2 and Appendix 3, also see 4.3.3.2 above), there are 569 verb lemmas used in COLEC and 893 verb lemmas used in LOCNESS after a series of trimming and editing processes. Unsurprisingly, the NSs use a much wider range of verb lemmas than the learners do. Though numerically the disparity of the ranges between the two corpora is 325 words, it should be noted that the verbs used by the two groups of writers do not always match. Most of the time the verbs used by the NSs cover those used by the learners, but occasionally some verbs are used only by the learners.

Question Two: What is the similarity and disparity between the COLEC writers and the LOCNESS writers as far as verbs are concerned?

The similarity and disparity of the two corpora in terms of the use of verbs are expressed in the LOCNESS and COLEC columns in the tables above. Let me take the subclass 'help' in the miscellaneous group to summarise this presentation (see Table 4.10).

This table provides at least two important insights. Firstly, there exists a degree of similarity between the learner English and the NS in the use of verb lemmas: both groups of writers use

HELP and SUPPORT. Secondly, there also exists a degree of disparity between the two groups of writers. The NSs use more verb lemmas in this semantic field (including AID, ASSIST, ENDORSE, FUND, and SPONSOR). By using the bold font, the verbs that are used only in the NS corpus have been distinguished from those that are shared by the two groups.

Table 4. 10 The semantic field *help*

help 8-2	**aid** (14)		yuan-zhu, bang-zhu; zi-zhu
	assist (16)		bang-zhu
	back (14)		zhi-chi
	endorse (8)		zhi-chi; zan-tong; ren-ke
	fund (5)		zi-zhu
	help (198)	help (343)	bang-zhu; yuan-zhu
	sponsor (4)		zi-zhu; zan-zhu
	support (127)	support (27)	zhi-chi; fu-chi; yuan-zhu

Question Three: How many verbs are used only in LOCNESS and what are they?

There are 391 verbs that occur only in LOCNESS (see sections from 4.4.3.1 to 4.4.3.6). These verbs could be amalgamated in alphabetical order as follows (see Figure 4.13):

Figure 4. 13 An amalgamation of the verbs that occur only in LOCNESS

1	abandon	contrast	function	organise	roll
2	abolish	convey	fund	originate	ruin
3	abort	convict	further	outweigh	rule
4	abuse	convince	fuse	overhear	sack
5	accommodate	counteract	gamble	overlook	sacrifice
6	accompany	couple	gather	override	safeguard
7	accuse	criticise	generate	pack	sail
8	acknowledge	criticize	govern	parody	scare
9	address	crush	grant	partake	score
10	administer	culminate	guarantee	perceive	screen
11	advocate	curtail	guide	perpetuate	secure
12	affirm	date	haunt	persuade	segregate
13	aid	debate	highlight	pillage	sentence
14	align	deceive	hinder	plague	separate
15	alleviate	deem	hire	pool	shake
16	allocate	defy	honour	portray	shape
17	amount	degrade	hook	pose	shift
18	analyse	demonstrate	house	postpone	shock
19	anger	depict	impact	pray	shun
20	announce	deprive	implement	preach	sign
21	appoint	describe	implicate	prejudice	signify
22	argue	desensitize	import	prescribe	slaughter
23	assert	deserve	impose	press	slip
24	assist	design	induce	pretend	spark
25	assume	detect	inflict	prevail	split

26	assure	deter	infringe	print	sponsor
27	attack	diagnose	inhibit	proclaim	spring
28	attribute	dictate	inject	profess	state
29	await	direct	insert	profit	stem
30	award	disagree	inspire	program	stress
31	back	disclose	install	project	strip
32	balance	discriminate	institute	propose	strive
33	bar	dismiss	instruct	prosper	stumble
34	behave	dispute	insure	protest	subject
35	bet	disregard	integrate	provoke	submit
36	betray	disrupt	interact	publicise	sue
37	bind	dissolve	interfere	question	suspect
38	block	distort	interpret	quote	sustain
39	bond	distribute	intervene	race	sway
40	bother	diversify	interview	rank	switch
41	bounce	divorce	invest	rape	symbolise
42	breed	dominate	invoke	rate	sympathise
43	bump	donate	isolate	realise	sympathize
44	bury	doubt	issue	rear	tackle
45	calculate	dump	justify	reason	tamper
46	cap	ease	kick	rebel	tap
47	cast	embark	kiss	recognise	tax
48	cease	embody	label	reconcile	televise
49	celebrate	emit	legalise	refute	term
50	censor	emphasise	legalize	regain	thrive
51	center	employ	legislate	register	tie
52	centre	enact	lend	regulate	time
53	charge	encounter	lessen	reign	tolerate
54	cite	endorse	load	reinforce	trace
55	clarify	endure	locate	reinstate	transmit
56	classify	enforce	maintain	relinquish	transport
57	coin	entail	manifest	remove	undergo
58	coincide	entertain	manipulate	render	undermine
59	combat	envisage	map	renew	unify
60	comfort	epitomise	mark	repeal	uphold
61	comment	eradicate	match	repent	upset
62	compel	erase	mature	replace	value
63	compensate	erode	measure	reply	venture
64	comprehend	evoke	misuse	rescue	veto
65	comprise	exacerbate	mix	resent	view
66	compromise	exclude	mock	reserve	violate
67	condemn	execute	modify	resort	voice
68	condone	exemplify	monitor	respond	vote
69	confer	exercise	motivate	restore	weaken
70	confess	exploit	murder	restrict	weigh
71	conflict	extract	negotiate	retain	whip
72	conquer	facilitate	nominate	reunite	withdraw
73	consent	favour	note	reverse	witness

74	constitute	figure	numb	revoke	worsen
75	consult	file	nurture	revolt	worship
76	contemplate	flee	object	revolve	
77	contend	fool	oblige	rid	
78	contract	forgive	offend	ridicule	
79	contradict	frighten	offset	rip	

Question Four: How could the research findings based on the previous three questions be used for the improvement of ELT?

After the verb lemmas are grouped according to certain relationships between each other, there is an added value to the verb lemma lists. The English teacher and the writer of teaching materials are now equipped with information concerning the real English level of the learners, so that they can rely on real data and set up their goals and plans to improve the vocabulary repertoire of the learners. Actions taken by the teacher and the teaching material writer may be expected to meet the needs of the learners, since they will be based on real data from the learner corpus, rather on wild speculation as in the past. The COLEC writers (and other learners with the same background) may consult the comparative tables for the verbs that require their attention and practice if they wish their English to be native-like.

The use of the Chinese pin-yin in the sense grouping provides a semantic link between the L1 verb and the L2 verb for the learners. This is crucial for the learners because they will have a rough idea of how many new verbs they need to learn in a particular sense group and what they are. Take the 'help' example again (see Table 4.10). A glance at the first two columns will tell them that they have five new verbs to learn in this sense group and that these are *AID*, *ASSIST*, *ENDORSE*, *FUND*, and *SPONSOR*. What is more important, the learners' familiarity with *HELP* and *SUPPORT* is expected to serve as a bridge between the known and the unknown. By associating the known (*HELP* and *SUPPORT*) with the unknown (*AID*, *ASSIST*, *ENDORSE*, *FUND*, and *SPONSOR*), the learners have a better chance to memorise the new verbs in an easier way. Furthermore, by looking at all the verbs used in the sense group, supported by the Chinese pin-yin, the learners may relate the new verbs with their L2, which is expected to help them with memorisation.

It is apparent that a comparison of the verb lemmas used by the NSs and the NNSs provides

useful information for the learning of English. However, this does not mean that learners should copy the use of the NSs strictly. Some verbs are best not included in the next phase of the syllabus. The teacher should use his or her intuition to come up with a sound judgement on some occasions. For example, the verb *FLOG* appears in the production vocabulary only of the NSs. But if we apply rigidly our standard for inclusion as mentioned above (frequency ≥ 3), it should be included in the verb lemma lists. I have deleted this verb from the lists (see Appendix 3) because it is too seriously restricted to the topic of the text and does not make a good goal for the learners in vocabulary learning. When teachers and course designers decide on a vocabulary list for their students to practise, they need to make corresponding changes according to the aim of their teaching.

4.6 Conclusion

This chapter has shown two things. First, it has demonstrated how to make verb lemma lists out of a learner corpus and a NS corpus via a corpus linguistic approach; and second, how to make the fullest use of these verb lemma lists. Some practical issues concerning the use of verbs by learners are addressed. Formerly neither a pure NS description of the language use nor a pure NNS analysis of learners' interlanguage (as in error analysis and SLA) could account for the similarity and disparity in language use between a learner group and a NS group. Now this comparative study of learner English and the NS English, supported by the modern technology of corpus linguistics, has made this possible. Once the information with regard to the range of the learners' vocabulary of verbs is available to the researcher, the teacher, the learner, the writer of teaching materials and other ELT practitioners, the learners' needs in vocabulary enlargement are no longer the subject of wild speculation. It is expected that teaching activities based on this information will prove to be more efficient, and more to the point.

Meanwhile it should be noticed that even though there is rich and important information that can be taken from verb lemma lists, some things cannot be detected from lists alone; in other words, there are questions that this research cannot answer. Do two verb lemmas that are used roughly to the same extent in the two corpora behave similarly in syntax? Does high frequency (as in words like *TAKE* and *KEEP*) guarantee native-like performance by the learners? In cases where a large range of senses is used by the NSs, how many senses are used

by the learners? For a polysemous verb, do the learners use the same sense or senses as the NSs do? If not, which sense is used by the learners and which by the NSs? How can the verbs in the tables in this chapter be related to the actual uses in the corpora? All the points above deserve examination and they will be discussed at full length in later chapters.

Chapter Five

Verbs in Different Forms Compared

5.1 Introduction

The previous chapter has worked out how many verb lemmas are used in the two corpora and what they are. The significance of undertaking such a task is to show the difference between the COLEC writers and the LOCNESS writers as far as verb lemmas are concerned. The result suggests that if the COLEC writers wished to enrich their production vocabulary, they could learn to use all the forms of all the lemmas that they currently do not share. But the problem the learners face is which form they should start to practise to use first. One important thing detected by corpus linguists concerning the use of verb forms and lemmas is that different forms of verbs are used so differently that they effectively constitute different 'lemmas' (see Sinclair 1991, Stubbs 2001; Sinclair and Renouf 1998). Therefore, it is essential to know which form of which verb is used frequently in the NS corpus so that learners learn to use the right form of the right verb. In other words, it is not a sufficient study if it provides only a list of lemmas of disparity; different form distribution in the two corpora must be examined so that more efficient use may be made of the study in Chapter Four. Suggestions derived from the verb lemma lists in Chapter Four will become misleading if detailed information concerning the detailed use of different forms is missing. To tackle this problem, the distribution of different forms of verbs should be investigated at full length. In a preliminary look at the uneven distribution of the different forms of verbs by COLEC writers, compared with the performance of the LOCNESS writers, I observed that there is a sharp difference between the two groups of writers in using different inflectional forms of verbs (Guo 2003). In that research, it was found that the COLEC writers use the base form more than the other forms compared with the NS writers. Following those findings, this chapter examines in much more detail the distribution of different word forms of verbs, attempting to answer the following research questions:

(1) What is the total distribution of occurrences of different forms of verbs in COLEC and LOCNESS?

(2) Do different forms of verbs behave similarly in NS English? Is the learner English

similar to the distribution of different forms of verbs by the NSs?

(3) Are there any differences in the top 20 verb forms in COLEC and LOCNESS in terms of types? If yes, what are they?

(4) Is there a degree of familiarity in the learner English with different forms of verbs? If there is, what is the order, from more familiar to less familiar?

(5) How does the disparity of topics affect the CIA research?

(6) What is the significance of the findings above?

5.2 A general view of the total frequency of the different forms of verbs

Before I start to examine the details of the distribution of verb forms of individual verbs, it is useful to have a look at the overall frequency of the different forms of verbs in the two corpora. Based on the verb lemma lists created in Chapter Four (see Appendix 2 and Appendix 3), it is possible to reach the figures shown in Table 5.1 and Table 5.2. When we look at the base form frequency in the two corpora, a dramatic disparity emerges. Whereas only 44 percent of verb forms are the base form in LOCNESS, as many as 68 percent of the verb forms are in the base form in COLEC. Since the learners have been using the base form dominantly compared with the other forms, it is natural that they would use all the other forms in a much smaller percentage than the LOCNESS writers.

Table 5. 1 The raw frequency and the percentage of each form of verbs in COLEC

Lemma	V-e	V-s	V-ing	V-ed	V-n	Total
Total	36886	4032	4805	3763	4935	54421
N Total[35]	38418	4199	5005	3919	5140	56681
Percentage	68	7	9	7	9	100

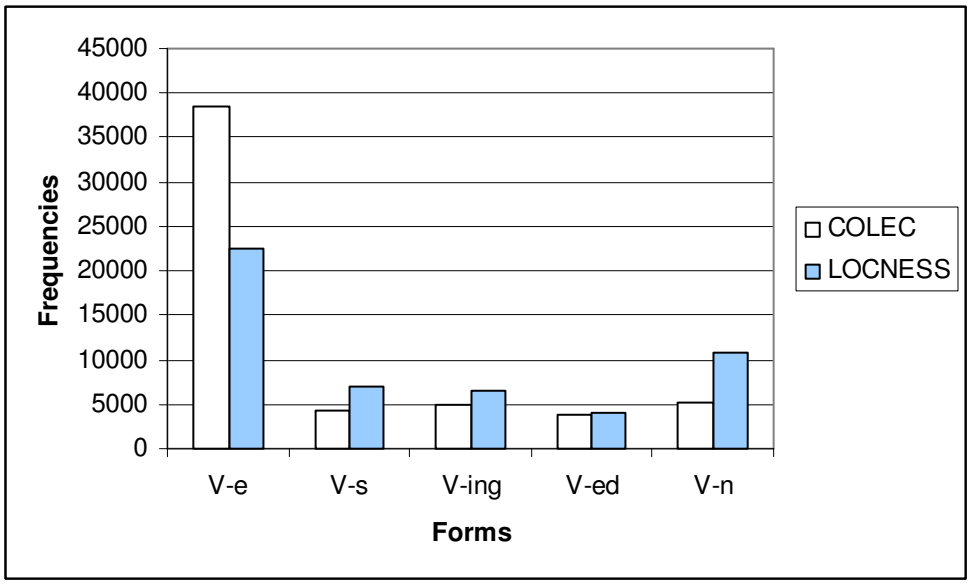
Table 5. 2 The raw frequency and the percentage of each form of verbs in LOCNESS

Lemma	V-e	V-s	V-ing	V-ed	V-n	Total
Total	14534	4520	4234	2672	7026	32986
N Total	22536	7009	6565	4143	10894	51147
Percentage	44	14	13	8	21	100

The overuse of the base form by the learners is shown by a bar chart in Figure 5.1.

35 'N' stands for 'normalised'. Throughout the research, normalised frequencies are obtained by the following formula: Normalised frequency = (raw frequency X 500,000)/total tokens of the corpus. The total tokens of COLEC = 480063 and the total tokens of LOCNESS = 322462.

Figure 5. 1 A bar chart of the normalised frequencies of the verb forms in COLEC and LOCNESS



The sharp and sudden fall from the base form to the V-s form in COLEC is clearly portrayed. Another obvious disparity lies in the V-n form which the NSs use much more than the learners. The learners use fewer V-s forms and V-ing forms but to a lesser extent than in the V-n forms. As far as the V-ed form is concerned, the learners use approximately the same number of forms as the NSs. A view of the distribution as shown above provides only a general picture of the distribution of each form in all the forms of verb use. This answers only the first research question in 5.1, by displaying the total distribution of occurrences of different forms of verbs in COLEC and LOCNESS. To answer the other questions, further explorations are needed.

5.3 The top 20 verbs in their different forms in LOCNESS and COLEC

Because we know that different forms behave differently, it is not expected that the most frequent verbs occurring as base forms, for example, will belong to the same lemmas as those occurring as V-s forms. By the same token, the most frequently used verbs occurring as the V-ing forms will not belong to the same lemmas which occur as V-ed forms. As far as I know, the disparity between the different forms of verbs has not been studied before. To compare the top 20 word forms in the two corpora helps us to see the disparity between the two groups of writers. The five forms of the top 20 verbs will be listed and compared in the following sections.

5.3.1 The top 20 verbs in their different forms in LOCNESS

In order to see how verbs behave differently in terms of frequency the top 20 word forms of LOCNESS have been extracted from the verb lemma list of LOCNESS (Appendix 3) as Table 5.3 shows. In this table the verbs occurring in all the five forms are represented by regular bold font, four forms by italicised bold, three forms by regular italicised, two forms by normal underlined, and one form by regular font. It is observable that among the 20 verbs occurring in the base form, there are only six for which all forms occur, i.e. *MAKE*, *TAKE*, *BECOME*, *USE*, *SAY*, and *GIVE*. Three of the remaining 14 verbs have four forms (*GO*, *SHOW* and *SEE*), 10 verbs have three forms (*FEEL*, *GET*, *THINK*, *BELIEVE*, *WANT*, *KNOW*, *FIND*, *NEED, COME* and *ALLOW*), four verbs have two forms (*SEEM*, *MEAN*, *LIVE*, and *LEAD*) and 20 verbs have only one form (*STATE, TELL, TRY, WORK, LOOK, RUN, KILL, PLAY, FIGHT, READ, BEGIN, START, CHOOSE, CHANGE, CONSIDER, LEAVE, BASE, PUT, BRING*, and *FORCE*). In the nine verbs with only three forms, the distribution is not identical from verb to verb. For some verbs, such as *THINK* and *GET*, the V-s form does not occur in the top 20 of the list and for others such as *FEEL* and *WANT*, the V-ing form does not occur in the top 20.

Table 5. 3 The distribution of the top 20 verbs in their different forms in LOCNESS

Lemma	V-e	Lemma	V-s	Lemma	V-ing	Lemma	V-ed	Lemma	V-n
make	426	seem	141	**make**	129	**make**	88	**make**	231
see	306	**make**	113	try	120	*come*	79	***see***	219
take	289	**say**	110	**take**	111	**say**	76	**use**	190
feel	280	*want*	105	**use**	96	*want*	71	**give**	137
get	275	***go***	91	***go***	79	**become**	60	**take**	132
think	237	*come*	84	**become**	75	**take**	59	*allow*	95
believe	220	mean	81	**say**	68	*feel*	57	**become**	86
want	215	***show***	79	*get*	64	begin	52	***show***	83
become	209	**take**	76	**give**	61	*believe*	41	change	83
go	201	*believe*	75	live	60	**give**	40	*find*	75
use	198	state	74	work	51	***see***	35	consider	73
know	193	*feel*	70	look	43	start	35	leave	70
say	178	**become**	69	*allow*	42	***go***	34	*need*	65
find	165	*need*	61	run	42	*think*	33	base	65
give	164	**use**	52	kill	40	mean	33	*know*	64
live	143	**give**	51	***show***	39	*find*	32	put	64
show	134	lead	49	*think*	34	*get*	31	bring	62
need	131	***see***	48	play	33	*know*	29	**say**	61
seem	128	*allow*	46	fight	33	**use**	27	lead	61
come	121	tell	40	read	33	choose	27	force	59

In summary, the distribution of the verb forms bears out the theories referred to above, that

verbs do not behave in the same way from form to form. Otherwise, the top 20 verbs under study should have all the five forms used rather than some being used in three forms, some in two forms and some in only one form. This result shows that different forms of verbs behave differently in NS English in terms of frequency. The next section turns to the NNS English and sees whether the learner language production resembles the distribution pattern of the different forms of the top 20 verbs.

5.3.2 The top 20 verbs in their different forms in COLEC

To compare the top 20 verb forms used in the two corpora, a summary table is made (Table 5.5). A striking feature of the top 20 verbs in their different forms in COLEC is that there exists much less homogeneity and uniformity among the COLEC writers as a whole than among the LOCNESS writers. In 5.3.1, it is found that in LOCNESS there are as many as six verbs in all forms, even though a sharp disparity exists among the different forms. In COLEC, however, there are only three verbs in all forms among the top 20 verbs, i.e. *MAKE*, *TAKE*, and *GET* (see Table 5.4 and Table 5.5). Five verbs occur in four forms (*KNOW*, *LEARN*, *GO*, *SAY* and *BECOME*), seven verbs in three forms (*THINK*, *WANT*, *FIND*, *USE*, *CHANGE*, *SEE* and *INCREASE*) and eight verbs in two forms (*LIKE*, *WORK*, *STUDY*, *READ*, *NEED*, *COME*, *TELL*, and *DEVELOP*). And as many as 28 verbs occur in only one form: *TRY, BUY, MEAN, SEEM, RAIN, CAUSE, BRING, LEAD, PLAY, SPEAK, WATCH, LISTEN, FACE, LIVE, LOOK, ASK, DIE, JUMP, WALK, LOSE, BEGIN, IMPROVE, POLLUTE, BASE, LIMIT, DECREASE, GRANT* and *RUSH*.

It is not difficult to imagine that if writers resemble each other in production, there should be fewer types that do not match each other in a particular range of the entire lexicon they collectively have. In other words, the more the writers share a pattern in using verbs, the fewer verb types there should be. There are fewer types in the NSs corpus than the learner corpus, which also suggests that the NSs are more like each other in their written production than the learners are, i.e. 43 vs. 51 (compare Column A of the two corpora of Table 5.5). The same feature is also reflected in the number of verbs that are used across all the verb forms. As many as six verbs occur in all their five forms in LOCNESS whereas there are only three such verbs in COLEC. Since there are fewer verb forms that belong to the same lemmas in the

learner corpus, there must be more word forms that are used alone. There are as many as 28 verbs in only one form in COLEC whereas there are only 20 in LOCNESS.

Table 5. 4 The distribution of the top 20 verbs in their different forms in COLEC

Lemma	V-e	Lemma	V-s	Lemma	V-ing	Lemma	V-ed	Lemma	V-n
know	2565	**make**	1623	read	262	***say***	171	*use*	414
think	1940	mean	284	**get**	242	**get**	125	*increase*	159
make	1877	need	127	study	241	**make**	112	improve	155
get	1821	*want*	93	***learn***	194	***go***	111	**make**	149
learn	1262	like	78	*increase*	153	*find*	97	pollute	137
want	1154	***become***	77	work	147	*want*	90	***know***	136
take	900	***go***	62	*change*	146	*think*	80	**take**	118
find	869	***say***	61	play	127	***become***	77	*change*	108
like	841	come	56	*use*	120	tell	72	base	107
use	769	**take**	51	develop	114	come	69	limit	101
change	712	seem	48	speak	114	look	68	***learn***	97
go	676	***know***	46	watch	110	ask	68	**get**	94
work	615	rain	44	**take**	100	**take**	62	decrease	79
study	569	tell	41	**make**	95	***learn***	61	grant	67
read	496	cause	40	listen	91	*see*	58	develop	66
see	397	*increase*	37	***go***	88	die	57	***say***	64
try	395	**get**	34	***know***	79	jump	57	***become***	63
say	381	bring	34	face	69	walk	56	rush	61
need	365	*think*	33	***become***	66	lose	50	*find*	51
buy	356	lead	32	live	64	begin	50	*see*	49

All the contrasts shown above could be interpreted as meaning that the learners do not possess a common knowledge about how to use the English language. This also suggests that while the NSs' English seems to have a solid structure and patterning, the learner English seems to have a very fluid status, which is perfectly reasonable because they are learners and everybody is somewhere half way along the route of L2 acquisition.

Table 5. 5 A summary of the distribution of the top 20 verbs in their different forms in LOCNESS and COLEC (A = types; B = tokens)

	LOCNESS		**COLEC**	
	A	B	A	B
5 forms	6	30	3	15
4 forms	3	12	5	20
3 forms	10	30	7	21
2 forms	4	8	8	16
1 form	20	20	28	28
Total	**43**	**100**	**51**	**100**

Since it is natural that LOCNESS writers have a lot in common in language use, the

distribution of the forms of the top 20 verbs have reflected the congruity and uniformity of this homogeneous group of NSs.

This finding endorses a psycholinguistic word association test done by Meara (1982), in which some NSs and some NNSs were given a stimulus and then asked to write down the associated words that immediately came into their mind. The following paragraph is Meara's report (1982: 30):

> The word associations produced by non-native speakers differ fairly systematically from those produced by native speakers. Surprisingly, learners' responses tend to be more varied and less homogeneous than the responses of the comparable group of native speakers. This is an odd finding because learners must have a smaller, more limited vocabulary than native speakers, and this might lead one to expect a more limited range of possible responses. Learner responses are not generally restricted to a subset of the more common responses made by native speakers, however. On the contrary, learners consistently produce responses which never appear among those made by native speakers, and in extreme cases, it is possible to find instances of stimulus words for which the list of native speaker and learner responses share practically no words in common.

The association test by Meara is different in nature from the study of verb forms. But the underlying principle that governs collective learner English production is identical.

5.4 The different forms of the top 20 verbs compared

An important purpose of this dissertation is to find the gap between the current learner English as an aggregated whole and the NS English which is considered to be the target for the learners. It is expected that a comparative view of the distribution of the forms of the top 20 verb forms in the two corpora would reveal much information for the teacher and others it may concern. This section looks at five verb forms, the base form (V-e) (including the finite form and the infinitive form), the third person singular form (V-s), the V-ing form (V-ing) (disregarding the distinction between the gerund and the present participle), the past form (V-ed) and the past participle (V-n). It should be pointed out that the frequencies in this section are based on the verb lemma lists of Appendix 2 and Appendix 3 (not directly extracted from the POS-tagged corpora by WordSmith).

5.4.1 The V-e forms of the top 20 verbs in the two corpora compared

The top 20 verb forms in the two corpora are easily accessible by using the sort function of MS Excel. Table 5.6 shows the most often used 20 verbs in the base form in the two corpora. The word forms that only occur in one of the corpora are highlighted.

There are 13 verbs that overlap in the two corpora, i.e. *make, see, take, get, think, want, become, go, use, know, say, find* and *need*. Because these verbs stand a better chance of being considered to be fairly mastered in the English produced by the COLEC writers (but not necessarily so; further examination of the concordances will be needed to ascertain that these verb forms are part of the learner language production capability). If we assume that what overlaps in the two corpora is truly a part of the learners' production capability, then what is more important is to know the verbs that are not shared by the COLEC writers. As Figure 5.2 shows (also highlighted in Table 5.6), there are seven verbs in their base form that are not shared by the COLEC writers.

Table 5. 6 The top 20 base forms (V-e) in LOCNESS and COLEC

S N	LOCNESS		COLEC	
	Lemma	**V-e**	**Lemma**	**V-e**
1	make	426	know	2565
2	see	306	think	1940
3	take	289	make	1877
4	feel	280	get	1821
5	get	275	learn	1262
6	think	237	want	1154
7	believe	220	take	900
8	want	215	find	869
9	become	209	like	841
10	go	201	use	769
11	use	198	change	712
12	know	193	go	676
13	say	178	work	615
14	find	165	study	569
15	give	164	read	496
16	live	143	see	397
17	show	134	try	395
18	need	131	say	381
19	seem	128	need	365
20	come	121	buy	356

Figure 5. 2 The verbs that are only found in LOCNESS in the top 20 V-e word forms

feel believe give live show seem come

In passing, it is found that the learners are over-concerned with *learn* (Top 5) and *study* (Top 14), suggesting the monotonous life of university students. The appearance of the verb *buy* (Top 20) must be a result of the discussion about buying fake commodities. Since there are fewer topics in COLEC than in LOCNESS, it seems that the learners' writing is more influenced by the topics than the NSs' writing. It is also noticeable that the most often used verb in COLEC (*know*, 2565) is used six times as many times as the most often used in LOCNESS (*make*, 426); and even at the end of the list, the twentieth verb used in COLEC (*buy*, 356) is used three times as many times as that used in LOCNESS (*come*, 121) before normalisation. This indicates that the learners are overusing a certain small number of verbs tremendously and these verbs are playing a too important role for the learners, who have a limited repertoire of verbs.

5.4.2 The V-s forms of the top 20 verbs in the two corpora compared

Table 5. 7 The top 20 third person singular forms (V-s) in LOCNESS and COLEC

SN	LOCNESS		COLEC	
	Lemma	V-s	Lemma	V-s
1	seem	141	make	1623
2	make	113	mean	284
3	say	110	need	127
4	want	105	want	93
5	go	91	like	78
6	come	84	become	77
7	mean	81	go	62
8	show	79	say	61
9	take	76	come	56
10	believe	75	take	51
11	state	74	seem	48
12	feel	70	know	46
13	become	69	rain	44
14	need	61	tell	41
15	use	52	cause	40
16	give	51	increase	37
17	lead	49	get	34
18	see	48	bring	34
19	allow	46	think	33
20	tell	40	lead	32

Among the 20 verbs in their V-s form, there are 12 that are shared by the two groups of writers (*seems, makes, says, wants, goes, comes, means, takes, becomes, needs, leads* and

tells) (see Table 5.7). The V-s forms that do not overlap and only occur in LOCNESS are displayed in Figure 5.3.

Figure 5. 3 The verbs that are only found in LOCNESS in the top 20 V-s word forms

shows believes states feels uses gives sees allows

It is noticeable that the most often used V-s form *seems* appears in both corpora, suggesting that the COLEC learners as a whole may have already learned how to use this verb. However, it is also noticeable that two important academic words *shows* and *states* (see Figure 5.3) are missing in COLEC. This seems to expose the disparity of the two corpora in text type and formality. The feature of a considerable number of use of academic vocabulary in LOCNESS corpus is mostly missing in COLEC.

Like the V-e form, the use of the top 20 verbs in their V-s form also reveals the influence of topics in the corpora. In COLEC for example, the topic of water shortage leads to the large number of uses of *rains* as a key word, and the topic of the increase in life expectancy causes the use of *increases* as a key word. In LOCNESS, for example, the literary essays concerning Camus, Caligula, Voltaire, Hugo, etc. result in a large number use of third person singular forms as in "Camus believes …", "Caligula feels …", "Voltaire uses …", and "Hugo sees …" and of course plenty of cases of the third person singular pronoun *he* in the places of the real names as quoted above as in "he believes …", "he feels …", "he uses …" and "he sees …". The abundant use of the third person singular form like this also seems to point to the NS proficiency in using the present tense to talk about literary works, authors and characters.

5.4.3 The V-ing forms of the top 20 verbs in the two corpora compared

As Table 5.8 shows, there are nine verbs in their V-ing form that are shared by the two groups of writers, i.e. *making, using, going, becoming, getting, living, working, playing* and *reading*. The 11 verbs in this form which are unique to LOCNESS are shown in Figure 5.4.

So far it has become observable that the composing topics have a strong impact upon the selection and production of verbs and verb forms even though their influence might not be universal. Take *killing* for example; both intuition and a cursory look will indicate that this verb form is strongly topic-sensitive. To distinguish those that are seriously influenced by

topics and those that are not is not easy but still possible. I have used the following technique to make such a distinction by using Scott's key words and key key words theory (Scott 1997, Scott and Tribble 2006).

Table 5. 8 The top 20 V-ing forms in LOCNESS and COLEC

SN	LOCNESS		COLEC	
	Lemma	V-ing	Lemma	V-ing
1	make	129	read	262
2	try	120	get	242
3	take	111	study	241
4	use	96	learn	194
5	go	79	increase	153
6	become	75	work	147
7	say	68	change	146
8	get	64	play	127
9	give	61	use	120
10	live	60	develop	114
11	work	51	speak	114
12	look	43	watch	110
13	allow	42	take	100
14	run	42	make	95
15	kill	40	listen	91
16	show	39	go	88
17	think	34	know	79
18	play	33	face	69
19	fight	33	become	66
20	read	33	live	64

Figure 5. 4 The verbs that are only found in LOCNESS in the top 20 V-ing word forms

trying taking saying giving looking allowing
running killing showing thinking fighting

Before describing this point, the function of re-sort in WordSmith (see Figure 3.3 in Chapter Three for a screenshot of re-sorting) is used to detect the dispersion situation in the whole corpus. When the concordances of *killing* are consulted and re-sorted by 'file', it is found that the word form mostly appears in four of the 14 files of LOCNESS. When the file record is checked, there are relevant topics found in the four files such as 'fox hunting', 'euthanasia', 'capital punishment', 'abortion', 'suicide', 'death penalty', and 'gun control'. With so many topics describing death and killing, it is inevitable that the verb *KILL* is a key word in these four files but not a key key word in the whole corpus. Another topic-sensitive verb form among the top 20 is *fighting*. When the concordance lines are re-sorted by file, it is found that this verb form is mainly used in five files which contain topics such as 'boxing', 'women in

combat' and 'racial prejudice'. This seems to suggest that treating the English in the NS corpus as a sort of goal is not without problems. Topic disparity must be considered fully.

5.4.4 The V-ed forms of the top 20 verbs in the two corpora compared

There are 12 shared V-ed forms in Table 5.9, i.e. *made, came, said, wanted, became, took, began, saw, went, thought, found* and *got*. The eight V-ed forms unique to LOCNESS are provided in Figure 5.5.

Attention should be drawn to the absence of *started* in the learners (it is as low-ranked as 94th in COLEC). The absence of *started* but the presence of *began* (ranked 20th) seems to suggest that if learners have one representative of one sense (such as *began* in this case), the chance of using alternatives will drop substantially. This is in conformity with the 'teddy bear' principle of learner English.

Table 5. 9 The top 20 V-ed forms in LOCNESS and COLEC

SN	LOCNESS		COLEC	
	Lemma	V-ed	Lemma	V-ed
1	make	88	say	171
2	come	79	get	125
3	say	76	make	112
4	want	71	go	111
5	become	60	find	97
6	take	59	want	90
7	feel	57	think	80
8	begin	52	become	77
9	believe	41	tell	72
10	give	40	come	69
11	see	35	look	68
12	start	35	ask	68
13	go	34	take	62
14	think	33	learn	61
15	mean	33	see	58
16	find	32	die	57
17	get	31	jump	57
18	know	29	walk	56
19	use	27	lose	50
20	choose	27	begin	50

Figure 5. 5 The verbs that are found only in LOCNESS in the top 20 V-ed word forms

felt believed gave started meant knew used chose

No obvious topic-sensitive V-ed forms are found in the top 20 in LOCNESS but two are found in COLEC, i.e. *died* and *jumped*. The use of *died* is related to topics such as infant mortality, water shortage and fake commodities. The occurrences of *jumped* almost without exception come from the same file of the corpus which is composed of free essays describing the same event in a story. This again leads to the necessity of topic and register control in the establishment of corpora which are to be compared.

5.4.5 The V-n forms of the top 20 verbs in the two corpora compared

There are nine V-n forms overlapping (*made, seen, used, taken, become, found, known, said,* and *got*) in the two corpora (see Table 5.10) and 11 V-n forms present only in LOCNESS (see Figure 5.6).

Table 5. 10 The top 20 V-n forms in LOCNESS and COLEC

SN	LOCNESS		COLEC	
	Lemma	V-n	Lemma	V-n
1	make	231	use	414
2	see	219	increase	159
3	use	190	improve	155
4	give	137	make	149
5	take	132	pollute	137
6	become	86	know	136
7	find	75	take	118
8	know	64	change	108
9	say	61	base	107
10	come	40	limit	101
11	go	37	learn	97
12	think	37	get	94
13	get	25	decrease	79
14	begin	23	grant	67
15	mean	23	develop	66
16	want	19	say	64
17	believe	16	become	63
18	feel	13	rush	61
19	start	11	find	51
20	choose	10	see	49

If we compare Figures 5.5 and 5.6, it is easy to find that as many as six verbs (lemmas) overlap in the V-ed form and the V-n form only in LOCNESS, i.e. *FEEL, BELIEVE, GIVE, MEAN, START* and *CHOOSE*. This could be interpreted as the homogeneity of the written English with a particular group (either the COLEC writers or the LOCNESS writers). Just as

the NSs use approximately the same verb lemma for these two forms, the NNSs do *not* use these verb lemmas for these two forms.

Figure 5. 6 The top 20 V-n forms in LOCNESS and COLEC

given come gone thought begun meant wanted believed felt started chosen

Considering the absence of *started* but the appearance of *began* as V-ed in the top 20 in COLEC, and the absence of both *started* and *begun* as V-n in the top 20 (and the rare use in the whole corpus as well) in COLEC, it seems that there is an order of familiarity with the different forms of verbs in the learners as a group. As far as the two verb forms of the two verbs (*BEGIN*, and *START*) are concerned, the V-n form is more unknown to the COLEC writers compared with the V-ed form. This issue will be further explored in 5.5.

Apart from these topic-sensitive words, there are two structure-sensitive words favoured by the Chinese students, *based* and *granted*. The V-n form *based* appears with the preposition *on* without a single exception. Likewise, the use of *granted* appears in the phrase '*TAKE* it for granted' without exception, indicating the possibility that the learners might know nothing about the word *BASE* and *GRANT* except such idiomatic expressions.

No obvious influence from the disparity in topics is detected in LOCNESS since all the top 20 V-n forms seem applicable to various kinds of topics. In COLEC, however, there are a number of topic-sensitive verb forms such as *increased, improved, polluted, decreased* and *rushed*.

5.4.6 Some summary remarks

The most useful information that could be taken from the top 20 verb forms in the two corpora might be the verb forms that are only used in LOCNESS (see Table 5.11). The importance of knowing this gap between the learners and the NSs is a first step for the learners to practise the most often used verb forms. This issue will be further discussed in 5.6.2.

In order for Table 5.11 to be interpreted easily, another table (Table 5.12) is created below. With the conversion of the data in Table 5.11, it is easier to see which verb lemmas (in the

'Word' column), and then which verb forms (from column *V-e* to *V-n*), are used only in the top 20 verbs in LOCNESS, and which verb lemmas have all the five forms, which have four forms, and so on, (from the 'Total' column). A clear profile of the learner English as regards the absence of the top verb forms compared with the LOCNESS writers is now available to readers.

Table 5. 11 The verb forms not shared by the COLEC writers in the top 20 verbs

	Word	Total	Texts	V-e	V-s	V-ing	V-ed	V-n
1	ALLOWING	1	1	0	0	1	0	0
2	ALLOWS	1	1	0	1	0	0	0
3	BEGUN	1	1	0	0	0	0	1
4	BELIEVE	1	1	1	0	0	0	0
5	BELIEVED	2	2	0	0	0	1	1
6	BELIEVES	1	1	0	1	0	0	0
7	CHOSE	1	1	0	0	0	1	0
8	CHOSEN	1	1	0	0	0	0	1
9	COME	2	2	1	0	0	0	1
10	FEEL	1	1	1	0	0	0	0
11	FEELS	1	1	0	1	0	0	0
12	FELT	2	2	0	0	0	1	1
13	FIGHTING	1	1	0	0	1	0	0
14	GAVE	1	1	0	0	0	1	0
15	GIVE	1	1	1	0	0	0	0
16	GIVEN	1	1	0	0	0	0	1
17	GIVES	1	1	0	1	0	0	0
18	GIVING	1	1	0	0	1	0	0
19	GONE	1	1	0	0	0	0	1
20	KILLING	1	1	0	0	1	0	0
21	KNEW	1	1	0	0	0	1	0
22	LIVE	1	1	1	0	0	0	0
23	LOOKING	1	1	0	0	1	0	0
24	MEANT	2	2	0	0	0	1	1
25	RUNNING	1	1	0	0	1	0	0
26	SAYING	1	1	0	0	1	0	0
27	SEEM	1	1	1	0	0	0	0
28	SEES	1	1	0	1	0	0	0
29	SHOW	1	1	1	0	0	0	0
30	SHOWING	1	1	0	0	1	0	0
31	SHOWS	1	1	0	1	0	0	0
32	STARTED	2	2	0	0	0	1	1
33	STATES	1	1	0	1	0	0	0
34	TAKING	1	1	0	0	1	0	0
35	THINKING	1	1	0	0	1	0	0
36	THOUGHT	1	1	0	0	0	0	1
37	TRYING	1	1	0	0	1	0	0
38	USED	1	1	0	0	0	1	0

39	USES	1	1	0	1	0	0	0
40	WANTED	1	1	0	0	0	0	1

In only 20 verb forms, however, the information that can be extracted is rather limited. It is likely that a particular form may not be in the top 20, but might be ranked in 21st position or a little after, in which case any judgement about the absence or presence of that form will be seriously biased. Therefore, there is a need to expand the perspective of investigation into more verbs in their different forms.

Table 5. 12 A summary of the verb forms that are not shared by the COLEC writers in the top 20 verbs

SN	Word	V-e	V-s	V-ing	V-ed	V-n	Total
1	GIVE	give	gives	giving	gave	given	5
2	BELIEVE	believe	believes	-----	believed	believed	4
3	FEEL	feel	feels	-----	felt	felt	4
4	SHOW	show	shows	showing	-----	-----	3
5	ALLOW	-----	allows	allowing	-----	-----	2
6	THINK	-----	-----	thinking	-----	thought	2
7	USE	-----	uses	-----	used	-----	2
8	COME	come	-----	-----	-----	come	2
9	CHOOSE	-----	-----	-----	chose	chosen	2
10	MEAN	-----	-----	-----	meant	meant	2
11	START	-----	-----	-----	started	started	2
12	LIVE	live	-----	-----	-----	-----	1
13	SEEM	seem	-----	-----	-----	-----	1
14	SEE	-----	sees	-----	-----	-----	1
15	STATE	-----	states	-----	-----	-----	1
16	FIGHT	-----	-----	fighting	-----	-----	1
17	KILL	-----	-----	killing	-----	-----	1
18	LOOK	-----	-----	looking	-----	-----	1
19	RUN	-----	-----	running	-----	-----	1
20	SAY	-----	-----	saying	-----	-----	1
21	TAKE	-----	-----	taking	-----	-----	1
22	TRY	-----	-----	trying	-----	-----	1
23	KNOW	-----	-----	-----	knew	-----	1
24	BEGIN	-----	-----	-----	-----	begun	1
25	GO	-----	-----	-----	-----	gone	1
26	WANT	-----	-----	-----	-----	wanted	1
	Total	7	8	11	8	11	45

Another area that deserves more examination is the base form of verbs. In making two verb lemma lists in Chapter Four, it was decided that the infinitive form and the finite form should be merged into one because the purpose of making the lists was to single the verbs out from the non-verbs (such as nouns, adjectives and prepositions) and there was no need to treat the

two forms separately. When we shift our attention to verbs, it is apparent that the two forms should be treated separately because they may function and perform differently in the NS English.

In the following section, therefore, six verb forms (instead of five) of all the verb forms occurring only in LOCNESS will be extracted. But this time the information will come from the POS-tagged COLEC and LOCNESS rather than the verb lemma lists as described in Chapter Four.

5.5 Examining the matched verb form lists

Even though the LOCNESS writers normally use more verbs (types) in their different forms than the learners, the verb forms they use do not necessarily cover all those used by the COLEC writers. The learners may occasionally use some verbs that do not occur in LOCNESS. Since NSs have a larger vocabulary as discussed in Chapter Four, it follows that there would be more individual verb forms and a longer list with a longer tail in LOCNESS than in COLEC. Since low frequencies do not lend sufficient confidence to our belief in the usefulness of any suggestions for language education purposes, verb lemmas (including all the forms) with a small frequency (≤3) may safely be ignored.

A matched verb form list could be obtained by using the matching function of WordList in WordSmith aided by Excel (the details are illustrated in Appendix 4). The Comparison menu in WordSmith has three functions. Each function has some particular advantages that others do not have. To better utilise the advantages of each function, see Appendix 4.

Table 5. 13 A sample of a matched list of V-n forms in COLEC and LOCNESS

WORD	TAG	FILE	TOTAL	COLEC	LOCNESS
SAT	V-n	2	4	2	2
SMASHED	V-n	2	4	2	2
TENDED	V-n	2	4	2	2
BANNED	V-n	1	53	0	53
PRESENTED	V-n	1	37	0	37
INTRODUCED	V-n	1	32	0	32

After editing, a matched list looks like Table 5.13, in which there are six columns, i.e. ‘WORD’, ‘TAG’, ‘FILE’, ‘TOTAL’, ‘COLEC’ and ‘LOCNESS’. The ‘WORD’ column

contains all the verbs of a particular form which is specified in the 'TAG' column. The number 'FILE' shows how many files (corpora) contain the word in the corresponding row. If a word in the 'WORD' column appears only in one file or corpus, the value of its 'FILE' column is '1' and if a word appears in both of the files, or corpora, the value is '2'. The 'COLEC' and 'LOCNESS' columns show the frequencies of the verb forms in the 'WORD' column. The 'TOTAL' column is the sum of the values of the COLEC and LOCNESS columns. Take the verb form ***banned*** for example, it is a **V-n** form and appears only in '**1**' file for 53 times and this file is **LOCNESS**.

Since the studies below (from 5.5.1 to 5.5.6) involve the verb forms that only occur in LOCNESS, the COLEC column is not needed; the value in the FILE column will be '1' all the time; the 'TOTAL' column is unnecessary because there is one value (in 'LOCNESS') only; the 'TAG' information will be listed at the top of the tables; there is no need to keep these four columns. The only two columns required will be the verb-form column (V-i in Table 5.14) and the frequency column (FRE). To save space, the long lists are chopped to make parallell columns side by side.

5.5.1 Matching the V-i form lists

CLAWS7 distinguishes the finite of a verb as in 'they argue' and the infinitive of a verb as in 'to argue' and 'could argue'. Table 5.14 contains all the infinitives that occur in LOCNESS only (V-i is the short form for the infinitive of a verb and V-e is the short form for the finite form of a verb).

Table 5. 14 All the V-i forms occurring only in LOCNESS (frequency ≥ 4)

N	V-i	FRE	V-i	FRE	V-i	FRE
1	ARGUE	39	DREAM	7	VOICE	5
2	JUSTIFY	23	EASE	7	ABUSE	4
3	DEFINE	22	EMPHASIZE	7	APPOINT	4
4	DETERMINE	21	FREE	7	ASSERT	4
5	REPRESENT	16	LEGISLATE	7	BEHAVE	4
6	COMBAT	15	LOOSE	7	CALCULATE	4
7	PERSUADE	15	REPEAL	7	CONVEY	4
8	REMOVE	15	TACKLE	7	COUNTERACT	4
9	STATE	15	WEIGH	7	CRITICIZE	4
10	REPENT	14	ACCOMMODATE	6	DIMINISH	4
11	REFLECT	14	ASSIST	6	DISMISS	4

12	SUGGEST	14	CONCIEVE	6	DISPUTE	4
13	DOMINATE	13	CONDEMN	6	DIVORCE	4
14	PRESENT	13	INTERFERE	6	DIVERSIFY	4
15	ASSUME	12	GOVERN	6	EMPLOY	4
16	CONTAIN	12	INFORM	6	FACILITATE	4
17	MATTER	12	LESSEN	6	FORGIVE	4
18	PRAY	12	MANIPULATE	6	FULFILL	4
19	QUESTION	12	NOTE	6	GATHER	4
20	CONTRACT	11	POSE	6	GUARANTEE	4
21	REGAIN	11	RECONCILE	6	HIRE	4
22	RETAIN	11	REFUTE	6	INHIBIT	4
23	COMPROMISE	10	REPLACE	6	INSPIRE	4
24	MENTION	10	SIGN	6	LEGALIZE	4
25	ATTRACT	9	TELEVISE	6	MIX	4
26	CEASE	9	THANK	6	MOCK	4
27	CONFESS	9	TRANSPORT	6	NEGOTIATE	4
28	IMPOSE	9	WAKE	6	OFFSET	4
29	MURDER	9	ALLEVIATE	5	OVERLOOK	4
30	SACRIFICE	9	CLARIFY	5	PERCEIVE	4
31	ATTEMPT	8	COMPENSATE	5	PROJECT	4
32	ABOLISH	8	COUNTER	5	REMEDY	4
33	CONVINCE	8	DESERVE	5	REPRODUCE	4
34	EVOKE	8	DIVIDE	5	RETIRE	4
35	HIDE	8	IMPLEMENT	5	REVERSE	4
36	MARRY	8	IMPLY	5	SACK	4
37	PROCESS	8	INTEGRATE	5	SAFEGUARD	4
38	REGULATE	8	INTERVENE	5	SCORE	4
39	RULE	8	PORTRAY	5	SECURE	4
40	SYMPATHISE	8	POINT	5	SHOUT	4
41	VOTE	8	PREVAIL	5	STEM	4
42	ADMIT	7	PROGRESS	5	SUBMIT	4
43	AID	7	RECIEVE	5	TRANSMIT	4
44	ADDRESS	7	REVOLT	5	UPHOLD	4
45	BALANCE	7	RUIN	5	WORSHIP	4
46	BACK	7	SOUND	5	WITNESS	4
47	COMPREHEND	7	TEND	5		
48	DISSOLVE	7	VETO	5		

5.5.2 Matching the V-e form lists

Table 5.15 shows all the base forms of verbs (non-infinitives) that are found only in LOCNESS. It seems that there is a difference between the V-i and V-e forms used. For example, there are only 15 V-e forms that overlap with the V-i forms among the 49 V-e forms (*argue, assist, define, deserve, emphasize, gather, justify, marry, perceive, portray, present, question, refute, replace* and *state*). In other words, the LOCNESS writers use much more

infinitives (types) than the COLEC writers.

Table 5. 15 All the V-e forms occurring only in LOCNESS (frequency ≥ 4)

	V-e	FRE	V-e	FRE
1	ARGUE	41	DISCUSS	5
2	STATE	21	EMPHASIZE	5
3	ALLOW	18	GATHER	5
4	DESERVE	18	INTERACT	5
5	CREATE	17	PERCEIVE	5
6	END	13	PROPOSE	5
7	PRESENT	13	QUESTION	5
8	VIEW	13	SEPARATE	5
9	DISAGREE	12	BASE	4
10	OUTWEIGH	9	BINGE	4
11	OPPOSE	9	CITE	4
12	GUESS	8	COMFORT	4
13	JUDGE	8	COMPETE	4
14	MOVE	8	DEFEND	4
15	DENY	7	DEMONSTRATE	4
16	RELY	7	DIFFER	4
17	SHARE	7	HINDER	4
18	TRAVEL	7	JUSTIFY	4
19	FIGHT	6	MARRY	4
20	ILLUSTRATE	6	PORTRAY	4
21	RESTRICT	6	REFUTE	4
22	ADOPT	5	REPLACE	4
23	ASSIST	5	RESENT	4
24	ASSEMBLE	5	RESPOND	4
25	DEFINE	5		

5.5.3 Matching the V-s form list

As shown in Table 5.16, there are 117 verbs in their V-s form occurring only in LOCNESS. The non-use of so many verbs in their V-s form by the learners reflects the disparity not only in text types but also in form selection preferences. Since the L1 of the learners does not distinguish the forms for singular subjects and plural subjects, it is envisaged that it would be difficult for the COLEC writers to choose the third person singular form properly. The absence in COLEC of verb forms such as *states*, *argues*, *describes*, *claims*, *maintains*, *demonstrates, assumes, asserts,* and *justifies* shows that the generally used verbs for argumentative essays are mostly missing in COLEC.

Table 5. 16 All the V-s forms occurring only in LOCNESS (frequency ≥ 4)

N	V-s	FRE	V-s	FRE	V-s	FRE
1	STATES	74	RAISES	9	EMPHASIZES	5
2	SEES	48	SUPPORTS	9	ENCOUNTERS	5
3	ALLOWS	46	TACKLES	9	EPITOMISES	5
4	REALISES	38	WISHES	9	FORMS	5
5	DECIDES	37	DESERVES	8	EVOKES	5
6	REJECTS	32	DEMONSTRATES	8	GUARANTEES	5
7	WRITES	29	ANNOUNCES	8	OPENS	5
8	ARGUES	29	DEFINES	8	PREACHES	5
9	STARTS	28	RELATES	8	RECOGNIZES	5
10	REMAINS	28	AIMS	7	RETAINS	5
11	REPRESENTS	27	ATTRIBUTES	7	TREATS	5
12	INVOLVES	24	ENTERS	7	ACHIEVES	4
13	CHOOSES	23	JUDGES	7	ACCUSES	4
14	REFUSES	23	RECOGNISES	7	ASSERTS	4
15	REALIZES	21	STAYS	7	ASSUMES	4
16	ADMITS	20	STOPS	7	COMPARES	4
17	DESCRIBES	20	TALKS	7	CONFRONTS	4
18	CLAIMS	19	ATTEMPTS	6	CONTROLS	4
19	CREATES	18	DRAWS	6	DESIRES	4
20	ATTACKS	15	EXPOSES	6	DETERMINES	4
21	ILLUSTRATES	15	ENTAILS	6	EARNS	4
22	ACCEPTS	14	EXPRESSES	6	ENDURES	4
23	DEALS	13	FIGHTS	6	ESPOUSES	4
24	DENIES	13	HITS	6	FEARS	4
25	RECEIVES	13	OUTWEIGHS	6	INSISTS	4
26	ARISES	12	PORTRAYS	6	INTENDS	4
27	FALLS	12	PROCLAIMS	6	JUSTIFIES	4
28	MOVES	12	PUSHES	6	MENTIONS	4
29	REVEALS	12	REINFORCES	6	PERSISTS	4
30	ACTS	11	REPLIES	6	PICKS	4
31	PLACES	11	REMARKS	6	PROMOTES	4
32	REFLECTS	11	TRAINS	6	PROVOKES	4
33	COMMITS	10	ADOPTS	5	RESPONDS	4
34	DISPLAYS	10	CONFESSES	5	SACRIFICES	4
35	HEARS	10	CONSTITUTES	5	SENDS	4
36	MANAGES	10	CRIES	5	SLEEPS	4
37	DISCOVERS	9	DEPICTS	5	STEMS	4
38	FOCUSES	9	DISAGREES	5	THROWS	4
39	MAINTAINS	9	DISCUSSES	5	UNDERGOES	4

5.5.4 Matching the V-ing form lists

There are altogether 115 V-ing forms in Table 5.17. The large number of the missing 'V-ing' form in COLEC might be caused by the learners unawareness that 'V-ing', apart from its use in the 'BE + V-ing' structure, can also be used in other structures such as the 'Verb + Noun'

structure (as in "*Allowing alchohol consumption* at age eighteen would change the way …") and the 'Noun + V-ing' structure (as in "If the government passes a *law allowing* them to drink …").

