



<http://waikato.researchgateway.ac.nz/>

## Research Commons at the University of Waikato

### **Copyright Statement:**

The digital copy of this thesis is protected by the Copyright Act 1994 (New Zealand).

The thesis may be consulted by you, provided you comply with the provisions of the Act and the following conditions of use:

- Any use you make of these documents or images must be for research or private study purposes only, and you may not make them available to any other person.
- Authors control the copyright of their thesis. You will recognise the author's right to be identified as the author of the thesis, and due acknowledgement will be made to the author where appropriate.
- You will obtain the author's permission before publishing any material from the thesis.

Department of Computer Science



Hamilton, New Zealand

# Supporting Collocation Learning

by  
Shaoqun Wu

This thesis is submitted in partial fulfilment of the requirements  
for the degree of  
Doctor of Philosophy in Computer Science  
at The University of Waikato

August 2010

© 2010 Shaoqun Wu







# Abstract

---

Collocations are of great importance for second language learners. Knowledge of them plays a key role in producing language accurately and fluently. But such knowledge is difficult to acquire, simply because there is so much of it.

Collocation resources for learners are limited. Printed dictionaries are restricted in size, and only provide rudimentary search and retrieval options. Free online resources are rare, and learners find the language data they offer hard to interpret. Online collocation exercises are inadequate and scattered, making it difficult to acquire collocations in a systematic way.

This thesis makes two claims: (1) corpus data can be presented in different ways to facilitate effective collocation learning, and (2) a computer system can be constructed to help learners systematically strengthen and enhance their collocation knowledge.

To investigate the first claim, an enormous Web-derived corpus was processed, filtered, and organized into three searchable digital library collections that support different aspects of collocation learning. Each of these constitutes a vast concordance whose entries are presented in ways that help students use collocations more effectively in their writing. To provide extended context, concordance data is linked to illustrative sample sentences, both on the live Web and in the British National Corpus. Two evaluations were conducted, both of which suggest that these collections can and do help improve student writing.

For the second claim, a system was built that automatically identifies collocations in texts that teachers or students provide, using natural language processing techniques. Students study, collect and store collocations of interest while reading. Teachers construct collocation exercises to consolidate what students have learned and amplify their knowledge. The system was evaluated with teachers and students in classroom settings, and positive outcomes were demonstrated.

We believe that the deployment of computer-based collocation learning systems is an exciting development that will transform language learning.



## Acknowledgements

---

Words cannot express my gratitude to my supervisor Prof. Ian Witten. A topic he suggested for my postgraduate study has grown into my PhD research and has become my passion. This would not have been possible without his generous support, great patience and consistent encouragement. He is also a friend and a mentor. He taught me to stay positive and keep on going. He helped me build confidence in myself and my research. Without him, I would not have become who I am today. Thank you for taking me as your student and giving me opportunities to challenge myself.

I am heartily thankful to Dr. Margaret Franken, for her support, enthusiasm, knowledge and guidance throughout my thesis. I really enjoyed the time we spent together: planning evaluations, writing research papers, and working with students. We made a good team!

I am grateful to my other supervisors, Dr. Dave Nichols and Dr. Tony Smith, for their invaluable knowledge.

I need to express my gratitude to Katherine Brown who has supported me in a number of ways. She helped me recruit evaluation teachers and students, and organize evaluations. She has been introducing my work to students and telling me how they loved it.

I am deeply grateful to Dr. John Brine who has supported me throughout my thesis with his patience and knowledge.

In my daily work I have been blessed with a friendly and cheerful group of fellow PhD students in the DL lab. Special thanks to Michael Walmsley for helping me carefully proofread this thesis.

I would like to show my gratitude to the Royal Society of New Zealand Marsden fund who founded my research.

Many thanks go to all evaluation participants from Pathways College and School of Education at the University of Waikato.

I would like to thank my family: my husband, Xiaofeng, for his understanding, support, and encouragement during my study, and my daughter, Shannon who turns eight months today, for being happy and healthy, and for giving me time to complete this thesis.

Lastly, I owe a great deal to CLS, the writing support tool I developed for this study, for helping me write this thesis and say thanks to people who made it possible.

# Table of Contents

---

Abstract .....	iii
Acknowledgements .....	v
List of Tables.....	xi
List of Figures .....	xiii
1. Introduction.....	1
1.1 Motivation.....	1
1.2 Thesis statement.....	3
1.2.1 Presenting corpus data for collocation learning .....	3
1.2.2 Constructing a collocation learning system .....	4
1.3 CLS: A collocation learning system .....	5
1.4 Contributions.....	7
1.5 Thesis structure .....	9
2. Background .....	11
2.1 What is collocation?.....	12
2.2 Collocation studies .....	13
2.3 Collocation learning .....	15
2.3.1 The importance of collocation learning .....	15
2.3.2 The difficulty of collocation learning .....	18
2.4 Collocation teaching.....	19
2.4.1 Teaching strategies.....	20
2.4.2 Collocation selection.....	23
2.4.3 Collocation activities.....	24
2.5 Resources .....	27
2.5.1 Collocation dictionaries .....	28
2.5.2 Concordancers.....	30
2.5.3 The British National Corpus .....	32
2.6 Corpus-based language learning .....	33
2.7 The Web corpus .....	35
2.7.1 Size.....	36
2.7.2 Representativeness .....	36

2.7.3	Cleanliness.....	38
2.8	Using the Web corpus .....	38
2.9	Computer-assisted collocation learning .....	41
2.9.1	Collocation exercises on the Web .....	42
2.9.2	Concordancer tools for teachers and students .....	45
3.	Presenting corpus data for collocation learning .....	49
3.1	Using Web text.....	50
3.1.1	Cleaning the data .....	52
3.1.2	Building contextual resources .....	54
3.2	The WEB PHRASES collection .....	55
3.2.1	Using the collection.....	56
3.2.2	Building the collection .....	58
3.3	The WEB PRONOUN PHRASES collection .....	59
3.3.1	Using the collection.....	60
3.3.2	Lexical resources .....	64
3.3.3	Building the collection .....	67
3.4	The WEB COLLOCATIONS collection.....	69
3.4.1	Limitations of collocation resources .....	70
3.4.2	Using the collection.....	72
3.4.3	Building the collection .....	75
3.4.4	Web collocations <i>vs</i> BNC collocations .....	77
4.	Extracting collocations for language learning.....	81
4.1	Extracting and evaluating collocations.....	82
4.1.1	Frequency .....	82
4.1.2	Hypothesis Testing .....	83
4.1.3	Mutual information.....	86
4.1.4	Comparison of measures .....	86
4.2	Determining candidate collocations .....	92
4.2.1	Syntactic tagging .....	93
4.2.2	Matching tagged <i>n</i> -grams against statistical patterns .....	96
4.2.3	Ranking the result.....	98
4.3	Investigating tagging errors .....	99

4.3.1	Tagging the BNC .....	99
4.3.2	Comparison with CLAWS .....	101
4.4	Evaluating extracted collocations .....	103
4.4.1	Baseline collocation data .....	104
4.4.2	Test data .....	105
4.4.3	Ranking the Web collocations .....	106
4.4.4	Quality and quantity of Web collocations.....	112
5.	Evaluating collocation resources with language learners .....	117
5.1	The WEB PRONOUN PHRASES collection .....	118
5.1.1	Participants and procedure .....	118
5.1.2	Results.....	119
5.2	The WEB PHRASES and WEB COLLOCATIONS collections .....	125
5.2.1	Designing a user guide .....	125
5.2.2	Participants and procedure .....	127
5.2.3	How students used CLS .....	128
5.2.4	Assessing CLS's potential .....	130
5.2.5	Discussion .....	134
6.	Constructing a collocation learning system .....	137
6.1	Supporting collocation learning .....	138
6.1.1	Creating learning material.....	139
6.1.2	Facilitating noticing, retrieval and generation .....	139
6.2	Building collocation-enriched collections.....	143
6.2.1	Adding texts .....	143
6.2.2	Identifying collocations.....	145
6.3	Using the collection.....	146
6.3.1	Searching and browsing collocations.....	146
6.3.2	Expanding collocation knowledge.....	148
6.3.3	Cherry-picking .....	149
6.4	Collocation activities.....	150
6.4.1	Collection-based activities .....	152
6.4.2	Dictionary-based activities.....	161
7.	Evaluating CLS .....	171

7.1	Collocation evaluation.....	171
7.1.1	Evaluation texts .....	172
7.1.2	Investigating collocations that are identified.....	172
7.1.3	Teacher's selection and judgment of collocations.....	174
7.2	Evaluating the “cherry-picking” facility.....	182
7.2.1	Background.....	182
7.2.2	Testing collocation knowledge.....	183
7.2.3	Collecting collocations .....	186
7.2.4	Results .....	187
7.2.5	Questionnaire.....	188
7.2.6	Discussion.....	190
7.3	Theoretical evaluation with language teachers.....	190
7.3.1	Background.....	191
7.3.2	Results .....	191
7.3.3	Discussion.....	193
8.	Conclusion.....	195
8.1	Presenting corpus data for collocation learning .....	196
8.2	Constructing a collocation learning system.....	198
8.3	Into the future .....	203
	References .....	205
	Appendix A Function word list .....	211
	Appendix B Penn Treebank tags .....	213
	Appendix C User guide .....	215
	Appendix D Evaluation of collocation resources .....	221
	Appendix E Keywords and collocations produced by students .....	227
	Appendix F Fill-in-Blanks exercises .....	231
	Appendix G Cherry basket .....	233
	Appendix H Cherry-picking questionnaire (student) .....	235
	Appendix I Cherry-picking questionnaire (teacher).....	239
	Appendix J Instructions for teachers .....	241
	Appendix K Teachers' discussions .....	245

# List of Tables

---

Table 1.1 Tools and resources used in the thesis .....	7
Table 2.1 Usage note for the word discretion .....	22
Table 2.2 Subphrase frequencies for have been found to be infected/polluted with ..	40
Table 3.1 Number of units in the n-gram corpus .....	52
Table 3.2 Number of n-grams in the Web Phrases collection.....	55
Table 3.3 Response time of the Web Phrases collection .....	58
Table 3.4 Number of pronoun phrases in Web Pronoun Phrases .....	60
Table 3.5 Patterns that follow the words love and hate .....	64
Table 3.6 Collocation types and examples.....	70
Table 3.7 Useful collocations from Collins and WebCorp .....	71
Table 3.8 Grouping templates and examples .....	76
Table 3.9 Example of index and dictionary file .....	76
Table 3.10 Collocation types with statistical data from two corpora.....	78
Table 3.11 Most frequent collocations of four types from two collections .....	78
Table 3.12 Web and British National Corpus entries for a collocation template .....	79
Table 3.13 Top ten cause + noun collocations in three concordances .....	79
Table 4.1 Frequency of four 2-grams.....	87
Table 4.2 Four 2-grams ranked by four measures .....	87
Table 4.3 Top 30 Web bigrams, ranked by five measures .....	89
Table 4.4 Top 30 Web bigrams, filtered by function words .....	90
Table 4.5 Top 30 BNC bigrams, ranked by five measures .....	91
Table 4.6 Top 30 BNC bigrams, filtered by function words .....	92
Table 4.7 Tag mapping between Penn Treebank and CLAWS5 .....	97
Table 4.8 Regular expressions for ten collocation types.....	98
Table 4.9 Categories of mismatched tags in full and five-gram context .....	100
Table 4.10 Percentage of matched tags in three experiments .....	101
Table 4.11 Examples of inconsistent tagging between OpenNLP and CLAWS .....	101
Table 4.12 Words used to filter five-grams .....	102
Table 4.13 Categories of mismatched tags between OpenNLP and CLAWS .....	103

Table 4.14 Number of collocations extracted from the Oxford Collocation Dictionary for Students of English .....	105
Table 4.15 Reasons why particular collocation types are not used in the evaluation.....	107
Table 4.16 Headwords that are not covered by Web collocations .....	107
Table 4.17 Number of collocations in the baseline and test data .....	107
Table 4.18 Precision at various recall values for three measures, Frequency, t-test and LLR .....	112
Table 4.19 Percentage of collocations that do not occur in Web Collocations .....	113
Table 4.20 Words whose class is difficult to determine.....	114
Table 5.1 Summary of the log data .....	122
Table 5.2 Samples extracted from student text.....	125
Table 5.3 System generated alternatives to errors .....	131
Table 5.4 Student changes to errors identified in their texts .....	132
Table 7.1 Collocation statistics for evaluation texts.....	173
Table 7.2 Problems associated with automatic collocation identification.....	174
Table 7.3 Statistics of collocations identified by teachers.....	175
Table 7.4 Collocations identified by teacher A, but not by the system .....	177
Table 7.5 Collocations identified by teacher B, but not by the system .....	178
Table 7.6 Collocations that the teachers did not approve.....	179
Table 7.7 Frequency of uncommon combinations in Web Phrases .....	181
Table 7.8 Collocations not approved by teacher A .....	181
Table 7.9 Statistics of collocations used in the evaluation .....	183
Table 7.10 Students' vocabulary test scores.....	184
Table 7.11 Number of keyword and collocations produced by students.....	185
Table 7.12 Results of Fill-in-Blanks tests .....	185
Table 7.13 Statistics of collocations collected by the students .....	187
Table 7.14 Collocations collected by students .....	188
Table 7.15 Use of collected collocations.....	189

## List of Figures

---

Figure 1.1 Architecture of CLS .....	6
Figure 2.1 Entry in the <i>Oxford Collocation Dictionary for Students of English</i> ..	30
Figure 2.2 Online concordancer at <a href="http://www.lextutor.ca">www.lextutor.ca</a> .....	31
Figure 2.3 Excerpt of a BNC XML document .....	33
Figure 2.4 Concordance data returned by WordCorp for the word <i>make</i> .....	39
Figure 2.5 Frequency plotted against phrase length.....	41
Figure 2.6 Collocation exercises at <a href="http://a4esl.org">a4esl.org</a> .....	43
Figure 2.7 Multiple choice exercise at <a href="http://www.better-english.com">www.better-english.com</a> .....	45
Figure 2.8 Collocation exercises on <a href="http://angelfire.com">angelfire.com</a> .....	47
Figure 3.1 CLS's collocation learning resources .....	50
Figure 3.2 Sample <i>n</i> -grams .....	52
Figure 3.3 Samples retrieved for <i>I was a little disappointed</i> .....	53
Figure 3.4 Searching facilities provided by WEB PHRASES .....	57
Figure 3.5 Searching facilities provided by WEB PRONOUN PHRASES.....	62
Figure 3.6 Browsing facilities provided by WEB PRONOUN PHRASES .....	65
Figure 3.7 Synonyms for <i>disappointed</i> retrieved from WordNet .....	67
Figure 3.8 Collocates of the word <i>make</i> .....	72
Figure 3.9 Searching facilities provided by WEB COLLOCATIONS .....	75
Figure 4.1 Parsing a document.....	94
Figure 4.2 Precision-recall curves .....	111
Figure 6.1 Collocation learning platform in CLS .....	138
Figure 6.2 Collection building interface: adding an article .....	143
Figure 6.3 Configuring collocation identification parameters .....	144
Figure 6.4 Example collection's "About" page .....	146
Figure 6.5 A document in the collection, with collocations highlighted .....	147
Figure 6.6 Collocation results when searching for the word <i>family</i> .....	148
Figure 6.7 Browsing by collocation type .....	149
Figure 6.8 Collocations related to <i>extremely high</i> .....	150
Figure 6.9 Picking cherries .....	151

Figure 6.10 Cherry basket .....	151
Figure 6.11 Fill-in-Blanks exercise .....	153
Figure 6.12 Common Alternatives exercise .....	155
Figure 6.13 Correcting Errors exercise.....	156
Figure 6.14 Multiple Choice exercise.....	157
Figure 6.15 Design interface for the Fill-in-Blanks activity .....	161
Figure 6.16 Collocation Guessing exercise .....	163
Figure 6.17 Collocation Dominoes exercise .....	164
Figure 6.18 Collocation Matching exercise.....	165
Figure 6.19 Related Words exercise .....	166
Figure 6.20 Design interface for the Collocation Guessing activity .....	169

# 1. Introduction

---

*You shall know a word by the company it keeps*

J.R. Firth, 1957

Why do language learners find it difficult to differentiate between words like *look*, *see* and *watch*, *injury* and *wound*, or *broad* and *wide*? Why do students who know many individual words still struggle to express complex ideas simply and precisely? Why are so many frustrated that they make little visible progress? How is it that native speakers communicate so much more effectively? The answers rest on the collocation knowledge of learners. It is the collocates of *look*, *see* and *watch*, *injury* and *wound*, or *broad* and *wide* that reveal their different shades of meaning, rather than their dictionary definitions (Conzett, 2000).

Complex ideas are hard to express unless one can use simple vocabulary in a range of collocations (Lewis, 1993). Hill (1999) points out that students with good ideas often lose marks, because they do not know the four or five most important collocates of a key word that is central to what they are writing about. Wray (2002) and Nesselhauf (2003) emphasize that collocations are particularly important for learners striving for a high degree of competence in a second language, because they enhance not only accuracy but also fluency.

## 1.1 Motivation

Studies suggest that an educated native speaker of English has a vocabulary of around 20,000 word families (Goulden et al., 1990). That is a large number, but still a manageable goal for the most determined and motivated learners. However, it pales into insignificance when compared with the total number of items—expressions, idioms, collocations—that native speakers have (Hill, 2000). Collocation knowledge is difficult to acquire simply because there is so much of it. Native speakers carry hundreds of thousands—possibly millions—of lexical chunks in their heads, ready to draw upon in order to produce fluent, accurate and meaningful language (Lewis, 1997). This presents a daunting challenge to language learners.

Teachers face great challenges in helping their students develop collocational competence. Classroom time is inadequate even for learning the basic vocabulary. In practice, collocation teaching is neglected (Farghal and Obeidat, 1995).

Collocation learning has been peripheral in the classroom for two principal reasons. First, grammar is the traditional focus of curriculum, especially in EFL teaching, because it is relatively easy to teach and assess. Second, identifying a set of useful collocations is a daunting task, and because of the limited resources at their disposal most teachers have to rely on intuition. This is challenging even for native speakers, let alone teachers whose mother tongue is not English (Gabrielatos, 2005). Collocation learning is a cumulative process that involves a great deal more than rote memorization. Students with limited study time will not learn appropriate collocations unless they are deliberately selected, prioritized, and incorporated into language material (Swan, 1996).

Resources like dictionaries and concordancers are useful tools for learning collocations. However, printed dictionaries are expensive, the number of collocations they provide is restricted by physical size, and consultation facilities are insufficiently flexible to meet all the needs of learners. Concordancers are among the most frequently used tools for exploring corpora, particularly with a view to examining collocation use. They allow students to obtain, organize, and study real-language data derived from corpora. However, not all concordance results are easily navigated and analyzed by learners. Information must be presented in a way that is both accessible and relevant to learners. They should provide sufficient and varied language data, in combinations that are flexible and generative.

Although the rise of computer-assisted language learning has brought a new dimension and dynamic into language learning, little research has been done on computer-assisted collocation acquisition. Online collocation exercises have several limitations. First, they are inadequate compared to the sheer size of collocation knowledge that learners need to acquire. Second, they are created by teachers who focus on particular topics, which may not be suitable for learners with different needs. Third, collocations are pulled out of their original context,

and scant attention is paid to their actual use in real language. Last but not least, online exercises are scattered throughout the Web, which makes it difficult for learners to study collocations in a systematic way.

## 1.2 Thesis statement

This thesis aims to address two issues: (1) to investigate how corpus data should be presented for collocation learning, and (2) to construct and evaluate a system that helps learners systematically strengthen and enhance their collocation knowledge.

### 1.2.1 Presenting corpus data for collocation learning

Conventional collocation resources like dictionaries and concordancers are either limited by physical size or offer language data that is hard for learners to interpret. This leads to our first hypothesis:

Corpus data can be processed and organized in different ways to help learners expand collocation knowledge.

Language corpora, defined by Meyer (2002) as collections of “texts or parts of texts upon which some general linguistic analysis can be conducted,” now feature prominently in the teaching and learning literature. However, they are of little use without properly organized and carefully designed access tools, because raw corpus data inevitably overwhelms ordinary learners.

Corpus data needs to be processed in order to meet the needs of learners with different language abilities and learning purposes. Processing involves both fragmentation and selection. Fragmentation builds subsets of corpus data—for example, subsets containing text that includes words in a particular wordlist, say the most frequent 5000 words. Selection extracts text that exhibits a particular language feature—for example, all sentences that start with a pronoun.

Once processed, corpus data needs to be organized so that learners can find what they want. Learners may seek prepositions that follow *is responsible (for)*; words that precede *but not least (last)*; adverbs that occur between *I am* and *aware (well, fully, also, and quite)*. Having learnt a new word like *difference*, students should

be able to find verbs that collocate with it (*make a difference, tell the difference, see the difference, understand the difference*), and to inspect collocations in real language and in different contexts in order to expand their knowledge.

### 1.2.2 Constructing a collocation learning system

Collocations need to be explicitly learnt. This leads to our second hypothesis:

For a given collection of language learning text, pedagogically valuable collocations can be automatically identified and incorporated into a learning environment that facilitates the key activities of *noticing, retrieval* and *generation*.

Word combinations that can be gleaned from text are not necessarily pedagogically valuable. Careful selection must be undertaken to ensure that the identified collocations are: (1) both common and important; (2) needed by the student population; (3) match the language ability level of a particular student group.

The quality of identified collocations reflects the performance of the underlying natural language processing tools and the algorithms used for extraction. They inevitably fall short of perfect accuracy. Furthermore, the special needs and language ability of particular student groups are difficult to quantify. Therefore, although the hypothesis specifies that the techniques should be *automatic*, we recognize that the identification process ultimately requires human judgment: language teachers must be given an opportunity to revise the identified items before they are presented to students.

Extracted collocations are of little value in themselves: they need to be explicitly learnt. This thesis recognizes the three processes that Nation (2001) summarizes as leading to lexical acquisition: noticing, retrieval and generation. Learning starts with *noticing*, which occurs when the learner deliberately pays attention to an item as part of the language, rather than as part of a message. It is affected by several factors: the salience and usefulness of the item, its presentation, the learner's interest and motivation, the learner's mindset—for example, focusing on individual words *vs.* larger chunks of language—and the learning environment.

*Retrieval* is the process of remembering an item. It involves three aspects. First, the item needs to be understood in the context in which it occurs. This might be by guessing its meaning from the context, looking it up in dictionaries, or constructing an interpretation by debating its meaning with peers or teachers. Second, the item's meaning must be retrieved when it is met in reading or listening. Third, the item must be used in circumstances that are semantically and pragmatically appropriate.

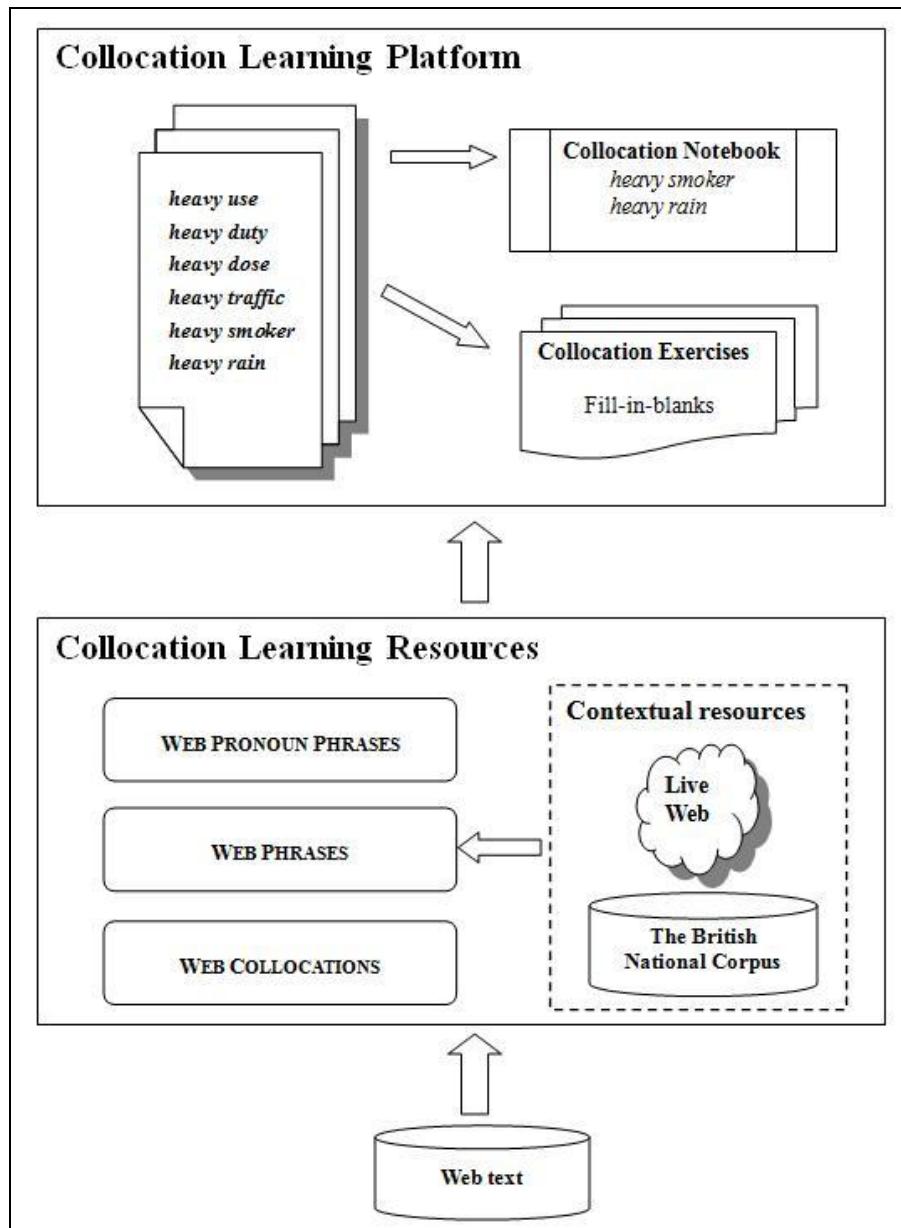
*Generation* is the process of enriching and stretching the learner's knowledge of an item. It occurs when the item is met in different forms and contexts. For example, the word *heavy* has different meanings when used in *heavy rain* and *heavy smoker*; its adverbial form is *heavily*. Generation can be achieved by incorporating material from various sources to create a rich contextual environment that enables learners to discover and analyze new meanings and multiple uses of collocations.

### 1.3 *CLS: A collocation learning system*

To investigate these claims, we have developed the CLS collocation learning system. It utilizes the Greenstone digital library software, which allows users to build large collections of documents and metadata and serve them on the Web (Witten et al., 2010). Each collection is equipped with a full-text index and metadata browsing facilities.

CLS comprises two components: collocation learning resources and a collocation learning platform. Figure 1.1 outlines its structure. In the lower part, CLS processes and organizes Web text into three collections:

- WEB PRONOUN PHRASES, containing phrases starting with pronoun words—*I, you, he, she, we, they* and *it*.
- WEB COLLOCATIONS, containing collocations organized by syntactic pattern, and
- WEB PHRASES, containing word sequences of up to five words.



**Figure 1.1 Architecture of CLS**

Learners use these resources to explore pronoun phrases and collocations and check their text against general usage on the Web. The British National Corpus (Section 2.5.3) and the live Web offer contextual information to help learners study these phrases in different contexts.

Tools and resources	Purpose
Greenstone software	Provides the server-client infrastructure for constructing CLS (Chapter 3 and 5).
Google <i>n</i> -grams	Built into three collocation learning resources (Section 3.1)
British National Corpus	Built into a Greenstone collection for providing contexts of <i>n</i> -grams (Section 3.1.2)
WordNet Roget's thesaurus Edinburgh Word Association thesaurus Lemma list	These four resources are all used when retrieving words related to or associated with a query term (Section 3.3.2)

**Table 1.1 Tools and resources used in the thesis**

In the upper part, CLS automatically identifies collocations in texts provided by teachers, and presents them in a way that attracts the attention of learners. While reading the text, learners collect collocations of interest and store them in a notebook. Collocation exercises that are automatically generated from the text and produced under teacher control help consolidate the collocations that learners have encountered. To amplify collocation knowledge, external resources are either linked to the identified collocations or offered as a help facility in exercises.

CLS is a substantial software system built on top of several existing components. Table 1.1 gives the resources and tools used in the thesis. CLS utilizes the client-server infrastructure provided by the Greenstone digital library software. It includes its own server and client components, implemented using the Java and Javascript technologies respectively. Both are substantial pieces of software: the server components comprise 120 Java classes and 40,000 lines of code, while the client components contain 55 Javascript files and 25,000 lines of code.

## 1.4 Contributions

The contributions made during this investigation are as follows.

### *Presenting corpus data for collocation learning*

- Three learning resources from Web text that allow learners to study pronoun phrases, to study collocations organized by syntactic pattern, and to check their text against general usage on the Web (Chapter 3).
- An algorithm that extracts collocations from Web text (Chapter 4).

- Comparative evaluation of five standard statistical measures for ranking collocations on Web and BNC bigrams (Chapter 4).
- An evaluation of collocations extracted from Web text with respect to those in the *Oxford Collocation Dictionary for Students of English* (Chapter 4).
- Assessment of the impact of restricted context on the accuracy of the tagging process (Chapter 4).
- Two user studies that investigated the effectiveness of the three collocation learning resources in supporting writing (Chapter 5).
- A publication in the *Computer Assisted Language Learning Journal*:  
Wu, S., Franken, M. and Witten, I.H. (2009). “Refining the use of the web (and web search) as a language teaching and learning resource.” *Computer Assisted Language Learning*, 22(3), 249-268, July.
- A publication in the *RECALL Journal*: Wu, S., Witten, I.H. and Franken, M. (2010). “Utilizing lexical data from a Web-derived corpus to expand productive collocation knowledge.” *ReCALL*, 22(01), 83-102, January.

#### *Constructing a collocation learning system*

- A collocation learning system that allows teachers or learners to build their own collections of text, and helps learners notice and collect collocations that have been identified in text and consolidate what they have learnt by doing a variety of exercises (Chapter 6).
- A user study that investigated the quality of automatically identified collocations against those manually selected by teachers (Chapter 7).
- A user study that examined the effectiveness of the collocation collection facility with students who are doing university study (Chapter 7).
- A user study that explored the strengths and limitations of CLS with language teachers (Chapter 7).
- A publication in the *Computer Assisted Language Learning Journal*:  
Wu, S., Witten, I.H. and Franken, M. (2010). “Supporting collocation learning with a digital library.” *Computer Assisted Language Learning Journal*, 23(1), 87-110, February.

Designing a computer-assisted language learning system is a complex task. CLS is based on teaching strategies that teachers use in classrooms, and language

acquisition theories that have been put forward by many researchers. The thesis investigates novel ways of constructing a collocation learning focused system, and has included five initial user studies that provide useful insights for understanding and further development of the system. Full evaluations of the educational effectiveness of CLS will be needed to assess its eventual effect on collocation learning, and here the design of the entire learning environment, including the goals, motivations and training of teachers and students, will play an important role. However, such evaluations are beyond the scope of the thesis.

### *1.5 Thesis structure*

The remainder of the thesis is structured as follows. Chapter 2 provides background by reviewing the definition of collocation and related studies, and discusses the importance and difficulty of collocation learning. It also introduces the strategies and activities that teachers adopt inside and outside classroom, and two main collocation resources—printed dictionaries and online concordancers. Then it explores how Web text is used for language learning, and reviews collocation exercises and facilities available on the Web.

To investigate the thesis's claims, we built collocation resources from Web text. Chapter 3 explains how it is processed and organized into three digital library collections, and demonstrates how to use the search and retrieve facilities they provide to study collocations in different contexts.

Then we focus on the algorithm used to extract collocations from Web text (Chapter 3). A comparison of five statistical measures was conducted, and Frequency was chosen to rank extracted collocations. The impact of restricted context that Web text provides on the accuracy of the part-of-speech tagger was assessed. The quality and quantity of extracted collocations were evaluated with respect to the *Oxford Collocation Dictionary for Students of English* dictionary.

To determine the effectiveness of the collocation resources, two evaluations were undertaken with students to support their general and academic writing (Chapter 5). Both suggest that these resources can help students improve their writing in

terms of correcting grammar and collocation errors, generating text and expanding text.

The CLS collocation learning platform is introduced in Chapter 6. We show how to create digital library collections and explain how collocations are automatically identified in the text. Then we demonstrate how students study, collect and store collocations and introduce eight collocation activities with which teachers create unlimited exercises to consolidate what students have learnt.

Three evaluations were conducted to assess the usefulness of CLS (Chapter 7). First, the quality of automatically identified collocations was examined by two teachers. Second, the facility that students use to collect collocations was tested in a Masters study course to illustrate its use in supporting academic writing. Third, four teachers were invited to explore CLS, and feedback was gathered for future development.

Chapter 8 concludes the thesis and discusses future work.

## 2. Background

---

Learning a second language is not an easy task. Teachers seek efficient ways of improving student performance, given the limited time they have to study the language. Many teachers have realized that grammar alone is not enough to help students achieve native-like proficiency. Students may have learnt the grammar to construct the sentence *he is a strong smoker*, but do they understand that *strong* and *smoker* do not go together? Research shows that a learner's collocation knowledge plays a key role in producing language fluently and accurately (Nation, 2001); and collocation learning has recently become a major focus of interest in second language learning.

Collocations are a common phenomenon in any language. However, providing a universal definition of what constitutes a collocation is difficult: different researchers take different views and adopt different approaches to suit their own purpose. For example, some restrict collocations to adjacent words, while others focus on non-consecutive fragments. This chapter reviews common definitions of collocation from the point of view of linguists, lexicographers, statisticians, and language teachers, and looks at five related studies—lexical, semantic, lexicographical, computational and structural.

Because collocation knowledge is difficult to acquire, teachers have developed many teaching strategies and activities to help students expand their collocation repertoire. Collocation dictionaries are available for students to check collocations they are uncertain of. Concordancers, a traditional tool of linguists to analyze corpus text, have been used by students to explore the language.

In recent years, researchers have turned their eyes to the use of Web text in collocation learning. This chapter discusses the Web as a corpus, and introduces projects that use its text and technologies to provide concordance data, and to extract collocations. The chapter concludes with a survey of collocation exercises and tools available on the Web.

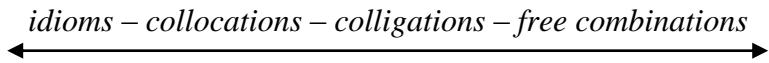
## 2.1 What is collocation?

The term “collocation” has many definitions in the literature. It is an unclear concept with various names: lexical items, prefabricated chunks, routinized formula, formulaic sequences, conventionalized language form, fixed or semi-fixed expressions and so on. Nattinger and Decarrico (1992) define a collocation as “a string of specific lexical items that co-occur with a mutual expectancy greater than chance.” Nation (2001) identifies collocations as “items which frequently occur together and have some degree of semantic unpredictability.” Benson et al. (1986) call them “fixed, identifiable, non-idiomatic phrases and constructions.” In the view of Lewis (1997), “Collocations are those combinations of words which occur naturally with greater than random frequency.” Sinclair (2004a) describes the phenomenon of collocation as “the choice of one word conditions the choice of the next, and of the next again.” In statistical terms, a collocation is two or more consecutive words with a special behavior (Manning and Schütze, 1999). From a language teacher’s point of view, collocations are “words which I think my students will not expect to find together” (Woolard, 2000).

Despite different views on collocations, the common problem that linguists and researchers face is how to delimit them from other types of word combination. In phraseology, there are three major classes of word combinations: idioms, collocations and free combinations. The two most widely accepted differentiation criteria are semantic opacity and collocation restriction, though different terms are used by different linguists. Semantic opacity, also called non-compositionality, is the extent to which the meaning of a phrase is not transparent from its constituents. Collocation restriction, also called substitutability or flexibility, is that the constituent words can be substituted by other words. Idioms—for example, *by and large* and *hell for leather*—are the most extreme examples of non-compositionality and non-substitutability. Collocations are characterized by limited compositionality and substitutability, for example, *pay attention/fees/bills*. Free combinations are freely compositional and substitutable—*buy a book/car/house* or *sell a book/car/house*. Some researchers also consider productivity (the form of a combination being structurally unique), and frequency

(free combination being the most frequent, while idioms are the least frequent) to differentiate idioms, collocations and free combinations more clearly.

Although different researchers use different distinction criteria, most admit that the boundary between the three categories is not clear cut. Nattinger and DeCarrico (1992) state that “instead of assuming a qualitative, either-or distinction between idiomatic language and regularly generated language, collocationists are more prone to see language on a cline, with completely invariant clusters at one end of the continuum, and free combining morphemes at the other, with all degrees of combinational flexibility in between.” They advocate Wood’s (1981) model of language patterns shown below.



Idioms, at one end of the continuum, are completely unpredictable and frozen in their meaning and form, while at the other end are free combinations. Collocations and colligations, in the middle, are somewhat predictable, but restricted to certain items. Colligations are generalizable classes of collocations, for which at least one construct is specified by category rather than as a distinctive lexical item (Nattinger and DeCarrico, 1992). An example is a verb of motion + directional particle such as *go off*, *chase up*, and *run away*.

## 2.2 *Collocation studies*

Since Firth coined the term “collocation” in the 1950s, researchers have taken different approaches to describing, categorizing and predicting collocations, focusing on different aspects of this phenomenon. This section introduces basic ideas of lexical, semantic, and lexicographical study. Each covers a wide field and thorough review is beyond the scope of this thesis. We provide more descriptions of computational and structural studies, because they are closely related to the thesis. These two studies are discussed separately; however, in practice they often overlap.

Lexical studies are based on the assumption that collocation words receive their meaning from the words they co-occur with. As explained by Palmer (1933), one

of the meanings of *night* is its collocability with *dark*, and of *dark*, collocation with *night*. Semantic studies investigate collocations on the basis of the semantic framework, and try to use semantic properties of lexical items to explain why these items collocate with only certain other items. When compiling collocation dictionaries, lexicographers need to decide how to define, select, and organize collocations. Different lexicographers adopt different approaches based on their definition of collocation, and their users and budget.

Computational studies use computers to scan the text of large corpora for collocations. Researchers restrict collocation units to comprise a specified word (the node) that co-occurs with a span of words on each side. Sinclair (1966) defines node, span and collocate as follows:

We may use the term node to refer to an item whose collocations we are studying and we may then define a span as the number of lexical items on each side of a node that we consider relevant to that node. Items in the environment set by the span we will call collocates.

He believes a span of four is adequate for any type of data. In practice, a wider or narrower span may be used for different purposes.

Not all words within a span of a particular word are of interest unless they co-occur at a frequency greater than chance would predict. Recently, statistical techniques have been employed to locate collocations. Church and Hanks (1989) propose an information-theoretically motivated measure—mutual information—that estimates how much one word tells us about the other, using the probability of observing X and Y together and the probability of observing X and Y separately. Manning and Schütze (1999) introduce hypothesis testing to assess whether two words occur together more often than chance. These measures will be discussed in Section 4.1.

Some researchers suggest that collocation is associated with structure and should be studied in structurally defined patterns. According to Mitchell (1971), the collocation *heavy drinker* follows the colligation pattern adjective + agentive noun. Renouf and Sinclair (1991) investigate collocations using the following framework:

<i>a + ? + of</i>	<i>an + ? + of</i>	<i>be + ? + to</i>	
<i>too + ? + to</i>	<i>for + ? + of</i>	<i>had + ? + of</i>	<i>many + ? + of + ?</i>

In the *BBI Combinatory Dictionary of English*, Benson et al. (1984) group collocations into grammatical and lexical categories that are further divided into different types by the grammatical and syntactic patterns they follow. A grammatical collocation consists of a dominant content word—noun, verb, adverb and adjective—and a preposition or grammatical structure. For example, *a pleasure to + infinitive* (e.g., *a pleasure to do it* and *a pleasure to meet you*) is a grammatical collocation of the noun + *to* + infinitive type. Lexical collocations are combination of verbs, nouns, adjectives, and adverbs that follow the pattern adjective + noun, noun + noun, verb + noun, etc. Consistent with this structural approach is the work of Justeson and Katz (1995) who use the AN, NN, AAN, ANN, NAN, NNN, NPN (A: adjective, N: noun, P: preposition) part-of-speech tag patterns for collocation filtering when extracting collocations from a corpus text.

### 2.3 Collocation learning

The importance of collocations in successful language learning was recognized as early as seventy years ago by Palmer (1933). However, learning them is not as straightforward as one might expect. This section looks at the importance and the difficulty of collocation learning.

#### 2.3.1 The importance of collocation learning

Collocation learning is important from a pedagogical view for many reasons. The following four are based on linguistic and pedagogical research.

##### 1. Language knowledge is collocation knowledge

Nation (2001) argues that language knowledge is collocation knowledge because the storage of chunks of language in long-term memory forms the basis of learning, knowledge and use. He supports Ellis's (2001) contention that language learning and use can be accounted for by association between sequences of words, without the need to refer to grammatical rules. A number of researchers (for

example, Arabski, 1979; Bahns and Eldaw, 1993; Marton, 1977) have pointed to the fact that many errors can be attributed to lack of correct and appropriate use of collocations. Knowledge of collocations can impact on a number of skills. Brown (1974), for instance, believes that oral production, listening comprehension and reading speed can be improved through an increase of their collocation knowledge.

## **2. Learning collocations is a natural way of learning a language**

When children start learning a language, they memorize and retrieve *want to go* as a whole unit *wanttogo*. Later, they learn to segment this previously unanalyzed unit and attach meanings to segmented pieces, whereby they learn to say *want to play*, *want to find*. Nattinger and Decarrico (1992) argue that adults do not go about the task in a completely different way. They suggest that in a relatively natural environment, all language learners seem to go through two stages: they memorize chunks of language in certain frequent and predictable social situations, and then they break these chunks down to construct sentences.

## **3. Collocation knowledge is important for developing both fluency and accuracy**

Why can native speakers communicate more quickly and efficiently than language learners? Hill (2000) explains that the vast repertoire of ready-made chunks that native speakers store in their head enables them to process and produce language at a much greater speed. When listening or reading, they recognize these chunks as units rather than processing everything word-by-word. Along the same lines, Pawley and Syder (1983) suggest that native speakers store most words both individually and in larger chunks. In order to achieve native-like selection and fluency, learners need to do the same thing—store units of language at phrase or clause length as chunks in memory. Lewis (1997) adds that prefabricated chunks allow learners to use expressions that they were unable to construct creatively from rules. Doing this should ease frustration and, at the same time, promote motivation and fluency.

Hill (2000) further emphasizes the importance of collocation knowledge in relation to developing accuracy of expression. Learners often use long, labored, clumsy sentences in speech and writing because they are unable to express

complex ideas lexically. In many cases, the unnatural sentences or phrases they produce can be replaced by collocations. For example, *his disability will continue until he dies* could be avoided if the learner learnt some adjective collocates of *disability*, such as *mental, physical, permanent, severe, and intellectual*: in this case, *he has a permanent disability*.