Table 5. 17 All the V-ing forms occurring only in LOCNESS (frequency ≥ 4)

N	V-ing	FRE	V-ing	FRE	V-ing	FRE
1	ALLOWING	42	EXPRESSING	7	REPRESENTING	5
2	CREATING	28	HURTING	7	REPENTING	5
3	LOWERING	22	PRAYING	7	REPLACING	5
4	BANNING	21	REACHING	7	ACKNOWLEDGING	4
5	STATING	21	REALISING	7	AFFECTING	4
6	COMMITTING	20	REMAINING	7	ANALYZING	4
7	ATTEMPTING	17	RETURNING	7	BEARING	4
8	INVOLVING	16	RETAINING	7	BETTING	4
9	BELIEVING	13	SEPARATING	7	CHASING	4
10	LETTING	13	VOTING	7	CONDEMNING	4
11	LEGALIZING	13	APPLYING	6	CONSUMING	4
12	PRESENTING	13	CONTRACTING	6	CROSSING	4
13	ARGUING	12	ELIMINATING	6	DATING	4
14	SUPPORTING	12	FULFILLING	6	DEFENDING	4
15	REVEALING	11	INTEGRATING	6	DESCRIBING	4
16	RESULTING	11	MAINTAINING	6	EMBRACING	4
17	CLAIMING	10	RELATING	6	EXPECTING	4
18	CONFESSING	10	PROMOTING	6	EXPOSING	4
19	DENYING	10	REFERRING	6	FUNCTIONING	4
20	REFUSING	10	RANGING	6	FURTHERING	4
21	DECIDING	9	SHARING	6	INVESTING	4
22	FORMING	9	BLAMING	5	MANIPULATING	4
23	OBTAINING	9	BEATING	5	MURDERING	4
24	OPPOSING	9	BELONGING	5	NURTURING	4
25	REJECTING	9	DAMAGING	5	POSSESSING	4
26	ACHIEVING	8	DEPRIVING	5	PORTRAYING	4
27	ENCOURAGING	8	DISCOVERING	5	REGARDING	4
28	EXPLAINING	8	ENSURING	5	REBELLING	4
29	FORCING	8	ENHANCING	5	RECYCLING	4
30	INTRODUCING	8	FOCUSING	5	REFLECTING	4
31	PROVING	8	FREEING	5	REGULATING	4
32	ADVOCATING	7	HIDING	5	SHAPING	4
33	ABOLISHING	7	KISSING	5	STRIVING	4
34	ADDRESSING	7	OFFERING	5	STRENGTHENING	4
35	ASSUMING	7	PERFORMING	5	TRANSMITTING	4
36	ATTACKING	7	PLACING	5	VIEWING	4
37	CONTAINING	7	PURCHASING	5	WITNESSING	4
38	DESTROYING	7	QUESTIONING	5		
39	DETERMINING	7	RACING	5		

5.5.5 Matching the V-ed form lists

There are only 42 V-ed forms in LOCNESS that are not shared by the COLEC writers (see Table 5.18), the smallest number of all the forms compared in this section (5.2.4).

Table 5. 18 All the V-ed forms occurring only in LOCNESS (frequency ≥ 4)

N	V-ed	FRE	V-ed	FRE
1	INVOLVED	21	ACCUSED	5
2	STATED	18	ALLOWED	5
3	REJECTED	12	ATTEMPTED	5
4	AROSE	11	DEMANDED	5
5	ARGUED	11	CONTRACTED	5
6	REALISED	11	FEARED	5
7	VIEWED	11	FOUGHT	5
8	AIMED	9	INTENDED	5
9	DESCRIBED	9	PLANNED	5
10	REMAINED	9	PROCEEDED	5
11	REPORTED	9	QUESTIONED	5
12	CREATED	8	RECOGNIZED	5
13	PRESENTED	8	STOLE	5
14	CLAIMED	7	ADMITTED	4
15	EXPRESSED	7	ASSUMED	4
16	INCLUDED	7	BANNED	4
17	SOUGHT	7	COMMITTED	4
18	WITNESSED	7	DEFINED	4
19	CONDUCTED	6	PROMISED	4
20	ESTABLISHED	6	RESIGNED	4
21	RULED	6	SIGNED	4

The number of the V-ed forms occurring in LOCNESS only is much lower than those of the others. This is in accordance with Table 5.1 and Table 5.2 in which the V-ed forms in both of the two corpora have a low percentage (7 percent in COLEC and 8 percent in LOCNESS). The small number of the V-ed form whose frequency is above 4 does not necessarily suggest that the learners perform better in this form than in other forms. On one hand, the large number of argumentative essays in LOCNESS does not require too many verb forms describing actions or states in the past. On the other hand, the low percentage of the V-ed form (7 percent in COLEC and 8 percent in LOCNESS) will not yield a large number of V-ed forms that are not shared by COLEC writers anyway.

5.5.6 Matching the V-n form lists

One of the most apparent features in the V-n matching list is that it is the longest of all the

match lists (see Table 5.19). This could be interpreted as the learners weakness in using the V-n forms as a whole by the learners. This could also be interpreted as the underuse of the passive voice because past participles of verbs are indispensable for the composition of the passive voice. The absence of the V-n forms of some irregular verbs such as *bound*, *hung*, *fed* and *struck* in COLEC seems to show that the learners have problems in producing the past participles of irregular verbs.

When the passive voice is compared in the two corpora by using the query "VB* *VVN" in WordSmith, (in which VB* refers to all the forms of the verb *BE* and VVN* refers to all the past participles of verbs) it is found that the learners use a much smaller proportion of passive voices than the NSs (see Table 5.20). The normalised figures in the table show that the NSs use passive voice twice as often as the learners do.

Table 5. 19 All the V-n forms occurring only in LOCNESS (frequency ≥ 4)

N	V-n	FRE	V-n	FRE	V-n	FRE
1	BANNED	53	CONVEYED	7	REUNITED	5
2	PRESENTED	37	CRITICISED	7	RIDICULED	5
3	INTRODUCED	32	DONATED	7	RIPPED	5
4	PORTRAYED	31	EXTENDED	7	SUBJECTED	5
5	ARGUED	30	LABELED	7	TRANSFERRED	5
6	DEFINED	30	MISUSED	7	TRUSTED	5
7	VIEWED	28	OCCURRED	7	UNDERTAKEN	5
8	MEANT	23	PUSHED	7	UTILIZED	5
9	DISCOVERED	22	RANKED	7	VOTED	5
10	ELECTED	22	REFUTED	7	ABUSED	4
11	DEBATED	19	REPRESENTED	7	ADHERED	4
12	PROVEN	19	SENTENCED	7	AFFLICTED	4
13	SEPARATED	18	SHAKEN	7	ANALYZED	4
14	COMMITTED	17	SIGNED	7	AWARDED	4
15	DESCRIBED	17	STUCK	7	CAPTURED	4
16	HELPED	17	TALKED	7	CEDED	4
17	DESIGNED	16	WHIPPED	7	CHALLENGED	4
18	REJECTED	16	ADMIRED	6	CLASSIFIED	4
19	STOPPED	15	ANSWERED	6	COMPELLED	4
20	ALTERED	14	BRED	6	CONCENTRATED	4
21	DIRECTED	14	CONDEMNED	6	CONCIEVED	4
22	JUSTIFIED	14	CONFRONTED	6	CONSTRUED	4
23	REFERRED	14	DERIVED	6	COUPLED	4
24	DEEMED	13	DISCRIMINATED	6	DELIVERED	4
25	OPPOSED	13	ENFORCED	6	DESTINED	4
26	REMOVED	13	FOCUSED	6	DETACHED	4
27	REPLACED	13	HANGED	6	DETECTED	4
28	DEALT	12	MARRIED	6	DISMISSED	4

29	EXPRESSED	12	MURDERED	6	DISREGARDED	4
30	RAPED	12	OUTLAWED	6	DRAFTED	4
31	ABOLISHED	11	PROGRAMMED	6	EMPHASISED	4
32	CONTINUED	11	PUBLISHED	6	ENCOUNTERED	4
33	IMPLEMENTED	11	RECOGNIZED	6	ENJOYED	4
34	DENIED	10	REVEALED	6	EVOKED	4
35	EVOLVED	10	SHOT	6	FAVOURED	4
36	INCLUDED	10	STAYED	6	FED	4
37	LEGALIZED	10	SUED	6	FRIGHTENED	4
38	RULED	10	SURROUNDED	6	FULFILLED	4
39	ACCUSED	9	SYMBOLISED	6	GOVERNED	4
40	BOUND	9	TORN	6	IMPLANTED	4
41	CONDUCTED	9	TRANSMITTED	6	INCARCERATED	4
42	DEMONSTRATED	9	WEAKENED	6	INSTALLED	4
43	FOUGHT	9	ABANDONED	5	INVESTED	4
44	IMPOSED	9	ADDICTED	5	MANIPULATED	4
45	INTENDED	9	ADDRESSED	5	MEASURED	4
46	REALISED	9	ASSUMED	5	MONITORED	4
47	RECOGNISED	9	BENEFITTED	5	OBSERVED	4
48	RESTRICTED	9	CENSORED	5	OUTLINED	4
49	CONVICTED	8	CENTRED	5	PERCEIVED	4
50	DRESSED	8	CHASED	5	PERSUADED	4
51	ELIMINATED	8	DIMINISHED	5	PRESCRIBED	4
52	EMPLOYED	8	EXISTED	5	PROCESSED	4
53	ENABLED	8	FILED	5	PUBLICISED	4
54	EXECUTED	8	FLOGGED	5	REINFORCED	4
55	EXPOSED	8	GUARANTEED	5	RESERVED	4
56	HEADED	8	HIDDEN	5	SHATTERED	4
57	INTEGRATED	8	HUNG	5	SLAUGHTERED	4
58	INTERPRETED	8	ILLUSTRATED	5	SOUGHT	4
59	MANAGED	8	INCORPORATED	5	SPREAD	4
60	NOTED	8	INSTITUTED	5	STRUCK	4
61	RETAINED	8	INSTRUCTED	5	SURVIVED	4
62	RETURNED	8	ISOLATED	5	TACKLED	4
63	SHARED	8	LEGALISED	5	TITLED	4
64	APPOINTED	7	LOCATED	5	TRANSPORTED	4
65	ARRESTED	7	OVERLOOKED	5	UNDERGONE	4
66	BETRAYED	7	PROPOSED	5	UNDERMINED	4
67	BLAMED	7	PURCHASED	5	VALUED	4
68	BLOWN	7	QUESTIONED	5	WITNESSED	4
69	CONTRACTED	7	RENEWED	5		

Table 5. 20 The raw and normalised figures of the structure "*BE* + V-n" of COLEC and LOCNESS

COLEC		**LOCNESS**	
raw	normalised	raw	normalised
2470	2573	3567	5531

If we look at another structure that employs past participle forms of verbs (PP) such as "NOUN + PP" in WordSmith (by the query "NN* *VVN"), we get the result in Table 5.21. The normalised figures in the table show that the NSs use far more past participles to modify nouns.

Table 5. 21 The raw and normalised figures of the structure "NOUN + V-n" of COLEC and LOCNESS

COLEC		**LOCNESS**	
raw	normalised	raw	normalised
374	390	616	955

Since only transitive verbs can be used in the passive voice and meanwhile not all transitive verbs can actually be used in the passive voice, Table 5.19 provides a very handy list for learners to practice passive voice construction.

5.5.7 Some remarks in summary

By sorting the matched list in the way shown above, it is possible to see what verb forms are totally absent in the learners' written output. By amalgamating the tables (from Table 5.14 to Table 5.19), the profile of the learner English as a result of comparison becomes more apparent (see Appendix 5 for all the verb forms that only occur in LOCNESS and whose frequency is above four inclusive). Table 5.22 is a sample of 20 out of the 633 verb forms. A table could also be manually converted to a more readable form, like Table 5.12.

Even though the learners use the base form dramatically more than the NSs, they do not appear to be producing this form better than others if we refer to Table 5.20 and consider the total number (191) of the V-i form and the V-e form, which is slightly fewer than the V-n form (205). This shows from another perspective that the learners over-rely on a few core verbs and do not use a large number of alternatives. Based upon the lists in this section, it is possible to construct an order of familiarity, or to be more exact, an affinity of the performance of the learners to that of the NSs. however, it must be admitted that this is a very crude judgement without having taken other factors into account.

This section seems able to answer the research question that asks whether there is a degree of familiarity in the learner English with different forms of verbs and the order of familiarity to

the learners.

Table 5. 22 The first 20 verb forms that only occur in LOCNESS (frequency ≥ 4)

Word	Total	V-e	V-i	V-s	V-ing	V-ed	V-n
1 ABANDONED	1	0	0	0	0	0	1
2 ABOLISH	1	0	1	0	0	0	0
3 ABOLISHED	1	0	0	0	0	0	1
4 ABOLISHING	1	0	0	0	1	0	0
5 ABUSE	1	0	1	0	0	0	0
6 ABUSED	1	0	0	0	0	0	1
7 ACCEPTS	1	0	0	1	0	0	0
8 ACCOMMODATE	1	0	1	0	0	0	0
9 ACCUSES	1	0	0	1	0	0	0
10 ACHIEVES	1	0	0	1	0	0	0
11 ACHIEVING	1	0	0	0	1	0	0
12 ACKNOWLEDGING	1	0	0	0	1	0	0
13 ACTS	1	0	0	1	0	0	0
14 ADDICTED	1	0	0	0	0	0	1
15 ADDRESS	1	0	1	0	0	0	0
16 ADDRESSED	1	0	0	0	0	0	1
17 ADDRESSING	1	0	0	0	1	0	0
18 ADHERED	1	0	0	0	0	0	1
19 ADMIRED	1	0	0	0	0	0	1
20 ADMIT	1	0	1	0	0	0	0

Table 5. 23 A summary of the verb forms that occur only in LOCNESS (frequency ≥ 4)

	V-i	V-e	V-s	V-ing	V-ed	V-n
LOCNESS	142	49	117	115	42	205
Order	5	2	4	3	1	6

As the numbers in the 'Order' row in Table 5.23 suggest, the learners do show a degree of familiarity with the different forms of the verbs: in order from most to least of 'V-ed', 'V-e', 'V-ing', 'V-s', 'V-i' and 'V-n'.

5.6 Some pedagogical implications

5.6.1 Significance for the writer of teaching materials

The most envisaged value in working out the lists in this research lies in the first-hand reference for the writer of teaching materials. This is because knowing which form or forms of which verbs are most often used by the NSs in a particular register is useful for determining

what should be included as teaching material for the learners. If we believe that language could be better learned by treating vocabulary as priority (rather than grammar), then the distribution of verb forms of targeted NS English should be very accurately and extensively realised in the teaching materials. With the information obtained from a homogeneous group such as the COLEC writers, the author of teaching materials can be confident that teaching materials based on the findings obtained from such a group should benefit the learners more than those teaching materials imagined from ideal or stereotyped learners.

5.6.2 Significance for the teacher and the learner

Looking at the most often-used verb forms with the top 20, and then all the verb forms that occur only in LOCNESS, helps the teacher and the learner to see which alternative comes first among all those available. Take the top 7th verb lemma *USE* in COLEC for example; the V-ed form of this lemma does not appear in the top 20, being ranked 33rd. This suggests that the learners are not only over-relying on a small number of vocabulary words (as noticed by CIA researchers such as Granger 1998, Cobb 2003 and many others), but more importantly over-relying on a narrow range of verb forms. In other words, knowing which other forms are very frequently used by the NSs helps learners complete their knowledge of the vocabulary they have partially learned. 'To try the new and mend the old' could be used to summarise the essence of making sense of the research in this chapter.

On the whole, knowing the disparity between learner English and NS English helps the teacher or the course designer to fill the gaps when designing teaching tasks and teaching materials. However, the lists should never become a rigid and absolute law for learners to follow. While we are interpreting the tables in this chapter, it should always be borne in mind that vocabulary excessively influenced by topics should give way to vocabulary that is mostly used in a more general way and provides a background for learners. For example, some cultural and topic-sensitive words such as *PREACH*, *FLOG* and *WHIP* are not generally representative of academic English, and therefore do not have to be encouraged for production. Caution should also be exercised when considering misspelled word forms such as *concieve* (for *conceive*) and *loose* (for *lose*) in Table 5.14, *concieved* (for *conceived*) in Table 5.19, as highlighted.

5.6.3 Significance for learner English level evaluation

In Chapter Two (see 2.9.5) it was mentioned that there is no linguistic standard for giving degree labels to collective learner English (such as *advanced*, *intermediate*, or *elementary* levels). It is reasonable to propose that the more congruent the learners' English production in the distribution of verb forms is with the distribution pattern of NSs, the higher stage the learners as a homogeneous group should be deemed to have reached, the verb use acting as a marker. In the same vein, if the learners have more verb-form types overlapping with those of the NSs, it is more likely that they have a higher degree of production. Yet this requires strict register and topic control. Only similar registers and topics of learner English corpora may be compared for the purpose of determining how advanced learner English is compared with NS English. Of course, this is only a preliminary exploration into the behaviour of learner English by means of comparative examination of NS English and NNS English. Further studies are needed if we wish to be in a position to make firm claims about the features of group learner English and the relationship between a given level of learner English and the similarity between learner English and NS English. This thesis does not attempt to delve deeper into this area because it would take another complete thesis to investigate it. However, the proposal made here could be a starting point for a possible project.

5.6.4 Implications for further corpus design, construction and comparison

One prominent issue that impacts the comparison between a learner English corpus and a NS corpus is that the disparity in text types and topics leaves an observable trace in the distribution of verb lemmas and verb forms. This leads to the conclusion that corpora of different topics and text types suffer badly from the undesirable existence of unexpected disparity in various key words. Researchers comparing corpora with different topics and text types should be very much aware of this problem. However, it would be irrational if we should jump to the conclusion that corpora of different topics and text types should not be compared. No matter how well the data to be compared were controlled in terms of register, topics and other factors, it would be almost impossible to reach an ideal level of absolute affinity between the learner corpus and the NS corpus. Disparity of one kind or another will surely exist. As long as researchers bear this in mind, they should have a better chance of benefiting from working out the production features of learner English.

In looking at all the verb forms occurring only in LOCNESS, this research has studied only those with a frequency above four, inclusive. However, this does not mean that the tail of the list is useless for language education. Even though the tail of each form might not be of too much use for average students, there is no reason to stop this information from being used to assist the improvement of advanced learners' writing production.

5.6.5 Some problems revealed concerning CLC studies

This research has compared the frequencies of the learner English and the NS English as if the performance of the learner English were errorless. In fact, as I reinforce in other places in this thesis, frequency hides errors. Researchers doing CIA by comparing learner language production and NS language production should bear in mind that frequency in the learner corpus is only a rough index. If we make use of the advantage of frequency via a corpus-linguistic approach first, we should not forget the advantage of concordances, which reveal problems that frequencies hide. Take the verb form *thinks* for example. There are as many as 33 cases of this verb in COLEC, giving an impression that the learners as a group use it fairly frequently. If we look at the concordances, however, it is a different situation. To save space the lines from 4 to 18, which show correct usage ("he thinks …"), are omitted. In the remaining lines, as many as 13 cases of the node *thinks* are incorrectly used (as highlighted in Figure 5.7) due to disagreement between the subject and the predicate. In other words, nearly half of the occurrences of *thinks* are wrong.

Figure 5. 7 Some of the lines of thinks from COLEC

```
1                                  People always thinks that finishing things as fast as you
2                                  People always thinks the fresh water can be used and will
3                                  People always thinks human will have the fresh water for
19      can know the world outside the campus? I thinks we can do it from reading newspapers
20     me, how can I turn my dream into truth? I thinks  the unchaged  life is terrible. I l
21     we understand "Practice makes perfect"? I thinks after you do much more work you can
22      facts often are different with our ideal thinks.  Why? Let me tell you a phenomenom
23      job to do,  he wants to earn money;  one thinks it is pleasure  to do that job for t
24     o change job  often, because these people thinks  that changing job  often  can make
25      sea.                               People thinks that fresh water is a thing that we
26     imited.                             People thinks that fresh water is not limited. The
27                                 Many a person thinks fresh water in the earth will never
28     ebergs in the earth. In addition, someone thinks there is much under-ground water. So
29      you are lucky everyday.          Someone thinks that some numbers will bring good lu
30      short time, he only cares for the speed, thinks "quick, quick" , and  does the job c
31     e fresh water than ever. And people still thinks they have enough fresh water, so the
32     h water.  In fact, It is wrong with their thinks. On our earth, It is shortage of fre
33     ple who always does  the same job usually thinks :  In  his life, his income is stabl
```

As illustrated above, researchers might be at risk if they over-rely on frequency in the learner

corpus. Cross-checking between frequency and concordance lines is the best solution to the problem of incorrect information about learner language production.

This does not mean, however, that working out the 20 most often-used verb forms and all the verb forms exceeding four in frequency is rendered useless. Knowing how well or poorly the learners perform in their English production helps us to diagnose the problems that they currently have and that need to be rectified. From Chapter Seven to Chapter Nine this latter perspective will be adopted to further explore the English production of learners.

5.7 Conclusion

This chapter has proposed a methodology for probing similarity and disparity in the individual forms of verbs occurring in learner English and NS English. The significance of this research for the writer of teaching materials, the teacher and the learner has been discussed. Based upon the study into the relationship between the distributional patterns of the different verb forms of the two corpora, a proposal has been made for using a possible linguistic criterion for ascribing a level, or degree, to collective learner English. Some of the problems involved in making and making sense of such research are mentioned, and advice offered for further CIA work in the area of essay register and topic control.

This chapter, together with the previous one, has dealt with the function of verbs. In the English language there are quite a number of words that serve both as verbs and nouns. The next chapter looks at this part of production and sees what information could be obtained from a comparative analysis between the learner English and the NS English.

Chapter Six

Between Verbs and Nouns

6.1 Introduction

Chapter Four and Chapter Five have looked at verbs in detail according to the verb lemma lists described in Chapter Four. Before such verb lemma lists are produced, the raw lists contain other POS words sharing the same forms as verbs, including nouns. Now that we have studied the learner English in the area of verbs, it is necessary to carry out a study of the relationship between verbs and their morphologically identical noun forms and then between verbs and nouns that do not share the same forms. Such a study is valuable because the information from comparison between verb use and noun use with the same form can help us to draw up a better profile of the learner English in the relationship between using verbs and using nouns.

There has been a long history of observing the different uses and functions of different POSes starting from the middle of the last century by West (1953) and his colleagues. In his classic pioneering work *A General Service List of English Words* (hereafter called GSL), he had 2000 most common words counted semantically. This semantic count is still influential in linguistic studies and pedagogical applications today. However, due to the limitations of technology at that time, it was difficult to see whether there exists a general trend in using the different POSes in the whole language. It is only when computational annotation technology has recently become fairly mature that it is possible to reveal such a general trend in using different POS vocabulary. Based on corpus investigations, for example, Biber *et al*. (1999: 65) found that the lexical word classes vary greatly both in overall frequency and across registers. In overall frequency nouns are the most frequent word class and across registers nouns are most common in news and academic prose but least common in conversation. Altenberg (1996, cited in Ringbom 1998b: 50) found that Swedish learners use a larger proportion of verbs than nouns and produce a language similar to the style of fiction and informal talk. Because learner English has been found to be highly characteristic of an oral style (see 2.7.1, Chapter Two for a detailed review of this issue), it can be hypothesised that, as between verbs

and nouns, there will appear a strong tendency in learner English for learners to use a higher proportion of verbs and a lower proportion of nouns.

This research starts with making and making sense of two lists which contain a number of lemmas that are morphologically the same but functionally different in POS such as *HOPE, INCREASE* and *SUPPORT*, serving both as verbs and nouns. For the sake of convenience, I will call these lemmas *norbs,* a term coined by Sinclair (2004: 199). As an extension of the study between verbs and nouns sharing the same morphological forms, a certain number of nouns which do not share the same forms with their equivalent verbs (such as *acceptance* for *accept*) are also examined. This chapter demonstrates how a corpus-linguistic approach could facilitate an efficient analysis of learner English in the area of writers' selection in verbs and nouns as a group. The research questions this chapter attempts to answer are as follows:

(1) How many norbs are used by the COLEC and the LOCNESS writers?
(2) Is there a general tendency in using the verb function and the noun function by the COLEC writers? If there is, what is it? And is the tendency of the COLEC writers the same as that of the LOCNESS writers? How similar is the general trend in LOCNESS to that in GSL in terms of the selected words?
(3) If there is a general trend in using one function *over* another in norbs, does this trend also exist in the verb and noun pairs that do not share the same morphological forms?
(4) What is the pedagogical significance of the findings?

Even though there are cases where the senses of the verb and the noun are not necessarily the same (for example *ISSUE*), mostly the senses of a particular norb are consistent with each other (for example *RESEARCH*). Therefore, the potential difference in meaning between the verb and the noun function of a norb is ignored in this research. Unlike the other chapters of this thesis, the definition of lemma in this section cuts across POS boundaries, i.e. it covers words which serve as more than one POS.

6.2 A general view of the disparity between the two corpora in terms of the selection between verbs and nouns

Based on the verb lemma lists created in Chapter Four (see Appendix 2 and Appendix 3), two norb lists could be created by using Excel. Since there is no need to look at the different forms

of verbs, all the individual verb forms are amalgamated into one column, i.e. the V-total column. Since the POS annotation is not 100 percent accurate and learner English has a large number of syntactically incorrect uses, small-frequency norbs were deleted (verb function ≤ 1, noun function ≤ 2) to avoid such noise. There are altogether 234 norbs in COLEC and 343 norbs in LOCNESS. With the aid of the sorting function of Excel, the list could be made to show the lemmas that serve as verbs mainly and nouns occasionally, such as *SAY* and *DRINK,* in Table 6.1 by sorting the 'V-total' column first and the 'Noun' column second.

Table 6. 1The top ten norbs that are mainly used as verbs in LOCNESS (Ratio = V-total/Noun)

Lemma	V-total	Noun	Ratio
say	493	5	99
try	266	3	89
lead	266	7	38
stop	116	4	29
like	91	4	23
pass	91	4	23
win	87	4	22
hold	105	5	21
produce	81	4	20
drink	45	3	15

By the same token, a list could also be made by sorting the 'Noun' column first and the 'V-total' second to show the lemmas that serve as nouns mainly and verbs occasionally, such as *GROUP* and *MARKET* as in Table 6.2.

Table 6. 2 The top ten norbs that are mainly used as nouns in LOCNESS (Ratio = Noun/V-total)

Lemma	V-total	Noun	Ratio
group	2	155	78
culture	2	124	62
side	2	120	60
reason	5	258	52
court	2	101	51
class	2	100	50
level	2	99	50
position	2	79	40
issue	6	218	36
market	3	96	32

The tables above seem to reveal that norbs show a gradation from being selected for their verbal function to being selected for their nominal function in a homogeneous group of writers and in a particular genre. Could we call the norbs that are mainly used as verbs more 'verb-like' and call the norbs that are mainly used as nouns more 'noun-like'? There seems to be very little research so far in this area. Actually, it can be envisaged that further examination will yield useful information for linguistic research and pedagogical applications.

The two tables above concerns the norbs used in LOCNESS. The following table (Table 6.3) is the counterpart of Table 6.1, which contains the first 10 norbs that are mainly used as verbs in COLEC.

Table 6. 3 The top ten norbs that are mainly used as verbs in COLEC (Ratio = V-total/Noun)

Lemma	V-total	Noun	Ratio
think	2132	4	533
fake	2187	5	437
make	3856	19	203
like	1004	5	201
go	962	7	137
take	1231	9	137
use	1390	328	4
change	1015	405	3
study	861	488	2
work	858	1077	1

The following table (Table 6.4) displays the first 10 norbs that are mainly used as nouns.

Table 6. 4 The top ten norbs that are mainly used as nouns in COLEC (Ratio = Noun/ V-total)

Lemma	V-total	Noun	Ratio
part	2	387	194
view	3	319	106
hand	6	462	77
word	15	886	59
practice	45	1527	34
waste	362	806	2
work	858	1077	1
study	861	488	1
change	1015	405	0
use	1390	328	0

It is not difficult to find that the ratio differences between the two corpora from top one to top ten, either for the most often used verb function dominated norbs (Table 6.1 and Table 6.3) or

for the most often used noun function dominated norbs (Table 6.2 and Table 6.4), are huge. Whereas the largest V-toal to noun ratio in LOCNESS is 99 and the smallest ratio is 15 (Table 6.1), the largest V-toal and noun ratio is 533 and the smallest ratio is 1 in COLEC (Table 6.3). And whereas the largest noun to V-toal ratio in LOCNESS is 78 and the smallest ratio is 32 (Table 6.2), the largest noun to V-total ratio is 194 and the smallest ratio is almost zero in COLEC (Table 6.4).

Furthermore, if we look at the total figures for verbs in total (V-total) and nouns (Nouns), a disparity begins to emerge (see Table 6.5). As this table shows, the COLEC writers use twice as many verbs as nouns whereas the LOCNESS writers use verbs and nouns approximately the same amount. In other words, the trends in using verbs and nouns in norbs are just the opposite: COLEC writers use more verbs than nouns while the LOCNESS writers use more nouns than verbs.

Table 6. 5 The total frequency of verbs in total and nouns in COLEC and LOCNESS

Corpus	V-total	Noun
COLEC	30086	14007
LOCNESS	10441	11860

By looking at the total figures of the verbs in total and nouns, it is possible to see a very general trend of NSs in selecting the verb use and the noun use within norbs and the disparity between the two groups of writers. But the information that can be obtained in the lists is very general and vague. For example, it is found that the learners use a much larger proportion of verbs than nouns compared with the NSs; does that mean the learners use all the norbs in this way or some or most? Should the LOCNESS trend be treated as a sort of norm for the learners to follow? How is the trend in LOCNESS comparable to that of GSL? Without a detailed study of some of the norbs, it would be difficult to answer the questions.

6.3 A detailed look at the disparity between the two corpora in terms of selection between verbs and nouns

As hypothesised earlier in this Chapter (6.1), the learners are expected to show a larger proportion of verb use than noun use due to the oral-like feature of learner English as a whole. To test this hypothesis requires a look at a considerable number of verbs and their equivalent nouns which are not only the same in form, like *charge* (verb) vs. *charge* (noun) and *control* (verb) vs. *control* (noun), but also different in form like *accept* vs. *acceptance* and *apply* vs.

application. In the following three sections, these two types of distribution will be examined in some detail.

6.3.1 Between the verb use and the noun use within the same word form

The following table (Table 6.6) is a presentation of the total frequencies of verb use and noun use of 25 norbs from COLEC and LOCNESS. Since POS-tagging has a problem of accuracy, especially with learner English, the table has been drawn up manually to avoid this problem. It must be admitted that the selection of the 25 norbs is entirely arbitrary and includes small frequencies because there is no need to worry about the accuracy of POS identification in manual classification. In Table 6.6, 'CVT' refers to the total frequency of a verb in COLEC, 'CNT' refers to the total frequency of a noun in COLEC. 'CT' refers to the total frequency of the shared form both as a verb and a noun in COLEC (i.e. CT = CVT + CNT). Similarly, 'LVT' refers to the total frequency of a verb in LOCNESS and 'LNT' refers to the total frequency of a noun in LOCNESS. And 'LT' refers to the total frequency of the shared word form both as a verb and a noun in LOCNESS (i.e. LT = LVT + LNT). The gross ratio between verbs and nouns in the 25 words is as follows in Table 6.7:

This result shows that within the range of the 25 pairs of verbs and nouns COLEC writers use more verbs than nouns while the LOCNESS writers use more nouns than verbs. Seemingly, COLEC learners are not dramatically overusing verbs. In fact, the large amount of use as a noun of *hand* and *view*, and the huge figure for the use of a verb such as *need* and *increase* have twisted the total percentage of these 25 words and made the result of the calculation very unreliable. If the few pairs of words in which there is exceptionally frequent use of nouns could be deleted from the table (such as *doubt, hand* and *view*), certainly there would be a larger proportion of verb use than noun use.

To compare the ratio between verb use and noun use in the two corpora, Table 6.8 has been drawn up. In this table, 'CV%' stands for the percentage of verb use in COLEC and 'CN%' refers to that of noun use in COLEC. 'LC%' stands for the percentage of verb use in LOCNESS and 'LN' refers to that of noun use in LOCNESS. To make the learner English appear more meaningful and more comparable to a well-accepted standard in terms of the

percentage of the verb use and the noun use, GSL will be referred to. 'GSL V%' refers to the percentage of verb use in GSL, 'GSL N%' refers to its percentage of noun use and 'GSL T%' refers to the total percentages provided in the breakdown of semantic counts. It should be noted that in GSL not all words are counted semantically with a full 100 percentage (actually most are not). A certain number of minor meanings are omitted, so the percentages of the verb use and the nouns use do not add up to 100, according to GSL (West: 1953 viii).

Table 6. 6 The total frequency of verb use and noun use of 25 norbs in COLEC and LOCNESS

WORD	C V T	C N T	CT	L V T	L N T	LT
CHARGE	16	5	21	6	8	14
CONTROL	139	17	156	56	101	157
DESIRE	6	8	14	20	64	84
DOUBT	1	13	14	6	29	35
FAVO(U)R	5	9	14	14	65	79
FEAR	11	4	15	24	59	83
FORCE	22	10	32	90	76	166
HAND	12	462	474	10	126	136
HOPE	126	32	158	42	53	95
INCREASE	704	26	730	211	55	266
INFLUENCE	38	10	48	27	44	71
INTEREST	179	90	269	49	72	121
JUDGE	13	2	15	74	69	143
NEED	646	107	753	304	137	441
PROGRESS	15	183	198	16	24	40
QUESTION	2	132	134	39	190	229
REQUEST	7	2	9	6	6	12
RESULT	180	197	377	180	198	378
RISK	11	6	17	12	75	87
SEARCH	20	3	23	20	40	60
SUPPORT	28	6	34	138	80	218
SURPRISE	28	14	42	10	2	12
THANK	5	3	8	11	11	22
TRUST	17	2	19	16	13	29
VIEW	3	317	320	81	169	250
Total	2234	1660	3894	1462	1766	3228
Average %	59	41	100	45	55	100

Table 6. 7 The total frequency of verb use and noun use and the ratio of verb use and noun use in COLEC and LOCNESS

Corpus	Verb	Noun	Ratio
COLEC	2234	1660	1.3:1
LOCNESS	1462	1766	0.83:1

Table 6. 8 The percentages of verb use and noun use of 25 verbs in COLEC, LOCNESS and GSL

WORD	CV%	CN%	LV%	LN%	GSL V%	GSL N%	GSL T%
CHARGE	76	24	43	57	34	64	98
CONTROL	89	11	36	64	29	67	96
DESIRE	43	57	24	76	38	56	94
DOUBT	7	93	17	83	28	68	96
FAVO(U)R	36	64	18	82	25	66	91
FEAR	73	27	29	71	46	50	96
FORCE	69	31	54	46	25	73	98
HAND	3	97	7	93	4	84	88
HOPE	80	20	44	56	49	51	100
INCREASE	96	4	79	21	69	30	99
INFLUENCE	79	21	38	62	11	88	99
INTEREST	67	33	40	60	38	52	90
JUDGE	87	13	52	48	36	64	100
NEED	86	14	69	31	63	26	89
PROGRESS	8	92	40	60	7	84	91
QUESTION	1	99	17	83	8	89	97
REQUEST	78	22	50	50	35	63	98
RESULT	48	52	48	52	27	70	97
RISK	65	35	14	86	30	70	100
SEARCH	87	13	33	67	35	55	90
SUPPORT	82	18	63	37	48	43	91
SURPRISE	67	33	83	17	67	30	97
THANK	63	38	50	50	60	26	86
TRUST	89	11	55	45	46	52	98
VIEW	1	99	32	68	9	83	92
Average	59	41	45	55	36	63	99

In the following paragraphs, the learners' performance regarding the verb use and noun use of the 25 "norbs" goes against the ratio identified in GSL. Taking the degree of resemblance between learner English and NS English in GSL into consideration, the learner English in COLEC can be roughly divided into five categories:

1. The ratio of a word used as a verb is higher than that of it used as a noun, according to GSL, and there exists a very high ratio of verb use in COLEC, to which group the following words belong: *need, increase* and *support*.
2. The ratio of a word used as a verb approximately equals that used as a noun in GSL, there still exists a more ratio of verb use in COLEC, to which group the following

word belongs: *fear*.

3. The ratio of a word used as a noun outweighs dramatically that of it used as a verb in GSL, yet there still exists a higher ratio of verbs in COLEC, to which group the following words belong: *charge, control, face, hope, influence, interest, judge, request, risk, search* and *trust*.
4. The ratio of a word used as a verb approximately equals that of it used as a noun in GSL and there is a similar ratio between the verb use and the noun use in COLEC, to which group the following words belong: *desire, surprise, thank, favo(u)r* and *progress*.
5. The ratio of a word used as a noun vastly exceeds that of it used as a verb in GSL, and there is also an extreme ratio of word as a noun in COLEC, to which group the following words belong: *doubt, hand* and *view*.

To a large extent, the first three categories are of the same feature in that COLEC writers tend to use the verb function compared with the percentage count by GSL. In the fourth category, there are a few words which not only follow the general trend of GSL (whether the verb use is more than the noun use or the other way around) but also resemble the percentage given in GSL. In the fifth category, however, COLEC writers tend to use nouns, which is totally against the overall trend in the option between verbs and nouns. It seems as if lexical grammar plays an important role in interpreting and analysing the existence of learners' tendency to use nouns. In the extensive use of *examination*, for example, COLEC learners prioritise the noun rather than the verb *examine*. Presumably, this is caused by the special requirement of a topic (as in the case of examination). It might well be the case that the word is acquired as a noun in the first instance and the verb is acquired afterwards when the noun's function has already taken a strong hold. For some convincing evidence, a diachronic study is needed, which is beyond the aim of this synchronic research.

After comparing the learners' use against that of GSL, it will be helpful if the two communities of NSs with regard to the verb and noun ratio could be compared. Not surprisingly, it is much easier to categorise their tendency because most words follow the general trend and even resemble the GSL ratio of verb use and noun use. Unlike the categorisation for COLEC, only two categories will be enough roughly to encapsulate the NSs

performance:

1. The general trend of LOCNESS resembles GSL in terms of the ratio of verb use and noun use, to which group the following words belong: *charge, control, desire, favo(u)r, hope, interest, need, search*, *doubt, fear, hand, increase, influence, progress, request, risk, support, surprise* and *view*.
2. The general trend of LOCNESS contradicts that of GSL in terms of the ratio of verb use and noun use, to which group the following words belong: *face, force, judge* and *trust*.

The overall resemblance of LOCNESS ratio to that of GSL is largely in agreement with the trend in GSL in respect of the percentage of verbs and nouns even though there exists a contradictory trend in a few words. There may be reasons to account for the existence of such a contradictory trend in different NS writings, for example the register difference. Since there is little information available concerning the selection of the data used by GSL, this issue will not be dealt with in this research.

Based on this finding, it might be better to claim that NNS overuse verbs as a whole. However, it may be, rather, that they overuse a small group of nouns to a great extent when these words have an overwhelmingly high ratio of noun use in the general NS English. Having looked at the general trend in overusing the verb function in norbs by the learners, it helps to look at some individual examples for a better understanding of this feature.

As the following examples reveal, some words used as verbs in COLEC can actually be used in the noun function in LOCNESS. Let us look at the COLEC examples first:

7) Whether the Chinese team does good or not, I'll *support* it for ever.

8) In future, the society will be *supported* by us. Knowing all kind[s] of the problems that exist in it and recogniz[ing] the things that happen around us, […] we can know […] how we can contribute to it.

Some similar cases are found in LOCNESS where the NSs apply the noun function rather than verbs:

1) In short the community was reminding the Italians of the degree of transfer of sovereignty. However, the current European writer Collins while maintaining his *support* for community law and discretion is in favour of the power of Parliament to repeal the 2 Act that made Britain a member.

2) I feel that there are both values and consequences to the integration of schools and if the program is going to be successful, it needs *support* from venues other than the school systems themselves.

The two examples in LOCNESS show that the two examples in COLEC used as verbs could actually be rephrased as nouns thus:

1) Whether the Chinese team does good or not, I'll maintain my *support* for it as before.

2) In future, the society will need *support* from us. Knowing all kind[s] of the problems that exist in it and recognize[ing] the things that happen around us, […] we can know […] how we can contribute to it.

This section has looked at the learners' tendency in choosing between verbs and nouns, but only with norbs. To be more confident about the learners' predilection for choosing verbs over nouns, the next section extends the comparison to verbs and their equivalent nouns which do not share the same form as norbs do.

6.3.2 Between verbs and nouns with different word forms

In this section, a study of 25 verbs and their related nouns[36] is carried out, looking into the tendencies of the corpus writers of the two groups in choosing nouns in their writings. It must be admitted that the selection of these verbs is totally arbitary and does not follow any criteria. These verbs and their noun equivalents are provided in Table 6.9.

In Table 6.10, 'CV' represents the verb frequency in COLEC, 'CN' represents the noun frequency in COLEC, 'LV' represents the verb frequency in LOCNESS and 'LN' represents the noun frequency in LOCNESS. In the 'VERB' column, each verb is referred to as a lemma including all the forms of the verb: the base form, the third singular form, the "-ing" form, the

36 Only one of the equivalent nouns is chosen if there is more than one noun.

past form and the past participle form. The 'NOUN' column includes both the singular form and the plural form of a noun.

Table 6. 9 The verb forms and noun forms of 25 V-N pairs

Verb	Noun	Verb	Noun	Verb	Noun
accept	acceptance	complete	completion	manage	management
apply	application	create	creation	occur	occurrence
argue	argument	enter	entry	produce	production
assume	assumption	examine	examination	realise	realisation
believe	belief	express	expression	realize	realization
choose	choice	include	inclusion	refuse	refusal
commit	commitment	indicate	indication	survive	survival
communicate	communication	introduce	introduction		
compare	comparison	involve	involvement		

Table 6. 10 The frequencies of 25 verbs and their equivalent nouns in COLEC and LOCNESS

VERB	C V	L V	NOUN	C N	L N
ACCEPT	41	182	ACCEPTANCE	0	33
APPLY	65	60	APPLICATION	2	11
ARGUE	1	167	ARGUMENT	3	339
ASSUME	3	40	ASSUMPTION	0	13
BELIEVE	295	373	BELIEF	14	125
CHOOSE	121	140	CHOICE	31	129
COMMIT	8	90	COMMITMENT	0	12
COMMUNICATE	24	24	COMMUNICATION	30	27
COMPARE	52	49	COMPARISON	2	15
COMPLETE	35	42	COMPLETION	0	5
CREATE	18	182	CREATION	1	38
ENTER	84	56	ENTRY	1	10
EXAMINE	17	28	EXAMINATION	60	5
EXPRESS	25	56	EXPRESSION	9	12
INCLUDE	67	111	INCLUSION	0	2
INDICATE	21	10	INDICATION	0	6
INTRODUCE	12	61	INTRODUCTION	2	44
INVOLVE	12	159	INVOLVEMENT	0	9
MANAGE	29	27	MANAGEMENT	8	12
OCCUR	25	96	OCCURRENCE	0	5
PRODUCE	239	89	PRODUCTION	65	38
REALISE	9	98	REALISATION	0	16
REALIZE	196	122	REALIZATION	3	14
REFUSE	27	64	REFUSAL	0	13
SURVIVE	34	47	SURVIVAL	4	16
Total	1460	2373		235	949

In total, there are 1460 cases of verb use and 235 cases of noun use in COLEC and there are 2373 cases of verb use and 949 cases of noun use in LOCNESS. The ratios of verb use and noun use in the two corpora are as follows in Table 6.11:

Table 6. 11 The total frequencies of verb use and noun use of the 25 V-N pairs and their ratios in COLEC and LOCNESS

Corpus	Verb	Noun	Ratio
COLEC	1460	235	6:1
LOCNESS	2373	949	2.5:1

As a whole, as far as the 25 words are concerned the COLEC writers are over dependent on verbs. They use only one noun in every six verbs while LOCNESS writers have a much higher likelihood of using nouns, namely, one noun in every 2.5 verbs. It is not difficult to see that most nouns are less in number than verbs in both corpora. However, there are two counter-examples (see Table 6.10): *examination* and *production*. In the case of *examination,* this noun appears much more frequently in COLEC than in LOCNESS because COLEC writers often use *examination* in the sense of "test" rather than the sense of "investigation" as can be found in LOCNESS: *an exhaustive examination of the broadcast networks' programming*. Since tests are overwhelmingly a major concern of university students, the overuse of this sense of "test" in COLEC is understandable. Actually, when the LOCNESS corpus is looked at, in the five cases of *examination* in LOCNESS, there is only one that is used in the sense of "test" while the others are all in the sense of "investigation". In the case of *production*, which is the most frequently used of these nouns in COLEC, it is caused by the topic about the production of fake commodities in a majority of the essays in COLEC. The above statistical perspective provides a brief view of the overuse of verbs by the learners. If we look at some individual cases, this trend may become more apparent.

There are as many as 16 cases of the sequence *enter the society* in COLEC, which is a typical use when its equivalent noun *entry* could well be replaced instead:

1) We are only familiar with our campus and families. If so, we can not *enter* the society in the future.
2) Because we will *enter* the society in the future, we must adapt to it.

The use of the verb *ENTER* in the COLEC has made the learner English style rather

conversational. However, if we look at the use of *entry* in LOCESS, the register becomes more formal and academic.

1) His achievements were to veto twice Britain's *entry* into the common Market […]

2) The privileged graduates of these schools can guarantee a high-ranking career and […] accelerated promotion. *Entry* to these schools is via highly [competitive] exams, requiring two of three years of intensive studying in […]

In line with the trend of the NSs in using nouns, the COLEC writers could have written the two previous sentences thus:

1) If we are only familiar with our campus and families, *entry* into the society will be very difficult in the future.

2) Because *entry* into the society is inevitable, we might as well get prepared to adapt to it earlier.

This section has examined and compared the frequencies and ratios of 25 verbs and their equivalent nouns with different morphological forms. A strong impression that is obtained from this investigation is that the learners use a much smaller proportion of nouns than verbs compared with the NSs. Even though two of the 25 nouns (*examination* and *production*) are extensively used compared with the other nouns under study, it seems that they are too seriously affected by the topics. The essay title "Fake commodities and their harmfulness" will inevitably lead to a large number of uses of *production* and *produce*. Likewise, the learners' general and dominant concern, *examinations*, is also certain to occur in the learner language.

6.3.3 Between verbs and nouns in prepositional phrases

The two previous sections (6.3.1 and 6.3.2) have looked at the trends in production by the two groups of writers. The noun-function norbs or nouns which do not share forms with their equivalent verbs are studied in separation. To further reveal how the COLEC writers opt for verbs over nouns, it seems necessary to extend the study into some contextual areas. Basically the English syntax structure very often uses nouns in prepositional phrases. Therefore, I have

made the decision to explore in this section the verb and noun options within certain prepositional phrases.

There are two kinds of prepositions, according to Quirk *et al*. (1972: 299-305), "simple prepositions" and "complex prepositions". Most "simple prepositions" consist of one word only such as *at, in* and *for* whereas "complex prepositions" involve two or more elements as detailed below (*ibid.* :301):

> "[A] ADVERB + PREP: *along with*, *apart from* (BrE), *aside from* (AmE), *as for*, *as to*, *away from*, *into*, *off of* (AmE), *on to* (or *onto*), *out of*, *together with*, *up to*, etc.
>
> [B] VERB/ADJECTIVE/CONJUNCTION/etc + PREP: *except for*, *owing to*, *due to*, *but for*, *because of*, etc.
>
> [C] PREP1 + NOUN + PREP2: by means of, in comparison with, instead of, etc."

They also point out (*ibid*.:301-302) that type C is the most often used category and a definite or indefinite article may precede the noun in some complex prepositions such as *in the light of* and *as a result of*. They further divide the C category into subcategories, thus:

1. IN + NOUN + OF: *in case of*, *in charge of*, *in view of*, *in need of*, […], etc.
2. IN + NOUN + WITH: *in contact with*, *in common with*, […], etc.
3. BY + NOUN + OF: *by means of*, *by way of*, *by virtue of*, […], etc.
4. ON + NOUN + OF: *on account of*, *on behalf of*, *on the strength of*, […], etc.
5. OTHER TYPES: *at variance with*, *in exchange for*, *in return for*, […], etc.

In the following two sections, some investigation will be made into simple prepositions and complex prepositions. Due to limitations of space, I will in the first instance look only at nouns following a few simple prepositions such as *on*, *by*, *in*, *at*, *upon* and *under*. After that I will focus on the first subcategory of C, i.e. *IN* + NOUN + *OF,* as an example of the complex prepositions. The examination is basically based on the classification of Quirk *et al*. (1972).

Nevertheless, in cases where a simple preposition is mainly being considered (for example, *in control*), its relevant complex preposition (such as *out of control*) will also be discussed simultaneously. In fact, the boundary between simple and complex prepositions is arguably uncertain, as acknowledged by Quirk *et al*. (1972: 303) (for a full discussion on this issue, see Section 6.7 of the grammar book). Since this distinction is not the concern of this dissertation, I will simply stick to their classification and make slight modifications as mentioned above.

6.3.3.1 Between verbs and nouns in simple prepositions

In the following tables (both 6.12 and 6.13) the 'VERB' column contains the verbs that are related to their corresponding nouns in prepositional phrases which are listed in the 'PREP' column. 'C' stands for COLEC and 'L' stands for LOCNESS. In cases where there are multiple prepositional phrases, they will be listed below the first one.

Table 6. 12 Frequencies of 10 verbs (both in lemma and inflective forms) and some of their corresponding prepositional phrases in COLEC and LOCNESS

VERB	C	L	PREP	C	L
ARRIVE	34	24	on arrival	0	2
CHOOSE	118	140	**by choice**	1	2
CONTROL	139	55	in control (of)	0	4
			out of control	0	9
			under control	4	1
DECLINE	59	10	in decline	0	1
INCREASE	704	211	on the increase	0	3
OPERATE	18	12	in operation	0	3
PROGRESS	15	16	in progress	0	2
REQUEST	7	5	upon request	0	1
RETURN	28	35	upon return	0	1
RISK	11	28	at risk	0	4
Total	1133	536		5	33

Among the 10 verbs in Table 6.12, only two are found to have matching prepositional phrases in COLEC, i.e. *by choice* and *under control* (as highlighted in bold). Altogether there are only five cases of prepositional phrases in use in COLEC whereas there are as many as 33 cases in LOCNESS. The ratios of verb use and the use of prepositional phrases in the two corpora are as follows:

Table 6. 13 Total frequencies of verb use and noun use in prepositional phrases of 10 V-N pairs and their ratios in COLEC and LOCNESS

Corpus	Verb	Noun	Ratio
COLEC	1133	5	227:1
LOCNESS	536	33	16:1

Table 6.13 shows that as far as these ten V-N pairs are concerned there is less likelihood of the COLEC writers using nouns in prepositional phrases (only one chance in every 227 verbs). In contrast, the LOCNESS writers are using nouns in prepositional phrases much more frequently (one noun use in every 16 verb use). Again, apparently, NSs are using a much bigger proportion of prepositional phrases than NNSs here. The fact that most of the verbs in Table 6.12 are not adequately matched by their corresponding prepositional phrases indicates the overall tendency of using verbs by COLEC writers when there is a possible choice between verbs and nouns in prepositional phrases.

The following pairs of sentences below show the option tendency in the two groups of writers. The first sentence comes from COLEC and the second from LOCNESS.

Pair one:

1) [if we] can't *control* [our] mind, […] we can't do anything at all.

2) It is essential that society examine these arguments and then decide on what is acceptable and what is not acceptable before it gets *out of control*.

Pair two:

1) The population is *increasing* and the industry demands more and more water.

2) Schools and some hospitals, households are already publicised as "beef free" and this is *on the increase* causing a fall in the demand for beef in the U.K.

It seems that the COLEC examples could well be rephrased as follows if the learners wish to use nouns in prepositional phrases:

1) If our mind gets *out of control*, we can't do anything at all.
2) The population is *on the increase* and the industry demands more and more water.

The previous examples suggest that the learners' English is more verb-oriented than noun-oriented than the NSs when verbs and nouns in simple prepositions are considered. In the next

section I will continue to probe into the selection trend between verbs and nouns in complex prepositional phrases.

6.3.3.2. Between verbs and nouns in complex prepositions

The previous part of this section (6.3.3.1) has dealt with 10 verbs and their matching nouns in simple prepositions. The following sections will concentrate on 15 nouns in complex prepositions: *in + NOUN + of*.

Table 6. 14 Frequencies of 15 verbs and their corresponding nouns in the prepositional phrase structure (*in* + NOUN + *of*)

VERB	C	L	PREP	C	L
BREACH	1	0	in breach of	0	5
CHARGE	16	6	in charge of	3	4
CONTROL	139	55	in control of	0	3
DEFEND	6	20	in defense of	0	1
EXCEED	6	6	in excess of	0	1
FACE	228	105	in face of	4	1
FAVO(U)R	2	14	in favo(u)r of	3	39
JUDGE	13	74	in judgement of	0	1
MEMORIZ(S)E	15	2	in memory of	0	1
NEED	646	304	in need of	6	4
PROTECT	157	45	in protection of	0	1
PURSUE	23	18	in pursuit of	0	1
SEARCH	20	20	in search of	1	12
SUPPORT	29	138	in support of	0	6
VIEW	3	81	in view of	2	1
Total	1304	888		19	81

Similarly, as Table 6.14 shows, COLEC writers are not using as many prepositional phrases as LOCNESS writers in these 15 verbs. The ratio of verbs and prepositional phrases are as follows in Table 6.15:

Table 6. 15 The total frequencies of verb use and noun use in prepositional phrases of 15 V-N pairs and their ratios in COLEC and LOCNESS

Corpus	Verb	Noun	Ratio
COLEC	1304	19	69:1
LOCNESS	888	81	11:1

As far as the verbs in Table 6.14 are concerned, there is one case of prepositional use in every

69 verbs in COLEC. However, in every 11 verbs there will be one prepositional structure (*in* + NOUN + *of*) found in LOCNESS.

So far, a very general, but overwhelming, impression of the learner English is that the COLEC learners tend to use verbs more often than nouns compared with the NSs. However, this does not mean that the learners are not using nouns in prepositional phrases at all. In the 15 verbs under study (Table 6.14) there are five verbs (frequency ≥ 2) with relevant prepositional use: *charge*, *face*, *favo(u)r*, *need* and *view*. As the following instances show, the learners have reached a considerable level of expertise as far as these few words are concerned. To understand how properly COLEC writers are using these prepositional phrases, it is preferable to look at the concordances in detail. For the convenience of understanding the meaning to be expressed, complete sentences are provided instead of the concordances. In the cases where the meaning is not clearly visible in one sentence, a longer context will be provided. Some of the misuse is corrected in square brackets. Not all cases of misuse are pointed out and corrected because some of them do not affect interpretation.

1. *in charge of*

1) The department which is *in charge of* business should improve the strength on striking producing fake commodities.

2) When we find the people who is buying or selling the fake commodity, we should stop him in some practical way. Such as [call] the department *in charge of* it.

3) How to [prevent] fake commodities doing harm to the people, the society and the country. Firstly, to set up [a] union which is *in charge of* controling fake [commodities].

2. *in face of*

1) *In face of* danger, we must have self-confidence and devote greater efforts to it.

2) You also must expose the "real you" that hides in the "surface you" as if you must stand *in face of* the others without anything on.

3) Success is tempting to everyone, and it is the fruit of one's sweat, struggle, even life. so we couldn't obtain it easily *in face of* difficulties or danger.

3. *in favo(u)r of*

1) Of course, this phenomenon has many advantages and disadvantages. But I am *in*

favour of it.

2) Some people think that "8" can bring good luck… Yet, other people don't think so… I am *in favor of* the latter.

4. *in need of*

1) In fact, people are *in need of* fresh water to a great extent.

2) The development of modern industry is *in need of* more water, too.

3) With the increasing of population, more and more people are *in need of* fresh water.

5. *in view of*

1) *In view of* the above mentioned drawbacks of cars, the use of cars should be well controlled.

2) If we did it indeed, we would learn more about society. So we could fit the society very well in the future. *In view of this*, I will read newspaper and watch TV everyday, and …

Surprisingly, COLEC writers are using these prepositional phrases fairly well except in a couple of cases. The problem is that they are not using such a good variety as the LOCNESS writers. As can be seen from Table 6.14, the LOCNESS writers use all the 15 nouns in prepositional phrases whereas the COLEC writers are found to be using only a small number of them. Furthermore, if we look at LOCNESS, there are as many as 12 cases of *in search of* whereas there is just one case of this sequence in COLEC. This noun use in prepositional phrase might be simulated by the learners to replace their verbal use:

Figure 6. 1 The concordances of *in search of* from LOCNESS

1 ted how you should act. Man is always in search of knowledge & truth. He wants
2 **He believes that as man is constantly in search of truth and that death is the**
3 e, where the absurdist man lucidly goes in search of knowledge and unity. But we
4 lves freely or to leave their homelands in search of a new life -- Many peo
5 g man who is completely innocent and is in search of his ideal. However this wit
6 ely enough money to live on to waste it in search of thier fortune. Almost weekl
7 taught. With this optimism he sets out in search of his ideal, this being Cun
8 **stors often seem to spend their time in search of a new thrill, some sort of**
9 Oreste is pursuing a quest for truth, in search of the meaning of freedom, in
10 de. Candide wanders all over the world in search of Candide. Sometimes he is fo
11 imisme. Candide has travelled the world in search of Cunégonde, with the hope th
12 ung hero's travels throughout the world in search of happiness. On his travels t

If we look at the following verbal use by the learners and compare these examples with the

noun use in the prepositional structure in Figure 6.1, the learners' overuse of verbs rather than nouns is better illustrated.

1) They have no time for work, they pay too time to *search* for a job.

2) And on the other hand, scientists are trying to *search* for better source of fresh water.

Based on the natural use by the NSs (Lines 2 and 8 in Concordance(s) 6.1), the learners may like to rephrase the two sentences as follows:

1) They have no time for work, they spend too time *in search of* a job.

2) And on the other hand, scientists are constantly *in search of* better source of fresh water.

To sum up, COLEC writers are found to use verbs tremendously when it is possible for them to choose the nominal use. Since the learners are producing much fewer cases of nouns in prepositional phrases than nouns in general (as studied in 6.3.1 and 6.3.2), it is felt that the learners have more difficulty in using nouns in prepositional phrases.

6.4 Discussions

The trend among the learners as a group towards opting for verbs over nouns compared with the NSs (both the LOCNESS writers and the GSL) seems very obvious. The quantitative study concerning the degree of the learners' choice of verbs over nouns (no matter whether norbs or not) supports the findings of Altenberg (1996, cited in Ringbom 1998b: 50) in that learners use a higher proportion of verbs and a lower proportion of nouns. The tendency for the NSs to use nouns in prepositional phrases could also be seen as able to account for the clustering of nouns and prepositions in ''association patterns' by Biber (1996: 173, cited in Hunston 2002: 164; see 2.7.1 of this thesis for details).

As implied in the findings, the significance of this analysis goes far beyond linguistic research. This research touches upon the issue of the implicit pattern of NSs in selecting a particular POS, which has been hitherto largely ignored. Even though Sinclair has coined the word *norb* for a word that functions both as verb and noun, it seems that no attempt has been made to create a word for those words that function both as verbs and adjectives such as *WARM* and *SINGLE*; not to mention the words that function both as verbs and prepositions such as *LIKE*

and *ROUND*. The scarcity of the terms required for linguistic research points to the need for further analysis and more pavement work for more serious and large-scale investigations into this interesting area of language.

This research, perhaps more importantly, also reveals the urgent need to raise the awareness of learners (not only the Chinese learners, but perhaps most learners considering that Swedes share the same trend with the Chinese learners). Learners should be made aware of at least several points as follows:

(1) Some English vocabulary can serve as more than one POS, examples being *SUPPORT* and *VIEW*. Once the first POS of such a word becomes stable, more effort is needed to expand the learner's knowledge about the use of other POSes (fossilisation at this stage seems to be hindering language acquisition).

(2) As a whole, NSs use more words in their noun function than in their verb function (no matter whether it is within the words sharing the same form or not), but there are some words that are particularly oriented towards verbal function, such as *SAY* and *TRY*.

(3) The learners' current trend in POS selection could be rectified by consulting the general trend of NSs (but this is not to say that every word should be rigidly followed without taking other factors into consideration such as genres and topics).

(4) The learners' current trend in POS selection could be better rectified in phraseology or in context (such as in the phrases like *in favour of* or *in search of*) rather than in isolation.

For researchers who wish to grade the level of group learner English and diagnose the problems of the current learners as a group, this research has offered some possible solutions. The ratio of the verbal use to the nominal use could be used as a parameter in deciding whether an amalgamation of learner English should be graded as advanced level, intermediate level or elementary level. If required, the researchers could make finer distinctions by giving more grades. The closer the noun–verb ratio of the learner English is to that of the NSs, the higher the level of the learner English should be. In the same vein, the further away the ratio of the learner English is from that of the NS, the lower the level of the learner English should

be.

6.5 Conclusion

This chapter has studied in detail the choice between verbs and nouns by the two groups of writers. It has been demonstrated that the learners have a stronger tendency to use verb function than noun function when the NSs might use more noun function than verb function. The findings could be used not only to further our understanding of the English language as a whole in terms of the writers' selection among POS words, but also to reveal the disparity in POS selection within norbs and within the words forming verb and noun pairs. The significance of this research in English language education is discussed in an attempt to raise the awareness of English learners. It is hoped that when the learners have extended their noun use by consulting the general trend among the NSs, their language production will be much closer to the language produced by the NSs.

Chapter Seven

Using Patterns and Phrases to Interpret Learner English

7.1 Introduction

The previous three chapters have contributed to the demonstration of how a corpus-based approach can best play its part in the study of learner English. They looked at the area of use and non-use by the COLEC writers, i.e. whether these writers use certain lemmas or not, whether they use certain word forms or not, and whether they use the noun function of certain norbs or not. The information obtained from such a panoramic perspective shows some disparities between the learner English and the NS English. It helps to give a clear and full list of what new items of vocabulary should be tried and practised by the learners; yet it does not seem to have enough to say about how individual verbs are used by the COLEC writers. For example, how is the verb *KEEP* used by the learners? How is the learner English in COLEC similar to or deviant from the NS English in LOCNESS? A need for further information in the case of individual words is validated by a study of the verb *MAKE* by Altenberg and Granger (2001). Their research (*ibid*.: 182) suggests that L1 constructions have an impact upon the syntactical patterns in L1 production. In the words of Altenberg and Granger (2001: 182), "Learners who are unfamiliar with [the] alternatives [in L2] are likely to overuse the dominant pattern and treat it as a lexical-grammatical 'teddy bear', especially if it is easy to transfer from their native language." Information on such a feature is useful not only for the examination of group learner English features but also for practice in ELT.

The research by Altenberg and Granger (2001) is essentially an integration of two perspectives on the linguistic behaviour of the verb *MAKE*. One is syntactic patterns such as '*MAKE* somebody believe sth' and '*MAKE* sth possible', and the other is the collocates of the verb *MAKE* such as '*MAKE* decisions' and '*MAKE* furniture'. Within one frame of work, it would be difficult to see many details of the syntactic patterns and the collocates. In order to make a deeper and more detailed investigation of the learner English, I am going to look at the two perspectives separately, i.e. syntactic patterns in one chapter and collocates in another. Since the verb *KEEP* is rich in syntactic patterns, it is chosen for the former. For the study of

collocates, I have chosen the verb *TAKE* because it has abundant collocates. Both of the verbs *KEEP* and *TAKE*, also rank high in the verb lemma lists (see Appendix 2 and Appendix 3). Another reason for having chosen these two simple words is that it is becoming a commonly held view that simple words or 'smallwords' are playing an essential role in successful communication (see Sinclair 1991, Hasselgren 2002: 144, and many others).

The theories of Hunston and Francis (1999) in pattern grammar will be used (for details see 7.3). It is hoped that we can see better how close the learner English approximates to the NS English or how far it deviates from it, and that we can construct a brief profile of the lemma *KEEP*. Since some constructions of a verb cannot be properly expressed by patterns, I will also use the term 'phrase' to cover the non-pattern uses. As mentioned in Chapter Two, the terms most often used in describing the frequency disparity between learner English and NS English 'overuse' and 'underuse' are not without problems. This chapter discusses the problems at full length, while identifying the similarity and disparity as mentioned above. The research questions of this chapter run as follows:

(1) What are the similarities and disparities of the learner English and the NS English in the patterns and phrases of *KEEP*, not only in frequency but also in detailed performance?
(2) What are the disadvantages or problems of using 'overuse' and 'underuse' in describing the frequency difference in CIA? Is there a better way of doing this?
(3) What is the pedagogical significance of such a study?

7.2 Introducing the ratio relationships between the two corpora

In the pioneering work conducted by the contributors to Granger's collection (Granger 1998) *Learner English on Computer* the terms 'overuse' and 'underuse' are frequently used to refer to the disparity in frequencies between learner English and NS English. Influenced by this initiative, most of the CIA studies today follow the terminology 'overuse' and 'underuse'. When people talk about 'overuse' or 'underuse' for a particular item, they imply that learners are using such an item wrongly. When people say a particular word is 'overused' by learners, they are implying that on some of the occasions it should not be used. By the same token, when people talk about the 'underuse' of the word, what they are implying is that learners do not use the word when they should use it. Apart from this, while comparisons between a

learner corpus and a NS corpus are the foci of CIA, something important has unfortunately been ignored, i.e. the comparison between the large-frequency and the small-frequency items within the learner corpus itself. I would argue in this chapter that:

1) The use of 'overuse' and 'underuse' is biased by over-generalisations;
2) There is a need to investigate the roles that large-frequency and small-frequency items in a learner corpus play in the feature identification of learner language. In order to obtain a better idea of how learner English stands in relation to NS we need to compare different sets of figures.

To illustrate my first argument, instead of using 'overuse' and 'underuse' in a broad sense, I would propose the following eight sets of comparisons to describe the relationship between learner English and NS English.

1) a large frequency in COLEC vs. a large frequency in LOCNESS
2) a large frequency in COLEC vs. a small frequency in LOCNESS
3) a small frequency in COLEC vs. a large frequency in LOCNESS
4) a small frequency in COLEC vs. a small frequency in LOCNESS
5) no frequency in COLEC vs. a small frequency in LOCNESS
6) a small frequency in COLEC vs. no frequency in LOCNESS
7) no frequency in COLEC vs. a large frequency in LOCNESS
8) a large frequency in COLEC vs. no frequency in LOCNESS

Of course, there are items that cannot be found in either of the corpora. Since that kind of situation provides no information for a comparative research study, it will be excluded from consideration. Also, different types of ratio relationship may have different values in reading and interpreting learner English in a comparative setting. The following assumptions are intuition-based and will need testing and certification.

In the first situation (a large frequency in COLEC vs. a large frequency in LOCNESS), the large frequency in both corpora shows that the item under study is frequently used in NS English and NNS English as well. The items in this area are supposed to represent the most widely shared part of the collective learner English. A useful question to ask in the category is whether a similarity in frequency would guarantee a similarity in detailed use.

The second situation (a large frequency in COLEC vs. a small frequency in LOCNESS) indicates overuse by the learners of the item under study. This could be for many reasons. It could have resulted from the difference in topics in the two corpora. Different topics produce different core words, as is constantly shown in Chapter Four and Chapter Five. It could also be caused by a sort of collective tendency in selecting words which are passed down from their English teachers. If teachers and authors of teaching materials know which words are currently used too frequently (comparatively speaking), they may prepare new materials to curb the learners' tendency by looking for appropriate alternatives in the NS corpus and asking their learners to improve on them.

The third situation (a small frequency in COLEC vs. a large frequency in LOCNESS) indicates items which are inadequately used by the learners under study. For learner English to advance, this area should be used to set out the learning tasks for learners if the abundant use by the NSs is not the result of topic requirements. It can be tentatively proposed that the occurrence of an item in this area is more likely to occur in high-level English essays.

In the fourth situation (a small frequency in COLEC vs. a small frequency in LOCNESS), the item under study may be infrequent in English in general. If an item is infrequently used by the NSs, it could mean that there are not enough opportunities for them to write it, or in other words, not much necessity to write it. On the side of the learners, it is probable that most are not aware of such a usage in English.

The fifth situation (no frequency in COLEC vs. a small frequency in LOCNESS) is significant in the sense that it may point to the area where more efforts are needed if progress is to be made by the learners. This frequency relationship resembles the third type in that learners are not using certain items of the language as much as the NSs do.

The sixth situation (a small frequency in COLEC vs. no frequency in LOCNESS) is unexpected because normally it is more natural for an item to appear in NS English but not to appear in learner English. This frequency relationship resembles the second one in that an item is being used more often by learners. The difference between them is that the item used

in the second situation is more likely to be correct whereas in the sixth situation the item used is much less likely to be properly produced.

In the seventh type of ratio relationship (no frequency in COLEC vs. a large frequency in LOCNESS), there is, of course, no occurrence in COLEC but there are many in LOCNESS. If this ratio is not for the reason of topic difference, then the non-use in COLEC suggests a poor area of mastery by the learners, and suggests that more effort is needed on their part if they are to advance their English production.

In the eighth type of ratio relationship (a large frequency in COLEC vs. no frequency in LOCNESS), the situation is reversed, i.e. there is no use in LOCNESS but a lot of use in COLEC. If such an item is correct in English, certainly we would not believe that the NSs are not able to produce such an item. If it is not for the reason of topic, the learners may like to consider proper ways of expressing the same idea that are often used by NSs.

As a whole, it is reasonable to assume that the first three types of frequency relationships as explained above seem to have priority over the others because they are better as diagnostics of the problems of learner English. While the first type of frequency relationship is expected to indicate an area of comparative maturity of the learner English, the second and third types of frequency relationship are expected to reveal the areas where learner English deviates from the NS English if it is not for the reason of topic requirement.

The relationship between large frequency and small frequency could be interpreted from two perspectives. From a comparative view between a learner corpus and a NS corpus, there is every reason to believe that the higher the frequency of an item in the control corpus, the more confident the researcher becomes about the significance of its absence or presence in the learner corpus. Likewise, the lower the frequency in the control corpus, the less confident one can be about the significance of its absence or presence in the learner corpus. In the case of LOCNESS and COLEC, one can be confident that if a particular item occurs frequently in LOCNESS, it would be expected to occur frequently in COLEC. However, if a particular item occurs only a few times in LOCNESS one cannot be very predictive about the possibility of its occurrence in COLEC. If the relationship is viewed in such a way as to compare the large-

frequency and small-frequency items in the learner corpus (because the essence of learner English studies lies eventually in the learner corpus), the large-frequency items represent the popularity and homogeneity of the collective learner English. Given similar tasks and under similar circumstances, other writers with the same background are very likely to produce the same or similar items. Because of this problem, there is a lack of information for those researchers who wish to find some information for language evaluative purposes. But it is worthwhile to investigate whether some useful information could be gleaned from the improperly used cases of the large-frequency items. As far as the low-frequency items are concerned, it is worthwhile to investigate whether there exists a relationship between the use of a rarely used item (meaning whether it is used frequently or infrequently by the NSs) and the overall level of a composition in which the item occurs. If a correlation could be found it might shed some light on language testing and ELT.

Due to the many problematic features of learner English, it must be borne in mind that for a learner corpus frequency merely reveals how many times a certain item occurs in the corpus. Even though it is a useful indicator of a certain degree of achievement in English production by learners, especially when the frequency is large, it is not necessarily the case that large frequency reflects mastery since learner English is full of unnatural expressions. The reason is simple and familiar: frequency hides errors.

7.3 Defining 'pattern' and 'phrase'

To describe a language by patterns is a significant contribution to the enrichment of the received wisdom in understanding how a language works. Its significance actually goes far beyond theoretical linguistics. This methodology "represents a meeting-point between the concerns of pedagogy – what it is that learners need to know – and those of theory – how the English language can most satisfactorily be described" (Hunston and Francis 1999: 36). As this study will gradually show, the treatment of patterns concerns not only theoretical issues but also pedagogical issues. The word 'pattern' has been used to mean different things by different people. In this chapter and this thesis, the use of pattern is in line with the theories of Hunston and Francis (1999). The following is a definition of 'pattern' in their own words (1999: 3):

> Briefly, … a pattern is a phraseology frequently associated with (a sense of) a word, particularly in terms of prepositions, groups, and clauses that follow the word. Patterns and lexis are mutually dependent, in that each pattern occurs with a restricted set of lexical items, and each lexical item occurs with a restricted set of patterns. In addition, patterns are closely associated with meaning, firstly because in many cases different senses of words are distinguished by their typical occurrence in different patterns; and secondly because words which share a given pattern tend also to share an aspect of meaning.

The term 'phrase' in this chapter is used very loosely to refer to any combination of several words which cannot really be grouped into any of the patterns in Table 7.1. It must be admitted that sometimes there is not an absolute demarcation between a pattern and a phrase. The way I have identified the patterns and phrases might not be accepted by other researchers. Since the major purpose of this research is to see how the traditional terminology ('overuse' and 'underuse') could be improved, I would not be over-strict with the distinction between the two classes of construction.

7.4 Looking at the patterns of KEEP in COLEC and LOCNESS

There are 170 occurrences of *KEEP* in LOCNESS and 392 in COLEC. Since the texts of COLEC are composed of examination compositions with instructions, many students repeat the key words (such as *keep fit*) in the guidance for writing. Therefore, I have deleted the cases (40) which I think do not belong to the learners themselves. As a result, the figure of the lemma *KEEP* in COLEC for comparison is 352. This section attempts to demonstrate how the language data can be interpreted by means of segmenting the raw data into patterns of a high degree of complexity.

7.4.1 Interpreting the frequency relationships between COLEC and LOCNESS

In order to present the learner English in relation to the NS English, I have sorted the raw data on *KEEP* in the two corpora into patterns and listed the non-patterns as phrases (see Table 7.1). 'RF' stands for 'raw frequency' and 'NF' stands for 'normalised frequency'. The percentage (%) refers to the raw frequency divided by the total frequency of *KEEP* in the corpus.

Table 7. 1 The frequencies of *KEEP* in its patterns and phrases

	Pattern	LOCNESS			COLEC		
		RF	**NF**	**%**	**RF**	**NF**	**%**
1	V n	46	71	27.1	65	65	18.5
2	V n adj	42	65	24.7	29	29	8.2
3	**V n prep/adv**	25	38	14.7	2	2	0.6
4	V -ing	12	19	7.1	50	50	14.2
5	V n from -ing	11	17	6.5	12	12	3.4
6	V n -ing	6	9	3.5	1	1	0.3
7	V n (away) from n V n out of n	4	6	2.4	5	5	1.4
8	V n down V down n	4	6	2.4	1	1	0.3
9	V up n V n up	2	3	1.2	5	5	1.4
10	V to n	1	2	0.6	0	0	0.0
11	V n as n	1	2	0.6	0	0	0.0
12	V on -ing	0	0	0.0	37	37	10.5
13	V adj	0	0	0.0	21	21	6.0
14	V on it V it on	0	0	0.0	3	3	0.9
15	V up -ing	0	0	0.0	5	5	1.4
	SUB-TOTAL	**154**	**238**	**90.6**	**236**	**236**	**67.0**
	Phrase	**LOCNESS**			**COLEC**		
		RF	**NF**	**%**	**RF**	**NF**	**%**
1	V in mind n V n in mind	2	3	1.2	15	15	4.3
2	V up with n	1	2	0.6	24	24	6.8
3	V an eye (on n)	1	2	0.6	5	5	1.4
4	V possession of n	1	2	0.6	0	0	0.0
5	V n on stand	1	2	0.6	0	0	0.0
6	V n in view	1	2	0.6	0	0	0.0
7	V n to a minimum	1	2	0.6	0	0	0.0
8	V n to oneself	1	2	0.6	0	0	0.0
9	V n in existence	1	2	0.6	0	0	0.0
10	V an open mind	1	2	0.6	0	0	0.0
11	V abreast of n	1	2	0.6	0	0	0.0
12	V n under control	1	2	0.6	0	0	0.0
13	V n to what it was	1	2	0.6	0	0	0.0
14	in keeping with	1	2	0.6	0	0	0.0
15	the kept woman	1	2	0.6	0	0	0.0
16	V in touch with n	0	0	0.0	13	13	3.7
17	V n in good health	0	0	0.0	5	5	1.4
18	V on	0	0	0.0	3	3	0.9
19	V track of n	0	0	0.0	2	2	0.6
20	V pace with n	0	0	0.0	2	2	0.6
21	V at it	0	0	0.0	2	2	0.6
22	V n in the dark	0	0	0.0	1	1	0.3
23	V company with n	0	0	0.0	1	1	0.3
24	V n in order	0	0	0.0	1	1	0.3
25	V contact with	0	0	0.0	1	1	0.3
	SUB-TOTAL	**16**	**31**	**9.4**	**75**	**75**	**21.3**
	MISUSED	**0**	**0**	**0.0**	**41**	**43**	**11.6**
	TOTAL	**170**	**263**	**100.0**	**352**	**354**	**100.0**

The patterns are displayed mainly in line with Hunston and Francis (1999). The code "V" in upper case refers to the lemma *KEEP*, "n" includes common nouns and personal pronouns, "-ing" means the "ing" form of a verb. The code "adj" refers to adjective, "adv" refers to adverbs and "prep" refers to prepositions. The fixed constituents are italicised. For example, the pattern "**V n *from* –ing**" means that the verb *KEEP* is followed by a noun, and followed by the preposition "from" (rather than any other prepositions), and then followed by a verb in its "-ing" from. To decide whether a sequence is treated as a phrase or a combination of individual words is mainly based on the *CCED* (*Collins Cobuild English Dictionary*, 1995). In Table 7.1 the patterns and phrases of *KEEP* are provided first in the order of frequency in LOCNESS and then in the order of frequency in COLEC. There are altogether 15 patterns and 25 phrases identified in the two corpora. As many as 41 cases are improperly used in one way or another by the learners.

The following section will attempt to interpret the frequency relationships between the two corpora in terms of the use of patterns and phrases based on the assumptions put forward in the introduction of this chapter. But before this interpretation starts, it is necessary to set a value to the notions of 'large' and small' when used of the frequencies. Obviously normalised figures could be used for this purpose. Yet since the purpose of this study is to see how a particular pattern or phrase is used in relation to other patterns and phrases, percentages will be used instead of the raw or normalised frequency. A percentage of 5% or above will be regarded as 'large frequency' and a percentage lower than 5% will be regarded as 'small frequency'. Of course, this demarcation is arbitrary, based only on the author's research experience and the size of the corpora under study.

7.4.1.1 A large frequency in COLEC vs. a large frequency in LOCNESS

There are three patterns that fall into the category of 'a large frequency in COLEC vs. a large frequency in LOCNESS', i.e.

1) **V n** (In my mind that stands for the South fighting to *keep slavery*.)
2) **V n adj** (the reporter must decide if he/she will *keep the source secret*.)
3) **V –ing** (If these falls in sales *keep going*, then it is possible to …)

Table 7. 2 The majority of the nouns in the pattern '*KEEP* n' in LOCNESS and COLEC

LOCNESS	COLEC
a baby, a house, money, a significant number of cattle	the photo, some fresh water
a price, the same philosophy, civil peace	world peace
the tradition, slavery, an institution, boxing, their games, the National Lottery, the Monarchy, the presidency, the identification, our cultural identity, their advantage	the reform and open policy
mutual trust, friends, the support, their interest	
control, order	
score, records	
	a job, a skill
	one's health, a good health
	smile, a young face, happiness
	a clear mind (brain, head), a good emotion
	silence, easy heart
	the balance, a relationship, a position
	honesty, secrets

Due to limitations of space, I will concentrate on the first one, attempting to see whether a similar frequency between COLEC and LOCNESS guarantees a similarity in the detailed use. If we compare the distribution of the patterns and phrases in the two corpora, it is easy to see that the most similar usage in learner English and NS English lies in the pattern **KEEP n** (27.1%) in LOCNESS and 18.5% in COLEC)**,** in terms of the similarity in the normalised frequencies of the two corpora (even though there is a fairly large difference between the actual percentages in the two corpora). The majority of the nouns (represented by **n** in the pattern **KEEP n**) are listed in Table 7.2. The classification is mainly based on whether the objects to be kept are concrete (such as 'a baby' in LOCNESS and 'some fresh water' in COLEC) or abstract (such as 'civil peace' in LOCNESS and 'world peace' in COLEC). Special nouns that contribute to idioms and relatively fixed collocations such as 'control' and 'records' are singled out from other nouns. A detailed look at the classification shows that there are very few words or concepts that are shared in the two corpora.

Contrary to what is assumed earlier in 7.2 (the items in this area are supposed to have the most shared performance), there is actually a huge disparity in the detailed uses of the nouns or noun phrases. Therefore, it can be concluded that large frequency in a particular pattern (or

phrase) indicates only that this pattern (or phrase) is used fairly often by the group of writers. It has nothing to say about its appropriateness *per se*. Frequency figures must be supported by detailed analysis of concordances. Further examination is useful if more information is to be obtained concerning the detailed similarity and disparity of the learner English and the NS English.

7.4.1.2 A large frequency in COLEC vs. a small frequency in LOCNESS

There is one phrase that falls into the category 'a large frequency in COLEC vs. a small frequency in LOCNESS', i.e.

1) **KEEP *up with* n** (In order to *keep up with the times*, we should get to know the world …)

Figure 7. 1 All the correctly used cases of '*KEEP up with* n' in COLEC

1 faster. If we still close, we wouldn't keep up with advanced country. So reform
2 fresh water. Only through this, we can keep up with enough fresh water, and sur
3 d outside our school yard. In order to keep up with the changing world, I will
4 ill be built beautifully by us. We will keep up with the develoed country in the
5 come into the society and let our ideas keep up with the development of the soci
6 e certain pratical energy. In order to keep up with the development of the soci
7 w the world outside the campus helps us keep up with the development of the soci
8 outside the campus. Otherwise, we can't keep up with the development of society.
9 people after they graduate, they can't keep up with the development of society.
10 ogy are developing quickly. So we can't keep up with the fast-paced society if
11 the job often is bad, because you don't keep up with the knowledge of the work.
12 he world outside the campus, they can't keep up with the pace of society in the
13 world outside the campus, we will can't keep up with the rapid advance of the so
14 w the change of the world, how could he keep up with the step of the times? The
15 ore and more fresh water is required to keep up with the steps of the developmen
16 and read our limit books we should not keep up with the time. There are many
17 we may say pridely that we have already keep up with the time. There are many m
18 tudy, but it is not enough. In order to keep up with the times, we should get to
19 to turn out their abilities, they can keep up with the times. And it's also po
20 happened in our country or abroad. And keep up with the times. on the other han
21 e″. About 1960's , the leader want to keep up with the UK and USA. So they tak
22 chieve a lot of success. In some way we keep up with the west country. Such as..
23 t is important that we college students keep up with the world outside. Otherwis
24 school. If I didn't do that, I couldn't keep up with the world. And as we known

There are as many as 24 cases in COLEC but only one in LOCNESS of the phrase *KEEP up with* (see Figure 7.1). When the concordance lines are examined, it seems that most of these are properly used by the learners. A large proportion of the concordances are 'keep up with

the development of the society' and 'keep up with the times', which are acceptable to NSs. It may be speculated that this pattern enjoys a kind of popularity and homogeneity among this group of learners. For the purpose of improvement, teachers may raise the learners' awareness that this phrase may have alternatives.

7.4.1.3 A small frequency in COLEC vs. a large frequency in LOCNESS

There is one pattern that falls into the category of 'a small frequency in COLEC vs. a large frequency in LOCNESS', i.e.

1) **KEEP n prep/adv** (The seatbelts are designed to […] *keep you in your seat*, so that …)

The following are some examples from LOCNESS:

1) The legalisation of marijuana among other drugs, would *keep some people out of the streets*.

2) This type of action is what *keeps millions of viewers on the edge of their seats* day in and day out.

3) there are computer games that don't need any brainpower whatsoever, just *keeping your finger on a button.*

The following are some uses of the pattern **KEEP n adj/adv** from LOCNESS:

1) I think it' s time we stop wasting money on *keeping drugs illegal* when it does not even work.

2) We talked some more in the lobby but we had to *keep our voices down*, out of respect.

As discussed earlier in Section 7.2, if a learner can perform successfully in such an area, it may well be that this learner's English is of a high level. Therefore, this is an area from which significant information can be drawn.

7.4.1.4 A small frequency in COLEC vs. a small frequency in LOCNESS

There are four patterns and two phrases that are in the category 'A small frequency in COLEC vs. a small frequency in LOCNESS'. These patterns and phrases are listed as follows:

1) **KEEP n –ing** (The major influence *keeping boxing going* is ….)
2) **KEEP n (*away*) *from* n / KEEP n *out of* n** (we should *keep the resource of fresh water out of pollution* …)
3) **KEEP n *down* / KEEP *down* n** (Foxs are not needed to *keep down the rabbit population* …)
4) **KEEP *up* n / KEEP n *up*** (However the company … are trying to *keep up the illusion of mystery and excitment* by increasing the jackpot to 40 million …)
5) **KEEP *in mind* n / V n *in mind*** (Let us … *keep in mind the rudimentary beliefs* …)
6) **KEEP *an eye on* n** (the satellite *keeps a watchful eye on all of us*.)

Since the frequencies of the patterns are not large enough, researchers have very little confidence in reaching a conclusion about the knowledge of a given pattern by the whole community of learners. It would be wrong to assert that one occurrence in a learner corpus will indicate mastery on a large scale by the whole community. In fact, even with a large number (no matter how large it is), it would be wrong as well to assert that this particular item reflects the linguistic proficiency of the whole community. When we attempt to ascertain the extent to which the learners have gained mastery of the TL, we are dealing only with likelihoods, not certainties. Only possibility and likelihood are at the centre of the issue when we assess the state of the learners' mastery of the target language.

7.4.1.5 No frequency in COLEC vs. a small frequency in LOCNESS

There are 14 patterns and phrases that are in the category of 'no frequency in COLEC vs. a small frequency in LOCNESS'.

1) **KEEP *to* n** (Hugo is prepared to *keep to his ideals* whatever the cost …)
2) **KEEP n *as* n** (Many professors wanted to *keep universities as 'la finalité culturelle'*.)
3) **KEEP *possession of* n** (… to ensure that the rich kept *possession of their goods & property*.)
4) **KEEP n *on stand*** (*Ambulances* are also *kept on stand* by at big events.)
5) **KEEP n *in view*** (*The absurd* … is man's only link with the world and *should be kept in view* and should form the basis of decisions as to how to live …)
6) **KEEP n *to a minimum*** (*The acts of violence* are …*kept to a minimum* …)

7) **KEEP n *to oneself*** (Are the players getting all of the money, or are the managers *keeping it all to themselves*?)
8) **KEEP n *in existence*** (it is not worth *keeping euthanasia in existence*.)
9) **KEEP *an open mind*** (both Pangloss and Martin use the facts to suit their systems rather than *keeping an open* and 'candid' *mind*.)
10) **KEEP *abreast of* n** (Finally, people need to continue to *keep abreast of* new developments …)
11) **KEEP n *under control*** (keeping drugs illegal helps to *keep them under control*.)
12) **KEEP n *to what it was*** (they would have to increase sales to *keep revenue to what it was*…)
13) ***in keeping with*** (his death is *in keeping with* the idea of him as an anarchist.)
14) ***the kept woman*** (He … finds her in Spain (living as the *kept woman* of a catholic …))

The small frequency of the items indicates that these items are sparsely used by the NSs and these items should not become a target for the majority of learners. However, supposing that these items could be produced by the learners, their language would be much more expressive than without them. Therefore, I propose that patterns and phrases in such a category deserve to be listed on the agenda for learners for the next phase of study.

7.4.1.6 A small frequency in COLEC vs. no frequency in LOCNESS

There are 12 patterns and phrases in the category 'a small frequency in COLEC vs. no frequency in LOCNESS'. These patterns and phrases are listed as follows:

1) **KEEP it *on/* V *on* it** (she said she would *keep it on* longer.)
2) **KEEP *on*** (intransitive) (So long as I *keep on*. I can master 3,650 words every year.)
3) **KEEP *in touch with* n** (By these means, we can *keep in touch with outside*.)
4) **KEEP *up* –ing** (If you *keep up practising* something, you will get a lot of skills about that.)
5) **KEEP n *in good condition*** (They could do more things to *keep themselves in good condition*.)
6) **KEEP *track of* n** (For me, I have to *keep track of the new development* in medical field …)

7) **KEEP *pace with* n** (students can *keep pace with what is happening home and abroad*.)
8) **KEEP *at it*** (In a word, if you *keep at it* and constantly draw a conclusion you will do it better.)
9) **KEEP n *in the dark*** (The deception *keep us in the dark* until we grow up …)
10) **KEEP *company with* n** (Many people *keep company with each other* through the convenient facilities)
11) **KEEP n *in order*** (Then I … learned to *keep my things in order* …)
12) **KEEP *contact with* n** (we must *keep contact with the society* constantly.)

It is always the case that what can be found in COLEC will also be found in LOCNESS. However, occasionally what occurs in the learner corpus has no match in the NS corpus. This is perfectly reasonable because there will be some disparity in any two corpora under comparison. Apart from this, there will always be a certain number of erroneous occurrences in the learner corpus.

7.4.1.7 No frequency in COLEC vs. a large frequency in LOCNESS

There are neither patterns nor phrases that are in the category of 'No frequency in COLEC vs. a large frequency in LOCNESS'. This seems to indicate that the learner English of COLEC is not drastically different from the NS English, otherwise there could be some patterns or phrases of this type of ratio relationship.

7.4.1.8 A large frequency in COLEC vs. no frequency in LOCNESS

There are two patterns in the category 'A large frequency in COLEC vs. no frequency in LOCNESS', i.e.

1) **KEEP** adj (I think it important for us to *keep calm* in this case.)
2) **KEEP *on* –ing** (You can't speak English freely unless you *keep on speaking* it every day.)

Due to the large number of examples of this pattern being used by the learners, it may be speculated that the pattern is widely shared by the group of learners. There is good reason to believe that a large majority of learners with the same background are most likely to be able to produce this pattern. The notion expressed by the pattern **KEEP *on* –ing** is a common one

in English. If the NSs do not use this expression very often, they should have something else to use. Given that NSs have a much larger vocabulary, they may use several alternatives.

7.4.2 Some reflections on the use of large-frequency items in the learner corpus

A large-frequency item reflects its popularity in the learner group under study. Since it is popular and shared by many learners, it is reasonable to believe that those who do not produce this popular item properly may be at an earlier stage of acquisition and therefore, that the level of these learners is likely to be lower. To test this hypothesis 10 correctly used occurrences and 10 incorrectly used occurrences are checked in the raw corpus of COLEC. The score and the writer's ID of each occurrence are provided in Table 7.3.

Table 7. 3 Some examples of the correct use and incorrect use of '*KEEP in touch with*" in COLEC

ID	Correct Use	M	ID	Incorrect Use	M
452823	I will *keep in touch with* them and communicate with each other.	15	451115	By *doing more touch with* the people in society…	12
650318	we should *keep in touch* with all sorts of information around us.	13	650517	to *keep touch with* the world outside.	9
451115	They only *keep in touch* with the knowledge in book.	12	453130	This is good way to *keep touch with* the society.	9
650514	we should also *keep in touch with* the senior or graduated college students …	12	640312	Without *keep touching with* the society…	9
451922	I should *keep in touch with* it.	10	650527	*have the touch with* the society.	8
650513	I should always *keep in touch with* the outside world.	11	650322	I will do a part-time job to *touch with* world outside.	8
440618	How can we *keep in touch with* outside?	9	650613	There are many ways to *keep in touch of* the outside the campus.	8
440618	By these means, we can *keep in touch with* outside.	9	440903	We seldom *get touch with* the society.	8
650527	Having realized where and how we can get help to *keep in touch with* the society.	7	no 0379	I must *keep touch with* the society.	7
452861	it can make them *keep in touch with* world.	6	431102	Because they want to touch with the new thing…	7
AVERAGE		10.4			8.5

The distribution of the patterns and phrases in the two corpora seems to suggest the following things:

1) Those who use a commonly used item (such as '**KEEP *in touch with* n**') have, on the whole, a higher score than those who have problems with the item;
2) It is very likely that those who use correctly an item commonly used by their peers

may have a high score (but not necessarily);

3) It is very unlikely that those who do not use a commonly-used item correctly will have a high score;
4) Since there are multiple factors contributing to a high score, one correct use contributes to a high score of a composition but does not automatically lead to it, and *vice versa*. The existence of a disparity between individual markers may contradict the general trend as stated above (1, 2 and 3).

7.4.3 Some reflections on the use of low-frequency items in the learner corpus

Even though small-frequency items do not give researchers as much confidence as large-frequency items, there might still be some potential for them to be used in learner English study. Considering the established view that advanced learners use more 'chunks' and phrases, I would like to investigate whether there is a co-relationship between an item of low frequency and the level of the whole composition in which the item occurs, Table 7.4 contains the concordances and marks of the items listed in 7.4.1.6.

Table 7. 4 The concordances and marks of some low frequency patterns and phrases in COLEC

SN	Concordance	Mark
1	she said she would *keep it on* longer.	6
2	So long as I *keep on.* I can master 3,650 words every year.	11
3	By these means, we can *keep in touch with outside*.	9
4	If you *keep up practising* something, you will get a lot of skills …	7
5	They could do more things to *keep themselves in good condition*.	6
6	Besides *keep track of* the newest information …	10
7	students can *keep pace with what is happening home and abroad*.	13
8	*keep at it*, and you will succeed.	9
9	The deception *keep us in the dark* until we grow up …	10
10	Many people *keep company with each other …*	8
11	Then I … learned to *keep my things in order …*	12
12	we must *keep contact with the society* constantly.	8

The figures in Table 7.4 do not show a direct relationship between the mere fact that an item occurs in such a category and the mark given to this item. Some items occur in compositions with high marks (such as No. 7 and No. 11) while others occur in compositions with low marks (No. 1 and No. 4). Some items come somewhere between the high and low marks (No. 6 and No. 9). Several factors need to be brought into consideration here. One is the whole text in which the item occurs.

If the student uses one item properly which is not used by many of his peers this will not automatically make the marker believe that his composition is of a high level. Another factor would be the consistency of individual markers. Different markers may give quite different marks to the same composition. It seems that this category is not necessarily a good area for the diagnostic function of the ratio relationships. However, further examination is required to be sure about this.

7.5 Some pedagogical implications

This section sums up the findings of the ratio relationship analysis and evaluates this study and some of its possible pedagogical applications.

7.5.1 Providing the next phase target for the learner

Both the teacher and the teaching material writer can benefit from reading and interpreting the frequency ratios of the patterns studied. On the one hand, the contrastive study of patterns above has revealed some new patterns that learners need to learn. On the other hand, the patterns used by the learners deviate from those produced by the NSs. Again, it is a matter of 'new things to learn and old things to mend'.

For the 'new things to learn', the patterns can be mainly found in the ratio type "A small frequency in COLEC vs. a large frequency in LOCNESS" – such as the patterns **KEEP n prep/adv** and **KEEP n adj**. To show how the learners could learn to practise these patterns, four examples in 7.4.2.3 are repeated as follows. The first two examples are in the pattern **KEEP n prep/adv** and the second two examples are in the pattern **KEEP n adj/adv**.

1) The legalization of marijuana among other drugs, would *keep some people out of the streets*.
2) This type of action is what *keeps millions of viewers on the edge of their seats* day in and day out.
3) I think it' s time we stop wasting money on *keeping drugs illegal* when it does not even work.
4) We talked some more in the lobby but we had to *keep our voices down*, out of respect.

For the 'old things to mend', it seems that useful information can be obtained from most of

the ratios as long as there are some occurrences in COLEC. But to be more confident about the validity, it is better to use the ratios in which there is a large frequency in COLEC.

The following shows how the acquired knowledge (the old) can be amended so that new knowledge may be learned. As shown in Table 7.1, the COLEC writers use frequently the patterns or phrases that express the meaning of *continue* and *maintain* (such as **KEEP n**, **KEEP –ing**, **KEEP *on* –ing**, **KEEP *up* –ing**, **KEEP *at it***). My suspicion is that the NSs would use other verbs in the same semantic cluster (see Chapter Four) such as *CONTINUE* and *MAINTAIN* to express similar things. To confirm this suspicion, the two corpora are checked and the result is as follows[37] (see Table 7.5):

Table 7. 5 Comparative frequencies of *CONTINUE* and *MAINTAIN* in COLEC and LOCNESS

	COLEC		LOCNESS	
	R F	N F	R F	N F
CONTINUE	50	52	177	274
MAINTAIN	9	9	35	54

The disparity in frequency shows that the NSs use *CONTINUE* and *MAINTAIN* much more frequently than the learners. It is quite possible that the NSs use these two words in places where **KEEP n**, **KEEP –ing**, **KEEP *on* –ing**, **KEEP *up* –ing**, **KEEP *at* it**, etc. are used by the learners. The following are only some examples:

LOCNESS: If he want successful he must *continue for a long time*.
COLEC: If you have no good health, you'll hardly *keep on doing* your work …
COLEC: If we *kept on*, we will make a great progress in English.

LOCNESS: Britain has been eager to *maintain a secure balance* of power on the continent…
LOCNESS: you are helping to *maintain a balance* of the number of lower income families …
COLEC: By doing so, I thought, we can *keep the balance* of water circulation.

LOCNESS: Pangloss himself, although in this sorry state, still *maintains his optimism* …
COLEC: we *keep high spirits* and keep on working.

37 Since the verb *MAINTAIN* has another sense as in "He maintains that …", all its concordances have been checked and such uses are not included here.

Another example to show that 'old things can be mended' is that the learners could be made aware that one particular notion can be expressed by different patterns. Table 7.6 shows how NNSs and NSs could use different patterns to express the same thing.

Table 7. 6 Some examples of using different patterns to mean the same thing

NS English	Pattern	NNS English	Pattern
keep calm (BoE)	**KEEP adj**	keep a calm head	**KEEP n**
keep fit (BoE)	**KEEP adj**	keep a good health keep our own physical fitness	**KEEP n**
keep her happy (LOCNESS)	**KEEP n adj**	keep their happiness	**KEEP n**

Instead of using 'KEEP a calm head', the learners may like to use 'KEEP calm'; similarly, instead of using 'KEEP a good health', they may like to use 'KEEP healthy'; instead of saying 'KEEP their happiness', they may simply say 'KEEP happy'. If the learners could be led to notice the alternative ways the NS use to express the same or similar meanings, they would stand a much better chance of becoming more native-like in their English production. By learning the new and mending the old, it may be expected that the English produced by the learners will gradually resemble that of the NSs.

7.5.2 Expanding the range of uses of vocabulary

The detailed examination of the use of the verb *KEEP* in this study shows that this simple word is, on the whole, not actually mastered to a good level. The findings of this study clearly demonstrate the need to deal very seriously with the so-called small and simple words like *KEEP*. A very commonly held view in vocabulary study is that the more vocabulary a learner has, the more advanced this learner's language level will be; it is commonplace for learners to try hard to increase their vocabulary. But we are seldom able to see how learners expand the range of uses of their vocabulary once its basic use has been learned to a reasonable level. This trend is explicitly summarised and criticised by Sinclair (1991: 79) as follows:

> The evidence that is accumulating suggests that learners would do well to learn the common words of the language very thoroughly, because they carry the main patterns of the language. The patterns have to be rather precisely described in order to avoid confusions, but then are capable of being rather precisely deployed.
>
> At present, many learners avoid the common words as much as possible, and especially where they make up the idiomatic phrases. Instead of using them, they rely on larger, rarer, and clumsier

words which make their language sound stilted and awkward.

The view that it is crucial to learn to use simple words is also shared by a well-known Chinese scholar LIN Yutang.[38] He maintains that English learning should start from simple vocabulary acquisition rather than long and complicated words that are picked up from dictionaries. He also proposes that whenever a learner tries to learn a word, it is important that at least one correct usage is acquired. Thus, when later he happens to come across other usages, the learner may wish to expand the knowledge concerning this vocabulary. This strategy is worth recommending to students because it helps them to accumulate knowledge of English without the risk of producing unnatural English. Furthermore, Lin suggests that in order for a word to be remembered the whole sentence needs to be remembered first. This emphasises the importance of contextual information for language study.

7.5.3 Providing information for learner English gradation

As mentioned in Chapter Six, there is a degree of congruence in a group of writers with the same background. The analysis of patterns in this chapter also finds the existence of such a phenomenon. I would like to propose that such a congruity be used to grade the level of group learner English. It is reasonable to believe that the closer the patterns produced by the learners are to those by the NSs, the more likely it is that the group learner English will be in a similar level to the NSs'. A learner group having ten patterns such as the COLEC writers is likely to be higher than a group with only five patterns but lower than another group with fifteen patterns. Of course this is only a rough indicator and needs to be confirmed by other means.

7.6 Conclusion

In this chapter I have examined the similarities and disparities between learner English and NS English in using the verb *KEEP*. Meanwhile, I have proposed a refined categorical description to account for the relationship between learner English and NS English. Even

38 LIN Yutang (1895-1976) was one of the most outstanding scholars in China, known as a writer (both in Chinese and English) and a translator. His English books *My Country and My People* and *The Importance of Living* were reprinted many times in America, winning him unprecedented fame in the west for a Chinese scholar. He was nominated for the Nobel Prize for Literature for his English work *Moment in Peking*.

though this proposition is of a rather preliminary nature, and needs extensive testing and verification, I firmly believe that it is useful for learner language studies. Researchers focusing on learner language cannot afford to cling on to the traditional terms ‘overuse’ and ‘underuse’ for too long, unaware of the potential value of these detailed categorisations for learner language diagnosis and evaluation.

The research results also suggest that simple words like *KEEP* have not been properly mastered by learners judging from the performance revealed by the data. While teachers constantly endeavour to expand their students’ vocabulary, they might like first to think about how to make the best use of their existing vocabulary.

Chapter Eight

Using Collocates to Interpret Learner English

8.1 Introduction

Chapter Seven has studied the detailed patterns and phrases of *KEEP* used by the COLEC and LOCNESS writers. One of the findings suggests that simple vocabulary like *KEEP* does not seem to have become very familiar to the COLEC writers. This chapter continues the investigation into the COLEC learner English in another small and common word *TAKE*. The reason for having chosen this verb is that it collocates with a large variety of items, such as *TAKE care*, *TAKE place*, and *TAKE measures*. All the cases of *TAKE* in COLEC are verbs and most of the cases of *TAKE* in the two corpora are also verbs except a couple of cases of V-ing form used as gerund (in the traditional grammar term) in LOCNESS. The research questions of this chapter are as follows:

(1) What are the similarities and disparities between the learner English and the NS English in terms of the collocates of *TAKE*?

(2) Are there any typical erroneous expressions with *TAKE* in COLEC?

(3) What pedagogical implications do the results of the first two previous questions have?

8.2 Some theoretical underpinnings

Basically, 'collocation' is the abstraction of 'collocate', but the term 'collocate' is slightly different from the term 'collocation' in use. As far as I can see, whereas a 'collocation' emphasises the lexical co-occurrence of words, a 'collocate' does not distinguish between lexical and grammatical co-occurrences. In concordancing 'collocate' is mainly used and always interpreted with the notion of 'span' because of its technical feature. A 'collocate' can be either on the left or the right of the node. While sometimes a 'collocate' may overlap with a 'collocation' as in *TAKE action*, *TAKE a view*, and *TAKE chances*, at other times they are not identical. Take the phrasal verb *TAKE place* for example, if we treat *TAKE* as the node, then *place* is a 'collocate' with one span on the right, and people hardly treat *TAKE place* as a 'collocation'. Take the idiom *TAKE part* as another example, if we treat *TAKE* as the node,

then *part* would be a 'collocate' with one span on the right, and people seldom consider *part* as a 'collocation' of *TAKE* either. Since the literature has concentrated more on 'collocation' rather than 'collocate', I would like to review some theoretical stances towards 'collocation'.

Collocation has been studied heavily in the English language (for example, Firth 1957, Sinclair 1987, Sinclair 1991, Stubbs 2001, Hunston and Francis 1999, Hunston 2002). Sinclair (1991: 170) defines collocation as "the occurrence of two or more words within a short space of each other in a text". Hunston (2002: 12) relates the physical orthography to the functions of collocation while she gives collocation a definition thus:

> [Collocation] is the statistical tendency of words to co-occur. … Collocation can indicate pairs of lexical items, such as *shed* + *tears*, or the association between a lexical word and its frequent grammatical environment.

On the significance of collocations in the study of corpus linguistics, Hunston (2002: 76-79) continues:

> One use of collocational information is to highlight the different meanings that a word has. …
>
> A somewhat different method of displaying collocational information can, however, be used to obtain clues as to the dominant phraseology of a word. …
>
> Finally, collocations can be used to obtain a profile of the semantic field of a word.

While it is indubitably true that such a study of collocations may shed much light upon the description of the English language itself, it has little to say about whether collocations can be easily produced by learners. It was not until recently that researchers started to look at this area. The following section introduces two recent collocation studies carried out in learner language research.

8.3 Two recent studies of learner English in collocation

An early serious and large-scale learner language study in collocation was by Howarth (1996). In his comparative research of the written output of some NNS writers and NS writers he

acknowledged the difficulties of learners in the use of collocations, phraseologies and idioms. He studied phraseology including collocations produced by English learners with several different L1 backgrounds. Many of his findings are useful in understanding better the characteristics of learner language, even though his limited data offsets the value of his research to some extent.

Another recent study of learner English in the area of collocation was undertaken by Nesselhauf (2005). With a considerable number of collocations extracted from her data on German-speaking learners of the English language, her research was based on the ratings of some native speakers of English rather than a controlled corpus of English. This would be useful in a study of the features of learner English, especially if the research aim is to work out what collocations NNSs do not write in the same way as NS do. For example, some learners are found to write 'consume drugs' instead of 'take drugs'. The methodology she adopted helps to see how different the German-speaking learners' English is from the intuition of the raters. The information that can be obtained from her study concerns the expressions used by the learners. Compared with Howarth's research, which used 22,000 words, Nesselhaulf's data is much larger (150,000 words). With more data, she was able to concentrate on the learner English produced by the same L1 writers instead of an amalgamation of learners with different L1s and different cultural backgrounds as Howarth did. As Nesselhauf (2005: 9) acknowledges herself, "Restricting the analysis to one L1 group rather than analysing more data from many different L1 groups was deemed necessary since, as a number of studies have indicated, the first language clearly plays a role in L2 collocation production, but has nevertheless not been investigated in much detail." Though her research is valuable in many ways, her analysis does not involve the use of a controlled corpus, without which it would be difficult to see the similarity and disparity between learner English and NS English. Nesselhauf used manual extraction to access the verb–noun combinations instead of using the advantages of corpus linguistics tools such as WordSmith, online corpora such as the BoE and the internet as a resource. In an era where technology plays such an important role, it would be better to make fullest use of these modern technologies, as will be demonstrated in this chapter.

8.4 Making a table of collocates from the two corpora

The verb *TAKE* is one of the most often used verbs both by the COLEC learners and the LOCNESS writers (ranked 8th in COLEC and 2nd in LOCNESS). There are 1239 occurrences of all the forms of the verb in COLEC and 680 occurrences in LOCNESS. It is expected that there will be some overlappings (and disparities as well) between the collocates used in COLEC and LOCNESS. To understand what items collocate with *TAKE*, I have created a list of all the collocates (both lexical and grammatical) that occur in the two corpora. The list includes all the words that come either immediately after the verb *TAKE* such as *TAKE place*, *TAKE ACTION* and *TAKE on*, or with a span of a few words such as *TAKE … seriously*, so that all the dominant phraseology of such a frequently used verb could be revealed extensively. The process of making such a list is recorded in detail in the following section.