In summary,

Learners of English as a foreign or second language, like learners of any language, have traditionally devoted themselves to mastering words—their pronunciation, forms, and meanings. However, if they wish to be able to acquire active mastery of English, that is, if they wish to express themselves fluently and accurately in speech and writing, they must learn to cope with the combination of words into phrases, sentences and texts.  
 (Benson et al., 1997)

#### **4. Collocation knowledge is important for improving complexity in both speech and writing**

Lewis (2000) suggests that teachers should encourage their students to see the value of and build up so called “islands of reliability”—formulaic chunks that often occur in fluent speech and academic writing. These can help learners convey the central meaning of what they wish to say, especially if it is complex. Academic writing texts, such as the one shown below, are rich in informational content and contain a high density of noun + noun and noun + *of* + noun phrases (in bold).

*The conceptual framework for the study was derived from an exploration of the research literature which focused on the general field of leadership, educational and the genre of teacher leadership.*

Good writing is characterised not only by accuracy, but also by complexity. This largely depends on the writer’s ability to construct noun phrases. However, learning noun phrases is entirely absent from—or overlooked in—regular EFL classes, including those for English for Academic Purposes (Lewis, 2000).

### 2.3.2 The difficulty of collocation learning

Studies show that an educated native speaker of English knows about 20,000 word families (see for instance, Goulden et al., 1990). However, the size of their mental lexicon—stored as prefabricated multi-word chunks—is larger than was first thought (Lewis, 1997). High-frequency words make up about 80% of the words in running text, and the first 2000 words cover almost 90% of what we say and write (Nation, 1997). It is those hundreds of millions of expressions, idioms, and collocations that make up the language of everyday use. The single most formidable task the learner faces is mastering a sufficiently large lexicon to achieve native-like fluency.

To make the situation more challenging, all lexical items, expressions or collocations, are arbitrary: they are conventionalized language that simply has been used for years. Very few of them were consciously learnt by native speakers. Learners, especially EFL students, do not have constant language exposure, as native speakers do. As a primary language source, they rely heavily on coursebooks from which many features of natural language have been removed (Lewis, 1997). As Wray (2000) states:

Gaining full command of a new language requires learners to become sensitive to the native speakers' preference for certain sequences of words over others that might appear just possible. From the bizarre idiom, through the customary collocation, to the turns of phrases that have no other apparent linguistic merit than that 'we just say it that way', the subtleties of a language may floor even the proficient non-native, not so much because of a non-alignment between interlanguage and target language forms, as because the learner lacks the necessary sensitivity and experience that will lead him or her unerringly away from all the grammatical ways of expressing a particular idea except the most idiomatic.

Learners have a tendency to translate word for word, and think of words that are definitional equivalents in the L1 (first language) and the L2 (second language). Teachers who speak the learner's L1 understand why they often make collocation

errors like *strong smoker* instead of *heavy smoker*, *powerful tea* (for *strong tea*), and *big rain* (for *heavy rain*). The study of Biskup (1992) on the English of Polish and German students confirmed the influence of L1 on the production of L2 collocations. Collocation is a notoriously challenging aspect of English productive use even for advanced learners (Bishop, 2004; Nesselhauf, 2003). Bahns and Eldaw (1993) investigated the productive knowledge of 58 German EFL students in translation and cloze tasks. The results show that collocations are the major cause of poor writing performance.

Collocation learning has been peripheral in the classroom, especially in EFL teaching. Teachers are under pressure from curricula that are traditionally grammar focused and exams that are used to evaluate their teaching performance. They have to decide how best to use the limited class time. For teachers whose mother tongue is not the target language, grammar and individual words are relatively easy to teach and assess. Learners tend to believe that single words are the units of meaning and, without adequate guidance, have no means of distinguishing useful collocations from the mass of possibilities. Consequently, they fail to notice collocations and even to understand their existence and importance (Bishop, 2004).

Another difficulty that teachers and learners face is that there are few resources for checking which collocation is correct by looking up it. Many non-native teachers still use out-of-date dictionaries rather than modern ones with many thousands of corpus-based examples. Few coursebooks address collocations, explicitly and most teachers are forced to rely on intuition (Conzett, 2000).

## *2.4 Collocation teaching*

Despite wide recognition of the importance of collocations in language learning, it remains largely unclear how they should be taught. This section looks at strategies and activities that teachers have developed to help their students explore collocations and retain them in long-term memory, and further expand and enrich their collocation repertoire. Then it discusses what kind of collocations researchers recommend should be selected and prioritized for learning.

### 2.4.1 Teaching strategies

Teaching collocations is difficult; therefore adopting effective strategies is important. General practice involves three aspects: awareness-raising, deliberate teaching, and recording and recycling.

#### **Awareness-raising**

Collocations are arbitrary. Many extremely useful collocations slip by unnoticed and are therefore not stored and reused by learners. Before beginning to notice this kind of language for themselves, learners need to be aware why we say *make an appointment* rather than *create an appointment*.

To draw attention to this phenomenon and help learners develop an understanding of the kinds of chunk found in texts, many researchers suggest that class time is better spent raising awareness and encouraging effective recording of collocations rather than concentrating on individual items (Woolard, 2000; Conzett, 2000; Lewis, 2000). Teachers can help students divide up texts containing familiar items into chunks and seek patterns in them. Chunking can take place while listening to stories or performing reading and writing tasks (Nation, 2001). Lewis (1997) adds that important collocations should be presented in the classroom, and students should be trained to learn them as a whole and break them into parts later. Conzett (2000) recommends selecting books that include many collocations and introducing them to students in certain contexts, training students to observe and note as many collocations as possible through reading and then reinforce them in writing. Woolard (2000) and Lewis (1997) suggest providing students with a selection of mis-collocations they have made in their production of language, to stress that not all individual words can be combined freely.

Teachers have developed many awareness-raising exercises to help students notice and select useful collocations. For example, Hill (2000) suggests that students underline all verb + noun collocations in a text, and take a common word and find as many collocates as they can. More activities will be described in Section 2.4.3. Given limited class time, it is important that teachers equip students with skills that enable them to study collocations by themselves outside the classroom. Computer concordancers are a useful consulting tool. Hoey (2000)

suggests that students use them to explore natural-occurring collocations, and study the same collocations in different text.

### **Deliberate teaching**

Collocations will not take care of themselves, and must be deliberately taught. Teachers should devote more class time to learning multi-word items rather than individual words, and recycle partially known words by actively introducing additional collocations to extend what students already know (Lewis, 2000). When teaching a new word, Hill (2000) encourages teachers to present some of its most common collocations at the same time, and further stresses that a new word—particularly a noun—should never be taught without giving a few common collocates. For example, when introducing the new word *storm* also teach *snow storm*, *dust storm*, *winter storm*, *thunder storm*, *desert storm*, and *tropical storm*. Lewis (2000) highlights the importance of this approach, as it helps students widen their understanding of what those words mean and—more importantly—how they are used.

Teaching collocations makes students more precise. Learners, especially lower level ones, tend to overuse common words such as *very* because of their limited storage of adverb modifiers. It is a good idea to introduce some common and useful modifiers when teaching an adjective or verb, for example, *completely*, *physically*, *mentally*, *emotionally* for *exhausted* and *heavily*, *strongly*, *deeply*, *easily*, *unduly* for *influenced by*.

Teaching collocations also helps students learn de-lexicalized words. De-lexicalized words such as *thing*, *way*, *get*, *take*, and *put* carry little or no meaning in themselves. In general, the more de-lexicalized a word, the wider its collocational range. It is important that these words are met, acquired and recorded in collocations. Teaching collocations of common de-lexicalized words is a far more productive way for learners to spend their time and energy than studying unusual new words (Lewis, 1997). For example, one of the best ways to make one's spoken English more natural is to learn expressions that use the verb *get*, such as *get a chance to*, *get a kick out of*, and *get around to*.

**Table 2.1 Usage note for the word *discretion***

<b>word</b>	<b>special context</b>	<b>collocations</b>	
discretion (n)	caution/privacy, authority, judgment	prepositions	<i>at your/someone's discretion</i>
		verbs	<i>exercise discretion</i> <i>handle something with discretion</i> <i>use discretion</i> <i>leave to somebody's discretion</i>
		adjectives	<i>complete/total/utmost discretion</i>
examples	<p><i>There are no service charges added to the bill. Tip at your discretion.</i></p> <p><i>He handled the private matter with complete discretion.</i></p> <p><i>The job applicants were hired at the discretion of the hiring committee.</i></p>		

Last but not least, for learners with specific learning purposes, teachers can select and introduce particular groups of collocations such as ones related to a topic, or ones for writing, such as *evidence suggests*, *recent findings support*, and *draw conclusions*.

### Recording and recycling

Teachers have developed many strategies to reinforce and consolidate what students have learnt. Recording and recycling are two. It is common practice for teachers to ask students to keep a notebook for writing down words they have encountered that they think are important for later review. Lewis (1997) and Conzett (2000) suggest that learners collect useful collocations day by day as they meet them in text and conversation, and carefully and systematically organize them with the help of dictionaries or other resources. They recommend arranging collocations in three ways:

- grammatically: noun + noun, adjective + noun, verb + noun
- by useful words: *do, make, get, speak*
- by topic: holiday, travel, work, interview.

Collocations can be indexed alphabetically and associated with complete expressions, usage notes, example sentences and other helpful information. Table 2.1 shows an entry for the word *discretion* suggested by Conzett (2000). It comprises the context in which the word commonly occurs, the prepositions,

verbs and adjectives it collocates with, and example sentences demonstrating the usage of these collocations.

It is unrealistic to expect learners to acquire a word that they have only encountered once. Recycling or repetition is a common strategy that teachers employ to help learners retain vocabulary in long-term memory. Recycling can occur through extensive reading and exposure to the target language outside the classroom. Teachers consciously recycle what their students have learnt by repeating certain kinds of activity that will be introduced in Section 2.4.3—for example, reviewing a collocation a few days after the initial encounter.

#### 2.4.2 Collocation selection

From the tremendous number of possibilities, how should collocations be selected for students to learn? Brown (1974) uses the notion of “normal” and “unusual” collocations, and recommends that “normal” ones be taught because they form the basis of “unusual” ones. However, he does not define what “normal” or “unusual” collocations are and implies that they are largely based on intuition. Other researchers propose frequency-based selection. Channell (1981) suggests that words should be presented with high-frequency collocates when they are first encountered by learners, while Nation (2001) adopts two main criteria—frequency and range. Attention is given to very frequent and immediately useful collocates, and then the range of related collocations taken from different contexts is dealt with. Yorio’s (1980) selection criteria are based on need, usefulness, productivity, currency, frequency and ease.

Lewis (1997) categorizes collocations in terms of strength and frequency. Strong collocations behave almost as single words, while weak ones are free combinations of common words. Collocations may be any combination of strong and frequent, strong and infrequent, weak and frequent, or weak and infrequent. He criticizes the use of frequency as the sole guide to strength and suggests that good collocations are those that occur more often than is statistically likely. Teachers need to be aware of both strength and frequency when selecting collocations.

Hill (2000) recommends drawing the learner's attention to collocations that follow particular syntactic patterns, such as adjective + noun, noun + noun, verb + adjective + noun, verb + adverb, adverb + adjective and verb + preposition + noun. He stresses the power of nouns in selecting collocations: identify key nouns in the text and then look for noun, verb and adjective collocations. He also suggests that teachers think of collocation on a spectrum, with weak and strong collocations at each end and medium-strength ones in the middle. It is those of medium-strength that are particularly important for learners, because they make up a large part of what we say and write every day. However, Hill (2000) does not describe how to differentiate them.

#### 2.4.3 Collocation activities

The book *Teaching Collocations* (Lewis, 2000), with contributions by practicing teachers and researchers, contains a large collection of activities designed for different teaching purposes, such as preparing essays, raising awareness, enhancing precision, and improving retention. This section introduces some typical activities that teachers use in the classroom. In practice, of course, they overlap. For example, some awareness-raising activities also serve to enhance precision.

##### **Preparing essays**

The ability to write good essays in another language is one of the most difficult, but important, skills that learners need to acquire. Writing requires good command of language, which demands productive knowledge in the extreme. Teachers often complain that learners lack ideas about what to write, while learners who have good ideas struggle to put them into words. Teaching collocations excels in this respect. Before writing, students brainstorm topic and essay-type-related collocations, where essays may be narrative, descriptive or argumentative. They start with collecting nouns strongly associated with the topic of the essay, and then look for verbs and adjectives that collocate with the noun, and then adverbs for verbs and adjectives.

When giving essay feedback, teachers provide collocation-oriented suggestions. One teacher (Hill et al., 2000) uses the following procedures:

- highlight clumsy phrases that can be replaced with collocations,
- give the essay back to students who then work on those phrases,
- provide the correct collocations if students were unable to produce them themselves, and
- give the essay back to students again for a final revision.

### **Raising awareness**

Exploring text is one of the most common awareness-raising activities. Students read an article and mark collocations in a text, forcing them to notice larger chunks rather than individual words. Teachers ask students to focus on collocations of particular syntactic patterns—for example, underlining nouns and then highlighting which verbs are used before them—or picking those of special interest.

The reverse version is to reconstruct the content of an article. After reading an article, one group of students writes down ten collocations. Another group reconstructs the original text based on the collocations the first group provides. This forces students to seek collocations that carry the main ideas of a text, and makes them more aware of collocations as an essential carrier of meaning.

Given the topic of an article, students compete to predict words they think will occur in it. This traditional pre-reading game is often played in the classroom to stimulate interest and facilitate comprehension before students begin reading. It can also serve as a retrospective activity, where students recall and review a list of expressions and collocations that are important for accurately expressing the ideas relevant to the article.

### **Enhancing precision**

To help students express ideas more precisely, teachers have developed many activities using collocation dictionaries.

*Find a better word* asks students to use a dictionary to find a better way to express each of these:

*a bad effect*

*a big effect*

*an effect that helps*

*an effect nobody expects    a very funny effect    an effect that put things right*

*Near synonyms* helps students differentiate between commonly confused word pairs such as *injury* and *wound*, *clothes* and *cloths*, *beside* and *besides*, or between words of similar meaning such as (1) *task, job, word, career, occupation, profession*; (2) *mistake, error, fault, problem, defect*. The difference between these words rests largely on the difference in their collocational fields.

*Correcting common mistakes* requires students to correct collocation mistakes in sentences. In the example *I was completely **disappointed** when I failed my exam*, students need to look up the word in bold, determine the possible collocates of *disappointed*, and pick ones that are most appropriate in the given context; in this case, *utterly* or *bitterly*.

*Alternative to very* asks students to find other words with a meaning similar to *very*, but stronger or more precise. For example, *very* can be used with the following adjectives. Students use a dictionary to look for alternatives.

*exhausted*

*encouraged*

*disorganized*

*unexpected*

*handicapped*

*recommended*

*disillusioned*

*prepared*

## Improving retention

Learning collocations is a daunting task. Teachers use game-like activities to help students maintain high motivation in the process of transferring what they have learnt to long-term memory.

Collocation domino games can be created using noun + noun collocations, as shown below, or other patterns such as noun + *of* + noun, verb + noun or adjective + noun:

*blank cheque — cheque book — book club — club sandwich — sandwich board — board room ...*

Teachers provide the first collocation (or the last, or both) and students fill in the rest, making the chain as long as possible. One variation is that students use words of other syntactic types, for example, *book a hotel* rather than *book club*, using a verb instead of a noun.

*Odd one out* asks students to delete the word that does not form a strong partnership with the given ones. In the example below, *smoker* is out because *strong smoker* is not a good collocation.

*STRONG language, smoke, accent, indication*

In *Collocation Guessing*, learners are given several verb or adjective collocates of a hidden noun word that must be guessed. For example,

*plain, dark, white, bitter, milk, bar of—chocolate;*

*huge, growing, profitable, export, domestic, black—market.*

Collocates of *chocolate* or *market* are presented one by one until the learner guesses the word. Learners can compete to see who needs the fewest hints.

In *Finding collocation partners*, given two lists of words, one containing adverbs and the other adjectives, students match parts of collocations to form strong adverb + adjective partnerships and then use them to fill in the blanks in the sentences given below. Here is an example:

List 1	List2
<i>carefully</i>	<i>situated</i>
<i>highly</i>	<i>overcrowded</i>
<i>dangerously</i>	<i>chosen</i>
<i>ideally</i>	<i>qualified</i>
<i>The disco was already</i>	<i>... when the fire started.</i>

## 2.5 Resources

Collocation resources are widely used both within and outside the classroom.

### 2.5.1 Collocation dictionaries

Printed dictionaries are traditional language resources for finding word definitions and common usage. With widespread recognition of the importance of collocations, modern general-purpose dictionaries pay more attention to collocations by including them as a part of word entries. For example, *The Oxford Advanced Learners' Dictionary* (OALD sixth edition 2000) contains about 10,000 collocations. However, this is rather a small amount compared to the sheer number of collocations in a language. Cowie (1981) criticizes the inconsistent presentation of collocations and suggests that more should be introduced in general pedagogical dictionaries. In recent years, several dedicated collocation dictionaries have emerged. They serve as reference tools that help users decide which collocations to use on their own. This section introduces four—three printed and one electronic—in terms of scope, intended users, organization, look-up method and illustrative examples.

The *BBI Combinatory Dictionary of English* (Benson et al., 1986) focuses on “essential grammatical and lexical recurrent word combinations.” The revised version (1997) contains 18,000 entries and 90,000 collocations, and claims that it covers material that cannot be found in existing dictionaries for second language learners. Collocations are divided into eight grammatical and seven lexical categories (Section 2.2). The words are alphabetically ordered and indexes are provided. Each word entry contains a few examples (one to three). However, presenting lexical and grammatical collocations together may confuse users, so this dictionary is more useful for academic learners who are familiar with the reference materials and for whom grammatical accuracy is a priority (Lewis, 2000).

The goal of the *LTP Dictionary of Selected Collocations* (Hill and Lewis, 1997) is to help intermediate and advanced learners to use words they already know more effectively. Collocations are grouped into noun, verb, adjective and adverb sections. The five most important collocation types are identified as: adjective + noun, verb + noun, noun + verb, adverb + adjective and verb + adverb. For each one, a headword is selected. Headwords, also called entry or index words, are the

words that are used to look up collocations. To find a collocation, use the noun if it comprises a noun, otherwise use adjective, verb and adverb in that order. Collocations containing common adjectives such as *good*, *bad*, *big*, and *small*, and adverbs such as *very*, *really*, *rather*, and *quite* are omitted. However, the commonness of a word is largely determined by the author's intuition. Collocations are presented in a simple list format. Examples are not available in this dictionary.

In recent years, many collocation dictionaries were compiled based on the study of large corpora: the *Oxford Collocation Dictionary for Students of English* (2009) and *Collins Cobuild's English Collocations*. The first is based on the 100 million words in the British National Corpus (Section 2.5.3) and covers over 150,000 collocations for 9,000 headwords. It aims to help students speak and write native-like English, and claims that except for totally free combinations and extremely idiomatic ones, a full range of collocations is included:

- fairly weak collocations: *see a film* and *an enjoyable holiday*
- medium-strong collocations: *see a doctor* and *direct equivalent*
- the strongest and most restricted collocations: *see reason* and *burning ambition*.

Figure 2.1 shows an excerpt of collocations for the noun word *cause*, grouped by word sense. In some cases, a brief explanation of the sense is given. For example, the noun *cause* has three senses (because of space restrictions, only two are shown). It can be used with (1) a list of adjectives such as *real*, *root*, *true*, indicated by ADJ., (2) the verb *discover*, *find*, *identify* in the form of verb + *cause*, and (3) the verb *be* and *lie* in the form of *cause* + verb. Examples are provided for some collocations.

The *Collins Cobuild's English Collocations* published on CD-ROM and based on the 200 million words in the Bank of English, provides 140,000 collocations and 2,600,000 examples. It defines collocations as frequent word combinations, including idioms, phrasal verbs, compounds, fixed phrases and grammatical patterns. To find collocations, the user selects a target word from a list of 10,000 words of English. Clicking that word brings up a screen displaying the twenty

OXFORD Collocations   dictionary for students of English	
<b>cause</b> noun	
<b>1</b> sb/sth that makes sth happen	
<b>ADJ.</b> <b>real, root, true, underlying</b> <i>the root cause of the problem</i>   <b>deeper, greater</b> <i>A greater cause for resentment is the discrepancy in pay.</i>   <b>biggest, chief, clear, fundamental, important, main, major, primary, prime, principal, significant</b>   <b>common</b> <i>Smoking is a common cause of premature death.</i>   <b>likely, possible</b>   <b>known, unknown</b>   <b>direct, indirect</b>   <b>immediate, initial</b>   <b>long-term, short-term</b>   <b>reasonable</b> <i>There is no reasonable cause to suspect an unnatural death.</i>   <b>contributory</b>   <b>hidden</b>   <b>social</b> <i>the social causes of ill health</i>	
<b>VERB + CAUSE</b> <b>discover, find, identify</b> <i>attempts to identify the immediate cause of the breakdown</i>	
<b>CAUSE + VERB</b> <b>be, lie in sth</b> <i>The real cause of the problem lies in the poor construction of the bridge.</i>	
<b>PHRASES</b> <b>cause and effect, the cause of death, due to/from/of natural causes</b> <i>He died of natural causes.</i>	
<b>2</b> reason	
<b>ADJ.</b> <b>good, great, real, reasonable</b>	

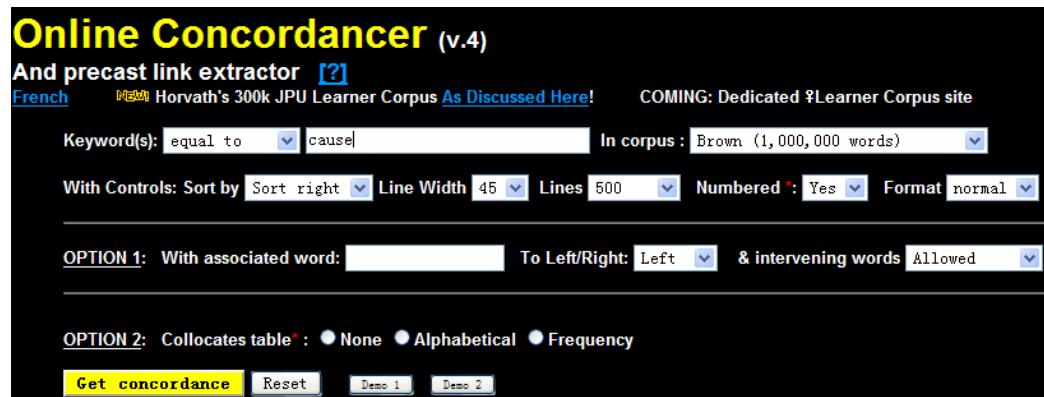
**Figure 2.1 Entry in the *Oxford Collocation Dictionary for Students of English***

most frequent collocates that occur on either side of it. Clicking a collocate shows twenty randomly selected examples in a typical concordance format. Each example can be expanded to show more contexts. Despite the large volume of text this dictionary is based on, it is disappointing that it only provides twenty collocates, which often include common words such as *any, own, and new*.

### 2.5.2 Concordancers

A concordancer is “a piece of software, either installed on a computer or accessed through a website, which can be used to search, access and analyse language from a corpus” (Peachey, 2005).

One accessible and user-friendly concordancer, shown in Figure 2.2a and available on the Web, is the *Compleat Lexical Tutor* from Université du Québec à Montréal (Cobb, n.d.). Using this tool, students can enter a word and explore what words are most likely to occur before or after it. They specify a keyword to search for, and select one of a number of different corpora to search in. They can also associate another word with the keyword, specifying a position—left, right or any. The search results are chunks of text (constrained by line width) that contain the



### (a) The interface

```

1 n of material by hypervelocity impacts would cause a deviation from a linear relationship. In t
2 he vagina as the woman lies on her back) may cause a great deal of distress in a virgin. But du
3 vert, he devoted himself to the anti-slavery cause. A group of young men influenced by him enro
4 6-inch torrent of water is released, it may cause a lot of comment as it passes through or by
5 is no substitute for those intangibles which cause a man to sacrifice part of his earning poten
6 utter away the surface atoms of the dust and cause a slow diminution in size, with a resultant
7 rather than cuts for the texture (cuts could cause air pockets under the glaze creating pinhole
8 ct to stand or fall only by the merits of my cause". All seven recognized that independence
9 pressure and pain signals may involuntarily cause all the vaginal muscles to contract in an ef
10 ay fail to elicit a conditioned reaction but cause an increased synchrony instead of the excita
11 ve his wife the following description of the cause and consequence of diarrhoea: "I have bin a
12 ing action on the empirical determination of cause and effect provides a toughness and bravado
13 ituation in which he can perceive no visible cause and effect sequence, he should be alert to i
14 mechanistic universe, governed by the laws of cause and effect, bound in chains of determinism
15 any line of action related to the concept of cause and effect. He bases his approach on the bel
16 es with virtually no understanding of social cause and effect. Small wonder, then, that we fear
17 Pp. 64-66. Haydn C. Covington argued the cause, and filed a brief for petitioner. H17 0550
18 e tiresome vocabulary of that lost and dying cause, and in the sprung syntax that is supposed t

```

### (b) Concordance entry

**Figure 2.2 Online concordancer at [www.lextutor.ca](http://www.lextutor.ca)**

keyword and, if specified, the associated word. Figure 2.2b shows the result of searching for the word *cause*, which is underlined. A line width parameter determines the size of the context that is displayed (here it is 45 characters).

More complex concordancers allow users to search using regular expressions or even discriminate between spoken and written language use. The British National Corpus website provides an example.<sup>1</sup> Users use the equal (=) character to restrict the search by part-of-speech, and braces { and } to enclose a regular expression. Unlike the previous example, the result comprises a list of complete sentences, each with an associated sequence number—for example, AA9—that links to a page displaying a surrogate of the document containing the sentence, including the title, author, publisher and total word count.

---

<sup>1</sup> <http://sara.natcorp.ox.ac.uk/lookup.html>

Support for learner use of corpora and concordancing is premised on the fact that exposure to a word in different contexts, both lexical and grammatical, allows learners to develop a greater sense of its meaning. Many features associated with using a concordancer to analyse and present word and collocation information may also lead to better retention of vocabulary items. Concordancing provides for multiple or repeated exposures, and in using a concordancer students are likely to be motivated by the need to use a word—one of the three components identified as part of Hulstijn and Laufer’s Involvement Load Hypothesis (2001). Hulstijn and Laufer suggest that the involvement load is high, and therefore students are more likely to learn and retain vocabulary items if the need for particular items is determined by the learner rather than the teacher. This is indeed the case if students are using a concordancer as a resource to help them improve their own writing, both to generate language items and to review ones they have already used.

### 2.5.3 The British National Corpus

The British National Corpus (BNC) contains a wide range of written (90million words) and spoken (10million words) British English language. The written text come from newspapers, specialist periodicals and journals, academic books and fiction, published and unpublished letters and memoranda, as well as school and university essays. The spoken text comprises orthographic transcriptions of conversations, and spoken language collected from business or government meetings, radio shows and phone-ins. The work of building this collection started in 1991 and lasted three years. The latest version, published in 2007, is distributed in XML format.

Figure 2.3 shows an excerpt of a written news article (indicated by *wtext* and *NEWS*) with the heading (marked by *<head>* element) given on the right side. Each segment is marked by an *<s>* element, which contains *<c>* elements for punctuations and *<w>* elements for words. They contain the following attributes:

- *c5* attribute: part-of-speech tag from the CLAWS5 tagset (Section 4.3.2),
- *hw* attribute: root form of the word, and
- *pos* attribute: simplified part-of-speech tag.

```

- <wtext type="NEWS">
- <div level="1" n="21-DEC-89 edition, page 18">
- <head>
  <s n="1">
    <w c5="ATO" hw="a" pos="ART">A</w>
    <w c5="NN1" hw="country" pos="SUBST">Country</w>
    <w c5="NN1" hw="diary" pos="SUBST">Diary</w>
    <c c5="PUN">:</c>
    <w c5="NPO-NN1" hw="eggleston" pos="SUBST">EGGLESTON</w>
    <w c5="NN1-VVB" hw="burn" pos="SUBST">BURN</w>
    <c c5="PUN">,</c>
    <w c5="NPO" hw="teesdale" pos="SUBST">Teesdale</w>
    <c c5="PUN">:</c>
    <w c5="NPO" hw="small" pos="SUBST">Small</w>
    <w c5="VVB" hw="burn" pos="VERB">burns</w>
    <w c5="CJT" hw="that" pos="CONJ">that</w>
    <w c5="VVB" hw="feed" pos="VERB">feed</w>
    <w c5="ATO" hw="the" pos="ART">the</w>
    <w c5="AJ0" hw="main" pos="ADJ">main</w>
    <w c5="NN2" hw="river" pos="SUBST">rivers</w>
    <w c5="PRF" hw="of" pos="PREP">of</w>
    <w c5="ATO" hw="the" pos="ART">the</w>
    <w c5="AJ0-NN1" hw="pennine" pos="ADJ">Pennine</w>
    <w c5="NN2" hw="dale" pos="SUBST">Dales</w>
    <w c5="VVB" hw="create" pos="VERB">create</w>
    <w c5="DTO" hw="some" pos="ADJ">some</w>
    <w c5="PRF" hw="of" pos="PREP">of</w>
    <w c5="ATO" hw="the" pos="ART">the</w>
    <w c5="AJ0" hw="wild" pos="ADJ">wildest</w>
    <w c5="NN2" hw="feature" pos="SUBST">features</w>
    <w c5="PRF" hw="of" pos="PREP">of</w>
    <w c5="ATO" hw="the" pos="ART">the</w>
    <w c5="NN1" hw="landscape" pos="SUBST">landscape</w>
    <c c5="PUN">.</c>
  </s>

```

*A country Diary:  
EGGLESTON BURN,  
Teesdale: Small burns  
that feed the main  
rivers of the Pennine  
Dales create some of  
the wildest features of  
the landscape.*

**Figure 2.3 Excerpt of a BNC XML document**

## 2.6 Corpus-based language learning

Corpus linguistics has moved beyond the realm of pure linguistics and become of interest to those involved in language teaching and learning. As Gabrielatos (2005) states, “Corpus has now become one of the new language teaching catchphrases, and both teachers and learners alike are increasingly becoming consumers of corpus-based educational products, such as dictionaries and grammars.”

Most corpora are based on particular domains, genres, or collections of certain types of document from which recurrent phrases and grammatical patterns can easily be retrieved (Stubbs and Barth, 2003). A corpus is therefore a particularly productive context in which to study collocations. Various kinds of corpora have been compiled for different study purposes; for example, multilingual, monolingual, parallel, aligned and learner corpora. Students can compare their language use with expert use by building vocabulary profiles for text written in their course assignments.

Peachey (2005) summarizes four ways of using a corpus in language learning:

1. exploring collocations, which helps students develop awareness of language patterns,
2. looking at errors, which helps students identify common language errors,
3. understanding different meanings, which helps students learn polysemic words, and
4. finding genuine examples, which exposes students to the language in authentic context.

Sinclair (2004a) adds three more dimensions of use:

5. analyzing semantic preferences, or co-occurrence of items that share semantic features,
6. exploring colligation, or co-occurrence of grammatical phenomena, and
7. discovering semantic prosody, or the positive or negative verbal environment in which an item commonly occurs.

Fuentes (2003) conducted a study using a corpus-based approach to improve student performance in oral business English presentations. The study used two types of corpora: academic, made up of written textbook material and articles introducing basic business concepts, and professional, comprising oral business reports and product reviews. Students were recruited for the experimental and control group and assigned the same oral presentation task. The experimental group participated in corpus-driven activities for two weeks, including identifying clusters and patterns, examining a glossary, and doing fill-in-the-gap exercises. The study confirmed the positive influence of corpus-based development, and found that learners produced more semi-technical business English collocations, non-business English clusters and technical compounds in their oral presentation.

Chambers and O'Sullivan (2004) investigated the importance of corpus consultation as a new type of literacy in the context of language learning. In their study, eight postgraduate students consulted concordancing tools to help improve their writing skills in French. Teachers underlined errors in the student's written text and placed an *x* to indicate basic inaccuracies such as gender, agreement, verb form etc. Then students were asked to correct the errors by consulting a concordancer and record changes as a direct result of this consultation. The study

shows that consultation helped students identify and correct basic errors like gender, agreement between nouns and adjectives, using capital letters in expressions such as *président de la République*, misspelling, and grammatical and lexical-grammatical patterning errors.

Despite the widespread adoption of corpus-based language learning, the application of computer corpora for language teaching is still a neglected area (Chambers and O'Riordan, 2006). Such corpora have three limitations. First, although their use has gained predominance in tertiary education, it is still conspicuously absent in secondary education and general ELT classes. Second, “a corpus is not a simple object” (Sinclair, 2004b), and most learners find it difficult to handle the complex information it provides. Third, texts in corpora tend to be of little interest to learners—they know nothing about the author of the message in a concordance line and their illocutionary intentions (Braun, 2005).

Chambers and O’Sullivan (2004) urge pedagogical mediation by teachers through the preparation of corpora that are meaningful to their students, making corpus data relevant for specific learning purposes and training students on corpus analysis and consultation skills. The importance of pedagogical mediation is echoed by Braun (2005), and Kaltenböck and Larcher (2005). The former proposes the use of small genre-specific corpora or corpora created by teachers and learners themselves, the incorporation of other data formats such as audio or video alongside the text, and the addition of annotations to facilitate multi-dimensional access to corpora content. Braun also suggests that corpora material should be complemented with comments and explanations, exploratory tasks and exercises, and study aids for learners and teachers. Kaltenböck and Larcher argue that learners should not just observe, but should be encouraged to read the corpus text and carry out language learning tasks that involve exchanging information with other learners who may have read similar material.

## 2.7 *The Web corpus*

The Web is a potentially useful corpus for language study because it provides examples of language that are contextualized and authentic. The most striking, and perhaps the most compelling, feature of the Web for language teachers, and

developers of teaching resources, is its size. However, this brings its own problems. Web content is heterogeneous in the extreme, uncontrolled and hence “dirty,” and exhibits features different from the written and spoken texts in other linguistic corpora. This section looks at these features in terms of size, representativeness, and cleanliness.

### 2.7.1 Size

The size of the Web far outstrips any existing corpus and grows on a daily basis. Kilgariff and Grefenstette (2003) show this in their comparison of frequencies of a set of English phrases. For example, the phrase *perfect balance* occurs in the British National Corpus 38 times, as opposed to 355,538 in Spring 2003 using AltaVista as the search engine, and 1,910,000 today (August 2010, using Google).

The continual addition of new text has drawbacks, however, for it makes individual search results inconsistent and unstable. Indeed, Biber and Kurjian (2007) observe that linguistic patterns found on the Web can vary radically—and seemingly randomly—from one search to the next. Therefore, when teachers set certain kinds of exercises involving direct Web search they cannot rely on what they will retrieve or know exactly what their students will see. This is a serious disadvantage.

### 2.7.2 Representativeness

Most corpora are based on particular domains, genres, or collections of certain types of documents from which recurrent phrases and grammatical patterns can easily be retrieved (Stubbs and Barth, 2003). However, this certainly cannot be said about the Web taken as a whole. More than a decade ago, Kessler, et al. (1997) characterized it as a large and heterogeneous domain. Since then it has grown many-fold in both size and diversity.

Biber and Kurjian (2007) recognize that identifying genre is an especially important consideration for linguistic research based on the Web, but acknowledge the difficulty of doing so. Search engines and other portals impose various taxonomic structures on Web items and resources. As Meyer (2002) notes, Yahoo categorizes documents and websites into fields such as *Arts and*

*Humanities* and *Science Education*, each having further subcategories—both in terms of the content itself, and of information sources such as journals or magazine articles. Similarly, Robb (2003) explores limiting searches to within particular educational domains using site names ending in *edu*, *ac.uk*, *edu.au* and *jp*. However, these categories are still broad and not particularly useful for language study.

Biber and Kurjian (2007) used the two categories *Home* and *Science*, with their respective subcategories, to explore linguistic differences amongst Web-based texts. They conclude that there is wide variation within each category and subcategory, and substantial overlap in the occurrence of a large number of linguistic features. In other words, the categories imposed by search engines reflect little or no consistency between the genres of the documents that fall under them.

To what extent does the text found on the Web resemble or differ from that in traditional hardcopy form? Meyer (2002) asks the question in this way: are electronic texts essentially the same as traditionally published written texts? Apart from online journals, newspapers, and advertising material, most of the text on the Web—for example, documents posted on personal home pages or constructed on blogs—has not been subjected to any editorial process. This is in clear distinction to traditional commercially published text, for which the economics of publishing dictate quality control mechanisms that affect and to some extent normalize the writing style.

According to Biber and Kurjian's (2007) study, identifiable Web-based text types include: personal, involved, stance-focused narration, persuasive/argumentative discourse, addressee-focused discourse, and abstract/technical discourse. Two of these types (personal, involved, stance-focused narration; and addressee-focused discourse) appear to be particular to the Web. Some features that characterize the former are: first person pronouns; mental verbs such as *think*; certainty adverbials such as *certainly*, *definitely*, *surely* and *undoubtedly*; *that*-clauses; the pronoun *it*; and past tense. Some that characterize the latter are: second person pronouns, progressive verbs, desire verb + *to*-clause (Biber and Kurjian, 2007).

The complexity and variety of Web text means that searches produce results that are anomalous with those obtained by searching corpora based on written material, which are necessarily focused and selected—and even with those based on spoken material.

### 2.7.3 Cleanliness

The Web contains a huge number of language errors such as grammatical and spelling mistakes, not to mention the use of unusual and less acceptable collocations. Kilgariff and Grefenstette (2003) describe it as a “dirty corpus.” This represents a rather serious constraint on its use for language learners, because a fundamental requirement for such texts is that they represent exemplary models of language. One response is to limit searches to impeccable sources (Robb, 2003). Robb describes how to use Project Gutenberg, a huge collection of e-texts of material that is out of copyright, particularly works of literature and texts of historical value (Robb, 2003).

## 2.8 *Using the Web corpus*

Because of its massive volume of natural text, researchers, teachers and learners are turning their attention to the Web. The fact that it is a rich source of data for linguistic analysis is evidenced by projects such as WebCorp (Renouf et al., 2007), an online web application, developed at Birmingham City University, and KWICFinder, a downloadable desktop application, developed by Fletcher (2005). Both work on top of a search engine and search the live Web for concordances that are similar to those derived from ordinary corpora. The newest version of KWICFinder switched from AltaVista to the Yahoo search engine. WebCorp initially utilized standard Web search engines such as Google, but its latest refinement allows users to choose a particular one—Google, AltaVista/Yahoo, Bing, Ask, Metacrawler, or Open Directory.

Users enter a word or phrase, and choose options such as concordance span, uppercase/lowercase, maximum number of web pages to retrieve, site domain, etc. Using the list of URLs returned by a search engine, the page content is retrieved and concordance lines are extracted and presented to users in HTML, plain text, or

<a href="http://makezine.com/">http://makezine.com/</a>
Document Dated: 2009/09/21 00:00:00 (metatag)
<a href="#">Plain Text</a> <a href="#">Word List</a> 1108 tokens, 650 types
<ul style="list-style-type: none"> <li>• are simple alterations you can <b>make</b> for a cakier or softer</li> <li>• to show you how to <b>make</b> Tops/Spinners using Tea Light</li> </ul>
<a href="http://blog.makezine.com/">http://blog.makezine.com/</a>
Document Dated: Unable to find date
<a href="#">Plain Text</a> <a href="#">Word List</a> 3274 tokens, 1326 types
<ul style="list-style-type: none"> <li>• to use your photos to <b>make</b> cool things such as a</li> <li>• you need this book. Please <b>make</b> sure you include your email</li> <li>• Download the project PDF to <b>make</b> this stunning photo mosaic where</li> <li>• the board. I decided to <b>make</b> things A LOT easier on</li> <li>• a metal bending tool to <b>make</b> these. A drill press is</li> <li>• Kelly Jensen: I want to <b>make</b> skywire7 commented on Book giveaway +</li> </ul>
<a href="http://www.amazon.com/Tipping-Point-Little-Things-Difference/dp/0316346624">http://www.amazon.com/Tipping-Point-Little-Things-Difference/dp/0316346624</a>
Document Dated: 2007/05/09 00:00:00 (author specified)
<a href="#">Plain Text</a> <a href="#">Word List</a> 5355 tokens, 1844 types
<ul style="list-style-type: none"> <li>• a small change that can <b>make</b> a big difference.' (p. 183)</li> <li>• thought and converstaion, and will <b>make</b> me look at life and</li> <li>• group that it takes to <b>make</b> that change. Gladwell's first example</li> <li>• little things or people can <b>make</b> a huge difference in our</li> <li>• Makes Me a Practical Optimist" <b>make</b> a difference and change the</li> <li>• provide? Your comments can help <b>make</b> our site better for everyone.</li> </ul>
<a href="http://well.blogs.nytimes.com/2009/09/16/what-sort-of-exercise-can-make-you-smarter/">http://well.blogs.nytimes.com/2009/09/16/what-sort-of-exercise-can-make-you-smarter/</a>
Document Dated: 2009/09/22 01:51:54 (server header)
<a href="#">Plain Text</a> <a href="#">Word List</a> 3658 tokens, 1539 types
<ul style="list-style-type: none"> <li>• best form of exercise to <b>make</b> you smarter? Walk to the</li> <li>• sure why people did not <b>make</b> this connections decades earlier. The</li> <li>• See, now I have to <b>make</b> my cats work out harder.</li> </ul>

**Figure 2.4 Concordance data returned by WordCorp for the word *make***

XML format. Figure 2.4 shows concordance data for the word *make* provided by WebCorp. It displays the URLs from which the Web pages were retrieved, the date of the page and a list of concordance lines containing the target word.

The developers of WebCorp believe that it offers text domains and types that are not available in other corpora: neologisms; newly-vogueish terms; rare or possibly obsolete terms; rare or possibly obsolete constructions; and phrasal variability and creativity. However, as Renouf et al. (2007) point out, the performance of WebCorp relies heavily on the underlying commercial search engines, and therefore:

- the amount of web text searched is limited,
- the speed of results is inhibited, and
- services such as word count statistics and wildcard search are unreliable and inconsistent.

Seretan et al. (2004) uses text snippets returned by the Google search engine to extract syntactic based collocations—e.g., adjective + noun and verb + noun.