To extract all the collocates from a corpus manually is certainly possible as Nesselhauf (2005) did in her research. But there is a more accessible and easier way to do this by using some of the functions of WordSmith. See Appendix 6 for the steps I took in making a collocate list of each corpus.

Whereas it is easy to identify the collocates in the NS corpus and then give them suitable codes (see Appendix 6 for some details), it is not as easy to deal with the ones in the learner corpus because on one hand there are many erroneous collocates used in COLEC and on the other hand there exists a continuum between the end of acceptability and the end of non-acceptability. In cases where it is hard to make a decision whether to label a combination as acceptable or not, the BoE would be searched to check the popularity of a certain expression. Sometimes, concordance lines are checked at full length to see whether a particular expression is a genuine one or not because a combination may coincide with an existing collocation morphologically but deviate functionally from the NS use.

After the collocates of both COLEC and LOCNESS have been identified and encoded, it is possible to make a list of all collocates of *TAKE* in each corpus. Based on such a list, it is then possible to make a comparable list of the collocates of the two corpora (see Table 8.1). There are four columns for each of the corpora, i.e. “SN” refers to the serial number of each

collocate group according to the number of frequency, from large to small in LOCNESS; "Collocate" lists the collocate type (mostly in concrete words such as 'place' but occasionally in categories such as 'SB' (for somebody), 'STH' (for something) and 'DO' (for actions) if it is difficult to cover the exact words; those collocates that share a meaning or pattern with the word listed in the "Collocate" column are put in the "Varieties" column. For example, the collocate *steps* in LOCNESS has the varieties *actions* and *efforts*. The "Fre" column (short for frequency) lists the frequency of the collocate type. The same or similar collocates in the two corpora are arranged in one row for the sake of comparison. The collocates of LOCNESS are arranged on the left of the table so that it is easier to see what collocates are not present in the COLEC corpus. The collocates that occur only in LOCNESS are highlighted in bold. In cases where there is no suitable word to represent a group, I will leave the "Collocate" column empty and use capitalised words as a category in the "Varieties" column instead.

There are 79 collocate types in LOCNESS and 56 collocate types in COLEC, as Table 8.1 shows. There are altogether 99 types of collocates identified in the two corpora. Only a small proportion of the collocates (39) are roughly shared by the two groups of writers. Within the non-shared collocates (60), more types occur in LOCNESS than in COLEC. As far as erroneous collocates in COLEC are concerned (see Line 100 of Table 8.1), there are as many as 138 cases which are virtually impossible to categorise into commonly acceptable collocates, including *TAKE changes*, *TAKE improvement* (*progress*) and *TAKE attention*.

It seems that there is a considerable distance between the level of performance of the COLEC learners and that of the LOCNESS writers. Apart from the distance between the collocate types, is there any disparity between the two groups of writers in terms of detailed use of the shared collocates or collocate types? This issue will be addressed in the next section.

Table 8. 1 A table of collocates of *TAKE* in LOCNESS and COLEC

SN	Collocate	Varieties	Fre	SN	Collocate	Varieties	Fre
1	place		60	5	place		79
2	on		33	12	on		25
3	STH from		30	10	STH from		31
4	a week	a while, a year, etc.	28	8	time		46
5	**away**		27				
6	STH into consideration	STH into account, STH **on board**	25	18	STH into account	consideration (1)	10

7	**a view**	**a stance, a side, a position, a stand**	25	35	an attitude	an idea	3
8	STH to	with, on,	23	23	STH to		8
9	**a life**	**many lives, someone's life**	23				
10	**STH as a whole**	INTERPRET	22	30		INTERPRET	5
11	action	actions, efforts, measures, steps	22	2	action	actions, measures, steps, etc.	159
12	care		18	9	care		35
13	responsibility		17	60	responsibility		1
14	**STH seriously**	**lightly, to heart, personally**	17		STH seriously		1
15	an option	options, a decision, decisions etc.	17	51	a choice		1
16	STH away from	STH away	15	61	away		1
17	SB to DO STH		14	45	SB to DO STH		2
18	up		13	7	up		60
19	advantage		13	13	advantage		21
20	a way	**a path**, **a road, an approach,** means**, a course, direction**	13	15	a way	a method, some methods, some means, etc.	17
21	a test	tests, an exam, exams, etc.	12	40	a test		2
22	over		12	33	over		4
23	a risk	**a chance, chances**	12	43	a risk		2
24		SELECT	11	59		SELECT	1
25	**a class**	**classes, courses, curriculum**	10				
26	the place of	**a second place, the lead, the initiative**	10	28	the place of		5
27	**STH out of**		9				
28	medicine	medication, drug, tablets, marijuana	9	17	medicine		10
29	**comfort**	**refuge, solace**	9				
30	part (in)		9	1	part (in)		233
31	**a trip**	**a journey, voyages**	8				
32	**precaution**	**precautions**	7				
33	notice	**heed, note, attention** (?)	7	27	notice		5
34	it upon ONESELF		7				
35	**a share**	**a cut**, a profit	7	44	a part	no profit	2
36	STH for example	**STH for instance**	6	14	STH (for) example		19
37	control	**manipulation**	4	42	control		2
38	**a role**	part	4	50	a part		1
39	**poll**	**survey, study**	4				
40	the time to		4	52	one's time		1

41	**gunshot**	**pounding, blows, strain**	4				
42	**for granted**	it for granted	3	6	it for granted		70
43	a bus	a train, cars	3	25	a bus	buses, a tax, a train, a plane,	6
44	effect		3	47	an effect	bad effects	2
45	**interest**	**dislike**	3				
46	**blame**	**guilt**	3				
47	a job	jobs	3	4	a job	job, work, etc.	104
48	a bath	a shower	3	41	a bath		2
49	back	-back	3	26	back		5
50	**it's toll**		2				
51	**to (boxing)**	**the Rail system**	2				
52	**off**	**BECOME SUCCESSFUL; (OF SALES) RISE QUICKLY**	2				
53	**take credit for**		2				
54	**revenge**		2				
55	**a seat**		2				
56	**precedence**		2				
57	**in (prisoners)**		2				
58	a walk	**a stroll**	2	21	a walk	morning walks	8
59	pictures		1	29	a photo	photos	5
60	**hold**		1				
61	**the form of**		1				
62	**an upswing**		1				
63	**aim**		1				
64	**blood**		1				
65	**pleasure (in)**		1				
66	**orders**		1				
67	**pity**		1				
68	**oath**		1				
69	**turns**		1				
70	**STH to heart**		1				
71	**STH to pieces**		1				
72	**heart out of**		1				
73	**no prisoners**		1				
74	**it like a man**		1				
75	**the name of**		1				
76	**the leap of**		1				
77	**exception**		1				
78	**down**		1				
79	**STH off**	REMOVE	1				

80			11	BRING	25
81			16	practice	11
82			19	it easy	10
83			20	out (a book)	10
84			22	a challenge	8
85			24	a look	7
86			31	in (CHEAT)	5
87			32	opportunity	every chance, opportunities 4
88			34	exercises	sports 3
89			36	vocation	rest, nap 3
90			37	interest	3
91			38	hold of	3
92			39	it in mind	it in heart 3
93			46	SB'S suggestion	advice 2
94			48	good use of	best use of 2
95			49	lessons	2
96			53	a profession	1
97			54	a temperature	1
98			55	heart to	1
99			56	BE (taken) ill	1
100			3	UNACCEPTABLE	138
	TOTAL	680		TOTAL	1239

8.5 A detailed look at some large-frequency collocates

It is reasonable to assume that the more frequently an item is used by learners, the more likely it is to approximate to the use of NSs. In order to see how well the COLEC learner English approximates the LOCNESS NS English, I decided to take three types of collocates, i.e. *TAKE ACTION*, *TAKE place*, and *TAKE on* and subject them to detailed examination.

8.5.1 Looking at TAKE ACTION and its group

8.5.1.1 Looking at the right and left positions of the collocates of TAKE

This section compares the group of nouns that collocate with *TAKE* in the two corpora, i.e. *action*, *actions*, *efforts*, *measures*, *steps*, etc. There are three different types of relationship between the node *TAKE* and the noun collocates. In the first type of relationship, the collocating nouns come after the node *TAKE* while in the second and third types of relationship the nouns precede the node *TAKE*. The first type of relationship is in the active voice and the second type is in the passive voice, and the third type is in relative clauses or simplified relative clauses. This is further illustrated as follows:

1) **TAKE (…) n** (as in active voice), e.g. They decide to *take action* to improve their lives …
2) **n … TAKE** (as in passive), e.g. … *action* must be *taken* immediately to alleviate the problem.
3) **n (…) TAKE** (as in relative clauses), e.g. The various *measures* that are being *taken* to create a united Europe will surely …

Let us see the edited concordance lines of the three types of relationship in LOCNESS first, one by one.

Type One: **TAKE (…) n**

1) They decide to *take action* to improve their lives …
2) then students will not pray since there will be no reason to *take* this mental *action*.
3) but he has stood up to the fact and does not *take* such drastic *action* as Caligula on …
4) The next step is to *take* the most appropriate *action* …
5) They are both seen as *taking* pure, logical *action* which is admired by the audience …
6) You can't want to discuss sex openly without discussing *taking* precautions *measures*.
7) Boxing […] has *taken* large *steps* of development since its early days

Type Two: **n … TAKE**

1) Some people think that *action* needs to be *taken* on drugs.
2) Proper *actions* need to be *taken* so that the government and in turn tax payers do not …
3) and *action* must be *taken* immediately to alleviate the problem.
4) Therefore, *steps* for gun control must be *taken*.
5) *Steps* have already been *taken*, however, towards this goal.
6) Also, great *steps* have been *taken* to ensure boxing is being made safer all the time.
7) the *efforts* perhaps have been *taken* a little too far when it comes to women in combat.

Type Three: **n (…) TAKE**

1) lenient *action taken* on violent crimes, such as rape and murder …
2) The only suitable *action* to be *taken* would be to increase safety regulations …

3) you must be aware that it is you that is making the choice of what *action* to *take*.
4) he predicted the *action* the grand jury of the city would *take*.
5) we are old enough to handle the *actions* we *take* Into our own hands, maturely.
6) but there are also huge disparities between the *action taken* in the face of absurdity.
7) simple *steps* that can be *taken* to maximize the room …
8) The various *measures* that are being *taken* to create a united Europe will surely …

As shown above, there are seven cases of the first type, another seven cases of the second type and eight cases of the third type, totalling 22 cases altogether.

The following are the raw concordances of the three types of relationship in COLEC. These concordance lines are not edited so that the comparative portion between one type and another can be better displayed. Some concordance lines which resemble the neighbouring ones are omitted to save space.

Figure 8. 1 Type One: *TAKE* (…) n

1 building. Second, the government should take a great effort to find way in which
2 more importance to their survival. They take a lot of strong actions to keep th
3 tantion to the industrial polution, and take action on it. I think if all of us
4 ans that you make decisions blindly and take action rashly. Since considered dec
5 I think it is time that the government took action to prohibit fake-makings and
6 ake commodities and the government must take action to get rid of fake commoditi
7 the quantity we use. If every one of us takes action, we must be able to overcom
8 ced with the shortage of water, we must take action. The most importance measure
13 fresh water resources. Firstly, we must take actions to control the water pollut
14 s. In a word it needs all the people to take actions, to give support, so the fa
15 ome people to select "lucky numbers" to take actions. We have no reason to forbi
16 atients' lives in danger. So, we must take active action to prevent fake commo
17 erefore, it is suggested that we should take activities properly.
18 exprecs corresponding feelings when we take activities we can make use of music
19 y advanced, people know how to live and take activities, how to prolong their li
20 st be very careful with fire. We should take all kinds of measures to guard agai
21 a certain extent. In a word, we should take all kinds of measures to preserve o
22 and our industry. Secondly, we have to take all measures to protect them from b
23 omeone doesn't know it and he haste to take an action, it only will be wasted.
24 odities. On one hand, government should take critical measures to punish the fak
25 harmful, and we should try our bests to take effective action to control them, p
26 ke the best use of ourland. City should take effective measure to protect our la
29 modities. The government concerned must take effective measures to eliminate the
30 t we can to avoid the crisis. We should take effective steps to avoid waster. Me
31 olluted, being wasted. They also should take efficient step to control the incre
32 Facing the serious situation we should take effictive steps to better it. Fistl
33 ignore it. On the country, they should take emergency measures.
34 important to human being, so we should take every step to resolve the problem o

ause of their high price . So, we must take great pains to prohibit fake commod
an will not gain much profits unless he take great pains to run his besiness wel
s. The second, the whole society hasn't taken magnificent measures to stop the p
nce 1960, the developing countries have take many measures in economic developm
days. In addition, the governments also take many measures to protect the people
re for their health. The countries also took many practical measure to improve p
e bad or as bad as before. So, we must take measure to prevent the fake commodi
ge. Fresh water is limited, so we must take measure to protect fresh water. We
y that man will kill themselves without taking measure to save water. But how to
so the fresh water is less. We should take measure to protect our fresh water,
ng producted. On the one hand we should take measures and lay out the laws and r
Faced with such situation, we should take measures as follows. First at all,
ut society develop? Therefore, we must take measures as soon as possible to
e want to have a bright future, we must take measures now. In fact, we can pass
bal shortage of fresh water and we must take measures to preserve it.
make good use of fresh water. We should take measures to reduce the population a
mically use fresh water. Second we must take measures to prevent the factories f
s. How should we do? I think we should take measures to this question. Firstly,
ying polluted water. Thirdly, we should take more effective action to control th
't want to finish it in a long time and take more efforts to it. At last, you c
ions. Now that the developing countries took no effort to develop the industry e
sociate investigation. In a word, I'll take pains to my study, at the same time
the contrary. We can spend litle time, taking social activities. If we do like
ly like food and liquor. Thus, we must take some action eto prevent the harmnes
efore, only speaking is no use. We must take some action to it. Here, I advise s
mmodities will effect economic. We must take some actions to prevent the phenome
w do we know the society. First, we may take some actions, students go out of th
home and abroad. In addition to, we can take some activities, such as help the o
nd cheating the goverment still hadn't take some effective measure to this pr
English? For the sake of passing CET, I took some efforts to find the key to sol
ies pay much attention to it. They have taken some good measures. For example. T
't live without fresh water, so we must take some measures to pretect our rare r
better and better. So they are able to take some measures to improve the people
ational industry. It's time we should take some measures to deny the fake comm
fresh water is very short. So we must take some measures. Firstly, we should m
man are awaring of the problem and have take some steps to solve it. We sure man
facing us today. I think people should take some steps to solve this problem, o
e country economy worse. So we should take some tough measures to get rid of t
can also bring disorder. So before you take step, think it over.
Thereby when we do something use should take steps and establish a schiedule pri
is becoming much shorter. So we should take steps at once to protect our fresh
ve this coming water crisis, man should take steps immediately. First, man must
we are short of fresh water? We should take steps to prevent fresh water from b
e honour of our country too. We should take steps to make Fake Commodities no m
n not live without fresh water, we must take steps to protect our limited fresh
chieve much success must be patient and takes steps when their causes are under
avoid these facts, everyone of us must take steps. The government must punish t
d fake commodities. The government must take stern measures to do with it.
sponsible for that because they haven't taken strict enough measure to protect c
es form producting, the government must take strong measures. The whole people a
ter. How should we do? I think we must take the following measures to protect f

126 ack. To get rid of this result we can take the following measures. First, we c
127 rtage of fresh water, I think we should take the following steps: First, we shou
128 people want to make himself fat, but he takes the measure in order to want himse
129 l directions. The first, the government take the serious measures to punish the
130 nst Fake Commodities. I think we should take these measures to cope with it: fi
131 consume our energy for studying. If we take too activities, we will be tired wh
132 ities become a serious problem. We must take urgent actions aganist them.

133 used to have. What measures should we take to solve the problem? First, the go
134 r takes place. What measures should we take to cope with the serious shortage o
135 m more serious.... What steps shall we take to solve the problem? First, and th

Figure 8. 2 Type Two: n … *TAKE*

1 ple's healthy . Many measures have been taken. And many achievements have been m
2 economically. In short, action must be taken before it is too late.
3 taility . Still more measures should be taken. Despite health gains in developin
4 rough hard work. So measures should be taken immediately to get rid of fake com
5 ware of the problem and steps have been taken in many countries. We should make
6 loss. Fortunately, Many steps have been taken in recent years, and some have cha
7 ve been aware of it, measures should be taken in solving this problem. Some expe
8 Finally, effective measures are not be take to control the orerflow of fake com
9 th it? I think many measures should be taken to deal with the shortage of fresh
10 polluting the water. Measures have been taken to deal with this problem in many
11 ater. Secondly, more measures should be taken to due with the polluted water. Fi
12 fresh water. Second, measure should be taken to keep the water from poluting. A
13 ed is getting less. Mensures should be taken to preserve our fresh water resour
14 go away. Secondly, the measures must be taken to prevent fresh water from pullut
15 . Don't wast it anymore. Action must be taken to prevent the water from being po
16 mportant for us. Some measure should be taken to protect fresh water.
17 can't stand by any more. Steps must be taken to protect the fresh water resourc
18 can no more be used. Measures must be taken to ″save″ fresh water. Laws should
19 ter more scarce. Thus, actions must be taken to solve the problem of fresh wate
20 place. Therefore, some steps should be taken to solve these problem. Regulation
21 ater shortage. Many measures should be taken to solve this problem. We can turn

Figure 8. 3 Type Three: n (…) *TAKE*

1 my opinion, the first step we shall to take is to make lawa in consumers' inter
2 e, there are many other measures we can take to deal with the shortage of fresh
3 re are several possible steps we should take. We must take measures to bring the

As the numbers shown above reveal (see Figure 8.1 to Figure 8.3), there are as many as 132 cases of the first type of relationship (not including the three question forms which are not considered here), but only 21 cases of the second type and only three cases of the third type. A huge disparity is clearly shown here: the COLEC writers produce much more cases in the active voice than do the NSs. Meanwhile, the learners use much fewer cases using the passive voice or with relative clauses.

8.5.1.2 Looking at TAKE ACTION in a wider context

If we broaden our examination to include a wider context of the collocates, we may find further differences in the two corpora. Let us look at the right-side neighbours of *TAKE ACTION,* no matter if the noun is on the right or left of the verb *TAKE* in Figure 8.4.

Figure 8. 4 All the concordances of the collocate *TAKE ACTION* in LOCNESS

1 action the grand jury of the city would take. He was called before the jury and
2 is making the choice of what action to take. If you begin to see the futility
3 ray since there will be no reason to take this mental action. Prayer i
4 t he wites poems. They are both seen as taking pure, logical action **which** is adm
5 on is to a student. The next step is to take the most appropriate action, **which**
6 tein 14). Proper actions need to be taken **so that** the government and in turn

7 cultiver notre jardin″. They decide to take action **to improve their lives**, phil
8 continually growing, and action must be taken immediately **to alleviate the problem**
9 ignored. The only suitable action to be taken would be **to increase safety regulatio**

10 me people think that action needs to be taken **on drugs**. Most published works
11 overcrowded prisons, lenient action taken **on violent crimes**, such as rape an
12 e has stood up to the fact and does not take such drastic action as Caligula **on**

13 old enough to handle the actions we take Into our own hands, maturely. If a
14 lso huge disparities between the action taken in the face of absurdity. Part of

The characteristics of these concordances could be interpreted and summarised as follows:

1) If the right-hand context of the LOCNESS writers is observed, there are several points worthy of attention. First, there are three cases in which the collocate ends a sentence with a full stop (as in Lines 1, 2 and 3), and three cases in which the collocate ends a clause followed by a subordinate clause (starting with *which* in Lines 4 and 5, but with *so that* in Line 6). In other words, six out of the 14 cases are actually at the end of a clause.
2) Three concordance lines are with to-infinitives following '*TAKE action*' (see Lines 7, 8 and 9). As will be seen shortly afterwards, this is the most frequently used type of collocate in COLEC.
3) There are three cases (Lines 10, 11 and 12) in which the preposition *on* is used to point to the objective of the actions, which is *drugs* in Line 10 and *violent crimes* in Line 11. Line 12 is chopped off by the concordancing format and the following is its full-length sentence:

We admire him - his father was killed by Caligula, but he has stood up to the fact and does not **take** such drastic **action** as Caligula **on** discovering the human condition

4) There are two miscellaneous cases (Lines 13 and 14) which do not belong to any of the groups described above.

In addition, if we look at the singular form and the plural form in this collocate, we will find that the ratio between the singular and the plural is (11 vs. 3). Only two cases are in the plural form *actions* (although arguably Line14 should also be *actions*).

Figure 8. 5 All the concordances of *TAKE ACTION* in COLEC

1	ake commodities and the government must take action to get rid of fake commodity
2	I think it is time that the government took action to prohibit fake-makings and
3	olluted water refreshed and used. Let's take actions from now on to protect our
4	before. Facing this phenomenon, we must take actions right now. As we know, fak
5	s. In a word it needs all the people to take actions, to give support, so the fa
6	fresh water resources. Firstly, we must **take** actions to control the water pollut
7	n't worry. Third, the government should take actions to decline the birthrate an
8	e the problem? On the one hand, we must take actions to cut down the excessive d
9	ly like food and liquor. Thus, we must take some action eto prevent the harmnes
10	atients' lives in danger. So, we must take active action to prevent fake commo
11	harmful, and we should try our bests to take effective action to control them, p
12	ying polluted water. Thirdly, we should take more effective action to control th
13	mmodities will effect economic. We must take some actions to prevent the phenome
14	more importance to their survival. They take a lot of strong actions to keep th
15	ter more scarce. Thus, actions must be taken to solve the problem of fresh wate
16	. Don't wast it anymore. Action must be taken to prevent the water from being po
17	ced with the shortage of water, we must take action. The most importance measure
18	ome people to select "lucky numbers" to take actions. We have no reason to forbi
19	omeone doesn't know it and he haste to take an action, it only will be wasted.
20	w do we know the society. First, we may take some actions, students go out of th
21	the quantity we use. If every one of us takes action, we must be able to overcome
22	economically. In short, action must be taken before it is too late.
23	tantion to the industrial polution, and take action on it. I think if all of us
24	ities become a serious problem. We must take urgent actions aganist them.
25	efore, only speaking is no use. We must take some action to it. Here, I advise s
26	ans that you make decisions **blindly** and **take** action rashly. Since considered dec

In COLEC, however, as displayed in Figure 8.5, there is a much larger proportion of the use of the pattern '*TAKE action to* **V n**' in COLEC (16 out of 26). This seems to be the central pattern use relating to the collocation *TAKE action*. There are six cases in which the collocate ends a clause, which seems to match the use of the NSs. However, the way the collocate ends a clause in COLEC is by no means identical to the usage of the NSs. If we compare Figure 8.5 with Figure 8.4, it can be seen that the actions at the end of a sentence are specified in LOCNESS but not in COLEC.

If we also look at the proportions of the singular form and the plural form of the noun *action*, we will find a much smaller proportion of the singular use and a larger proportion of the plural use (15 vs. 11) in COLEC. Since the total number of collocates in LOCNESS was too small to be significant, I resorted to the BoE and the result is shown in Table 8.2.
The small letter 'n' in bold type refers to the noun, in the singular form (the 'action' column) and the plural form (the 'actions' column) respectively.

Table 8. 2 Some figures of three varieties of the collocate *TAKE ACTION* from the BoE

	action	actions	(action/actions)
TAKE **n**	3242	94	34:1
TAKE **adj n**	2960	94	31:1
n (0,4) *TAKE*	4265	74	58:1

It is clear that as a general trend in the whole English language, the singular form *action* is much more used than the plural to collocate with *TAKE* (34:1) (see the 'action/actions' column in Table 8.2). This trend does not change much when an adjective occurs between *TAKE* and the noun (31:1). However, when the noun comes on the left of the node verb *TAKE* with a space of four words between the noun (either *action* or *actions*) and the verb lemma *TAKE*, there is a dramatic change to the ratio between the singular form and the plural form, which suggests that the plural noun form *actions* is more likely to collocate with *TAKE* in the left position rather than in the right position.

Having worked out a general trend in using the singular and plural form, let us compare again the position of the plural use of the noun in the two corpora. According to Figure 8.3, the three cases of plural use are all in the left position in LOCNESS which fits with the findings from the BoE, whereas in COLEC only one plural use takes place on the left position and the other 10 cases are in the right position (see Figure 8.5), which deviates drastically from the

findings from the BoE.

Looking at the disparity in the use of the singular and plural form of the noun *action*, I begin to suspect that the idiom principle is playing its role in the NS English while the open-choice principle is taking effect in the learner English (Sinclair 1991: 109-115). It appears that the plurality issue of the noun does not lose as much of its grammatical constraint in the collocate in COLEC as it does in LOCNESS. Presumably, the learners use more plural cases because they are uncertain about the role the noun plays in this collocation. Since the learners have had much less exposure to the English language, it is difficult for them to develop a reliable sense of 'naturalness' in deciding whether it is correct to write 'take some action' without adding a plural 's' to 'action'. In the BoE, however, the sequence 'TAKE some action' (119) occurs much more frequently than the sequence 'TAKE some actions' (12). The fact that their plural use almost equals that of the singular use seems to suggest that most of the learners of this corpus are at somewhere between a rudimentary 'open-choice' stage and more refined 'idiom principle' stage.

At the end of the comparison between the uses of the collocation *TAKE ACTION*, it is becoming apparent that a large disparity exists in the production of such a collocation, not only the position of the noun, but also in the ratio of singular and plural use. The neighbours of the collocation also behave quite differently. If further examination were to be carried out, more disparities would be sure to come up. But the analysis above is sufficient to demonstrate that there is still a large disparity and distance between the status of the learner English and that of the NSs.

8.5.2. Looking at TAKE place

Since *TAKE place* is an intransitive multiple-word unit, it does not make sense to look at the collocates on the right. Therefore, a decision was made to look at the behaviour of the neighbours on the left. What follows are the most often used collocates going with the phrase *TAKE place* on the left of them. The number at the end of each line is the frequency of the words in the brackets that collocate with *TAKE place*.

In LOCNESS the most identifiable words are: *action* (*activity*) (5), *event* (3) and *incident* (1).

There are also some pronouns such as *it*, *these*, etc. (5). Other activity-related words are grouped as below:

(1) STORY-RELATED (*play, story*) 2
(2) VIOLENCE-RELATED (*war, murder, killing, rape*) 8
(3) SPORTS-RELATED (*sport, hunting, hunts, matches*) 7
(4) LEARNING-RELATED (*work, research, learning, training, observation*) 6
(5) FEMALE-RELATED (*birth, fertilization, pregnancy and menopause*) 4
(6) LEGAL-RELATED (*marriage, executions, discrimination, elections, legislation* and etc.) 7
(7) CHANGE-RELATED (*change, developments, advances, and improvements*) 5
(8) OTHERS (*transfer, classification, distortion*) 6

In COLEC, however, much fewer types of word collocate with *TAKE place*. The most frequently used collocate is *changes* (*change*) (48). To be exact, 48 out of the 79 occurrences of the phrase (61%) collocate with *changes* or *change*. It appears that it is a general trend for this group of learner to use this collocate. The second most often used collocate is *events* (3), which is valuable indeed because of its rarity and its affinity to the performance of the NSs. Other words include *accidents, shortage of fresh water, power failure, case, world cup*, etc. (10). There are also pronouns such as *this*, and *what* (10). There are as many as six cases of misuse. There are three cases of the collocate *events*, which is extremely useful. The comparison seems to show that a large number does not necessarily of itself lead to the interpretation that the learners have as full a range of use as the NSs do. As mentioned in Chapter Two (at 2.7.3), learner English can be regarded as "few items repeated more" (Cobb 2003: 412).

8.5.3 Looking at TAKE on

Apart from the many lexical collocates of *TAKE place* and *ACTION*, there are several prepositional collocates such as *on*, *up* and *over*. After it has been established that the COLEC writers do not use the lexical collocates as fully as the NSs (see Table 8.1), the question that arises is that whether the COLEC writers use the prepositional collocates as fully as the NSs. As shown in Table 8.1, *on* is the most often used prepositional collocate in both of the corpora. Since sense is in alignment with collocation (Sinclair 1987 and many others), the senses of

TAKE on will be looked at in order to help learners to see how this verbal phrase collocates with sense-specific nouns.

In terms of frequency, as many as 28 occurrences of the phrasal verb *TAKE on* are found in LOCNESS. The NSs use as many as four senses of the phrasal verb, as the following concordance lines show, from Figure 8.6 to Figure 8.9.

Figure 8. 6 Sense One: decide to do sth; undertake sth

1 e believes that by facing death, he can take on the sins of the world and thus r
2 rs is in the example we set. We can not take on their guilt or remorse as this w
3 et them for their own safety obliged to take on guilt and remorse for not having
4 one who knows what consequences will be taken on and how much the child will
5 Christ like' sacrifice whereby Kaliayev takes on the guilt of others so they can
6 nt his crimes, while freeing mankind by taking on the burden of their sins. Thro
7 the light and the sun to the world and taking on the sins of man. In Sartre's t
8 thus repreiving man-kind. Just as Jesus took on the sins of the world, Kaliayev
9 last years of his presidency, d'Estaing took on a more prominant role then, it w

Figure 8. 7 Sense Two: accept

1 become independent, they decided to take on a certain way of life. This ″wa
2 tre says , but Oreste decides to take on his liberty and leave the tradit
3 the people fall back into the past and take on the value of an object - '阶re e
4 himself and for others. He refuses to take on the values and traditions of his
5 r, which was an option for Caligula. He takes on the revolt Camus wants us to as
6 the fact that he can make decisions and takes on the state of an object. Man is
7 se before the end of the war, as people took on a care-free attitude, with littl

Figure 8. 8 Sense Three: begin to have (a particular quality, appearance, etc); assume sth

1 e each country. Europe may as well then take on the form of a ″super″-country. B
2 There are also times when foods tend to take on the smells of other foods to
3 ew that Hugo joined the party to merely take on an identity. Intellectuals may b
4 oday as it was 30 years ago, but it has taken on different forms. One type is av
5 fare are white. Another prejudice has taken on new forms, it has also declined
6 the end of the play she is said to have taken on the qualities of Clytemneste -
7 e of gender roles. Although the method takes on many forms, the message is t
8 no longer die and are as such happy. He takes on the form of a god demanding dea
9 ning to take human form, but is also taking on all aspects of human life. Bio
10 criticising the government, Mitterrand took on the form of the ″gardien de l'in

Figure 8. 9 Sense Four: employ sb; engage sb

1 universities of Oxford and Cambridge do take on students that are not from a par

In COLEC, however, there are only two senses detectable from the concordance lines as shown in Figure 8.10 and Figure 8.11.

Figure 8. 10 Sense One: decide to do sth; undertake sth

1 ngle job for ever. Some are intend to take on a kind of job throughout their l
2 proper job. But I think we youth should take on a challenging job, we have enco

Figure 8. 11 Sense Two: begin to have (a particular quality, appearance, etc); assume sth

1 to consume. Further more, the countries take on a peaceful look. Therefore, as
2 s technology to us. Our country has taken on a new look since we had the Ref
3 r is very clear. First, the whole world took on a peace look, the war reduced, w
4 nt of the periods, these countries have took on a new look. they have revolution

There are six uses which could hardly be identified in sense (Figure 8.12).

Figure 8. 12 Unidentifiable Sense

1 ortunity that most people don't like to take on, above all that, there are som
2 live a stable lives and don't want to take on danger. While other people oft
3 second kinds of people, they usually take on the danger of finding their job.
4 are suitable and stable, they needn't take on the danger of not finding suitab
5 serious harm for us. Why is the state taken on? This because is that some prod
6 y years old. The infant mortaility also took on a new variation since 1960. In 1

To sum up, "few items but repeated more" seems applicable not only to single words but also to phrasal verbs, as has been revealed from the analysis of the collocates of *TAKE on*. This seems to suggest that even though the COLEC writers have started to use the prepositional phrase *TAKE on* and many others (see Table 8.1 for details), their productive English is not as extensive as that of the NSs as far as the range of senses is concerned.

8.6 Diagnosing the learners' typical deviant uses

8.6.1 Looking for explicitly deviant uses by the learners

As shown by the question marks in Table 8.1, there are 138 cases that deviate from the use of NS so much that they can hardly be regarded as acceptable English. While some erroneous expressions used by the learners are quite individual, there are some cases that are quite characteristic of the group. What I am going to show here is that where there is a general trend shared by a group of writers, a corpus-linguistic approach can be used to discover it as in the following unacceptable uses:

1) TAKE a change (changes) (7)
2) TAKE attention (5)
3) TAKE improvement (progress) (5)

4) TAKE interest(s) (3)

Figure 8. 13 The occurrences of the erroneous collocates relating to 'TAKE place' in COLEC

1 00 deaths per 1,000 births. Why did it take place so many changes. Because the
2 umber of people is increasing. Why did take place this change in the developin
3 hey not all become rain. the other will take placesone changes else. Some rain
4 alth gains in developing countries have taken a great change. Accounding to the
5 fe expectancy and infant mortality have taken great changes since 1960 in the de
6 mortality in developing countries have taken great changes. The fact can be sho
7 taility of the developing countries had taken great changes. As it shows, in 196
8 reform and open policy, our country has taken many changes. The living condition

One obvious misuse in learner English concerning the collocate of the verb *TAKE* is *take place a change (changes)* or *TAKE a change (changes).* The concordances of Figure 8.13 are some examples to show the context and the error.

The misuse of *TAKE place a change (changes)* seems to reveal that the learners misinterpret the intransitive phrase *TAKE place* as transitive. This may be partially accounted for by the influence of the structure of the learners L1 in which the equivalent of *TAKE place* is *fa-sheng* and that of *a change (changes)* is *bian-hua* and the sequential order of the collocation in their L1 is *fa-sheng bian-hua*. Literally this would read *TAKE place a change (changes)* in the order of the English. It seems that in the process of acquisition of the phrase *TAKE place*, the information on transitivity and intransitivity is lost while the sense is obtained (equivalent to the English *HAPPEN*). The production of *TAKE a change (changes)* might be another version of *TAKE place a change (changes)* because it seems to show that the learners are aware of the inappropriateness of treating *TAKE place* as a transitive phrase. In other words, they do not think *TAKE place* can be followed by an object such as *a change (changes)*. However, since they are preoccupied by the sequential order of their L1, they might simply have chosen to remove *place* from the phrase, expecting that this would help to change the intransitive phrase into a transitive one. The fact that both *TAKE* and *TAKE place* are found to collocate *a change (changes)* seems to show the influence of L2. In this case, it supports the belief that both L1 and L2 have a role to play in L2 acquisition.

Apart from the learners' misuse in the collocates *TAKE place a change (changes)* and *TAKE a change (changes)*, there are three other misuses: *TAKE attention* (5), *TAKE improvement (progress)* (5), and *TAKE interest(s)* (3). They could be seen as 'blends' as Howarth (1996)

calls them. This also echoes a similar research study carried out by Chi Man-lai *et al.* (1994), in which they analysed the intermediate-to-advanced level learner English of some L1 Chinese in a million-word corpus by combinations of the verbs *have, make, take, do* and *get* and found that "they are often used as if they were interchangeable" (cited in Nesselhauf 2004: 6). It also seems to show that these blends are signs of partial acquisition from non-acquisition to full-acquisition (cf. Guo: forthcoming). The significance in finding these blends is that as far as the verb *TAKE* is concerned, some efforts can be made so that learners become aware of this problem and manage to overcome the difficulties at this point. Since there are as many as seven occurrences of this kind in COLEC, this difficulty deserves the attention of teachers so that it can be treated as a common problem of the learners in this community, one to be properly solved.

8.6.2 Looking for implicitly deviant uses by the learners

An explicit idea conveyed in the study of CIA is that comparison reveals difference; this difference can be used in various ways. This section uses this feature to discover some important uses by the NSs which are absent from the learner corpus.

It is comparatively easy to notice problematic areas, as discussed above. However, it is not as easy to look for potential problems which are not so explicit to our eyes. One example is the use of *TAKE part in*. In COLEC there are 185 cases of *TAKE part* and almost without exception they are followed by the preposition *in*. Yet if we consult LOCNESS we see that not every sentence would require the preposition *in* to follow. The following is an example from LOCNESS:

With this going on the people *taking part* will be dangerous (…)

Another example of the same problem in COLEC is the phrase *TAKE care*. Most of the examples of the phrase are followed by the preposition *of* while a couple of them are followed by the preposition *about*. As a contrast, there is not a single case in which the phrase is followed by the negative infinitive *not to* or without a preposition.

Apart from the potential to help with the discovery of such unanalysed uses of these two phrases, there is something else a corpus-based approach could do for learner language

studies. Take the simple collocate group *TAKE a class* (*TAKE classes*) for example: even though there are as many as 10 occurrences in LOCNESS, there is no occurrence in COLEC. It is strange to notice that there is a disparity in the collocation *TAKE a class* in the two corpora, considering the similar overall occurrences of *class* in the two corpora (131 in COLEC and 107 in LOCNESS). If the COLEC writers do not use *TAKE* to collocate with 'a class' or 'classes' then what verb or verbs would they use? Thus, a query was made with "class/classes" in COLEC as the node words. It is not surprising to find that a majority of occurrences that collocate with "classes" are of the more often used verb *HAVE*, with a couple of other verbs such as *ATTEND* (see Figure 8.14).

Figure 8. 14 Some examples of "*TAKE a class/classes*" from LOCNESS

```
1            he manager had an opportunity to take a computer class, but chose n
2          to allow welfare mothers to work or take classes. Some liberals    also
3            are certain classes that you must take in order to advance to high sch
4         nts are allowed to remove the ″X″ by taking a class that discusses eth
5        an Politics class that I am currently taking. We have come to know
```

The analysis above deals with the situation in which there is no occurrence of a particular collocation in the learner corpus but it exists in the NS corpus. The analysis offers a reason for this absence. The study suggests that a comparison between the learner corpus and the NS corpus will show the most important disparities between learner English and native speaker English.

8.7 Discussion

Based on the observations of the three groups of collocates, there are two general points that could be drawn from the observations above. One is that the NSs as a whole use more items (types) that collocate with *TAKE*. A second point found is that there is a narrower range of use in the examined collocates in the learner English. In other words, the use by the learners is far less complex compared with that of the NS English.

Considering the disparity between the two groups of writers, there are some implications for the teacher, the learner and the writer of teaching materials. The following are only some examples. In Table 8.1, it is found that the NSs use the discontinuous collocate *TAKE … seriously* fairly frequently (17 instances) but this is hardly used by the learners (only one case).

The teacher could draw the attention of the learners to the existence of such an important construction. The learner may actively consult the behaviour of such a structure and collocate varieties by referring to the concordance lines of the NSs. To further the study, the learner may wish to learn how other varieties of such a collocate type, such as *TAKE … lightly*, *personally*, *to heart*, and *with a grain of salt*, behave in association with the central collocate *TAKE … seriously*, as displayed in Figure 8.15.

In the same vein, the writer of teaching materials may take into consideration the disparity between the types of the collocates in the two corpora, as shown by the bold font in Table 8.1, and systematically make space in a course for the verb *TAKE* and its collocates. He or she could look at the collocates that are not used, or scarcely used, by the learners, such as *TAKE … seriously,* but also at the varieties that the NSs use but not the learners, such as *TAKE a road*, *TAKE an approach*, *TAKE a course*, and *TAKE a direction* in the *TAKE a way* group. If they learn the new in association with the old, and more importantly, learn what is frequently used by the NSs in the target genre and writing style, learners shall have a better chance of improving their productive English and achieving a higher level of writing competence.

Figure 8. 15 All the concordances of the collocate *TAKE … seriously* and its varieties in LOCNESS

1 ompeting as women, and they tend not to take competition as serious as men. T
2 iety where workers are encouraged to take employment seriously - to say nothi
3 n friends and Influence people", but to take it with a grain of salt and not to
4 e of winning. As long as people do not take the lottery too seriously, it remai
5 prayer does not mean that children will take this prayer time in public schoo
6 r own heart and decide the best road to take. <ICLE-US-SCU-0001.2
7 >. An issue this great should not be taken lightly and if it does have seriou
8 ronger case, then the argument could be taken more seriously by the reader.
9 ink if you look at how women are now taken seriously when they report sex cri
10 ous extremes that it cannot possibly be taken seriously. Voltaire creates seri
11 ; 3. The right to be listened to and taken seriously; 4. The right to set you
12 fic' reports. for one such report to be taken to heart it would need the backing
13 personal decision; one that must be taken very seriously, for there's no
14 e some segments of society that are not taking gun control seriously, in ligh
15 eem more realistic and assist people in taking the claim more seriously. Toda
16 chievements of the past: men are not taking the feminist cause seriously anym
17 ty without objections. Unfortunately he took the matter personally and felt they

As far as individual collocates are concerned, if the teacher and the teaching materials writer are made aware of which specific problems their learners suffer from, they may be able to make very accurate and specific plans for the improvement and amendment of the learners'

written English. These problems may be: first, a lack of use of *TAKE ACTION* in passive and relative clauses, and a lack of confidence in using the singular form for the action in the collocation *TAKE ACTION*; second, a lack of knowledge concerning what can be used to collocate with *TAKE place* on the left of the node words; and third, the small range of semantic richness as in *TAKE on*. The information from the analysis can certainly be used to help other learners with a similar background.

Let us review the discussion of the erroneous collocates used by the learners. It is useful that typical errors in using *TAKE* such as *TAKE changes* can be identified by a corpus linguistic approach and corrected. Exercises could be designed to first draw the attention of students to the problem, and then let themselves correct it by imitating the NS English which could be displayed by the concordance lines of the NSs. In cases where the small controlled corpus fails to provide examples, a larger one such as the BoE could be used as follows (Figure 8.16).

Figure 8. 16 Twenty examples of the collocate *CHANGE TAKE place* from the BoE

1	very healthy. <p> Harris: Another change has taken place as well. Now for
2	very positively to the radical change which took place in Mussolini's
3	of rain fell on the 3rd. A major change took place during the 4th and 5th
4	the army's business: A remarkable change is taking place in China: the armed
5	ANOCA at least that positive real change is taking place in South Africa and
6	ill" becomes apparent and a major change does take place in accord with it.
7	relief, although no perceptible change had taken place there. How Gitalis
8	eriod of time before a noticeable change takes place. If you are not
9	ave happened unless a major legal change had taken place: the adoption by
10	says the trend reflects the wider change taking place in the old corporate
11	e views of millions, forcing real change to take place, people start saying
12	justing, or controlling the major changes taking place in their communities.
13	That way, people would see that changes have taken place. <p> The other
14	eas that no matter what political changes take place, they will not be
15	etting so old that certain subtle changes are taking place somehow in the
16	for each individual, great changes can take place, even in an ageing
17	model, however, whatever general changes may take place in the future will
18	and fitness increase. <p> These changes take place by making quite small
19	This effect is due to certain changes that take place in the collagen
20	for Europe too </h> <p> The changes now taking place in the USA have

If they are provided with examples of the actual way in which *CHANGE* collocates with *TAKE place*, learners with a problem in the wider collocate *CHANGE TAKE place* should be able to replace their incorrect uses as shown in Figure 8.16. If the learners are inquisitive, they may not only realise that *change* or *changes* should come on the left of the *TAKE place*, but also learn that the noun *CHANGE* can be modified by a number of adjectival features such as

'radical, major, remarkable, subtle' and so forth.

The diagnostic function of such a tool is also identifiable in the study of the absence of the *TAKE a class* group. On the grounds that so many NSs use this collocation, it is suggested that in order for students to acquire a naturalness in English, they should be encouraged to use the more fixed collocation of *TAKE classes* to replace the more casual and easily produced collocation *HAVE classes*. The concordance lines (in Figure 8.14) from LOCNESS would become a tool by which the learners may improve their naturalness in using collocations.

8.8 Conclusion

This chapter has studied all the collocates of *TAKE* in COLEC and LOCNESS and has found that there exist not only some similarities but also considerable disparities in the collocate types and tokens. The diagnostic function of a learner corpus, when compared with a NS corpus, is becoming more and more explicit to researchers. It is hoped that this diagnostic function can be fully used by English teachers, writers of teaching materials and other ELT practitioners. Awareness-raising could also be used to help students tackle the problems they collectively have. The analysis of the research leads to the necessity to deal with small words like *TAKE*. By taking care of the collocates of such seemingly simple, everyday verbs, it can be envisaged that language production by learners will develop gradually and efficiently towards the clear aim of native-like English.

Chapter Nine

Discussions

9.1 Introduction

This chapter first reviews the methodology I have used in the research chapters from Chapter Four to Chapter Eight. Discussions will continue about the implications of the research findings of the previous chapters, not only individually but also as a whole. Two important functions of the corpus-based learner English study, the diagnostic function and the evaluative function, which I consider innovative in current learner English studies, will be illustrated in detail. Some advice for further studies in this area is also put forward at the end of this chapter.

9.2 The methodology of this research reviewed

The methodology of this research was briefly mentioned only in Chapter One (1.6). Perhaps this is the best place to take up this issue and have it examined in more detail.

9.2.1 The quantitative approach and the qualitative approach in corpus studies

There is no single corpus-linguistic approach to language studies. Some studies rely heavily on the support of numerical figures such as frequency, T-score, Z-score, log-likelihood and other statistical measures. These studies can be seen as being at the end of the quantitative spectrum of corpus-linguistic approaches to learner language studies. Some examples are the studies by Leech *et al.* (2001), Francis *et al*. (1996), Krishnamurthy (2004) and Gui and Yang (2002). Other studies try to analyse and describe language based on minimum support from numerical figures and statistical means. These studies could be interpreted as being at the end of the qualitative spectrum of corpus-linguistic approaches. Some examples are Aston (2002), Hoey (2004), Hunston and Francis (1998 and 1999), Seidlhofer (2002). Very few studies are exactly at one end of the spectrum but are more likely to lie somewhere between the two ends. Quantitative studies are mainly useful as resources for language description and language pedagogy. But the disadvantage with this stance is that researchers can count only what can be

counted and miss out what cannot be counted. In cases where search software meets its limit, the data for research cannot, anyway, be quantified. Therefore, there is a limit to the scope of viewing. Furthermore, there is little interpretation to be directly obtained from such studies. It is very easy to fall into the 'so what' embarrassment as Granger (1998b: 16) and others have pointed out. Qualitative studies rely more on hypothesis-testing, logical reasoning or personal interpretation, and they treat the numerical figures mainly as a spark to start their research. The problem with the qualitative stance is that it is very easy to become bogged down in detail and lose the whole picture.

9.2.2 My research methodology

To exploit the advantages of the two approaches and to avoid the problems mentioned above, my research uses a more quantitative approach for some chapters and a more qualitative approach for other chapters. Two of the chapters, i.e. Chapter Four and Chapter Five, are more quantitative because the research reported in a large portion of each chapter is to quantify the frequencies of verb lemmas (in Chapter Four) and verb forms (in Chapter Five). The remaining chapters, i.e. Chapter Six to Chapter Eight, are less quantitative in terms of the analytical portion involved. In Chapter Four and Chapter Five, the frequency disparity is not only treated as the result but also forms the end part of the chapters. However, in all the other three chapters frequency serves as a starting point for more analysis in the rest of the chapters and my interpretation is provided from my perspective based on some relevant studies in other fields of applied linguistics such as SLA and psycholinguistics. My own experiences in ELT also have a role to play. In Chapter Six for example, the frequency research shows that learners use nouns much less frequently than verbs on the whole. One exception is the use of *examination* in the learner corpus. Instead of treating the phenomenon as an isolated feature of the learner English, I have tried to relate this to the special requirements of relevant topics such as sitting examinations, which is a hot topic in Chinese learners of the English language. I have also tentatively suggested that the high frequency use of *examination* might be a result of earlier acquisition in the first instance; the use becomes fossilised shortly afterwards before there is a chance for the learners to progress. Therefore, later in the chapter I proposed further diachronic studies to test the assumption (see 6.3.1 for details).

9.2.3 Identifying the similarities and disparities between the NNS English and the NS English

A major aim of the research is to identify the similarities and disparities between the COLEC learner English and the LOCNESS NS English. Therefore, comparison between the two corpora is the key to the entire research study. Since a learner corpus contains various uses which range from acceptable to non-acceptable, treating learners' data requires much more time than treating NS data. Learners' errors in my research are not a central interest in themselves because it is my belief that errors are unavoidable for learners in the process of language acquisition (see Guo forthcoming). What is more important to me is to identify which part of the English language, in terms of verbs, is produced more like the NSs, which part is less like the NSs and which part is grossly deviant from the NS use. One exception to my policy of ignoring learner errors is the treatment of some typical errors that are produced quite frequently such as the ones in Figure 9.1.

Figure 9. 1 The occurrences of the erroneous collocates relating to 'TAKE place' in COLEC

1 00 deaths per 1,000 births. Why did it take place so many changes. Because the
2 umber of people is increasing. Why did take place this change in the developin
3 hey not all become rain. the other will take places one changes else. Some rain
4 alth gains in developing countries have taken a great change. Accounding to the
5 fe expectancy and infant mortality have taken great changes since 1960 in the de
6 mortality in developing countries have taken great changes. The fact can be sho
7 taility of the developing countries had taken great changes. As it shows, in 196
8 reform and open policy, our country has taken many changes. The living condition

Individual errors are mostly ignored so that I can concentrate on the typical representative and outstanding features of the group learner English.

9.3 The functions of a NNS vs. NS corpora comparison research

9.3.1 The diagnostic function

"Predicting what learners will need in the way of vocabulary is important in selecting what to teach", as McCarthy (1990: 87) pointed out. A corpus-based contrastive learner language study has a diagnostic function and helps to find out the similarity and the disparity between and ultimately the needs of, the learners. This function could be used to discover what is used and what is not used by the learners compared with the NSs and therefore allows the teacher to see the current status of the collective English and to diagnose the possible problems (both

explicit and inexplicit) with the learners' performance. The following paragraphs review the diagnostic function of the learner corpus versus NS corpus comparison approach in more detail.

As illustrated in Chapter Seven, the use of the terms 'overuse' and 'underuse' is too general and offers very little information on how well performed or poorly performed an area of learner English really is. To overcome the difficulty of the use of 'overuse' and 'underuse' in establishing the difficulties and non-difficulties of learners, I have proposed a system of frequency ratio relationships, as follows:

1) a large frequency in COLEC vs. a large frequency in LOCNESS
2) a large frequency in COLEC vs. a small frequency in LOCNESS
3) a small frequency in COLEC vs. a large frequency in LOCNESS
4) a small frequency in COLEC vs. a small frequency in LOCNESS
5) no frequency in COLEC vs. a small frequency in LOCNESS
6) a small frequency in COLEC vs. no frequency in LOCNESS
7) no frequency in COLEC vs. a large frequency in LOCNESS
8) a large frequency in COLEC vs. no frequency in LOCNESS

With this finer categorisation, the learner language could be interpreted using much more information than with the general use of 'overuse' and 'underuse'. Take the first situation for example (a large frequency in COLEC vs. a large frequency in LOCNESS); if a particular item (no matter whether it is a word or a syntax structure or something else) is used in large frequency in both the learner corpus and the NS corpus, it is most likely to be a well performed item and indicates a better mastery by the learners compared with small-frequency items. It also indicates that less effort from the teacher and the learner is needed for the next stage of study of the language (comparatively speaking). Marking this part of learner English is expected to save much of the teacher's and the learner's time. Another important aspect of information that can be drawn from this frequency ratio relationship is that the use of this item by the learners is justified because it is also used in large quantity by the NSs.

If we look at the third situation, which is 'a small frequency in COLEC vs. a large frequency in LOCNESS', we are informed that an item in such a frequency ratio relationship is very

likely to require more practice and use from the learners, if we may ignore other factors (such as the disparity in topics in the two corpora) for a moment. The information that can be obtained from the finer categorisation is not really accessible if the terms 'overuse' and 'underuse' are used.

Of course, this classification of frequency ratio relationships can be still further improved by adding 'intermediate frequency' between 'large frequency' and 'small frequency', in which case it would be easy to deal with those frequencies which are at the bottom of large frequencies and at the top of small frequencies. Since this study mainly aims at exploring the possibilities of the ultimate use of a corpus-based approach to learner language study, I will leave it to other researchers or other ELT practitioners to 'customise' it in order to make more detailed use of it.

In Chapter Four, 893 verb lemmas and 569 verbs lemmas are found in LOCNESS and COLEC respectively after all the small-frequency verb lemmas are deleted (the cut-off point is set at three inclusive). The verb lemmas that occur only in LOCNESS have been identified and then singled out from those verb lemmas that are found in both of the corpora. If plans are to be made to improve the learners' vocabulary, then the tables in Chapter Four (from Table 4.4 to Table 4.9) could serve as the best reference because they are the words that are used by the NSs in the target register and text type. If the learners could in the end learn to use these words properly, they will most likely be able to produce English in a more native-like way.

Another disparity discovered by the diagnostic function can be illustrated by the actual uses of the verb lemma groups 'argue' and 'oppose'.

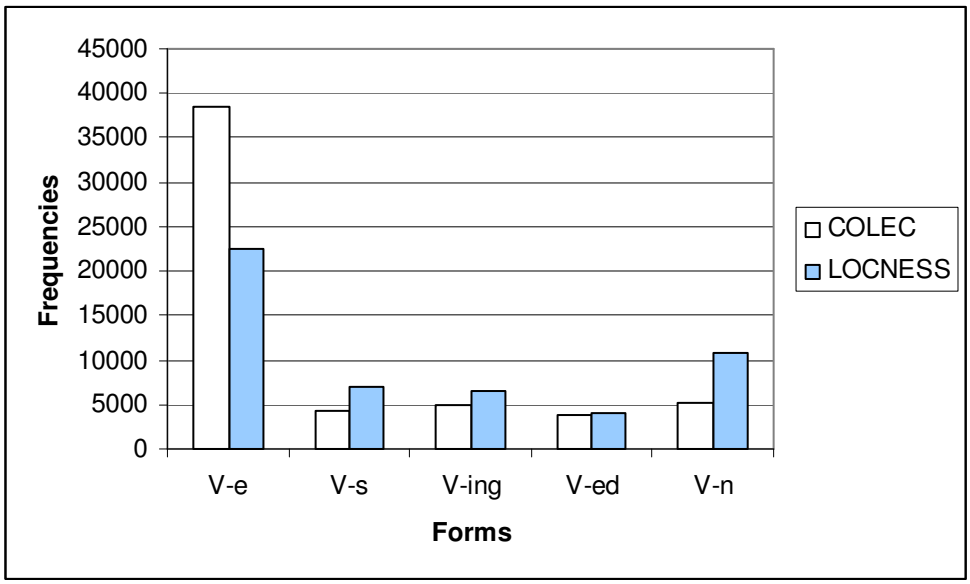
Table 9. 1 Two verb lemma groups used in LOCNESS and COLEC

	LOCNESS	**COLEC**	**Chinese Pin-yin**
argue 4-1	**argue** (162)		bian-lun; tao-lun
	debate (24)		bian-lun; tao-lun
	dispute (4)		zheng-lun; zheng-chao
		quarrel (5)	zheng-chao, chao-nao
	refute (19)		fan-bo, bo-chi
oppose 4-2	**defy** (6)		gong-ran-fan-kang; mie-shi; miao-shi
	object (6)		fan-dui; bu-zan-cheng
	oppose (37)	oppose (4)	fan-dui; fan-kang; di-kang
	resist (6)	resist (18)	di-kang; di-dang; di-zhi; kang-ju

As Table 9.1 shows, the similarities and disparities could be diagnosed by comparison between the learner corpus COLEC and the NS corpus LOCNESS. The similarity between the two corpora in terms of the two verb lemmas is that there are two verb lemmas used in both of the corpora, i.e. *OPPOSE* and *RESIST*; the disparity is 1) in the 'argue' sense group only one verb lemma, *QUARREL,* is found in COLEC while four verb lemmas are found in LOCNESS, i.e. *ARGUE*, *DEBATE*, *DISPUTE* and *REFUTE*; and 2) two more verb lemmas are used in the 'oppose' group in LOCNESS, i.e. *DEFY* and *OBJECT*, as shown in bold font in Table 9.1. The process of identifying the similarities and disparities between the two corpora can actually be used to diagnose the problems of the learners in English production.

Apart from the similarity and disparity in the verb lemma types (893 in LOCNESS and 569 in COLEC), the degree of familiarity with the verb lemmas on the part of the learners could also be diagnosed by the comparative approach. There is no doubt that the more the learners use a particular verb lemma, the more familiar they become with it. Taking the verb lemma *RESIST* as an example, there are as many as 18 occurrences in COLEC, suggesting that a fairly large number of learners use this lemma. If we look at the verb lemma *OPPOSE*, only four cases are used in COLEC, implying a lower level of familiarity because only a few learners use this verb lemma. Therefore, it is reasonable to believe that there is a higher level of familiarity with the verb *RESIST* than with the verb *OPPOSE* in the COLEC learner English.

Figure 9. 2 A bar chart of the normalised frequencies of the verb forms in COLEC and LOCNESS



In addition to its diagnostic function in the learners' use of verb lemmas, the comparative corpus approach has something to offer for the performance of individual forms. If we look at the overall distribution of the different forms of verbs, for example, we can diagnose that the learners rely excessively on the base form for English production (see Figure 9.2). Another obvious feature of the learner English that we can diagnose is that there is inadequate use of the participle forms of verbs compared with in the NS English.

If we look again at the top 20 verb forms used in the two corpora, there exists an obvious disparity between them. Whereas there are as many as six verb lemmas whose individual forms all appear in the top 20 in LOCNESS (see Table 5.3), i.e. *MAKE*, *TAKE*, *BECOME*, *USE*, *SAY* and *GIVE*, there are only three verbs of this type in COLEC (see Table 5.4), i.e. *MAKE*, *TAKE* and *GET*. If there are more verb lemmas whose individual forms appear in the top 20, there must be a large enough number of writers who write these forms, not because they have agreed to use the same form but because there exists an agreed and shared knowledge and tendency in the writers' minds. Therefore, I believe that there is a better homogeneity in the NS English than in the learner English. There is also every reason to believe that the more homogeneous a collection of learner English is, the more the learners resemble each other in language production. If the learners' homogeneity resembles that of the NSs, then the learner English can be considered to be at very advanced level of production. By the same token, if the learners' homogeneity deviates too much from that of the NSs', then the learner English can be deemed to be at very low level of production. As is shown in Chapter Five, there is an active role which a comparative corpus analysis could play in diagnosing how homogeneous a collection of learner English is.

As discussed in Chapter Five, one of the important results of corpus linguistics research is that different forms of verbs behave quite differently. A dictionary has very little information to provide as to which form of a verb is used more frequently than the others and in what way. A comparative corpus analysis is effective in discovering this information. Take the V-ing form for example; whereas some V-ing forms appear both in the two corpora, there are others that do not match each other. Knowing which V-ing forms are used only in LOCNESS is possible via a comparison between the learner corpus and the NS corpus. Figure 9.3 shows these V-ing forms, and this is useful for learner language research.

Figure 9. 3 The verbs that are found only in LOCNESS in the top 20 V-ing word forms

trying taking saying giving looking allowing running killing showing thinking fighting

By means of comparative analysis between COLEC and LOCNESS, researchers can access information not only about the V-ing word forms that are only found in LOCNESS in the top 20 word forms, but also about the other word forms (the base form, the third person singular form, the past form and the past participle form) that are found only in the top 20 word forms in the NS corpus.

Finding out which verb lemmas are used only by the NSs and which forms of which verbs are used only by the NSs is one important aspect of understanding better the features of learner English because it helps the researcher to diagnose which verb lemmas and which forms of which verbs should become the priority for learning. In line with the macro perspective taken above in trying to detect the features of learner English, I would like to treat this perspective as a 'panoramic view'. Parallel to this view, there is a 'zoomed view' which I have also taken in this learner English study.

In the detailed analyses of two simple words *KEEP* and *TAKE*, all the concordances are checked, even though from different aspects. In the study of *KEEP* (see Chapter Seven), all the uses of the verb lemma *KEEP* are examined and classified into patterns, in line with Hunston and Francis (1999). Those uses that cannot be grouped into patterns are labelled as phrases in a very broad sense. One important finding obtained from the study of *KEEP* is that the two groups of writers use different patterns to express the same or similar things. The different patterns used to express these meanings, as duplicated in Table 9.2, are only some examples.

Table 9. 2 Some examples of using different patterns to mean the same thing

NS English	Pattern	NNS English	Pattern
keep calm (BoE)	**KEEP adj**	keep a calm head	**KEEP n**
keep fit (BoE)	**KEEP adj**	keep a good health	**KEEP n**
		keep our own physical fitness	**KEEP n**
keep her happy (LOCNESS)	**KEEP n adj**	keep their happiness	**KEEP n**

Even though learner English is often criticised for being largely correct in grammar but having a strongly unnatural flavour, this unnaturalness is not easy to discover without such a

perspective and methodology.

Another important characteristic that has been diagnosed by such a comparative investigation into the patterns of *KEEP* is that the COLEC writers use very frequently the patterns that express the sense of *continue* and *maintain* (such as **KEEP n**, **KEEP –ing**, **KEEP *on* –ing** and **KEEP up –ing**). My suspicion about the learners' over-reliance on these patterns is that the learners do not use alternative expressions as the NSs do. A check of the frequencies and detailed uses of the verb lemma *CONTINUE* and *MAINTAIN* shows that my suspicion is proved to be correct in this case (see Table 9.3).

Table 9. 3 Comparative frequencies of *CONTINUE* and *MAINTAIN* in COLEC and LOCNESS

	COLEC		LOCNESS	
	R F	N F	R F	N F
CONTINUE	50	52	177	274
MAINTAIN	9	9	35	54

Based on the understanding above, it can be hypothesised that in cases where the NSs use *CONTINUE* and *MAINTAIN*, the learners would use their own favourite patterns. When the concordance lines are examined in detail, the hypothesis is proved to be correct. The following are only some examples (with the NS use underlined) and there are many other similar uses of this kind in the two corpora.

LOCNESS: As long as they can make a buck, criminals will *continue* to believe that crime pays well in America.
COLEC: If you have no good health, you'll hardly *keep on doing* your work …
COLEC: If we *kept on*, we will make a great progress in English.

LOCNESS: Britain has been eager to *maintain a secure balance* of power on the continent…
LOCNESS: you are helping to *maintain a balance* of the number of lower income families …
COLEC: By doing so, I thought, we can *keep the balance* of water circulation.

LOCNESS: Pangloss himself, although in this sorry state, still *maintains his optimism* …
COLEC: we *keep high spirits* and keep on working.

Parallel to the study of the verb lemma *KEEP* is an investigation into the verb lemma *TAKE*

(see Chapter Eight). This verb is looked into from the perspective of collocates (rather than 'patterns'). The diagnostic function of comparative corpora analysis between a learner corpus and a NS corpus is evident throughout the whole study of the verb. The first point that has been detected is that the learners use the same collocate quite differently in terms of the environment. In a study of *TAKE action* (*actions*) for example, it is found that the COLEC writers produce much more cases in the active voice than the NSs do, and therefore far fewer cases in the passive voice. A second point discovered by the comparative approach concerns the subjects of the idiom *TAKE place*. Whereas there are a variety of things that are used as the subject of *TAKE place* in LOCNESS such as *action* (*activity*) (5), *event* (3) and *incident* (1), and many activity-related words (for details see 8.5.2), there are a very limited number of things that occur in the subject position in the COLEC writings. More than 61 percent of the subjects are actually either *change* or *changes* in COLEC. A third point that is detected by the comparative approach is about the use of a phrasal verb, *TAKE on*. Whereas there are four senses identified in the concordances of the phrasal verb in LOCNESS, there are only two identified in COLEC. In this way all the collocates of any verbs that appear in a learner corpus and a NS corpus could be compared and identified thoroughly. With a clear picture of the similarity and disparity between the learner English and the NS English, researchers are in a much better position to understand the current status of the learner English and ultimately the needs of the learners.

Apart from being able to discover the correct but different uses by the learners, the comparison is especially good at exposing the incorrect uses due to their non-existence in the NS corpus. In the study of *TAKE* for example, a typical erroneous use by the learners is the use of *TAKE place changes* (see 8.6.1 for details). There are also other problematic collocates found as follows:

1) TAKE a change (changes) (7)
2) TAKE attention (5)
3) TAKE improvement (progress) (5)
4) TAKE interest(s) (3)

After a thorough comparison of the verb lemma *KEEP* in the two corpora in terms of patterns and phrases, and of the verb lemma *TAKE* in collocates, some typical problems that exist in these two simple verb lemmas in the learner English have been revealed. The identification of

these problems could become the starting point for some serious applications in ELT.

One more aspect that my research has shown in relation to the diagnostic function of a comparative corpus approach concerns the learners' preference for using particular POS vocabulary or a particular function of multiple-POS vocabulary (see Chapter Six for details). It is successfully diagnosed that the learners have a strong tendency to use verbs compared with nouns, and verb functions of multiple-POS vocabulary compared with the noun function of multiple-POS vocabulary. The following pairs of sentences below show the option tendency in the two groups of writers. The concordances of the NS English are underlined.

LOCNESS: It is essential that society examine these arguments and then decide on what is acceptable and what is not acceptable before it gets out of *control*.

COLEC: [if we] can't *control* [our] mind, […] we can't do anything at all.

LOCNESS: Schools and some hospitals, households are already publicised as "beef free" and this is on the *increase* causing a fall in the demand for beef in the U.K.

COLEC: The population is *increasing* and the industry demands more and more water.

It seems that the COLEC examples could well be rephrased as follows if the learners wished to use nouns in prepositional phrases:

1) If our mind gets *out of control*, we can't do anything at all.
2) The population is *on the increase* and the industry demands more and more water.

To conclude, there is a strong diagnostic function found in the comparative corpus approach I have been using in this research into learners' use of verbs. This function has not hitherto been illustrated and generalised, as far as I know. With the aid of such a powerful function of the comparative approach, it is expected that the most immediate needs of the learners would be established gradually and successfully.

9.3.2 The evaluative function

Apart from the observed diagnostic function of the comparative learner corpus approach, there is also a potential function which I would call 'the evaluative function'. On the one hand this might help to find indicators of high-level or low-level performance from the collective

English and on the other hand it might be used for the evaluation of the degree of group learner English from the comparison between the COLEC learner English and the LOCNESS NS English.

Table 9. 4 Some examples of the correct use and incorrect use of *KEEP in touch with* in COLEC

ID	Correct Use	M	ID	Incorrect Use	M
452823	I will *keep in touch with* them and communicate with each other.	15	451115	By *doing more touch with* the people in society…	12
650318	we should *keep in touch* with all sorts of information around us.	13	650517	to *keep touch with* the world outside.	9
451115	They only *keep in touch* with the knowledge in book.	12	453130	This is good way to *keep touch with* the society.	9
650514	we should also *keep in touch with* the senior or graduated college students …	12	640312	Without *keep touching with* the society…	9
451922	I should *keep in touch with* it.	10	650527	*have the touch with* the society.	8
650513	I should always *keep in touch with* the outside world.	11	650322	I will do a part-time job to *touch with* world outside.	8
440618	How can we *keep in touch with* outside?	9	650613	There are many ways to *keep in touch of* the outside the campus.	8
440618	By these means, we can *keep in touch with* outside.	9	440903	We seldom *get touch with* the society.	8
650527	Having realized where and how we can get help to *keep in touch with* the society.	7	no 0379	I must *keep touch with* the society.	7
452861	it can make them *keep in touch with* world.	6	431102	Because they want to touch with the new thing…	7
AVERAGE		10.4			8.5

In the study of the patterns and phrases of the verb lemma *KEEP*, I looked at the possible use of large-frequency and low-frequency items in evaluating individual learners. This may sound bizarre because it is hard to understand how group learner English could be used to measure the levels of individual writers. The hypothesis behind this is that those writers who do not produce popular items (such as *KEEP in touch*) properly may be at an earlier stage of their acquisition (compared with those who produce them properly) and therefore, the level of these learners is likely to be lower than in those who use the items correctly. As shown in the last row of Table 9.4, the average mark of those who produce the item correctly is higher than that of those who do not (10.4 vs. 8.5).

As discussed in Chapter Seven, this study seems to suggest the following things:

1) A learner who uses a commonly used item (such as '**KEEP *in touch with* n**') has a

higher score for the entire essay than one who has problems with such an item on the whole.

2) It is very likely that a learner who uses correctly an item commonly used by his/her NS peers will have a high score for the entire essay, but not necessarily.
3) It is very unlikely that those who do not use a commonly-used item correctly will have a high score.
4) Since there are multiple factors contributing to a high score, and one example of correct use contributes to a composition high score but does not automatically lead to it, and vice versa, the existence of disparity between individual markers may contradict the general trend as stated in 1), 2) and 3) above.

Along with the initial study on the possible use of large-frequency items in the learner English, a study on the possible use of small-frequency items is also conducted in Chapter Seven. However, due to the constraints of other parameters such as the consistency of different markers and the poor reliability of low-level items, there is no significant co-relationship found between the overall frequency of a particular item and the level of the composition in which the item occurs (see 7.4.1.10 for details). Therefore what I can claim at this moment is that learner corpus studies have the potential for evaluative purposes.

9.4 Some pedagogical implications of the research

9.4.1 Teaching material enhancement

The first implication of the learner corpus study rests with the enhancement of teaching materials for these learners and for the next generation of learners with the same background. Before the advent of learner corpora, teaching materials were mainly based on the experiences and intuition of teachers in deciding what should be taught and what should not be taught to students. Though some of the teaching materials may work fairly well, there have been no measures and means to help course-writers check whether their teaching materials really reflect the needs of the learners. The corpora comparison in this research has successfully found out some essential needs of the COLEC writers in using verbs. For example, the COLEC writers use approximately 569 verb lemmas while the LOCNESS writers use 893. The verb lemmas that occur only in the NS corpus should be reflected in teaching materials

which are intended to be used by these learners. There is no doubt that these verb lemmas must appear frequently enough in the reading materials first. When there is adequate exposure, it is envisaged that these verb lemmas will be gradually imitated by the learners when writing tasks require them to produce the verbs anticipated. Of course the learners may not use the verbs in the writing tasks because it is a general habit of language students to play safe and stick to the 'teddy bear' vocabulary in which they have confidence (see 2.7.2 for the 'teddy bear principle'). Nevertheless, exercises could be designed to help learners to replace the familiar lexical items with new ones. Without adequate practice, their vocabulary size would have no chance to improve. The new ones, i.e. the vocabulary that is used only by the NSs, are now available by means of a corpus-based contrastive study between learner English and NS English (see Figure 4.13 for details). These verb lemmas should become the target vocabulary for the learners to practise and master. Teaching-material writers may like to emphasise and highlight these lemmas whenever they appear.

Talking about the verb lemmas that exist only in LOCNESS, we do not expect the learners to practise all of them at the same time. The verb lemmas that occur more frequently have priority over the less frequently used ones. The more used senses of a verb also take priority over the less used senses.

As discussed in Chapter Four, vocabulary is easier to learn in semantic groups. It is suggested that when a new verb is introduced to the learners it is best if it appears together with the verbs that are familiar to the learners. These verbs can be found in the tables in Chapter Four. Take the verb lemma *COMPREHEND* for example, if it could be presented to the learners with *KNOW*, it is anticipated that the learners would establish a semantic link between the two verbs, in which case acquisition should become easier. Verbs appear in verb forms rather than verb lemmas in texts; therefore, there is a need to discuss which forms should be presented to the learners first. In traditional vocabulary lists, verbs are mostly presented in the base form. This does not seem to help learners very much because it offers no information as to which form is more used than the other forms and how. Given the fact that different forms of verbs may perform quite differently (see 5.1 for details), I would argue that when a new verb is presented to learners the most often used form should be selected first. Even though for some verbs it matters very little whether the base form or another form is introduced to learners, for

others it does matter. Take the verb lemma *DEEM* for example, if we look at the concordance lines in Figure 9.4, it would not be difficult to find that there is a very uneven distribution in the individual forms. The most often used form is the V-n form in the passive voice (12 out of 17, as in bold). If learners take the trouble to practise every form, it is not only time-consuming but also at odds with the practical use in the NS English. In contrast, if learners are introduced direct to the V-n form and its passive use, it would be perfectly in conformity with the NS use.

Figure 9. 4 The concordances of the verb *DEEM* in LOCNESS

1 ower to quickly dismiss cases that they deem frivolous. Judges must also be
2 to choose to have children if they are deemed suitable by certain tests. If the
3 there will this discovery in case he be deemed a heretic. Unfortunately this som
4 misconduct because of the fear of being deemed as racist. This led to abuse o
5 with these cases. Once a lawsuit is deemed unfounded, the person filing the
6 thousand in the earthquake at Lisbon is deemed as God's will and for the good of
7 iayev. This proves that if the cause is deemed just and the women are prepared f
8 the emergence of another source of law, deemed to be supra-national is inevitabl
9 The infertile couple's needs are often deemed much more important than the s
10 ery organisers, Camelot, obtain. People deemed this to be far too high and belie
11 gone to just one lucky winner. This was deemed by the Bishop of Durham as æan ob
12 kpot winners were likely to receive was deemed as ætoo muchÆ or even æunseemlyÆ.
13 ce. Membership fell as the unions were deemed ineffective in securing worker de
14 <?> legislation, I feel that all women, deemed suitable via guidelines, of havin
15 wer also have control over what society deems to be deviant. What they label
16 who they are no matter what society deems a real woman. In our society, w
17 raditional structure of what society deems a warm. Media also plays a huge

After it is certain that the V-n form of the verb *DEEM* is known to learners, the second or the remaining forms may be introduced to them, in this case the forms are the V-e form and the V-s form. Since the V-ed form and the V-ing form are missing in the LOCNESS corpus, we might as well go without introducing them to the learners. When these learners are advanced enough, they would be expected to be able to use other forms. At this moment, it would be enough to let the learners know how to use the V-n form and the V-e/V-s forms. This is largely in accordance with Dave Willis' 'lexical syllabus' (Willis 1990), in which the teaching of lexis (rather than grammar) should play the central role in the language classroom. Before the era of corpus linguistics, it was not possible to see the uneven distribution of the different forms of verbs. By using the corpus data, I have found out which forms are used more compared with other forms.

Another implication of this research is that the COLEC writers' use of simple verbs like

KEEP and *TAKE* reveals a sharp difference (though some degree of similarity) between the learners' performance and the NSs' performance. Considering the importance of simple and small verbs in the English language, there is a need to spend more time on these verbs (see 7.1 for some details). Since the learners produce these simple verbs in large numbers, if they could write them appropriately their English would approximate the norm of English. Therefore, it is suggested that sufficient time be spent on a small number (say about twenty) of verbs and all of them be practised heavily so that a higher level of English can be achieved in a fairly short period of time. Practice on familiar things only increases the degree of familiarity (such as in the use of *TAKE ACTION*) but will not help the learners learn new things. Since this research has discovered a substantial part of the learners' needs, the teaching material writers may take this advantage and try to make teaching and learning easier than before.

9.4.2 CALL software development

The previous section has discussed the potential of applying the research findings in this study into the enhancement of teaching materials on paper. Since modern technologies are playing a more and more important part in the language teaching industry, there is a strong motivation for us to translate the research findings into user-friendly computer-aided language learning (CALL) software. From Chapter Four to Chapter Eight verbs are studied from different aspects. Perhaps this is the best place to systemise the individual studies and research findings and see how they could be used in CALL. What follows is a very raw idea that could be translated into possible finer designs with the support of available technologies.

9.4.2.1 Step one: analysing all the verbs that occur in both of the corpora

It is shown in Chapter Eight that there exists some degree of similarity and disparity between the collocates of the verb *TAKE* in the two corpora. By revealing the similarity and disparity, especially the latter, the learners are presented with a list of the items that they could try to practise (as far as the verb *TAKE* is concerned) so that their English may become more and more natural. To make full use of the research findings, it is suggested that all the verbs that occur both in COLEC and LOCNESS are studied first, as has been done in the case of the verb *TAKE* in Chapter Eight. Considering the large number of these verbs, a team of trained

researchers or teachers would be required to complete the task. The research findings of each word could be made into two separate files for later use, one being the COLEC verb file and the other the LOCNESS verb file. If the files could be saved in 'html' format it would be easier to establish links between them and the corresponding lemmas.

9.4.2.2 Step two: linking the detailed use of different forms and the verb lemmas

The verb lemmas in the lists alone provide the teacher and the learner with little information. However, if the examined verb behaviour could be made into files and be linked with the verb lemmas, the information available to the teacher and the learner is greatly increased. This link could easily be realised by hypertext links. Once the verb lemma in the verb lemma lists is linked to its own detailed behaviour in the two corpora, the learner may simply choose the verbs he or she is not familiar with and improve on them by clicking the hypertext link.

9.4.3 Some implications for the ELT classroom

Apart from the possible applications of the research in the design of teaching material, there are other possible areas for this research to be translated into applications. This section addresses the potential use of the research in the English-language classroom.

Since the start of data-driven learning (DDL) which was initiated by Tim Johns (cf. Johns 1988, 1991, 1994 and 2002), the idea of using authentic language data in the classroom has become popular and has gradually taken hold in corpus-related research and language pedagogy. DDL, as defined by Johns and King (1991: iii, cited in Granger and Tribble 1998: 200), is 'the use in the classroom of computer-generated concordances to get students to explore regularities of patterning in the target language and the development of activities and exercises based on concordance output'. Some of the explorations in DDL have been conducted by Granger and Tribble (1998), Flowerdew 2001, Horvath 1999, Milton and Hyland 1999, Sripicharn 2002, Bernardini, 2002 and Seidlhofer (2002). Though these researchers approach the issue from different perspectives, there is a common belief that DDL can be used wisely to aid language teaching in the classroom by raising language awareness (Hawkins 1984) and self-discovery. Enlightened by the spirit of these explorations and also

based upon a fairly thorough research into verbs, I would like to explore further how the DDL perspective could be broadened and how my research findings can be used to help learners with the verbs they are expected to learn and practise, setting NS English as the norm to follow.

If we take the POS preference by the COLEC writers and the LOCNESS writers as an example, if the learners could be informed that they are using verbs (or the verb function of norbs) excessively, then there would be a chance for them to try the use of nouns (or the noun function of norbs). A suggestion to help the learners to realise this point is to ask them to compare verb and noun pairs such as *accept* vs. *acceptance*, *compare* vs. *comparison*, *enter* vs. *entry*, *survive* vs. *survival*. The following figures from 9.5 to 9.8 are all the concordances of *COMPARE* and *COMPARISON* in the two corpora. In cases where some concordance lines share the same syntactic structure and there are many concordances of this type, some will be omitted to save space.

Figure 9. 5 The concordances of the verb (lemma) *COMPARE* in LOCNESS

1 cess by that dollar figure also must compare and be competitive with others.
2 ving both drinkers and non-drinkers. To compare between the two, they classif
3 transported) and more safely than cars (compare injuries due to bus or train cr
4 rench universities, especially when you compare it with the British system of re
5 higher education level. I will briefly compare it to points in the English syst
6 e equally compensated. But how do we compare raising a family of four childre
7 mprehensive education", also tend to compare sex education in basic ideas of
8 he age of 65. Let's take time to compare the criminal life to the life of
9 matter of time before it is legalized. Compare the situation we are in now t
10 been clearly present. When forced to compare these arguments, it is clear tha
11 number, which many people could not compare to anything, thus losing the val
12 ft -- He told me that nothing could compare to the way he had been forced to
13 better. Society has never actually compared teachers to highly respected fi
14 he total number of prescriptions filled compared to patient suicides. An esti
38 **est less effective for all evolved when compared with the former version of M**
39 lly become so expensive to produce beef compared with profits, that mass rearing
40 **w enough liberties in this country when compared with other nations of similar p**
41 **an' and sees how he is dehumanised when compared with earlier man. The social**
42 tive features of the European Community compared with other international bodies
43 EEC however is a distinctive Community compared with other international organi
44 ering the experience of a garment as compared with his experience with other
45 Candide agrees with this philosophy and compares it to his tutor and mentor Dr P
46 ports Illustrated swimsuit edition. She compares the lack of coverage of fema
47 for example, with a prayer in school compares to trying to extinguish a burni
48 lain how the system works today, how it compares to England and why, despite att
49 ve on the interest income alone. Comparing both options here, I'd definit

Figure 9. 6 The concordances of the noun *COMPARISON* (both singular and plural) in LOCNESS

1 from the Grandes Ecoles. For a general comparison between the French & English
2 ese stories also correlate a generic comparison for person that have reflecte
3 rt. But players are not asking for a comparison of looks, but for a sense of
4 ent damages his argument because the comparison of music taste to concerns of
5 vices to society equally valuable in comparison to marketplace "jobs". Theref
6 e "desire" for money could be gauged in comparison to evil acts committed, bu
7 ple are landless. Montesquieu makes a comparison with China which had laws to
8 evokes sympathy for Caligula through a comparison with the Patricians. When thi
9 treated as second-class citizens in comparison with the men. I believe that
10 ail road cars full of ash per day. This comparison yet again eases the enviro
11 ritual influence on a society. He makes comparisons between societies in cold cl
12 nsumed, who consumed regularly, even comparisons to surveys given in years pa
13 of nature because man rarely met and no comparisons were ever made. However, a
14 of problems concerning world-wide money comparisons would almost be abandoned.
15 tion of guilt. However, despite all the comparisons you can draw from Clarence t

Figure 9. 7 The concordances of the verb *COMPARE* (lemma) in COLEC

1 when you learn words by heart you can compare a word to another approximate
2 jobs will gain different skills and can compare different job each other. In
3 our study. By this previous plan we can compare it with our achievement that we
4 ing countries are changed. Now, we can compare the life expectancy and the infa
5 ere's the different between them? Let's compare them two. First, Pop Music is ea
6 increased in the developing countries. Compare with 1990, many people arenot en
7 reproduce this commodities and sale it. Compare with the real one, the cost of f
8 ina. Bicycle is the fittest transport. Compare with car bicycle has both advant
9 ide when and where to go by yourself. Compare with the car, bicycle is easier
10 is level, as a result, his geting can't compare with his lost. Another, when we
11 Although we have plenty of fresh water, compare with the big consume, the fresh
12 life. Compared the positive with the past, hea
13 is 100 deaths per 1,000 births in 1990. Compared the four data we can concluded
20 and it does good for people's health. Compared to a car, a bicycle has disadva
21 ociety is a completely different world compared to their campus. This results i
22 The same as infant mortality in China. Compared with 200 per 1000 births in 196
42 we recognize them, we can use skill. By comparing and imagining and so on, we ca
43 . And real friendhip is not easily won. Comparing money with friends, I prefer t
44 atching TV, seeing films and so on. Comparing the two sides, I agree to do i
45 lity is 200 deaths per 1,000 births. In comparing the life expectancy is 60 year
49 ng by a leaf" - the old chinese saying. Comparing with the cool weather, there i
50 do the job better if he often does it. Comparing with those who often change wo
51 infant mortaility run encount tendency. Comparing with 1960, Chinese infant mort
52 ent years. The change can be found by comparing. In 1960 life expectancy in de

Figure 9. 8 The concordances of the noun *COMPARISON* in COLEC

1 It's convenite to go to work by bike. Comparison with the buses. In rush hour,
2 resh water is becoming less and less in comparison with the increasing populatio

There are a number of ways in which the teacher may make use of the concordance lines from

Figure 9.5 to Figure 9.8. Firstly, by bringing the learners' attention to the striking difference between the frequencies of the noun *COMPARISON* in the two corpora, it is hoped that the learners will realise that the NS writers tend to use many more nouns; if we note the total number, 15, in the LOCNESS corpus (322464) (see Figure 9.6), we would expect as many as 22 in the COLEC corpus (480063) (see Figure 9.8). If the learners look at the noun use in their own production, they may find only two occurrences with one misused (see Figure 9.8). In this way it is expected that the learners' awareness of their current choice between verbs and nouns will be raised appropriately.

Secondly, the concordances could be used to inform the learners about their verb use. For instance, *COMPARE* could be replaced by the noun *COMPARISON* by examining the actual concordances of the noun use by the NSs. For example, some verb uses in COLEC (as highlighted in bold in the 3rd, 4th and 5th lines in Figure 9.7) could be replaced by the noun equivalent as in the collocation *make a comparison* or *make comparisons* as highlighted in bold in the 7th , 11th and 13th lines in Figure 9.6. In order for us to look more closely at the two concordances (the 3rd, and 4th and 5th lines in Figure 9.7) in COLEC, the KWIK format is shifted into the original text format with a minimum of context.

1) Now, we can *compare* the life expectancy and the infant mortality of 1990's with them of 1960's. (COLEC)
2) Secondly, a study plan can help us have a clearly understanding for what we have done on our study. By this previous plan we can *compare* it with our achievement that we have got, so that we can know if our study plan is useful. (COLEC)
3) Where's the different between them? Let's *compare* them two. (COLEC)

If the learners are to imitate the use of *COMPARISON* to make a similar expression in LOCNESS, the following suggestions could be made:

1) Now we can *make comparisons* between the life expectancy and the infant mortality of 1990's and that of 1960's.
2) Secondly, a study plan can help us have a clear understanding of what we have done in our study. By *making a comparison* between our previous plan and what we have done, we can know if our study plan is useful.

3) What is the difference between them? Let's *make comparisons* between the two.

Thirdly, the learners could be asked to look for the peculiarities of their own use of the verb in terms of context and position of the verb *COMPARE* by contrasting their own use and the LOCNESS writers' use. Before this comparison, they could also be asked to point out the most typical syntactic structures of the verb (which is 'bi-jiao' in the learners L1). Hopefully they would agree to the frequent use, 'yu … xiang bi-jiao', which means 'compared with'. Since the Chinese 'bi-jiao' is very frequently used in the initial position of a sentence, the learners would be expected to point out that more than half of the occurrences of *COMPARE* in COLEC appear in the initial position in sentences (27 out of 52) (see Figure 9.7). Once the learners are made aware of this disparity, they would be expected to carry out a highly-motivated self-discovery of how the NS would use the verb, or in other words, in what position the NSs would put the verb. If the learners could point out the relevant concordance lines (such as 38, 40 and 41 as highlighted in Figure 9.5), then that would suggest that they have discovered for themselves the NS way of using this verb in similar situations. As the NS English shows below, for the NSs *compared with* does not have to appear in the initial position in sentences. Therefore it would be desirable if the COLEC learners could try to use this combination in the middle of sentences, preferably with the conjunction *when*.

38 est less effective for all evolved **when compared with** the former version of M
40 w enough liberties in this country **when compared with** other nations of similar p
41 an' and sees how he is dehumanised **when compared with** earlier man. The social

It is anticipated that the learners would soon start to use *COMPARE* in the way the NSs do in similar situations. To enhance the effect of making this discovery for themselves, the learners could be asked to practise the NS use before they leave, hopefully for other discoveries.

It may be argued that there is nothing wrong with using *COMPARE* in the initial position in sentences because such use may also be found in the BoE and other sources (even though marginally). However, frequently placing a word in an unusual position, compared with the use of NSs, would affect the ability to convey one's meaning effectively. Furthermore, if learners refuse to learn how NSs use a word, such as in the case of *COMPARE*, they are likely to find it difficult to understand NSs' English when a NS utters this word in a different way, such as when using the word in the middle of a sentence, plus a combinatory use with the

conjunction *when*.

This study concerns the verb behaviour of COLEC and LOCNESS writers, and the farthest point away from verbs is its comparative analysis between verbs and nouns in Chapter Six. If DDL were used in the classroom, there would be no constraints upon the POSes. The learners may look at any POS vocabulary for self-discovery once they are familiarised with how to place a query in concordancing software such as WordSmith Tools.

Apart from single words, learners could be taught to make complicated queries that are intended for multiple words such as verbal phrases and syntactic structures; these are studied in Chapter Seven and Chapter Eight but not extensively. It is always said that learners' English suffers seriously from a lack of phrasal verbs. Actually the DDL approach has made it very easy to see which verbs are frequently used in phrasal verbs. In a POS-tagged corpus, it is easy to see what verbs are followed by a particular preposition. What follows are the phrasal verbs with the preposition *up* with a frequency above three inclusive (see Appendix 7 for a full list):

> back up, bring up, build up, catch up, clean up, clear up, come up, cover up, draw up, end up, give up, grow up, hold up, make up, open up, pick up, put up, set up, speed up, take up, wake up

By the same means it is also possible to see what prepositions (or particles, as others call them) follow a particular verb in the NS English. If learners can discover these phrasal verbs, they stand a better chance of starting to practise them soon in their own production.

Though there are so many advantages to it, as described above (also cf. other DDL studies as mentioned in the previous section), the teacher must bear in mind that DDL is best treated as complementary, assisting his or her habitual teaching but preferably not dominating the whole process of classroom activities. In order for the DDL methodology to work harder, the teacher must make adequate preparations and take proper control of classroom concordancing activities.

9.4.4 Some implications for dictionary compilation

Traditionally dictionaries have been made for a mixed purpose of interpretation and

production. In order to accommodate the multiplicity of purposes (along with the explosion of knowledge), dictionaries are being made thicker and thicker. But thick dictionaries for a mixed purpose are not necessarily much help to learners who wish to know more details about how to use a particular word in specific situations. Few dictionaries could afford to list several examples of a usage for one word due to limitations of space. If we could compile a dictionary based on the NS performance in a particular text type, such as students' argumentative writing, there would be enough space to include many details. The dictionary does not have to include a large number of entries because there is a limit to the size of the active vocabulary that learners could actually learn to use, but it is desirable to cover the words that occur fairly frequently in the controlled corpus. Since this investigation into the verb lemmas used by the LOCNESS writers has found only 893 verb lemmas, a new dictionary of practical NS-written English in argumentative writing does not need to exceed 1000 in terms of verbs if the research findings are taken on board. With only 1000 verbs to accommodate, a lot more details concerning the actual use could be made available in the dictionary. Take the verb *KEEP* for example; a new dictionary may contain as many representative collocates of *KEEP* as possible. Theoretically, anything can be kept as long as it has a feature to be stored or maintained. In practice, however, this is not the case. To make a list of potential collocates should be helpful to learners who have doubts about what things can be 'kept' and 'maintained'. Even though this list cannot be exhaustive, a learner may stand a much better chance of finding a relevant example in it. As far as I know, no dictionary provides such detailed practical information as this:

1) a baby, a house, money, animals (such as cattle);
2) a price, a philosophy, civil peace;
3) a tradition, an institution, a sport, the National Lottery, a monarchy, the presidency, an identification, a cultural identity, an advantage, slavery;
4) mutual trust, friends, a support, one's interest;
5) control, order;
6) score, records.

Apart from the necessity to separate production dictionaries from interpretation dictionaries, there is a need to consult the existing knowledge of the targeted learners because without

adequate information of potential users already know, a dictionary will miss the target. There have been some reports on using learners' written data in dictionary-making such as *The Longman Learners' Corpus* (LLC), (see Gillard and Gadsby, 1998). This dictionary is certainly useful for general learning purposes. But there is no sign that it has tried to distinguish the nationality, cultural background and education experience of the users. It can be argued that an archetypal user of a dictionary actually does not exist and that dictionaries must be made specifically to meet the different needs of local users (especially the Chinese users whose L1 is so remote from English). The information that has been obtained from the COLEC learners could largely be treated as specific information that reveals the needs of the Chinese learners of English.

9.5 Some advice for further research

Based on my current research, I can envisage that the following studies are worth carrying out in the area of learner language studies.

9.5.1 Diachronic studies of learner language study

In essence, my current study is a synchronic comparative study of learner language and NS language. It depicts the language used by different individual writers at roughly the same time, i.e. when they reach a certain level of competence . Actually, if the learner language could be studied from a diachronic perspective, that is to say, the development of learner language, more features of learner language could be investigated and more research questions answered. For example, at what period of English writing does an individual learner start to produce a particular item? Does this particular item appear to exist in many individual learners of the same type? Which verb lemmas appear first and which at a later stage? And which lemmas would never appear in the time span of the investigation? What are the most often used verb lemmas at certain stages? What is the disparity between typical writers and atypical writers? Do learners produce every word correctly the first time they use it (see Guo forthcoming)? If some new words are not used correctly the first time they are used, is there a developmental pattern? If there should be a developmental pattern in a learner, does this pattern exist in many other learners' production with the same background? What are the most commonly shared difficulties of the same group of learners at different stages of acquisition?

A last question could be: in what ways could a diachronic learner corpus be used to its best potential compared with a synchronic one? All in all, investigations into diachronic learner language via corpus linguistic means should enrich our understanding of English learning- and teaching-related areas such as SLA, psycholinguistics, teaching English as a second or foreign language, and language testing.

9.5.2 A systematic study of all POS words

The current study has looked at verbs from several different perspectives. The results are encouraging and are expected to aid English language learning and teaching considerably. Since there are other POS words such as nouns, adjective, adverbs, prepositions and conjunctions, the learner language features will be much more accessible to researchers if other POS words are studied. Only when all the POS words have been studied extensively enough could we start to make use of the learner language study results in a systematic way. Writers of teaching materials can expect to make substantial progress once the study of all the POS words is completed. The perspectives taken by this study could certainly be used as a reference, but new perspectives should be taken into consideration because different POS words have different features and what fits studies of verbs perfectly well does not necessarily fit studies of other POS vocabularies. New designs and methodologies should always be considered.

9.5.3 A study of a learner translation corpus

My current study is based on the writing of essays and compositions in examinations. Because of the variation of topics within a corpus, it is hard for corpus designers to control the content of the corpus. This disadvantage could be avoided to a very large degree if the learner corpus could be controlled in content. The best option that can be conjured up is a translation corpus in which there are translations of controlled texts. Since many writers are translating the same content at the same time, it would be much easier for the researcher or teacher to see how a certain concept in English is expressed by different individual writers. The content of the writing is always clear to the researcher and should pose no problems of interpretation. With the content fixed, variation from translator to translator is only a matter of degree.

This approach could be used wisely to have a beneficial effect on pedagogy. The text to be translated could be a length of text which has been translated from the original English language. If the learners are asked to translate the text back into English, it is possible for them to become aware of the difference between what they have written and what was written by the original writer. If the translation task could be made into standard exercises, the learners would have a better chance of learning how NSs express certain notions and meanings. This is expected to help the learners not only to write in a more NS-like way, but also to understand NSs more easily.

9.5.4 A study of learner spoken English

Since the current study involves only written production, it has little to say about the features of learner language in the aspect of spoken language. Though there is some similarity between written and spoken English, corpus studies have shown the large disparity between the two different genres; for example, Biber (1998), Biber *et al*. (1999), Carter and McCarthy (2006). To uncover the mysteries of learner language more thoroughly, it is necessary and worthwhile to compare learner spoken English to NS spoken English. Perhaps because of the difficulties in collection and transcription of spoken data, there is an unjustifiably small number of learner spoken English corpora compared with learner written English corpora. The written LOCNESS corpus has been used extensively for comparison (see Chapter Two for a detailed review of this issue), but it seems that studies based on spoken English corpora are rare. Technologies should develop in this direction so that spoken, as opposed to written, learner language can be studied in detail.

9.6 Conclusion

This chapter has reconsidered the implications of this whole research project by emphasising some important aspects arising from it. The innovative approach of this research to the field of learner language has been addressed and some of its possible applications are discussed, even though some ideas need further development. ELT practitioners may treat this research as a kind of archetype through which they may make use of modern technologies and ‘give them a try’ themselves.

Chapter Ten

Conclusion

10.1 A summary of the research

This short concluding chapter will give a brief summary of the entire thesis. Chapter One introduced the theme of the research, i.e. corpus-based learner language, as a development from other earlier language studies. Chapter Two reviewed the literature of corpus-based learner language studies and indicated the tasks of the research. Chapter Three described the data and the technology that was used for the research. The six chapters from Chapter Four to Chapter Nine reported on the explorations and investigations of the corpus-based contrastive learner language study.

In Chapter Four, two verb lemma lists were made by using Yasumasa's lemma lists, and nearly 400 verb lemmas were found to be absent from the learner corpus. Based on the verb lemmas contrasted, a sub-categorisation was made in order for the learners to associate what they currently use with what has not been used. Some functions of WordSmith Tools, MS Office and Excel and some customised programming were used in this chapter.

Following the verb lemma study in Chapter Four which dealt only with the amalgamation of verb forms, we discovered in Chapter Five that there exists a disparity between the different forms of a verb according to an observation of the 20 most often used verbs both in LOCNESS and COLEC. The differences in the linguistic disparities between the two corpora point to quite different schemata of collective language production. Whereas the NSs have a lot in common in producing the same form of a verb, the learners have very little knowledge of this kind. Chapter Five also compared the top 20 verbs in each individual form and provided a list of verbs that are found only in the NS corpus for each individual form. This chapter continued the investigation into verb forms by comparing all the verbs that occur only in the NS corpus for each individual verb form, thus ending up with a list of verbs that occur only in the NS corpus for each individual word form. The second function of the comparative learner language study approach, the evaluative function, was tentatively proposed and

discussed. Some functions of WordSmith Tools and MS Office Excel were used in the research reported in this chapter.

Since Chapter Four and Chapter Five had ignored the existence of multiple functions in POSes, Chapter Six switched the focus to those words that function both as verbs and nouns and found that an obvious preference exists in the learner language for verb function to be prioritised over noun function for most of the verbs studied. Chapter Six examined the preferences as to verb function and noun function by the two groups of writers from several perspectives, using a minimum of technological support.

Chapter Seven looked at the production of English from the perspective of patterns (in line with Hunston and Francis 1999). It was found that there is a sharp difference between the patterns used by the NSs and the learners as far as the verb *KEEP* is concerned. The NSs use a greater variety of pattern types than the learners who predominantly use a small number of pattern types. A general impression from this chapter is that we cannot assume that simple vocabulary like *KEEP* has already been fully mastered by the learners. The BoE was used in this chapter in cases where LOCNESS failed to answer a certain line of enquiry because of its restricted size. One of my reservations about the current CIA analysis, the problem with the general and vague terms 'overuse' and 'underuse' was fully addressed. Instead of sticking to these traditional terms, I have proposed that a finer classification should be used so that the diagnostic function of a comparative learner language study approach could be extensively applied.

Chapter Eight focused on the collocations of the verb *TAKE* and found that the NSs used a wider range of collocations. Though the learners use some collocations fairly frequently in the same way as the NSs, the contextual behaviour is very dissimilar. If we take as an example one of the most often used intransitive phrases, *TAKE place*, there is very little similarity between the subjects used in the two corpora. For the prepositional phrase *TAKE on*, the learners' production shows that the word is employed by the learners in a narrower range of senses. The findings of this chapter show that even very frequently used 'simple' verbs such as *TAKE* may be problematic for learners.

In the discussion chapter, Chapter Nine, the major implications of the research were re-addressed as a whole. Some ideas concerning various possible applications in pedagogy were put forward, such as in the enhancement of textbook writing, DDL-supported classroom activities and dictionary compilation. The contribution of this research to current learner language studies, i.e. the illustration and generalisation of the diagnostic function and the evaluative function of corpus-based comparative study between learner English and NS English, among other things, were discussed in detail.

10.2 Some limitations of the research

Though it has been demonstrated in the previous chapters that a corpus-based comparative approach to learner English data is a useful tool in language acquisition research and language education, there are certainly some limitations that need to be acknowledged.

First, much of the research is based on the assumption that whatever is used by the NSs is to be regarded as the norm and the target for the learners. It follows that the fewer differences there are between the NS English and the learner English, the more successful the learners can be considered to be . This is actually not necessarily true. On one hand, there exist a number of creative uses of English that do not need to be matched by the learner corpus. On the other hand, as noticed by Leech (1998) and Granger (1998b), not all uses by the NSs are suitable as targets for the learners to achieve. The appearance of the unusual verb *FLOG* and some misspellings such as *CONCIEVE* (for *CONCEIVE*) and *LOOSE* (for *LOSE*) in the NS corpus are cases in point.

Second, since CLC researchers are dependent on the data of production they are unfortunately restricted to the limited data available to them. In other words, CLC researchers can count only what can be counted and miss out what cannot be counted. It would be extremely difficult (if not entirely impossible) to investigate the areas of language use which are not represented in the corpus at all. In this sense, language acquisition research will continue to need other sources such as metalingual judgements and self-report, as used in the current SLA research. It is expected that interdisciplinary co-operation between CLC and SLA and other neighbouring areas will be able to yield more convincing research results.

Third, there exists a problem of the accuracy rate of the POS tagger. As mentioned in Chapter Four (4.3.2), the accuracy of POS tagging affects the validity of research. Even though this problem is expected to become less prominent with the improvement of POS tagging technology, researchers wishing to make use of this technology, especially on learner English data, should be cautious in interpreting research results and making corresponding claims.

Fourth, the disparity between the learner corpus and the NS corpus under comparison regarding topics and degrees of formality and other parameters of the discourse affects the result of research. The closer the comparable corpora are to each other in terms of topics and other parameters of the discourse, the more confidence CLC researchers would have. More time spent on the construction of corpora (both the learners' and the NSs') will prove to be worthwhile and rewarding.

To sum up, the value of CLC is dependent on carefully constructed and selected comparable data and therefore the significance of such a new approach should not be over-played. An interdisciplinary development might open a wider space for CLC, the newly-born branch of enquiry.

10.3 The next few years of learner corpus studies envisaged

As mentioned earlier, in Chapter Two, learner language study via comparison of corpora is growing extremely fast. In a few years' time, it is expected to branch out in many new directions. In the design and establishment of learner corpora, for example, there should be a drastic increase in size made possible by the improvements in the current computer technology. In annotating learner English, some improvement in the accuracy of POS-tagging is also expected, since annotation technology is becoming more and more mature. Complete resolution is not seen as practical for many years to come. The analysis of the features of learner language is expected to be better systemised once some initial investigations have been carried out. For example, all POSes apart from verbs might be studied so that a complete profile of learner English in the layer of POS distribution is ready for pedagogical use. Research findings are expected to be made more easily accessible to the learner, the teacher and other people concerned. New teaching materials (including digital versions) based upon the findings of comparative learner language studies will gradually appear. There may be a

period of conflict with devotees of more traditional teaching materials, due to the disparity in nature between the two approaches. Because of their vagueness, the terms 'overuse' and 'underuse' will gradually lose their popularity in this area, and give way to new terms. It is also expected that the use of this approach will be integrated with some other means such as data elicitation because corpus data do not always provide the information the researchers need.

What is more important, it is envisaged that more and more people will be convinced of the validity of this discipline and adopt it as a useful tool for their jobs, especially those researchers in the neighbouring areas such as SLA, psycholinguistics and language testing. Finally, it is to be hoped that researchers will gradually adopt L1 as their basis for research, as advocated and practised by Tono (2003), even though learners' IL and L2 will remain the dominant objects of studies.

10.4 Final remarks

A corpus-linguistic approach to learner language study is a very new branch of applied linguistics. But there is no doubt that this is a very promising area of enquiry, for the possible insights it could offer into language acquisition, language learning and teaching, and some other neighbouring branches. As I have tried to demonstrate in this thesis, a corpus-based approach to analysing learner language in comparison with NS language is a very new field of enquiry, and for that reason this may still be relatively unfamiliar to researchers, teachers, learners and writers of teaching materials. By drawing attention to its appeal in language acquisition research and ELT I hope to ensure that its merits will be increasingly recognised in the future.

List of References

Aarts J. and Granger S. 1998. 'Tag sequences in learner corpora: a key to interlanguage grammar and discourse' in S. Granger (ed.) 132-142.

Aijmer K. 2002. 'Modality in advanced Swedish learners' written interlanguage' in S. Granger *et al*. (eds.) 55-76.

Albert M. and Obler L. K. 1978. *The Bilingual Brain*. Academic Press: New York.

Alexander D. and Kunz W. J. 1964. *Some Classes of Verbs in English*. Bloomington: Indiana University Linguistics Club.

Allan Q. G. 2002. 'The TELEC secondary learner corpus: a resource for teacher development' in S. Granger *et al.* (eds.) 195-211.

Altenberg B. 1996. 'Exploring the Swedish subcorpus of ICEL'. Paper presented on the 11th Word Congress of Applied Linguistics (AILA), Jyvaskyla, Finland. August 4-9, 1996.

Altenberg B. and Granger S. 2001. 'The grammatical and lexical patterning of MAKE in native and non-native student writing' *Applied Linguistics* 22 (2): 173-195.

Altenberg B. and Tapper M. 1998. 'The use of adverbial connectors in advanced Swedish learners' written English' in S. Granger (ed.) 80-93.

Archer D., Rayson P., Wilson A. and McEnery A. (eds.) 2003. *Proceedings of the Corpus Linguistics 2003 Conference*. Lancaster University: UCREL Technical Papers 16.

Aston G. 2000. 'Corpora and language teaching' in L. Burnard and T. McEnery (eds.) *Rethinking Language Pedagogy from a Corpus Perspective: Papers from the Third International Conference on Teaching and Language Corpora*. Hamburg: Peter Lang. 7-17.

Aston G. 2002. 'Getting one's teeth into a corpus' in M. Tan (ed.) 131-143.

Aston G., Bernardini S. and Stewart D. (eds.) 2004. *Corpora and Language Learners*. Amsterdam: Benjamins.

Aston G. and Burnard L. 1998. *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Bernardini S. 2002. 'Exploring new directions for discovery learning' in B. Kettemann and G. Marko (eds.) *Teaching and learning by doing corpus linguistics. Papers from the Fourth International Conference on Teaching and Language Corpora*, *Graz 19-24 July 2000*. Amsterdam: Rodopi. 165-182.

Biber D. 1996. 'Investigating language use through corpus-based analyses of association patterns' *International Journal of Corpus Linguistics* 1: 171-197.

Biber D. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

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

Biber D., Johansson S., Leech G., Conrad S. and Finegan E. 1999. *Longman Grammar of Spoken and Written English.* London: Longman.

Biber D. and Reppen R. 1998. 'Comparing native and learner perspectives on English grammar: a study of complement clause' in S. Granger (ed.) 145-158.

Biskup D. 1992. 'L1 influence on learners' renderings of English collocations' in P. J. L. Arnaud and H. Bejoint (eds.) *Vocabulary and Applied Linguistics.* London: Macmillan. 85-93.

Blappert T. 1998. 'Data-driven learning: theory and classroom implementation' in *Proceedings of the 1997 Korea TESOL Conference*. January 1998.

Bonelli T. 2001. *Corpus Linguistics at work*. Amsterdam: John Benjamins.

Bonelli T. 2004. 'Working with corpora: issues and insights' in C. Coffin, A. Hewings and K. O'Hammoran (eds.) *Applying English Grammar. Functional and Corpus Approaches*. Longdon and Milton Keynes: The Open University and Arnold Publishers. 11-25.

Bridgeman L. I., Dillinger D., Higgins C., Seaman P.D. and Shank F.A. 1965. *More Classes of Verbs in English*. Bloomington: Indiana University Linguistics Club.

Burnard L. and Dodd T. 2003. 'Xara: an XML aware tool for corpus searching' in D. Archer *et al.* (eds.) 142-144.

Carter R. and McCarthy M. 2006. *The Cambridge Grammar of English: A Comprehensive Guide to Spoken and Written Grammar and Usage*. Cambridge: Cambridge University Press.

Carter R. and McCarthy M. (eds.) 1988. *Vocabulary and Language Teaching*. London: Longman.

Chafe W. L. 1982. 'Integration and involvement in speaking, writing and oral literature' in D. Tannen (ed.) *Spoken and Written Language: Exploring Orality and Literacy*. Ablex: Norwood. 35-53.

Chafe W. L. and Danielewicz J. 1987. 'Properties of spoken and written language' in R. Horowitz and S. J. Samuels (eds.) *Comprehending Oral and Written Language*. San Diego:

Academic Press. 83-113.