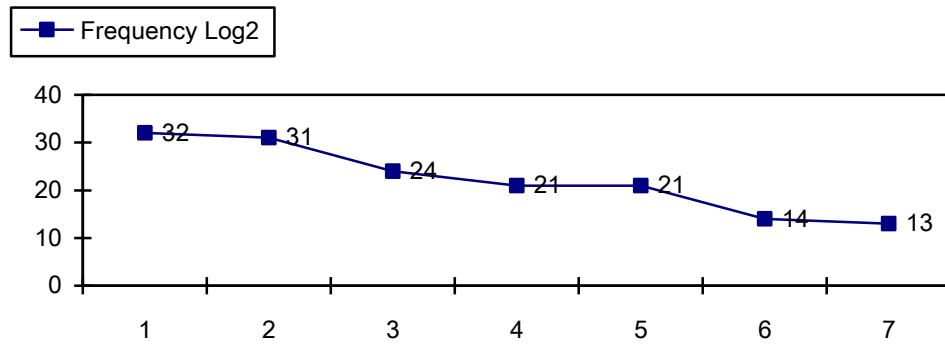
**Table 2.2 Subphrase frequencies for *have been found to be infected/polluted with***

No. of Words	fragment	Google hits	log <sub>2</sub>
1	have	3,040,000,000	32
2	have been	1,870,000,000	31
3	have been found	11,900,000	24
4	have been found to	2,030,000	21
5	have been found to be	1,850,000	21
6	have been found to be infected	15,200	14
7	have been found to be infected with	9,370	13
6	have been found to be polluted	1,140	10
7	have been found to be polluted with	300	8

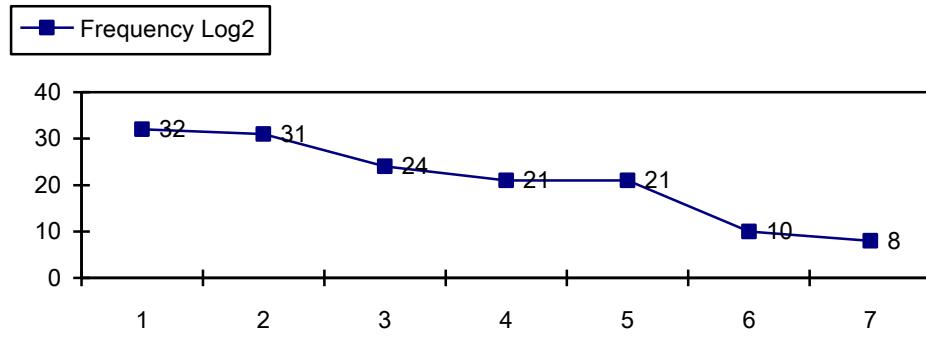
Using a word as the query term, from 100 to 1000 snippets are retrieved and parsed by a syntactic parser to identify bigrams matching particular syntactic patterns. Candidate bigrams and their associated frequency are extracted and ranked by the statistical measures proposed by Manning and Schütze (1999). Seretan et al. evaluated extracted collocations with a human judge and non-native students, and against the BBI dictionary (Benson et al., 1986), but drew no clear conclusions on the performance of their approach. Nevertheless, they recognize several limitations of using snippets for extracting collocations:

- the amount of data obtained from the Web is restricted by the search engine provider (1000 queries per day and 1000 results for a given query),
- the system cannot be used online as a Web application due to the lengthy processing and parsing time, and
- snippets contain many repeated and incomplete sentences, which affects the quality of the collocations.

Shei (2008) used the occurrence counts of successively truncated subsequences retrieved from Google to identify formulaic sequences. He devised a visual tool that represents the frequency of sequential word combinations, and their subsequences. Figure 2.5 shows the frequency lines *have been found to be infected with* and *have been found to be polluted with*, drawn from frequencies



(a) *have been found to be infected with*



(b) *have been found to be polluted with*

**Figure 2.5 Frequency plotted against phrase length**

given in Table 2.2. Frequencies inevitably become smaller as more words are included. Shei asserts that a sequence may be considered formulaic if the frequency line remains stable when new words are added, and that learners could use this to guide their choice of collocation. For instance, *have been found to be infected with* is more formulaic than *have been found to be polluted with* because the frequency line of the former is flatter than that of the latter.

## 2.9 Computer-assisted collocation learning

In recent decades, advanced computer and information technologies have unleashed the power of computers in language learning. The unprecedented growth of the Internet and ubiquity of personal computers has provided opportunities to augment, or even replace, face-to-face teaching by generating

learning activities that are readily accessible to learners outside the classroom. The Web has become a popular and effective place to learn foreign languages. Learners can benefit from the wealth of free resources, such as practice exercises, language courses and language analysis tools like concordancers.

With the advent of intelligent language learning systems, artificial intelligence technologies are used to deal with language problems. According to Bowerman (1993), the first intelligent CALL (Computer-Assisted Language Learning) system was produced to check answers to comprehension questions by using syntactic and semantic knowledge. Recently, intelligent CALL systems have shown a growing reliance on natural language processing research (Debski, 2003). Natural language parsers provide linguistic analyses of written language by representing the syntactic and, sometimes, semantic structure of sentences, and tagging words with their part-of-speech. Although these systems have been criticized for being unable to account for the full complexity of human language (Salaberry, 1996), they have, however, been used to capture interesting fragments or aspects of a given language. For example, Dodigovic (2005) has explored the use of natural language processing technology in developing a program to raise awareness of errors in language production. This program was designed to help Chinese and Indonesian students improve their academic writing by reducing grammatical errors.

Computer language activities have become popular ways of helping learners practise and improve their English, but those for collocation learning are rare and inadequate. This section describes some activities and tools that are either dedicated to or can be used for supporting collocation learning.

### 2.9.1 Collocation exercises on the Web

Surprisingly, there are few collocation exercises on the Web, as opposed to millions of vocabulary and grammar ones. For example, *a4esl.org*, one of the most popular English learning websites, hosts hundreds of language exercises contributed by teachers around the world, of which only two are collocation exercises, each containing ten questions.

**Collocations**  
10 Questions by Jim Papple

rotten, bad, free-range, Easter, scrambled

- 1 cards
- 2 eggs
- 3 hotels

[ Start Again ]    Question's Value: 30    Game Points: 0    [ End Quiz ]

**(a) Adjective + noun collocation exercise**

**Collocations Related to Colour**  
8 Questions by Jim Papple

After the roller coaster-ride his face was completely \_\_\_ of colour.

- 1 assembled
- 2 drained
- 3 flooded

[ Start Again ]    Question's Value: 30    Game Points: 0    [ End Quiz ]

**(b) Collocation exercise related to colour****Figure 2.6 Collocation exercises at *a4esl.org***

Collocation exercises, normally presented as complementary material for vocabulary study, often take the form of quizzes, puzzles, fill-in-blanks, matching, permutation, or games. They are created by teachers who prepare questions, answers and explanations and make them available on language learning websites. Exercises that provide instant performance feedback in an attractive way can be generated with the help of tools like the *Hot Potatoes* software.<sup>2</sup> In general, exercises are scattered across different websites, and the material that they offer lack context, and are limited and fixed by the designer. Teachers may not find them particularly useful for their students and for different teaching purposes.

---

<sup>2</sup> <http://hotpot.uvic.ca>

The two collocation exercises in Figure 2.6 were contributed by a language teacher to *a4esl.org*. In the first, given a list of adjectives or verbs at the top and three nouns below, the student must select the noun that combines best with all the adjectives and verbs—the answer here is *eggs*. The answer (correct or not) is revealed instantly once the student clicks a noun. The second exercise focuses on collocations related to colour, such as *drained of colour*, *wear yellow*, *in the pink*, *shade of purple*, *added colour* and so on.

*www.better-english.com* provides 15 business collocation exercises in multiple choice format. Each one contains 20 questions, focusing on a particular group of nouns or adjectives. A question consists of one or two sentences and a set of choices (from 5 up to 15) that are either nouns or adjectives. The exercise shown in Figure 2.7 asks the student to choose the noun that fits the context presented in each question. The noun is removed from the question text and the student chooses one from a dropdown list that is the same for all questions. The answers and scores are given when the student clicks the *check* button at the bottom of the page. The words in the dropdown list appear to be random; in this case, there is no obvious pedagogical explanation for studying *belief*, *bill*, *blunder*, *bias*, *blame*, *benefit*, *behaviour*, *blow*, *battle*, *beach*, *blast*, and *bitterness* together.

The eleven collocation exercises offered by *angelfire.com* take the form of drag-and-drop, matching, and gap filling. Figure 2.8a shows a drag-and-drop exercise implemented using Macromedia Flash technology in which adjectives in a sentence are replaced by dashed lines. The student drags an adjective from the right side and drops it onto a dashed line. The score is given automatically once all the questions are answered. The gap filling exercise shown in Figure 2.8b, asks students to choose a word from the dropdown list to fill in the gap in a sentence. A hint is given, in this case *something that seems to exist although it may not*, by clicking the question mark button. Exercises focusing on two-word collocations are shown in Figure 2.8c, in which students match the words on the right side to those on the left to form valid collocations, such as *heavy traffic*, *narrow margin*, and *closely-guarded secret*.

1. I don't see any point in continuing to compete in the Japanese market. We're fighting a losing _____.	blame
2. I need to get away from everything and lie back and relax on an unspoilt _____ somewhere.	??? belief bill blunder bias
3. I'm appalled by your callous _____. That is no way for a responsible company to act.	blame
4. He continues to cling to the _____ that he can do everything. But he really needs to delegate.	benefit behaviour blow battle beach blast bitterness
5. I'm happy to say that we have derived considerable _____ from working with your company.	???
6. They show a deep-rooted _____ against buying foreign products.	???
7. Their claim really stretches the bounds of _____. I'm sure it's false.	???
8. Cleaning up the pollution will be expensive. Who is going to foot the _____?	???
9. The strike last year created great _____ between the strikers and the non-strikers.	???
10. The public inquiry absolved my company from any _____.	???
11. When it exploded I got caught in the _____ and was thrown 50 feet.	???
12. It took me a while to get over such a devastating _____.	???
13. When I was made redundant I was given a big cheque to soften the _____.	???
14. I think we have committed a monumental _____ here. How can we put it right?	???
15. somebody is going to have to shoulder the _____ for what has gone wrong.	???
16. I'm going to a _____ concert in aid of the flood victims in Africa.	???
17. I'm phoning up concerning your outstanding _____. When can you settle it?	???

**Figure 2.7 Multiple choice exercise at [www.better-english.com](http://www.better-english.com)**

### 2.9.2 Concordancer tools for teachers and students

Instead of relying on existing corpora and concordancers, teachers and students alike can build and analyze their own corpora with the help of concordance tools. There are many such tools on the Web—for examples, MonoConc, WordSmith, Xaira, Kfngram, and AntConc. They share similar functionality. This section introduces the basic functionalities of AntConc (Anthony, 2006), a freeware, multiplatform application, designed specifically for learners in a classroom context.

AntConc was originally designed for use in a technical teaching class at Osaka University. It allows students to create their own mini-corpora to study and analyze field-specific text. It has undergone several upgrades based on requests and feedback from teachers. Students build a corpus by downloading and scanning text from research articles, or partial text such as titles, abstracts and sections, organizing them into files in plain text, HTML or XML format, and then uploading those files into the system. Once the corpus is constructed, users can use any of five separate tools. The Concordance tool retrieves concordance data for a search term, which can be a word, phrase or regular expression. The View

There is an ..... shortage of food in the country.	acute
John's becoming a vegetarian was a ..... change.	growing
This can't continue; it's an ..... situation.	drastic
Drug use is a ..... problem in many countries.	intolerable
That regime has implemented ..... measures to control the rioting.	full-scale
What you have just said is ..... nonsense.	utter
The government is launching a ..... investigation.	preliminary
Your idea seems very .....	far-fetched.
The ..... data tend to support your conclusions.	harsh
Politicians are very aware of ..... opinion.	public

A Spellmaster.Com Game - code and design ©Frank McAree 2001

### (a) Drag-and-drop exercise

1. There is an  [?] contradiction between God's omniscience and man's free will.
2. The report said the senator's health was a  [?] secret.  
serious  
utter  
apparent
3. Senior citizens said the new drug was ..... something that seems to exist although it may not  [?] cost of medicine.
4. Most people don't like to drive in  [?] OK
5. He said he had an  [?] dislike of any form of censorship.
6. There was  [?] unemployment at the height of the Depression.
7. The bill to reduce the Federal Budget passed the House by a  [?] margin.

### (b) Gap filling exercise

heavy  
intense  
massive  
apparent  
narrow  
utter  
closely-guarded  
ruling  
exorbitant  
serious

traffic  
class  
margin  
doubt  
cost  
nonsense

time  
effect  
dislike  
secret  
contradiction  
unemployment

**(c) Matching two word collocation exercises**

## **Figure 2.8 Collocation exercises on *angelfire.com***

File tool looks at how the search term is used in a particular file. The Concordance Search Term Plot tool shows how the search term is distributed in the corpus. The Wordlist and Keyword List tools generate and examine the statistics—such as frequency, rank and “keyness”—of all words that occur in the corpus. The Word Clusters tool shows multi-word units and their frequency.

Anthony (2006) introduces a set of procedures that teachers may follow when using AntConc to help students study word appropriateness. Outside the class, students collect texts in their discipline from the Web or other resources to build their own corpora. Then they write a short text and note down the words or phrases they feel uncomfortable with. In class, students use the Concordancer tool to search for the words and phrases highlighted, and examine the results in terms of frequency of occurrence and distribution across corpus text to identify the appropriateness or inappropriateness of the search term. They are then encouraged to use thesauri or dictionaries to find alternatives to inappropriate ones and test them again. Teachers are advised not to provide answers, but instead to give suggestions or feedback on the search results.

### 3. Presenting corpus data for collocation learning

---

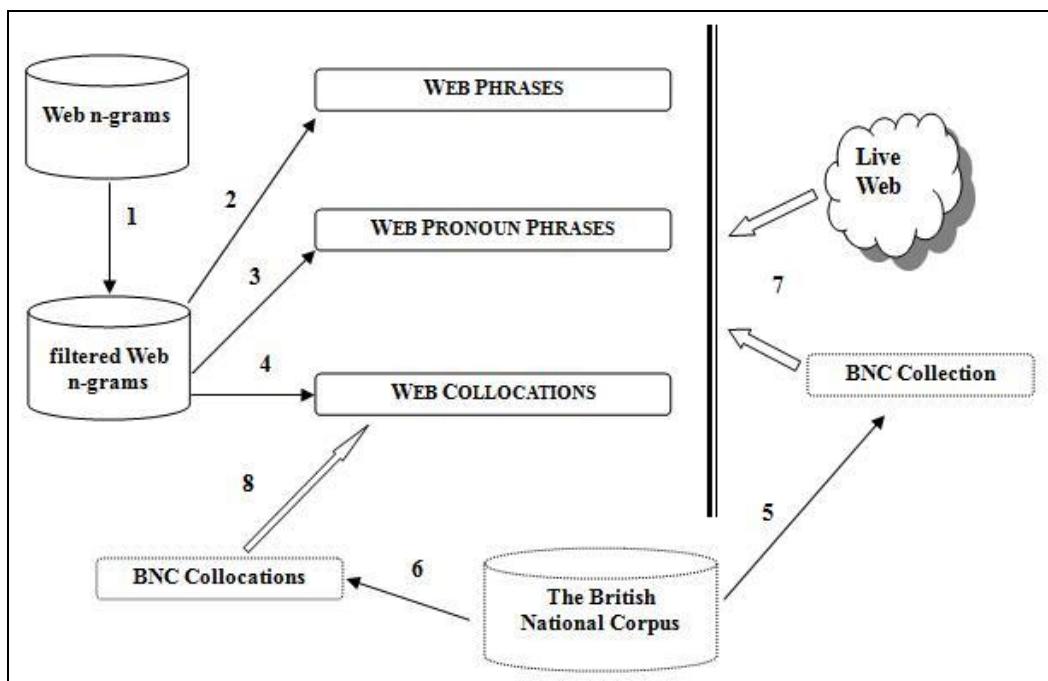
In recent years, researchers have begun to exploit large corpora for language teaching and learning (Yoon, 2008). In fact the potential of corpora as a resource in language learning has been evident to researchers and teachers since the late 1960s (Chambers, 2005). However, collocation learning resources derived from corpus data are either limited in coverage or lack learner-friendly interfaces. This thesis explores the use of Web text for collocation learning. The Web has unique features shared by no other corpus. It is potentially useful for language study because it provides a virtually unlimited number of examples of language that are both contextualized and authentic.

CLS (Collocation Learning System, introduced in Section 1.3) provides collocation learning resources that make use of a trillion word tokens of Web text, summarized in the form of *n*-grams and made available by Google. Figure 3.1 shows the structure; the numbers in the description below refer to the numbered arrows. CLS filters Web text (1) and uses the Greenstone digital library software to organize, design and build three searchable collections from different parts of the information, and serve them on the Web. It creates three primary collections:

- WEB PHRASES (2)
- WEB PRONOUN PHRASES (3)
- WEB COLLOCATIONS (4).

Two secondary collections are built from the text of the British National Corpus (Section 2.5.3): the BNC collection (5) and the BNC Collocations collection (6).

The three primary collections, WEB PHRASES, WEB PRONOUN PHRASES, and WEB COLLOCATIONS, are enriched in different ways. First, they are linked to the BNC collection and to the live Web (7). Collocations within WEB COLLOCATIONS are compared with the BNC collocations (8). For each of the three primary collections, the Greenstone digital library system's searching and browsing facilities are tailored to support collocation learning.



**Figure 3.1 CLS's collocation learning resources**

This chapter describes how to use and build these collections. They are evaluated in Chapters 4 and 5.

### 3.1 Using Web text

The Web has often been used in linguistics research to corroborate intuitions about the frequency of individual words, collocations, phrasal verbs, and idioms. As many researchers have noticed (Seretan et al., 2004; Chinnery, 2008; and Shei, 2008), it is a particularly valuable source of information about collocations because it provides text on contemporary issues and authentic examples of current and emerging language usage. The studies described in Section 2.8 have explored the potential of the Web as a corpus for learners. They use the content of web pages and text snippets from search engine hits to generate concordance data, and discover collocations in context.

However, this approach is limited. First, search providers do not allow unrestricted access from programs (as opposed to people) and in some cases prohibit it altogether. Although arrangements can sometimes be made with search engine companies for limited experimental usage for research purposes, these are

restricted to a certain number of queries per day, which would be insufficient to support concordance-style services on a satisfactory, scalable basis. Second, features of the Web itself make it less than suitable for language learning and teaching in raw form. These include its massive size, and the fact that it contains many items that are potentially confusing or misleading for learners, such as non-word character strings, website names and grammatical errors. A third, more minor, problem is that the frequency counts that search engines return for words and phrases are only approximate, though they are probably a good enough indication for language learning purposes.

Instead of relying on live Web searches to generate collocation and concordance data, we work with an off-line corpus generated and supplied by Google. This contains short sequences of consecutive words, called “*n*-grams,” along with their frequencies. Unigrams comprise one word; bigrams two; tri-grams three; and so on. The corpus contains all of these up to and including five-grams. Using this resource is an innovation that mitigates some of the constraints associated with the Web as corpus. It also provides a sound basis for operating scalable services that use Web text as a resource for language teaching and learning.

This off-line corpus is a vast set of word *n*-grams in the English language, along with their frequencies. The text was collected by Google in January 2006 from publicly accessible Web pages. The corpus was generated from approximately one trillion word tokens of text—a staggeringly large body of natural English. *N*-grams that occur fewer than 40 times were discarded (by Google, before publishing the corpus). Even so, the material comprises approximately 90 GB of text files.

Table 3.1 summarizes its size. The number of *n*-grams increases as *n* grows beyond 1, peaks at *n*=4, and then begins to decay. Figure 3.2 shows a few of these lines in the raw data files supplied by Google. They are simple: each *n*-gram occupies a line:

word\_1 <space> word\_2 <space>... word\_n <tab> *count*

where *count* is the number of occurrences of this *n*-gram.

**Table 3.1 Number of units in the  $n$ -gram corpus**

Tokens	1,024,908,267,229	$10^{12}$
Sentences	95,119,665,584	$0.95 \times 10^9$
Unigrams	13,588,391	$0.014 \times 10^9$
Bigrams	314,843,401	$0.3 \times 10^9$
Trigrams	977,069,902	$1.0 \times 10^9$
Four-grams	1,313,818,354	$1.3 \times 10^9$
Five-grams	1,176,470,663	$1.2 \times 10^9$

I ASKED FOR !</S>	53
I ASKED FOR A SO	67
I ASKED FOR I SAW	52
I Asked For , Inspirational	40
I Asked For It </S>	52
I Asked For Love </S>	66
I Asked For More Butter	318
I Asked For Reinforcements ,	77
I Asked For That robb06	926
I asked for ?</S>	1072
I asked for Anonymous --	339
I asked for Christmas .	61
I asked for Courage ...	80
I asked for a 12	170
I asked for a 2	51
I asked for a </S>	237
I asked for a <UNK>	130
I asked for a >	71
I asked for a CD	83
I asked for a Coke	75
I asked for a Mgr	49
I asked for a river	163
I asked for a roll	43
I asked for a room	1395
I asked for a ruling	55
I asked for a sample	183

**Figure 3.2 Sample  $n$ -grams**

### 3.1.1 Cleaning the data

It is necessary to clean up this corpus in order to make it suitable for language learning. This process has the useful side benefit of reducing its massive size to more manageable proportions.

Like the Web itself, the  $n$ -grams are messy. They include many non-word character strings, website names and grammatical errors. While the first two can easily be removed, it is virtually impossible to eliminate grammatical errors.

### Web samples

- ➡ Kristy Lee Cook: 'I was a little disappointed to go' Updated | Comment | Recommend ... best performance so far, so I was a little disappointed to go this soon because ...
- ➡ Oklahoma City Marriott, Oklahoma City: I was a little disappointed - Visit TripAdvisor, your source for the web's best unbiased reviews of hotels and vacations, ...
- ➡ I was a little disappointed in the sharpness of. User: My Threads. Flat view. Navigation: ... I was a little disappointed in the sharpness of. Posted by: OTD ...
- ➡ Keeping Delphi ... I was a little disappointed that the preview webinar this ... It was at least an opportunity for some more Q&A and a couple of ...
- ➡ With apologies to those who have downloaded what I wrongly claimed was the " ... I was a little disappointed that the preview webinar this morning was little ...
- ➡ The Apple Developers conference is just that, a developers conference. ... I was a little disappointed because all of a sudden people were pirating ...
- ➡ On Monday I said I was a little disappointed that Michael Chabon's The Yiddish Policemen's Union won the Hugo Award ... Rant, io9 commenters rule, michael ...
- ➡ " 05 Spelltech Syrah, Contra Costa. 2006 Alamos Malbec, ... I was a little disappointed, ... So I was a little disappointed by this bottle at first, though it ...
- ➡ 55 of 57 people found the following review helpful: I was a little disappointed, August 8, 2003 ... But beyond that I was a little disappointed in the book. ...

### (a) From the Web

### BNC samples

- ➡ 'All that's fine,' I said, though I wasn't particularly interested in the vows of a child who had just gone to boarding school. 'You say that the paintings have been handed down. Is that all there is to the story?' I was a little disappointed. 'Will they ever be worth anything?'
- ➡ I was a little disappointed by the grip even on wet and damp rock by the rubber-cleated and stud-pattern sole. However I found they performed excellently on steep and wet grass.
- ➡ 'The second twin didn't cry straight away. Before it could, I cut the cord and took it into the ante-room. Then it cried. It was another girl. I was a little disappointed, but I could only hope that Celia was still a bit hazy from the drugs. I went back and told Lilian the second twin, a boy, had died because the cord was round its neck. She accepted it.'
- ➡ I am one of those who welcome without any reservation the citizens charter that has been brought forward by the Government. I was a little disappointed, although perhaps not surprised, at the grudging welcome for some of the suggestions from the Labour party. Many of the suggestions made by the Government in the citizens charter give us an opportunity to contribute our own ideas to what should go into the citizens charter and for those ideas to expand and to grow as a result of the kernel in the programme that the Government have produced.
- ➡ 'Dear Fatal,' writes Philip Saunders from North Devon, 'I'm writing to thank you for the excellent screen wipes. I have to say that I was a little disappointed at first when mine failed to absorb all the rain from a really wet windscreen, but I found that when held edge-on, this handy, blue tool proved most effective at scraping the snow and ice from my car. I am now left with only two questions: What is that little metal sliding bit for? Oh, and what were those funny tissue things?'

### (b) From the BNC

**Figure 3.3 Samples retrieved for *I was a little disappointed***

Deficiencies in natural language processing technology makes analysis difficult and somewhat unreliable, but—more importantly—the fact that no context is available beyond the neighboring few words makes accurate parsing impossible in principle (we discuss this problem in the next Chapter).

Nevertheless, great improvements can be made by cleaning up the text. CLS uses a wordlist derived from the BNC to remove non-words and website names, and discards all word sequences that include words not in this list.<sup>3</sup> This reduces the volume of text by 30% and yields a much tidier corpus. However, it has the effect of removing sequences containing neologisms (often ones coined since the BNC was constructed), notably, for example, the word *google*. Of course, it would be trivial to add such terms to the wordlist.

### 3.1.2 Building contextual resources

For language learners, *n*-grams have the intrinsic limitation that context is lost when they are removed from the original text. Context has long been recognized as crucial for vocabulary learning (Nagy, 1997). The remedy adopted by this thesis is to reconstruct suitable contexts from two sources and present them to users on demand.

The first source is the Web. Whenever a language learner requests the context of a particular *n*-gram, CLS connects to a search engine, uses the words as a phrase query and retrieves sample texts. The Yahoo search engine is used because Google imposes some limitations, and disables automatic queries from computer programs other than Web browsers. Yahoo has no obvious disadvantages in terms of the quality of text snippets retrieved for a particular search.

The second source is the BNC. The BNC text is split into paragraph units and built into a searchable collection using the Greenstone digital library software. Whenever the learner asks to see examples of a particular *n*-gram in context, we arrange for Greenstone to search the collection for occurrences and display the relevant paragraphs.

---

<sup>3</sup> <http://www.lexically.net/downloads/version4/downloading%20BNC.htm>

**Table 3.2 Number of  $n$ -grams in the WEB PHRASES collection**

<b>unique words</b>	<b>two-grams</b>	<b>three-grams</b>	<b>four-grams</b>	<b>five-grams</b>
145,000	14 million	420 million	500 million	380 million

Figure 3.3a and b show samples retrieved from the Web and BNC respectively for the phrase *I was a little disappointed*. The contemporary nature of the snippets in Figure 3.3a is apparent from the fact that two of the eight report the feelings of an unsuccessful 2008 American Idol contestant. Many more examples of this phrase are available on the Web and can be obtained by clicking the *next* button at the bottom of the page. The phrase has ten BNC hits, of which five are shown in Figure 3.3b. They tend to be more coherent than the Web snippets, and are presented in a fuller context.

Both sources have limitations, and the two are somewhat complementary. The BNC provides far fewer examples, the number declining rapidly for longer sequences. In many cases there are none at all—even for items that occur reasonably frequently on the Web. For example, *I was very disappointed in* occurs 1,560,000 times on Web, but not at all in the BNC.<sup>4</sup> On the other hand, the Web text, being extracted from individual Web pages rather than the aggregations in the  $n$ -gram corpus, is often unclean, incomplete and repetitive.

### 3.2 *The WEB PHRASES collection*

The WEB PHRASES collection is built on Shei's (2008) pioneering work (Section 2.8), which allowed users to study particular words and phrases to check whether and to what extent the text they have written represents common usage. It allows free exploration of word combinations, unconstrained by grammatical class. Table 3.2 shows the number of  $n$ -grams in this collection. Users can study the words that most commonly follow, precede, or occur between particular words or phrases. Frequency is interpreted as some indication of the representativeness or authenticity of the sequence. If the frequency is zero, that text does not appear in

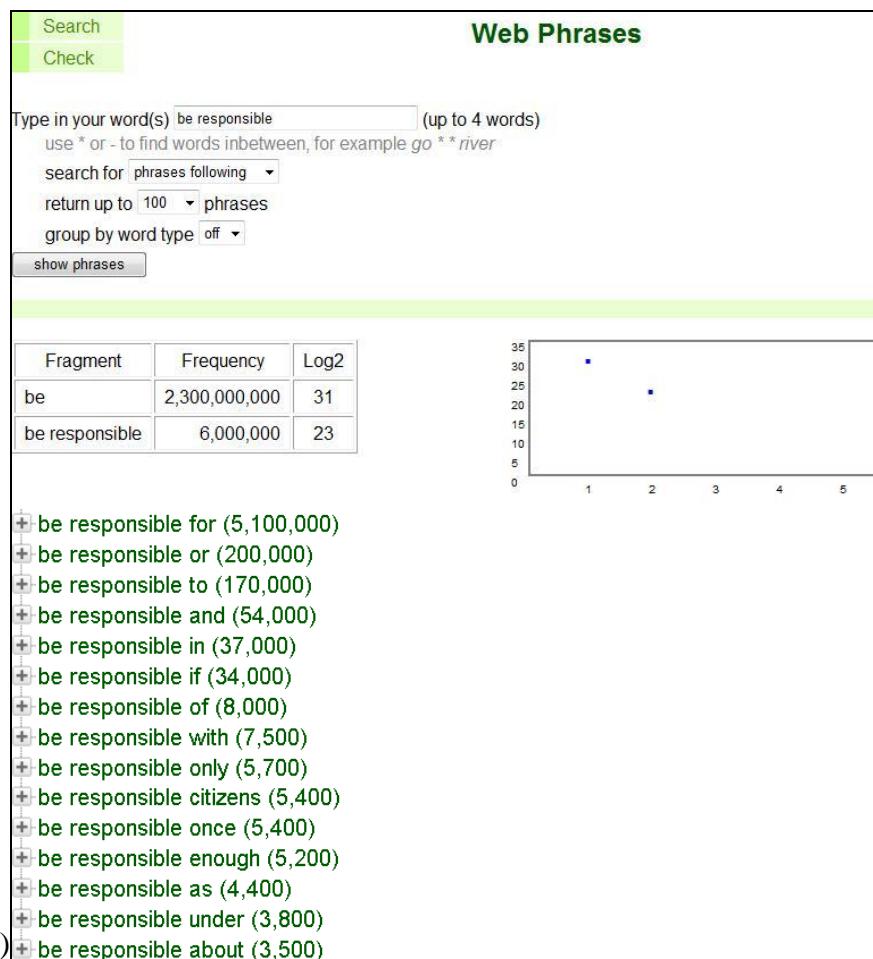
---

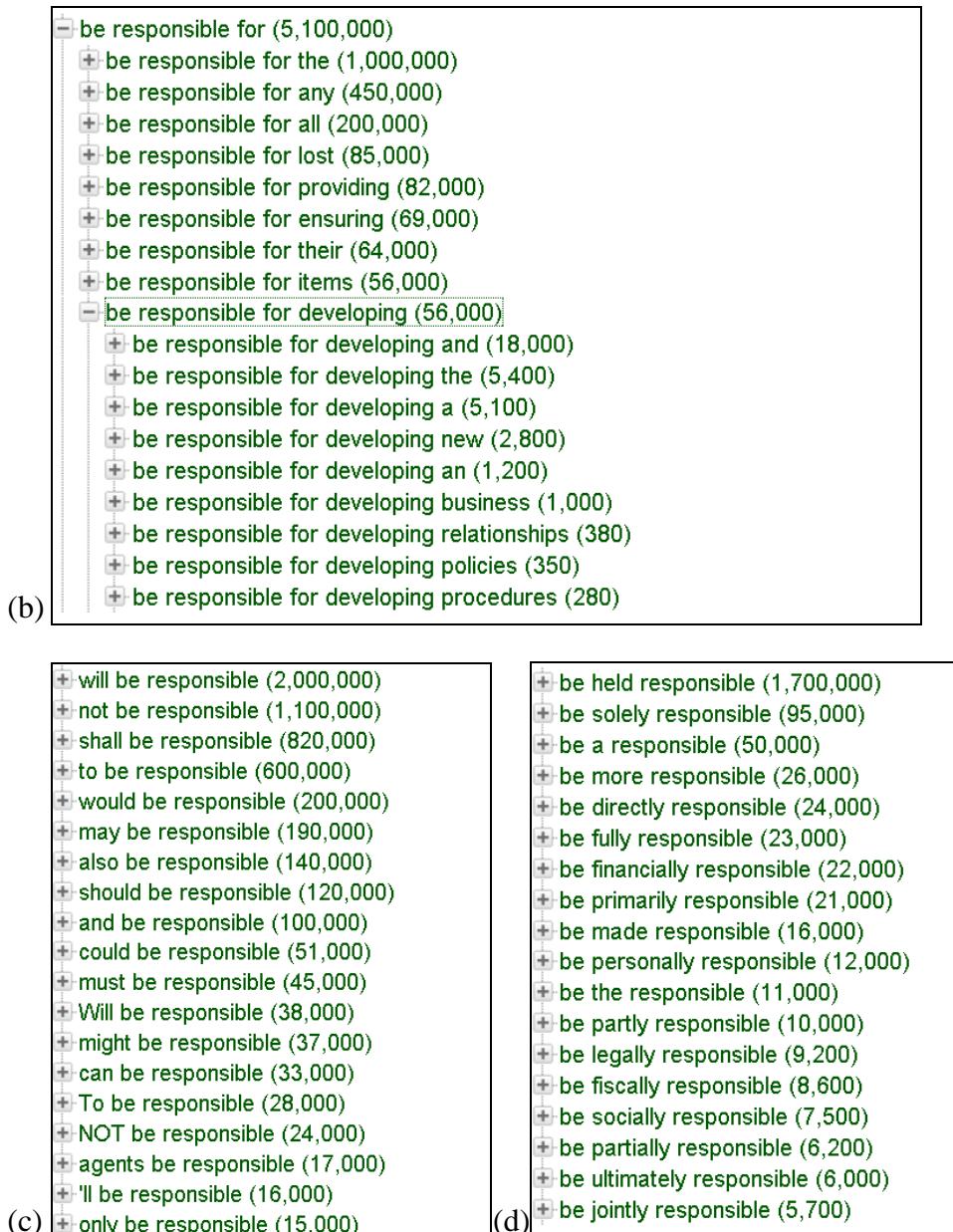
<sup>4</sup> Retrieved using the Google search engine on August 15, 2010.

the collection. This might be good news for creative and confident writers, but for most language learners it is a negative reflection on what they have written.

### 3.2.1 Using the collection

The interface allows users to determine the words that most commonly follow a particular word or phrase. Figure 3.4a illustrates this for the phrase *be responsible*. The interface contains three parts. A statistical table gives the frequency count for the query word or phrase. On the right is a graph that indicates visually how the frequency reduces as words are added: frequency is represented by its logarithm for ease of visualization. Below is an expandable tree that displays associated phrases in reverse frequency order, along with their frequency count.





**Figure 3.4 Searching facilities provided by WEB PHRASES**

The most frequent words following *be responsible* are *for*, *or*, *to*, *and*, etc. Clicking *be responsible for* and *be responsible for developing*, the tree expands and displays the phrases associated with these phases, as shown in Figure 3.4b, and the table and graph update accordingly. A phrase can be expanded up to five words, or until no further extensions occur in the collection. At that point, samples of text that use the phrases can be retrieved from the Web and from the BNC by clicking the appropriate icon.

**Table 3.3 Response time of the WEB PHRASES collection**

query words	response time (seconds)
<i>be responsible</i> (phrases following)	1.5
<i>be responsible</i> (phrases preceding)	1.5
<i>be * responsible</i>	9
<i>be ** responsible</i>	25
<i>be *** responsible</i>	30
<i>be * responsible * the</i>	20

Users can search backwards by specifying the *phrases preceding* option. As shown in Figure 3.4c, one can browse around words that precede *be responsible*. Most of them are modal verbs—*will*, *shall*, *would*, etc. Furthermore, a wild card character (\*), which stands for any word, can be used in the search. Figure 3.4d shows the adverbs (*solely*, *directly*, *fully*, etc.) that are associated with *be \* responsible*. Further asterisks can be added, for example, *be \*\* responsible*, *be \*\*\* responsible*, and *be \* responsible \* the*, each one indicating a wild card word.

The *return up to* option allows users to determine the number of phrases to return (the default is 100). The bigger the number, the longer it takes to obtain the search result. Common words like *the*, *a*, *of*, and *to* often make it hard for users to glean useful language patterns. To address this problem, the *group by word type* option allows users to determine what words of a particular part-of-speech—preposition, verb, noun, adjective, etc—follow or precede a phrase.

### 3.2.2 Building the collection

This collection consists of two copies of the filtered Google *n*-grams: one in natural order and the other in reverse order. The first allows users to look up the frequency of a particular word or phrase, and the phrases that follow them. The second supports the *phrases preceding* option discussed in the previous section, and is generated by reversing each *n*-gram—e.g., *a good day* becomes *day good a*—and re-sorting alphabetically. To achieve a reasonable response time, two steps are applied:

1. split the original files into smaller files, each containing 10,000  $n$ -grams, and
2. build search indexes.

The search indexes comprise two kinds of file: dictionary file and index file. The original  $n$ -grams are grouped by the number  $n$  and stored in separate files that contain 10 million  $n$ -grams each. For each  $n$ , the dictionary files are generated by splitting an original file into 1000 smaller files ( $10,000,000/10,000$ ). Each entry in the index file occupies one line: the last  $n$ -gram of a dictionary file, the name of the dictionary file.

It should be noted that the index structure is not optimised: indexes are rudimentary and 10,000  $n$ -grams per dictionary file is somewhat arbitrary. Table 3.3 shows the response time for retrieving 100 phrases using a browser for the query words described in the previous section on a computer with 3GHZ CPU, 1GB RAM and 10Mb Internet connection. The collection responds reasonably well to the first two queries. However, the index structure is not really designed for wild card (\*) searching. For example, searching for *be \* responsible* involves:

1. retrieve dictionary files that contain three-grams that start with the word *be*,
2. identify those that also end with the word *responsible*,
3. sort them by frequency, and
4. return the top 100 phrases (the user can alter this number).

Efficient indexes are needed to support wild card searching, but this was not pursued in this research because of time constraints.

### *3.3 The WEB PRONOUN PHRASES collection*

The WEB PRONOUN PHRASES collection contains a large number of pronoun phrases, that is, ones that contain *I, he, she, you, they, we, and it*. Table 3.4 shows the number of phrases in this collection: 570,000 in total and an average of 80,000 for each pronoun. It is designed to help language learners express what they think, feel and do. Students might answer a simple question like “How are you today?”

**Table 3.4 Number of pronoun phrases in WEB PRONOUN PHRASES**

<i>I</i> -phrase	102,000
<i>he</i> -phrase	75,000
<i>she</i> -phrase	49,000
<i>you</i> -phrase	88,000
<i>they</i> -phrase	79,000
<i>we</i> -phrase	63,000
<i>it</i> -phrase	110,000
total	566,000

factually (“My head aches”), or perfunctorily (“OK”). But they find it hard to go beyond simple declarative statements and talk about their feelings in greater depth.

Part of the reason is that learners have not experienced enough of the language to express themselves in the first person in ways that sound natural. As Moskowitz (1978) notes, curricular material tends to focus on facts and everyday transactions, only rarely touching on vocabulary that is appropriate for communicating more subjective aspects of everyday life. To help remedy this she advocates integrating a humanistic approach to language teaching with a planned curriculum to promote self-actualization and self-esteem, so that students can express themselves meaningfully in the first person. Another part of the reason is that fluency does not blossom from a comprehensive lexicon of difficult words, nor even from familiarity with the most common ones. Instead, it requires an internalized repertoire of phrases and expressions composed of words used in everyday life (Lewis, 1993).

### 3.3.1 Using the collection

There are three ways for learners to examine the usage of a word: phrases that contain it, phrases that precede it, and phrases follow it. These are discussed below. Then we describe the browsing operations that are built into the collection.

#### Phrases containing a particular word

Suppose the learner wants to write a personal statement—an *I*-phrase—to express disappointment. Figure 3.5a shows the search results for the word *disappointed*. It

shows *I*-phrases that contain the word *disappointed* in inverse frequency order, grouped by tense—past, present perfect, present, and future. Each phrase is assigned tense metadata during the collection building process.

Clicking the phrase or the image icon that follows the frequency retrieves samples from the Web and the BNC respectively (Section 3.1.2). The most common sentence begins *I was a little disappointed* (47,000 occurrences), a past tense usage; the second begins *I was a bit disappointed* (29,000 occurrences). Both of them involve the hedges *a little* and *a bit*, which is useful pragmatic, as well as grammatical and lexical, information.

Search for

the word(s)  in   phrases

[disappoint](#) [disappoints](#) [disappointing](#) [disappointment](#) [disappointments](#)

- synonyms: [noun](#), [adjective](#), [verb or adverb](#)
- antonyms: [noun](#), [adjective](#), [verb or adverb](#)
- related words: [noun](#), [adjective](#), [verb or adverb](#)
- associated words

**Simple Past (14)**

- ▶ [I was a little disappointed ... \(47274\)](#)
- ▶ [I was a bit disappointed ... \(29676\)](#)
- ▶ [I was disappointed with the ... \(15092\)](#)
- ▶ [I was very disappointed with ... \(13676\)](#)
- ▶ [I was very disappointed in ... \(11662\)](#)
- ▶ [I was disappointed in the ... \(11455\)](#)
- ▶ [I was disappointed that the ... \(8413\)](#)
- ▶ [I was disappointed by the ... \(8148\)](#)

**Present Perfect (3)**

- ▶ [I've never been disappointed ... \(11847\)](#)
- ▶ [I have not been disappointed ... \(11198\)](#)
- ▶ [I have never been disappointed ... \(6218\)](#)

**Simple Present (11)**

- ▶ [I am disappointed that the ... \(8865\)](#)
- ▶ [I am very disappointed in ... \(7669\)](#)
- ▶ [I'm disappointed in you ... \(6817\)](#)
- ▶ [I am very disappointed with ... \(6666\)](#)

(a)

**Verb(Past Tense) (2)**

- ▶ I was disappointed ... (97787)
- ▶ I was not disappointed ... (11775)

**Verb(Present Tense) (1)**

- ▶ I am disappointed ... (74355)

**Verb(Past Tense) + Adverb (9)**

- ▶ I was very disappointed ... (50583)
- ▶ I was really disappointed ... (12093)
- ▶ I was so disappointed ... (10577)
- ▶ I was extremely disappointed ... (8251)
- ▶ I was quite disappointed ... (6983)
- ▶ I was somewhat disappointed ... (5764)
- ▶ I was rather disappointed ... (4423)
- ▶ I was pretty disappointed ... (3856)
- ▶ I was also disappointed ... (3321)

(b)

**Preposition: with (3)**

- ▶ ...disappointed with ... (58952)
- ▶ ...disappointed with the ... (23454)
- ▶ ...disappointed with this ... (4217)

**Preposition: in (4)**

- ▶ ...disappointed in ... (54874)
- ▶ ...disappointed in the ... (20578)
- ▶ ...disappointed in you ... (7791)
- ▶ ...disappointed in this ... (4516)

**Subordinating Conjunction: that (8)**

- ▶ ...disappointed that the ... (21134)
- ▶ ...disappointed that I ... (9214)
- ▶ ...disappointed that we ... (5047)

**Wh-adverb (2)**

- ▶ ...disappointed when ... (11187)
- ▶ ...disappointed when I ... (5006)

(c)

**Figure 3.5 Searching facilities provided by WEB PRONOUN PHRASES**

More information on the query term appears above the search results: links to family words, synonyms, antonyms from WordNet, and related words from Roget, each grouped by part-of-speech, and to associated words from the Edinburgh thesaurus. We discuss these in Section 3.3.2.

If more than one term is typed into the search box, phrases containing each one are presented under the various categories in the search results. Quotation marks can be used to signify that the query should be treated as a phrase. It is interesting and often instructive to lengthen a chosen phrase word by word and see how the popular contexts change.