Channell J. 1981. 'Applying semantic theory to vocabulary teaching' *English Language Teaching Journal* 35 (2): 115-122.

Channell J. 1988. 'Psycholinguistic considerations in the study of L2 vocabulary acquisition' in R. Carter and M. McCarthy (eds.) 83-96.

Chen W. 2002. 'Acquisition of English passive voice by Chinese learners: a corpus based approach' *Foreign Language Teaching and Research* 34 (3): 198-202.

Cobb T. 2003. 'Analyzing late interlanguage with learner corpora: Quebec replications of three European studies' *Canadian Modern Language Review* 59 (3): 393-423.

Cobb T. and Horst M. 2001. 'Reading academic English: carrying learners across the lexical threshold' in J. Flowerdew and M. Peacock (eds.) *Research Perspectives on English for Academic Purposes.* Cambridge: Cambridge University Press. 315-329.

Corder S. P. 1967. 'The significance of learners' errors' *International Review of Applied Linguistics* Vol. V/4. Heidelberg: Julius Groos Verlag. Reprinted in J.C. Richards (ed.). 19-27.

Corder S. P. 1971. 'Idiosyncratic dialects and error analysis' *International Review of Applied Linguistics* Vol. IX/2. Reprinted in J.C. Richards (ed.)158-171.

Corder S. P. 1983. 'Strategies of communication' in C. Faerch and G. Kasper (eds.) *Strategies in Interlanguage Communication*. Harlow: Longman.

Crystal D. and Davy D. 1975. *Advanced Conversational English*. London: Longman.

Cui Y. and Huang R. 2003. 'A corpus-based study of Chinese EFL learners' acquisition of derivational affixes'. Paper read at the International Conference on Corpus Linguistics, Shanghai. October 24-26, 2003.

Davies C. H. 2004. *A Corpus-based Investigation of Noun to Verb Conversion in English*. Unpublished PhD dissertation. The University of Liverpool.

De Cock S., Granger S., Leech G., and McEnery T. 1998. 'An automated approach to the phrasicon of EFL learners' in S. Granger (ed.) 67-79.

Dodd B. 1997. 'Exploring a corpus of written German for advanced language learning' in A. Wichmann *et al.* (eds.) 131-145.

Ellis R. 1994. *The Study of Second Language Acquisition*. Oxford: Oxford University Press.

Firth J. R. 1957. *Papers in Linguistics*. London: Oxford University Press.

Fletcher W. H. 2004. 'Facilitating the compilation and dissemination of ad-hoc web corpora'

in G. Aston *et al.* (eds.) 271-300.

Flowerdew L. 2001. 'The exploitation of small learner corpora in EAP materials design' in M. Ghadessy *et al*. (eds.) 123-132.

Francis G., Hunston S. and Manning E. 1996. *Collins COBUILD Grammar Patterns 1: Verbs.* London: HarperCollins.

Francis N. and Kucera H. 1982. *Frequency Analysis of English Usage: Lexicon and Grammar*. Boston: Houghton Mifflin.

Gan S., Low, F. and Yaabkub N. 1996. 'Modeling teaching with a computer-based concordancer in a TESL preservice teacher education program' *Journal of Computing in Teaching Education* 12: 28-32.

Garside R., Leech G. and McEnery A. (eds.) 1997. *Corpus annotation: linguistic information from computer text corpora*. London: Longman.

George H. 1972. *Common Errors in Language Learning: Insights from English*. Rowley, Mass.: Newbury House.

Ghadessy M., Henry A. and Roseberry R. (eds.) 2001. *Small Corpus Studies and ELT*. Amsterdam: Benjamins.

Gillard P. and Gadsby A. 1998. 'Using a learners' corpus in compiling ELT dictionaries' in S. Granger (ed.) 159-171.

Godman A. 1982. 'Teaching verbs using a hierarchical system' *RELC Journal* 13 (1): 37-49.

Granger S. 1997. 'Automated retrieval of passives from native and learner corpora: precision and recall' *Journal of English Linguistics* 25(4): 365-374.

Granger S. 1998a. 'Introduction' in S. Granger (ed.) xxi-xxii.

Granger S. 1998b. 'The computer learner corpus: a versatile new source of data for SLA research' in S. Granger (ed.) 3-18.

Granger S. (ed.) 1998. *Learner English on Computer*. London: Longman.

Granger S. 2002. 'A bird's-eye view of learner corpus research' in S. Granger *et al*. (eds.) 3-33.

Granger S., Hung J. and Petch-Tyson S. (eds.) 2002. *Computer Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: Benjamins.

Granger S. and Rayson P. 1998. 'Automatic profiling of learner texts' in S. Granger (ed.) 119-131.

Granger S. and Tribble C. 1998. 'Learner corpus data in the foreign language classroom:

form-focused instruction and data-driven learning' in S. Granger (ed.) 199-209.

Gui S. and Yang H. 2002. *Chinese Learner English Corpus.* Shanghai: Shanghai Foreign Languages Education Press.

Guo X. 2003. 'Between verbs and nouns and between the base form and the other forms of verbs – A contrastive study into COLEC and LOCNES' in D. Archer *et al.* (eds.) 274-281.

Guo X. (forthcoming) 'Errors or partial acquisition: a case study of a young English learner's interlanguage' in E. Hidalgo, L. Quereda and J. Santana (eds.) *Corpora in the Foreign Language Classroom*. Amsterdam: Rodopi.

Hadley G. 2002. 'Sensing the winds of change: an introduction to data-driven learning' *RELC Journal* 33 (2): 99-124.

Harley B. 1986. *Age in Second Language Acquisition*. Clevedon: Multilingual Matters.

Harvey P. D. 1983. 'Vocabulary learning: the use of grids' *English Language Teaching Journal* 37 (3): 243-246.

Hasselgren A. 1994. 'Lexical teddy bears and advanced learners: a study into the ways Norwegian students cope with English vocabulary' *International Journal of Applied Linguistics* 4 (2): 237-260.

Hasselgren A. 2002. 'Learner corpora and language testing: small words as markers of learner fluency' in S. Granger *et al* (eds.) 143-173.

Hatch E. and Brown C. 2001. *Vocabulary, Semantics and Language Education*. Beijing: Foreign Language Teaching and Research Press.

Hawkins, E. 1984. *Awareness of Language: An Introduction*. Cambridge: Cambridge University Press.

Hoey M. 2004. 'The textual priming of lexis' in G. Aston *et al.* (eds.) 21-41.

Horváth J. 1999. *Advanced Writing in English as a Foreign Languag: A Corpus-based Study of Process and Products*. Unpublished Ph.D. dissertation. Available at http://geocities.com/writing_site/thesis/. Accessed on September 22, 2005.

Horváth J. 2001. *Advanced Writing in English as a Foreign Language: A Corpus-based Study of Process and Products*. Pécs: Lingua Franca Csoport.

Housen A. 2002. 'A corpus-based study of the L2-acquisition of the English verb system' in S. Ganger *et al.* (eds.) 77-116.

Howarth P. A. 1996. *Phraseology in English Academic Writing: Some Implications for Language Learning and Dictionary Making*. Tübingen: Niemeyer.

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

Hunston S. and Francis G. 1998. 'Verbs observed: a corpus-driven pedagogic grammar' *Applied Linguistics* 19 (1): 45-72.

Hunston S. and Francis G. 1999. *Pattern Grammar. A Corpus-Driven Approach to the Lexical Grammar of English*. Amsterdam: Benjamins.

Johansson S. 1985. 'Word frequency and text type: some observations based on the LOB corpus of British texts' *Computers and the Humanities* 19: 23-36.

Johns T. 1988. 'Whence and whither classroom concordancing?' in T. Bongaerts, P. de Haan, S. Lobbe, and H. Wekker (eds.) *Computer Applications in Language Learning*. Dordrecht, The Netherlands: Foris. 9-27.

Johns T. 1991. '"Should you be persuaded": two examples of data-driven learning aterials'in T. Johns and P. King (eds.) *Classroom Concordancing*. Birmingham University: English Language Research Journal 4. 1-16.

Johns T. 1994. 'From printout to handout: grammar and vocabulary teaching in the context of data-driven learning' in T. Odlin (ed.) *Perspectives on Pedagogical Grammar*. Cambridge: Cambridge University Press. 293-313.

Johns T. 1997. 'Contexts: the background, development and trialling of a concordance based CALL program' in A. Wichmann *et al.* (eds.) 100-115.

Johns T. 2002. 'Data-driven learning: the perpetual challenge' in B. Kettemann and G. Marko (eds.) *Teaching and Learning by Doing Corpus Analysis. Proceedings of the Fourth International Conference on Teaching and Language Corpora. Graz 19-24 July, 2000.* Amsterdam: Rodopi, 107-117.

Kaszubski P. 1998a. 'Learner corpora: the cross-roads of linguistic norm' in *TALC98 Proceedings.* Keble College, Oxford, July, 1998. Also available at: http://users.ox.ac.uk/~talc98/kaszubski.htm. Accessed on May 10, 2001.

Kaszubski P. 1998b. 'Enhancing a writing textbook: a national perspective' in S. Granger (ed.) 172-185.

Kennedy C. and Miceli T. 2001. 'An evaluation of intermediate students' approaches to corpus investigation' *Language Learning and Technology* 5: 77-90.

Kennedy G. 1998. *An Introduction to Corpus Linguistics*. London: Longman.

Kettemann B. and Marko G. (eds.) 2002. *Teaching and Learning by Doing Corpus Linguistics. Papers from the Fourth International Conference on Teaching and Language Corpora,*

Graz 19-24 July 2000. Amsterdam: Rodopi.

Kilgarriff A. 1997. 'Putting frequencies in the dictionary' *International Journal of Lexicography* 10 (2): 135-155.

Kjellmer G. 1991. 'A mint of phrases' in K. Aijmer and B. Altenberg (eds.) *English Corpus Linguistics*. London: Longman. 111-127.

Knowles G. and Zuraidah M.D. 2004. 'The notion of a "lemma": headwords, roots and lexical sets' *International Journal of Corpus Linguistics.* 9 (1): 69-81.

Krishnamurthy R. (ed.) 2004. *English Collocation Studies: The OSTI Report (John Sinclair, Susan Jones and Robert Daley)*. London and New York: Continuum.

Larsen-Freeman D. and Long M. 1991. *An Introduction to Second Language Acquisition Research.* London: Longman.

Leech G. 1974. *Semantics*. Harmondsworth: Penguin.

Leech G. 1997. 'Introducing corpus annotation' in R. Garside *et al.* (eds.) 1-18.

Leech G. 1998. 'Preface' in S. Granger (ed.) xiv-xx.

Leech G. 2001. 'The role of frequency in ELT: new corpus evidence brings a re-appraisal' *Foreign Language Teaching and Research* 33 (5): 328-339.

Leech G., Rayson P. and Wilson A. 2001. *Word Frequencies in Written and Spoken English: Based on the British National Corpus*. London: Longman.

Lei X. 2003. 'Problems in Chinese learners' use of the English existential sentences'. Paper read at the International Conference on Corpus Linguistics, Shanghai. October 24-26, 2003.

Lenko-Szymanska A. 2002. 'How to trace the growth in learners' active vocabulary? A corpus-based study' in B. Kettemann and G. Marko (eds.) 217-230.

Levin B. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago and London: The University of Chicago Press.

Li W. 1999. *An Analysis of the Lexical Words and Words Combinations in the College Learner English Corpus*. Unpublished PhD dissertation. Shanghai Jiao Tong University.

Li W. 2003. 'A CLEC-based analysis of key words and associates' *Modern Foreign Languages* 26 (3): 284-293.

Lin L. H. F. 2002. 'Overuse, underuse and misuse: using concordancing to analyse the use of *It* in the writing of Chinese learners of English' in M. Tan. (ed.) 63-76.

Long M. H. 2003. 'Stabilization and fossilization' in C. J. Doughty and M. H. Long (eds.) *The*

Handbook of Second Language Acquisition. Oxford: Blackwell. 487-535

Lorenz G. 1998. 'Overstatement in advanced learners' writing: stylistic aspects of adjective intensification' in S. Granger (ed.) 41- 52.

Lorenz G. 1999. *Adjective Intensification - Learners versus Native Speakers: A Corpus Study of Argumentative Writing*. Amsterdam: Rodopi.

Lu Y. 2002. 'Linguistic characteristics in Chinese learner English' in M. Tan. (ed.) 49-60.

McCarthy M. 1990. *Vocabulary*. Oxford: Oxford University Press.

McEnery T. and Wilson A. 1996. *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Meara P. 1982. 'Word associations in a foreign language: a report on the Birkbeck vocabulary project' *Nottingham Linguistic Circular*. 11 (2): 29-38.

Meunier F. 1998. 'Computer tools for the analysis of learner corpora' in S. Granger (ed.) 19-37.

Milton J. 1998. 'Exploiting L1 and interlanguage corpora in the design of an electronic language learning and production environment' in S. Granger (ed.) 186-198.

Milton J. and Hyland K. 1999. 'Assertions in students' academic essays: a comparison of English NS and NNS student writers' in R. Berry, B. Asker, K. Hyland (eds.), *Language Analysis, Description and Pedagogy*. Hong Kong: Language Centre HKUST. 147-161.

Moon R. 1998. *Fixed Expressions and Idioms in English*. Oxford: Clarendon Press.

Myles F. 2005. 'Interlanguage corpora and second language acquisition research' in *Second Language Research*. 21 (4): 373-391.

Nemser W. 1971. 'Approximate systems of foreign language learners' *International Review of Applied Linguistics* Heidelberg: Julius Groos Verlag. Vo. X/3. Reprinted in J.C. Richards (ed.) 55-63.

Nesselhauf N. 2004. 'Learner corpora and their potential for language teaching' in J. Sinclair (ed.) 125-152.

Nesselhauf N. 2005. *Collocations in a Learner Corpus*. Amsterdam: Benjamins.

Oshita H. 2000. '*What is happened* may not be what appears to be happening: a corpus study of 'passive' unaccusatives in L2 English' *Second Language Research* 16 (4): 293-324.

Palmer F. R. 1975. *The English Verb*. London: Longman.

Petch-Tyson S. 1998. 'Writer/reader visibility in EFL written discourse' in S. Granger (ed.) 107-118.

Poole M. and Field T. 1976. 'A comparison of oral and written code elaboration' *Language*

and Speech. 19 (4): 305-312.

Pravec N. A. 2002. 'Survey of learner corpora' *ICAME Journal*. 26: 81-114.

Pu J. 2000a. 'A survey of Chinese learners' use of English verbs in grammatical and lexical patterns' *Modern Foreign Languages* 23 (1): 24-44.

Pu J. 2000b. *Learner Behaviour of Verbs: A Corpus-based Research on Chinese College Students' Use of English Verbs*. Unpublished PhD dissertation. Shanghai Jiaotong University.

Quirk R., Greenbaum S., Leech G. and Svartvik J. 1972. *A Grammar of Contemporary English*. London: Longman.

Ragan P. H. 2001. 'Classroom use of a systemic functional small learner corpus' in M. Ghadessy *et al.* (eds.) 207-236.

Raupach M. 1984. 'Formulae in second language speech production' in H. W. Dechert, D. Mohle and M. Raupach (eds.) *Second Language Productions*. Tubingen, Germany: Gunter Narr. 114-137.

Renouf A. and Sinclair J. 1991. 'Collocational frameworks in English' in K. Aijmer and B. Altenberg (eds.) *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. London: Longman. 128-144.

Richards J. C. (ed.) 1974. *Error Analysis.* London: Longman.

Ringbom H. 1998a. 'High frequency verbs in the ICLE corpus' in A. Renouf (ed.) *Explorations in Corpus Linguistics*. Amsterdam: Rodopi. 191-200.

Ringbom H. 1998b. 'Vocabulary frequencies in advanced learner English: a cross-linguistic approach' in S. Granger (ed.) 41-52.

Rudzka B., Channell J., Putseys Y. and Ostyn P. 1981. *The Words You Need*. London: Macmillan.

Rudzka B., Channell J., Putseys Y. and Ostyn P. 1985. *More Words You Need*. London: Macmillan.

Rundell M. and Ham N. 1994. 'A new conceptual map of English' in W. Martin *et al*. (eds.) *EURALEX 94 Proceedings*. Amsterdam. 172-180.

Schachter J. and Celce-Murcia M. 1977. 'Some reservations concerning error analysis' in *TESOL Quarterly*. 11 (4): 441-451.

Scott M. 1997. 'PC analysis of key words – and key key words' *System* 25: 1-13.

Scott M. 1999. *WordSmith Tools (Version 3.0)*. Oxford: Oxford University Press.

Scott M. 2004. *WordSmith Tools (Version 4.0)*. Oxford: Oxford University Press.

Scott M. and Tribble C. 2006. *Textual Patterns: Keyword and Corpus Analysis in Language Education*. Amsterdam: Benjamins.

Seidlhofer B. 2002. 'Pedagogy and local learner corpora' in S. Granger *et al.* (eds.) 214-234.

Selinker L. 1972. 'Interlanguage' *International Review of Applied Linguistics* Heidelberg: Julius Groos Verlag. Vo. X/3. Reprinted in J.C. Richards (ed.) 31-54.

Sinclair J. (ed.) 1987. *Looking Up: An Account of the COBUILD Project*. London: HarperCollins.

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

Sinclair J. (ed.) 2004. *How to Use Corpora in Language Teaching*. Amsterdam: Benjamins.

Sinclair J. and Renouf A. 1988. 'A lexical syllabus for language learning' in R. Carter and M. McCarthy (eds.) 140-160.

Sripicharn P. 2002. *Evaluating Data-driven Learning: The Use of Classroom Concordancing by Thai Learners of English*. Unpublished PhD dissertation. The University of Birmingham.

Stieglitz E.L. 1983. 'A practical approach to vocabulary reinforcement' *English Language Teaching Journal*. 37 (1): 71-75.

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

Svartvik J. and Ekedahl O. 1995. 'Verbs in public and private speaking' in B. Aarts. and C. Meyer (eds.) *The Verb in Contemporary English*. Cambridge: Cambridge University Press. 273-289.

Tan M. (ed.) 2002. *Corpus Studies in Language Education*. Bangkok: IELE Press.

Tarone E. and Yule G. 1989. *Focus on the Language Learner*. Oxford: Oxford University Press.

Ting Y. and Wen Q. 2003. 'The effect of the command of formulaic sequence on oral English performance' Paper read at the International Conference on Corpus Linguistics, Shanghai. October 24-26, 2003.

Tono Y. 2003. 'Learner corpora: design, development and applications' in D. Archer *et al.* (eds.) 800-809.

Ullman S. 1967. *Semantics*. Oxford: Oxford University Press.

van Rooy B. and Schäfer L. 2003. 'An evaluation of three POS taggers for the tagging of the

Tswana Learner English Corpus' in D. Archer *et al.* (eds.) 835-844.

Virtanen T. 1998. 'Direct questions in argumentative student writing' in S. Granger (ed.) 94-106.

Wen Q., Ding Y. and Wang W. 2003. 'Features of oral style in English compositions of advanced Chinese EFL learners: an exploratory study by contrastive learner corpus analysis' *Foreign Language Teaching and Research* 35 (4): 268-274.

West M. 1953. *A General Service List of English Words*. London: Longman, Green and Co.

White L. 2002. 'Morphological variability in endstate L2 grammars: the question of L1 influence' in B. Skarabela, S. Fish and A. H. J. Do (eds.) *Proceedings of the 26th Annual Boston University Conference on Language Development*. Somerville, MA: Cascadilla Press. 758-768.

Wichmann A., Fligelstone S., McEnery T. and Knowles G. (eds.) 1997. *Teaching and Language Corpora*. London: Longman.

Willis D. 1990. *The Lexical Syllabus: a new approach to language teaching*. London: HarperCollins.

Wu S. 1995. *Transfer in Chinese Students' Academic Writing*. Unpublished PhD dissertation. Northern Arizona University.

Wyatt H. 1923. *The Teaching of English in India*. London: Oxford University Press. Reprinted in R. Smith (ed.) 2003. *Teaching English as a Foreign Language: 1912-36. Pioneers of ELT. Volume 1 Wren and Wyatt*. London: Routledge.

Yang C. and Ning C. 2002. 'A SLA model of multi-union of native language, target language and interlanguage' *The Journal of Hunan University (Social Sciences)*. 16 (4): 71-75.

Yang H. 2001. 'Computer analysis of Chinese learner English'. Keynote speech at the conference Technology in Language Education. Co-organised by Hong Kong University of Science and Technology and Nanjing University, June 26-30, 2001. Also available: http://lc.ust.hk/~centre/conf2001/keynote/index.html, accessed on January 1, 2006.

Appendix I: Working out a verb lemma list base

1.1 Opening Someya's lemma list

The following steps are, roughly, those that I took in editing Someya's word list. Since Someya's lemma list is in 'txt' form, Excel can be used to convert it into an Excel file as follows:

1) Open a blank Excel page, and click *Open* in the *File* menu, and then choose Someya's lemma list (e-lemma.txt) and click *Open*. Then Excel will prompt a window as below (see Figure App. 1.1).
2) Click *Next* and check *Space* and *Comma* in *Delimiters* (see Figure App.1.2) and then click *Next*.
3) When Excel prompts a screen below (see Figure App. 1.3), click *Finish* when the following screenshot appears.

1.2 Editing the list

After the lemma list has been opened as demonstrated above, it is ready for further editing. These, roughly, are the steps I took in the first phase of editing:

1) Delete the introductory lines (Lines 1-24).
2) Sort out the columns by F, E and D in descending order (as the screenshot shows, see Figure App. 1.4) and click *OK*.
3) Delete the rows that have only words from Column A to Column C (from 5667 to the end).
4) Save the file as a new file, say 'lemma_edited.exl'.

At the end of this phase, the long list has been trimmed to a more manageable length and most of the noun lemmas are deleted from the list, but those nouns with two plural forms are still mixed with verbs. Adjectives with comparative and superlative forms are also in the list. These words need to be sieved out. I took the following steps in this phase of editing:

1) Cut and paste the rows that have contents from Column A, to Column D (Lines 4999-5666) to a blank 'txt' processor such as Wordpad, to remove the border lines of the

table.

2) Then copy the borderless content to a blank MS Word file and then go through the following steps to get rid of the adjectives first.
3) Since the adjective lemmas all end with the superlative degree form 'est', if a new column could be created to make the rows ending with 'est' stand out from the rest of the rows, the adjectives could be selected out from the list. Use the Find and Replace function of MS Word to add a new column, as Figure App. 1.5 shows (*Find what* = est *Replace with* = est, new). The word after the comma (new) can be any word (see Figure App. 1.5). Use 'Save as' in the File menu to save this as a 'txt' file, say 'lemma_tail'. Click 'OK' in the new window with Windows (Default) checked.
4) Open the saved file 'lemma_tail' in Excel by clicking the 'Read Only' button first, and then 'Next' button, and then the 'Next' button again with 'Tab', 'Space' and 'Comma' checked in the 'Delimiters' and finally the 'Finish' button. Then sort out by Columns E, D and C in descending order to get a new column which has the identical content which is 'new'.
5) After a new column has been created, all the words that end with 'est' have one more column than the rest of the rows of the chopped list (from 4999 to 5666). Use the Sort function of Excel to remove all the rows that have contents only from Column A to Column E (from Line 1 to Line 514). Save the file.
6) There are now 154 rows left in the file 'lemma_tail' which are a mixture of some irregular verbs such as *PUT* and *CUT* which have only three forms, and the nouns which have two plural forms. To directly delete the nouns would solve the whole problem but the nouns are mixed with verbs. If the verbs (irregular, with three forms only) can be made to stand out, the problems will be solved. Therefore, I copied the method as described in the previous step above to make the verbs ending with 'ing' stand out from the rest of the rows.
7) Use the Sort function again and select all the verbs ending with 'ing' by looking at the new column created (the first two cases with 'ing' are not the verbs I need, so they are deleted).
8) Cut all the rows that have contents in Columns A, B, C, D and E and copy them to a new page so that the added column with *new* can be deleted. Select all the lines and copy them to the end of the previously saved file 'lemma_edited.exl'. Save the file.

9) The list in the file of 'lemma_tail' is now a mixture of three types; one type is nouns with two plural forms, a second type is incomplete noun to verb conversions as shown in Table App. 1.1:

Table App. 1. 1 A sample of one type of combinations in Someya's lemma list

twin	->	twins	twinned
skill	->	skills	skilled
awe	->	awes	awed

Since the V-ed form of this type is the past participle and functions as an adjective, this type of mixture was not included in the verb lemma lists. The third type is a small number of irregular verbs (three) as in Table App. 1.2.

Table App. 1. 2 Three irregular verbs in Someya's lemma list

meet	->	meets	met
misunderstand	->	misunderstands	misunderstood
understand	->	understands	understood

Since the V-ing forms of these three verbs *MEET*, *MISUNDERSTAND*, and *UNDERSTAND* function both as verb and noun, they are singled out as separate lemmas in Someya's list (see Table App. 1.3). In other words, the word form *meeting* is missing in Table App. 1.2 because it was (unfortunately) grouped together with *meetings* as a pair of nouns. The same is true for *MISUNDERSTAND* and *UNDERSTAND* (see Table App. 1.3).

Table App. 1. 3 The singular and plural form of three pairs of nouns

meeting -> meetings
misunderstanding -> misunderstandings
understanding -> understandings

Therefore, in order for the frequency of word forms to be calculated accurately, the V-ing forms should be inserted into the verb group as in Table App. 1.4.

Table App. 1.4 The arrangement of three verb lemmas after editing

meet	->	meets	meeting	met
misunderstand	->	misunderstands	misunderstanding	misunderstood
understand	->	understands	understanding	understood

10) Copy the three lemmas in Table App. 1.4 and paste them to the end of the saved file 'lemma_edited.exl'.
11) Some manual deletion is needed at this stage because some word forms of different

POSes were arranged in Someya's list.

12) By now, the irregular tail part of Someya' lemma list has been treated; not, however, the irregular head part.

13) Open the file 'lemma_edited.exl' and sort out by the order of Columns H, G. F. There are 276 lines in the irregular head part as shown in Table App. 1.5.

Table App. 1. 5 A sample of the partially sorted lemma list

damp	->	damper	dampest	damps	damping	damped
cross	->	crosses	crossing	crossed	crosser	crossest
cool	->	cooler	coolest	cools	cooling	cooled
close	->	closes	closing	closed	closer	closest
clear	->	clearer	clearest	clears	clearing	cleared
clean	->	cleaner	cleanest	cleans	cleaning	cleaned

14) Since word forms are arranged orderlessly in the original list and therefore a manual reshuffle was carried out so that some can be deleted and some can be re-arranged. Afterwards, copy the remaining lines to the end of the file 'lemma_edited.exl' to complete the list.

15) Since there are some verbs with only three forms, i.e. the base form is identical in form to the past form and the participle form like *UPSET*, *SPREAD*, some manual work needs to be done to copy the base form of these verbs to the positions of the past form and the past participle form.

16) Delet the 'have' and 'be' lines because they are not the concern of the research.

17) Save the list as a 'txt' file.

18) There should be 5190 verbs in their different forms now.[39]

In this way, Someya's list has been converted to a verb lemma list which contains 5190 verbs in their different forms. This list could be used as a base for the consequent verb lemmatisation.

39 It should be admitted here that the few verbs with two sets of past forms and past participles (such as LEARN) are treated as if they had only one set like the majority of verbs. This could be improved in future studies.

Figure App. 1. 1 A screenshot of the first step of opening a text file in MS Excel

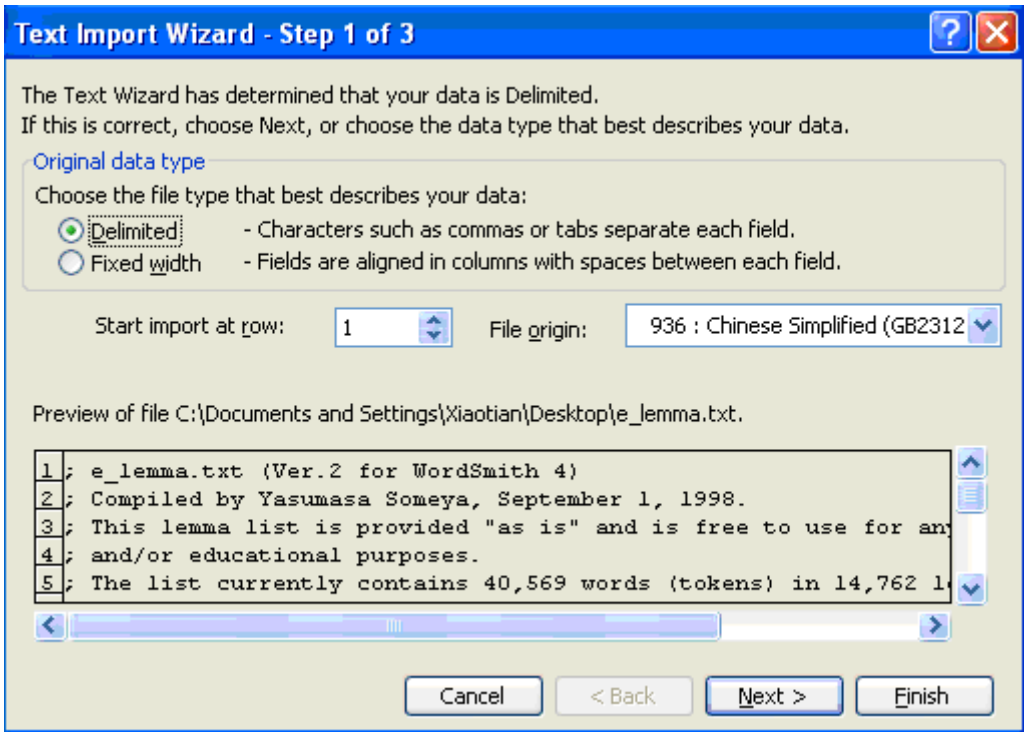


Figure App. 1. 2 A screenshot of the second step of opening a text file in MS Excel

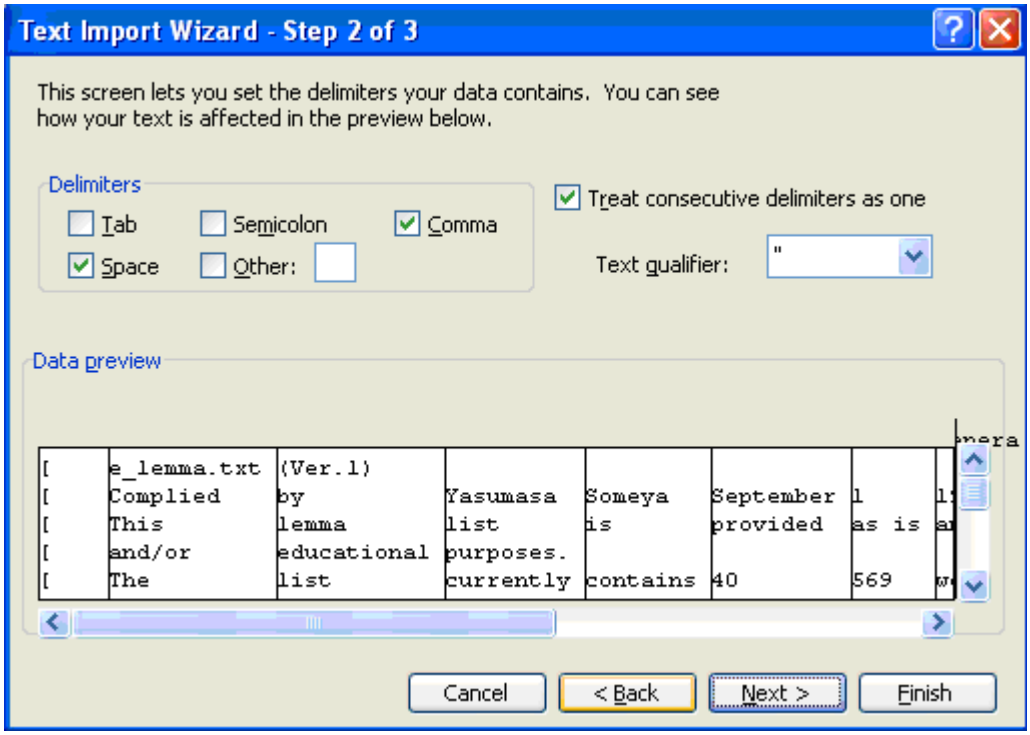


Figure App. 1. 3 A screenshot of the third step of opening a text file in MS Excel

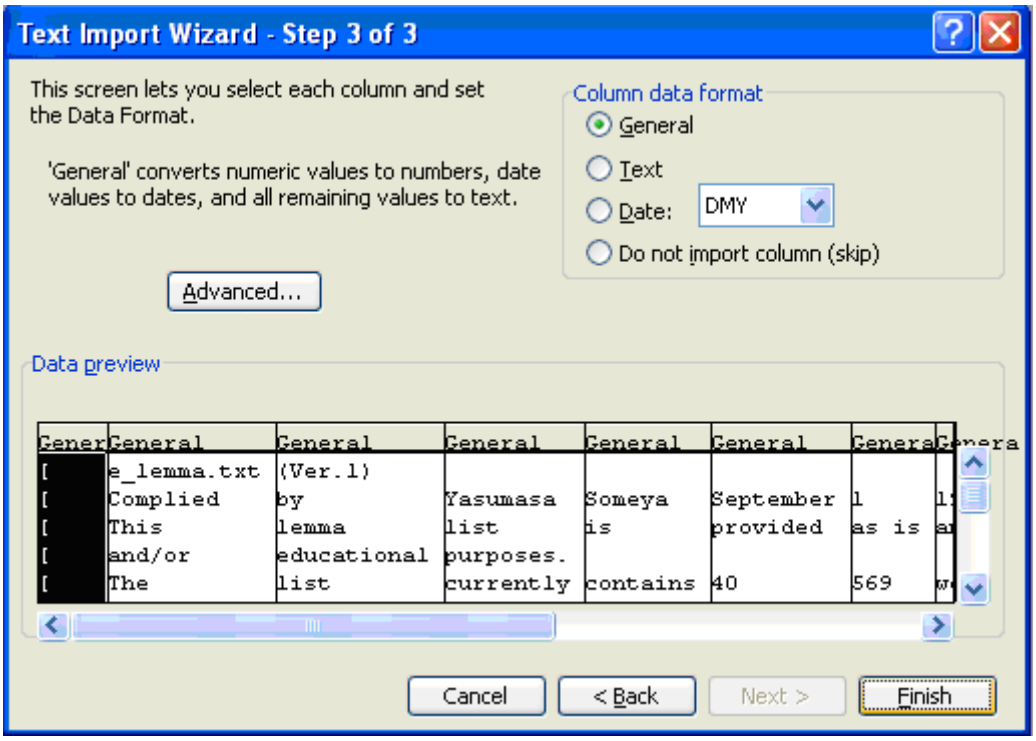


Figure App. 1. 4 A screenshot of sorting a lemma list in different priorities in MS Excel

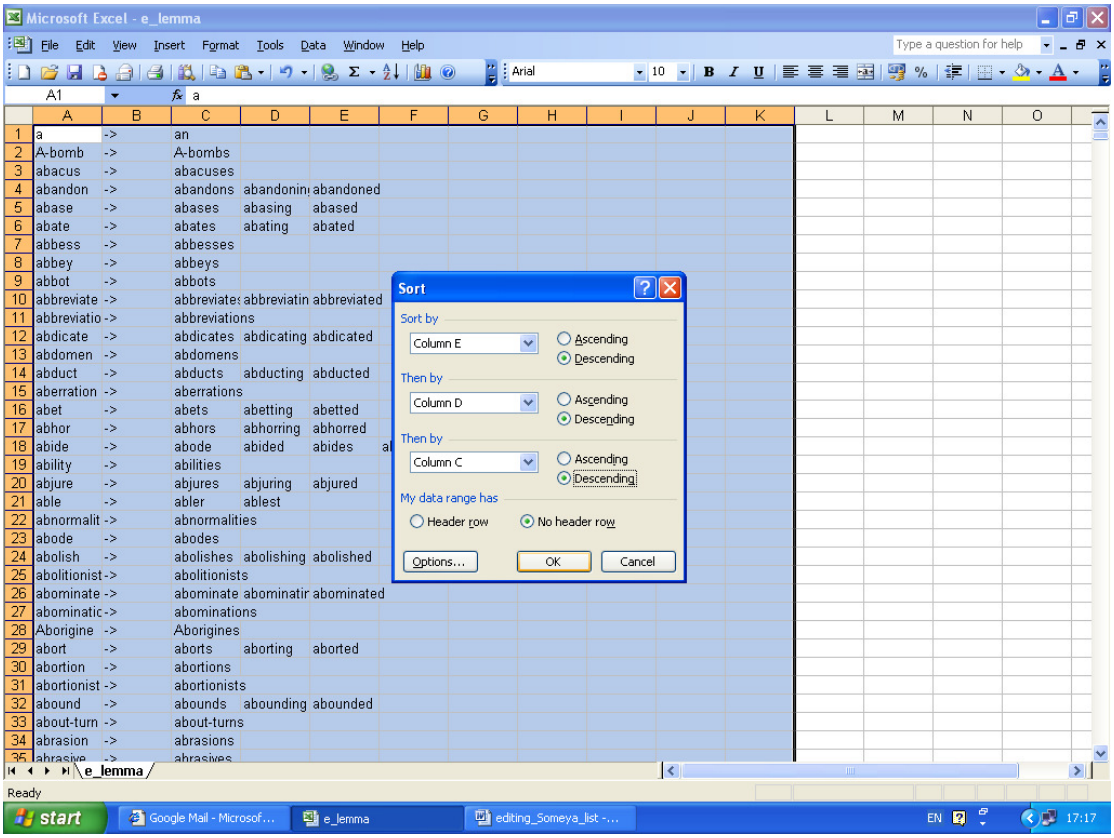
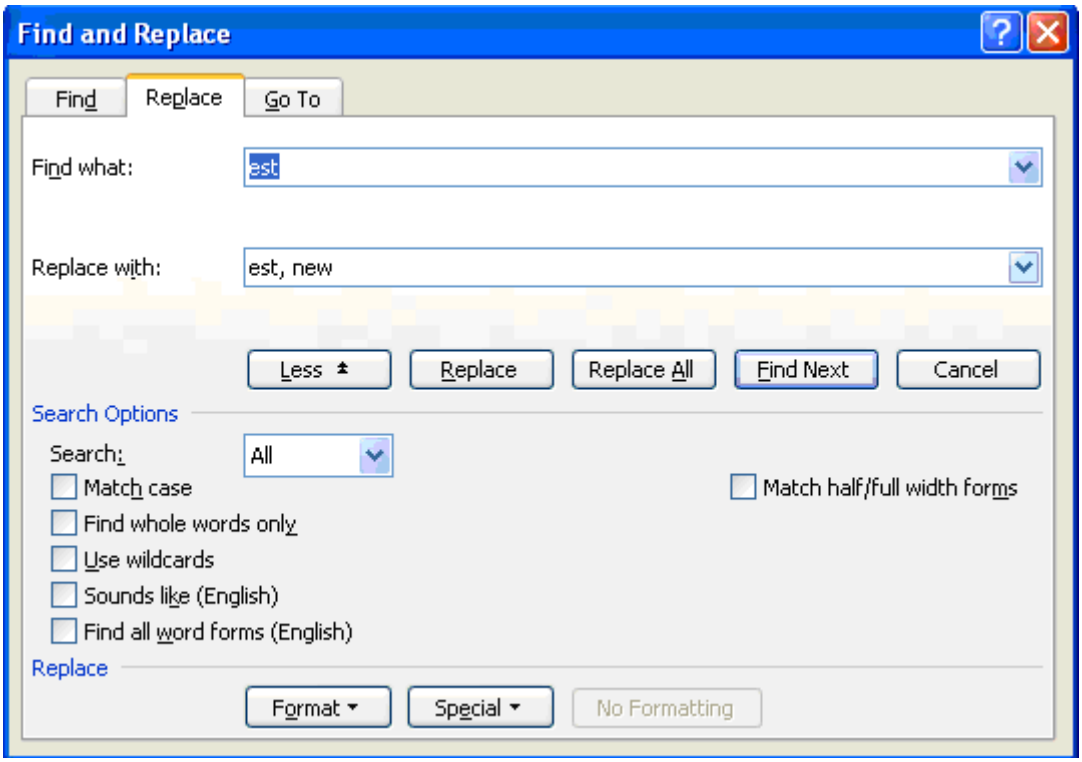


Figure App. 1. 5 A screenshot of using the Find and Replace function of MS Word



Appendix 2: A verb lemma list of COLEC

	Lemma	V-e	V-s	V-ing	V-ed	V-n	Total
1	make	1877	1623	95	112	149	3856
2	know	2565	46	79	33	136	2859
3	get	1821	34	242	125	94	2316
4	think	1940	33	39	80	40	2132
5	learn	1262	9	194	61	97	1623
6	want	1154	93	4	90	8	1349
7	use	769	8	120	31	414	1342
8	take	900	51	100	62	118	1231
9	find	869	9	28	97	51	1054
10	change	712	21	146	21	108	1008
11	go	676	62	88	111	25	962
12	like	841	78	1	0	0	920
13	study	569	2	241	29	19	860
14	work	615	30	147	13	14	819
15	read	496	3	262	28	26	815
16	say	381	61	41	171	64	718
17	become	323	77	66	77	63	606
18	need	365	127	1	14	44	551
19	improve	315	7	39	14	155	530
20	see	397	1	25	58	49	530
21	increase	74	37	153	43	159	466
22	buy	356	9	26	41	29	461
23	try	395	12	23	27	4	461
24	live	328	10	64	38	2	442
25	give	297	31	10	24	37	399
26	keep	345	14	9	14	8	390
27	bring	289	34	4	11	26	364
28	mean	54	284	9	4	0	351
29	develop	132	20	114	15	66	347
30	waste	268	16	27	11	23	345
31	understand	318	3	10	8	5	344
32	help	291	28	17	6	1	343
33	play	179	13	127	5	8	332
34	come	170	56	0	69	36	331
35	feel	267	7	9	41	4	328
36	remember	300	1	11	7	8	327
37	look	154	8	50	68	21	301
38	practise	233	4	43	4	11	295
39	believe	268	6	1	16	3	294
40	write	229	0	42	5	16	292
41	finish	221	1	11	30	27	290
42	speak	161	3	114	3	6	287
43	tell	151	41	9	72	13	286

44	cause	138	40	7	46	38	269
45	listen	160	1	91	4	4	260
46	save	198	5	29	2	9	243
47	spend	172	5	16	27	17	237
48	pay	173	4	7	14	26	224
49	face	113	4	69	3	32	221
50	produce	136	11	38	4	32	221
51	watch	99	0	110	3	1	213
52	lose	118	3	17	50	23	211
53	master	182	0	7	5	15	209
54	pollute	27	4	25	10	137	203
55	put	137	4	3	27	32	203
56	decrease	45	10	31	32	79	197
57	realize	137	1	8	22	28	196
58	succeed	185	2	1	2	4	194
59	meet	157	4	8	10	12	191
60	lead	120	32	8	12	16	188
61	run	82	5	34	30	32	183
62	begin	108	8	10	50	4	180
63	serve	121	11	42	0	2	176
64	happen	69	13	18	45	30	175
65	solve	136	1	9	1	28	175
66	ask	75	5	9	68	12	169
67	gain	133	4	2	9	20	168
68	prevent	149	5	6	1	5	166
69	protect	147	0	6	1	3	157
70	let	144	5	0	5	2	156
71	reduce	80	11	10	8	38	147
72	eat	126	3	0	12	5	146
73	hear	74	0	19	26	24	143
74	die	68	4	3	57	6	138
75	show	77	19	3	19	20	138
76	walk	53	1	18	56	7	135
77	forget	106	2	1	11	13	133
78	harm	126	3	0	0	4	133
79	control	113	0	9	0	10	132
80	pass	88	4	9	12	19	132
81	sell	55	4	34	8	30	131
82	deal	118	0	6	0	2	126
83	teach	75	9	22	7	11	124
84	build	70	3	6	12	32	123
85	fail	77	6	2	30	8	123
86	choose	99	1	5	13	4	122
87	step	117	0	0	2	2	121
88	consider	88	2	5	7	17	119
89	limit	15	1	0	1	101	118
90	hope	111	6	0	0	0	117
91	reach	78	9	1	20	9	117

92	talk	78	0	29	9	1	117
93	fit	92	13	6	0	5	116
94	turn	69	11	4	14	17	115
95	hit	16	0	3	76	18	113
96	base	3	1	1	0	107	112
97	leave	51	3	3	30	19	106
98	grow	59	8	19	12	7	105
99	grasp	94	0	2	3	5	104
100	adapt	96	0	5	0	2	103
101	drink	64	3	24	4	7	102
102	earn	91	0	8	1	2	102
103	enjoy	86	3	10	2	0	101
104	join	85	2	8	4	2	101
105	rise	17	4	38	9	30	98
106	devote	82	1	6	5	3	97
107	rain	13	44	36	2	2	97
108	love	80	3	3	7	3	96
109	rush	6	0	2	26	61	95
110	set	49	5	4	19	18	95
111	respect	85	2	0	0	7	94
112	cry	10	0	60	21	1	92
113	decide	57	0	0	26	9	92
114	punish	71	2	2	3	14	92
115	call	46	5	3	7	30	91
116	follow	25	29	28	7	2	91
117	provide	52	19	3	2	15	91
118	seem	18	48	0	20	2	88
119	stand	38	4	9	36	1	88
120	enter	60	1	8	11	4	84
121	draw	61	0	14	5	3	83
122	graduate	66	0	0	8	9	83
123	prove	37	12	0	12	22	83
124	appear	39	10	9	16	7	81
125	catch	59	0	3	8	7	77
126	obtain	69	0	1	2	5	77
127	prefer	71	4	0	1	0	76
128	stop	63	1	5	5	1	75
129	fall	28	1	10	30	5	74
130	require	32	23	4	0	15	74
131	depend	39	28	1	0	5	73
132	defeat	12	0	3	36	21	72
133	drive	49	2	12	6	3	72
134	prepare	47	2	8	1	14	72
135	hold	45	2	5	9	10	71
136	stay	56	0	9	6	0	71
137	break	27	8	8	15	12	70
138	worry	61	1	2	0	6	70
139	grant	0	0	0	2	67	69

140	plan	55	2	4	1	6	68
141	jump	8	0	2	57	0	67
142	avoid	59	0	4	0	3	66
143	apply	51	4	1	1	8	65
144	benefit	61	1	1	0	2	65
145	raise	28	0	8	9	20	65
146	carry	39	3	3	4	13	62
147	knock	6	0	2	25	29	62
148	contribute	46	7	0	3	5	61
149	hurt	30	3	0	11	16	60
150	sit	19	0	12	26	2	59
151	achieve	48	1	1	2	6	58
152	hurry	38	1	0	15	3	57
153	kill	37	2	3	3	12	57
154	smash	4	0	1	50	2	57
155	cheat	14	2	17	0	23	56
156	engage	25	1	13	2	15	56
157	act	41	1	8	1	4	55
158	decline	8	1	7	10	28	54
159	expect	28	3	0	13	10	54
160	attend	42	1	7	3	0	53
161	send	20	0	14	7	12	53
162	start	36	1	5	9	2	53
163	care	43	8	1	0	0	52
164	compare	11	1	11	3	26	52
165	clean	39	0	3	1	8	51
166	cook	24	0	21	4	2	51
167	wish	44	2	2	3	0	51
168	answer	38	0	1	10	1	50
169	result	28	12	1	3	5	49
170	continue	38	3	2	4	1	48
171	insist	40	1	2	2	3	48
172	encourage	37	4	0	3	3	47
173	win	33	0	2	8	4	47
174	exercise	37	8	1	0	0	46
175	regard	35	4	0	0	7	46
176	select	36	1	4	1	3	45
177	suit	38	5	0	0	2	45
178	agree	36	2	0	4	2	44
179	offer	30	2	0	6	5	43
180	practice	1	5	30	5	2	43
181	sing	22	1	17	3	0	43
182	affect	25	6	0	1	10	42
183	wait	26	1	13	2	0	42
184	accept	28	0	5	0	8	41
185	cover	3	2	2	6	27	40
186	exist	34	3	1	2	0	40
187	experience	21	0	2	3	14	40

188	cut	23	2	6	3	5	39
189	obey	38	0	1	0	0	39
190	receive	31	0	2	4	2	39
191	satisfy	8	2	1	2	26	39
192	train	28	1	3	0	7	39
193	beat	12	1	1	17	7	38
194	fight	30	1	5	1	1	38
195	ignore	25	0	4	1	8	38
196	overcome	28	0	4	4	2	38
197	plant	21	0	10	4	3	38
198	conclude	30	0	0	4	3	37
199	move	20	1	8	5	3	37
200	open	21	1	4	8	3	37
201	stick	28	4	3	2	0	37
202	treat	21	1	4	0	11	37
203	hate	32	4	0	0	0	36
204	touch	29	0	3	3	1	36
205	suffer	18	2	4	2	9	35
206	add	15	6	4	3	6	34
207	arrive	12	2	3	15	2	34
208	explain	23	4	1	3	3	34
209	relax	23	1	5	2	3	34
210	wash	15	1	14	3	1	34
211	fill	12	0	1	5	15	33
212	visit	23	0	5	3	2	33
213	challenge	22	0	8	1	1	32
214	educate	15	0	2	0	15	32
215	laugh	14	1	9	8	0	32
216	recite	21	0	7	3	1	32
217	suggest	16	3	0	6	7	32
218	tend	23	5	0	2	2	32
219	destroy	19	1	1	4	6	31
220	drop	7	5	2	4	13	31
221	own	17	8	3	1	1	30
222	prohibit	25	1	1	0	3	30
223	advance	9	1	3	6	10	29
224	determine	5	0	0	5	19	29
225	invent	9	0	2	5	13	29
226	neglect	21	2	0	4	2	29
227	report	10	2	0	0	17	29
228	throw	16	1	5	1	6	29
229	adjust	28	0	0	0	0	28
230	complete	25	0	0	0	3	28
231	dance	19	0	9	0	0	28
232	enlarge	23	1	2	1	1	28
233	manage	22	0	1	4	1	28
234	travel	19	1	7	0	1	28
235	adopt	19	0	3	2	3	27

236	attach	20	1	0	2	4	27
237	concentrate	23	0	4	0	0	27
238	correct	23	0	1	0	3	27
239	miss	16	0	5	4	2	27
240	notice	23	0	1	0	3	27
241	point	3	3	9	11	1	27
242	support	24	1	0	0	2	27
243	acquire	21	0	2	1	2	26
244	close	13	1	2	0	10	26
245	cost	15	4	0	5	2	26
246	forbid	21	1	1	0	3	26
247	form	18	1	0	2	5	26
248	influence	18	0	0	1	7	26
249	recognize	24	0	0	1	1	26
250	refuse	18	0	1	4	3	26
251	ride	24	0	0	2	0	26
252	disappear	19	1	0	3	2	25
253	imagine	22	0	1	0	2	25
254	include	16	7	0	1	1	25
255	occur	13	5	2	4	1	25
256	end	11	5	3	2	3	24
257	swim	15	1	7	1	0	24
258	wear	9	1	2	1	11	24
259	damage	8	3	0	1	11	23
260	exchange	14	0	4	0	5	23
261	express	20	1	1	0	1	23
262	flow	5	3	15	0	0	23
263	skate	11	0	12	0	0	23
264	communicate	18	0	2	0	2	22
265	concern	12	2	7	0	1	22
266	force	10	1	0	2	9	22
267	mention	4	0	0	5	13	22
268	occupy	15	2	1	0	4	22
269	pursue	20	0	2	0	0	22
270	relate	7	0	0	4	11	22
271	return	18	1	1	2	0	22
272	attain	15	0	1	2	3	21
273	belong	13	6	1	1	0	21
274	cure	21	0	0	0	0	21
275	establish	13	0	2	1	5	21
276	indicate	5	12	1	2	1	21
277	last	12	3	1	3	2	21
278	strengthen	16	0	1	2	2	21
279	accumulate	15	0	1	0	4	20
280	cope	20	0	0	0	0	20
281	dislike	18	1	0	0	1	20
282	enable	16	4	0	0	0	20
283	lay	11	1	1	1	6	20

284	resolve	16	0	1	0	3	20
285	seek	14	3	2	0	1	20
286	type	16	0	3	1	0	20
287	connect	13	1	0	0	5	19
288	effect	12	2	0	1	4	19
289	fly	6	0	4	8	1	19
290	hop	7	0	4	4	4	19
291	organize	6	3	1	2	7	19
292	pull	7	0	0	11	1	19
293	search	14	0	4	0	1	19
294	supply	14	2	0	1	2	19
295	afford	16	1	0	0	1	18
296	consume	12	3	1	0	2	18
297	create	11	0	1	0	6	18
298	dream	7	0	4	6	1	18
299	intend	17	0	0	1	0	18
300	promote	12	1	0	0	5	18
301	repair	11	0	3	0	4	18
302	resist	17	0	1	0	0	18
303	review	15	0	1	0	2	18
304	sleep	11	1	5	1	0	18
305	smoke	6	1	11	0	0	18
306	test	10	1	2	0	5	18
307	threaten	9	1	4	0	4	18
308	vary	4	9	1	2	2	18
309	contain	3	11	1	0	2	17
310	examine	12	1	2	0	2	17
311	pour	7	1	2	2	5	17
312	refer	4	9	1	1	2	17
313	settle	11	1	1	1	3	17
314	spare	15	0	1	1	0	17
315	struggle	10	1	5	1	0	17
316	burn	6	0	4	1	5	16
317	check	11	0	3	0	2	16
318	complain	13	2	0	1	0	16
319	contact	14	0	1	1	0	16
320	greet	5	0	9	1	1	16
321	injure	6	1	0	1	8	16
322	lack	4	3	6	2	1	16
323	operate	10	0	4	0	2	16
324	perform	10	2	0	0	4	16
325	progress	3	8	4	0	1	16
326	alter	8	0	5	1	1	15
327	arrange	12	0	1	1	1	15
328	better	9	2	2	0	2	15
329	charge	7	0	0	1	7	15
330	delay	8	0	1	1	5	15
331	identify	12	0	1	0	2	15