### **Phrases preceding a particular word**

Given a word, learners can study language patterns that frequently precede it. In the pull-down menu near the top of Figure 3.5b, *Phrases preceding* has been selected, and in this case the search results are grouped by words that appear in the preceding context. They show that the most common sentence structure with *disappointed* takes the form *be + disappointed*, and again the past tense is most common. The hedges *very*, *really*, *so*, *extremely*, *quite*, *somewhat*, *rather*, and *pretty* are often used in this context.

### **Phrases following a particular word**

This allows users to explore what words and phrases follow a particular word. Figure 3.5c shows that the prepositions *with* and *in* commonly follow *disappointed*, and that *disappointed* is often followed by *that-* and *when-clauses*. These indicate useful sentence structures that learners can employ when they want to express disappointment about something.

Table 3.5 contrasts the patterns that follow the words *love* and *hate* (obtained by the same method but, for succinctness, displayed in tabular form rather than as screenshots). This not only reveals what people commonly love or hate, but also helps learners choose appropriate words when they want to express similar feelings.

**Table 3.5 Patterns that follow the words *love* and *hate***

<b>love</b>	<b>hate</b>
<b>you</b>	<b>you</b>
... love you ... 246236	... hate you ... 20825
... love you I love ... 106610	... hate you I hate ... 8165
... love you so much ... 97106	... hate you so much ... 6675
... love you because I ... 43553	
... love you more than ... 41664	... hate the... 20010
... love you and I ... 37987	... hate the fact that ... 17283
... love you forever ... 34593	... hate the idea of ... 11524
<b>the</b>	... hate the thought of ... 6754
... love the fact that ... 69154	... hate the way you ... 6502
... love the way you ... 62343	
... love the idea of ... 49511	<b>myself</b>
... love the smell of ... 40600	... hate myself and want ... 16722
<b>these</b>	... hate myself for losing ... 6322
... love these shoes so ... 35925	<b>him</b>
	... hate him ... 6893

## Browsing

Figure 3.6a shows the beginning of the list of *I*-phrases. Interestingly, *think* is the most frequent word that follows *I*, and the next four most frequent verbs are *have*, *know*, *want* and *like*. Figure 3.6b displays the language patterns that are associated with *think* in the first person context.

For the structure corresponding to the first person singular pronoun followed by a verb (*I* + verb), *think* is retrieved as the most frequent verb. This corroborates the findings of Biber and Kurjian (2007) that frequently occurring linguistic features associated with personal narrative texts on the Web are the first person pronoun *I*, mental verbs such as *think*, and *that*-clauses. It also aligns with Biber et al.'s (1999) earlier finding that the most frequent lexical bundle in conversation consists of a subject pronoun (first person) and a verb phrase to express a personal opinion, such as in the phrases *I think that* and *I think he*.

### 3.3.2 Lexical resources

When searching a digital library collection, learners often find it difficult to formulate query terms because of their limited vocabulary. CLS uses external

browse  in

(a)

- + 1. think (64,000,000)
- + 2. have (52,000,000)
- + 3. know (39,000,000)
- + 4. want (29,000,000)
- + 5. like (28,000,000)
- + 6. 'm (26,000,000)
- + 7. had (24,000,000)
- + 8. am (19,000,000)
- + 9. going (17,000,000)
- + 10. was (14,000,000)
- + 11. get (14,000,000)
- + 12. thought (13,000,000)
- + 13. see (13,000,000)
- + 14. got (12,000,000)
- + 15. believe (12,000,000)
- + 16. hope (12,000,000)
- + 17. need (10,000,000)
- + 18. sure (9,600,000)
- + 19. do (9,400,000)

(b)

- 1. think (64,000,000)
  - I do not think I (1,600,000)
  - I do not think it (1,600,000)
  - I do not think that (1,500,000)
  - I do not think so (750,000)
  - I do not think the (700,000)
  - I do not think you (660,000)
  - I do not think we (640,000)
  - I cannot think of (610,000)
  - I think it would be (580,000)
  - I do not think there (570,000)
  - I think it's a (530,000)
  - I do not think they (470,000)
  - I do not think he (420,000)
  - I do not think this (360,000)
  - I think this is a (350,000)
  - I think it is a (300,000)
  - I do not think anyone (270,000)
  - I think I'm going (230,000)

**Figure 3.6 Browsing facilities provided by WEB PRONOUN PHRASES**

databases—WordNet, Roget’s thesaurus, the Edinburgh Word Association thesaurus and Yasumasa Someya’s lemma list—to retrieve words related to or associated with a particular query term. In WordNet, nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms, each expressing a distinct concept.<sup>5</sup> Roget is a widely used thesaurus; the online version contains 15,000 words.<sup>6</sup> The Edinburgh Word Association thesaurus contains word association strengths derived experimentally from human subjects.<sup>7</sup> Yasumasa Someya’s lemma list contains about 15,000 entries.<sup>8</sup> We downloaded these resources and developed computer programs to incorporate them into the collection.

In the interface shown in Figure 3.5, related lexical information of the query word appears above the search results. The first line is words from the lemma list: *disappoint, disappoints, disappointing, disappointment, and disappointments* are provided for the word *disappointed*. Clicking one of these words changes the query word to that word. Below are the links to synonyms, antonyms from WordNet, related words from Roget and associated words from Edinburgh. Clicking one brings up a page containing a list of words retrieved from the corresponding resource. Figure 3.7 shows the synonyms for *disappointed* retrieved from WordNet.

Each resource is filtered to remove words and phrases that do not appear in the collection. This eliminates usage that rarely occurs, and prevents learners from becoming overwhelmed with choice. For example, WordNet contains these synonyms for the adjective *sad*:

*bad, bittersweet, depressing, depressive, gloomy, saddening, doleful, mournful, heavyhearted, melancholy, melancholic, pensive, wistful, tragic, tragical, tragicomic, tragicomical, sorrowful, deplorable, distressing, lamentable, pitiful, sorry*

Only those in boldface, a small minority, actually occur in the collection. On the other hand, all three WordNet antonyms—*glad, joyful* and *good*—occur. Related words from Roget that appear in the collection include *unpleasant, unacceptable,*

---

<sup>5</sup> <http://wordnet.princeton.edu>

<sup>6</sup> <http://www.gutenberg.org/etext/10681>

<sup>7</sup> <http://www.eat.rl.ac.uk>

<sup>8</sup> [http://www.lexically.net/downloads/e\\_lemma.zip](http://www.lexically.net/downloads/e_lemma.zip)



**Figure 3.7** Synonyms for *disappointed* retrieved from WordNet

*touching, troublesome, fearful, hard* and *cutting*, while associated words in the Edinburgh database include *happy, unhappy, bad, cry, death, girl, glad* and *me*.

### 3.3.3 Building the collection

To build the collection, CLS identifies five-grams that commence with the pronouns *I, he, she, you, they, we*, and *it*. Five-grams are used because these provide the largest context. Then the identified pronoun phrases are placed into a Greenstone digital library collection.

#### Selecting pronoun phrases

Two selection steps are applied:

1. select five-grams that start with a pronoun word, and
2. discard grammatically incorrect sequences.

In step 2, surviving 5-grams are parsed into phrases by the OpenNLP chunker (Section 4.2.1), and suspect ones discarded. OpenNLP uses the Penn Treebank tagset (Section 4.2.1)—producing, for example, for the five-gram *I asked for a room*

[NP I/PRP] [VP asked/VBD] [PP for/IN] [NP a/DT room/NN]

As will be discussed further in Section 4.2.1, square brackets indicate phrases, at the beginning of which is a phrase-level tag that identifies the syntactic role of the phrase. This fragment contains the noun phrase (NP) *I*, the verb phrase (VP) *asked*, the preposition phrase (PP) *for*, and the noun phrase (NP) *a room*. Word-level tags follow each word and convey tense and number information: *I* is a proper noun

(PRP), *asked* is a past-tense verb (VBD), *for* is a preposition (IN), *a* is a determiner (DT), and *room* is a singular noun (NN).

Tagged sentences are matched against a regular expression that specifies a noun phrase (NP), followed by a verb phrase (VP), optionally preceded by adverbial phrases (ADVP); and may optionally end with a noun, prepositional (PP), adverb, adjective (ADJP), particle (PRT) phrase or clause (SBAR).

### **Creating the collection**

Greenstone works with a basic unit of *document*.<sup>9</sup> Documents consist of *sections*, and Greenstone accommodates hierarchies of sections—typically chapters, sections, subsections, etc.—of arbitrary depth. Searching can be at both the document and the section level.

Making each pronoun phrase a separate document would yield a collection with 5.8 million documents; organizing them as separate sections of the same document would yield a single document with 5.8 million sections. Both are undesirable for performance reasons. As a compromise, the pronoun phrases were grouped based on the leading pronouns and then the first adjective and verb encountered. For example, *I was a little disappointed* and *I was disappointed in the* are placed in the same file, along with all other *I*-phrases that have *disappointed* as the first adjective. The smallest documents correspond to rare words and contain just one section. The largest have many thousands of sections, which again impacts search performance, but CLS truncates them to the 100 most frequent pronoun phrases containing that adjective and verb. This yielded 57,000 documents with an average of about 10 pronoun phrases each.

Greenstone has a scheme of “plugins” that allows it to deal with different document formats in an extensible manner. CLS uses a custom plugin to process files that contain lists of pronoun phrases, treating each one as an independent document. It extracts metadata corresponding to *frequency*, *word type* and *tense*. For the last two, the plugin identifies the nouns, verbs, adjectives, adverbs and

---

<sup>9</sup> Documents may contain text or multimedia, though the latter does not concern us here.

prepositions in each pronoun phrase and associates them with that document as metadata.

This collection has a hierarchical browser that allows users to browse by wordlist (Figure 3.6) and see the pronoun phrases in which any particular word appears was created. For each pronoun word, e.g., *I, he, she, you, they, we*, and *it*, four wordlists were generated and sorted into inverse frequency order:

1. all words regardless of type
2. main verbs
3. main adjectives
4. modal words.

### *3.4 The WEB COLLOCATIONS collection*

The WEB COLLOCATIONS collection allows learners to study common word combinations that are organized by syntactic pattern. The total number of collocations, and the number of words the collection covers are shown in Table 3.6. CLS targets ten collocation types that involve nouns, adjectives, verbs and adverbs. The first six patterns are adopted from the work of Benson, et al. (1986). The other four are noun + noun, adverb + verb, verb + *to* + verb, and verb + adjective from the *Oxford Collocation Dictionary for Students of English*. To make full use of five-grams, four types are extended to include further items of potential use for learners. These extensions are also shown in Table 3.6.

To help learners correctly use nouns and verbs: (1) determiners and possessive pronouns, e.g., *the, a, any, some*, and *his*, that precede nouns are included, for example, *make a difference*, and (2) prepositions and adverbs that follow verbs are included, for example, *switch off the lights*. To enrich and expand collocation knowledge, adjective modifiers that precede nouns are included so that learners can not only study *cause irritation*, and *pose a threat*, but also *cause skin/eye/throat/stomach irritation*, and *pose a serious/significant/direct/real/immediate threat*. To help learners understand that some verb forms are more dominant than others—for example, *time goes on* is far more common than *time is going on*—noun + verb collocations are further divided based on the form that the verb takes.

**Table 3.6 Collocation types and examples**

collocation type	example	collocations	words
verb + noun(s)	<i>make appointments</i>		
includes:			
verb + noun + noun	<i>cause liver damage</i>	8,700,000	54,000
verb + adjective + noun(s)	<i>take annual leave</i>		
verb + preposition + noun(s)	<i>result in the dismissal</i>		
verb + adverb	<i>apologize publicly</i>	200,000	11,000
noun + noun	<i>a clock radio</i>	4,200,000	53,000
noun + verb			
includes:	<i>the time comes</i>		
noun + verb with present tense	<i>the time is running out</i>	1,200,000	34,000
noun + be + present participle	<i>the time is spent on</i>		
noun + be + past participle			
noun + <i>of</i> + noun	<i>a bar of chocolate</i>	7,800,000	40,000
adjective(s) + noun(s)	<i>a little girl</i>		
includes:			
adjective + noun + noun	<i>a solar energy system</i>	6,300,000	56,000
adjective + adjective + noun(s)	<i>a beautiful sunny day</i>		
verb + adjective			
includes:	<i>make available</i>		
verb (incl. phrasal) + adjective	<i>take up more</i>	91,000	9,800
verb + noun + adjective	<i>take it easy</i>		
verb + <i>to</i> + verb	<i>cease to amaze</i>	440,000	11,000
adverb + verb	<i>beautifully written</i>	500,000	13,000
adverb + adjective	<i>seriously addicted</i>	200,000	10,000

### 3.4.1 Limitations of collocation resources

There are two kinds of resources dedicated to collocation study: online collocation tools and printed collocation dictionaries. Online resources are rare and limited. The *Collins Collocation Sampler*<sup>10</sup> and *WebCorp's Collocation Profile* are the only ones we have encountered. The first, based on 56 million words of contemporary written and spoken text, allows learners to search for collocations of a particular word. The result, shown in Figure 3.8a, is a list of words occurring on either side of the target word, along with the frequency of individual words and

<sup>10</sup> <http://www.collins.co.uk/Corpus/CorpusSearch.aspx>

**Table 3.7 Useful collocations from Collins and WebCorp**

<b>Collins collocations</b>	<b>WebCorp collocations</b>
<i>make sure, make up, make sense, make decisions, make easier</i>	<i>make money, make sure, make sense</i>

combinations, and significance scores are calculated using either the *t-test* or the mutual information measure (Section 4.1).

The *WebCorp’s Collocation Profile* of a word, shown in Figure 3.8b, is generated using the content of web pages (up to 500 pages) returned by the Google search engine. It displays collocates on either side of the target word within a four-word span and their frequency. In Figure 3.8b, the word *money* co-occurs with *make* 252 times, mostly on the right-hand side (233 vs. 19), and 193 times within one word span, e.g., *make money*. The information shown in the figure is primarily intended for lexicographers or applied linguists. It seems less useful for language learners: out of 24 Collins and 20 WebCorp collocations, only those shown in Table 3.7 seem plausible. The entry *difference* and *decision* of Collins may mislead learners into thinking that *make difference* and *make decision* are correct forms.

Printed collocation dictionaries are designed for students to look up collocations that have been carefully selected by lexicographers. Given limited space, lexicographers have to determine which headwords and their collocations to include. In most cases, only one syntactic class is covered for multiple-class words—for example, an entry might be included for the verb *cause*, but not the noun *cause*. Even when the syntactic class is covered, there may be a difference between singular and plural nouns. The learner may assume that collocations of a singular noun apply to its plurals as well, or vice versa, but this is not always true. For some nouns, both forms are appropriate and depend on the context—*make a decision* and *make decisions*—but for others, one is more dominant—for example, *make a living* is 7,000 times more frequent than *make livings*.<sup>11</sup>

---

<sup>11</sup> Calculated using the hits returned by the Google search engine on February 25, 2010.

Collocate	Corpus Freq	Joint Freq	Significance
to	1104731	21137	95.611666
sure	13160	2725	50.550177
it	494702	6544	40.836931
you	421797	5795	39.830074
can	113012	2522	35.478756
up	104073	2268	33.308750
will	111798	1998	28.315586
a	973489	8987	27.534409
your	91112	1708	26.886880
difference	4193	720	25.809225
would	97660	1681	25.397263
them	76140	1369	23.520336
sense	9766	647	22.921226
that	526293	5005	22.016354
more	94468	1443	21.696900
we	191233	2278	21.482956
could	59556	1092	21.239985
they	224821	2525	20.942193
feel	16421	634	20.907439
decisions	2160	417	19.727705
any	50225	911	19.282722
easier	2733	405	19.235042
does	21081	600	18.857437
decision	6988	418	18.206159

Word	Total	L4	L3	L2	L1	R1	R2	R3	R4	Left Total	Right Total
money	252	4	6	7	2	193	24	11	5	19	233
new	101	2	4	2	2	2	2	85	2	10	91
IE	84					83		1	0	0	84
open	83						82	1		0	83
tabs	82							82	0	0	82
re	81		81							81	0
sure	62	2	1	2	1	52	1	3		6	56
GNU	59		2		56				1	58	1
online	58		3	1		1	46	4	3	4	54
way	55	3	5	34	3		1		9	45	10
command	55	5	5	4	11	7	12	9	2	25	30
USB	54	1					53		1	0	53
Key	52		1					51	1	0	51
Sawed-off	51					51				0	51
file	47	6	11	6	5		12	3	4	28	19
use	47	8	4	1	5	6	16	5	2	18	29
blog	44	4	5		1	1	4	4	25	10	34
PDT	44	1	41	2						44	0
rule	41	8	8	6	3	1	3	7	5	25	16
sense	40		4			26	5	1	4	4	36

(a) Collins Collocation Sampler

(b) WebCorp collocation profile

Figure 3.8 Collocates of the word *make*

Consecutive and non-consecutive collocations are given together, which may confuse the learner unless proper illustrative examples are given. When following a verb, some nouns must be preceded by an article, some must not, and some can be accepted either way. For example, we normally say *cause offence* and *commit an offence*. Adverbs can occur on either side of a verb. Sometimes, one side is dominant—for example, *heavily influenced by* is more native-like than *influenced heavily by*. It is hard to give such information in a printed dictionary. Space restrictions also make it difficult to provide even a few examples.

### 3.4.2 Using the collection

To look up collocations, the user simply types in the word of interest. The collection retrieves all collocation types associated with the query word and lets the user choose one to continue with. Figure 3.9a shows the result of searching for the word *cause*. First, the collocation types are grouped by word class. In this case, *cause* can be used as verb and noun. The verb section contains six collocation types related to the verb *cause*, while the noun section is dedicated to the noun *cause*. Beside each collocation type is the most frequent example of it. Clicking

one, say *cause problems*, brings up a collocation page like that shown in Figure 3.9b. It displays more collocations of this type, sorted in inverse frequency order and presented in two columns, along with the frequency and links that retrieve samples from the BNC and the live Web. The *next* button at the bottom brings up the next page containing more collocations.

The user can (1) restrict the level of vocabulary displayed in the result by specifying a wordlist (Section 3.4.3), (2) decrease or increase the number of collocations to return per page, (3) only include collocations whose frequency falls below a particular value by adjusting a frequency cut-off, and (4) decide whether to group collocations. The first three are straightforward; we discuss the fourth in more detail.

Collocations can be grouped together to allow users to inspect variants of a collocation; it also helps minimize confusion caused by partial collocations. Collocations like *a beautiful skin* and *cause different side*, which should be *a beautiful skin color* and *cause different side effects* respectively, are called partial collocations, and their occurrence is due to constraints on the length of *n*-grams. The Grouping option only has an effect on the four collocation types that are extended (Table 3.6): verb + noun, noun +verb, adjective + noun and verb + adjective. It groups collocations according to a template consisting of the main parts of a collocation type. The templates, and examples of their use, are given in Table 3.8. *Cause problems*, the most common *cause* + noun collocation, has 285 variants, and *cause serious problems*, *cause unpredictable problems*, and *cause major problems*, are grouped under the *cause* + *problems* template.

Collocations can be compared. To do this, the user enters two words. CLS retrieves the collocations associated with these words, groups common and different ones together, and presents them side by side. Figure 3.9c shows the result of comparing the verbs *speak* and *tell*, in the verb + noun type. They have 11 out of 100 collocations in common, the most frequent being *speak on behalf of* and *tell millions of*. *Speak* and *tell* can both be used with *truth*, *someone*, *everyone*, *anyone*, etc.

Type in your word:

show the most frequent 1000 words

return up to 40 collocations per page

with frequency cut-off 1000

grouping off

**cause used as Verb**

- Noun + cause: [this site may cause](#)
- Adverb + cause: [otherwise cause](#)
- cause + Adjective: [cause undesirable](#)
- cause + Noun: [cause problems](#)
- cause + Adverb: [cause primarily](#)
- cause + Verb: [cause to occur](#)
- Verb + cause: [expected to cause](#)

**cause used as Noun**

- Adjective + cause: [the leading cause](#)
- Adjective + cause: [the second leading cause](#)
- Adjective + cause: [single largest preventable cause](#)
- cause + be + Gerund or Present Participle: [cause was riding on](#)
- cause + Noun: [cause determination](#)
- Noun + cause: [the root cause](#)
- cause + of + Noun: [cause of death](#)
- Noun + of + cause: [law of cause](#)

(a)

<b>cause used as Verb: cause + Noun</b>				
cause problems	2,100,000	UK	cause actual results	1,900,000 UK
cause damage	1,400,000	UK	cause harm	860,000 UK
cause injury	590,000	UK	cause cancer	580,000 UK
cause confusion	410,000	UK	cause death	400,000 UK
cause people	330,000	UK	cause trouble	290,000 UK
cause results	260,000	UK	cause pain	250,000 UK
cause to your system	250,000	UK	cause harmful interference	250,000 UK
cause a problem	230,000	UK	cause disease	230,000 UK
cause irritation	220,000	UK	cause an increase	210,000 UK
cause a delay	200,000	UK	cause drowsiness	180,000 UK
cause interference	180,000	UK	cause birth	180,000 UK
cause permanent damage	170,000	UK	cause changes	160,000 UK
cause a denial	160,000	UK	cause the medicine	160,000 UK
cause the system	150,000	UK	cause serious problems	150,000 UK
cause side effects	140,000	UK	cause birth defects	140,000 UK
cause any problems	140,000	UK	cause infection	140,000 UK
cause some people	130,000	UK	cause offence	130,000 UK
cause dizziness	130,000	UK	cause serious injury	120,000 UK
cause serious damage	120,000	UK	cause a lot of	120,000 UK
cause this image	120,000	UK	cause a denial of	110,000 UK

(b)

&gt;&gt; next

Common collocates between **speak** and **tell**:

speak the truth	230,000	UK	tell the truth	920,000	UK
speak for anyone	63,000	UK	tell anyone	820,000	UK
speak unto thee	17,000	UK	tell thee	110,000	UK
speak with people	13,000	UK	tell people	1,300,000	UK
speak for the rest	13,000	UK	tell the rest	33,000	UK
speak with your doctor	11,000	UK	tell your doctor	360,000	UK
speak of things	11,000	UK	tell things	26,000	UK
speak for the whole	11,000	UK	tell the whole	170,000	UK
speak a lot	10,000	UK	tell a lot	63,000	UK

Different collocates of **speak** and **tell**:

speak on behalf of	300,000	UK	tell millions of	2,400,000	UK
speak the language	230,000	UK	tell a friend	850,000	UK
speak on behalf	180,000	UK	tell the difference	770,000	UK
speak a language	170,000	UK	tell the story	660,000	UK
speak about curriculum	120,000	UK	tell the story of	580,000	UK
speak the same language	100,000	UK	tell the world	430,000	UK
speak a word	98,000	UK	tell your friends	420,000	UK
(c) speak the language of	85,000	UK	tell a story	410,000	UK

**Figure 3.9** Searching facilities provided by WEB COLLOCATIONS

### 3.4.3 Building the collection

Collocations are extracted from five-grams and then organized into a digital library collection. The extraction process is fully explained in Section 4.2. This section focuses on how to organize the extracted collocations to facilitate searching and retrieving.

The collection consists of index and dictionary files that are built for each collocation type and each constituent word of a collocation. A collocation type has two to four index files, each corresponding to a particular position in a collocation. For example, noun + noun has two index files, say *i0* and *i1*; where *i0* is for the first noun and *i1* for the second. The verb + noun and adjective + noun types have four index files because they are extended to include more components (see Table 3.6). Each word in an index file occupies one line: the word, the name of the dictionary file, and the most common collocation. A dictionary file contains all collocations of a particular word in a particular position, with their frequencies.

Table 3.9 shows excerpts from index and dictionary files: *i0* is the index file for the first words of adjective + noun collocations and *c029* is the dictionary file of the adjective *front*.

**Table 3.8 Grouping templates and examples**

collocation type	template	collocation examples	template example
verb + noun	a verb word + a noun word	<i>cause serious problems</i> <i>cause unpredictable problems</i> <i>cause major problems</i>	<i>cause problems</i>
adjective + noun	an adjective word + a noun word	<i>bright sunny day</i> <i>beautiful sunny day</i> <i>warm sunny day</i> <i>hot sunny day</i>	<i>sunny day</i>
noun + verb	a noun word + a verb word	<i>time is spent on</i> <i>time is spent in</i> <i>time will be spent on</i>	<i>time spent</i>
verb + adjective	a verb word + an adjective word	<i>make it easy for</i> <i>make it easy to</i> <i>make them easy to</i> <i>make things easy for</i> <i>make life easy for</i>	<i>make easy</i>

**Table 3.9 Example of index and dictionary file**

<i>i0 (index file)</i>			<i>c029 (dictionary file)</i>	
word	dictionary file	most common collocation	collocation	frequency
<i>front</i>	c029	<i>the front page</i>	<i>the front page</i>	970964
<i>broken</i>	c041	<i>a broken link</i>	<i>the front door</i>	939981

To cater for students with different language abilities, sub-collections are generated from three language learning wordlists:

- The most frequent 1000 words in English (West, 1953)
- The most frequent 3000 words, including the above 1000 words (West, 1953)
- The most frequent 3570 words: 3000 words from above plus the 570 most popular academic words (Coxhead, 1998).

Researchers in language learning distinguish four kinds of word: high-frequency, academic, technical and low-frequency. Many studies have been conducted on identifying high-frequency words from different corpora, grouping them into

frequency-based lists like the most frequent 1000, and 2000 words (West, 1953; Hwang and Nation, 1995). West's *General Service List of English Words* contains around 2000 headwords (West, 1953).<sup>12</sup> High-frequency words make up about 80% of words in running text. Academic words are ones that are common in different kinds of academic text, covering about 9% of running words in such texts (Nation, 2001). The most popular academic word list is Coxhead's *Academic Word List*, containing 570 headwords (Coxhead, 1998).<sup>13</sup> Technical words are ones that are closely related to a topic and subject area, making up 5% of text. Low-frequency words cover about 5% of text, and form the largest group.

The sheer number of vocabulary that learners need to acquire demands different strategies for each category of word. Because of their paramount importance, high-frequency words become the primary goal of vocabulary study. For each sub-collection, a wordlist is used to filter out collocations whose constituent words are not in that wordlist. Each sub-collection has its own set of indexes and dictionary files.

### 3.4.4 Web collocations vs. BNC collocations

The extraction algorithm described in Section 4.2 was applied to the BNC in order to compare Web collocations with ones extracted from the BNC. The results underscore the massive and diverse nature of Web collocations. Table 3.10 shows the total number of collocations, the number of headwords, and the average number of collocations for each headword of each collocation type. For each collocation type, the headword (in bold) is somewhat arbitrarily selected to give some idea of how many collocations there are for a particular word.

As the table shows, 2–9 times more collocations were extracted from Web five-grams than from the BNC, and the number of collocations available for a particular headword increases accordingly. The top three types have more than ten million examples, containing 50,000 to 80,000 headwords. Even the smallest—verb + *to* + verb—contains 170,000 collocations. The most frequent Web collocation is *constitutes acceptance of* (95,000,000 times), while the most

---

<sup>12</sup>available at [http://www.lextutor.ca/freq/lists\\_download/](http://www.lextutor.ca/freq/lists_download/)

<sup>13</sup>available at [http://www.lextutor.ca/freq/lists\\_download/](http://www.lextutor.ca/freq/lists_download/)

**Table 3.10 Collocation types with statistical data from two corpora**

collocation type	Web collocations			British National Corpus		
	collocations	headwords	collocations/headword	collocations	headwords	collocations/headword
verb + noun(s)	<b>20,000,000</b>	<b>72,000</b>	277	1,700,000	64,000	27
noun + verb	6,600,000	92,000	71	800,000	27,000	30
adjective(s) + noun(s)	<b>19,000,000</b>	<b>80,000</b>	237	2,800,000	84,000	33
noun + noun	8,500,000	70,000	121	1,000,000	39,000	26
adverb + adjective	510,000	20,000	25	75,000	13,000	6
adverb + verb	1,300,000	20,000	65	180,000	12,000	15
noun + of + noun	<b>14,000,000</b>	<b>50,000</b>	280	1,200,000	41,000	29
verb + adverb	870,000	19,000	45	190,000	9,000	21
verb + adjective	230,000	16,000	14	37,000	6,600	6
verb + to + verb	170,000	9,500	17	90,000	6,200	15

**Table 3.11 Most frequent collocations of four types from two collections**

British National Corpus				Web collocations			
verb + noun	adjective + noun	noun + noun	noun + of + noun	verb + noun	adjective + noun	noun + noun	noun + of + noun
<i>take place</i>	<i>last year</i>	<i>interest rates</i>	<i>point of view</i>	<i>constitutes acceptance of</i>	<i>private message</i>	<i>web site</i>	<i>kinds of items</i>
<i>took place</i>	<i>first time</i>	<i>health care</i>	<i>sort of thing</i>	<i>make money</i>	<i>valid steam resource</i>	<i>home page</i>	<i>top of page</i>
<i>shook his head</i>	<i>same time</i>	<i>trade union</i>	<i>the end of year</i>	<i>have access</i>	<i>online review</i>	<i>credit card</i>	<i>period of time</i>
<i>do anything</i>	<i>last night</i>	<i>trade unions</i>	<i>way of life</i>	<i>share your thoughts</i>	<i>new window</i>	<i>email address</i>	<i>point of view</i>
<i>take part</i>	<i>great deal</i>	<i>member states</i>	<i>cup of tea</i>	<i>change your orders</i>	<i>respective owners</i>	<i>industry news</i>	<i>amount of time</i>
<i>said nothing</i>	<i>last week</i>	<i>car park</i>	<i>period of time</i>	<i>find answers</i>	<i>huge selection</i>	<i>business headlines</i>	<i>years of age</i>
<i>go home</i>	<i>local authorities</i>	<i>health service</i>	<i>couple of years</i>	<i>sell all kinds of</i>	<i>same time</i>	<i>review share</i>	<i>selection of books</i>
<i>see pp</i>	<i>recent year</i>	<i>income tax</i>	<i>parts of country</i>	<i>assumes all responsibility</i>	<i>wide range</i>	<i>payment methods</i>	<i>seller of item</i>
<i>take advantage of</i>	<i>young people</i>	<i>poll tax</i>	<i>end of month</i>	<i>make changes</i>	<i>real estate</i>	<i>search engine</i>	<i>bottom of page</i>
<i>had nothing</i>	<i>same way</i>	<i>labour market</i>	<i>time of year</i>	<i>take place</i>	<i>registered trademark</i>	<i>customer support</i>	<i>terms of use</i>

**Table 3.12 Web and British National Corpus entries for a collocation template**

<b>collocation</b>	<b>Web</b>	<b>BNC</b>	<b>examples</b>
<i>cause + problems</i>	285	56	<i>cause serious problems, cause major problems</i>
<i>cause + damage</i>	257	54	<i>cause permanent damage, cause significant damage</i>
<i>cause + harm</i>	147	24	<i>cause irreparable harm, cause no harm</i>
<i>cause + injury</i>	90	14	<i>cause physical injury, cause substantial injury</i>
<i>cause + death</i>	68	14	<i>cause sudden death, cause premature death</i>

**Table 3.13 Top ten *cause + noun* collocations in three concordances**

<b>Web collocations</b> 36,000 collocations		<b>British National Corpus</b> 2360 collocations		<b>Compleat Concordancer</b> 54 collocations	
<b>samples</b>	<b>frequency</b>	<b>samples</b>	<b>frequency</b>	<b>samples</b>	<b>frequency</b>
<i>cause problems</i>	2,100,000	<i>cause problems</i>	160	<i>cause problems</i>	5
<i>cause actual results</i>	1,900,000	<i>cause trouble</i>	71	<i>cause suffering</i>	4
<i>cause damage</i>	1,300,000	<i>cause damage</i>	48	<i>cause damage</i>	2
<i>cause harm</i>	850,000	<i>cause difficulties</i>	40	<i>cause offence</i>	2
<i>cause injury</i>	580,000	<i>cause cancer</i>	34	<i>cause death</i>	2
<i>cause cancer</i>	580,000	<i>cause injury</i>	32	<i>cause distress</i>	2
<i>cause confusion</i>	400,000	<i>cause death</i>	28	<i>cause a great increase</i>	2
<i>cause death</i>	410,000	<i>cause confusion</i>	27	<i>cause another war</i>	1
<i>cause trouble</i>	280,000	<i>cause harm</i>	23	<i>cause deactivation</i>	1
<i>cause pain</i>	250,000	<i>cause offence</i>	22	<i>cause a deviation</i>	1

frequent one in the BNC is *last year* (7670 times). On average, the most frequent Web collocations in each type occur 33 million times, while 76% of BNC collocations occur only once.

Table 3.11 shows the ten most frequent collocations of each type. The Web collocations are commonly found on Web pages, particularly commercial sites, such as *sell all kinds of, respective owners, credit card* and *kinds of items*. There is only one common collocation in the first two types (*take place*), two in noun + of + noun (*period of time* and *point of view*), and none in noun + noun.

Web collocations demonstrate great diversity in the language patterns they represent. For example, there are 285 variants of *cause problems*, including *cause serious problems, cause major problems* and *cause unpredictable problems*. The BNC contains only 56, half of which occur only once. Table 3.12 gives four more examples. While the sheer volume of examples could present a challenge for less

proficient learners, we believe it is valuable for advanced learners who wish to expand their range of collocation phrases for expressing propositions in precise and authentic ways.

As a final example, we include results from the *Compleat Concordancer* (Section 2.5.2). Table 3.13 shows the top ten *cause* + noun(s) collocations from three collections: WEB COLLOCATIONS, the BNC and the *Compleat Concordancer*. The first contains 36,000 collocations; the second 2360, of which 84% occur once and 8% twice, and the third 54, most of which appear just once. Interestingly, *cause problems* is the most frequent entry in all three cases. Upon further examination, it seems that *cause* is used mostly in a negative sense and associated with problems, damage, death, and so on.

## 4. Extracting collocations for language learning

---

The previous chapter investigated how to capitalize on the vast amount of human-generated text readily available on the Web by building the WEB COLLOCATIONS collection, which is designed to overcome the limitations of other collocation resources. It contains a massive volume of collocations, organized by syntactic pattern and ranked by a statistical measure. Its interface allows learners to seek collocations by specifying any constituent word, and to compare the collocates of two words to see which they have in common and which distinguish them. The present chapter explains how collocations are extracted from Web five-grams and ordered for presentation to the user.

The procedure adopted for extracting collocations has two components: a statistical measure by which collocations are ranked for presentation, and the selection of candidate collocations according to a predetermined set of syntactic patterns (as discussed in Chapter 3). The first component turns out to be extremely simple: in Section 4.1 we conduct a comparative evaluation of five measures on Web and BNC bigrams that supports the use of plain frequency for ranking. The second component meets a significant obstacle: not only are automated parsing techniques error-prone, but the problem is exacerbated by the restricted context that five-grams provide for determining the part of speech of their constituent words. This thesis research uses an open source part-of-speech tagging tool (OpenNLP). Section 4.3 assesses the impact of restricted context on its accuracy by comparing the results of tagging text in full context with that obtained when the context is restricted to five-grams. It also evaluates the tagger in a different way: by comparing its performance on five-grams with the result of another automatic tagger, namely the one used to produce the British National Corpus, on the full-context text.

Finally, Section 4.4 evaluates the quality and quantity of the WEB COLLOCATIONS collection with respect to the *Oxford Collocation Dictionary for Students of English*.

## 4.1 Extracting and evaluating collocations

Extracting collocations from a corpus of text generally involves five steps:

1. extract a set of candidate collocations from the corpus,
2. calculate a statistical score for each one,
3. rank candidates according to the scores,
4. select a predetermined number of the top candidates for manual inspection,  
and
5. identify the true collocations manually.

Candidate collocations are often word  $n$ -grams—usually bigrams. In the simplest case, the first step involves considering all pairs of consecutive words in the corpus as candidate collocations. However, linguistic analysis is sometimes applied to identify candidates that follow particular syntactic patterns, e.g., adjective + noun, or verb + noun. That is the method adopted in this thesis, and we return to it in Section 4.2.1. In the second step, there are several possibilities for the statistical score, and these are discussed below. The remaining steps are self-explanatory. Note that, in general, steps 1–4 serve to identify a set of likely collocations, from which the true collocations are selected manually in step 5 using human judgement.

This section examines three statistical approaches for ranking collocations: frequency, hypothesis testing and mutual information (Manning and Schütze, 1999). A preliminary comparative evaluation is conducted on Web and BNC raw bigrams and bigrams from which function words have been removed. Their performance on collocations that are filtered by syntactic patterns—which is the candidate selection method used in this thesis—will be discussed in Section 4.4.3.

### 4.1.1 Frequency

Frequency of occurrence is the simplest method of ranking. However, it does not work well because the  $n$  best collocations tend to be overwhelmed by small structural expressions involving function words alone. Nevertheless, Justeson and Katz (1995) obtain surprisingly accurate results using a simple heuristic: restrict collocation candidates to certain syntactic patterns, such as adjective + noun, noun

+ noun, adjective + adjective + noun, etc. This method has been widely adopted because of its simplicity.

#### 4.1.2 Hypothesis Testing

Ranking by frequency works well on syntactically filtered data. However, high frequency can be accidental. Hypothesis testing is a statistical technique to assess whether or not something is a chance event. It is based on the *null hypothesis* that the occurrence of two adjacent words  $w_1$  and  $w_2$  is independent, in which case their probability of coming together can be estimated as:

$$H_0: P(w_1 w_2) = P(w_1)P(w_2).$$

Word probabilities are calculated using the maximum likelihood estimate:

$$P(w) = \frac{f_w}{N},$$

where  $f_w$  is the frequency of word  $w$  and  $N$  is the total number of tokens in the corpus. The statistical likelihood that the event would occur if  $H_0$  were true is computed, and  $H_0$  is rejected if the likelihood falls below a certain threshold and retained otherwise. Widely used statistical tests are the *t-test*, the log-likelihood ratio, and Pearson's  $\chi^2$  test.

##### The *t-test*

The *t-test* calculates the difference between the observed and expected means, scaled by the variance of the data:

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{\sigma^2}{N}}}$$

where  $\bar{x}$  is the observed mean,  $\sigma^2$  the observed variance,  $N$  the sample size, and  $\mu$  the expected mean of the data. If the  $t$  score is large enough, the null hypothesis of independence can be rejected with a certain confidence. Assume that the  $t$  score of *powerful tea* is 0.9998 in a corpus. The value is not larger than 2.756, a critical value for a confidence level of  $\alpha = 0.005$ ,<sup>14</sup> so we cannot reject that *powerful tea*

---

<sup>14</sup> [http://en.wikipedia.org/wiki/Student%27s\\_t-distribution](http://en.wikipedia.org/wiki/Student%27s_t-distribution)

does not form a collocation. Manning and Schütze (1999) suggest that the level of significance (i.e. 2.756) itself is less useful, and the *t-test* should be used to rank collocations because a language—if compared with a random word generator—is regular so that few completely unpredictable events happen.

For ranking collocations, this method can be extended to use proportions or counts. That is:

$$\begin{aligned}\mu &= p(w_1) \times p(w_2) \\ \bar{x} &= p(w_1 w_2) \\ \sigma^2 &= p(w_1 w_2)(1 - p(w_1 w_2)) \approx p(w_1 w_2),\end{aligned}$$

where the  $p$ 's are occurrence probabilities estimated from the data. From this, it is easy to obtain:

$$t = \sqrt{f_{w_1 w_2}} \left(1 - \frac{f_{w_1} f_{w_2}}{N f_{w_1 w_2}}\right).$$

The score is high if the occurrence of the word pair is greater than would be expected by chance alone, which indicates the frequency-based nature of this method. For pairs with the same occurrence frequency, the score is greater if the occurrence of either or both words is low. Thus collocations composed of rare words are ranked higher than those of common words.

### **Log-Likelihood Ratios: (LLR)**

This method compares the hypothesis of dependence between the words with the hypothesis of independence, and estimates how much more likely one is than the other. The two hypotheses are defined as:

$$\text{Hypothesis 1: } P(w_2 | w_1) = p = P(w_2 | \neg w_1),$$

that is, the occurrence of the second word ( $w_2$ ) is independent of the occurrence of the first ( $w_1$ ), and

$$\text{Hypothesis 2: } P(w_2 | w_1) = p_1 \neq p_2 = P(w_2 | \neg w_1),$$

that is, the occurrence of the second depends on that of the first—which is good evidence for a significant collocation. Based on the assumption of binomial distributions, the likelihoods of Hypothesis 1 and 2 are:

$$L(H_1) = b(c_{12}; c_1, p) b(c_2 - c_{12}; N - c_1, p)$$

$$L(H_2) = b(c_{12}; c_1, p_1) b(c_2 - c_{12}; N - c_1, p_2),$$

where

$$b(k; n, x) = \binom{n}{k} x^k (1-x)^{n-k}, \quad p = \frac{c_2}{N} \quad p_1 = \frac{c_{12}}{c_1} \quad p_2 = \frac{c_2 - c_{12}}{N - c_1},$$

$c_1, c_2, c_{12}$  are the occurrence counts of  $w_1$ ,  $w_2$  and  $w_1 w_2$  respectively, and  $N$  is the number of tokens in the corpus.

The logarithm of the likelihood ratio  $\lambda$  is then:

$$\log \lambda = \log \frac{L(H_1)}{L(H_2)}.$$

In practice,  $-2 \log \lambda$  is used instead of  $\log \lambda$ , because it is asymptotically  $\chi^2$  distributed and can therefore be used to test the null hypothesis.

### Pearson's chi-square test ( $\chi^2$ )

The method based on the chi-squared distribution compares the observed frequency with the expected frequency for each possible outcome and rejects the null hypothesis if the difference is large. The chi-square statistic is defined as:

$$\chi^2 = \sum_i^n \frac{(O_i - E_i)^2}{E_i},$$

where  $O_i$  is the observed frequency,  $E_i$  the expected frequency, and  $n$  the number of possible outcomes of an event.

For two-word collocations, this becomes

$$\chi^2 = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11} + O_{12})(O_{11} + O_{21})(O_{12} + O_{22})(O_{21} + O_{22})},$$

where  $N$  is the number of words in the corpus,  $O_{11}$  the occurrence frequency of both words,  $O_{22}$  the non-occurrence frequency of both words,  $O_{12}$  the occurrence frequency of the first word but not the second, and so on.

### 4.1.3 Mutual information

Church and Hanks (1990) and Church et al. (1991) propose the use of mutual information, an information-theoretically motivated measure, for collocation discovery. A widely used formulation is the pointwise mutual information (PMI). Fano (1961) defines the mutual information between events  $x'$  and  $y'$  as follows:

$$I(x', y') = \log_2 \frac{P(x'y')}{P(x')P(y')} = \log_2 \frac{P(x' | y')}{P(x')} = \log_2 \frac{P(y' | x')}{P(y')},$$

in other words, the amount of information that the occurrence of the event represented by  $[y']$  provides about the occurrence of the event represented by  $[x']$ . To compute the PMI score, maximum likelihood estimates are used to calculate the probabilities.