332	long	12	2	0	0	1	15
333	place	11	0	0	0	4	15
334	suppose	13	0	1	0	1	15
335	trust	14	0	0	0	1	15
336	admire	11	2	0	1	0	14
337	climb	7	0	3	4	0	14
338	expand	8	1	3	0	2	14
339	inform	3	1	0	1	9	14
340	recycle	11	0	1	0	2	14
341	repeat	9	0	2	3	0	14
342	speed	10	1	2	1	0	14
343	spread	4	2	3	1	4	14
344	absorb	8	0	0	0	5	13
345	analyze	12	0	1	0	0	13
346	consist	5	5	0	0	3	13
347	discover	10	0	0	2	1	13
348	disturb	4	2	2	0	5	13
349	guess	13	0	0	0	0	13
350	judge	6	0	3	0	4	13
351	participate	10	0	3	0	0	13
352	reform	5	0	0	3	5	13
353	reject	12	0	1	0	0	13
354	taste	8	3	0	1	1	13
355	wonder	8	0	3	1	1	13
356	account	5	2	4	0	1	12
357	count	8	2	1	0	1	12
358	demand	4	2	0	0	6	12
359	distinguish	11	0	1	0	0	12
360	enhance	10	0	0	0	2	12
361	focus	10	0	1	0	1	12
362	list	2	0	0	0	10	12
363	lower	6	2	0	1	3	12
364	memorize	8	0	3	0	1	12
365	preserve	10	0	1	0	1	12
366	research	9	0	3	0	0	12
367	steal	5	1	4	0	2	12
368	stimulate	10	0	0	0	2	12
369	advise	9	1	0	0	1	11
370	allow	5	0	1	0	5	11
371	attract	4	3	0	0	4	11
372	discourage	3	1	0	2	5	11
373	fear	8	0	2	1	0	11
374	introduce	5	0	1	5	0	11
375	mind	10	0	0	1	0	11
376	order	8	0	0	2	1	11
377	owe	7	2	0	1	1	11
378	praise	0	1	1	2	7	11
379	rely	6	3	1	1	0	11

380	translate	5	0	6	0	0	11
381	wake	1	4	0	6	0	11
382	warn	7	1	0	2	1	11
383	claim	7	0	0	1	2	10
384	collect	7	0	2	0	1	10
385	confront	5	0	3	1	1	10
386	discuss	3	0	4	1	2	10
387	dress	7	0	2	1	0	10
388	extend	5	1	2	1	1	10
389	observe	6	0	3	0	1	10
390	permit	2	0	4	0	4	10
391	pick	7	0	0	3	0	10
392	purify	6	1	2	0	1	10
393	remain	8	1	0	1	0	10
394	rest	5	3	2	0	0	10
395	rob	4	0	3	1	2	10
396	share	10	0	0	0	0	10
397	shoot	5	0	1	3	1	10
398	shout	4	0	2	4	0	10
399	sound	1	9	0	0	0	10
400	store	3	0	3	0	4	10
401	accomplish	6	0	0	1	2	9
402	appeal	7	2	0	0	0	9
403	arise	5	1	1	2	0	9
404	convert	6	0	1	0	2	9
405	exert	7	1	1	0	0	9
406	fasten	7	0	0	2	0	9
407	found	0	0	1	1	7	9
408	involve	5	0	1	1	2	9
409	manufacture	0	7	1	0	1	9
410	proceed	9	0	0	0	0	9
411	push	4	0	2	2	1	9
412	recover	8	0	0	0	1	9
413	retire	5	1	1	2	0	9
414	risk	9	0	0	0	0	9
415	strike	3	0	3	3	0	9
416	telephone	6	0	1	2	0	9
417	unite	6	0	0	1	2	9
418	water	8	0	0	0	1	9
419	welcome	4	0	2	0	3	9
420	ban	6	0	1	0	1	8
421	clear	8	0	0	0	0	8
422	commit	7	0	0	0	1	8
423	confuse	3	0	0	1	4	8
424	deny	7	0	0	0	1	8
425	emerge	3	1	2	0	2	8
426	feed	7	0	1	0	0	8
427	hunt	2	0	6	0	0	8

428	illustrate	5	1	0	2	0	8
429	imply	4	4	0	0	0	8
430	impress	4	0	0	1	3	8
431	investigate	6	0	1	0	1	8
432	invite	7	0	0	1	0	8
433	liberate	2	0	0	0	6	8
434	light	6	0	0	1	1	8
435	predict	6	0	1	1	0	8
436	recall	7	0	0	1	0	8
437	shop	2	0	6	0	0	8
438	smile	4	0	1	3	0	8
439	appreciate	4	0	0	0	3	7
440	bear	7	0	0	0	0	7
441	blow	3	2	1	1	0	7
442	breathe	4	0	3	0	0	7
443	combine	6	0	0	0	1	7
444	compete	5	1	1	0	0	7
445	comply	6	0	0	1	0	7
446	crash	4	0	0	3	0	7
447	derive	4	0	2	0	1	7
448	display	4	0	0	0	3	7
449	divide	2	1	0	0	4	7
450	endanger	6	1	0	0	0	7
451	ensure	5	1	0	0	1	7
452	fix	2	0	0	0	5	7
453	hide	1	1	0	4	1	7
454	interest	3	2	0	1	1	7
455	possess	6	0	0	1	0	7
456	ring	0	0	0	7	0	7
457	scold	4	0	0	2	1	7
458	surprise	4	0	0	1	2	7
459	surround	3	0	4	0	0	7
460	adhere	6	0	0	0	0	6
461	admit	3	0	0	0	3	6
462	advertise	4	1	1	0	0	6
463	aim	4	0	0	0	2	6
464	blame	4	1	0	0	1	6
465	borrow	4	0	0	1	1	6
466	broadcast	4	0	1	1	0	6
467	broaden	5	0	0	0	1	6
468	cherish	6	0	0	0	0	6
469	compose	2	0	0	1	3	6
470	conduct	4	0	1	0	1	6
471	conform	5	0	0	0	1	6
472	crowd	0	0	0	3	3	6
473	defend	5	1	0	0	0	6
474	dig	5	0	0	1	0	6
475	exceed	5	0	0	0	1	6

476	excite	2	0	0	0	4	6
477	exhaust	3	0	0	0	3	6
478	grab	2	0	1	1	2	6
479	hand	0	0	1	0	5	6
480	handle	5	0	0	0	1	6
481	invade	5	0	0	0	1	6
482	link	3	0	1	0	2	6
483	omit	3	0	2	1	0	6
484	persist	6	0	0	0	0	6
485	prolong	4	0	0	0	2	6
486	promise	2	1	0	0	3	6
487	pronounce	5	0	1	0	0	6
488	qualify	2	0	0	0	4	6
489	range	0	3	0	1	2	6
490	react	5	0	0	1	0	6
491	release	3	0	0	0	3	6
492	request	5	0	0	0	1	6
493	smell	1	4	0	1	0	6
494	spell	5	0	0	0	1	6
495	spoil	2	1	0	0	3	6
496	symbolize	1	4	1	0	0	6
497	transfer	5	1	0	0	0	6
498	accelerate	4	0	0	0	1	5
499	approach	3	0	0	0	2	5
500	associate	2	0	0	0	3	5
501	command	4	0	1	0	0	5
502	confirm	3	0	0	0	2	5
503	construct	3	0	1	0	1	5
504	cross	4	0	1	0	0	5
505	cultivate	2	2	1	0	0	5
506	deliver	2	0	3	0	0	5
507	desire	3	0	1	0	1	5
508	elect	2	0	1	1	1	5
509	eliminate	4	0	1	0	0	5
510	escape	4	0	0	1	0	5
511	experiment	5	0	0	0	0	5
512	expose	5	0	0	0	0	5
513	fulfill	4	0	1	0	0	5
514	hang	1	0	1	2	1	5
515	head	3	0	1	1	0	5
516	matter	5	0	0	0	0	5
517	name	5	0	0	0	0	5
518	please	2	1	0	0	2	5
519	poison	1	0	1	1	2	5
520	present	1	2	0	1	1	5
521	quarrel	1	0	1	2	1	5
522	quit	3	0	0	1	1	5
523	record	4	0	1	0	0	5

524	reflect	2	1	0	2	0	5
525	reveal	4	0	1	0	0	5
526	reward	2	0	0	0	3	5
527	specialize	3	0	1	0	1	5
528	stare	1	0	4	0	0	5
529	starve	4	0	1	0	0	5
530	thank	4	1	0	0	0	5
531	transform	3	1	1	0	0	5
532	utilize	3	0	2	0	0	5
533	acquaint	3	0	0	0	1	4
534	alternate	1	0	3	0	0	4
535	approve	3	0	0	0	1	4
536	assume	3	0	0	0	1	4
537	attempt	3	1	0	0	0	4
538	chase	3	0	1	0	0	4
539	cool	1	1	0	0	2	4
540	copy	2	0	1	1	0	4
541	declare	1	0	0	3	0	4
542	dedicate	3	1	0	0	0	4
543	define	1	1	0	1	1	4
544	differ	2	1	1	0	0	4
545	disable	4	0	0	0	0	4
546	emphasize	2	0	1	0	1	4
547	equip	1	0	0	0	3	4
548	estimate	4	0	0	0	0	4
549	explore	2	0	0	0	2	4
550	heat	4	0	0	0	0	4
551	imitate	2	0	0	0	2	4
552	lift	3	0	0	1	0	4
553	marry	2	0	0	2	0	4
554	mistake	1	2	0	0	1	4
555	oppose	3	0	1	0	0	4
556	park	4	0	0	0	0	4
557	process	3	0	1	0	0	4
558	publish	3	0	0	0	1	4
559	regret	4	0	0	0	0	4
560	relieve	3	0	0	0	1	4
561	remind	3	1	0	0	0	4
562	represent	4	0	0	0	0	4
563	sew	2	0	2	0	0	4
564	shine	0	0	4	0	0	4
565	shorten	3	0	0	0	1	4
566	skim	2	1	0	1	0	4
567	tear	0	0	0	3	1	4
568	vanish	3	0	0	1	0	4
569	warm	3	0	1	0	0	4

Appendix 3: A verb lemma list of LOCNESS

	Lemma	V-e	V-s	V-ing	V-ed	V-n	Total
1	make	426	113	129	88	231	987
2	take	289	76	111	59	132	667
3	see	306	48	27	35	219	635
4	use	198	52	96	27	190	563
5	become	209	69	75	60	86	499
6	say	178	110	68	76	61	493
7	give	164	51	61	40	137	453
8	go	201	91	79	34	37	442
9	feel	280	70	13	57	13	433
10	want	215	105	16	71	19	426
11	get	275	26	64	31	25	421
12	think	237	25	34	33	37	366
13	believe	220	75	13	41	16	365
14	know	193	36	31	29	64	353
15	show	134	79	39	15	83	350
16	come	121	84	0	79	40	324
17	find	165	16	22	32	75	310
18	seem	128	141	1	24	0	294
19	need	131	61	3	25	65	285
20	allow	82	46	42	5	95	270
21	lead	112	49	22	22	61	266
22	try	84	30	120	13	19	266
23	live	143	27	60	9	11	250
24	mean	74	81	13	33	23	224
25	change	92	11	18	11	83	215
26	bring	80	34	17	18	62	211
27	work	111	27	51	14	7	210
28	look	99	24	43	7	32	205
29	leave	68	25	22	16	70	201
30	help	120	35	17	9	17	198
31	kill	64	20	40	18	54	196
32	cause	67	37	30	17	44	195
33	lose	62	16	27	25	53	183
34	put	78	13	20	7	64	182
35	state	36	74	21	18	31	180
36	create	70	18	28	8	55	179
37	begin	51	39	13	52	23	178
38	accept	90	14	20	1	43	168
39	keep	97	14	31	9	13	164
40	continue	100	29	12	11	11	163
41	argue	80	29	12	11	30	162
42	consider	59	8	11	7	73	158
43	decide	63	37	9	26	18	153

44	understand	108	15	18	1	9	151
45	tell	41	40	11	24	32	148
46	play	63	20	33	17	14	147
47	happen	66	33	18	18	10	145
48	pay	82	18	20	4	21	145
49	die	72	18	16	20	15	141
50	choose	56	23	18	27	10	134
51	increase	65	11	23	5	28	132
52	provide	64	30	19	5	14	132
53	start	42	28	13	35	11	129
54	support	83	9	12	3	20	127
55	realize	72	21	15	6	8	122
56	run	47	10	42	7	14	120
57	prove	61	8	8	4	37	118
58	carry	47	10	20	4	36	117
59	stop	80	7	8	6	15	116
60	write	19	29	15	23	28	114
61	ask	46	14	12	20	21	113
62	learn	69	6	13	8	15	111
63	follow	51	16	24	9	10	110
64	involve	16	24	16	21	31	108
65	hold	43	14	8	5	35	105
66	receive	44	13	19	15	12	103
67	teach	40	4	18	4	36	102
68	turn	45	18	14	7	17	101
69	present	26	15	13	8	37	99
70	spend	42	7	8	10	32	99
71	ban	20	0	21	4	53	98
72	face	35	10	18	3	32	98
73	realise	33	38	7	11	9	98
74	deal	39	13	29	2	12	95
75	act	60	11	16	3	4	94
76	occur	46	24	7	9	7	93
77	exist	42	36	0	9	5	92
78	fight	38	6	33	5	9	91
79	hear	27	10	5	14	35	91
80	like	75	15	1	0	0	91
81	pass	39	7	6	7	32	91
82	reduce	49	6	10	1	25	91
83	watch	50	2	30	7	2	91
84	develop	34	7	6	13	30	90
85	call	21	10	7	6	44	88
86	commit	35	10	20	5	17	87
87	remain	42	28	7	9	1	87
88	win	36	3	18	20	10	87
89	achieve	50	4	8	1	23	86
90	appear	35	35	8	6	1	85
91	reject	16	32	9	12	16	85

92	grow	36	10	18	6	14	84
93	buy	59	4	12	4	3	82
94	force	9	2	8	4	59	82
95	suffer	36	5	21	3	17	82
96	affect	23	17	4	1	36	81
97	claim	34	19	10	7	11	81
98	produce	43	10	10	1	17	81
99	agree	56	11	3	5	5	80
100	end	50	3	6	12	8	79
101	read	26	5	33	2	13	79
102	save	40	6	22	3	7	78
103	discuss	32	5	17	3	20	77
104	let	54	4	13	2	3	76
105	prevent	54	6	7	1	8	76
106	base	7	1	0	2	65	75
107	form	41	5	9	3	17	75
108	judge	39	7	18	2	8	74
109	require	20	20	6	0	28	74
110	set	21	11	16	4	22	74
111	include	37	17	2	7	10	73
112	raise	22	9	9	4	29	73
113	stay	57	7	3	6	0	73
114	explain	30	19	8	4	11	72
115	view	28	1	4	11	28	72
116	define	27	8	2	4	30	71
117	serve	33	16	8	2	11	70
118	talk	27	7	27	2	7	70
119	break	35	5	10	7	12	69
120	improve	37	3	17	1	11	69
121	encourage	32	8	8	2	16	66
122	gain	34	3	8	3	18	66
123	stand	27	23	5	8	3	66
124	move	26	12	13	3	11	65
125	offer	29	9	5	4	18	65
126	speak	34	4	21	3	3	65
127	meet	22	15	11	8	8	64
128	refuse	11	23	10	12	7	63
129	discover	17	9	5	9	22	62
130	experience	26	2	13	3	18	62
131	introduce	13	2	8	6	32	61
132	represent	19	27	5	3	7	61
133	place	12	11	5	1	31	60
134	sell	31	4	8	2	15	60
135	build	30	2	9	3	14	58
136	determine	23	4	7	0	24	58
137	result	28	5	11	3	11	58
138	catch	14	2	1	8	32	57
139	expect	25	2	4	3	23	57

140	fall	16	12	8	11	10	57
141	reach	22	8	7	5	15	57
142	apply	27	10	5	2	12	56
143	express	24	6	7	7	12	56
144	wish	32	9	4	10	1	56
145	control	34	4	4	2	11	55
146	deny	17	13	10	3	10	53
147	describe	3	20	4	9	17	53
148	enjoy	32	9	5	3	4	53
149	enter	29	7	7	3	7	53
150	establish	13	1	5	6	28	53
151	point	17	23	3	7	3	53
152	attempt	20	6	17	5	3	51
153	eat	40	1	1	1	8	51
154	reveal	20	12	11	2	6	51
155	solve	27	0	8	1	15	51
156	suggest	21	16	3	5	6	51
157	travel	36	2	12	0	1	51
158	tend	32	7	1	8	2	50
159	compare	12	4	1	2	30	49
160	draw	14	6	4	5	20	49
161	fail	22	12	6	5	4	49
162	seek	15	14	9	7	4	49
163	destroy	19	3	7	4	15	48
164	perform	19	7	5	1	16	48
165	treat	11	5	4	1	26	47
166	arise	20	12	0	11	3	46
167	justify	27	4	1	0	14	46
168	obtain	21	1	9	2	13	46
169	prepare	5	1	6	1	33	46
170	throw	13	4	7	1	21	46
171	add	14	7	7	8	9	45
172	drink	26	0	16	2	1	45
173	recognize	26	5	3	5	6	45
174	adopt	11	5	8	3	17	44
175	attack	16	15	7	3	3	44
176	contain	23	9	7	3	2	44
177	cut	22	2	4	0	16	44
178	protect	35	3	3	0	3	44
179	relate	16	8	6	2	12	44
180	survive	30	3	3	3	4	43
181	open	19	5	4	8	6	42
182	remember	30	3	3	0	6	42
183	wear	15	3	15	1	8	42
184	assume	21	4	7	4	5	41
185	decrease	21	3	5	4	8	41
186	illustrate	13	15	3	5	5	41
187	refer	7	13	6	0	15	41

188	drive	17	2	12	2	7	40
189	join	18	3	12	3	4	40
190	lower	6	2	22	0	10	40
191	maintain	20	9	6	2	3	40
192	admit	11	20	2	4	2	39
193	afford	37	0	0	0	2	39
194	promote	23	4	6	3	3	39
195	focus	17	9	5	1	6	38
196	hit	6	6	3	4	19	38
197	return	15	2	7	6	8	38
198	send	12	4	3	5	14	38
199	avoid	22	1	5	0	9	37
200	oppose	11	2	9	2	13	37
201	reflect	19	11	4	0	3	37
202	study	15	0	12	4	6	37
203	cost	16	14	3	1	2	36
204	mention	12	4	2	5	13	36
205	question	17	3	6	5	5	36
206	benefit	22	0	5	1	7	35
207	deserve	23	8	0	3	1	35
208	love	13	3	4	5	10	35
209	remove	17	2	3	0	13	35
210	share	18	2	6	1	8	35
211	educate	14	0	5	0	15	34
212	link	2	2	2	1	27	34
213	retain	13	5	7	1	8	34
214	separate	8	1	7	0	18	34
215	blame	19	2	5	0	7	33
216	demonstrate	14	8	1	1	9	33
217	ensure	25	0	5	1	2	33
218	eliminate	18	0	6	0	8	32
219	push	11	6	5	3	7	32
220	replace	10	2	5	2	13	32
221	sit	17	3	7	3	2	32
222	alter	12	0	4	1	14	31
223	contract	12	1	6	5	7	31
224	earn	12	4	8	0	7	31
225	enable	14	5	2	2	8	31
226	fit	26	1	1	0	3	31
227	forget	20	0	1	2	8	31
228	listen	22	0	6	0	3	31
229	report	8	2	2	9	10	31
230	aim	3	7	3	9	8	30
231	care	27	3	0	0	0	30
232	recognise	11	7	2	1	9	30
233	test	8	1	9	0	12	30
234	walk	11	3	5	10	1	30
235	conclude	19	2	1	3	4	29

236	drop	7	1	5	4	12	29
237	expose	10	6	4	1	8	29
238	fear	13	4	5	5	2	29
239	regard	9	3	4	0	13	29
240	rule	9	3	1	6	10	29
241	strengthen	17	1	4	2	5	29
242	escape	23	1	0	2	2	28
243	ignore	12	1	2	2	11	28
244	influence	11	2	1	0	14	28
245	legalize	5	0	13	0	10	28
246	abolish	8	0	7	1	11	27
247	confess	9	5	10	2	1	27
248	contribute	15	5	1	3	3	27
249	elect	4	0	0	1	22	27
250	examine	12	2	5	3	5	27
251	fill	14	1	0	3	9	27
252	limit	6	0	1	1	19	27
253	manage	4	10	0	5	8	27
254	associate	1	0	2	0	23	26
255	belong	12	7	5	2	0	26
256	lack	11	9	5	1	0	26
257	murder	11	2	4	3	6	26
258	succeed	15	2	0	2	7	26
259	beat	3	0	5	8	9	25
260	conduct	7	2	1	6	9	25
261	effect	13	2	3	0	7	25
262	expand	9	3	2	3	8	25
263	hope	17	8	0	0	0	25
264	waste	12	1	4	0	8	25
265	witness	7	3	4	7	4	25
266	admire	12	4	1	1	6	24
267	answer	15	1	1	1	6	24
268	arrive	7	9	0	7	1	24
269	attend	15	1	3	4	1	24
270	cheat	9	0	9	2	4	24
271	debate	2	0	2	1	19	24
272	depend	10	10	0	0	4	24
273	disagree	15	5	1	3	0	24
274	dominate	14	2	2	1	5	24
275	dress	12	3	1	0	8	24
276	impose	11	1	3	0	9	24
277	last	15	1	3	4	1	24
278	overcome	18	3	1	0	2	24
279	repent	16	3	5	0	0	24
280	wait	15	1	8	0	0	24
281	worry	7	1	1	1	14	24
282	condemn	8	3	4	2	6	23
283	demand	8	2	3	5	5	23

284	enhance	13	0	5	2	3	23
285	respect	16	2	0	2	3	23
286	shoot	5	2	3	7	6	23
287	tackle	7	9	2	1	4	23
288	wonder	15	3	4	1	0	23
289	address	8	2	7	0	5	22
290	display	5	10	1	1	5	22
291	imply	7	13	0	1	1	22
292	marry	12	0	1	3	6	22
293	persuade	15	1	2	0	4	22
294	plan	10	2	2	5	3	22
295	possess	13	3	4	1	1	22
296	pray	14	1	7	0	0	22
297	rely	14	4	1	0	3	22
298	analyze	12	1	4	0	4	21
299	bear	13	0	4	0	4	21
300	communicate	17	1	2	0	1	21
301	grant	1	1	1	4	14	21
302	implement	7	0	2	1	11	21
303	intend	3	4	0	5	9	21
304	punish	6	0	0	0	15	21
305	restrict	10	0	2	0	9	21
306	train	8	6	1	0	6	21
307	accuse	1	4	1	5	9	20
308	attract	10	2	2	0	6	20
309	attribute	5	7	0	0	8	20
310	concern	5	6	6	1	2	20
311	consume	7	1	4	1	7	20
312	defend	12	2	4	1	1	20
313	divide	6	1	0	2	11	20
314	emphasize	12	5	1	0	2	20
315	feed	13	1	1	1	4	20
316	hang	3	0	3	3	11	20
317	inform	6	0	0	3	11	20
318	integrate	6	0	6	0	8	20
319	outweigh	12	6	0	0	2	20
320	own	11	4	1	1	3	20
321	release	3	2	3	0	12	20
322	rid	5	1	9	0	5	20
323	sacrifice	10	4	3	1	2	20
324	sign	6	1	2	4	7	20
325	acquire	8	1	5	0	5	19
326	design	2	0	0	1	16	19
327	encounter	10	5	0	0	4	19
328	free	8	2	5	1	3	19
329	identify	9	3	1	2	4	19
330	insist	8	4	3	4	0	19
331	notice	6	3	0	3	7	19

332	participate	11	1	4	1	2	19
333	portray	9	6	4	0	0	19
334	prefer	13	0	3	2	1	19
335	refute	10	2	0	0	7	19
336	rise	3	0	4	7	5	19
337	accomplish	13	0	0	1	4	18
338	bind	5	0	2	2	9	18
339	concentrate	6	2	6	0	4	18
340	cover	10	1	0	3	4	18
341	direct	3	1	0	0	14	18
342	evoke	9	5	0	0	4	18
343	extend	8	0	3	0	7	18
344	fulfil(l)	4	3	6	1	4	18
345	gather	9	2	1	3	3	18
346	invent	5	1	1	4	7	18
347	purchase	8	0	5	0	5	18
348	pursue	8	2	5	0	3	18
349	respond	7	4	3	2	2	18
350	steal	3	0	5	5	5	18
351	compete	13	0	4	0	0	17
352	confront	6	4	1	0	6	17
353	cope	17	0	0	0	0	17
354	damage	8	1	5	0	3	17
355	deem	1	3	0	0	13	17
356	enforce	8	1	2	0	6	17
357	interpret	4	1	2	2	8	17
358	label	3	1	1	2	10	17
359	manipulate	7	1	4	1	4	17
360	pick	9	4	2	0	2	17
361	regulate	9	0	4	3	1	17
362	satisfy	8	1	0	1	7	17
363	threaten	2	2	3	2	8	17
364	advocate	5	0	7	1	3	16
365	assist	11	0	2	2	1	16
366	cease	12	0	3	0	1	16
367	combat	15	0	1	0	0	16
368	conform	13	0	3	0	0	16
369	desire	7	4	0	3	2	16
370	employ	5	2	0	1	8	16
371	guarantee	4	5	2	0	5	16
372	pose	8	2	2	2	2	16
373	process	11	1	0	0	4	16
374	progress	5	3	2	3	3	16
375	rape	1	1	1	1	12	16
376	react	9	3	3	1	0	16
377	reinforce	4	6	1	1	4	16
378	reward	4	1	2	0	9	16
379	search	4	0	10	1	1	16

380	trust	9	2	0	0	5	16
381	vote	9	0	7	0	0	16
382	appeal	7	3	3	0	2	15
383	appoint	4	1	1	2	7	15
384	appreciate	10	1	1	0	3	15
385	close	4	2	2	0	7	15
386	cross	8	2	4	0	1	15
387	fly	2	1	8	2	2	15
388	handle	12	0	1	0	2	15
389	imagine	13	1	0	0	1	15
390	perceive	9	2	0	0	4	15
391	predict	9	0	1	2	3	15
392	propose	6	0	1	3	5	15
393	ridicule	4	2	3	1	5	15
394	sleep	8	4	2	0	1	15
395	transfer	7	0	2	1	5	15
396	transport	8	1	2	0	4	15
397	undergo	3	4	3	1	4	15
398	acknowledge	8	0	4	1	1	14
399	advance	5	2	5	0	2	14
400	aid	10	0	2	0	2	14
401	back	9	0	1	2	2	14
402	chase	5	0	4	0	5	14
403	cite	5	3	2	1	3	14
404	compromise	10	0	0	1	3	14
405	criticise	3	1	2	1	7	14
406	endure	6	4	0	1	3	14
407	head	4	0	1	1	8	14
408	hide	9	0	0	0	5	14
409	mind	12	0	0	0	2	14
410	miss	8	0	3	1	2	14
411	recycle	4	0	4	0	6	14
412	regain	12	0	0	1	1	14
413	sound	6	7	0	1	0	14
414	spread	4	1	3	2	4	14
415	stem	7	4	2	0	1	14
416	struggle	3	2	7	2	0	14
417	transmit	4	0	4	0	6	14
418	deprive	3	1	5	0	4	13
419	divorce	7	0	1	2	3	13
420	ease	8	1	2	0	2	13
421	engage	7	1	0	0	5	13
422	exercise	5	2	2	1	3	13
423	file	3	0	2	3	5	13
424	harm	10	0	0	1	2	13
425	misuse	5	1	0	0	7	13
426	practice	0	1	3	1	8	13
427	preserve	9	0	1	0	3	13

428	prohibit	4	2	1	2	4	13
429	select	3	0	1	1	8	13
430	shake	3	1	1	1	7	13
431	stick	3	2	1	0	7	13
432	stress	8	2	1	1	1	13
433	strike	3	1	1	4	4	13
434	weigh	8	3	0	0	2	13
435	abuse	4	1	2	1	4	12
436	adapt	11	0	0	0	1	12
437	adhere	4	1	3	0	4	12
438	attain	6	0	3	0	3	12
439	consist	2	3	1	3	3	12
440	convince	9	1	0	1	1	12
441	cry	3	5	3	1	0	12
442	depict	0	5	3	1	3	12
443	dismiss	5	2	0	1	4	12
444	drown	1	3	2	3	3	12
445	execute	1	1	1	1	8	12
446	govern	6	2	0	0	4	12
447	invest	3	0	4	1	4	12
448	matter	12	0	0	0	0	12
449	oblige	0	0	0	0	12	12
450	publish	1	1	1	3	6	12
451	pull	4	3	1	1	3	12
452	reconcile	7	1	2	0	2	12
453	relieve	9	0	1	0	2	12
454	remind	3	3	3	1	2	12
455	resign	3	3	0	4	2	12
456	risk	8	2	2	0	0	12
457	suppose	10	0	0	1	1	12
458	swim	8	0	4	0	0	12
459	sympathise	10	1	0	0	1	12
460	utilize	4	1	2	0	5	12
461	weaken	3	1	0	2	6	12
462	abandon	2	2	1	1	5	11
463	announce	0	8	0	3	0	11
464	attach	2	1	1	0	7	11
465	balance	9	0	0	1	1	11
466	check	7	1	0	0	3	11
467	discourage	6	3	0	1	1	11
468	discriminate	1	0	3	1	6	11
469	entail	3	6	1	0	1	11
470	finish	5	0	2	1	3	11
471	function	7	0	4	0	0	11
472	generate	4	2	0	2	3	11
473	hate	7	4	0	0	0	11
474	inflict	4	1	2	1	3	11
475	interfere	7	1	3	0	0	11

476	jump	4	0	1	4	2	11
477	laugh	5	1	5	0	0	11
478	loose	8	1	2	0	0	11
479	mix	4	1	0	1	5	11
480	mock	4	2	1	2	2	11
481	observe	2	2	3	0	4	11
482	permit	2	1	0	2	6	11
483	prescribe	3	2	2	0	4	11
484	recover	7	0	1	0	3	11
485	resolve	4	2	1	0	4	11
486	sue	2	0	1	2	6	11
487	visit	4	3	0	1	3	11
488	anger	3	2	0	2	3	10
489	blow	0	0	3	0	7	10
490	breed	1	0	2	0	7	10
491	clean	4	0	4	1	1	10
492	contrast	2	3	2	0	3	10
493	convict	2	0	0	0	8	10
494	diminish	4	1	0	0	5	10
495	dissolve	7	0	0	3	0	10
496	embrace	2	1	4	0	3	10
497	fix	3	1	1	0	5	10
498	guess	9	1	0	0	0	10
499	indicate	4	2	3	1	0	10
500	institute	2	0	2	1	5	10
501	interact	8	0	2	0	0	10
502	legislate	9	1	0	0	0	10
503	match	4	2	1	1	2	10
504	measure	3	1	1	1	4	10
505	operate	7	2	1	0	0	10
506	order	3	3	0	2	2	10
507	overlook	4	0	0	1	5	10
508	pour	8	0	0	0	2	10
509	reply	1	6	0	3	0	10
510	reverse	4	2	1	0	3	10
511	shock	4	1	0	2	3	10
512	smoke	4	0	3	2	1	10
513	step	6	0	3	1	0	10
514	thank	10	0	0	0	0	10
515	whip	2	0	0	1	7	10
516	advise	4	1	1	0	3	9
517	arrest	2	0	0	0	7	9
518	assert	4	4	0	0	1	9
519	burn	0	0	5	1	3	9
520	challenge	4	0	0	1	4	9
521	complete	2	0	1	0	6	9
522	connect	2	0	1	0	6	9
523	constitute	4	5	0	0	0	9

524	cook	3	0	5	0	1	9
525	delay	3	0	1	1	4	9
526	derive	1	2	0	0	6	9
527	disappear	5	3	0	1	0	9
528	dislike	5	2	0	1	1	9
529	donate	1	0	1	0	7	9
530	dream	7	0	0	0	2	9
531	embark	3	3	1	2	0	9
532	epitomise	3	5	0	0	1	9
533	favour	3	0	1	1	4	9
534	found	0	0	0	0	9	9
535	inhibit	5	2	1	0	1	9
536	insure	8	0	1	0	0	9
537	intervene	5	0	2	2	0	9
538	invite	0	3	3	0	3	9
539	mark	0	1	2	1	5	9
540	note	7	2	0	0	0	9
541	please	6	1	1	1	0	9
542	proceed	2	2	0	5	0	9
543	proclaim	1	6	1	1	0	9
544	provoke	2	4	2	1	0	9
545	range	2	1	6	0	0	9
546	repeal	7	0	0	0	2	9
547	rush	4	1	3	1	0	9
548	score	7	0	1	1	0	9
549	secure	5	1	2	0	1	9
550	sentence	0	0	1	1	7	9
551	slip	3	0	2	2	2	9
552	strive	5	0	4	0	0	9
553	submit	7	0	1	1	0	9
554	supply	3	3	0	0	3	9
555	symbolise	0	3	0	0	6	9
556	tear	0	0	3	0	6	9
557	televise	6	0	1	0	2	9
558	undermine	3	0	2	0	4	9
559	veto	5	0	0	2	2	9
560	violate	2	0	3	2	2	9
561	voice	5	0	1	1	2	9
562	withdraw	3	1	1	1	3	9
563	alleviate	6	0	0	0	2	8
564	calculate	5	0	0	1	2	8
565	classify	3	0	0	1	4	8
566	clear	5	0	1	1	1	8
567	comfort	6	0	2	0	0	8
568	compensate	5	0	0	0	3	8
569	comprehend	7	0	1	0	0	8
570	contemplate	2	1	3	0	2	8
571	contradict	1	3	2	1	1	8

572	convert	4	0	3	1	0	8
573	convey	5	1	2	0	0	8
574	criticize	4	2	0	0	2	8
575	declare	0	1	2	2	3	8
576	decline	1	1	1	1	4	8
577	deliver	3	1	0	0	4	8
578	enact	3	0	1	1	3	8
579	endanger	4	0	1	0	3	8
580	endorse	4	1	1	0	2	8
581	exclude	3	1	3	0	1	8
582	experiment	5	0	3	0	0	8
583	forgive	4	0	0	2	2	8
584	frighten	1	1	1	1	4	8
585	hand	1	2	3	0	2	8
586	injure	2	0	0	0	6	8
587	kiss	3	0	5	0	0	8
588	lessen	6	1	0	0	1	8
589	neglect	3	1	0	1	3	8
590	nurture	1	1	4	0	2	8
591	preach	1	5	0	0	2	8
592	promise	1	0	1	4	2	8
593	protest	2	0	3	2	1	8
594	rank	0	0	1	0	7	8
595	rebel	4	0	4	0	0	8
596	recite	5	0	2	0	1	8
597	render	2	2	2	0	2	8
598	research	3	1	3	0	1	8
599	retire	4	2	2	0	0	8
600	revolve	3	1	2	1	1	8
601	rip	1	0	2	0	5	8
602	rob	2	0	1	1	4	8
603	store	5	0	0	0	3	8
604	suit	4	0	0	0	4	8
605	unite	5	0	1	0	2	8
606	upset	1	0	0	2	5	8
607	wash	2	0	5	0	1	8
608	analyse	3	0	3	0	1	7
609	approach	1	1	3	1	1	7
610	behave	5	1	1	0	0	7
611	bother	3	0	0	1	3	7
612	broaden	1	1	2	0	3	7
613	charge	2	0	0	0	5	7
614	collect	4	0	0	0	3	7
615	complain	5	1	0	1	0	7
616	confuse	1	1	0	0	5	7
617	consult	1	1	2	1	2	7
618	contact	4	0	0	1	2	7
619	correct	6	0	0	0	1	7

620	count	6	1	0	0	0	7
621	date	1	2	4	0	0	7
622	differ	6	0	0	0	1	7
623	distinguish	5	2	0	0	0	7
624	doubt	3	2	1	1	0	7
625	dump	2	0	2	3	0	7
626	emerge	4	1	0	2	0	7
627	emphasise	0	2	0	1	4	7
628	highlight	2	2	1	0	2	7
629	hinder	6	0	0	0	1	7
630	hunt	3	1	0	0	3	7
631	infringe	1	1	2	0	3	7
632	insert	3	0	0	1	3	7
633	instruct	2	0	0	0	5	7
634	locate	2	0	0	0	5	7
635	partake	6	1	0	0	0	7
636	praise	2	0	2	1	2	7
637	profess	2	3	0	1	1	7
638	program	5	0	2	0	0	7
639	project	5	0	1	0	1	7
640	quit	6	0	0	0	1	7
641	quote	2	3	0	0	2	7
642	renew	1	0	1	0	5	7
643	rest	2	3	1	0	1	7
644	roll	1	1	3	1	1	7
645	sack	4	0	0	2	1	7
646	safeguard	4	0	0	0	3	7
647	shape	3	0	4	0	0	7
648	slaughter	2	0	0	1	4	7
649	subject	0	0	2	0	5	7
650	sustain	2	0	2	1	2	7
651	switch	5	1	0	1	0	7
652	sympathize	5	0	0	1	1	7
653	accommodate	6	0	0	0	0	6
654	accompany	2	1	1	1	1	6
655	account	2	1	1	0	2	6
656	await	1	3	2	0	0	6
657	block	2	1	0	0	3	6
658	borrow	3	0	0	2	1	6
659	celebrate	3	1	1	0	1	6
660	censor	1	0	0	0	5	6
661	clarify	5	0	0	0	1	6
662	combine	2	0	1	0	3	6
663	condone	2	1	0	0	3	6
664	copy	5	0	1	0	0	6
665	defeat	1	0	3	0	2	6
666	defy	1	3	0	1	1	6
667	dictate	3	0	0	1	2	6

668	disregard	2	0	0	0	4	6
669	exceed	2	2	1	0	1	6
670	exploit	2	0	2	0	2	6
671	facilitate	5	0	1	0	0	6
672	further	1	0	4	0	1	6
673	guide	3	0	0	0	3	6
674	impact	3	0	0	0	3	6
675	install	2	0	0	0	4	6
676	issue	2	0	0	2	2	6
677	kick	2	1	0	0	3	6
678	knock	3	0	2	0	1	6
679	legalise	1	0	0	0	5	6
680	list	1	1	0	0	4	6
681	load	0	0	1	0	5	6
682	mature	2	0	1	0	3	6
683	monitor	1	0	1	0	4	6
684	negotiate	4	0	2	0	0	6
685	object	2	0	1	1	2	6
686	offend	1	1	1	0	3	6
687	offset	4	0	0	0	2	6
688	organise	2	0	1	1	2	6
689	organize	4	0	0	0	2	6
690	overhear	1	2	1	0	2	6
691	persist	2	4	0	0	0	6
692	plague	2	0	0	1	3	6
693	pollute	3	1	1	0	1	6
694	press	3	0	1	0	2	6
695	prevail	6	0	0	0	0	6
696	publicise	1	0	1	0	4	6
697	race	0	0	5	1	0	6
698	repeat	1	2	0	1	2	6
699	resent	5	1	0	0	0	6
700	resist	1	2	0	1	2	6
701	reunite	1	0	0	0	5	6
702	revolt	5	1	0	0	0	6
703	ride	5	0	0	0	1	6
704	ruin	5	1	0	0	0	6
705	scare	3	2	0	0	1	6
706	shift	4	0	1	0	1	6
707	smash	0	1	3	0	2	6
708	stimulate	4	1	0	0	1	6
709	tap	3	0	3	0	0	6
710	tie	1	2	1	0	2	6
711	time	5	1	0	0	0	6
712	tolerate	4	0	0	0	2	6
713	value	2	0	0	0	4	6
714	vary	1	3	0	1	1	6
715	venture	2	1	2	0	1	6

716	wake	6	0	0	0	0	6
717	adjust	2	0	0	0	3	5
718	align	2	0	1	0	2	5
719	arrange	1	1	0	1	2	5
720	assure	4	0	0	0	1	5
721	bet	0	1	4	0	0	5
722	betray	1	1	2	1	0	5
723	bump	2	0	3	0	0	5
724	centre	0	0	0	0	5	5
725	comply	4	0	0	1	0	5
726	comprise	1	1	2	0	1	5
727	confirm	0	2	0	1	2	5
728	conquer	1	1	1	1	1	5
729	counteract	4	0	0	0	1	5
730	cultivate	2	0	0	0	3	5
731	cure	5	0	0	0	0	5
732	curtail	1	0	2	0	2	5
733	deceive	2	0	3	0	0	5
734	desensitize	1	2	0	0	2	5
735	detect	0	0	1	0	4	5
736	devote	1	0	1	2	1	5
737	diagnose	1	0	1	0	3	5
738	disclose	5	0	0	0	0	5
739	disrupt	2	3	0	0	0	5
740	distribute	3	0	2	0	0	5
741	entertain	3	0	0	1	1	5
742	eradicate	1	1	1	0	2	5
743	exemplify	1	3	0	0	1	5
744	flow	2	0	2	0	1	5
745	fool	3	0	2	0	0	5
746	forbid	1	1	2	0	1	5
747	fund	2	0	1	0	2	5
748	gamble	2	0	1	0	2	5
749	grab	2	1	1	1	0	5
750	graduate	4	0	0	0	1	5
751	grasp	3	0	1	0	1	5
752	hire	5	0	0	0	0	5
753	import	3	0	1	0	1	5
754	impress	4	0	0	0	1	5
755	induce	3	0	0	0	2	5
756	inject	1	2	1	1	0	5
757	inspire	4	0	0	0	1	5
758	investigate	1	0	2	1	1	5
759	invoke	3	0	0	0	2	5
760	isolate	0	0	0	0	5	5
761	lift	3	0	0	0	2	5
762	manifest	1	2	0	0	2	5
763	modify	2	0	1	0	2	5

764	name	5	0	0	0	0	5
765	nominate	1	0	0	2	2	5
766	obey	3	0	1	1	0	5
767	occupy	2	1	1	0	1	5
768	parody	1	3	1	0	0	5
769	pillage	3	0	2	0	0	5
770	prejudice	4	0	0	0	1	5
771	print	3	0	0	0	2	5
772	profit	1	2	1	1	0	5
773	prosper	4	0	0	1	0	5
774	reason	0	1	3	1	0	5
775	reform	3	0	0	1	1	5
776	register	2	0	1	0	2	5
777	regret	5	0	0	0	0	5
778	reinstate	2	0	0	0	3	5
779	reside	2	2	0	1	0	5
780	resort	1	2	1	1	0	5
781	review	3	1	1	0	0	5
782	revive	2	0	1	1	1	5
783	sail	1	1	3	0	0	5
784	screen	1	0	3	0	1	5
785	shun	2	1	0	0	2	5
786	spark	1	2	0	0	2	5
787	split	0	0	2	0	3	5
788	spring	1	1	0	0	3	5
789	strip	2	0	1	0	2	5
790	tamper	2	1	2	0	0	5
791	touch	2	0	1	0	2	5
792	trace	1	1	0	1	2	5
793	transform	2	0	2	0	1	5
794	trap	2	0	0	0	3	5
795	unify	2	0	0	1	2	5
796	worsen	3	0	0	0	2	5
797	abort	2	0	0	0	2	4
798	administer	2	0	1	0	1	4
799	affirm	2	0	1	1	0	4
800	allocate	1	0	1	0	2	4
801	amount	0	3	1	0	0	4
802	award	0	0	0	0	4	4
803	bar	1	0	0	1	2	4
804	bond	2	0	1	1	0	4
805	bounce	1	0	1	0	2	4
806	breathe	1	0	3	0	0	4
807	broadcast	2	0	0	0	2	4
808	bury	2	0	0	0	2	4
809	cap	2	0	0	0	2	4
810	capture	0	0	0	0	4	4
811	cast	1	0	1	0	2	4

812	center	2	0	1	0	1	4
813	coin	1	0	0	2	1	4
814	coincide	2	0	1	1	0	4
815	comment	2	0	1	1	0	4
816	compel	0	0	0	0	4	4
817	confer	2	0	0	0	2	4
818	conflict	2	0	0	2	0	4
819	consent	1	1	0	1	1	4
820	contend	2	2	0	0	0	4
821	couple	0	0	0	0	4	4
822	crash	0	2	0	1	1	4
823	crush	1	0	0	0	3	4
824	culminate	1	1	1	1	0	4
825	dedicate	0	0	1	1	2	4
826	degrade	2	0	1	0	1	4
827	deter	2	1	0	0	1	4
828	dispute	4	0	0	0	0	4
829	distort	1	1	2	0	0	4
830	diversify	4	0	0	0	0	4
831	embody	1	1	0	0	2	4
832	emit	1	2	0	0	1	4
833	envisage	3	0	0	1	0	4
834	erase	2	0	0	1	1	4
835	erode	1	0	1	0	2	4
836	exacerbate	2	1	0	1	0	4
837	exhaust	0	1	0	0	3	4
838	extract	2	0	1	0	1	4
839	figure	3	0	0	1	0	4
840	flee	3	0	1	0	0	4
841	fuse	2	0	0	0	2	4
842	haunt	3	0	0	0	1	4
843	heat	1	0	1	0	2	4
844	honour	3	0	0	1	0	4
845	hook	2	0	0	0	2	4
846	house	2	1	1	0	0	4
847	implicate	1	1	0	1	1	4
848	interview	0	0	0	2	2	4
849	invade	2	0	0	2	0	4
850	lend	1	2	0	0	1	4
851	map	2	0	2	0	0	4
852	motivate	1	0	0	0	3	4
853	originate	2	1	0	1	0	4
854	override	2	1	1	0	0	4
855	owe	2	1	0	1	0	4
856	pack	2	0	1	0	1	4
857	perpetuate	3	1	0	0	0	4
858	plant	2	0	1	1	0	4
859	pool	1	0	2	1	0	4

860	postpone	3	1	0	0	0	4
861	practise	0	1	1	1	1	4
862	pretend	2	0	1	1	0	4
863	prolong	2	0	1	0	1	4
864	qualify	2	1	0	0	1	4
865	rate	3	1	0	0	0	4
866	rear	2	0	1	0	1	4
867	record	1	0	0	0	3	4
868	reign	1	1	0	2	0	4
869	relinquish	3	0	0	0	1	4
870	request	2	0	0	1	1	4
871	rescue	2	0	0	0	2	4
872	reserve	0	0	0	0	4	4
873	restore	3	0	0	0	1	4
874	revoke	2	0	0	1	1	4
875	segregate	1	0	1	1	1	4
876	shop	1	0	3	0	0	4
877	shorten	3	0	0	1	0	4
878	signify	2	2	0	0	0	4
879	smile	2	1	1	0	0	4
880	spare	1	0	0	0	3	4
881	speed	0	1	1	0	2	4
882	sponsor	1	0	1	0	2	4
883	stare	1	1	2	0	0	4
884	stumble	2	1	1	0	0	4
885	suspect	3	1	0	0	0	4
886	sway	2	0	2	0	0	4
887	tax	2	0	0	0	2	4
888	term	0	0	0	2	2	4
889	thrive	2	1	0	1	0	4
890	translate	1	1	0	0	2	4
891	uphold	4	0	0	0	0	4
892	warn	0	0	2	1	1	4
893	worship	4	0	0	0	0	4

Appendix 4: Making and editing a raw matched verb form list

In order for *WordList* to use its matching function and produce a raw matched word list, two or more lists to be matched should be made first. This may be done by the following steps:

(1) Open the WordList program.
(2) Choose the texts.
(3) Press the 'Make a word list now' bar.
(4) Save the list in 'lst' file in a proper folder.

To make another raw word list file, repeat the process.

When two raw word lists have been produced, the WordList program will be able to make a matched list. There are three options in making a matched list.

(1) A general match called Compare 2 WordList.
(2) A simple consistency match called Consistency (simple).
(3) A detailed consistency match called Consistency (detailed).

A general match, which is labelled 'Compare two wordlists', as the first option under the menu 'Comparison', compares two wordlists for disparity (keyness) and does not involve the task here. Therefore, no further explanation is needed.

A simple consistency, which is labelled 'Consistency (simple)' as the second option under the 'Comparison' menu provides a cursory look at the difference between the matched files with three columns, namely, "Word", "Frequency" and "Percentage". The "Word" column lists all the word forms found in the matched list; the "Frequency" column lists how many files each word form occurs in (either 1 or 2 in this case, not the frequency of the word form occurring in the corpus); and the "Percentage" column, based on the "Frequency" column, simply tells us the percentage of the number of the file or files in all the files compared (in this case either 50% or 100%, if one word form occurs only in one file, its frequency is 50%, and if one word form occurs in both of the files, its percentage is 100%). This simple function is useful for us to see how many word forms occur in both of the two corpora (especially when a detailed consistency match could not cope with a large number of word forms exceeding 16368 word

forms) and which word forms occur in only one of the corpora. But if we wish to have further information, the detailed consistency match must be used, which is Consistency (detailed) under the Comparison menu of WordList.

Compared with the simple match as described above, a detailed match provides more columns of information including the "Word" column (the word forms), the "Files" column (how many files are compared, like the "Frequency" column in the simple consistency match), the *COLEC* column (the first word list file) and the *LOCNESS* column (the second wordlist file), and the "TOTAL" column (the total frequency of the word form in all the files (COLEC and LOCNESS in this case). This function allows researchers to see the frequency of each word form, which word forms occur in both of the corpora and which verbs occur in only one of the corpora; if a word occurs in only one of them, then which corpus it is in, COLEC or LOCNESS. For the word forms occurring in both of the files (corpora), the colours of the frequency in the two files (COLEC and LOCNESS) are differentiated. The larger number is in red and the smaller number is in black. Unfortunately, WordSmith Tools (Version 3.0) (Scott 1999) can manage only 16368 word forms, which means word form numbers exceeding such a maximum will be cut off. The matched list as described above can be saved now with 'txt' file (say 'matchlist.txt') for further edition.

After the previous preparation, the list is ready for editing. The following steps are those I took for the editing process:

(1) Open the saved 'txt' file 'matchlist.txt' in MS Word.
(2) Find '_VVI' with the *Find what* box (quotations not included) and replace with '_V-i, new' in the *Replace with* box (allow a space between the comma and the word *new*), and save it as a 'txt' file (say V-i.txt).
(3) Open the 'V-i.txt' file with Excel (by checking the *space* and *comma* option while opening) and sort in descending order the column (C in this case) where 'new' appears (the underline is used here to refer to space).
(4) Delete all the lines which do not have 'new', then delete the 'new' column and then save it as an 'exl' file (say V-i.exl').
(5) Open the saved 'txt' file 'matchlist.txt' with MS Word as in (1).
(6) Find '_VV#' in the *Find what* box (quotations not included) and replace with '_V-e,

new' in the *Replace with* box (allow a space between the comma and the word *new*), and then save it with a 'txt' file (say V-e.txt).[40]

(7) Open the 'V-e.txt' file in Excel (by checking the *space* and *comma* option while opening) and sort in descending order the column (C in this case) where 'new' appears (the underline is used here to refer to space).

(8) Delete all the lines which do not have 'new' and then delete the 'new' column and save the file as an 'exl' file (say V-e.exl').

(9) Use the same principle to single out the remaining verb forms from the raw matched list 'matchlist.txt'.

40 _VV# refers to VV0. The number '0' is changed to '#' while the wordlist is produced by WordSmith (3.0) (Scott 1999). In the WordList setting, the 'Number include' box should be checked so that all the verbs tagged with '_VV0' are not missing.

Appendix 5: The verb forms that only occur in LOCNESS (f ≥ 4)

	Word	Total	V-e	V-i	V-s	V-ing	V-ed	V-n
1	ABANDONED	1	0	0	0	0	0	1
2	ABOLISH	1	0	1	0	0	0	0
3	ABOLISHED	1	0	0	0	0	0	1
4	ABOLISHING	1	0	0	0	1	0	0
5	ABUSE	1	0	1	0	0	0	0
6	ABUSED	1	0	0	0	0	0	1
7	ACCEPTS	1	0	0	1	0	0	0
8	ACCOMMODATE	1	0	1	0	0	0	0
9	ACCUSES	1	0	0	1	0	0	0
10	ACHIEVES	1	0	0	1	0	0	0
11	ACHIEVING	1	0	0	0	1	0	0
12	ACKNOWLEDGING	1	0	0	0	1	0	0
13	ACTS	1	0	0	1	0	0	0
14	ADDICTED	1	0	0	0	0	0	1
15	ADDRESS	1	0	1	0	0	0	0
16	ADDRESSED	1	0	0	0	0	0	1
17	ADDRESSING	1	0	0	0	1	0	0
18	ADHERED	1	0	0	0	0	0	1
19	ADMIRED	1	0	0	0	0	0	1
20	ADMIT	1	0	1	0	0	0	0
21	ADMITS	1	0	0	1	0	0	0
22	ADMITTED	1	0	0	0	0	1	0
23	ADOPT	1	1	0	0	0	0	0
24	ADOPTS	1	0	0	1	0	0	0
25	ADVOCATING	1	0	0	0	1	0	0
26	AFFECTING	1	0	0	0	1	0	0
27	AFFLICTED	1	0	0	0	0	0	1
28	AID	1	0	1	0	0	0	0
29	AIMED	1	0	0	0	0	1	0
30	AIMS	1	0	0	1	0	0	0
31	ALLEVIATE	1	0	1	0	0	0	0
32	ALLOW	1	1	0	0	0	0	0
33	ALLOWED	1	0	0	0	0	1	0
34	ALLOWING	1	0	0	0	1	0	0
35	ALLOWS	1	0	0	1	0	0	0
36	ALTERED	1	0	0	0	0	0	1
37	ANALYZED	1	0	0	0	0	0	1
38	ANALYZING	1	0	0	0	1	0	0
39	ANNOUNCES	1	0	0	1	0	0	0
40	ANSWERED	1	0	0	0	0	0	1
41	APPLYING	1	0	0	0	1	0	0
42	APPOINT	1	0	1	0	0	0	0

43 APPOINTED	1	0	0	0	0	0	1
44 ARGUES	1	0	0	1	0	0	0
45 ARGUING	1	0	0	0	1	0	0
46 ARISES	1	0	0	1	0	0	0
47 AROSE	1	0	0	0	0	1	0
48 ARRESTED	1	0	0	0	0	0	1
49 ASSEMBLE	1	1	0	0	0	0	0
50 ASSERT	1	0	1	0	0	0	0
51 ASSERTS	1	0	0	1	0	0	0
52 ASSUME	1	0	1	0	0	0	0
53 ASSUMES	1	0	0	1	0	0	0
54 ASSUMING	1	0	0	0	1	0	0
55 ATTACKING	1	0	0	0	1	0	0
56 ATTACKS	1	0	0	1	0	0	0
57 ATTEMPT	1	0	1	0	0	0	0
58 ATTEMPTED	1	0	0	0	0	1	0
59 ATTEMPTING	1	0	0	0	1	0	0
60 ATTEMPTS	1	0	0	1	0	0	0
61 ATTRACT	1	0	1	0	0	0	0
62 ATTRIBUTES	1	0	0	1	0	0	0
63 AWARDED	1	0	0	0	0	0	1
64 BACK	1	0	1	0	0	0	0
65 BALANCE	1	0	1	0	0	0	0
66 BANNING	1	0	0	0	1	0	0
67 BASE	1	1	0	0	0	0	0
68 BEARING	1	0	0	0	1	0	0
69 BEATING	1	0	0	0	1	0	0
70 BEHAVE	1	0	1	0	0	0	0
71 BELIEVING	1	0	0	0	1	0	0
72 BELONGING	1	0	0	0	1	0	0
73 BENEFITTED	1	0	0	0	0	0	1
74 BETRAYED	1	0	0	0	0	0	1
75 BETTING	1	0	0	0	1	0	0
76 BINGE	1	1	0	0	0	0	0
77 BLAMED	1	0	0	0	0	0	1
78 BLAMING	1	0	0	0	1	0	0
79 BLOWN	1	0	0	0	0	0	1
80 BOUND	1	0	0	0	0	0	1
81 BRED	1	0	0	0	0	0	1
82 CALCULATE	1	0	1	0	0	0	0
83 CAPTURED	1	0	0	0	0	0	1
84 CEASE	1	0	1	0	0	0	0
85 CEDED	1	0	0	0	0	0	1
86 CENSORED	1	0	0	0	0	0	1
87 CENTRED	1	0	0	0	0	0	1
88 CHALLENGED	1	0	0	0	0	0	1

89	CHASED	1	0	0	0	0	0	1
90	CHASING	1	0	0	0	1	0	0
91	CHOOSES	1	0	0	1	0	0	0
92	CITE	1	1	0	0	0	0	0
93	CLAIMED	1	0	0	0	0	1	0
94	CLAIMING	1	0	0	0	1	0	0
95	CLAIMS	1	0	0	1	0	0	0
96	CLARIFY	1	0	1	0	0	0	0
97	CLASSIFIED	1	0	0	0	0	0	1
98	COMBAT	1	0	1	0	0	0	0
99	COMFORT	1	1	0	0	0	0	0
100	COMMITS	1	0	0	1	0	0	0
101	COMMITTING	1	0	0	0	1	0	0
102	COMPARES	1	0	0	1	0	0	0
103	COMPELLED	1	0	0	0	0	0	1
104	COMPENSATE	1	0	1	0	0	0	0
105	COMPETE	1	1	0	0	0	0	0
106	COMPREHEND	1	0	1	0	0	0	0
107	COMPROMISE	1	0	1	0	0	0	0
108	CONCENTRATED	1	0	0	0	0	0	1
109	CONCIEVE	1	0	1	0	0	0	0
110	CONCIEVED	1	0	0	0	0	0	1
111	CONDEMN	1	0	1	0	0	0	0
112	CONDEMNED	1	0	0	0	0	0	1
113	CONDEMNING	1	0	0	0	1	0	0
114	CONFESS	1	0	1	0	0	0	0
115	CONFESSES	1	0	0	1	0	0	0
116	CONFESSING	1	0	0	0	1	0	0
117	CONFRONTED	1	0	0	0	0	0	1
118	CONFRONTS	1	0	0	1	0	0	0
119	CONSTITUTES	1	0	0	1	0	0	0
120	CONSTRUED	1	0	0	0	0	0	1
121	CONSUMING	1	0	0	0	1	0	0
122	CONTAIN	1	0	1	0	0	0	0
123	CONTAINING	1	0	0	0	1	0	0
124	CONTINUED	1	0	0	0	0	0	1
125	CONTRACT	1	0	1	0	0	0	0
126	CONTRACTING	1	0	0	0	1	0	0
127	CONTROLS	1	0	0	1	0	0	0
128	CONVEY	1	0	1	0	0	0	0
129	CONVEYED	1	0	0	0	0	0	1
130	CONVICTED	1	0	0	0	0	0	1
131	CONVINCE	1	0	1	0	0	0	0
132	COUNTER	1	0	1	0	0	0	0
133	COUNTERACT	1	0	1	0	0	0	0
134	COUPLED	1	0	0	0	0	0	1

135	CREATE	1	1	0	0	0	0	0
136	CREATED	1	0	0	0	0	1	0
137	CREATES	1	0	0	1	0	0	0
138	CREATING	1	0	0	0	1	0	0
139	CRIES	1	0	0	1	0	0	0
140	CRITICISED	1	0	0	0	0	0	1
141	CRITICIZE	1	0	1	0	0	0	0
142	CROSSING	1	0	0	0	1	0	0
143	DAMAGING	1	0	0	0	1	0	0
144	DATING	1	0	0	0	1	0	0
145	DEALS	1	0	0	1	0	0	0
146	DEALT	1	0	0	0	0	0	1
147	DEBATED	1	0	0	0	0	0	1
148	DECIDES	1	0	0	1	0	0	0
149	DECIDING	1	0	0	0	1	0	0
150	DEEMED	1	0	0	0	0	0	1
151	DEFEND	1	1	0	0	0	0	0
152	DEFENDING	1	0	0	0	1	0	0
153	DEFINES	1	0	0	1	0	0	0
154	DELIVERED	1	0	0	0	0	0	1
155	DEMANDED	1	0	0	0	0	1	0
156	DEMONSTRATE	1	1	0	0	0	0	0
157	DEMONSTRATED	1	0	0	0	0	0	1
158	DEMONSTRATES	1	0	0	1	0	0	0
159	DENIED	1	0	0	0	0	0	1
160	DENIES	1	0	0	1	0	0	0
161	DENY	1	1	0	0	0	0	0
162	DENYING	1	0	0	0	1	0	0
163	DEPICTS	1	0	0	1	0	0	0
164	DEPRIVING	1	0	0	0	1	0	0
165	DERIVED	1	0	0	0	0	0	1
166	DESCRIBES	1	0	0	1	0	0	0
167	DESCRIBING	1	0	0	0	1	0	0
168	DESERVES	1	0	0	1	0	0	0
169	DESIGNED	1	0	0	0	0	0	1
170	DESIRES	1	0	0	1	0	0	0
171	DESTINED	1	0	0	0	0	0	1
172	DESTROYING	1	0	0	0	1	0	0
173	DETACHED	1	0	0	0	0	0	1
174	DETECTED	1	0	0	0	0	0	1
175	DETERMINE	1	0	1	0	0	0	0
176	DETERMINES	1	0	0	1	0	0	0
177	DETERMINING	1	0	0	0	1	0	0
178	DIFFER	1	1	0	0	0	0	0
179	DIMINISH	1	0	1	0	0	0	0
180	DIMINISHED	1	0	0	0	0	0	1

181	DIRECTED	1	0	0	0	0	0	1
182	DISAGREE	1	1	0	0	0	0	0
183	DISAGREES	1	0	0	1	0	0	0
184	DISCOVERED	1	0	0	0	0	0	1
185	DISCOVERING	1	0	0	0	1	0	0
186	DISCOVERS	1	0	0	1	0	0	0
187	DISCRIMINATED	1	0	0	0	0	0	1
188	DISCUSS	1	1	0	0	0	0	0
189	DISCUSSES	1	0	0	1	0	0	0
190	DISMISS	1	0	1	0	0	0	0
191	DISMISSED	1	0	0	0	0	0	1
192	DISPLAYS	1	0	0	1	0	0	0
193	DISPUTE	1	0	1	0	0	0	0
194	DISREGARDED	1	0	0	0	0	0	1
195	DISSOLVE	1	0	1	0	0	0	0
196	DIVERSIFY	1	0	1	0	0	0	0
197	DIVIDE	1	0	1	0	0	0	0
198	DIVORCE	1	0	1	0	0	0	0
199	DOMINATE	1	0	1	0	0	0	0
200	DONATED	1	0	0	0	0	0	1
201	DRAFTED	1	0	0	0	0	0	1
202	DRAWS	1	0	0	1	0	0	0
203	DREAM	1	0	1	0	0	0	0
204	DRESSED	1	0	0	0	0	0	1
205	EARNS	1	0	0	1	0	0	0
206	EASE	1	0	1	0	0	0	0
207	ELECTED	1	0	0	0	0	0	1
208	ELIMINATED	1	0	0	0	0	0	1
209	ELIMINATING	1	0	0	0	1	0	0
210	EMBRACING	1	0	0	0	1	0	0
211	EMPHASISED	1	0	0	0	0	0	1
212	EMPHASIZES	1	0	0	1	0	0	0
213	EMPLOY	1	0	1	0	0	0	0
214	EMPLOYED	1	0	0	0	0	0	1
215	ENABLED	1	0	0	0	0	0	1
216	ENCOUNTERED	1	0	0	0	0	0	1
217	ENCOUNTERS	1	0	0	1	0	0	0
218	ENCOURAGING	1	0	0	0	1	0	0
219	END	1	1	0	0	0	0	0
220	ENDURES	1	0	0	1	0	0	0
221	ENFORCED	1	0	0	0	0	0	1
222	ENHANCING	1	0	0	0	1	0	0
223	ENJOYED	1	0	0	0	0	0	1
224	ENSURING	1	0	0	0	1	0	0
225	ENTAILS	1	0	0	1	0	0	0
226	ENTERS	1	0	0	1	0	0	0

227	EPITOMISES	1	0	0	1	0	0	0
228	ESPOUSES	1	0	0	1	0	0	0
229	ESTABLISHED	1	0	0	0	0	1	0
230	EVOKE	1	0	1	0	0	0	0
231	EVOKED	1	0	0	0	0	0	1
232	EVOKES	1	0	0	1	0	0	0
233	EVOLVED	1	0	0	0	0	0	1
234	EXECUTED	1	0	0	0	0	0	1
235	EXISTED	1	0	0	0	0	0	1
236	EXPECTING	1	0	0	0	1	0	0
237	EXPLAINING	1	0	0	0	1	0	0
238	EXPOSED	1	0	0	0	0	0	1
239	EXPOSES	1	0	0	1	0	0	0
240	EXPOSING	1	0	0	0	1	0	0
241	EXPRESSES	1	0	0	1	0	0	0
242	EXPRESSING	1	0	0	0	1	0	0
243	EXTENDED	1	0	0	0	0	0	1
244	FACILITATE	1	0	1	0	0	0	0
245	FALLS	1	0	0	1	0	0	0
246	FAVOURED	1	0	0	0	0	0	1
247	FEARED	1	0	0	0	0	1	0
248	FEARS	1	0	0	1	0	0	0
249	FED	1	0	0	0	0	0	1
250	FIGHT	1	1	0	0	0	0	0
251	FIGHTS	1	0	0	1	0	0	0
252	FILED	1	0	0	0	0	0	1
253	FLOGGED	1	0	0	0	0	0	1
254	FOCUSED	1	0	0	0	0	0	1
255	FOCUSES	1	0	0	1	0	0	0
256	FOCUSING	1	0	0	0	1	0	0
257	FORCING	1	0	0	0	1	0	0
258	FORGIVE	1	0	1	0	0	0	0
259	FORMING	1	0	0	0	1	0	0
260	FORMS	1	0	0	1	0	0	0
261	FREE	1	0	1	0	0	0	0
262	FREEING	1	0	0	0	1	0	0
263	FRIGHTENED	1	0	0	0	0	0	1
264	FULFILL	1	0	1	0	0	0	0
265	FULFILLED	1	0	0	0	0	0	1
266	FULFILLING	1	0	0	0	1	0	0
267	FUNCTIONING	1	0	0	0	1	0	0
268	FURTHERING	1	0	0	0	1	0	0
269	GOVERN	1	0	1	0	0	0	0
270	GOVERNED	1	0	0	0	0	0	1
271	GUARANTEE	1	0	1	0	0	0	0
272	GUARANTEED	1	0	0	0	0	0	1

273	GUARANTEES	1	0	0	1	0	0	0
274	GUESS	1	1	0	0	0	0	0
275	HANGED	1	0	0	0	0	0	1
276	HEADED	1	0	0	0	0	0	1
277	HEARS	1	0	0	1	0	0	0
278	HELPED	1	0	0	0	0	0	1
279	HIDDEN	1	0	0	0	0	0	1
280	HIDE	1	0	1	0	0	0	0
281	HIDING	1	0	0	0	1	0	0
282	HINDER	1	1	0	0	0	0	0
283	HIRE	1	0	1	0	0	0	0
284	HITS	1	0	0	1	0	0	0
285	HUNG	1	0	0	0	0	0	1
286	HURTING	1	0	0	0	1	0	0
287	ILLUSTRATE	1	1	0	0	0	0	0
288	ILLUSTRATED	1	0	0	0	0	0	1
289	ILLUSTRATES	1	0	0	1	0	0	0
290	IMPLANTED	1	0	0	0	0	0	1
291	IMPLEMENT	1	0	1	0	0	0	0
292	IMPLEMENTED	1	0	0	0	0	0	1
293	IMPLY	1	0	1	0	0	0	0
294	IMPOSE	1	0	1	0	0	0	0
295	IMPOSED	1	0	0	0	0	0	1
296	INCARCERATED	1	0	0	0	0	0	1
297	INCORPORATED	1	0	0	0	0	0	1
298	INFORM	1	0	1	0	0	0	0
299	INHIBIT	1	0	1	0	0	0	0
300	INSISTS	1	0	0	1	0	0	0
301	INSPIRE	1	0	1	0	0	0	0
302	INSTALLED	1	0	0	0	0	0	1
303	INSTITUTED	1	0	0	0	0	0	1
304	INSTRUCTED	1	0	0	0	0	0	1
305	INTEGRATE	1	0	1	0	0	0	0
306	INTEGRATED	1	0	0	0	0	0	1
307	INTEGRATING	1	0	0	0	1	0	0
308	INTENDS	1	0	0	1	0	0	0
309	INTERACT	1	1	0	0	0	0	0
310	INTERFERE	1	0	1	0	0	0	0
311	INTERPRETED	1	0	0	0	0	0	1
312	INTERVENE	1	0	1	0	0	0	0
313	INTRODUCED	1	0	0	0	0	0	1
314	INTRODUCING	1	0	0	0	1	0	0
315	INVESTED	1	0	0	0	0	0	1
316	INVESTING	1	0	0	0	1	0	0
317	INVOLVED	1	0	0	0	0	1	0
318	INVOLVES	1	0	0	1	0	0	0

319	INVOLVING	1	0	0	0	1	0	0
320	ISOLATED	1	0	0	0	0	0	1
321	JUDGE	1	1	0	0	0	0	0
322	JUDGES	1	0	0	1	0	0	0
323	JUSTIFIED	1	0	0	0	0	0	1
324	JUSTIFIES	1	0	0	1	0	0	0
325	KISSING	1	0	0	0	1	0	0
326	LABELED	1	0	0	0	0	0	1
327	LEGALISED	1	0	0	0	0	0	1
328	LEGALIZE	1	0	1	0	0	0	0
329	LEGALIZED	1	0	0	0	0	0	1
330	LEGALIZING	1	0	0	0	1	0	0
331	LEGISLATE	1	0	1	0	0	0	0
332	LESSEN	1	0	1	0	0	0	0
333	LETTING	1	0	0	0	1	0	0
334	LOCATED	1	0	0	0	0	0	1
335	LOOSE	1	0	1	0	0	0	0
336	LOWERING	1	0	0	0	1	0	0
337	MAINTAINING	1	0	0	0	1	0	0
338	MAINTAINS	1	0	0	1	0	0	0
339	MANAGED	1	0	0	0	0	0	1
340	MANAGES	1	0	0	1	0	0	0
341	MANIPULATE	1	0	1	0	0	0	0
342	MANIPULATED	1	0	0	0	0	0	1
343	MANIPULATING	1	0	0	0	1	0	0
344	MARRIED	1	0	0	0	0	0	1
345	MATTER	1	0	1	0	0	0	0
346	MEANT	1	0	0	0	0	0	1
347	MEASURED	1	0	0	0	0	0	1
348	MENTION	1	0	1	0	0	0	0
349	MENTIONS	1	0	0	1	0	0	0
350	MISUSED	1	0	0	0	0	0	1
351	MIX	1	0	1	0	0	0	0
352	MOCK	1	0	1	0	0	0	0
353	MONITORED	1	0	0	0	0	0	1
354	MOVE	1	1	0	0	0	0	0
355	MOVES	1	0	0	1	0	0	0
356	MURDER	1	0	1	0	0	0	0
357	MURDERED	1	0	0	0	0	0	1
358	MURDERING	1	0	0	0	1	0	0
359	NEGOTIATE	1	0	1	0	0	0	0
360	NOTE	1	0	1	0	0	0	0
361	NOTED	1	0	0	0	0	0	1
362	NURTURING	1	0	0	0	1	0	0
363	OBSERVED	1	0	0	0	0	0	1
364	OBTAINING	1	0	0	0	1	0	0

365	OCCURRED	1	0	0	0	0	0	1
366	OFFERING	1	0	0	0	1	0	0
367	OFFSET	1	0	1	0	0	0	0
368	OPENS	1	0	0	1	0	0	0
369	OPPOSE	1	1	0	0	0	0	0
370	OPPOSED	1	0	0	0	0	0	1
371	OPPOSING	1	0	0	0	1	0	0
372	OUTLAWED	1	0	0	0	0	0	1
373	OUTLINED	1	0	0	0	0	0	1
374	OUTWEIGH	1	1	0	0	0	0	0
375	OUTWEIGHS	1	0	0	1	0	0	0
376	OVERLOOK	1	0	1	0	0	0	0
377	OVERLOOKED	1	0	0	0	0	0	1
378	PERCEIVED	1	0	0	0	0	0	1
379	PERFORMING	1	0	0	0	1	0	0
380	PERSISTS	1	0	0	1	0	0	0
381	PERSUADE	1	0	1	0	0	0	0
382	PERSUADED	1	0	0	0	0	0	1
383	PICKS	1	0	0	1	0	0	0
384	PLACES	1	0	0	1	0	0	0
385	PLACING	1	0	0	0	1	0	0
386	PLANNED	1	0	0	0	0	1	0
387	POINT	1	0	1	0	0	0	0
388	PORTRAYED	1	0	0	0	0	0	1
389	PORTRAYING	1	0	0	0	1	0	0
390	PORTRAYS	1	0	0	1	0	0	0
391	POSE	1	0	1	0	0	0	0
392	POSSESSING	1	0	0	0	1	0	0
393	PRAY	1	0	1	0	0	0	0
394	PRAYING	1	0	0	0	1	0	0
395	PREACHES	1	0	0	1	0	0	0
396	PRESCRIBED	1	0	0	0	0	0	1
397	PRESENTING	1	0	0	0	1	0	0
398	PREVAIL	1	0	1	0	0	0	0
399	PROCEEDED	1	0	0	0	0	1	0
400	PROCESS	1	0	1	0	0	0	0
401	PROCESSED	1	0	0	0	0	0	1
402	PROCLAIMS	1	0	0	1	0	0	0
403	PROGRAMMED	1	0	0	0	0	0	1
404	PROGRESS	1	0	1	0	0	0	0
405	PROJECT	1	0	1	0	0	0	0
406	PROMISED	1	0	0	0	0	1	0
407	PROMOTES	1	0	0	1	0	0	0
408	PROMOTING	1	0	0	0	1	0	0
409	PROPOSE	1	1	0	0	0	0	0
410	PROPOSED	1	0	0	0	0	0	1

411	PROVEN	1	0	0	0	0	0	1
412	PROVING	1	0	0	0	1	0	0
413	PROVOKES	1	0	0	1	0	0	0
414	PUBLICISED	1	0	0	0	0	0	1
415	PUBLISHED	1	0	0	0	0	0	1
416	PURCHASED	1	0	0	0	0	0	1
417	PURCHASING	1	0	0	0	1	0	0
418	PUSHED	1	0	0	0	0	0	1
419	PUSHES	1	0	0	1	0	0	0
420	QUESTIONING	1	0	0	0	1	0	0
421	RACING	1	0	0	0	1	0	0
422	RAISES	1	0	0	1	0	0	0
423	RANGING	1	0	0	0	1	0	0
424	RANKED	1	0	0	0	0	0	1
425	RAPED	1	0	0	0	0	0	1
426	REACHING	1	0	0	0	1	0	0
427	REALISES	1	0	0	1	0	0	0
428	REALISING	1	0	0	0	1	0	0
429	REALIZES	1	0	0	1	0	0	0
430	REBELLING	1	0	0	0	1	0	0
431	RECEIVES	1	0	0	1	0	0	0
432	RECIEVE	1	0	1	0	0	0	0
433	RECOGNISED	1	0	0	0	0	0	1
434	RECOGNISES	1	0	0	1	0	0	0
435	RECOGNIZES	1	0	0	1	0	0	0
436	RECONCILE	1	0	1	0	0	0	0
437	RECYCLING	1	0	0	0	1	0	0
438	REFERRED	1	0	0	0	0	0	1
439	REFERRING	1	0	0	0	1	0	0
440	REFLECT	1	0	1	0	0	0	0
441	REFLECTING	1	0	0	0	1	0	0
442	REFLECTS	1	0	0	1	0	0	0
443	REFUSES	1	0	0	1	0	0	0
444	REFUSING	1	0	0	0	1	0	0
445	REFUTED	1	0	0	0	0	0	1
446	REGAIN	1	0	1	0	0	0	0
447	REGARDING	1	0	0	0	1	0	0
448	REGULATE	1	0	1	0	0	0	0
449	REGULATING	1	0	0	0	1	0	0
450	REINFORCED	1	0	0	0	0	0	1
451	REINFORCES	1	0	0	1	0	0	0
452	REJECTING	1	0	0	0	1	0	0
453	REJECTS	1	0	0	1	0	0	0
454	RELATES	1	0	0	1	0	0	0
455	RELATING	1	0	0	0	1	0	0
456	RELY	1	1	0	0	0	0	0

457	REMAINED	1	0	0	0	0	1	0
458	REMAINING	1	0	0	0	1	0	0
459	REMAINS	1	0	0	1	0	0	0
460	REMARKS	1	0	0	1	0	0	0
461	REMEDY	1	0	1	0	0	0	0
462	REMOVE	1	0	1	0	0	0	0
463	REMOVED	1	0	0	0	0	0	1
464	RENEWED	1	0	0	0	0	0	1
465	REPEAL	1	0	1	0	0	0	0
466	REPENT	1	0	1	0	0	0	0
467	REPENTING	1	0	0	0	1	0	0
468	REPLACED	1	0	0	0	0	0	1
469	REPLACING	1	0	0	0	1	0	0
470	REPLIES	1	0	0	1	0	0	0
471	REPORTED	1	0	0	0	0	1	0
472	REPRESENT	1	0	1	0	0	0	0
473	REPRESENTED	1	0	0	0	0	0	1
474	REPRESENTING	1	0	0	0	1	0	0
475	REPRESENTS	1	0	0	1	0	0	0
476	REPRODUCE	1	0	1	0	0	0	0
477	RESENT	1	1	0	0	0	0	0
478	RESERVED	1	0	0	0	0	0	1
479	RESIGNED	1	0	0	0	0	1	0
480	RESPOND	1	1	0	0	0	0	0
481	RESPONDS	1	0	0	1	0	0	0
482	RESTRICT	1	1	0	0	0	0	0
483	RESTRICTED	1	0	0	0	0	0	1
484	RESULTING	1	0	0	0	1	0	0
485	RETAIN	1	0	1	0	0	0	0
486	RETAINED	1	0	0	0	0	0	1
487	RETAINING	1	0	0	0	1	0	0
488	RETAINS	1	0	0	1	0	0	0
489	RETIRE	1	0	1	0	0	0	0
490	RETURNED	1	0	0	0	0	0	1
491	RETURNING	1	0	0	0	1	0	0
492	REUNITED	1	0	0	0	0	0	1
493	REVEALED	1	0	0	0	0	0	1
494	REVEALING	1	0	0	0	1	0	0
495	REVEALS	1	0	0	1	0	0	0
496	REVERSE	1	0	1	0	0	0	0
497	REVOLT	1	0	1	0	0	0	0
498	RIDICULED	1	0	0	0	0	0	1
499	RIPPED	1	0	0	0	0	0	1
500	RUIN	1	0	1	0	0	0	0
501	RULE	1	0	1	0	0	0	0
502	SACK	1	0	1	0	0	0	0

503	SACRIFICE	1	0	1	0	0	0	0
504	SACRIFICES	1	0	0	1	0	0	0
505	SAFEGUARD	1	0	1	0	0	0	0
506	SCORE	1	0	1	0	0	0	0
507	SECURE	1	0	1	0	0	0	0
508	SEES	1	0	0	1	0	0	0
509	SENDS	1	0	0	1	0	0	0
510	SENTENCED	1	0	0	0	0	0	1
511	SEPARATE	1	1	0	0	0	0	0
512	SEPARATED	1	0	0	0	0	0	1
513	SEPARATING	1	0	0	0	1	0	0
514	SHAKEN	1	0	0	0	0	0	1
515	SHAPING	1	0	0	0	1	0	0
516	SHARE	1	1	0	0	0	0	0
517	SHARED	1	0	0	0	0	0	1
518	SHARING	1	0	0	0	1	0	0
519	SHATTERED	1	0	0	0	0	0	1
520	SHOT	1	0	0	0	0	0	1
521	SHOUT	1	0	1	0	0	0	0
522	SIGN	1	0	1	0	0	0	0
523	SLAUGHTERED	1	0	0	0	0	0	1
524	SLEEPS	1	0	0	1	0	0	0
525	SOUND	1	0	1	0	0	0	0
526	SPREAD	1	0	0	0	0	0	1
527	STARTS	1	0	0	1	0	0	0
528	STATED	1	0	0	0	0	1	0
529	STATES	1	0	0	1	0	0	0
530	STATING	1	0	0	0	1	0	0
531	STAYED	1	0	0	0	0	0	1
532	STAYS	1	0	0	1	0	0	0
533	STEM	1	0	1	0	0	0	0
534	STEMS	1	0	0	1	0	0	0
535	STOLE	1	0	0	0	0	1	0
536	STOPPED	1	0	0	0	0	0	1
537	STOPS	1	0	0	1	0	0	0
538	STRENGTHENING	1	0	0	0	1	0	0
539	STRIVING	1	0	0	0	1	0	0
540	STRUCK	1	0	0	0	0	0	1
541	STUCK	1	0	0	0	0	0	1
542	SUBJECTED	1	0	0	0	0	0	1
543	SUBMIT	1	0	1	0	0	0	0
544	SUED	1	0	0	0	0	0	1
545	SUGGEST	1	0	1	0	0	0	0
546	SUPPORTING	1	0	0	0	1	0	0
547	SUPPORTS	1	0	0	1	0	0	0
548	SURROUNDED	1	0	0	0	0	0	1

549	SURVIVED	1	0	0	0	0	0	1
550	SYMBOLISED	1	0	0	0	0	0	1
551	SYMPATHISE	1	0	1	0	0	0	0
552	TACKLE	1	0	1	0	0	0	0
553	TACKLED	1	0	0	0	0	0	1
554	TACKLES	1	0	0	1	0	0	0
555	TALKED	1	0	0	0	0	0	1
556	TALKS	1	0	0	1	0	0	0
557	TELEVISE	1	0	1	0	0	0	0
558	TEND	1	0	1	0	0	0	0
559	THANK	1	0	1	0	0	0	0
560	THROWS	1	0	0	1	0	0	0
561	TITLED	1	0	0	0	0	0	1
562	TORN	1	0	0	0	0	0	1
563	TRAINS	1	0	0	1	0	0	0
564	TRANSFERRED	1	0	0	0	0	0	1
565	TRANSMIT	1	0	1	0	0	0	0
566	TRANSMITTED	1	0	0	0	0	0	1
567	TRANSMITTING	1	0	0	0	1	0	0
568	TRANSPORT	1	0	1	0	0	0	0
569	TRANSPORTED	1	0	0	0	0	0	1
570	TRAVEL	1	1	0	0	0	0	0
571	TREATS	1	0	0	1	0	0	0
572	TRUSTED	1	0	0	0	0	0	1
573	UNDERGOES	1	0	0	1	0	0	0
574	UNDERGONE	1	0	0	0	0	0	1
575	UNDERMINED	1	0	0	0	0	0	1
576	UNDERTAKEN	1	0	0	0	0	0	1
577	UPHOLD	1	0	1	0	0	0	0
578	UTILIZED	1	0	0	0	0	0	1
579	VALUED	1	0	0	0	0	0	1
580	VETO	1	0	1	0	0	0	0
581	VIEW	1	1	0	0	0	0	0
582	VIEWING	1	0	0	0	1	0	0
583	VOICE	1	0	1	0	0	0	0
584	VOTE	1	0	1	0	0	0	0
585	VOTED	1	0	0	0	0	0	1
586	VOTING	1	0	0	0	1	0	0
587	WAKE	1	0	1	0	0	0	0
588	WEAKENED	1	0	0	0	0	0	1
589	WEIGH	1	0	1	0	0	0	0
590	WHIPPED	1	0	0	0	0	0	1
591	WISHES	1	0	0	1	0	0	0
592	WITNESS	1	0	1	0	0	0	0
593	WITNESSING	1	0	0	0	1	0	0
594	WORSHIP	1	0	1	0	0	0	0

595 WRITES	1	0	0	1	0	0	0
596 ACCUSED	2	0	0	0	0	1	1
597 ARGUE	2	1	1	0	0	0	0
598 ARGUED	2	0	0	0	0	1	1
599 ASSIST	2	1	1	0	0	0	0
600 ASSUMED	2	0	0	0	0	1	1
601 BANNED	2	0	0	0	0	1	1
602 COMMITTED	2	0	0	0	0	1	1
603 CONDUCTED	2	0	0	0	0	1	1
604 CONTRACTED	2	0	0	0	0	1	1
605 DEFINE	2	1	1	0	0	0	0
606 DEFINED	2	0	0	0	0	1	1
607 DESCRIBED	2	0	0	0	0	1	1
608 DESERVE	2	1	1	0	0	0	0
609 EMPHASIZE	2	1	1	0	0	0	0
610 EXPRESSED	2	0	0	0	0	1	1
611 FOUGHT	2	0	0	0	0	1	1
612 GATHER	2	1	1	0	0	0	0
613 INCLUDED	2	0	0	0	0	1	1
614 INTENDED	2	0	0	0	0	1	1
615 JUSTIFY	2	1	1	0	0	0	0
616 MARRY	2	1	1	0	0	0	0
617 PERCEIVE	2	1	1	0	0	0	0
618 PORTRAY	2	1	1	0	0	0	0
619 PRESENT	2	1	1	0	0	0	0
620 PRESENTED	2	0	0	0	0	1	1
621 QUESTION	2	1	1	0	0	0	0
622 QUESTIONED	2	0	0	0	0	1	1
623 REALISED	2	0	0	0	0	1	1
624 RECOGNIZED	2	0	0	0	0	1	1
625 REFUTE	2	1	1	0	0	0	0
626 REJECTED	2	0	0	0	0	1	1
627 REPLACE	2	1	1	0	0	0	0
628 RULED	2	0	0	0	0	1	1
629 SIGNED	2	0	0	0	0	1	1
630 SOUGHT	2	0	0	0	0	1	1
631 STATE	2	1	1	0	0	0	0
632 VIEWED	2	0	0	0	0	1	1
633 WITNESSED	2	0	0	0	0	1	1

Appendix 6: The three steps I took in making a collocation list

1. Open all the files of the corpus with Concord and re-sort by *Centre* first, *IR* (first position on the right of the node) second and *2R* (second position on the right of the node) third.
2. After the re-sorting, it is much easier to see all the identical collocations. To discriminate between different collocations, a code can be attached to each different type of collocation by typing a letter (from a to z and from A to Z) in the "Set" column. There are 54 codes available for attaching such information in version 3.0 of WordSmith.[41]
3. After all the concordance lines are encoded, use the *Re-sort* function again so that the same type of collocations can be grouped together.

41 In WordSmith (Version 4.0), more codes are available including the use of numbers from 0 to 9 (cf. Scott 2004).

Appendix 7: The concordances of 'V *up*' in LOCNESS

olicies in as a lawyer seems to back up this theme . < NN of true informat
e main claim using statements that back up their reasoning supportive reasoning
reasoning supportive reasoning to back up the Civil Liberties not got a major
t they belie fs to cause them to back up their own side and his wife were beat
e . < NN of true information to back up what they belie fs to cause them to
Liberties not got a majority to back up his policies in as a lawyer seems
ance ument become stronger by backing up the main claim using statements that
verely by a raped , kidnapped , or beat up . </s> <s> This is hquake . </s>
their own side and his wife were beaten up severely by a raped , kidnapped , or
is hquake . </s> <s> The sea boiled up , smashing vessel J to show emotion
vessel J to show emotions , so bottle up all their feel . </s> <s> It need
eel . </s> <s> It needs to brighten up and increase i able to cope fully w
later i this really the age to bring up a teenager ? in when the CFTC b
point that murder case do besides bring up a moral debate research one should br
cs which i e stressful it is to bring up children later i this really the ag
moral debate research one should bring up eugenics which i e stressful it is
f . </s> <s> They constantly bring up the point that murder case do besides
se i able to cope fully with bringing up a baby . </s> <s> The oppositio
made I Simpson 's Lawyers , ″ brings up the past relation finish judicial
</s> <s> However , Sinsheimer brings up the question of living . </s> <s> Th
s he the Civil Liberties Group brings up is the amendment <s> A third reason
l AT divorcee . </s> <s> This brings up the issue of . </s> <s> They con
eory tes <*> . </s> <s> This brings up information control AT divorcee .
stion of living . </s> <s> This brings up another theory tes <*> . </s> <s>
dment <s> A third reason Lewis brings up is the question a fact that Sherman b
tion finish judicial business brings up the issue of ″ Caligula ″ , Camus
</s> <s> The opposition also brings up a lie made I Simpson 's Lawyers ,
ssue of ″ Caligula ″ , Camus brings up the questions he the Civil Libertie
the question a fact that Sherman brings up in his article </s> <s> However ,
teenager ? in when the CFTC broke up as it had b ng in their background
n-malicious intentions being brought up short by the ues are bound to be b
t by the ues are bound to be brought up and discussed m N away to Athens to
<q . </s> <s> Children are brought up to repent and </s> <s> Even children
crut AT children of Argos are brought up in this atmosph <s> Even the childr
sph <s> Even the children are brought up in guilt , <q . </s> <s> Children
ssed m N away to Athens to be brought up there by a n owards violence , chil
out bonded families and well brought up children . </s> VH heard the sayin
n owards violence , children brought up to 'worship' ido y died was the iss
ppositio one has always been brought up to know about Americans have been bro
know about Americans have been brought up . </s> <s> Thus n-malicious inten
had b ng in their background brought up for public scrut AT children of Arg
s> s . </s> <s> The opinions brought up during the confe tory . </s> <s> I
confe tory . </s> <s> I was brought up to respect Sout bonded families an
hip' ido y died was the issue brought up again . </s> <s> s . </s> <s> The
and </s> <s> Even children are brought up to feel remorse Few negative aspect
orse Few negative aspects are brought up by the oppositio one has always be
</s> VH heard the saying , ″ buckle up . ″ </s> <s> When to hours a day bu
ression of N of each show is to build up the topic to a T stream of proposal

c to a T stream of proposals to build up an unprecedented N of God . </s>
ophy o Camus does , therefore , build up an impression of N of each show is
</s> <s> When to hours a day building up their strength the opposite stance an
strength the opposite stance and builds up the philosophy o Camus does , there
y fight and Caligula the story is built up through interaction government will s
ep a s cation the football team built up faithful . </s> <s . </s> <s> Thi
dented N of God . </s> <s> He built up step-by-step a s cation the footbal
<s . </s> <s> This damage has built up every fight and Caligula the story is
nteraction government will start to buy up surpluses . </s> <s d States . </
</s> <s d States . </s> <s> We came up with the idea t <s> When the dogs
ea t <s> When the dogs finally catch up with the fox , APPGE friends . </s
, APPGE friends . </s> <s> I caught up with him later e and Egisthe too ar
in remorse and it is okay to get caught up in the heat o </s> <s> In the art
im later e and Egisthe too are caught up in remorse and it is okay to get caug
t o </s> <s> In the article ″ Check Up or Check Out ″ w fighting for acts
s is to industry will have to clean up its production championship picture c
ck Out ″ w fighting for acts to clean up our lakes &; , the only way to clean
up our lakes &; , the only way to clean up sports is to industry will have t
production championship picture cleared up a little aft DD misconceptions are
t DD misconceptions are being cleared up and more realis X prove that they a
realis X prove that they are clearing up there act and use the ever increasing
t and use the ever increasingly clogged up roads . </s> <s at their numbers m
soon as of biological parents can come up with more concrete hour before the
e concrete hour before they can come up and see you . am sure farmers can com
and see you . am sure farmers can come up with some idea s> <s> Surely the U
ething mor ore point or views to come up with a solution , the couple have to
th a complete in which morals have come up more often than engineering is consta
money VD Cleveland a chance to come up with a package to try so hard to come
p with a package to try so hard to come up with a complete in which morals have
. </s> <s at their numbers most come up as soon as of biological parents can
th a solution , the couple have to come up with the money VD Cleveland a chan
idea s> <s> Surely the U.K. can come up with something mor ore point or vie
n than engineering is constantly coming up for discussion i is some grand sca
governme likely the party will cover up his death and . </s> <s> Sometimes I
scussion i is some grand scale cover up by our governme likely the party w
s death and . </s> <s> Sometimes I cut up all the vegetables were successfiil w
p all kinds of what else the press drag up . </s> <s> Schools are n’t sitting a
ables were successfiil without dragging up all kinds of what else the press drag
-testing company would have to draw up rules for a co nd tradition involve
> <s> They . </s> <s> They met to draw up tough drug-testing company would
Schools are n’t sitting around drawing up prols and con constitution which was
ls and con constitution which was drawn up between Michel De the th Republic
Michel De the th Republic was drawn up as a compromise the Faure reforms wer
compromise the Faure reforms were drawn up . </s> <s> They . </s> <s> They met
co nd tradition involved and dressing up . </s> <s> As France from NATO , and
lans for France and benefits , and drew up a detailed set Surgeon General , and
</s> <s> As France from NATO , and drew up plans for France and benefits , and d
detailed set Surgeon General , and dug up as much dirt on . </s> <s> So Ala
p on death row . and the toxics may end up in the water tab orrectly in the en
water tab orrectly in the end may end up being very successf , the metals s
rimination a n murder so they can end up on death row . and the toxics may end

one el prejudice may not actually end up in discrimination a n murder so the
to hurt individuals . </s> <s> You end up hurting someone el prejudice may no
ry profi PGE glasses , they would end up on the floor and s but must buy in
successf , the metals should not end up in the landfills no sporting event sh
landfills no sporting event should end up with opponents not integrated , who w
ost of hat aircraft brakes would end up to be very profi PGE glasses , they
</s> < in aggresive feelings and end up killing someone . people who are frie
letely f f criminals who actually end up in jail . </s> < in aggresive fee
ng someone . people who are friends end up fighting and often , and Gerald McCle
ng someone heir driving and they end up hitting an innocen nced to death ,
g and often , and Gerald McClelland end up in a wheelchair intention , they end
opponents not integrated , who will end up paying most of hat aircraft brakes
in a ja fighting and often times end up wanting to hurt individuals . </s> <
g an innocen nced to death , they end up waiting in a ja fighting and often
up in a wheelchair intention , they end up murdering someone heir driving and
up depriving them to be so simple ended up to be a big convenient store ended up
ow he is full of hope but has ended up having a leg under close scrutiny it
the nation laws and theories and ended up creating a new to be the best and end
creating a new to be the best and ended up using steroids t go two heavyweight
ids t go two heavyweight boxers ended up fighting at a who went to college end
fighting at a who went to college ended up dropping out to , but in reality he e
dirt on . </s> <s> So Alabama ended up being the nation laws and theories an
d with that Michael Watson has ended up how he is full of hope but has en
d up to be a big convenient store ended up in an almost f P . </s> <s> Michae
f P . </s> <s> Michael Watson ended up in a wheelchair </s> <s> But many wo
opping out to , but in reality he ended up depriving them to be so simple ended
eelchair </s> <s> But many women ended up dissatisfied with that Michael Wat
ving a leg under close scrutiny it ends up being completely f f criminals who
floor and s but must buy in , fatten up and slaughter th to choose because
ghter th to choose because they fill up so quickly at a Market , the debate f
quickly at a Market , the debate flared up again , but ov J quote should have
ov J quote should have been followed up with a strong , but he quickly follow
with a strong , but he quickly follows up the statement w dealt with , it wo
atement w dealt with , it would free up courts to deal by Boston College and
day , the , Many of these women gave up there children for become superior
urts to deal by Boston College and gave up pts. in the sm . </s> <s> She nev
e and although , so even if people gave up beef today , the , Many of these w
in the sm . </s> <s> She never gave up hope and although , so even if people
the morning journalist is forced to get up in court and tel not have the energ
rt and tel not have the energy to get up four times every States , it has only
is sickbed report when a person gets up in the morning journalist is forced t
ildren for become superior - he gets up from his sickbed report when a per
freedom (ch is too important to give up . </s> <s> Allen B O them are prep
> Allen B O them are prepared to give up their car to , do not want to give up
ld for GE willingness to die and give up . </s> <s> Clamence do not wish to b
e up their car to , do not want to give up personal liberty s . </s> <s> How
<s> Clamence do not wish to boldly give up their sovereignty a lot of people did
erty s . </s> <s> However , to give up the products now Britain be prepared
r government 's brother refuses to give up his now foolishl d hope that he wil
hysical s II the parents as they give up their child for GE willingness to d
ir sovereignty a lot of people did give up beef products , being told to just gi

roducts now Britain be prepared to give up their government 's brother refuses t
his reign of <s> and that he would give up his if it we with it They are givi
now foolishl d hope that he will give up his reign of <s> and that he would gi
beef products , being told to just give up trying . </s> <s> meant the poor
</s> <s> meant the poor had to give up their freedom (ch is too important
times every States , it has only given up some of it 's with life . </s> <s>
it 's with life . </s> <s> He gives up all metaphysical s II the parents a
est o DD is simmilar to Christ giving up his life so were evildoers for giving
p their children wiser than he , giving up his dream of , like a priest giving u
his if it we with it They are giving up on the rest o DD is simmilar to Chr
p his life so were evildoers for giving up their children wiser than he , giving
ife constraint involving the giving up of some of one s the convicted crim
io victim any good . </s> <s> Giving up one 's life constraint involving
up his dream of , like a priest giving up his reconciliatio victim any good .
of one s the convicted criminal goes up for continuous appe ady in Rabkin '
owing it as children get older and grow up males are accepte s . </s> <s> The
be resentful <s> He is looking to grow up and seeks to do t place for orphans
g violence . </s> <s> As children grow up , they learn </s> <s> I wo n't s
cepte s . </s> <s> The children grow up watching violence . </s> <s> As chil
y learn </s> <s> I wo n't say grow up or become an a > <s> Females , as t
ome an a > <s> Females , as they grow up , are accepted a test tube baby ma
> This an XX want my daughter to grow up thinking she co be extremely effici
prob </s> <s> The child may grow up to be resentful <s> He is looking to
e accepted a test tube baby may grow up with identicty prob </s> <s> T
eks to do t place for orphans to grow up . </s> <s> In the place for children
> <s> In the place for children to grow up . </s> <s> This an XX want my daug
> <s> ildren have proven that growing up in one provides TO read when we wer
> <s> hought to our children growing up in the nineties . </s> <s> A chil
nineties . </s> <s> A child growing up with the knowled world . </s> <s>
us appe ady in Rabkin 's book Growing Up Dead . </s> <s> hought to our chi
ed world . </s> <s> Females growing up in their teen r had the advantage o
e s example for children when growing up . </s> <s> They remorse as he had gr
r teen r had the advantage of growing up in a family </s> <s> He starts out g
provides TO read when we were growing up . </s> <s> The s example for child
family </s> <s> He starts out growing up sheltered in the NN to every young
the NN to every young person growing up today . </s> <s> ildren have prove
their I that the children have grown up without them an GE experiences , I
whe </s> <s> Others may have grown up with crime around novel . </s> <s
ere but he . </s> <s> If one has grown up washing their I that the children
</s> <s> They remorse as he had grown up elsewhere but he . </s> <s> If one h
them an GE experiences , I have grown up in an age whe </s> <s> Others may
ound novel . </s> <s> Candide grows up in this novel one child in four grows
ll its NN is banned and no-one grows up knowing it as children get older and
p in this novel one child in four grows up in poverty . < AT child because wh
y . < AT child because when it grows up and all its NN is banned and no-on
co be extremely efficient for heating up left overs and betrayl . </s> <s> Af
<s> The N these reasons seem to hold up , but one by so therefore they get ho
<s> They D evidence they do not hold up well . </s> <s> The N these reason
ew ″ regar R , and feels he must hold up the mirror of h that sterotypes do
ror of h that sterotypes do n't hold up . </s> <s> They D evidence they d
s and betrayl . </s> <s> After holding up a mirror to acts as the saviour - hol

mirror to acts as the saviour - holding up a new ″ regar R , and feels he must
but one by so therefore they get hooked up in the drug b VI these beautiful cr
g b VI these beautiful creatures hung up on walls for sh was because he want
lls for sh was because he wanted join up with the other p amelot , are tryin
e other p amelot , are trying to keep up the illusion of will be able to keep
up the illusion of will be able to keep up with the rate o <s> Pro-life advoc
o <s> Pro-life advocates have lined up in front of lunch . </s> <s> Before
ront of lunch . </s> <s> Before lining up in the routine should be encouraged t
he routine should be encouraged to link up more . </s> <s> Fo e places for pe
s> <s> Fo e places for people to lock up bikes . </s> <s> I used and books
s> <s> I used and books can be looked up on computers . their own councils mad
uropean Council of Ministers , made up by various minister European Centr
which shows argues that society is made up of functional parts N or observable
inister European Central Bank , made up of prominant econom IO the European
onom IO the European Community , made up from the European , the European Parl
's , w bring about a population made up entirely of well-bu s> <s> Infectio
nally integ CS the number was ″ made up ″ . </s> <s> Accordi test people
on computers . their own councils made up of staff , teachi H ″ to a single c
, teachi H ″ to a single country made up of the European Council of Minist
-bu s> <s> Infectious agents are made up of bacteria , vi gy' ; 'nigology' i
l parts N or observable level is made up of rationally integ CS the number
teria , vi gy' ; 'nigology' is a made up word which shows argues that society
European , the European Parliament made up by MEP 's , w bring about a populat
componen ces of the people that make up their ratings ; say that the undemo
Accordi test people 's genetic make up to find the pos hese included w/ Pa
s house ystem for commuters who make up a large bulk o orld does not necess
the pos hese included w/ Passion make up three main componen ces of the peop
ings ; say that the undemocratic make up of this house ystem for commuters
bulk o orld does not necessarily make up for the bad . of the murderer can
e bad . of the murderer can not make up for the loss of hey where subsudise
loss of hey where subsudised to make up the loss , now time , and helped to m
he loss , now time , and helped to make up for the large nu t he may have been
he large nu t he may have been making up his whole conf of a profit , someti
conf of a profit , sometimes marking up their prices way she does n't measure
p their prices way she does n't measure up , therefore a that she would rather m
therefore a that she would rather meet up with Kaliayev in out more about this
h Kaliayev in out more about this mixed up world I came redundant , it has opene
of these in which a person can open up and be vulnerable so they do n't open
p and be vulnerable so they do n't open up and express their she is willing to o
her heart to body part . </s> <s> Open up more of these in which a person ca
nd express their she is willing to open up her heart to found it impossible to p
world I came redundant , it has opened up areas of them world around us is open
reas of them world around us is opening up , the avalibility sons . </s> <s>
bility sons . </s> <s> If she opens up her heart to body part . </s> <s> Op
er heart to found it impossible to pass up such an offer r where bikes can be
younger sis sit in a boat , or pick up a bat . </s> < o sit on the couch
equivalent of away ; The dogs will pick up the scent and c ated the questions
r it had t humans being able to pick up the equivalent of away ; The dogs wil
t . </s> < o sit on the couch , pick up the remote control his brother and th
emote control his brother and then pick up a gun in a could skip work to pick up
k up a gun in a could skip work to pick up a free Thanksgivin C wait for the d

anksgivin C wait for the dogs to pick up a scent . </s> <s> The industry be
</s> <s> The industry began to pick up after it had t humans being able t
an offer r where bikes can be picked up and used and d , then raced home a
nd d , then raced home after picking up the kids from mom gets home she picks
p the kids from mom gets home she picks up my younger sis sit in a boat , or
nd c ated the questions start popping up for mom and or the beef until sale
mom and or the beef until sales push up again . </s> <s> It ace , unions c
s> <s> It ace , unions could also put up candidates for the without having to
andidates for the without having to put up a front , friend because teachers
friend because teachers have to put up with a lot in has been started she ro
th a lot in has been started she rounds up the dirty clothes t bother , take y
clothes t bother , take years saving up or adopt - but would probably have
pt - but would probably have to sell up . </s> <s> As a <s> The Abb from Pri
our own hours multinational company set up labs is an und ch as Dallas , could
that if yo ctually , the flag was set up the day after t PHS is bleeding and
nages and I the weaker sex , men set up double standards in N in less time
up a fund for an , which have been set up . </s> <s> These in Grande Ecole
standards in N in less time , you set up your own hours multinational company
bs is an und ch as Dallas , could set up a fund for an , which have been set u
o protec , or job search programs set up by the Famil figured the world wa
s> These in Grande Ecole will be set up for life , not League Baseball has se
following rules is an organisation set up in to protec , or job search progr
for life , not League Baseball has set up the following rules is an organisatio
the Famil figured the world was set up so that if yo ctually , the flag wa
<s> As a <s> The Abb from Prigord sets up a hoax to rob and more employers are
x to rob and more employers are setting up discussion groups discriminating . <
idelin M go to the states for setting up orphanages and I the weaker sex ,
s discriminating . </s> <s> By setting up protective guidelin M go to the sta
er t PHS is bleeding and quite shaken up , but he wi children , may not eve
t he wi children , may not even show up until years could still possibly sho
until years could still possibly show up after , , <s> <*> . These thought
, , <s> <*> . These thoughts spark up the issue on v N always has interes
g up on the lack of their ratings speak up and stop the c . </s> <s> This dev
n v N always has interest in speaking up on the lack of their ratings speak up
he c . </s> <s> This device has sped up one of our m rplane has also basica
r m rplane has also basically speeded up the whole postal ting up left overs
ostal ting up left overs and speeding up the process of he bottom . </s> <s
of he bottom . </s> <s> This speeds up the preparation mily concerns whic
h ce er negative feelings have sprung up simultaneously with worth it for Hugo
ation mily concerns which had sprung up in the th ce er negative feelings h
neously with worth it for Hugo to stand up for his views Women the courage to st
or his views Women the courage to stand up for their right may be your turn to
their right may be your turn to step up in front of which they so ardently st
n front of which they so ardently stick up for . </s> <s> By by the gutter pres
. </s> <s> By by the gutter press stir up public opinion an AT least impracti
p the case I to hunt . </s> <s> To sum up , fox hunting is . </s> <s> This
an AT least impractical - for summing up the woes and basis today . </s> <s>
oes and basis today . </s> <s> Summing up the case I to hunt . </s> <s> To sum
the sport would tasks that used to take up a lot of time many people will take u
</s> <s> T are very expensive , take up a lot of space he had been able to
up a lot of time many people will take up the lottery withou I the electoral

304 ce on DZ not weigh anything nor take up space . </s> <s> T are very expen
305 Hoederer 's offer college early to take up the sport would tasks that used to ta
306 st th) . </s> <s> Also , buses take up less space on DZ not weigh anythin
307 lot of space he had been able to take up Hoederer 's offer college early to ta
308 he meats o n 'Les Justes ' were taken up again by Camus . </s> <s> Caligula t
309 e Mythe d the proportion of it taken up by the meats o n 'Les Justes ' were
310 ng is . </s> <s> This idea is taken up in Le Mythe d the proportion of it
311 Camus . </s> <s> Caligula thus , takes up revolt against th) . </s> <s> Als
312 ou I the electoral system and themake up of the House Genetic engineering how
313 use Genetic engineering however throws up its own moral Most of this-water is t
314 ts own moral Most of this-water is tied up in glaciers , ic lfillment of these
315 ally speeded up the whole postal ting up left overs and speeding up the proces
316 rs , ic lfillment of these needs took up of majority of welfare-to-work pr
317 of welfare-to-work program can turn up jobs for most is time for people t
318 stand tha uch a great feeling to wake up in the morning ″ , Britons will wake
319 for most is time for people to wake up and understand tha uch a great feel
320 up in the morning ″ , Britons will wake up and see their morning she will wake u
321 up and see their morning she will wake up and see herself , I would like to wei
322 and see herself , I would like to weigh up both sides of - while Penn State whip
323 on Oregon - , can come home can whip up something in minut AT person with A
324 oth sides of - while Penn State whipped up on Oregon - , can come home can wh
325 n minut AT person with AIDS will wind up paying for expensi homes tend to g
326 expensi homes tend to get so wrapped up in the size