PMI is a good measure of independence, but a bad one of dependence, because the latter relies on the frequency of individual words. For bigrams with the same frequency, those with low-frequency words receive a higher score than those with high-frequency ones.

### 4.1.4 Comparison of measures

Two experiments were conducted to compare the result of these measures using Web and BNC bigrams. Bigrams containing non-word strings, website names, and words with a mix of upper- and lower-case were removed. All words were converted to lower case, because some proper nouns like *united states* occur as frequently in lower-case as they do in upper-case. In each case bigrams were ranked separately by the five measures discussed above, and the top 100 were examined.

To illustrate the behavior of these measures, four 2-grams are chosen. Their frequency is given in Table 4.1, and their rankings according to the other four measures in Table 4.2. The first three have similar overall frequencies, but the frequency of individual words varies. *Strong*, *heavy* and *wind* are 10-20 times more frequent than *rainfall* and 200 times more frequent than *mutatis* and *mutandis*. *mutatis mutandis* is a interesting pair. The frequency of the combination and individual words are almost the same (111,000 vs. 113000 vs. 101,000). In

	<b>Frequency</b>	<b>w<sub>1</sub></b>	<b>w<sub>2</sub></b>
<i>heavy rainfall</i>	114,000	24,000,000	2,600,000
<i>strong wind</i>	110,000	51,000,000	21,000,000
<i>mutatis mutandis</i>	101,000	111,000	113,000
<i>compunctionous visitings</i>	531	1509	2396

**Table 4.1 Frequency of four 2-grams**

<b>t-test</b>	<b>LLR</b>	<b>X<sup>2</sup></b>	<b>PMI</b>
<i>heavy rainfall</i>	338	<i>mutatis mutandis</i> 3,153,000	<i>mutatis mutandis</i> $4 \times 10^{11}$
<i>mutatis mutandis</i>	319	<i>heavy rainfall</i> 1,182,000	<i>compunctionous visitings</i> $3 \times 10^{10}$
<i>strong wind</i>	318	<i>strong wind</i> 491,000	<i>heavy rainfall</i> $5 \times 10^7$
<i>compunctionous visitings</i>	23	<i>compunctionous visitings</i> 18,285	<i>strong wind</i> $2 \times 10^6$

**Table 4.2 Four 2-grams ranked by four measures**

other words, we can be virtually certain that *mutandis* will occur next if we are told that *mutatis* is the current word. *compunctionous visitings* is a rare combination that is composed of two very infrequent words.

Table 4.2 also shows the corresponding scores. As explained by Manning and Schütze (1999), other than ranking, these scores are less useful in themselves. *T-test* and LLR both demonstrate the ability to identify combinations composed of rare words. LLR performs slightly better than *t-test*, because it ranks *mutatis mutandis* higher than *heavy rainfall*. They both fail to discover *compunctionous visitings*, which reflects their frequency-based nature. In contrast,  $\chi^2$  and PMI both serve well in picking up rare words combinations like proper nouns and technical terms. PMI particularly excels in discovering combinations of low frequency words like *compunctionous visitings*.

Table 4.3 shows the top 30 Web bigrams according to each of the five measures. With one exception (*the same*) for Frequency and *t-test*, and another (*rights reserved*) for LLR, the first three columns contain bigrams that are entirely composed of function words (*the*, *in*, *with*, etc). Function words and their combinations are extremely common in English. For example, the top bigram of *the* occurs 2,700 M times in this collection; its components *of* and *the* occur 12,000 M and 19,000 M times respectively. Sophisticated methods like *t-test* and LLR seem no better than Frequency in handling these extreme cases. In fact, Frequency and *t-test* share 28 and 88 bigrams in the top 30 and 100 respectively. In contrast, the majority of the last two columns are rare word combinations.

To eliminate the interference of function words, bigrams containing them were removed; Table 4.4 shows the result and Appendix A contains a list of function words used. Frequency and *t-test* share the same top 30 bigrams—in fact there are only two differences in the top 100.  $\chi^2$  and PMI are not shown in this table because they produce exactly the same set of bigrams as before.

Considering the high frequency of Web bigrams, can similar results be obtained from a corpus of more modest size, say several hundred million words? Table 4.5 and Table 4.6 show the top 30 BNC bigrams, with and without filtering. Compared to Frequency, both *t-test* and LLR show slightly better performance on unfiltered BNC bigrams: three and five interesting pairs, respectively (in bold). Moreover, Frequency and *t-test* share only 22 and 67 BNC bigrams of the top 30 and 100 respectively. In contrast,  $\chi^2$  and PMI exhibit the same behavior as they do on Web bigrams.

With filtering, Frequency and *t-test* share 29 and 97 bigrams in the top 30 and 100 respectively, which is similar to what the Web bigrams share (30 and 98 respectively).

In conclusion, there is no “best” measure. The situation depends on what kinds of word combinations are sought: general collocations, technical terms, or extremely rare combinations. Sometimes, even the simplest method—Frequency—achieves good results. Section 4.4 evaluates the WEB COLLOCATIONS collection against a commercial collocation dictionary, and further investigates the performance of the five measures in order to select the best one for ranking the collection for the purposes of language learning.

**Table 4.3 Top 30 Web bigrams, ranked by five measures**

Frequency	<i>t-test</i>	LLR	$\chi^2$	PMI
<i>of the</i>	<i>of the</i>	<i>of the</i>	<i>mutatis mutandis</i>	<i>siliconing siliconing</i>
<i>in the</i>	<i>in the</i>	<i>in the</i>	<i>cropmark cropmarks</i>	<i>filinto filinto</i>
<i>to the</i>	<i>to the</i>	<i>will be</i>	<i>wisdens wisdens</i>	<i>telexing telexing</i>
<i>on the</i>	<i>on the</i>	<i>do not</i>	<i>constitutes acceptance</i>	<i>wisdens wisdens</i>
<i>for the</i>	<i>for the</i>	<i>on the</i>	<i>endoplasmic reticulum</i>	<i>chancier chanciest</i>
<i>and the</i>	<i>to be</i>	<i>to be</i>	<i>bona fide</i>	<i>crinolined crinolined</i>
<i>to be</i>	<i>is a</i>	<i>has been</i>	<i>exclusio alterius</i>	<i>compunctions visitings</i>
<i>is a</i>	<i>will be</i>	<b>rights reserved</b>	<i>ipsum dolor</i>	<i>trencherman trenchermen</i>
<i>with the</i>	<i>from the</i>	<i>does not</i>	<i>respective owners</i>	<i>lobworm lobworms</i>
<i>from the</i>	<i>with the</i>	<i>can be</i>	<i>selfsame costliness</i>	<i>incompetences incompetences</i>
<i>by the</i>	<i>do not</i>	<i>to the</i>	<i>slothful encrustation</i>	<i>demitted demitting</i>
<i>at the</i>	<i>at the</i>	<i>have been</i>	<i>antidisestablishmentarianism antidisestablishmentarianism</i>	<i>bossiest bossily</i>
<i>of a</i>	<i>by the</i>	<i>is a</i>	<i>brickfield brickworks</i>	<i>brrrm brrrm</i>
<i>in a</i>	<i>is not</i>	<i>such as</i>	<i>rights reserved</i>	<i>spumed spumes</i>
<i>will be</i>	<i>as a</i>	<i>may be</i>	<i>retroental fibroplasia</i>	<i>charladies charlady</i>
<i>that the</i>	<i>in a</i>	<i>is not</i>	<i>cryogenic magnetometer</i>	<i>exclusio alterius</i>
<i>do not</i>	<i>can be</i>	<i>for the</i>	<i>superoxide dismutase</i>	<i>pyrethrums pyrethrums</i>
<i>is the</i>	<i>it is</i>	<i>as well</i>	<i>raths outrabe</i>	<i>tetchily tetchiness</i>
<i>to a</i>	<i>with a</i>	<b>the same</b>	<i>et al</i>	<i>anesthetise anesthetised</i>
<i>is not</i>	<i>that the</i>	<i>should be</i>	<i>myocardial infarction</i>	<i>chirrups chirrupy</i>
<i>for a</i>	<i>has been</i>	<i>can not</i>	<i>supplied argument</i>	<i>retroental fibroplasia</i>
<i>with a</i>	<i>of a</i>	<i>from the</i>	<i>followings unread</i>	<i>demythologise demythologised</i>
<i>as a</i>	<i>of this</i>	<i>did not</i>	<i>nolo contendere</i>	<i>bathtowels bathtowels</i>
<i>of this</i>	<i>and the</i>	<i>more than</i>	<i>ending soonest</i>	<i>extemporisation extemporise</i>
<i>it is</i>	<i>does not</i>	<i>at the</i>	<i>nolle prosequi</i>	<i>petitio principii</i>
<i>can be</i>	<i>for a</i>	<i>you can</i>	<i>carbonic anhydrase</i>	<i>peristyles peristyles</i>
<i>has been</i>	<b>the same</b>	<i>it is</i>	<i>petitio principii</i>	<i>disarmer disarmers</i>
<b>the same</b>	<i>can not</i>	<i>the the</i>	<i>prima facie</i>	<i>cerecloth cerement</i>
<i>does not</i>	<i>have been</i>	<i>with the</i>	<i>avenged sevenfold</i>	<i>circularisation circularise</i>
<i>can not</i>	<i>may be</i>	<i>as a</i>	<i>substantia nigra</i>	<i>chubbily chubbiness</i>

**Table 4.4 Top 30 Web bigrams, filtered by function words**

<b>Frequency</b>	<b>t-test</b>	<b>LLR</b>
<i>rights reserved</i>	<i>rights reserved</i>	<i>rights reserved</i>
<i>web site</i>	<i>web site</i>	<i>et al</i>
<i>et al</i>	<i>et al</i>	<i>respective owners</i>
<i>private message</i>	<i>private message</i>	<i>private message</i>
<i>real estate</i>	<i>real estate</i>	<i>constitutes acceptance</i>
<i>new window</i>	<i>new window</i>	<i>real estate</i>
<i>home page</i>	<i>home page</i>	<i>web site</i>
<i>respective owners</i>	<i>respective owners</i>	<i>new window</i>
<i>site map</i>	<i>site map</i>	<i>sponsored listing</i>
<i>official time</i>	<i>official time</i>	<i>site constitutes</i>
<i>sponsored listing</i>	<i>sponsored listing</i>	<i>stay informed</i>
<i>credit card</i>	<i>credit card</i>	<i>credit card</i>
<i>constitutes acceptance</i>	<i>constitutes acceptance</i>	<i>mailing list</i>
<i>site constitutes</i>	<i>site constitutes</i>	<i>per cent</i>
<i>email address</i>	<i>mailing list</i>	<i>supplied argument</i>
<i>mailing list</i>	<i>email address</i>	<i>official time</i>
<i>please contact</i>	<i>please contact</i>	<i>make sure</i>
<i>health care</i>	<i>make sure</i>	<i>find answers</i>
<i>make sure</i>	<i>health care</i>	<i>de la</i>
<i>same time</i>	<i>same time</i>	<i>email address</i>
<i>de la</i>	<i>de la</i>	<i>health care</i>
<i>return policy</i>	<i>return policy</i>	<i>valid stream</i>
<i>per cent</i>	<i>per cent</i>	<i>site map</i>
<i>find answers</i>	<i>find answers</i>	<i>payment details</i>
<i>stay informed</i>	<i>stay informed</i>	<i>stream resource</i>
<i>payment details</i>	<i>payment details</i>	<i>return policy</i>
<i>high school</i>	<i>high school</i>	<i>home page</i>
<i>search engine</i>	<i>search engine</i>	<i>methods accepted</i>
<i>business days</i>	<i>business days</i>	<i>review helpful</i>
<i>supplied argument</i>	<i>supplied argument</i>	<i>wide range</i>

**Table 4.5 Top 30 BNC bigrams, ranked by five measures**

Frequency	T-test	LLR	$\chi^2$	PMI
<i>of the</i>	<i>of the</i>	<i>of the</i>	<i>mutatis mutandis</i>	<i>supertonic mediant</i>
<i>in the</i>	<i>in the</i>	<i>it be</i>	<i>supertonic mediant</i>	<i>continuities discontinuities</i>
<i>it be</i>	<i>it be</i>	<i>the the</i>	<i>numerus clausus</i>	<i>closures redundancies</i>
<i>to the</i>	<i>there be</i>	<i>there be</i>	<i>continuities discontinuities</i>	<i>contributors demonstrators</i>
<i>be a</i>	<i>on the</i>	<i>in the</i>	<i>closures redundancies</i>	<i>amendments additions</i>
<i>on the</i>	<i>have be</i>	<i>per cent</i>	<i>nolle prosequi</i>	<i>discounts exemptions</i>
<i>have be</i>	<i>at the</i>	<i>the be</i>	<i>teachta dala</i>	<i>revaluations devaluations</i>
<i>and the</i>	<i>be a</i>	<i>the of</i>	<i>contributors demonstrators</i>	<i>descriptions interpretations</i>
<i>to be</i>	<i>from the</i>	<i>on the</i>	<i>debito justitiae</i>	<i>performs delivers</i>
<i>there be</i>	<i>by the</i>	<i>the and</i>	<i>sese seko</i>	<i>pyroxenes amphiboles</i>
<i>for the</i>	<i>for the</i>	<i>the to</i>	<i>vrye weekblad</i>	<i>amphiboles micas</i>
<i>be the</i>	<i>with the</i>	<i>have be</i>	<i>herri batasuna</i>	<i>bollocks knackers</i>
<i>at the</i>	<i>will be</i>	<i>a the</i>	<i>amendments additions</i>	<i>boobs knockers</i>
<i>by the</i>	<i>with a</i>	<i>the a</i>	<i>discounts exemptions</i>	<i>projectors screens</i>
<i>that the</i>	<i>i have</i>	<b><i>the same</i></b>	<i>retroletal fibroplasia</i>	<i>kisses caresses</i>
<i>with the</i>	<i>to the</i>	<i>the in</i>	<i>revaluations devaluations</i>	<i>disconnecting reconnecting</i>
<i>of a</i>	<i>as a</i>	<i>more than</i>	<i>meeney miney</i>	<i>airplanes starships</i>
<i>from the</i>	<b><i>the same</i></b>	<i>at the</i>	<i>descriptions interpretations</i>	<i>geeks crapping</i>
<i>he be</i>	<b><i>the first</i></b>	<b><i>at least</i></b>	<i>skrid mvj</i>	<i>interceptions corrections</i>
<i>i be</i>	<i>he have</i>	<i>rather than</i>	<i>abundante cautela</i>	<i>ushers usherettes</i>
<i>in a</i>	<i>one of</i>	<i>of of</i>	<i>performs delivers</i>	<i>fantails lionheads</i>
<i>they be</i>	<i>i be</i>	<i>of be</i>	<i>inprint screenprinter</i>	<i>sells abhors</i>
<i>with a</i>	<i>per cent</i>	<i>from the</i>	<i>pyroxenes amphiboles</i>	<i>syllabuses syllabi</i>
<i>as a</i>	<i>can be</i>	<i>be be</i>	<i>amphiboles micas</i>	<i>unglamourous coaches</i>
<i>will be</i>	<i>they be</i>	<b><i>further far</i></b>	<i>miglior fabbro</i>	<i>flid pranny</i>
<i>have a</i>	<i>would be</i>	<i>such as</i>	<i>bollocks knackers</i>	<i>wallets marts</i>
<i>he have</i>	<i>for a</i>	<b><i>number of</i></b>	<i>boobs knockers</i>	<i>widows widowers</i>
<i>i have</i>	<b><i>part of</i></b>	<b><i>part of</i></b>	<i>projectors screens</i>	<i>infl plu</i>
<i>for a</i>	<i>which be</i>	<i>if you</i>	<i>requiris circumspice</i>	<i>taxes tips</i>
<i>have to</i>	<i>to be</i>	<i>by the</i>	<i>kisses caresses</i>	<i>doodad doohickey</i>

**Table 4.6 Top 30 BNC bigrams, filtered by function words**

Frequency	t-test	LLR
per cent	per cent	per cent
year old	year old	prime minister
take place	take place	united states
prime minister	prime minister	year old
local authority	local authority	local authority
same time	same time	et al
united states	united states	take place
long term	long term	northern ireland
new york	new york	united kingdom
look like	look like	new york
make sure	make sure	long term
et al	et al	soviet union
united kingdom	united kingdom	same time
northern ireland	northern ireland	working class
young man	young man	co operation
labour party	labour party	trade union
working class	working class	see pp
soviet union	soviet union	labour party
world war	world war	wide range
long time	trade union	make sure
trade union	co operation	interest rate
co operation	great deal	world war
great deal	interest rate	great deal
interest rate	see pp	middle class
young people	young people	look like
see pp	long time	young man
year later	year later	managing director
large number	large number	european community
wide range	wide range	hewlett packard
old man	high level	large number

## 4.2 Determining candidate collocations

The previous section began with a five-step procedure for extracting collocations from a corpus of text and evaluated the performance of five statistical ranking criteria on bigrams extracted from the Web and BNC collections. Here we describe how this general procedure is adapted for use in this thesis. First, our focus is on collocations for language learning, and for this purpose it is extremely helpful to know their syntactic structure—and to extend the analysis beyond bigrams to useful short phrases of different lengths. Second, the procedure must be adapted for use with the raw material from which we extract collocations, namely the Web *n*-gram data. This means that step 5, which involves human

intervention, is unfeasible because of the massive volume of collocations, which number in the millions.

Syntactic tagging is an important component of the method that we use to identify collocations. We adopt OpenNLP, an open source part-of-speech tagging tool. The restricted context available in the Web  $n$ -gram collection inevitably increases the number of errors produced when tagging. To minimize these, we use the largest available  $n$ -grams ( $n=5$ ), parse them, and extract candidates that match particular syntactic patterns.

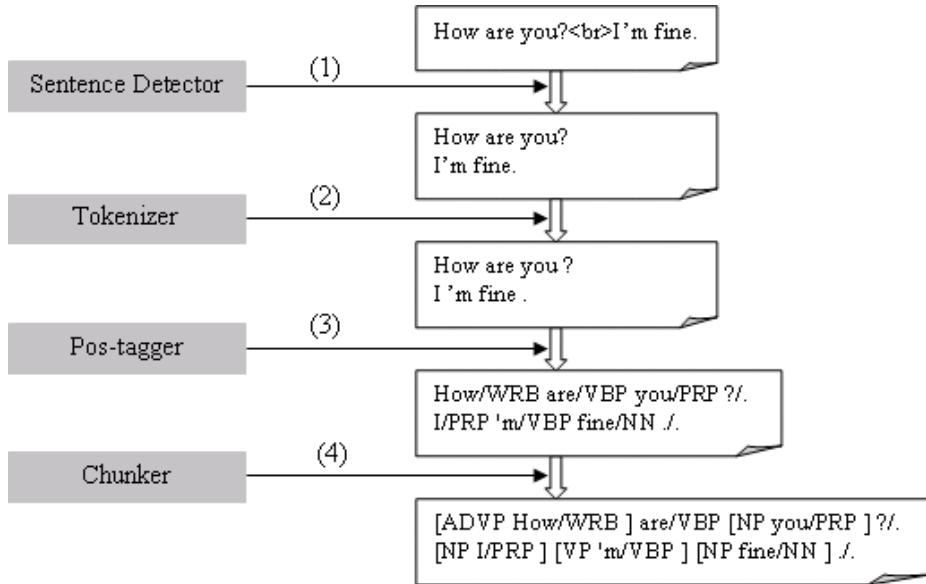
The process used to extract collocations is summarized in five steps:

1. assign part-of-speech tags to five-grams,
2. match tagged five-grams against syntactic patterns,
3. discard “dirty” collocations,
4. calculate a statistical score for each one, and
5. rank collocations for presentation to the user.

In step 1, the OpenNLP tagger is used to assign part-of-speech tags to five-grams. Then, in step 2, the tagged five-grams are compared against regular expressions that specify the syntactic patterns that were introduced and justified in Chapter 3, and those that match are extracted as candidate collocations. Some extracted collocations are messy because they contain a haphazard mix of upper- and lower-case letters, unconventional single-character words (other than the article *a* or pronoun *I*) such as *time t*, *p values*, and *m sections*, or repeated words such as *part part*, *pain pain* and *man man*; Step 3 discards these “dirty” collocations because they are not useful for learning. Step 4 calculates a statistical score, and step 5 presents the results to the user without any manual selection. Below we discuss steps 1, 2 and 5 in more detail.

#### 4.2.1 Syntactic tagging

Throughout this thesis research we use the OpenNLP package for syntactic tagging. Released under GNU Lesser General Public license (available at [opennlp.sourceforge.net](http://opennlp.sourceforge.net)), this is a collection of Java-based natural language



**Figure 4.1 Parsing a document**

learning tools that perform sentence detection, tokenization, part-of-speech tagging, and chunking.

Parsing involves the four steps illustrated in Figure 4.1. The first detects sentence boundaries and splits the input into individual sentences. Then sentences are converted into tokens. The tokenizer separates punctuation: for example, *you?* becomes two distinct tokens *you* and *?*. It also detects contractions, that is, shortened forms in which a subject and an auxiliary verb, or an auxiliary verb and *not*, are combined into a single word, and splits them into two parts—for example, *I'm*, *we're*, *you'd*, *can't*. The result of these two steps on the text “*How are you? I'm fine.*” is the following eight tokens:

How	are	you	?	I	'm	fine	.
-----	-----	-----	---	---	----	------	---

Next, the tagger performs tagging: it assigns a part-of-speech tag to each word. These tags begin with a letter that conveys the basic class and follow it with letters that qualify the class. For example, *N...* and *V...* indicate noun and verb; *NN* and *VBP* signal a singular noun and a non-third-person singular present verb. OpenNLP’s tagger adopts the Penn Treebank tagset (see Appendix B). It comprises three levels: word, phrase and clause; we use only the word-level tags. Finally, the chunker assigns non-overlapping phrase and clause tags.

OpenNLP utilizes the Maxent package, which implements the Maximum Entropy method for constructing statistical models for classification tasks.<sup>15</sup> It consults a tag dictionary that contains words and their associated part-of-speech tag(s), and a statistical model for applying part-of-speech tags to each token in a sentence. The model is trained on text from the *Wall Street Journal*, and the *Brown Corpus*. The more similar the text under investigation is to the training text, the more accuracy the tagger can achieve.

OpenNLP's dictionary contains only 16,200 words. This is completely inadequate for tagging Web five-grams, even ones filtered by the BNC wordlist, which contains 253,000 words. Thus we were obliged to produce a larger dictionary, which was done by importing words and their part-of-speech tags from the BNC. We did not retrain OpenNLP's statistical model, partly because training should be conducted on a large set of pre-tagged training text of similar character to Web text, which is practically not available, but mainly because improving the tagger lies beyond the scope of the thesis. More words and tags were added into the dictionary based on the BNC's wordlist.

Three steps were applied when compiling the new dictionary:

1. produce a mapping between the OpenNLP and CLAWS tagsets,
2. for existing words, keep the old tags and add new ones as necessary, and
3. add new words and their associated tags.

The BNC corpus has been tagged automatically by CLAWS, a dedicated general-purpose grammatical tagger. The wordlist we adopted contains CLAWS part-of-speech tags for each word. However, they cannot be directly used because OpenNLP and CLAWS employ different tagsets, and so a mapping between the two is needed. CLAWS will be discussed in Section 4.3.2; here we focus on its tags.

CLAWS uses the CLAWS5 tagset,<sup>16</sup> which contains 62 different tags. As mentioned earlier, OpenNLP uses the Penn Treebank tagset, with 39 tags. The latter tags are more general than that of the former. For example, CLAWS5

---

<sup>15</sup> <http://maxent.sourceforge.net/>

<sup>16</sup> <http://ucrel.lancs.ac.uk/claws5tags.html>

includes dedicated tags for the verbs *be*, *have*, *do*, whereas Penn Treebank treats these the same as other verbs. In CLAWS5, the word *to* can be either a preposition (PRP) or an infinitive marker (TO), while Penn Treebank makes no such distinction.

In most cases, there is a straightforward one-to-one mapping between the tags in the two sets, as shown Table 4.7. However, in some cases it is necessary to map two or more tags into a single tag in the other. If a tag occurs in one tagset only, the closest corresponding tag in the other set is chosen: examples are WP and PNQ in Table 4.7. If there is no corresponding tag in the other set, the tag UN (unknown) is used, as with the tags LS and UN in the first column of Table 4.7. Finally, tags that are overly specific in one set are mapped to the corresponding more general tag in the other. For example, VHD signifies the past tense of *have* in CLAWS5 and is mapped to VBD—a general past tense verb—in Penn Treebank.

In step 2, only nouns, verbs, adjectives, adverbs and pronouns were considered because OpenNLP’s original dictionary provides sufficient coverage of closed-class words such as prepositions, conjunctions, and determiners. A word may be present in OpenNLP’s dictionary, but a new meaning may occur in the BNC list. For example, JJR (comparative adjective) may be added to *better* and NN (noun) to *cut*, if they are not in the original dictionary.

The resulting dictionary contains 173,535 words, ten times larger than the original one.

#### 4.2.2 Matching tagged *n*-grams against statistical patterns

In step 2 of the procedure to extract collocations, set out at the beginning of this section, the tagged five-grams are compared against ten regular expressions, shown in Table 4.8, defined for syntactic patterns in Table 3.6. For example, the pattern for *verb + noun* is:

$$\begin{aligned} & \text{word/VB[DZP]?} + (\text{word/IN})? + (\text{word/DT})? + (\text{word/JJ})? + (\text{word/NN[S]})? \\ & + (\text{word/NN[S]})^* \end{aligned}$$

**Table 4.7 Tag mapping between Penn Treebank and CLAWS5**

<b>Penn Treebank</b>	<b>CLAWS5</b>	<b>definition</b>
JJ	AJ0	adjective
JJR	AJC	comparative adjective
JJS	AJS	superlative adjective
DT	AT0	article
RB, RBR, RBS	AV0	adverb
RP	AVP	adverb particle
WRB	AVQ	wh-adverb
CC	CJC	coordinating conjunction
IN	CJS	subordination conjunction
IN	CJT	the conjunction that
CD	CRD	cardinal number
PRP\$	DPS	possessive determiner form
DT, PDT	DT0	general determiner
WDT	DTQ	wh-determiner
EX	EX0	existential THERE
UH	ITJ	interjection or other isolate
NN	NN0, NN1, PNI	neutral noun and single noun
NNS	NN2	plural noun
NNP, NNPS	NP0	proper noun
PRP	PNP, PNX	personal noun
WP	PNQ	wh-pronoun
POS	POS	the possessive 's or '
IN	PRF, PRP	preposition
LS	UN	unknown
UN	PUL,PUN,PUQ,PUR,NULL,ORD	unknown
TO	TO	to
FW	UNC	Foreign words
VB,VBP	VBB, VDB,VNB, VVB,VBI,VDI, VHI,VVI	verb, base form
VBD	VBD,VDD,VND,VVD	verb, past tense
VBG	VBG,VDG,VNG,VVG	verb, gerund or present participle
VBN	VBG,VDG,VNG,VVG	verb, past participle
VBZ	VBZ,VDZ,VNZ,VVZ	verb, 3 <sup>rd</sup> person singular present
RB	XX0	the negative NOT or N'T
SYM	ZZ0	Symbol

A verb + noun collocation must begin with a verb (VB), which could be in base, past, or present form, followed by an optional preposition (IN), an optional article (DT), an optional adjective (JJ), a compulsory noun (NN) and optional nouns. Patterns that match any of the ten regular expressions are grouped by collocation type; ones that do not match are discarded.

**Table 4.8 Regular expressions for ten collocation types**

collocation type	regular expression
verb + noun(s) includes: verb + noun + noun verb + adjective + noun(s) verb + preposition + noun(s)	word/VB[DPZ]? + (word/IN)? + (word/DT)? + (word/JJ)? + (word/NN[S]?) + (word/NN[S]?)*
verb + adverb	word/VB[DPZ]? + (word/IN PR)? + word/RB
noun + noun	(word/DT)? + (word/NN[S]?) + (word/NN[S]?)
noun + verb includes: noun + verb with present tense noun + be + present participle noun + be + past participle	1. (word/DT)? + (word/NN) + (word/VBZ VBP) + (word/IN PR)? 2. (word/DT)? + (word/NN) + (is was are were) + (word/VBG) + (word/IN PR)? 3. (word/DT)? + (word/NN[S]?) + (is was are were) + (word/VBN) + (word/IN PR)?
noun + of + noun	(word/DT)? + (word/NN[S]?) + of + (word/DT)? + (word/NN[S]?)
adjective(s) + noun(s) includes: adjective + noun + noun adjective + adjective + noun(s)	(word/DT)? + (word/JJ) + (word/JJ)* + (word/NN[S]?) + (word/NN[S]?)*
verb + adjective includes: verb + adjective verb + noun + adjective	word/VB[DPZ]? + (word/IN PR)? + (word/NN[S]?)? + (word/JJ)
verb + <i>to</i> + verb	word/VB[DPZ]? + <i>to</i> + word/VB
adverb + verb	word/RB + word/VB[DPZ]? + (word/IN PR)?
adverb + adjective	word/RB + (word/JJ)

#### 4.2.3 Ranking the result

In printed dictionaries, collocations are organized by syntactic pattern and ordered in various ways. Some dictionaries show the most frequent or idiomatic ones first; others use arbitrary ordering. Given a list of collocations derived from the Web *n*-gram data, our goal is to present good collocations at the top of the list and relegate poor ones to the bottom. To accomplish this, we tested the five standard statistical measures introduced in Section 4.1 and selected the best for ranking extracted collocations, as explained in Section 4.4.3. It turned out to be a particularly simple one—plain frequency of occurrence.

### 4.3 Investigating tagging errors

Despite extensive research, all taggers make errors, and OpenNLP is no exception. It is simply not possible to obtain perfect results because of the complexity and ambiguity of human language. These systems rely on context and predefined rules to infer part-of-speech tags for each word—for example, whether *cut* is a verb or a noun in a given context. And because of the restricted context, errors inevitably occur more frequently when the input is five-grams. This section investigates to what extent this restricted context affects the performance of the OpenNLP tagger.

#### 4.3.1 Tagging the BNC

The first experiment compares the performance of the OpenNLP tagger in full context with the restricted context imposed by five-grams. The procedure is:

1. tag the BNC text in full context,
2. extract tagged five-grams,
3. extract raw five-grams (untagged),
4. tag raw five-grams,
5. compare five-grams tagged in steps 2 and 4, and
6. count the unmatched tags.

To obtain baseline data, the BNC text was tagged by OpenNLP, and five-grams were extracted. The corpus contains both written and transcribed spoken text, but the latter was not used because the mis-pronunciations and unplanned repetition it contains—for example *I er, mean, I mean*—present a great challenge to taggers (Leech et al., 1994). Furthermore, Web five-grams are unlikely to contain such text, given their written nature.

In steps 3 and 4, five-grams were extracted from the BNC and tagged in isolation.

The tags assigned in the two contexts were compared one by one, and unmatched ones were counted. Among the total of 54,000,000 tags, there were 82% matches. Unmatched tags were organized into 332 categories. Table 4.9 shows the most common 17, each of which accounts for at least 1% of the total. In this table, *verb* vs. *noun* means that a word was classed as a *verb* in full context, but marked as a *noun* in five-gram context, or vice versa. The remaining 315 categories, which

**Table 4.9 Categories of mismatched tags in full and five-gram context**

mismatched tag category	percentage
verb vs. noun	21.5%
past tense verb vs. past participle verb	16.4%
adjective vs. noun	12.4%
adjective vs. adverb	6.7%
wh-determiner vs. preposition	4.2%
preposition vs. adverb	3.8%
adjective vs. past participle	3.7%
particle vs. preposition	3.4%
noun vs. adverb	2.3%
numeral, cardinal vs. Noun	2.2%
adverb vs. particle	1.7%
pre-determiner vs. determiner	1.4%
-ing form of verb vs. adjective	1.3%
noun vs. modal	1.3%
verb vs. adjective	1.2%
noun vs. proper noun	1.1%
preposition vs. determiner	1.1%
other	13.9%

together account for 13.9% of the total mismatched tags, were merged into a single *other* category.

The fact that 82% of the tags match indicates that the context provided by five-grams is generally sufficient for tagging purposes. However, the wide variety of mismatched tags (332 categories) suggests that context does play an important role in part-of-speech tagging, particularly when determining whether a word is a verb or a noun, a past tense or past participle verb, and an adjective or a noun. These three categories account for half the mismatched tags, and result in mistakenly assigned collocation types. Collocations of the form noun + noun, noun + verb, verb + noun, and adjective + noun are particularly prone to tagging errors caused by the restricted context. Consequently, some collocations are assigned to the wrong category, or the same collocation is assigned to two different categories. For example, *time lags* is marked as both a *noun + verb* and a *noun + noun* collocation.

**Table 4.10 Percentage of matched tags in three experiments**

<b>words removed</b>	<b>five-grams discarded</b>	<b>accuracy</b>
none	0	55.3%
(1) words that are marked as unknown	15.8%	68.2%
(2) cardinal numbers, foreign words and pronouns		
<i>same, few, fewer, such, many, either, whose, what, to, that, where, when</i>	22%	77.2%

**Table 4.11 Examples of inconsistent tagging between OpenNLP and CLAWS**

<b>word</b>	<b>OpenNLP</b>	<b>CLAWS</b>
<i>550kg</i>	cardinal number	noun
<i>1954s</i>	noun	cardinal number
<i>someone</i>	noun	cardinal number
<i>s11</i>	noun	foreign words
<i>voce</i>	noun	foreign words
<i>de</i>	noun	foreign words
<i>India</i>	noun	proper noun
<i>Omphalos</i>	proper noun	noun
<i>February</i>	noun	proper noun

### 4.3.2 Comparison with CLAWS

The second experiment assessed the accuracy of the OpenNLP tagger against another standard. We did not have access to a large hand-tagged corpus to use as a gold standard. However, as mentioned earlier, the BNC corpus has been tagged by CLAWS, although no post-editing was undertaken to correct tagging errors. Some words have dual tags (like VVB-NN1) indicating that the tagger was unable to determine which category is correct, with sufficient confidence. CLAWS is undoubtedly a more advanced and accurate tagger than OpenNLP, and is claimed to achieve an error rate of 1.15% and an ambiguity rate of 3.75% in the tags it assigns (Leech and Smith, 2000).

The procedure for this experiment was:

1. extract five-grams tagged by CLAWS,
2. generate untagged five-grams,
3. re-tag them using OpenNLP,
4. compare the tags assigned in steps 1 and 3, and
5. count the unmatched tags.

**Table 4.12 Words used to filter five-grams**

	<b>words</b>	<b>OpenNLP</b>	<b>CLAWS</b>
Group 1	<i>same, few, fewer, such, many</i>	adjective or adverb	determiner
	<i>either</i>	conjunction	adverb or determiner
	<i>whose</i>	possessive wh-pronoun	wh-determiner
	<i>what</i>	wh-pronoun	wh-determiner
Group 2	<i>to</i>	TO	infinitive marker or TO
	<i>that</i>	conjunction, wh-determiner, determiner	conjunction, or determiner
	<i>where, when</i>	wh-adverb	wh-adverb or conjunction

In step 1, five-grams were extracted from the BNC's written text. These had been tagged by CLAWS in full context. Five-grams containing ambiguous tags, which account for about 20% of the total, were discarded. Then untagged versions of the five-grams were generated by stripping all tags, and these were retagged by OpenNLP. Step 4 of the above procedure is to compare the tags assigned by the two systems using the mapping in Table 4.7. The percentage of matching tags is given in Table 4.10, for each of three cases.

First, 55.3% of the tags match, without any processing or filtering.

Second, the two taggers yield inconsistent results for cardinal numbers, foreign words and pronouns, so all five-grams containing any such words, or any words tagged as "unknown," are removed. Table 4.11 gives examples of inconsistent tagging. This process discards 15.8% of the five-grams, and of the remainder, 63.1% of the tags match correctly.

Third, the five-grams were filtered by removing those containing the words shown in Table 4.12. These extremely common words are ambiguous with regard to syntactic class and therefore particularly prone to tagging inconsistency. Table 4.12 divides them into two groups. The first group—*same, few, fewer, such, many, either, whose, what*—are consistently assigned different tags by the two taggers. For example, OpenNLP treats *same* as an adjective, but according to CLAWS it is a determiner. The second group contains words for which the taggers have specialized tags. For example, OpenNLP indiscriminately applies the TO tag to any instance of the word *to*, whereas CLAWS assigns the TO tag to the infinitive *to* (as in *I want to go*) and to the preposition tag *to* (as in *vans raced to the side*).

**Table 4.13 Categories of mismatched tags between OpenNLP and CLAWS**

mismatched tag category	percentage
verb <i>vs.</i> noun	14.0%
adjective <i>vs.</i> noun	11.5%
past tense verb <i>vs.</i> past participle	9.9%
particle <i>vs.</i> preposition	7.7%
noun <i>vs.</i> adverb	7.6%
adjective <i>vs.</i> determiner	6.4%
adjective <i>vs.</i> adverb	6.1%
single nouns <i>vs.</i> plural nouns	5.8%
preposition <i>vs.</i> adverb	4.7%
adverb <i>vs.</i> particle	4.5%
adverb <i>vs.</i> determiner	4.4%
-ing form of verb <i>vs.</i> adjective	4.2%
noun <i>vs.</i> adverb	1.7%
adverb <i>vs.</i> EX	1.3%
verb <i>vs.</i> preposition	1.3%
noun <i>vs.</i> modal	1.0%
other	7.9%

Removing these twelve ambiguous words discards a further 22% of the five-grams, and 77.2% of the remaining tags match correctly.

Table 4.13 shows what percentage of the final errors (case 3 of Table 4.10) is accounted for by the most common tag mismatches. The top 16 mismatch types are shown; the remaining 35, which individually account for less than 1% of errors, are merged into the *other* category. The top three mismatches are *verb vs. noun*, *adjective vs. noun* and *past tense verb vs. past participle*. This is consistent with the results of Section 4.3.1. Earlier, there were 332 possible types of mismatch (Table 4.9 shows the most common 17), whereas here there are only 51 (the 16 shown in Table 4.13, plus 35 others). The discrepancy between the two figures is attributed to the way the two tagsets are mapped, and to the filtering operation that has been applied here. Tagging inconsistency between the two taggers adds considerable complexity to the experiments.

#### 4.4 Evaluating extracted collocations

The primary obstacle to evaluating WEB COLLOCATIONS is finding an authoritative database to serve as baseline data. The *Collins Collocation Sampler* (Section 3.4.1) seems ideal, but its output is restricted to 100 collocates regardless of word type.

The online *Compleat Concordancer* (Cobb, n.d.) is one of the best on the Web, and free to use, but is based on a collection of rather small corpora ranging from 80,000 to 4M words. After investigation, we decided to build the baseline data from the *Oxford Collocation Dictionary for Students of English* (OCDSE). It is based on a relatively large corpus—the BNC Corpus—and contains about 150,000 collocations for 9,000 headwords, organized into eleven collocation types (Section 2.5.1).

#### 4.4.1 Baseline collocation data

Table 4.14 shows the number of collocations contained in this dictionary. For each type it gives the number of headwords, the number of collocations, and some examples. Adjective + noun collocations constitute the largest group (37.5%), followed by verb + noun (19.2%), adverb + adjective (7.0%), and so on. It is unclear how this dictionary was generated: automatically, manually or both? Upon further investigation, it was found to include some arguable collocations—such as *19<sup>th</sup> century*, *\$20 reward*, *children's book* and *men's loo*.

The dictionary contains about 185,000 collocations in all, considerably more than the 150,000 that it claims. Only adjective + noun, noun + noun and adverb + adjective collocations, comprising 52% of the total, were used as baseline data because the other types are unsuitable for the reasons given in Table 4.15. Moreover, a further 6000 were discarded because they contain:

1. more than two words: *hormone replacement therapy*, *credit card number*, *social security system*
2. numbers: *19<sup>th</sup> century*, *10% share*, *500 workforce*, *10 chance*, and *\$20 reward*
3. proper nouns: *Argentinian nationality*, *AIDS diagnosis*, *Ashkenazi Jew*, *NATO country*, *Asian elephant*
4. hyphenated words: *full-time diploma*, *good-looking man*, *world-class player*
5. possessive nouns: *children's book*, *men's loo*, *artist's model*.

**Table 4.14 Number of collocations extracted from the *Oxford Collocation Dictionary for Students of English***

collocation type	headwords	collocations	example
adjective + noun	4997	69362 (37.5%)	<i>vague recollection</i>
verb + noun	4529	35516 (19.2%)	<i>keep the promises</i>
noun + preposition or preposition + noun	3584	12475 (6.7%)	<i>in press, position on</i>
noun + verb	1846	8091 (4.4%)	<i>plot unfolds</i>
noun + noun	2100	12283 (6.6%)	<i>plot development</i>
adverb + verb or verb + adverb:	1436	10144 (5.5%)	<i>directly recruit, recruited specially</i>
verb + <i>to</i> + verb:	749	3539 (1.9%)	<i>try to recruit</i>
verb + preposition:	1076	3027 (1.6%)	<i>recruit as</i>
adverb + adjective:	1450	13006 (7.0%)	<i>awfully careful</i>
verb + adjective	1464	7605 (4.1%)	<i>be + careful</i>
adjective + preposition	689	1121 (0.61%)	<i>careful about</i>
phrases	2791	8850 (4.8%)	<i>a plot of land</i>

Collocations containing more than two words are discarded because they have arbitrary lengths (three to four), and two-word collocations form the vast majority (91%). Those containing particular types of words—numbers, hyphenated words, proper and possessive nouns—were discarded because they are not included in the WEB COLLOCATIONS collection. These operations reduced the baseline data by a factor of two, to 88,000 collocations. These were divided into three types—adjective + noun, noun + noun and adverb + adjective—and grouped by headword.

#### 4.4.2 Test data

For each of the three collocation types, test data was extracted from the WEB COLLOCATIONS collection and organized by headword. The 16 headwords from the baseline collection shown in Table 4.16 are not covered by WEB COLLOCATIONS due to

- tagging errors: *sick* and *multinational* could be nouns, but are not recognized by OpenNLP

- inconsistent word class assignment between OCDSE and OpenNLP:  
*discredit, lord, and yes* are treated as nouns by OCDSE, but not by OpenNLP.

To make the baseline and testing data comparable and consistent, words shown in Table 4.16 and their collocations are removed from the former. Table 4.17 gives the size of the two data sets (headwords are in bold). The largest group—adjective + noun—covers 4,234,318 Web collocations with 870 per headword, which is almost 66 times larger than the 62,919 OCDSE collocations with 13 per headword.

#### 4.4.3 Ranking the Web collocations

For each collocation in the test data, the five statistical scores discussed in Section 4.1 were computed, and the collocations ranked accordingly. Then precision–recall curves were generated. Precision—a measure of fidelity—is computed as the number of Web collocations that are baseline collocations, divided by the total number of Web collocations. Recall—a measure of completeness—is computed as the number of Web collocations that are baseline collocations, divided by the total number of baseline collocations.

$$\text{Precision} = \frac{|\{\text{baseline\_collocations}\} \cap \{\text{web\_collocations}\}|}{|\{\text{web\_collocations}\}|}$$

$$\text{Recall} = \frac{|\{\text{baseline\_collocations}\} \cap \{\text{web\_collocations}\}|}{|\{\text{total\_baseline\_collocations}\}|}$$

For example, the word *happy* has 28 adverb + adjective baseline collocations and 646 Web collocations. Of the top 10 Web collocations (as ranked by a particular measure), 4 are baseline collocations and 6 are not. For this measure, precision at this point is 80% (8/10) and recall is 28.5% (8/28). Precision-recall curves are generated by varying the cut-off value (10 in the above) and plotting precision against recall.

Each headword of a collocation type is associated with a list of collocations; thus a precision-recall curve can be generated for each headword. However, this is unhelpful because there would be over a thousand curves for each measure. Instead, we average the recall and precision scores for each headword to generate a single curve for each measure. Separate evaluations are conducted for adjective

**Table 4.15 Reasons why particular collocation types are not used in the evaluation**

collocation type	reason for discarding
verb + noun	There are many collocations of non-consecutive words, and the number of constituent words that are included is inconsistent.
noun + verb	The verb can be in different forms based on the preceding noun—for example, <i>the moment arrives</i> —but only the base form, ( <i>arrive</i> ) is given in the dictionary.
verb/noun/adjective + preposition	The evaluation focuses on collocations consisting of content words: verbs, nouns, adjectives, and adverbs.
adverb + verb or verb + adverb	The position of adverbs is not indicated clearly in the dictionary. They can occur on either side of a verb, sometimes both.
verb + <i>to</i> + verb	There are few examples of this type, and the length of collocations varies from two to four words.
phrases	arbitrary length and form

**Table 4.16 Headwords that are not covered by Web collocations**

collocation type	tagging errors	inconsistency of word class assignment
adjective + noun	<i>sick, multinational</i>	<i>discredit, lord, yes</i>
<b>noun1</b> + noun2	<i>lunatic, cymbal</i>	<i>yes</i>
adverb + adjective	<i>adjust, acquainted,</i> <i>adjourn, bonkers</i>	<i>set, misplaced, bothered, shattered</i>

**Table 4.17 Number of collocations in the baseline and test data**

collocation type	headword	OCDSE collocations	average	Web collocations	average
adjective + noun	4863	62,919	13	4,234,318	870
<b>noun1</b> + noun2	2048	11,836	5.8	1,459,283	712
adverb + adjective	1420	11,385	8	24,9147	175

+ noun, adverb + adjective, and noun + noun collocation types. Figure 4.2 shows the precision-recall curves that result of these three collocation types, where  $n$  is 1 to 100 (recall rate becomes stable once  $n$  reaches 100).

We can immediately discard the  $\chi^2$  and PMI measures because their precision-recall scores lie below those for the other three measures across all types. For the adjective + noun collocation type, Table 4.18 shows the precision at 10%, 35% and 60% recall for the Frequency, *t-test* and LLR measures (the largest figure in each row is in bold type). The performance of these three measures is extremely

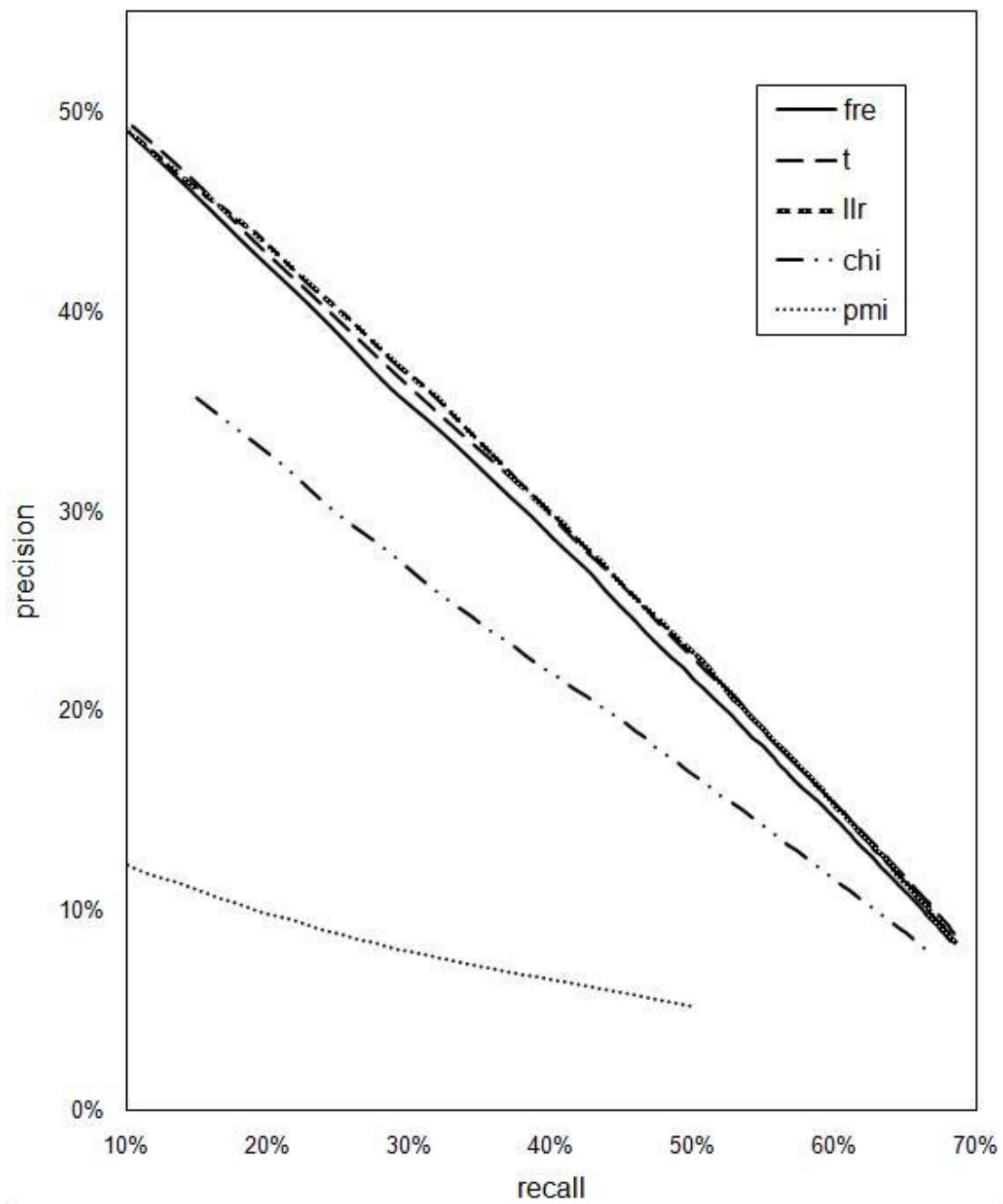
close. None of them outperforms the others, but *t-test* is marginally better on average. In general, as these figures show, one measure may be better than the others at some points, but the difference is small.

The noun + noun curve shows a surprising result that has not to our knowledge been observed in other, similar experiments: Frequency and *t-test* outperforms LLR. Again, the difference does not seem to be significant.

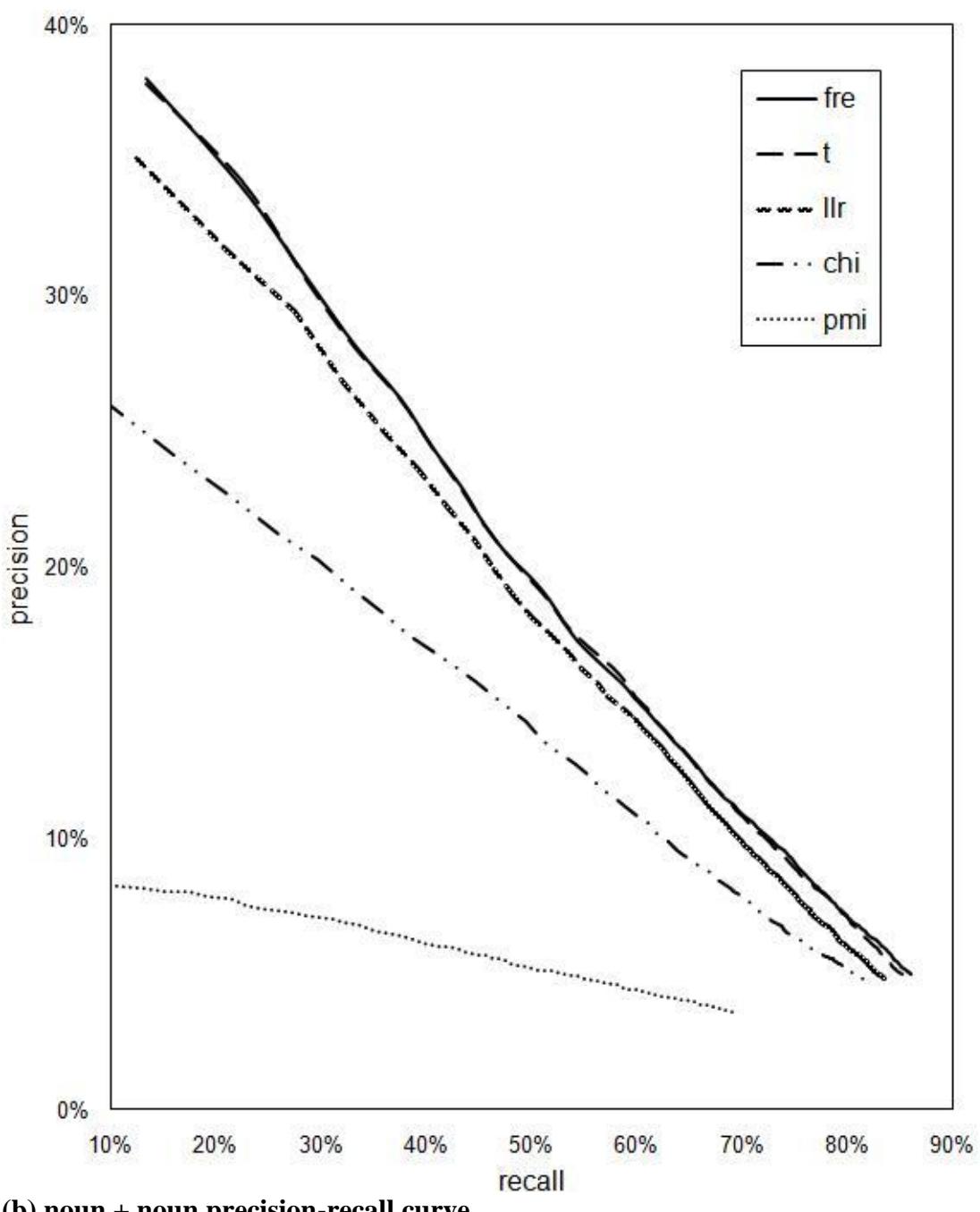
The adverb + adjective curve presents a rather interesting picture. The performance of  $\chi^2$  is strong at the beginning, but drops sharply in comparison when recall exceeds 25%. Frequency and *t-test* have a slow start, but catch up with LLR once recall reaches 45%.

In summary,  $\chi^2$  and PMI are unsuitable for ranking collocations for the purposes of second language learning because collocations for learning are common and frequent in nature. Frequency and *t-test* exhibit similar behavior across all three collocation types, which reflects the frequency-biased nature of *t-test*. Their relatively poor performance on adverb + adjective collocations at low recall values is attributed to an overwhelming number of collocations involving a small group of adverbs—*very*, *quite*, *always*, *pretty*, *just*, *more*, *most*—that are extremely frequent and can partner with almost any adjective. Unlike most collocation dictionaries, OCDSE does include collocations containing such adjectives, but not all possible ones. LLR delivers good and consistent performance. However, the difference between it and Frequency and *t-test* is small.

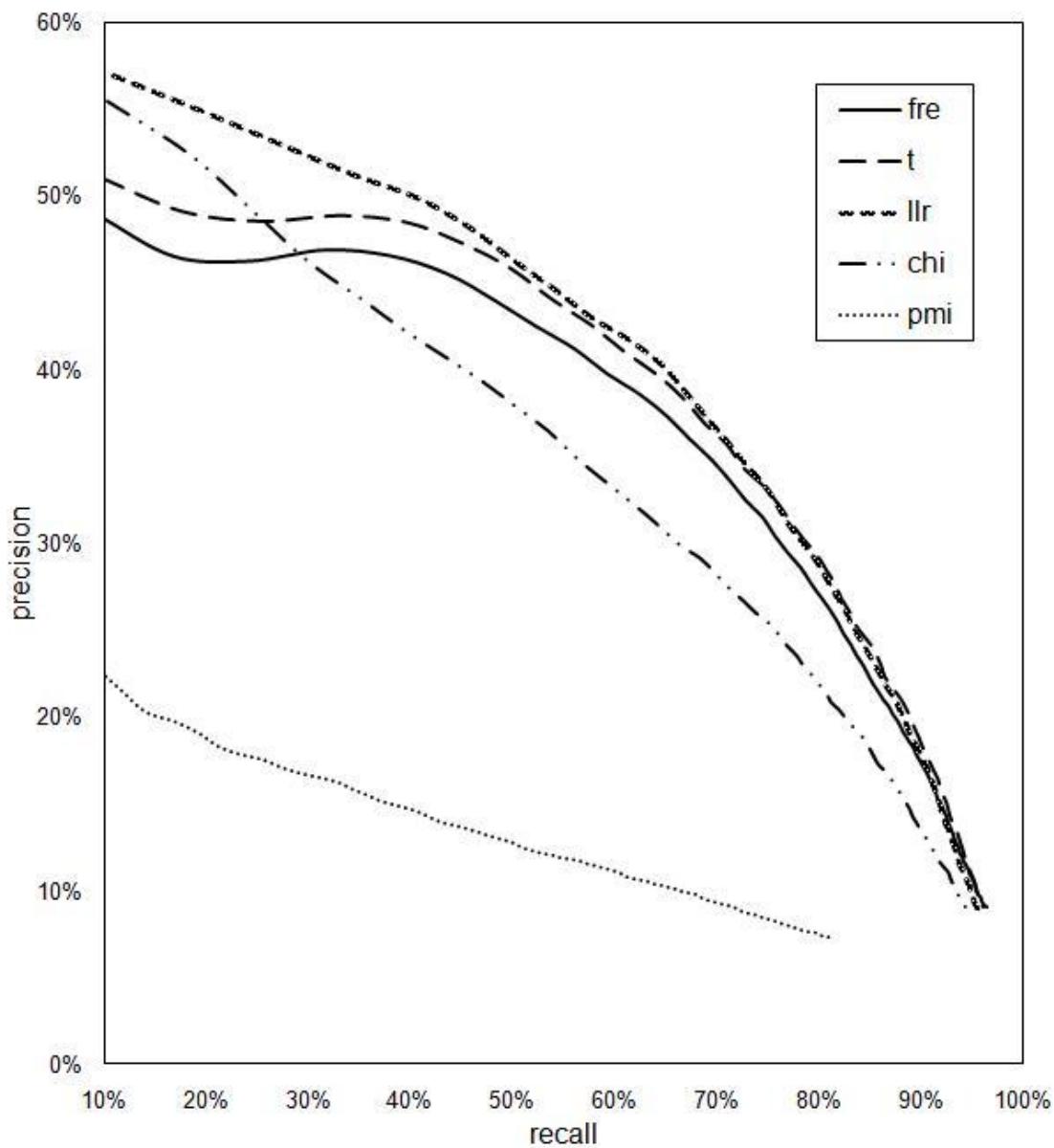
One explanation of the better performance of Frequency on noun + noun, which has heretofore been unobserved, is that the datasets used in previous experiments are relatively small, containing one to several hundred million words. The frequency of individual words and word combinations is low compared to those in the WEB COLLOCATIONS collection. For example, a relatively frequent pair, *community care*, occurs 653 times in BNC, and 240,000 times in WEB COLLOCATIONS.



(a) adjective + noun precision-recall curve



(b) noun + noun precision-recall curve



(c) adverb + adjective precision-recall curve

Figure 4.2 Precision-recall curves

**Table 4.18 Precision at various recall values for three measures, Frequency, *t-test* and LLR**

recall	Frequency	<i>t-test</i>	LLR
10%	50.13%	<b>50.52%</b>	50.11%
35%	31.89%	32.48%	<b>32.75%</b>
65%	10.82%	<b>11.01%</b>	10.96%
average	30.95%	<b>31.34%</b>	31.27%

Moreover, noun + noun collocations are less likely than the other two types to be overwhelmed by extremely common words. Our results are confirmed by other research. Krenn and Evert (2001), in a case study on extracting PP-verb collocations found that no statistical measures are significantly better than Frequency. A PP-verb is a verb that requires preposition phrase complements, e.g., *He ventured into the cave*, but not *He ventured*. The results of Wermter and Hahn (2006) on both general collocation and technical term extraction indicate that statistical sophistication does not pay off, compared with a simple frequency measure.

A more pragmatic reason for preferring Frequency is that it allows comparison between collocations having different lengths, such as *make efforts* and *make a difference*—whereas the other methods use different statistical formulas to calculate ranking scores and it is not clear that these produce comparable results. The length of Web collocations varies from two to five words. For all these reasons, we decided to use plain frequency to rank Web collocations.

#### 4.4.4 Quality and quantity of Web collocations

This section compares Web and baseline collocations in term of quantity and quality. Table 4.17 shows that there are far more Web collocations than baseline collocations, ranging from a factor of 20 for adverb + adjective to 120 for noun + noun. But what about quality? Ideally, the WEB COLLOCATIONS collection should cover all baseline collocations. We investigate this below.

The precision-recall curves in Figure 4.2 shows that 96% of the top 100 adverb + adjective baseline collocations are present in WEB COLLOCATIONS, while the

**Table 4.19 Percentage of collocations that do not occur in Web Collocations**

	<b>noun + noun</b>	<b>adverb + adjective</b>	<b>adjective + noun</b>
percentage	4% (480/11846)	9% (1063/11385)	22.6% (14260/62919)

number drops to 86% and 68% for noun + noun and adjective + noun respectively. In fact, the WEB COLLOCATIONS collection does not cover all baseline collocations. The adverb + adjective type achieves an outstanding recall rate in the top 100 due to the relatively small average number of collocations compared to that of other types (175 *vs.* 712 and 870) (Table 4.17). One explanation of a lower recall rate on adjective + noun is that this type is particularly prone to inconsistency of word class assignment, which will be discussed in more detail below.

The percentage of baseline collocations that do not occur in Web Collocations is given in Table 4.19. For example, 480 baseline noun + noun collocations are not covered by Web Collocations, which is 4% of the total number of baseline noun + noun collocations.

Three factors contribute to uncovered collocations:

1. low frequency
2. tagging errors
3. inconsistency of word class assignment.

Recall that Web collocations are extracted from five-grams. If the frequency of a baseline collocation is low, the chance of it occurring in five-grams is correspondingly low. Of uncovered collocations whose Web frequency is less than 1000, an appreciable fraction of them do not occur in five-grams: 72% (344/480) for noun + noun, 23% (240/1060) for adverb + adjective, and 9% (1338/14260) for adjective + noun.

The situation is worst for noun + noun collocations (72%). This could be improved by including 2- and 3-grams in the extraction process, but the restricted context would introduce more tagging errors. Given the small size of this group and their low frequency, this approach was not investigated further.

**Table 4.20 Words whose class is difficult to determine**

*engaged, paid, charming, disapproving, scared, exhausting, degrading, bound, impressed, concerned, flavoured, detached, baffled, loaded, trained, flattering, deprived, inclined, missing, aggrieved, composed, married, muddled, qualified, informed, situated, embarrassed, reassuring, constipated, assured, worrying, united, decayed, restricted, charged, excited, bewildered, patterned, confused, frightened, emaciated, engrossed, alarmed, fixed, opposed, patronizing, handicapped, preoccupied, preoccupied, amazed, contrived, pained, relieved, scattered, embarrassing, encouraging, exposed, organized, deformed, inflated, lacking, pleased, disturbed, startled, educated, exhausted, insulated, disposed, groomed, deserted, submerged, distracted, subdued, surprised, shaken, reserved, tired, neglected, mistaken, determined, puzzled, suited, amused, related, tailored, attached, terrified, shocked, bemused, isolated, charred, settled, orientated*

Factors 2 and 3 above interact, and it is sometimes difficult to tell which is to blame for a particular collocation. Tagging errors are caused both by limitations of the underlying tagger and the restricted context of five-grams, and result in mistakenly assigned word classes and therefore wrongly categorized collocation types (Section 4.3). For example, we could say that OpenNLP wrongly assigns *primarily engaged* to the adverb + verb type because in the baseline collocations it is tagged as adverb + adjective.

However, determining the classes of words like *engaged* is difficult, and contentious even for linguists. Other words for which this is the case are shown in Table 4.20. Unlike adjectives like *beautiful* and *happy*, these words can also be used as gerunds (e.g., *degrading*) or past participle verbs (e.g., *engaged*). However, not all gerunds and past participles are adjectives. Leech and Svartvik (1975) distinguish between adjectives and participles according to whether they can be modified by the adverb *very*—in which case the former is clearly an adjective. To maintain a reasonable degree of consistency when tagging the BNC, certain semantic criteria are used to differentiate adjectives and participles, and adjectives and nouns. (These criteria are specified in BNC’s *Word Class Tagging guideline*.<sup>17</sup>) For example, Leech et al. (1994) point out that there is no universal standard to determine the appropriate tag for *washing* in *washing machine*—noun, verb or adjective. They call for a tagging standard that can be used to determine what is an appropriate tag in a given context, and argue that “only if this [standard]

---

<sup>17</sup><http://www.natcorp.ox.ac.uk/docs/bnc2guide.htm>

is specified independently by an annotation scheme, can we feel confident in judging whether the tagger is ‘correct’ or ‘incorrect’.”

The words listed in Table 4.20 are tagged as adjectives in the baseline collocations, but as gerunds or past participles by OpenNLP. Consequently, their collocations could be categorized as adverb + adjective or adverb + verb depending on the context. In this experiment, they are identified as verbs in the WEB COLLOCATIONS collection, either gerund or past participle, resulting in 78% (831/1063) uncovered adverb + verb collocations.

The relatively low coverage of adjective + noun collocations (22.6% uncovered) is attributed to inconsistency of word class assignment between OCDSE and OpenNLP. For example, *car import*, *coal import*, *energy import*, *food import*, and *oil import* are adjective + noun collocations in the baseline collocations because *car*, *coal*, *energy*, *food*, and *oil* are treated as adjectives. However, they are noun + noun collocations according to OpenNLP because it classes these words as nouns. Out of 14,260 uncovered noun + noun collocations, 86.7% (12357) include such words.

In conclusion, the three factors discussed above affect the quality and quantity of WEB COLLOCATIONS. Some are more dominant than others, depending on the collocation type. However, WEB COLLOCATIONS contain most of the OCDSE collocations. Low frequency and tagging errors could be overcome if pre-tagged Web *n*-grams were available. Inconsistency of word class assignment between different collocation resources is a difficult problem that has no easy solution, and users need to be advised of this issue. In particular, the interface to the collocation system should help them by suggesting that they consult other collocation types where appropriate.



## 5. Evaluating collocation resources with language learners

---

Chapter 3 explored the use of Web text as a resource for collocation learning, and described three collections that were created to serve that purpose. With the WEB PRONOUN PHRASES collection, learners explore word sequences associated with pronouns: ones starting with the word *I* appear to be particularly productive. With the WEB COLLOCATIONS collection, learners study collocations organized by syntactic pattern. With the WEB PHRASES collection, learners check word sequences against general usage on the Web. In order to provide a realistic context of use, we recruited language learners who were attending an English language programme at Waikato Pathways College, which prepares international students for university study, for evaluating these collections.

The study focuses on the use of the three collections to support writing tasks. The strength of corpus-based activities is that they can provide students with rich lexico-grammatical information, which is very important in L2 writing. For writing, learners need information not only about vocabulary and grammatical forms, but also about multi-word sequences such as collocations, synonyms, idioms, syntactic patterns and lexical phrases. Several researchers have documented evidence of the challenges faced by relatively proficient second language learners in their use of formulaic sequences in a way that is both authentic and native speaker-like (Farghal and Obiedat, 1995; Howarth, 1998). It is this very aspect of L2 writing (Yoon and Hirvela, 2004) that has been exploited in the present study to support writing.

Two evaluations were conducted. The first involved twelve participants in a general intermediate language class. They were asked to write short descriptions of themselves and their family in order to elicit personal pronoun use. In the second, eight students from an IELTS<sup>18</sup> writing preparation class participated. Each wrote an essay and then used the WEB PHRASES and WEB COLLOCATIONS

---

<sup>18</sup> International English Language Testing System: <http://www.ielts.org/>

collections to correct errors highlighted by teachers. Use of the system was recorded in detail, and the search and retrieval data was analyzed alongside the texts the students wrote. The study tracks the way in which students formulated search queries and how they made use of the search results in the texts they wrote, and investigates the impact of the use of CLS on their writing, identifying its strengths and limitations.

### *5.1 The WEB PRONOUN PHRASES collection*

How useful is the WEB PRONOUN PHRASES collection for supporting writing in the context of self-expression?

#### 5.1.1 Participants and procedure

Twelve language students participated, six females and six males aged from 19 to 40 years. They were native speakers of six different languages (Korean, Argentinean, Colombian, Chinese, Dutch and Japanese). Their ability in grammar, reading, speaking, and writing had been assessed by the college. Grammar and reading were tested by the Oxford entry test, which yields two scores for each skill. Writing and speaking were tested by a writing task and interviews with teachers, who gave scores for each. The four scores were combined in order to allocate students to different classes. All participants were from the same intermediate class. However, as their teacher observed, their abilities varied greatly—for example, some excelled in speaking, but performed poorly in writing and vice versa. Despite our efforts to ensure uniformity, the evaluation still included participants who had a wide range of writing ability. To compare their ability before and after using the system they were asked to write a 150–200 word description of themselves the day before the evaluation.

The evaluation was conducted during a 2-hour session in a computer lab at the University of Waikato. In the first half hour, it was explained how the WEB PRONOUN PHRASES were gathered and what the system does. Then students were asked to prepare a personal profile of themselves for a home-stay family, including their background, family, interests, likes and dislikes, and any other things that they thought would make them seem interesting.

They wrote on paper, and in order to track changes they were instructed not to erase errors, but to cross them out or rewrite above the text. They were encouraged to bring dictionaries and use them, because the system did not check spelling. They were asked to circle any text fragments that the system had helped them generate or improve. Finally, they were allowed to seek help from their teacher and the researcher at any time regarding how to use the system, and for any other queries they had about their texts.

Each student was given an anonymous identifier, and their use of the system was recorded in detail and written to a log file. The log data included:

- the search terms entered
- the pronoun phrases used in the search, one of *I, we, they, she, he, or it* phrases
- synonyms or antonyms, related words, associated words that were looked up
- the retrieved samples, whether from the Web or the BNC.

Data was recorded sequentially, with a timestamp to make it easy to trace the sequence of each student's work and to make a connection between search results and use in their texts.

### 5.1.2 Results

Students using the system adopted one of two strategies. Most finished their writing first and then used it to check text they were uncertain of. Some (three) used the system to help generate text by finding the correct usage of a word and suggesting suitable sentence structures. The students produced fairly short texts, averaging 20 sentences per essay and 11 words per sentence. Grammatical errors, incorrect sentence structures, and incomplete sentences were scattered throughout their work. Because of the constraints of the topic—themselves and their family—and their limited language ability, their writing exhibited a narrow range of vocabulary and few idiomatic expressions. For example, the four most common words used were *like, come, want* and *live*. Sentence structure was simple and basic. Most sentences began with a pronoun, followed by the main verb and a noun or prepositional phrase. Feelings and emotions were expressed in a rather

plain way; linguistic boosters or hedges were rarely used. Although they were encouraged to write about their family, most just described themselves.

Little is known about how students make use of corpus-based resources like the WEB PRONOUN PHRASES collection, so this section first looks at the search strategies the students adopted. These have a significant impact on how much they were able to benefit from CLS. We focus on search term selection and refinement, and the use of the search results.

One obstacle for students to make effective use of the system is to find the right word to start with. For example, to express their likes or dislikes, they tend to choose simple and direct words such as *like*, *love*, *hate*, *hobby*, *movie*, *sport*, or *travel*, while more advanced students may use *enjoy*, *favorite*, *desire*, etc. On failing to retrieve what they want, students adopted four approaches to refine search terms:

1. change the word form—use plurals, other forms of a verb, or adverbs instead of adjectives,
2. explore lexical resources for synonyms, related or associated words to find alternatives,
3. use dictionaries or ask the teacher, and
4. simply give up and move on to the next section of text.

With respect to the use of the search results, some students always examined the phrases retrieved in the sample text before using them, while others barely looked at this functionality. Most students made direct use of what they had received, resulting in text that might be either appropriate or inappropriate. Some search results were modified before being incorporated into the text, including changing the word form, for example from *want* to *wanted*, or substituting one word for another, from *I really enjoy this movie* to *I really enjoy this sport*. Finally, in some cases, no apparent use was made of the search result.

Table 5.1 summarizes the log data. For each of the 12 students it shows the number of sentences in their text, the number of searches they launched, the number of times sample text on the Web or the BNC was viewed, and the number of lexical resources, i.e., synonyms and related words, that were viewed. The last

two columns give the positive or negative changes the students made to the text when using the system.

A total of 267 searches were conducted, 95% of which were for *I*-phrases (phrases that begin with the word “I”), ranging from 8 to 45 per student with an average of 22. Most searches used content words as queries to find phrases containing relevant words. For example, participants would search for *student*, *study* and *university* to describe their student status, or *like*, *love* and *hobby* to talk about their personal interests. In a few cases, students searched for function words such as *been*, *will*, *why*, *when*, *again*, *also*, *for*, *with*. It is not clear whether they were trying to learn the usage of these words, or use them to find phrases related to time or explanations, because these searches resulted in few follow-up activities such as looking up samples or use in the text.

Students evidently used CLS actively, for searches outnumbered the sentences generated. Except for the first student, the number of searches correlates well with the amount of text produced, and also, with rare exceptions, with the number of look-ups on the Web or the BNC. It is encouraging to see that the students tried to understand samples in context before using them. Surprisingly, most samples viewed came from the Web rather than the BNC—perhaps because the latter snippets tend to be lengthy paragraphs, and students were under time pressure to finish their essay. We found no instances of students using retrieved information unsuccessfully if they had extensively consulted contextual resources.

The evidence of the number of times lexical resources were consulted—in most cases five or fewer—paints a different picture. The logs reveal unexpected searches for words such as *and* and *will*, which suggests that some students did not understand the nature of these resources. However, students 6 and 7, whose writing skills were the best amongst all participants, used them extensively. This indicates that more advanced learners are more likely to explore alternative language usage.

**Table 5.1 Summary of the log data**

	sentences	searching	samples (web or BNC)	lexical resources (synonyms/collocations)	positive uses	negative uses
1	40	14	6	3	3	0
2	29	32	32	5	7	0
3	26	15	0	0	5	1
4	25	32	24	2	8	2
5	21	29	24	5	8	2
6	19	39	9	25	3	0
7	18	45	29	20	12	2
8	16	14	19	1	6	0
9	15	12	5	2	4	0
10	9	13	13	12	4	1
11	9	8	5	0	2	0
12	8	14	12	3	3	0
<i>total</i>	235	267	178	78	65	8

What impact did CLS have on the students' work in terms of text generation and revision? Their text was inspected manually and 73 uses are identified. A "use" is identified based on:

1. the student indicated use of the system by circling the text,
2. there was no evidence of such language usage in the text the student produced the previous day, and
3. log data confirmed that the altered text was suggested by the system.

The first criterion provides strong evidence of use, but in many cases students forgot to circle the text and consequently the second criterion was used as well. (For the second criterion, recall that students were asked to write two pieces of text: the first without using the system and the second during the evaluation the following day.) Here it is important to differentiate errors from mistakes. Students make language *errors* when they appeared to have little, or no knowledge, of the relevant linguistic feature—for example, one wrote *we want do something*, rather than *we want to do something*, because he did not know the correct usage of the verb *want*. Students make language *mistakes* when they write the wrong thing despite knowing the rules: in this case they are capable of recognizing the mistake and fixing it themselves. Mistakes were discarded if there was evidence of correct use elsewhere in the text.

It is important to note that a “use” of the system does not necessarily guarantee that the result is correct. The student might misinterpret the samples the system provides, resulting in a negative use. For example, one student changed *I like eat Taiwan’s snack* to *I would like to eat Taiwan’s snack* after searching for the word *like*. Unfortunately, in the original context the first version, although grammatically incorrect, is nevertheless more appropriate. This negative use is attributed to the student misunderstanding the meaning of the modal form *I would like to*. Moreover, *I would like to* is the dominant usage of the verb *like* and therefore accounts for most of the search results, which confused that student.

A positive use is a correct use of the search result in a text, leading to correct grammar, better sentence structure, and idiomatic, natural expressions such as *it would be better to*, *I enjoyed it a lot*, and *I wish I could*.

There were 65 positive and 8 negative uses, which means that every 3½ searches resulted in a use, 90% of which were positive. Most negative uses were due to inadequate pragmatic knowledge of a language expression, for example, the difference between *my friend was performing* and *my friend was going to perform*, or *I was singing* and *I have been singing*.

Now let us examine what students used CLS for. Uses are grouped into four categories:

1. checking grammar
2. generating text
3. expanding text
4. confirming text.

Table 5.2 gives samples extracted from student text for each category. In the first, students used the system to help correct grammar errors, find the right preposition, correct verb forms, and use conjunctions correctly. CLS provides a wealth of examples of usage of common verbs such as *go*, *want*, *continue* and *live*, which resulted in many corrections. One student even changed *I’ve been in NZ since four month ago* to *I’ve been in NZ since April* on searching for *been*.

In the second category, some students constructed sentences based on samples they found in the collection. They either used them directly or modified them to

suit their need. For example, the sentence *I enjoyed spending time with my close friend* stemmed from the *I*-gram *I enjoyed spending time with*. In one particular text there were seven idiomatic expressions such as *I wish I could*, *I think it is important to*, and *is very good at*. The original version of this text was mostly made up of simply structured sentences and showed no evidence that the student knew these expressions. This student told us that she could write a text in different ways by using the phrases found in the system.

Some students found it difficult to make their writing interesting and colorful because of their limited stock of vocabulary and idiomatic expressions. In the third category, many students made efforts to expand the text using samples provided by CLS. A common strategy involved the use of language boosters and hedges, including adverbials such as *very*, *really*, *so much*, *a lot*; expressions such as *I thought it would be better*; and collocations such as *born and raised*, and *absolutely beautiful*.

The fourth category is use of CLS to confirm text that has been written. A student's original text may show that they know the language features in question, but they may nevertheless consult the system for confirmation. For example, one student searched for the word *best*, and then checked the sample *I did my best to*—despite the fact that he had already used it correctly.

## **Discussion**

This evaluation suggests that the WEB PRONOUN PHRASES collection is a valuable resource for language learning, particularly in helping students to express themselves in richer and more native-like ways. While the variety of search strategies used may be in part due to unfamiliarity with the system—a factor shared by all participants—there also appeared to be individual differences that may be explained by different levels of proficiency. Vocabulary size has a significant impact on the extent to which students can make good use of CLS, because they must know the word before they can use the system, but often have only a vague idea of what they are seeking. The most successful students tried out 46 unique words as compared to the average of 18. The results also show that proficient learners can use the collection to generate text as well as revise it, but

**Table 5.2 Samples extracted from student text**

<b>category</b>	<b>original</b>	<b>new</b>
checking grammar	<i>I was born Seoul I went performance hall for singing I've been in NZ since four month ago I want find a good job</i>	<i>I was born in Seoul I went to performance hall for singing I've been in NZ since April I want to find a good job</i>
generating text		<i>I graduate from the music school My sister is very good at cooking I wish I could become a social worker I have developed interest in movies I think it is important to learn English I can travel all over the world</i>
expanding text	<i>I am close to them It is a beautiful place It is hard to speak English I thought to find another home-stay</i>	<i>I was born and raised in Taiwan I am very close to them It is an absolutely beautiful place It is really hard to speak English I thought it would be better to find another home stay</i>
confirming text	<i>I did my best to study English I cannot afford to lose more time</i>	

the limited vocabulary knowledge of less proficient learners restricts them to revisions. However, most student text demonstrated positive effects at the lexical, grammatical and perhaps most saliently the pragmatic level.

## 5.2 *The WEB PHRASES and WEB COLLOCATIONS collections*

Two types of evaluation were conducted to assess the utility and effectiveness of the WEB PHRASES and WEB COLLOCATIONS collections, and the way in which they can be used to improve text by generating useful language examples. First, to discover the potential to offer correct, appropriate and accessible alternatives, we used CLS to resolve errors in student writing. Then we asked students to use it in conjunction with a user guide, so that we could evaluate the use they made of CLS and how it affected their textual revisions.

### 5.2.1 Designing a user guide

A user guide was designed based on samples of student text included as exemplars in the IELTS *Specimen Materials Handbook* (IELTS, 1997). We created five kinds of exercise by analyzing typical errors that students make, and

relating them to the possibilities that CLS can help resolve. Appendix C gives the full guide. Here we provide a brief description.

First, CLS can be used for essay preparation. Given a topic, say *nuclear power*, students can find appropriate vocabulary in two ways. They can collect useful noun + noun, adjective + noun or noun + *of* + noun phrases using topic-related keywords like *nuclear*, *weapons*, *energy*, *benefits*, *threat*, *disadvantages*, and *solutions*. They can also learn what verbs are commonly associated with those words, and their correct usage. For example, in English, we say *pose a threat*, not *give threat*; *the benefits outweigh the disadvantages*, not *we outweigh the benefits and disadvantage*; *find solutions*, not *examine about the solutions*.

Second, learners tend to reuse particular words repeatedly throughout their essays. A typical example is overuse of the verb *rise* or *decline* in the IELTS task that asks for a description of changes and trends in an input text, graph, table or diagram. Examining collocations of words like *shares* or *prices* will quickly yield alternatives such as *jump*, *soar* and *surge*; or *drop*, *fall*, *slump*, *slip* and *plunge*.

Third, learners often misunderstand the usage of a word, and overgeneralize common words like *have*, *do*, *make*, *take*, and *give*. As a result, odd word combinations or idiosyncratic word choices are scattered throughout their writing. Examples are: *cultivate their children with*, *reinforce the income*, *deep interests*, *give threat*, *the city must have another solutions*. The WEB COLLOCATIONS collection can help learners make more accurate or appropriate choices of words and word sequences. For example, students look up the nouns that follow *cultivate*, or find verbs that are commonly associated with *solutions*.

Fourth, learners also find it difficult to boost or hedge statements by adding adverbs. Suppose one wants to add extra strength to the sentence *We will all benefit from it*. Searching *benefit \* from* in the WEB PHRASES collection yields *greatly*, *directly*, *significantly*, *enormously* and *immensely*. Or consider how adverbs are used to describe feelings appropriately and precisely. If one wishes to express disappointment, the WEB PHRASES collection provides a wide range of modifiers, from *extremely*, *deeply*, *bitterly*, *pretty*, *quite* to *rather*, *somewhat*, *just*, *slightly*.

Finally, exercises were designed to demonstrate how to use CLS to correct grammatical errors. Misused prepositions and ill-formed verbs were two dominant grammatical errors in the sample text: for example, *The government must be responsible of their welfare*, *They have increased day to day* and *this problem would resolve a little*. Those errors can be corrected by searching WEB PHRASES for *must be responsible*, *increased day \* day* and *this problem would*.

### 5.2.2 Participants and procedure

We worked with teachers in our institution's language support centre to recruit participants. The study targeted students who were involved in the IELTS writing preparation class. Nine students, three females and six males, from 18 to 30 years old and native speakers of five different languages (Chinese, Japanese, Arabic, Koran, and Chilean) participated in the evaluation.

During the first session, the students were given an IELTS argument writing task<sup>19</sup> selected by their teacher as part of their normal class programme. They were asked to write a response to the task within the usual 40 minute time allocation. However, contrary to normal practice, they were asked not to use dictionaries.

After this, an experienced teacher and we both examined the students' writing, highlighting aspects of the texts that we felt needed improvement and revision. It should be noted that while these were labeled as 'errors,' in many cases they are examples of not quite acceptable words or word sequences. While these seem to be vague criteria, as guidance, two areas were suggested for focus:

1. grammatical errors, e.g., incorrect use of verb forms and prepositions, misused plurals and articles, and missing verbs
2. lexical errors, e.g., wrong or inappropriate word combinations, particularly those involving noun + verb, verb + noun, adjective + noun and noun + noun combinations.

---

<sup>19</sup> The task was: Historical art has more cultural value than modern art. Discuss both sides of this argument and give your opinion.

Consistent with the approach taken by Chambers and O'Sullivan (2004), the text was highlighted at the phrase level. For example, in the student's text below the brackets [ ] indicate phrases identified as needing to be revised.

*Some famous museums have become [one the most powerful attractions] to [reinforce the income] for a particular country.*

The teacher and we met to compare marked sections of text. When agreement was reached, additional marking was added, where appropriate, to help students focus on particular parts of the highlighted phrases. For example, in the following text the words *powerful* and *reinforce* were underlined to assist students in searching for collocations, and the symbol ^ was inserted to indicate a missing element.

*Some famous museums have become [one ^ the most] [powerful attractions] to [reinforce the income] for a particular country.*

A second two-hour session, began with an initial 30 minutes in which students received a more detailed explanation of CLS. We demonstrated how to search for phrases and collocations, and look up examples from the BNC and Web using the material in the user guide. Because of time constraints, one error was randomly picked from their text and used to show how to correct it with the help of the system. Finally, the texts with errors highlighted were returned to the students, who then revised their text on their own, focusing particularly on the marked-up sequences. Help related to how to use the system was provided by the teacher and researcher. The student's actions were logged automatically for later analysis. A third session was available for students who needed more time to complete their revisions.

### 5.2.3 How students used CLS

This section looks at how students used CLS, including how they formulated query terms and made use of the search result. The log data demonstrated active use of the system for checking marked errors, with five queries per error on average. Most focused on correcting errors by replacing the highlighted words with alternatives found using the system. Students gave up on unresolved errors after a few unsuccessful attempts and moved on to the next. Two students chose to rewrite the text, using the system to help generate new phrases.

With two exceptions, students tended to consult the WEB PHRASES collection more frequently than WEB COLLOCATIONS, which may be attributed to the former's relatively straightforward interface. When looking up collocations, students often typed in more than one word, and sometimes even included prepositions. It seemed that they did not understand the structure of this collection and what it can offer. However, there was intensive and effective use from two students who issued four times more collocation queries than the average. They made several mistakes at the beginning, but became more comfortable after a few trials.

When formulating query terms, most students used the words in marked phrases as clues. For example, given *reinforce the income* and *powerful attractions*, the phrases preceding *income* and *attractions* were sought. This approach can be effective only if one part of the phrase is wrong. Some students chose incorrect search terms even if they were highlighted, for example using *fancy* for *fancy and good position*. In some cases, formulating queries could be challenging. For *they can be comparing with wine*, one student tried *comparing* and *comparison*, and then gave up. Using *be \* with* generates *be used with*, *be associated with*, *be dealt with* and so on, which may help students induce the right answer. However, they need to be trained to use this approach. When <sup>^</sup> is indicated in marked phrases, most students used \* in queries. Finally, some students used the whole marked phrased as a query (maybe they were expecting the system to correct them automatically), then removed words one at a time if no satisfactory results were found.

In the case of more than one alternative word or phrase being given, how did students make their choices? Advanced students chose more precise words, while others tended to use the more frequent ones, despite the fact that the less frequent ones may make a better text. On the other hand, some students clearly knew about using frequency as a clue. For example, working on *in the other hand*, one student searched for phrases preceding *other hand*, the system yielded *the other hand* (12,000,000 times), and *on other hand* (47,000 times). He checked out both alternatives, and finally used *on the other hand* (6,800,000 times), which is the top hit for *the other hand*. Some students were confused when the marked phrase

appeared in the search result. For example, one student did not make any changes when he discovered that both *point to* and *point out* are common phrases, although the latter was more appropriate according to the original text, and their correct usage was suggested in samples from the Web and BNC. Finally, it is disappointing that only three students looked up examples of the search results before using them in their text, which may be due to both time constraints and limited training.

#### 5.2.4 Assessing CLS's potential

This section looks at the changes the students made to their text. First, we used CLS to check the errors ourselves with the aim of establishing baseline data. The evaluation was conducted by myself—a second language learner.

The errors were classified into six types of structure: noun phrase, verb + noun, noun + verb, prepositional phrase, phrasal verb or verb + preposition, and verb + complement. Another large group of errors were classified as grammatical errors because they involved morpheme omission or error. Table 5.3 summarizes the counts of these errors and gives examples of acceptable alternatives generated by the system. Appendix D gives the full results.

In total, 108 errors of all types were identified across the texts. CLS was able to generate correct and appropriate alternatives for 95 (88%) of the cases. Focusing specifically on lexical non-grammatical errors, the success rate is higher, with 82 corrections in 88 errors (94%). Errors associated with noun phrases (adjective + noun, and noun + *of* + noun), together with errors in the verb + noun pattern, were the most frequent (63 errors). Combining sequences involving preposition use—preposition phrases, phrasal verbs and verb + preposition—there were 15 errors, a smaller but still substantial number, of which only two were not resolved. Grammatical errors represent a large group (20), but in contrast to the success of the system with lexical errors, relatively few grammatical errors were resolved.

**Table 5.3 System generated alternatives to errors**

	counts			examples	
	total	resolved	unresolved	student text	system generated alternatives
Noun phrase	36	34	2	<i>contemporary arts building</i>	<i>contemporary art gallery</i>
				<i>a fancy and good position</i>	<i>a unique position</i>
				<i>the most important steps of our evolution</i>	<i>stages of evolution</i>
				<i>a element of a national spirit</i>	<i>an expression of national spirit</i>
				<i>important events in their times</i>	<i>events of that time</i>
Verb + noun	27	25	2	<i>reinforce the income</i>	<i>increase the income</i>
Noun + verb	3	3	0	<i>the essay favour</i>	<i>I favour</i>
				<i>the profound influence created by</i>	<i>the profound influence exerted by</i>
Preposition phrase	8	7	1	<i>in the other hand</i>	<i>on the other hand</i>
Phrasal verb; verb + preposition	7	6	1	<i>play an important role on</i>	<i>play an important role in</i>
Grammatical errors	20	13	7	<i>more likely to be preserve</i>	<i>more likely to be preserved</i>
Verb + complement	4	4	0	<i>the argument may be true</i>	<i>the argument may be valid</i>
Adverb use	3	3	0	<i>are aware of a lot</i>	<i>are fully aware of</i>
total	108	95	13		

**Table 5.4 Student changes to errors identified in their texts**

	total	successful changes	successful changes	success rate (%)
Noun phrase	36	27	9	75
Verb + noun	27	16	11	59
Noun + verb	3	2	1	67
Preposition phrase	8	5	3	62
Phrasal verb; verb + preposition	7	6	1	85
Grammatical errors	20	11	9	55
Verb + Complement	4	3	1	75
Adverb use	3	2	1	67
Total	108	73	35	67

Table 5.4 indicates how the students used CLS. It shows the number of sequences that were marked up; the number of successful changes that the students made; the number of unsuccessful changes that led to anomalous and grammatically incorrect text; and the success rates. For instance, in the case of errors associated with noun phrases of the adjective + noun and noun + *of* + noun forms, 36 sequences were marked up, of which the students changed 27 (75%) successfully and 9 (25%) unsuccessfully. The high success rate indicates the willingness and ability of students to use CLS to revise their work.

Adjective + noun and noun + *of* + noun both showed a consistent and relatively high success rate. In most cases, students used the correct main noun, but picked inappropriate adjectives and modifying nouns, resulting in strange combinations—for example, *main culture value*, *powerful attractions*, *classical artifacts*, *numerous of countries*, *a great deal of museum*, *these sort of arts*, and *popularity of modern technology*. Students obtained good results on this kind of error, but the success rate declined when both parts were wrong. As an encouraging example, one student changed *modern art's appearing to the development of modern art*.

Using the wrong verbs accounted for the majority of verb + noun errors. The students (1) chose verbs that do not go with the following noun, e.g., *save the history*, *afford citizens more entertainment*, and *balance their consciousness*, (2) overgeneralized common verbs, e.g., *have an assumption*, and (3) chose imprecise

verbs, e.g., *know clearly about their culture*. One student tried alternative nouns, changing *spend their tour* to *spend their holiday* instead of *take a tour*. The relatively low percentage of correct changes (59%) indicates that verb + noun is challenging—changing the verb may alter the meaning of the whole sentence. For example, one student changed *paid an attention* to *draw an attention* without changing the rest of the text accordingly. CLS can give the verbs that are most frequently associated with a particular noun, but it is up to the student to pick an appropriate one. Some students chose ones that they were most familiar with regardless of context, which was not necessarily the best choice. Sometimes they chose one that made a good verb + noun combination but did not fit the context.

Students performed well (85%) on the verb + preposition category, probably owing to, the many useful examples that the system provides. For instance, they changed *play an important role on* to *play an important role in*, *give priority for* to *give priority to*, and *is famous with* to *is famous for*.

The result of successful changes in the grammatical errors is largely consistent with the success rate in other categories, though slightly lower at 55%. Students made five kinds of error:

1. wrong verb form: *is influence by*, *are deserve*, and *arts are comparing*
2. missing article: *contain wide range of*
3. missing auxiliary verb: *people who interested in*
4. misused plural and singular: *many century ago*
5. misspelled sentence adverb: *now a day*, and *in the mean while*.

It is not straightforward to use the system to correct verb form errors. Take *is influence by* as an example. The query *is \* by* gives a list of past particle verbs between *is* and *by*, but *is influenced by* is not among the 100 top hits. Students need to figure out by themselves that the past participle of *influence* should be used instead of the base form. The student who made this error tried *is influence by*, *influence by*, *influence \* by*, and then gave up. Errors related to missing articles and auxiliary verbs, and misused plurals and singulars need to be marked explicitly—for example, *contain ^ wide range of*, *people who ^ interested in*,

*many century ago*—to help students produce a correct query. Errors of the last kind are difficult to fix because the adverbs were misspelled.

For the other categories, there is too little data to give a sense of the pattern of changes. However, some changes were successful. For example, in the category verb + complement, *society has become more increasing fascinating* was changed to *society has become more accepting*; and *has made the society become more valuable* was changed to *has made the society become more open and liberal*. In the adverb category, changing *modern people strongly claim that* to *modern people legitimately claim that* indicates the potential of CLS to provide students with native-speaker-like examples.

Of the 108 marked sequences, we could only identify 95 changes, whether successful or unsuccessful, that the students had made to their text. The remaining 13 marked sequences were not used in the revised version—in other words, they were abandoned. This represents a type of avoidance strategy. It happened in particular with one student, who discarded the seven sequences and rewrote substantially different text from her draft. The log data showed that the students actually did some work on all 13 sequences, but gave up after a few unsuccessful attempts. Sequences that were removed were treated as unsuccessful changes, although sometimes they improved the text. In total, the student success rate was 67% (73/108); 70.5% if grammatical errors are excluded. Compared with our assessment of what is possible using CLS, the students achieved a 77% (73/95) success rate on their own. This provides a strong indication of their willingness and ability to use it for revising their text.

### 5.2.5 Discussion

One of the major limitations of this study is the time allocated to the evaluation. An in-depth study to capture the perceptions and strategies of students while using CLS is clearly needed. Nonetheless, we can make the following observations. When use of CLS resulted in a modification to the text, the alteration was most often an improvement, although some local changes did not necessarily produce better text overall. The system certainly has potential for helping students make correct and more appropriate word choices, and thereby generate better and more

native-like word sequences or collocations. The frequency-based phrases that it provides help students focus on the actual usage of particular words, including nuances that are generally left unarticulated in language teaching.

As we worked through the student's writing, we noticed the low volume of noun phrases. In particular, occurrences of noun+ *of* + noun were limited to quantification words such as *number*, *a great deal* and *lot*. In fact, this particular phrase type is prominent in academic writing, and we believe CLS will help students improve collocation knowledge in this respect.



## 6. Constructing a collocation learning system

---

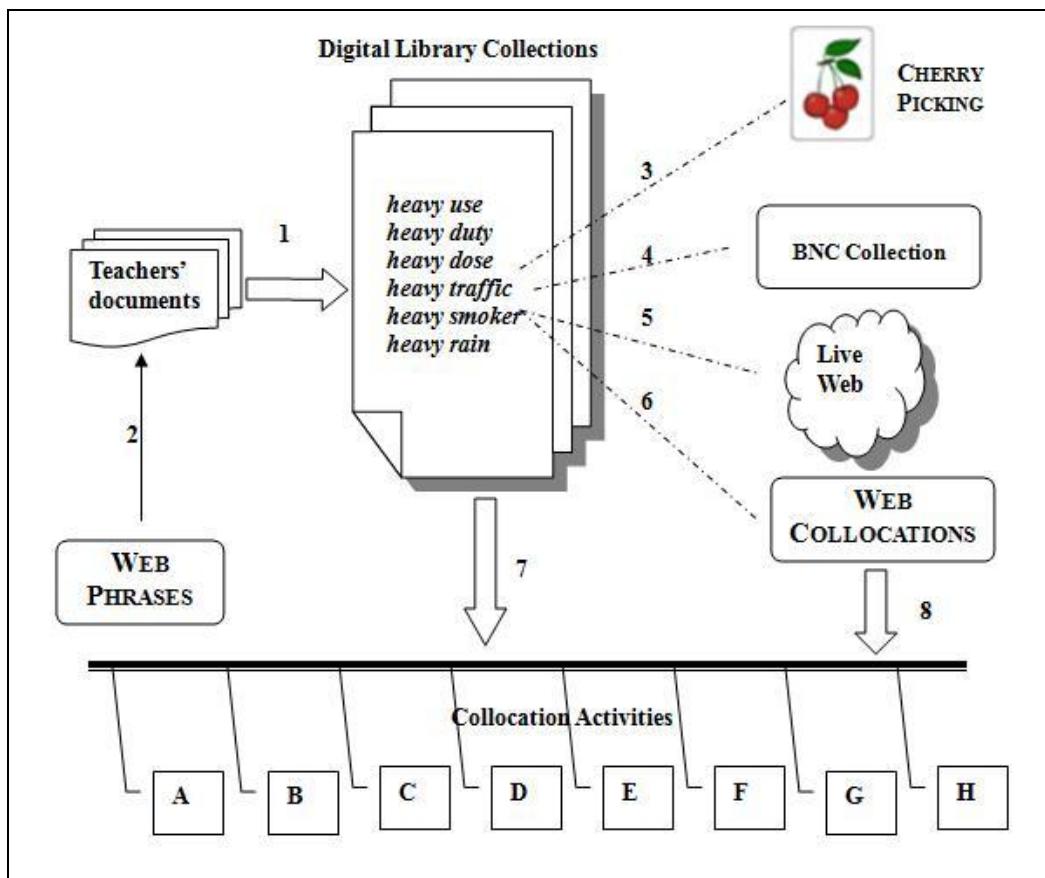
Extensive knowledge of collocations is a key factor that distinguishes learners from fluent native speakers. The sheer number of collocations that learners need to learn demands three things:

- constant language exposure,
- study of the most common and most important collocations, or ones for special purposes such as business, sports, news, and
- effective learning strategies.

The first requires a language environment in which learners can meet salient patterns repeatedly in naturally-occurring contexts. The second implies the careful selection of patterns of high priority and greatest relevance for learners from authentic text produced by people in actual communication situations. The third demands an organized and systematic study in a pedagogically enhanced environment.

This chapter describes the CLS collocation learning platform, which is outlined in Figure 6.1. The design is guided by collocation teaching strategies of *noticing*, *retrieval* and *generation* developed by teachers and researchers and summarized by Nation (2001) (Section 2.4.1). Articles such as those that teachers have prepared for their students are built into a digital library collection (1) and augmented with automatically identified collocations that are filtered using Web frequency drawn from the WEB PHRASES collection (2). While reading the articles, the learner's attention is attracted to highlighted collocations in context, and they study and collect collocations (3).

Learners expand and enrich their knowledge by examining related items retrieved from the WEB COLLOCATIONS collection (6), and by studying exemplary text in the British National Corpus (4) and live samples from the Web (5).



**Figure 6.1** Collocation learning platform in CLS

We have developed eight collocation activity types (A-H) that allow learners to practise collocations of newly learnt or partially known words, and convince them that learning collocations is a powerful way to improve their fluency and accuracy. For each one, teachers can generate an unlimited numbers of exercises, tailored to their classes, from the content of the collections they have built (7) or the WEB COLLOCATIONS collection (8), using a specially created interactive exercise design interface. Some activities are game-like, to help learners maintain high motivation; others mimic traditional collocation learning activities that teachers have developed for classroom use.

### 6.1 Supporting collocation learning

This section sketches how CLS supports collocation learning by allowing teachers (or learners) to build a collection of readings that are relevant to their study. It automatically extracts important collocations from readings and presents them

alongside the text. Students read the material to gain a degree of familiarity with particular collocations, study them in different contexts, and record ones that interest them. Then they undertake various learning activities based on the same material, presented in the form of exercises. Psychological conditions are incorporated into the system design to help learners notice important collocations, develop language sensitivity, and transfer from short- to long-term memory.

### 6.1.1 Creating learning material

Instead of using corpora available on the Web, teachers can use their own material to build CLS collections. CLS organizes and presents this in a way that helps students pay attention to the wealth and density of collocations. Lewis (2000) suggests that teachers should choose the right kind of text for their students because different genres exhibit different collocational characteristics. He emphasizes the importance of selecting materials that are suitable for particular groups of students, and for particular purposes. For example, subject-specific collections give the opportunity to encounter texts that exhibit particular patterns of both word choice and grammar. For example, student knowledge of business language is greatly enriched by basing learning on a corpus of business reports and product reviews (Fuentes, 2003).

To avoid overwhelming students, teachers control collection size simply by importing the right amount of material into the system. Material can come from conventional sources such as textbooks, newspapers, the Internet, and teachers themselves. Teachers can also associate a language level with a particular text and, later on, direct each student's attention to the level suitable for their ability.

For this thesis, we have built and evaluated three collections using three kinds of text: general reading articles, academic English, and abstracts extracted from doctoral theses. These collections and their different collocational features will be discussed in Chapter 7.

### 6.1.2 Facilitating noticing, retrieval and generation

Section 2.4.1 described classroom strategies that teachers adopt when helping students build up collocation knowledge: awareness-raising, deliberating learning,

and recording and recycling. These strategies are in line with three general psychological processes summarized by Nation (2001) that lead to words being remembered. We assume that learning collocations requires the same processes as learning words does because collocations should be learnt as a single unit rather than putting individual words together. Considering all these, CLS is designed to facilitate the process of noticing, retrieval and generation.

No noticing, no learning. The first process is to encourage learners to pay an attention to an item as part of the language rather than as part of a message. Nation suggests that noticing occurs when students deliberately study a word by looking it up in a dictionary, guessing its meaning from context, and negotiating its meaning with peers or teachers. Or the teacher highlights a word on the blackboard, and gives its definition, synonyms, or translation into the first language. Noticing is also affected by other factors such as salience and usefulness of the item, and the learner's interest and motivation.

The second process, retrieval, helps students retain a word in memory so that its form and meaning can be retrieved when they meet it while listening or reading (receptive knowledge) or use it in speaking or writing (productive knowledge). Meeting a word several times and at frequent intervals is an effective way to help students strengthen their memory of it (Nation, 2001). Activities that facilitate repetition include reading the same text several times, and doing follow-up exercises that force students to reuse what they have learnt.

The third process, generation, helps students meet or use a word in different forms or contexts. For example, the teacher provides different sentence samples or a range of collocations associated with that word, or asks students to use it in a new sentence context, or brainstorm collocations themselves. Moreover, the teacher encourages and trains students to use concordancers to study the word in real language.

CLS supports these three processes. It automatically extracts collocations that follow the syntactic patterns given in Table 3.6 from text provided by teachers and highlights them in the original context. Teachers control which patterns to focus on, because some might be of particular interests to particular groups of

students—for example, adjective + noun, noun + noun and noun + *of* + noun collocations for students doing university study. Students read the original article and the article with collocations highlighted in a separate interface. Searching and browsing facilities allow students to access extracted collocations by the words they contain or by their collocation type.

CLS employs two ways to help learners remember a collocation: repetition and use. Learning activities that it provides allow teachers (or students) to create exercises using the same material that students read, to gradually increase familiarity with its collocations. Typical word usage and salient collocations are recycled in different types of exercise to expose learners to them repeatedly. For example, sentences containing collocations of the commonly confused words *broad* and *wide* can be used in a reconstruction “fill-in-blanks” exercise that asks learners to form a valid collocation, while the same data can be used in a “correcting common mistakes” exercise that asks learners to identify and correct words that do not form strong partnerships. Exercises can be constructed to foster receptive or productive knowledge by making the answers available or forcing students to provide their own.

Repetition also occurs when learners are asked to record and organize collocations that they think are useful for an essay assignment or oral presentation. Bates (1989) introduces the idea of “berry-picking” to model the behavior of real users of information retrieval systems: choosing juicy documents from the briar patch. We adapt this as “cherry-picking” to describe how students can gather useful collocations while reading an article, or when searching and browsing collocations. Cherries grow in twos and threes, which reinforces the idea of collocation.

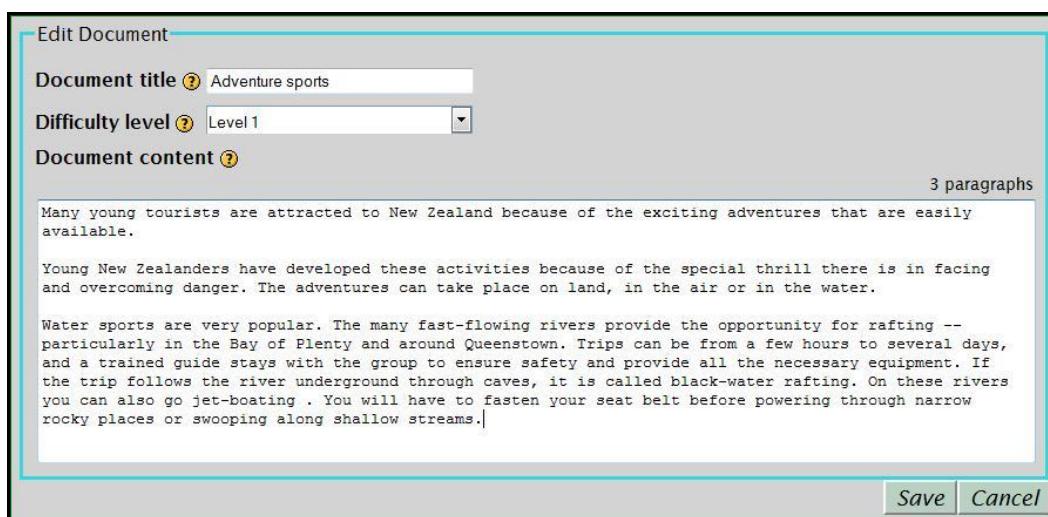
CLS links to external material to illustrate collocations in different contexts, enriching the learner’s collocation knowledge and promoting generative use. Currently, students can look at text samples extracted from the BNC and the Web itself, or examine related collocations retrieved from the WEB COLLOCATIONS collection. Other resources (not implemented for this thesis) such as online dictionaries or thesauri could also be incorporated into the system.

Generation can also be achieved through participating in collocation activities. CLS supports two kinds of activity: collection-based and dictionary-based. They stem from traditional classroom activities, serve different teaching purposes, and complement each other. There is a wide range of possible activities. For demonstration purposes, we have implemented four for each kind, chosen because:

1. they are common and popular in the classroom,
2. exercises, including questions and answers, can be automatically constructed, and
3. answers can be checked by the computer.

The collection-based activities are Fill-in-Blanks, Common Alternatives, Correcting Errors and Multiple Choice. As exercise material, they use collocations identified from the text and the text itself—which could be individual sentences or entire articles. When a sentence is used, the preceding and following sentence are also provided as context. WEB COLLOCATIONS and WEB PHRASES are incorporated into these activities. Collocations from the former serve as hints for students when doing exercises, and frequency associated with the latter is used to give scores to students in the Common Alternatives activity.

The dictionary-based activities are Collocation Guessing, Collocation Dominos, Collocation Matching and Related Words. These make use of collocations from the WEB COLLOCATIONS collection, and allow teachers to design exercises to expand the student's collocation knowledge for particular words. For example, teachers create exercises that ask students to seek other adjectives that strongly collocate with the word *adventure* after they have learnt *exciting adventure*, or exercises that help students differentiate the words *wound* and *injury*. To make this kind of activity more interesting, fun factors are added to the design: Collocation Guessing and Collocation Dominos mimic the *tetris* and *dominos* games respectively.



**Figure 6.2 Collection building interface: adding an article**

## 6.2 *Building collocation-enriched collections*

This section describes how to build a collocation-enriched collection using the Greenstone digital library software. For demonstration purposes, we used a dozen short articles of general interest, in which the available metadata are titles and difficulty level.<sup>20</sup> The standard Greenstone system allows such a corpus to be built into a digital library collection, equipped with a full-text index and metadata browsing facilities. We have enhanced the system to allow collection building through a Web browser and added a process to automatically identify collocations in the text and organize them to support collocation searching, browsing and learning. This section introduces the collection building procedure, focusing on how collocations are identified in given documents.

### 6.2.1 Adding texts

Building a collection involves five steps:

1. provide a collection name and a description,
2. upload the texts,
3. configure collocation identification parameters,
4. select collocation activities, and
5. create the collection.

---

<sup>20</sup> These articles are from the University of Waikato Pathway College's IELTS course.

Extract collocations from documents

<b>Collocation Types</b>	<input checked="" type="checkbox"/> Noun + Verb <input checked="" type="checkbox"/> Noun + Noun <input checked="" type="checkbox"/> Noun + <i>of</i> + Noun <input checked="" type="checkbox"/> Adjective + Noun <input checked="" type="checkbox"/> Adverb + Adjective <input checked="" type="checkbox"/> Adverb + Verb <input checked="" type="checkbox"/> Verb + Noun <input checked="" type="checkbox"/> Verb + Adverb <input checked="" type="checkbox"/> Verb + Verb <input checked="" type="checkbox"/> Verb + Adjective
<b>Enable Cherry Picking</b>	<input checked="" type="radio"/> Yes <input type="radio"/> No
<b>Use Web Data</b>	<input checked="" type="radio"/> Yes <input type="radio"/> No
<b>Frequency Value</b>	1000
<b><input checked="" type="checkbox"/> Collocation Fill-in-blanks</b> Learners fill in a blank to form a valid word combination. Type: Individual Instructions:	
<b><input checked="" type="checkbox"/> Common Alternatives</b> Learners find more words that go with the target word. Type: Individual Instructions:	

**Figure 6.3 Configuring collocation identification parameters**

The first step is straightforward. For the second, Figure 6.2 shows the interface for adding texts to the collection. The teacher provides the title (*Adventure sports*), selects a pre-defined level (beginner, intermediate and advanced) or specifies their own (as in this case *Level 1*), and then cuts and pastes the text into the box below. Clicking the *save* button uploads the text and brings up a blank form for the next document.

Once uploading is finished, the teacher configures the collocation identification parameters through the interface shown in Figure 6.3. She specifies (1) the collocation types (Table 3.6) that the system looks for in the text (the default is all ten types), (2) whether to allow “cherry-picking” (Section 6.3.3), (3) whether to use frequency from the WEB PHRASES collection to filter the collocations that are identified, (4) the frequency cut-off value below which collocations will be discarded (see Section 6.2.2). This parameter allows teachers to control the collocations they want their students to focus on—for example, the most frequent ones—and to discard collocations that do not occur in the WEB PHRASES collection because they are likely to be incorrect or infelicitous.

In step 4, the teacher selects which of the four collection-based and four dictionary-based collocation activities are to be associated with this collection. In the final step, she builds the collection if she satisfies with what she has done, or returns to the previous steps to make changes.

### 6.2.2 Identifying collocations

The process for identifying collocations is:

1. split the text into sentences,
2. assign part-of-speech tags to words,
3. match tagged word sequences against a set of syntactic patterns,
4. discard sequences that do not occur in the WEB PHRASES collection,
5. associate sample text with the collocations that have been identified, and
6. build search indexes and browsing structures.

In steps 1 and 2, OpenNLP is used to split the text into sentences and assign part-of-speech tags to words. In step 3, tagged words are matched against regular expressions defined for each collocation type (Table 4.8). Step 4 matches the sequences that are identified in the text against the WEB PHRASES collection and discards ones that do not appear or whose frequency falls below the specified frequency cut-off value. This step can be disabled, which might be desirable if collocations are expected to contain neologisms (such as the word *google*) that do not appear in the BNC and have therefore been omitted from WEB PHRASES (Section 3.1.1). We also use the frequency recorded in WEB PHRASES for ranking collocations when presenting them to students, to help them prioritize learning.

Whenever a collocation is identified, its sentence and the one before and after are extracted and associated with it in step 5. These allow students to study collocations in context rather than as isolated items, and are used in the learning exercises described below. To facilitate searching and browsing, step 6 builds indexes on the constituent words of each collocation, and creates browsing structures that group collocations by the words they contain, and by their type (Table 3.6).

As explained in Section 4.3, the process of identifying collocations is not perfect. Chapter 7 reports on a comparison of automatically identified collocations with those manually marked by teachers.

**About this collection:**

A collection of a dozen reading articles, ranging from 600 to 800 words each, aimed at international students in New Zealand studying an IELTS course. The material was prepared by the University of Waikato Pathways College.

**Language activities:**

These activities are particularly related to collocation learning. Collocations can be defined as words that commonly occur together, for example *heavy smoker*, *strong tea*, and *make efforts*.

<b>Multiple Choice</b>	<b>exercises</b>	<b>create an exercise</b>	
Learners choose a right word for a given word or phrase.			
<b>Type:</b> Individual			
<b>Correcting Errors</b>	<b>exercises</b>	<b>create an exercise</b>	
Learners must correct the word choice errors in a text.			
<b>Type:</b> Individual			
<b>Fill-in-blanks</b>	<b>exercises</b>	<b>create an exercise</b>	
Learners fill in a blank to form a valid word combination.			
<b>Type:</b> Individual			
<b>Common Alternatives</b>	<b>exercises</b>	<b>create an exercise</b>	
Learners find more words that go with the target word.			
<b>Type:</b> Individual			

**Figure 6.4 Example collection’s “About” page**

### 6.3 Using the collection

Figure 6.4 shows a collection built from the dozen short articles from the Waikato Pathways College’s IELTS course. This “About” page displays the collection’s title, description, and a list of learning activities. The *search* button allows users to seek documents and collocations containing particular words or phrases; they can also browse documents by title and difficulty level, and browse collocations by word and collocation type. Here we focus on collocation-related facilities.

#### 6.3.1 Searching and browsing collocations

Three ways are provided to access the collocations: in the context of an article; locating partners of a particular word; and browsing collocations by word and type. As in any digital library, users can find articles by searching or browsing, and display them. Here, an alternative version of each article is provided with

The screenshot shows a web page with a header 'Stamp Collecting for Beginners' and two tabs: 'Original' and 'Collocations'. The 'Collocations' tab is active. The main text discusses stamp collecting as a hobby, mentioning various items like coins, post cards, Coca Cola bottles, phone cards, and buttons. It highlights the 'extremely high' prices of some rare stamps, noting that they can be collected for free through the post. The text also describes buying new stamps from local post offices and finding them at stationery shops and specialist dealers. It mentions auction sites on the internet and local stamp clubs. The word 'extremely high' is underlined and followed by four small icons.

People around the world **love collecting things**; coins, **post cards**, Coca Cola bottles, **phone cards**, buttons – you name it, someone, somewhere collects it. Stamp collecting is one of the most **popular hobbies** in the world. It is both relaxing and fun, and, despite the **extremely high** 📈GBP prices of some of the world's rarer stamps, you can **collect stamps** for free, as they **come through the post** every day. Even buying **new stamps** from the **local post office** is **relatively inexpensive**. Many people get more stamps by asking friends and relatives to save any special ones that they get in the mail, and if you have more than one of **the same stamp**, then it is easy to **swap with friends**.

As well as the **post office**, stamps can often be bought at **stationery shops**. These often have **special collectors'** packs of **overseas stamps**, and they are generally not too expensive. You can also **find stamps** at **specialist shops** run by **stamp dealers**. These dealers will often have catalogues listing, for example, all the stamps ever sold in New Zealand. They also **list the price** of each stamp, should you **wish to buy** one, but these rarer stamps are generally quite expensive. Increasingly, **stamp collectors** looking for more 'hard-to-find' stamps will turn to **auction sites** on the internet. For instance, in New Zealand, a check on the Trademe website revealed thousands of **stamp items** for sale, with 1040 in the Waikato region alone.

Probably **the easiest way to start a stamp collection** is by buying (or better still, being given) **an old collection**. Shops like the New Zealand Post sell special 'Stamp Collector's Beginner Packs' which contain some stamps plus all the things that you **need to start** your collection. Joining **a local stamp club** is also a **good idea**, as you will **find people** there who can show you how to organize and **present your collection**. Most **big towns** and cities will have a **stamp club** – there may even be a club in your school or college. These clubs usually have **regular meetings** where members can **share information** about stamps and **swap stamps**. More **serious collectors** may want to **show off** their collections at **stamp competitions**.

**Figure 6.5 A document in the collection, with collocations highlighted**

collocations highlighted, to help students notice them and study their context. In the example shown in Figure 6.5, collocations related to stamp collecting—*collect stamps*, *new stamps*, *overseas stamps*, *stamp dealers*, *start a stamp collection*, *stamp club*, *stamp items*, *swap stamps*, *stamp competitions*—stand out from the rest of text, attracting the student's attention. The collocation *extremely high* (in the third line of text) has been clicked to reveal four small icons. Their function is described in Section 6.3.2.

From the first button at the top of Figure 6.4, *Search*, users can seek collocations in the collection that contain a particular word. Figure 6.6 shows the beginning of the result for the word *family*, sorted by frequency in the WEB PHRASES collection. The context of each occurrence—here there are five instances of the first collocation, *family members*—are gathered together to acquaint learners with different usage. For *family*, the most dominant collocation types are noun + noun and noun + of + noun: *family members*, *family history*, *family tree*, *family relationships*, *generations of family*, *side of a family*, and *encouragement of family*.

The screenshot shows a search interface with the following details:

- Search for:** collocations that contain the word **family**
- Results:** there are 20 collocation(s) matched the query **family**
- Top Result (family members):**
  - Icon:** A red rose icon.
  - Text:** family members
  - Icons:** Three small icons: a red rose, a car, and the British flag.
  - Count:** 4,800,000
  - Description:** In many cultures a further distinction is made between **family members** on the father's side and on the mother's side. American Indian languages such as Crow and Omaha use different terms for relatives, depending on which side of the family they belong to.
  - Notes:**
    - In some eastern languages, for instance Thai and Chinese, there is clear differentiation among **family members** according to age. The Thai language uses the prefixes "pi" and "nong" to show whether someone is older or younger than the speaker.
    - However, when it comes to words for describing family relationships, English is quite simple. Even though English does make some linguistic distinction between the sexes of family relationships (brother - sister, uncle - aunt), it does not usually specify differences in the age or social status of **family members**. Nor does it show which side of a family the members belong to.
    - The best place to start is by talking to living relatives. If you have older **family members** you can talk to, you may be able to save time by starting with them, and ignore any younger relatives. Interview older relatives to find out any of their family stories, particularly about memories of their childhood.
    - If possible, try to collect old letters, addresses, photographs, plus any educational or military records and certificates that may still exist. Old books like bibles and diaries can sometimes include names of **family members** and important dates. Any scrap books or old newspaper clippings may also provide vital clues in your search for more information.
- Second Result (family history):**
  - Icon:** A red rose icon.
  - Text:** family history
  - Icons:** Three small icons: a red rose, a car, and the British flag.
  - Count:** 1,500,000

**Figure 6.6 Collocation results when searching for the word *family***

Collocations are organized by word and type to facilitate browsing, invoked by the *Browse* button in Figure 6.4. When browsing by word, an alphabetic selector leads to the word in question—clicking the letter *f*, followed by the word *family*, obtains the collocations shown in Figure 6.6. Browsing by type retrieves all collocations of a particular type. Figure 6.7 shows some verb + noun examples: *take advantage of*, *take into account*, *lose weight*, *save time*, etc.

### 6.3.2 Expanding collocation knowledge

The last three of four small icons shown alongside the selected collocation (*extremely high*) in Figure 6.5, and the three icons shown after each collocation in Figure 6.6 and Figure 6.7, present additional resources associated with it. The first shows related items from the WEB COLLOCATIONS collection described in Chapter 3; the others retrieve relevant text samples from the Web and the BNC respectively.

The first function, invoked by clicking the second of the four little icons in Figure 6.5 or the first of the three icons in Figure 6.6 and Figure 6.7, opens a popup window showing different collocations that have the same first and last word

	<a href="#">titles</a>	<a href="#">collocations by word</a>	<a href="#">collocations by type</a>
<b>Noun Verb</b>			
 <a href="#">Noun of Noun</a>			
 <a href="#">Verb Adverb</a>			
 <a href="#">Verb Noun(s)</a>			
 <a href="#">take advantage of</a>	  	6,100,000	
○ Typically, someone seeking to cut back from full-time employment will opt to work somewhere between 50% and 80% of a full-time workload. These workloads can also be applied so that more staff are working during busier periods and both employers and employees can then <u>take advantage of</u> reduced staff work hours during quieter periods.			
 <a href="#">take into account</a>	  	2,500,000	
○ The garments needed to be affordable, and easily available in all sizes. It was also important to <u>take into account</u> the different seasons, so that variations could be made to the uniforms. Students needed to be comfortable in both summer and winter.			
 <a href="#">lose weight</a>	  	1,900,000	
○ Postal worker Richard O'Brien hit the headlines in August when it was revealed that the then 177 kg (390 lbs) parcel sorter was forced to take sick leave to <u>lose weight</u> .			
 <a href="#">save time</a>	  	1,300,000	
○ The best place to start is by talking to living relatives. If you have older family members you can talk to, you may be able to <u>save time</u> by starting with them, and ignore any younger relatives. Interview older			

**Figure 6.7 Browsing by collocation type**

respectively. Figure 6.8 gives the output for *extremely high*: the 20 most frequent related WEB COLLOCATIONS, sorted by frequency. For the first word they include *extremely important*, *extremely difficult*, *extremely low*, *extremely useful*, and so on; for the last we see *relatively high*, *unusually high*, *fairly high*, and *consistently high*. The *more ...* button at the bottom leads the user to a page on which more of these collocations can be found.

### 6.3.3 Cherry-picking

Figure 6.9 shows the cherry-picking interface that is launched by the two-cherry icon that follows the collocation in Figure 6.5 (also seen before each collocation in Figure 6.6 and Figure 6.7). In this case, *collect stamps* has been chosen because the article is about stamp collecting. The selected collocation is added to the student's personal cherry basket. They can optionally assign it to a category or categories, or add a new category—say “stamp collecting”—for it, then assign the collocation to it. The default is to leave it uncategorized. Students can pick

The screenshot shows a web page titled "Stamp Collecting for Beginners". At the top right are two buttons: "Original" and "Collocations", with "Collocations" being highlighted. Below this is a text paragraph about stamp collecting, mentioning "love collecting things", "popular hobbies", and "extremely high" prices. To the left of the main text is a table titled "related collocations" with two columns of words and their meanings. To the right is a large block of text describing stamp collecting.

	related collocations	
buying r	extremely important	relatively high
asking f	extremely high	extremely high
one of tl	extremely difficult	unusually high
	extremely low	fairly high
As well	extremely useful	consistently high
collector	extremely easy	particularly high
at speci	extremely rare	sufficiently high
the star	extremely helpful	exceptionally high
but thes	extremely popular	really high
'hard-to-	extremely small	especially high
Tradem	more...	more...

People around the world **love collecting things**; coins, **post cards**, Coca Cola bottles, **phone cards**, buttons – you name it, someone, somewhere collects it. Stamp collecting is one of the most **popular hobbies** in the world. It is both relaxing and fun, and despite the **extremely high** prices of some of the world's **most valuable stamps**, many people get more stamps by **mail**, and if you have more than **1000**, these often have **special** **values**. You can also **find stamps** in **catalogues** listing, for example, all **stamps**, should you **wish to buy** one, or **stamp collectors** looking for more. In New Zealand, a check on the **Waikato** region alone.

**Figure 6.8 Collocations related to *extremely high***

collocations from the articles shown in Figure 6.5, or from the search results page (Figure 6.6), or from pages reached by browsing collocations (Figure 6.7).

Figure 6.10 shows a student's personalized cherry basket. It displays collocations that the user has picked and placed into two categories: *family history* and *career*. Students can study items in the basket using the three icons described in Section 6.3.1. They can also delete collocations, move them into different categories, create new categories and delete old ones, and print the basket to take home—or send it to friends (the *Print friendly* button).

#### 6.4 Collocation activities

Built into the collocation learning system are four collection-based and four dictionary-based collocation activities. These are more accurately called activity *types*, for within each one a virtually unlimited number of exercises can be created by the teacher or learner, using an exercise design interface described below, from documents in the collections teachers or learners have built or from the WEB COLLOCATIONS collection.

The collection's “About” page (Figure 6.4) displays a list of activities that the teacher has selected when building the collection, including a brief description of each one, two buttons (*exercises*, *create an exercise*), and an icon depicting a person. The first button presents a list of exercises that have already been created;

The screenshot shows a web page titled "Stamp Collecting for Beginners". At the top, there are two tabs: "Original" and "Collocations", with "Collocations" being the active tab. The main content area contains text about stamp collecting, mentioning "love collecting things" like coins, post cards, Coca Cola bottles, phone cards, and buttons. It also discusses collecting stamps from the local post office or specialist shops, and buying new stamps online. A modal dialog box is overlaid on the page, prompting the user to "Add a collocation into the cherry basket". Inside the dialog, there is a text input field with placeholder text "Add this collocation into", a checkbox labeled "no category" (which is checked), a checkbox labeled "stamp collecting" (which is also checked), and a "Add Collocation" button.

**Figure 6.9 Picking cherries**

Cherry Basket					Print friendly
	career				
	efficient work environment				2713
	job offers				205656
	strong CV				462
	communication skills				1883369
	personal attributes				55851
	makes the best impression				518
	job spells				1801
	family history				
	family relationships				221449
	family tree				1131164
	family stories				104700
	ancestry chart				567
	direct ancestors				14164

**Figure 6.10 Cherry basket**

newly created ones are added automatically when the teacher saves them. The second button allows students or casual visitors to create (and use) temporary exercises with all the functionality of ones supplied by teachers, but these do not appear in the exercise list. For this they use precisely the same interface as teachers, described below. Only registered users—typically teachers—can create exercises that persist, and they must first log in using the “person” icon.

Each activity is associated with an exercise design interface through which teachers select materials for their students, create exercises at different levels of language difficulty, provide answers where necessary, and apply quality control to the automatically generated exercise content. First they must determine the purpose of the exercise and select material accordingly. Then they preview the questions that CLS provides, and remove unsuitable ones. For some activities answers are taken from the original text, while for others they are generated by the system. The latter is cheap but potentially unreliable, and teachers may wish to correct the system's suggested answers before the exercise is used.

Each activity also comes with a set of exercise parameters with which teachers design different exercises for their students for different teaching purposes. Each parameter has a default value that kicks in automatically if that parameter is not specified, so that CLS can always generate a default exercise. Moreover, the values of some parameters are picked randomly, so that a different exercise is obtained each time.

Below we introduce each activity individually, focusing on interface and design considerations. Because collection-based activities and dictionary-based activities have a slightly different set of parameters and exercise design interface, they are described separately.

#### 6.4.1 Collection-based activities

The four collection-based activities are Fill-in-Blanks, Common Alternatives, Multiple Choice and Correcting Errors. Correcting Errors exercises use whole documents; the others use sentences.

##### **Fill-in-Blanks**

Fill-in-Blanks exercises involve a set of collocations and their associated sentences. Constituents of the collocations are selectively removed from the text, and the learner is asked to choose the word that completes each collocation.

**Fill-in-Blanks**

Score: 0 out of 10 How to play | Summary report

save (1)   take (3)   lose (1)   eliminates (1)   encourage (1)   share (1)   play (1)   spend (1)

1. Typically, someone seeking to cut back from full-time employment will opt to work somewhere between 50% and 80% of a full-time workload. These workloads can also be applied so that more staff are working during busier periods and both employers and employees can then \_\_\_\_\_ advantage of reduced staff work hours during quieter periods.

2. The garments needed to be affordable, and easily available in all sizes. It was also important to \_\_\_\_\_ into account the different seasons, so that variations could be made to the uniforms. Students needed to be comfortable in both summer and winter.

3. Postal worker Richard O'Brien hit the headlines in August when it was revealed that the then 177 kg (390 lbs) parcel sorter was forced to take sick leave to \_\_\_\_\_ weight.

4. The best place to start is by talking to living relatives. If you have older family members you can talk to, you may be able to \_\_\_\_\_ time by starting with them, and ignore any younger relatives. Interview older relatives to find out any of their family stories, particularly about memories of their childhood.

5. Most big towns and cities will have a stamp club -- there may even be a club in your school or college. These clubs usually have regular meetings where members can \_\_\_\_\_ information about

**Check answers**

**Figure 6.11 Fill-in-Blanks exercise**

Figure 6.11 shows one such exercise, which focuses on finding the right verb for a noun. The missing verbs are given at the top of the exercise panel. When chosen, they disappear from this list—except for words that occur more than once, in which case the occurrence count (in parentheses) is decremented. Below is a list of items with target verbs omitted and the remainder of the collocation rendered in italics. The learner completes a collocation by dragging a word from the top and dropping it into place, where it appears in blue; the move can be undone by clicking the word. When the *Check Answer* button at the lower left is clicked, correctly formed collocations remain, but the offending word is removed from incorrect ones and reinstated at the top of the panel. The light bulb beside each collocation signifies a hint, and clicking it retrieves relevant items from the WEB COLLOCATIONS collection. For example, the hint for *advantage of* includes *added advantage of*, *gain a competitive advantage*, *create a competitive advantage*, *offer a tremendous advantage*, *get the advantage of*, and *see the advantage of*.

This activity works well for sets of words that share similar meanings but have different usage. Learners are frequently confused by common words—*make* and *do*, *speak* and *tell*, *see* and *look*—and find it difficult to understand their

differences by consulting dictionaries. Studying collocations is an effective way to help learners distinguish a word's various shades of meaning. The presentation in Figure 6.11 reinforces receptive rather than productive knowledge, but teachers can select a version in which the missing verbs are not shown at all but must be typed in by the learner. This reinforces productive knowledge, and is far more challenging.

### Common Alternatives

To add strength to adjectives, learners tend to use the word *very*, but in specific contexts there are usually more precise qualifiers that perform the same function. When describing someone as very beautiful, alternatives such as *really*, *truly*, *stunningly* and *incredibly* spring quickly to the mind of a native speaker, and are usually preferred. These alternatives can be found in the WEB COLLOCATIONS collection—in this case, a quick search finds 100 adverbs with frequency exceeding 1000, all of which are appropriate. The Common Alternatives activity helps elicit and expand this knowledge. Given a target word along with some collocation examples, learners are asked to enter as many collocations as possible—and their choices are scored.

Figure 6.12 shows an exercise that focuses on nouns commonly associated with the verb *reduce*. To get learners started, they are given some sample collocations: three from the original text—in this case *reduce stress*, *reduce heat loss* and *reduce fighting*—and one from the WEB COLLOCATIONS collection—here, *reduce the risk of*. The first three, from an article in the library that the teacher may already have asked students to read, refreshes their memory of this word. The other is the most frequent *reduce* + noun item in WEB COLLOCATIONS, and is intended to help students think of other common ones. The icons that follow each collocation allow students to retrieve text samples from the Web and the BNC.

Learners type a word or phrase into the text box and press the Enter key, at which point the system checks it. For example, *reduce more* would be invalid because this exercise requires nouns, or a phrase that contains a noun. If it is valid, the input text, preceded by the word *reduce*, is sought amongst *n*-grams of the same length in the WEB PHRASES collection. If it is found, the associated frequency is

**Figure 6.12 Common Alternatives exercise**

retrieved from that collection and used as a score. The learner is notified if the collocation is invalid or the phrase does not appear amongst WEB PHRASES; otherwise it is displayed along with its score and the two standard icons for further exploration (Web and BNC). In Figure 6.12 the user has already entered *reduce costs*, *reduce poverty*, and *reduce the possibility of*, for a total score of 10,181.

Competitive factors make this activity compelling. Learners can be connected to work on the same exercise and see each other's scores. This challenges them to outwit one another, and encourages them to discover more collocations.

### Correcting Errors

Unlike the preceding activities, Correcting Errors exercises are created from full documents rather than excerpts. Correcting language errors is a relatively difficult task because of the ambiguity of language, so to provide as much context as possible, the entire document is given. The teacher first chooses a document and several target collocation types, and then decides whether learners will work on the first or last constituent word. CLS replaces these words with infelicitous choices that learners must correct.

**Correcting Errors**

### The Truth About Career Beliefs

Fresh college graduates starting out on their careers are often confused by the conflicting information about their work and careers: "College grades are more important than experience." "My parents know best." "If I put my CV\* on the internet, the job offers will appear flooding in." Unfortunately, it is not always so easy for graduates to sort out the good information from bad, but knowing the truth about these common mistruths can help improve stress and assist in finding the right career.

When applying for jobs, the most important thing is to be realistic in what to expect. It is not always the most qualified person who gets the job, but rather the person who uses the best impression. A strong impression starts with a strong CV, and a strong CV gets you a job interview. Once you have made it to the job interview, then there are many other ways to impress, for example through personal attributes such as enthusiasm, confidence and honesty, as well as through networking and communication skills. The interview is the chance for you to prove that you are the best candidate for the job.

Although choice of college majors and grades can be important for some jobs, this does not mean that you have to match your major to a particular job, or that applicants with slightly lower grades will be ignored. When choosing a subject to study, probably the best advice is to give a subject that you like. You will have a chance to make more knowledge about different jobs through internships or later studies. Grades show that someone has the ability to study and learn, but it is equally important that that person also has strengths in other areas, such as leadership or technical skills.

Many new graduates worry too much about their first job. It is worth bearing in mind that most new graduates only stay in their first job for between 1 and 3 years. You are not a prisoner in your job, so if the first job doesn't

**Figure 6.13 Correcting Errors exercise**

Figure 6.13 shows an example, *The Truth About Career Beliefs*, which focuses on collocations of the verb + noun type and asks learners to find the right verb for the noun. Target collocations are underlined, and incorrect words colored in blue. Clicking a blue word brings up a box into which the student types a new word. The answer is checked when the learner presses the Enter key or moves to another word. Correct entries are changed to black, while incorrect ones remain blue. The hint icon (light bulb) shows more collocations, retrieved using the target collocation's first and last words. For example, the first set of hints for *improve stress* include *improve the accuracy of*, *improve performance*, and *improve the lives of*; while the second set includes *reduce stress*, *cope with stress*, and *handle stress*. To make them more relevant, the collocations adapt to what the learner has entered—if the learner changes *improve stress* to *decrease stress*, the collocations of *improve* are replaced by those of *decrease*.

### Multiple Choice

Multiple Choice exercises, comprising a question and a set of choices—typically four—from which the correct answer must be selected, are widely used language drills for learning vocabulary. We tailor this activity to collocation learning by

**Multiple Choice**



1. However, the factor that has probably had an impact on *the \_\_\_\_\_ number* of employees is that of being able to work a part-time schedule. Traditionally, a part-timer worked a half workload or less.
 

greatest     best     finest     shortest
  
2. This means that it is easier to fit work around school schedules, for example. People who may be otherwise unable to work a full-time position because of other commitments can still play *an \_\_\_\_\_ part* in the workforce. At the same time they can still bring in extra income for the family.
 

extra     entire     additional     active
  
3. People who may be otherwise unable to work a full-time position because of other commitments can still play an active part in the workforce. At the same time they can still bring in \_\_\_\_\_ income for the family. In addition, job-sharing also has benefits for the employer.
 

extra     large     additional     external

**Figure 6.14 Multiple Choice exercise**

using sentences containing particular collocations as questions, with one collocation part missing. Four choices, including the correct one, are shown to students, who must select one that forms a valid collocation.

Figure 6.14 shows an exercise that asks students to complete adjective + noun collocations. The collocation is rendered in italics, and one part is missing: learners must select the correct choice. When the *Check answer* button at the bottom of the screen (not shown) is clicked, the learner's correct choices are inserted into the blanks, while incorrect ones are left so that they can continue working on them. As with other activities, clicking the light bulb brings up further related collocations.

### Exercise parameters

For each of the four exercise types described above, the exercise content is selected by determining a few parameters that control the material retrieved by CLS. All have default values, and if no configuration is necessary a complete exercise can be generated with a couple of clicks of the mouse. Here are the principal parameters.

**Collocation type** determines what types of collocation are to be used, selected from a drop-down list that shows the ten types in Table 3.6 (multiple selections are possible). Learning can be enhanced by tailoring collocation types to the teacher's goals and the student's ability.

**Collocation position** specifies either the first or the last word of collocations. For example, in Fill-in-Blanks learners may be asked to specify *make* in        *an effort*, or *effort* in *make an \_\_\_\_\_*. Based on their objectives, teachers set either component as the target. Here, the first word would be an appropriate choice if the focus is on learning verbs associated with the noun *effort*.

**Hint** determines whether learners can receive extra help while doing the exercise. The WEB COLLOCATIONS collection is used as the source of hints. Given the example        *an effort*, a hint displays the 20 most frequent verbs that collocate with the noun *effort*.

**Number of sentences** determines the size of the exercise, in terms of how many questions are posed to learners. For the Correcting Errors activity, which does not use individual sentences, the teacher instead specifies **Document title** to determine which document to use.

**Contains words**, specific to the Fill-in-Blanks activity, allows teachers to design exercises focusing on particular words. If specified, only collocations that contain those target words are used. For example, teachers can create an exercise specifically to help students differentiate the commonly confused words *do* and *make*.

### Providing answer candidates

For two of the four exercise types described above, candidate answers are generated automatically. In Correcting Errors, the original words are replaced with incorrect ones, and in Multiple Choice, there are three incorrect choices for each question. It is not easy to find words that are incorrect yet plausible. Here we examine how CLS reduces the teacher's burden by providing a list of candidates. When creating an exercise, teachers can determine which of these to use, or provide their own candidates.

For each collocation, 20 candidates are generated during the collection building process. They are not randomly chosen. Rather, they must (1) somehow fit the context, (2) be of the correct form, and (3) not form a valid collocation. As an example of the second criterion, if a past tense verb or plural noun is used in the original text, the same must be true of each candidate. For the third, if the target collocation is *make a complaint*, candidates such as *file*, *lodge*, *resolve*, and *investigate* are not selected because they collocate strongly with *complaint*.

The process involves three steps, corresponding to the three criteria described above. We explain it using the example sentence

*Some of these communities have made a great effort to **improve this situation** by running special classes ...*

where *improve this situation* is the target collocation and *improve* is the target word. First, the preceding text, *effort to*, is used to locate verbs that somehow fit the context. CLS consults the WEB PHRASES collection and retrieves verbs that follow *effort to*. Using just two words as context generally yields a satisfactory list of candidates. Next, the candidates are tagged and discarded if their tag does not match that of the target word—in this case, *improve* is a verb in base form (recall that words of collocations are tagged when the collection is built). Finally, to remove candidates that form good collocations with *this situation*, the five-word phrase that encloses *improve this situation* is extracted from the original text, yielding *to improve this situation by*. Then verbs extracted in the second step are used to replace *improve*, and discarded if the resulting phrase does occur in the WEB PHRASES collection. In this example, the following “incorrect” candidates might be chosen:

- *to assure the situation by*
- *to present the situation by*
- *to develop the situation by*
- *to promote the situation by*
- *to maintain the situation by.*

### Exercise design interface

The interface for the Fill-in-Blanks activity, shown in Figure 6.15, is used to illustrate the exercise design process for all four exercise types. At the top, teachers enter a name for the exercise, and, optionally, select a category. Categories can be used to create exercises at different levels of difficulty, and new ones added if desired. The next panel is for exercise parameters, where the teacher selects a collocation type and, if desired, enters a word or words that must appear in all collocations—*take* and *make*, in this case.

The next panel gives the number of sentences to choose from, and is automatically updated following any parameter change. For example, this collection includes 180 sentences that contain verb + noun collocations, but this changed to 16 in the interface when the words *make* and *take* were entered, because only 16 sentences include those words. In the next panel, the teacher decides how many sentences to use in the exercise, whether learners have to guess the first or last word of collocations, and whether hints are allowed. The buttons underneath, Preview, Display, Print and Save, allow teachers to review the sentences and collocations that have been chosen, try out the exercise just as a student would, print it on paper, and save it for students to use. The last three are self-explanatory; we look at the first in more detail.

All exercise content is determined automatically based on the parameters specified. However, teachers may not be satisfied with what they see because (1) the question text may contain complicated structures or difficult vocabulary items that could hinder learning; (2) students may have already mastered some collocations that have been retrieved; (3) there may be errors in collocations (e.g., a noun + noun type may be marked as verb + noun); or (4) the items may be unsuitable for other reasons. During the preview process teachers apply quality control, discarding unsatisfactory questions and modifying the automatically generated answers or replacing them with their own.

The screenshot shows the 'Collocation Fill-in-Blanks' activity setup in CLS. At the top, there are 'Create Exercises' and 'List Exercises' buttons. The main title is 'Collocation Fill-in-Blanks'. Below it, the 'Exercise name' is set to 'take and make'. A dropdown for 'Select a category' shows 'none' with an option to 'Add a category'. Under 'Exercise parameters', 'Collocation Type' is set to 'Verb+Noun' and 'Contains words (separated by comma)' is set to 'take,make'. The 'Number of sentences to choose from' is set to 16. Under 'Activity parameters', 'Number of sentences' is set to 10, 'Collocation position' is set to 'first word', 'Show missing words' is set to 'Yes', and 'Hint' is set to 'Web collocations'. Below these settings are four buttons: 'Review', 'Print', 'Display', and 'Save'. The main content area displays three sentence fragments for editing:

- 3. Typically, someone seeking to cut back from full-time employment will opt to work somewhere between 50% and 80% of a full-time workload. These workloads can also be applied so that more staff are working during busier periods and both employers and employees can then \_\_\_\_\_ advantage of reduced staff work hours during quieter periods.
- 4. With the shift to a more knowledge-based economy, it is acknowledged that people are the most important asset of any company. In order to achieve the most efficient work environment, it is vital that employers \_\_\_\_\_ action to better manage workers' time. One area in which employers are doing this is by creating more flexible workplaces to meet their employees' needs.
- 5. Changing jobs every 3 to 5 years is commonplace now, and is no longer considered job-hopping. Experience with a number of different employers can be

For each sentence, there are 'Discard' and 'Show answer' buttons at the bottom.

**Figure 6.15 Design interface for the Fill-in-Blanks activity**

#### 6.4.2 Dictionary-based activities

CLS allows teachers to create four kinds of dictionary-based activities: Collocation Guessing, Collocation Dominoes, Collocation Matching and Related words. In these activities, teachers provide the words they want their students to focus on and create exercises using the content of the WEB COLLOCATIONS collection.

### Collocation Guessing

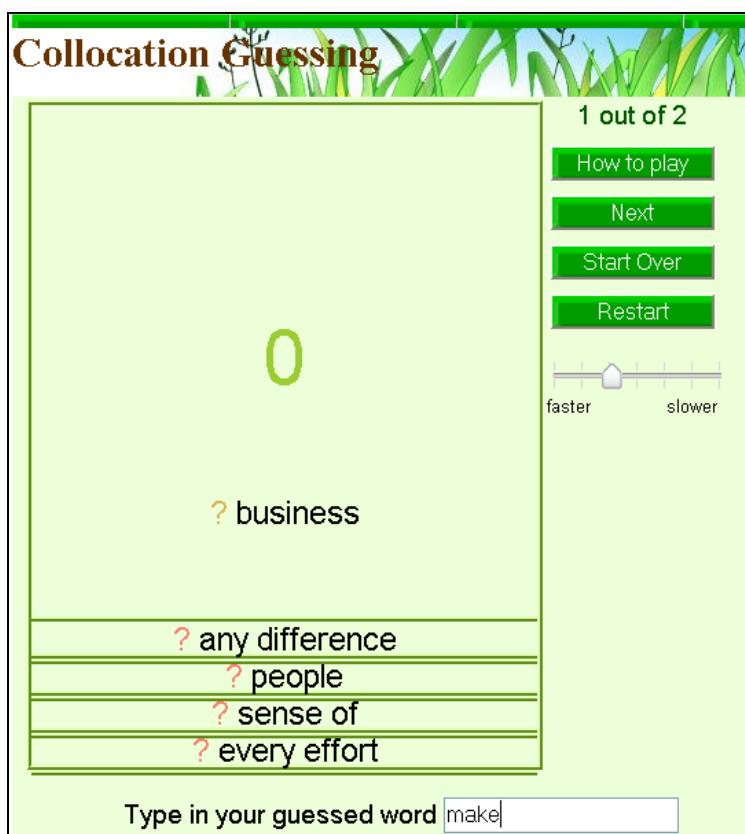
For the Collocation Guessing activity, the teacher chooses a target word and a number of associated collocations. CLS removes the target word from the collocation text, and then reveals the remaining text gradually to the learner who must guess the target word. For example, given the following words

*plain, dark, white, bitter, milk, bar of,*

the learner must guess the word that collocates with all of them. The answer is obvious to chocolate lovers.

The interface, shown in Figure 6.16, mimics the *tetris* game. One game comprises a word and a set of collocations; an exercise could contain more than one game. Collocation bricks are presented in the panel on the left side, and learners use the buttons on the right side to control the progress of the game. When the game starts, collocation bricks with the target word replaced with a question mark drop down one by one from the top of the game panel. Another follows as soon as the previous one reaches the bottom of the panel. Learners type in guesses continuously. The game is over when the correct word is given or collocations run out. Bonus points are awarded based on the number of collocations the learner has seen so far, at any time the learner can restart the current game, restart the whole exercise, or move on to the next game. The slider bar adjusts the speed at which the collocation bricks drop.

To create an exercise, the designer provides one or more target words. Using more than one word allows for creating subject or topic related exercises; alternatively exercises focus on a particular collocation type or a range of collocation types. Taking the word *make* as an example: if verb + noun were chosen, collocations used could be *make money, make use of, make every effort*; if all collocation types were used, they could be *make sure, make up, actually make, make money*. Both are good ways to enrich collocation knowledge.



**Figure 6.16 Collocation Guessing exercise**

### Collocation Dominoes

This activity mimics the *dominoes* game: the last word of the previous collocation becomes the first word of the next collocation. There is an example of collocation dominoes:

*bank cheque — cheque book — book club — club sandwich — sandwich board — board room ...*

Figure 6.17 shows an exercise created using the starting word *turn* and the noun + *of* noun type. The words that form the dominoes are given at the top of the panel. They are cut out from the dominoes and replaced with boxes in two alternating colors—the same color pair contains the same word. The first and last words are revealed to the learner, who drags and drops the words into boxes. Once one box is filled by the learner, the system automatically fills in the other one. A move can be undone by clicking the collocation text. The incorrectly formed collocations are changed back to boxes when the *Check Answer* button is clicked.

**Figure 6.17 Collocation Dominoes exercise**

English word classes are highly flexible, so verbs can be used as nouns or nouns as adjectives. Many learners, even advanced ones, may feel uncomfortable using noun + noun combinations and this activity can help them understand that these are standard English. The designer can decide whether collocation dominoes are open or closed. In the latter case, all words in dominoes are given to the learner, who puts them in the correct positions; in the former, any words can be filled in so long as the dominoes are complete. In either case the designer chooses a starting word, the maximum length of dominoes, and the collocation type.

### Collocation Matching

The Collocation Matching activity selects a set of collocations, normally from the same collocation type, splits each collocation into left and right part, and mixes the left and right parts separately. For example, *the secretary of state, course of action, hundreds of dollars* might be presented as:

<i>the secretary</i>	<i>of action</i>
<i>hundreds</i>	<i>of state</i>
<i>course</i>	<i>of dollars</i>

Learners must rematch them.

Figure 6.18 shows an exercise created using six quantification words: *grain, drop, slice, sheet, chuck, and bar*. The words and their associated nouns are split,

grain of bread  
drop of time  
a sheet of water  
a chunk of glass  
another slice of truth  
bar of chocolate

How to play   Start Over   Check Answer   New Exercise

**Figure 6.18 Collocation Matching exercise**

shuffled and placed in the left and right columns. Learners are asked to match a quantification word and a noun so that together they form a strong partnership by dragging and dropping the words on either side. At any point, they can restart the current exercise, check the answer, or start a new exercise that uses the same set of quantification words but with a different set of associated nouns.

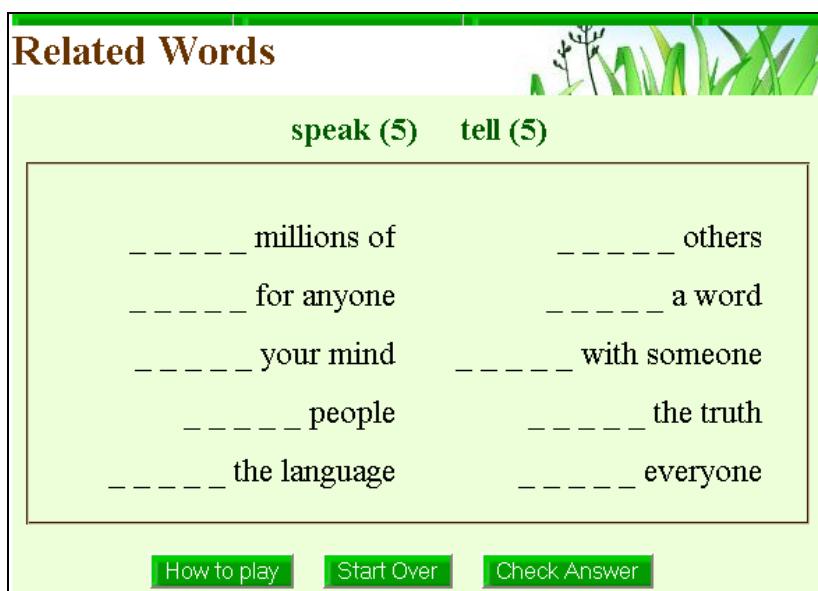
Picking collocations thematically can help learners practise particular groups of collocations, which adds extra value to this kind of activity. For example, exercises might be based on quantification words as in Figure 6.18, or certainty adverbials such as *certainly*, *definitely*, *surely* and *undoubtedly*.

### Related Words

The Related Words activity picks several related words and a number of their associated collocations, removes the related words from the collocation texts, and mixes the remaining collocation texts. For example:

pay make  
\_\_\_\_\_ the bill, \_\_\_\_\_ efforts, \_\_\_\_\_ the debt, \_\_\_\_\_ a difference

Learners are asked to choose the right word to complete a collocation, e.g., *pay the bill*, *make efforts*, *pay the debt* and *make a difference*.



**Figure 6.19 Related Words exercise**

Figure 6.19 depicts an exercise for the *speak* and *tell* pair, which are given at the top of the exercise panel. The number following a word indicates how many times this word can be used, and decrements each time it is used. Underneath is a list of collocations with related words replaced by dashed lines. They are grouped into two columns. The learner drags and drops a word onto a dashed line to complete a collocation, or undoes the move by clicking the collocation text. When the *Check Answer* button is clicked, the correctly formed collocations stay, but incorrect ones revert to their original state.

This activity works well with sets of words that share similar meaning but have different usage. Learners are often confused by a group of common words, and find it difficult to understand their differences just by looking them up in dictionaries. Studying their collocations is one effective way to help learners distinguish them. Some examples are *make* and *do*, *speak* and *tell*, and *see* and *look*.

### Exercise parameters

Again, each of the four exercise types described above is controlled by a set of parameters whose values are chosen by the user.

**Target words** are used to retrieve collocations that appear in an exercise. This parameter gives the designer control over the focus or purpose of an exercise. If absent, CLS randomly picks some words from a wordlist (see below). It is important to note that randomly generated words may not suitable for some activities. For example, words used for the Related Words activity should be somehow related as the activity name suggests.

**Collocation type** is the same as in collection-based exercises. Some types are particularly suitable for certain activities, such as noun + noun, and noun + *of* + noun for Collocation Dominoes. Different groups of students may experience difficulty in learning different collocation types, thereby a carefully selected type can effectively enhance learning.

**Sub-collection type** controls which sub-collection is used to retrieve collocations. Recall that we build three wordlist-based sub-collections that each consists of words in a particular wordlist (Section 3.4.3). This allows designers to control the level of vocabulary used in an exercise. For example, it is rarely a good idea to ask beginners to practise on academic words.

**Wordlist** is for generating random words. It is effective only if the Target Words parameter is not specified, and is used with the Collocation Type parameter if available. Suppose Wordlist is set to 1000 and the collocation type is noun + noun. The target words are randomly picked from the most frequent 1000 nouns extracted from all noun + noun collocations, and otherwise from the standard 1000 wordlist described in Section 3.4.3.

**The number of collocations to use** determines the size of an exercise. In Collocation Guessing, the larger the number the easier the exercise, because learners are able to see more hints. For the other exercises, balance is necessary because learners may be overwhelmed by the information presented. Learning is an accumulative process: sometimes less is more.

**How to select collocations** determines whether to use the most frequent collocation(s) or to select one or some randomly from the *n* best collocations (see below).

The ***n*** best controls the number of collocations to be used as candidates, from which one or more collocations are randomly picked for an exercise (see below).

### Selecting collocations for an exercise

Finally, we address the question of how to create an exercise that uses the best group of collocations and also allows learners to practice a variety of collocations associated with a particular word. This section describes two collocation selection principles that apply to all activities, although the specific algorithms vary slightly from one to another.

#### 1. The *n* best collocations

Most words—particularly common ones such as *take*, *make*, *cause*—have many collocations that can be grouped together by frequency range. The top group of one or two collocations is normally at least twice as frequent as the others. A second group with various numbers of collocations follows, and so on. It is important for learners to study collocations in the first group, but also in the second or third groups in order to expand their collocation knowledge. We select the *n* best collocations for a word and randomly pick one for each exercise. This explains why learners can practice different groups of collocations by clicking the *New Exercise* button. The value of *n* (default *n*=5) is given by the exercise designer and should be adjusted according to the frequency or usage of a particular word, or to the language ability of students. A general rule is to use a high value for common words or more advanced students.

#### 2. The most common collocations

In Collocation Matching and Related Words, learners match or differentiate collocations of two or more words. It is always possible that two words share the same group of collocates, e.g., *speak the truth* and *tell the truth*. Which is the best one to use? One option is to use both, which may not be desirable. Another is to select the strongest—in this case, *tell the truth*, because it is more frequent than *speak the truth*. Since a collocation is randomly picked for an exercise—the first principle—learners still have the chance to practise on *speak the truth* when another collocation is chosen for *tell*.

**Collocation Guessing**

Exercise name: make and take

Select a category: none

Configure parameters

Target words:  Entered by me  Randomly generated

Enter the target (or starting) word(s): make.take

Generate random words from: top 100

Number of random words: 10

Collocation Type: Verb+Noun

Sub-collection Type: The top 1000 words

Collocation position: 0

Number of collocations to use: 10

How to select collocations:  Randomly select from the n best  Always use the most common

n best: 15

	make	take	
make	<ul style="list-style-type: none"> <li>our way</li> <li>things</li> <li>a good decision</li> <li>any decisions</li> <li>no difference</li> <li>those changes</li> <li>best use of</li> <li>no sense</li> <li>a real effort</li> <li>any money</li> </ul>	<ul style="list-style-type: none"> <li>make our way</li> <li>make things</li> <li>make a good decision</li> <li>make any decisions</li> <li>make no difference</li> <li>make those changes</li> <li>make best use of</li> <li>make no sense</li> <li>make a real effort</li> <li>make any money</li> </ul>	<input checked="" type="checkbox"/> <input checked="" type="checkbox"/>
		<input type="button" value="Discard"/>	
take	<ul style="list-style-type: none"> <li>this effect</li> <li>necessary steps</li> <li>decisive action</li> <li>account of</li> <li>any part</li> <li>less time</li> <li>care of</li> <li>a long look</li> <li>my place</li> <li>advantage of</li> </ul>	<ul style="list-style-type: none"> <li>take this effect</li> <li>take necessary steps</li> <li>take decisive action</li> <li>take account of</li> <li>take any part</li> <li>take less time</li> <li>take care of</li> <li>take a long look</li> <li>take my place</li> <li>take advantage of</li> </ul>	<input checked="" type="checkbox"/> <input checked="" type="checkbox"/>
		<input type="button" value="Discard"/>	

Figure 6.20 Design interface for the Collocation Guessing activity

### Exercise design interface

The interface for dictionary-based activities resembles that for collection-based activities. We use the interface for Collocation Guessing, shown in Figure 6.20, as an example. Suppose a teacher creates an exercise that asks students to differentiate the words *make* and *take*. She specifies (1) several target words

(*make, take*) by entering them in the input box, or uses randomly generated ones by choosing a wordlist and the number of words to generate; (2) the collocation type to use (*Verb + Noun*); (3) the sub-collection type (*the top 1000 words*); (4) collocation position, i.e. which collocation constituent to practise (*the first word, e.g., the verb*); (5) the number of collocations to use in this exercise (*use 10 collocations*); and (6) how to select collocations (*randomly select from the 15 best*).

When the *Preview* button is clicked, CLS retrieves collocations from the WEB COLLOCATIONS collection that match the criteria specified. The teacher removes a word and its associated collocations from the exercise by clicking the *Discard* button or brings it back by clicking *Undo* button. Particular collocations can be discarded by unchecking the check box following them. For example, either *make a good decision* or *make any decisions* might be removed because they are similar.

## 7. Evaluating CLS

---

Chapter 6 described a collocation learning platform which automatically identifies collocations in text provided by teachers and students, using natural language processing techniques, and uses them to enhance the presentation of the text and also as the basis of exercises, produced under teacher control, that amplify collocation knowledge.

Three kinds of evaluation were conducted: identification of collocations by CLS vs. teachers, student use of the “cherry-picking” facility, and a theoretical evaluation of the potential use of CLS for improving student writing. We recruited three kinds of evaluator: senior language teachers, trainee teachers, and university students. Six teachers from the University of Waikato participated in the evaluation. One, from the School of Education, specializes in teacher training and computer-assisted language learning. Another is responsible for designing and organizing the online learning system for Waikato Pathways College, which specializes in academic literacy support for students. The remaining four are former language teachers, all of whom are currently studying to further their careers. These teachers helped to recruit students for participation in the student evaluations.

This chapter looks at automatically identified collocations to see whether teacher views of useful collocations coincide with those identified by CLS. Then it describes a trial of the “cherry-picking” facility in supporting academic writing. Finally, four trainee teachers were invited to examine the system, discuss its strengths and limitations, and explore its possible classroom use.

### *7.1 Collocation evaluation*

To assess the quality of collocations that CLS identifies in a given text, we examined ones generated from fifteen randomly selected articles and compared them with those manually identified by teachers.

### 7.1.1 Evaluation texts

To create baseline data, three types of text were used:

1. general English reading text: 20 articles originally published in *Password*, a magazine for new speakers of English<sup>21</sup>
2. IELTS reading text: 12 reading articles aimed at international students studying an IELTS course in New Zealand
3. academic reading text: 12 abstracts from PhD theses, prepared by a teacher as reading material for her students' Masters study.

For each text type, five articles were randomly chosen and given to teachers. The general English text contains 1911 running words, with an average length of 382 words per article. The IELTS and academic articles are slightly longer, with an average length of 772 and 628 words respectively. Baseline collocations were extracted using the algorithm described in Section 6.2.2.

Table 7.1 gives the number of collocations generated by the system, organized by syntactic pattern: 122 from the general English text and three times more from the other two. Adjective + noun and verb + noun collocations are the most common in the general English text. Noun + noun, adjective + noun and verb + noun form the majority in the IELTS and academic articles. Furthermore, the academic text contains a high proportion of noun + *of* + noun collocations. The verb + adjective pattern is least common across all three texts. The IELTS text contains slightly more verb + verb, adverb + adjective, verb + adverb, and adverb + verb collocations.

### 7.1.2 Investigating collocations that are identified

Before being given to teachers, collocations were manually examined and three problems were identified: tagging errors, partial collocations and incorrect chunking. Table 7.2 shows the counts and some examples. As discussed in Section 4.3, tagging errors are unavoidable in any natural language processing. They occur more frequently in academic text (14 errors), where sentence structures tend to be complex and often comprise multiple clauses and complex

---

<sup>21</sup> <http://www.password.org.nz/>

**Table 7.1 Collocation statistics for evaluation texts**

	<b>general</b>	<b>IELTS</b>	<b>academic</b>
number of articles	5	5	5
running words	1911	3862	3140
words per article	382	772	628
noun + noun	12	78	96
noun + verb	9	36	52
noun + of + noun	8	31	48
adjective + noun	30	144	119
verb + noun	45	88	53
verb + verb	10	15	3
adverb + adjective	1	8	4
verb + adverb	2	7	3
verb + adjective	1	1	2
adverb + verb	4	6	3
total	122	414	383
average per article	24	82	76

noun phrases. In the examples shown in Table 7.2, the verbs *highlight* and *light* were incorrectly identified as noun and adjective respectively. No errors were encountered in the general English text, mainly due to the simplicity of sentence structure—most sentences comprise only one clause. All tagging errors relate to multi-class words—14 mis-identified nouns and verbs, and five adjectives and verbs.

Partial collocations occur across all three texts, but particularly in the IELTS text. They are caused by incompleteness of the syntactic patterns defined for each collocation type. According to the patterns given in Table 3.6, *description of teaching* matches noun + of + noun, *examines the teaching* matches verb + noun and *children learn* matches noun + verb. These are problematic. The verb + noun pattern is intended to discover collocations like *make a difference*, *save time*, and *cause problems*. However, it also captures partial collocations like *prevent heat*, *make students*, and *keep light*.

Incorrect chunking, which may or may not be the result of tagging errors, occurs when a chunk crosses the natural boundary of a clause. It occurs frequently in collocations containing nouns. In example 1 of the third row of Table 7.2, *the strategies families* is mistakenly identified as a noun + noun collocation, despite the fact that *families* is the subject of the second clause—*families used to promote*

**Table 7.2 Problems associated with automatic collocation identification**

	general	IELTS	academic	examples
tagging errors	0	5	14	<p>1) <i>The descriptions and explanations reported in this study <b>highlight</b> the complexities of teachings.</i></p> <p>2) <i>If possible, you should have a separate light switch for every light, to prevent having to <b>light</b> unused areas.</i></p>
partial collocations	2	19	6	<p>1) <i>description of teaching (and learning processes)</i></p> <p>2) <i>examines the teaching (and learning processes)</i></p> <p>3) <i>children learn (and develop literacy expertise)</i></p> <p>4) <i>prevent heat (escaping)</i></p> <p>5) <i>make students (more aggressive)</i></p> <p>6) <i>keep light (clean)</i></p>
incorrect chunking	1	3	3	<p>1) <i>the present research focused on <b>the strategies families</b> used to promote home language learning in oral and written form.</i></p> <p>2) <i>some people make <b>a living researching</b> the family histories of others.</i></p>

*home language learning in oral and written form.* In example 2, *a living researching* is captured as a noun + noun collocation because *researching* is marked as noun.

### 7.1.3 Teacher's selection and judgment of collocations

Two teachers volunteered to scrutinize automatically identified collocations. Teacher A is an Education Faculty member involved in teaching graduate classes and postgraduate supervision. She specializes in discourse analysis and also has a specific interest in computer-assisted language learning and its potential for supporting academic literacy development. Teacher B is a learning support senior tutor, online coordinator and teacher at Waikato Pathways College and specializes in the teaching language online.

For each text type, these two teachers were given five articles (on paper) to mark collocations that they thought were worth learning. They were free to highlight any word sequences, of any length. The collocations they identified were counted and categorized manually. Those identified by CLS, but not by the teachers (not

**Table 7.3 Statistics of collocations identified by teachers**

<b>Teacher A</b>		si	tm	ac	so	rp	ac and rp
General text	122	54	31 (25%)	91	72 (79%)	84%	
IELTS	414	230	167 (40%)	246	201 (82%)	89%	
Abstracts	383	243	212 (55%)	171	136 (80%)	91%	
average							88%
<b>Teacher B</b>		si	tm	ac	so	rp	ac and rp
General text	122	71	41 (34%)	81	13 (16%)	44%	
IELTS	414	226	170 (41%)	244	73 (30%)	59%	
Abstracts	383	213	182 (48%)	201	89 (44%)	71%	
average							58%

si: system-identified collocation, tm: teacher-marked-collocation, ac: agreed-collocation, so: system-only-collocation. rp: reapproved-collocation.

including tagging errors, partial, and incorrect chunking related collocations), were given back to the teachers to review.

Table 7.3 summarizes the number of collocations (1) that were identified by the system (system-identified collocation, si), (2) that were marked by the teachers (teacher-marked-collocation, tm), (3) that teachers and the system agreed on (agreed-collocation, ac), (4) that were identified solely by the system (system-only-collocation, so), and (5) that were reapproved by the teachers (reapproved-collocation, rp) after reviewing.

There are overlaps in the system identified collocations. Take *provide a critical examination* as an example; there are two system-identified-collocations: *provide a critical examination* (verb + noun) and *a critical examination* (adjective + noun). If the teacher highlighted *provide a critical examination*, two teacher-marked-collocations were counted. If just *a critical examination* was marked, one teacher-marked- and one system-only-collocation were counted. If *provide a critical examination* was not marked at all, two system-only-collocations were counted.

The two teachers identified a similar number of IELTS collocations (230 vs. 226). Teacher A marked more academic collocations (243 vs. 214), and teacher B more general collocations (71 vs. 54), indicating their different interests and experience. CLS identified twice as many collocations as the teachers did. However, they did not always agree with it. Teacher A approved 55% of the collocations in academic

text, and the rate dropped below 50% for IELTS and general text. She reapproved about 80% of collocations after reviewing. In contrast, teacher B's approve and reapprove rates were fairly low, with an average of 40% and 30% respectively. She particularly disapproved of the collocations the system identified (16%) in the general text. In total, teacher A's approve rate was 88%, and teacher B's was 58%.

Table 7.4 and Table 7.5 summarize the collocations identified by teachers, but not by the system. Teacher A covered 19 categories and teacher B 12; they share ten (in bold). 23% of teacher A's collocations are in the form of preposition + noun—a preposition, optional adjectives, plus one or more nouns. 35% are phrases with various length; some are complete sentences such as *what are you waiting for*, *what most of us don't realize*, and *what will you tell them?*

Of teacher B's collocations, 82% take the form of preposition + noun, noun + proposition, verb + preposition, and verb-*ing* + noun. The verb + preposition (or particle) pattern is also called phrasal verb: “an English verb followed by one or more particles where the combination behaves as a syntactic and semantic unit.”<sup>22</sup> She pointed out that correct use of phrasal verbs such as *take off*, *cool down* and *bottom up* always present great challenges to language learners.

Other collocations not covered by CLS are:

- noun phrases: noun + *and* + noun, noun + *or* + noun, noun + noun + *and* + noun, and noun + adjective + noun
- verb phrases: verb-*ing* + *to* + verb, verb-*ing* and verb-*ing*, verb-*ing* + noun
- noun + *to* + verb
- adjective + *to* + verb
- verb + *and* + verb.

---

<sup>22</sup>defined by WordNet 3.0 at <http://wordnetweb.princeton.edu/perl/webwn?s=phrasal%20verb>

**Table 7.4 Collocations identified by teacher A, but not by the system**

<b>type</b>	<b>count</b>	<b>example</b>
<b>preposition + noun</b>	26	<i>at home; in childhood; in particular; over a year; throughout the year; at the rate of ; during busier periods; during the day; cross multiple sites; in school and community; over a year; from the sun; with similar use; towards the sun; during the warmer summer months; in the past decade; half the population; in many countries; at the same time; during the school day; in many cases; during quieter years; on the internet; despite the wishes; in side the safety of; about your problems</i>
<b>noun + preposition</b>	4	<i>the link between; attitude towards; opportunity for; positive interactions between</i>
noun + preposition + noun	3	<i>goals for future; immigrants in a new country; study for a qualification</i>
<b>verb + preposition</b>	4	<i>lead to; transfer to; prefer to; travel by car</i>
<b>noun + and + noun</b>	3	<i>documentation and analysis; beliefs and attitude; reading and writing</i>
<b>noun + noun + and + noun</b>	5	<i>formation; reproduction and transformation; data collection and analysis; home, school and community</i>
noun + adjective + noun	1	<i>energy efficient home</i>
noun + noun + noun	1	<i>data collection tools</i>
noun + or + noun	1	<i>major roads or airports</i>
possessive noun + noun	2	<i>children's development; parent's perceptions</i>
adjective + and + adjective	1	<i>oral and written</i>
verb + adverb + adjective	1	<i>become more aggressive</i>
verb + to + verb + and + verb	1	<i>learn to read and write</i>
<b>verb-ing + to + verb</b>	1	<i>declining to confirm</i>
verb-ing and verb-ing	2	<i>attracting and retaining; stimulating and nurturing</i>
<b>verb-ing + noun</b>	6	<i>making connections; creating opportunities; knowing the truth; hitting the beaches; leading to an improvement; keeping bees</i>
<b>adjective + to + verb</b>	4	<i>necessary to control; reluctant to change; willing to sacrifice; keen to attract staff</i>
<b>noun + to + verb</b>	2	<i>the ability to speak fluently; the first thing to consider</i>
<b>phrases</b>	39	<i>in relation to; the extent to which; despite the fact that; what is expected of you; from the point of view; when not in use; we don't know why; one of the main advantages; what are you waiting for; in line with; there had been no consultation; more likely to; for up to seven days; what most of us don't realize; if possible; when not in use; a trend which suggests; particularly among; it is acknowledged; it is vital that sb take action; there are many pressures on; because of other commitments; one of the main advantages; worry too much about; it is worth bearing in mind; get off to a good start; by word of mouth; what are you waiting for; swing to and fro like a; have no place in their busy lives; it is easy to connect to; what will you tell them?; the culture is different; life is hard; it will be better than; say them out loud; I don't have much confidence; New Zealand life; for at least five days</i>
<b>total</b>	110	

**Table 7.5 Collocations identified by teacher B, but not by the system**

Type	count	example
<b>preposition + noun</b>	17	<i>in particular; in conclusion; with the view; for fun; throughout the year; at the same time; in winter; in summer; in use; for example; in the workforce; in school; on land; in the bush; for the future; from memory; for the experience</i>
<b>noun + preposition</b>	13	<i>the connection between; interrelationship between; interaction between; consultation with; year ago; the battle against; a ban on; a stain on; benefit for; parents with children; speed of up to; centuries ago; instructions in English; important links between</i>
<b>adjective + preposition</b>	4	<i>significant in; useful for; deemed fit to; enthusiastic about</i>
<b>verb + preposition (or particle)</b>	32	<i>trying out; growing up; contributed towards; focus on; draw from; bottom up; grow up with; made from; heat up; cool down; take off; resulting in; weighing in; heading for; going up; showing up as; steer somebody towards; add up to; take off; cut back from; confused by; grow up to; get off to; relying on; opting for; to live with; provided by; race along; starting with; drift down; connect to the past; go back</i>
<b>noun + and + noun</b>	8	<i>home and school; beliefs and attitude; bilingualism and multilingualism; ideas and beliefs; day and times; learning and development; documentation and analysis; beliefs and attitude</i>
<b>noun + (and )+ noun + (and) noun</b>	4	<i>assimilation and accommodation adaptation; information and communication technologies; home language and culture; bilingualism and language learning</i>
<b>noun + to + verb</b>	1	<i>freedom to decide</i>
<b>adjective + to + verb</b>	1	<i>keen to point out</i>
<b>verb-ing + to + verb</b>	1	<i>racing to find</i>
<b>verb and verb</b>	2	<i>describes and explains; read and write</i>
<b>verb-ing + noun</b>	15	<i>making connections; creating opportunities; knowing the truth; applying the job; improving your English; creating possibilities; hitting the beaches; taking its toll; showing sign of; working from home; approaching retirement; applying for jobs; bearing in mind; searching on the Internet; keeping bees</i>
<b>phrase</b>	19	<i>in relation to; in terms of; what constituted success; in light of; the extent to which; thought that goes into; a build up of; back at work; because of ; for all concerned; per hour; swing to and fro; Roman alphabet; nineteenth century; United Kingdom; United States; on the other hand; life is hard; out loud</i>
<b>total</b>	117	

**Table 7.6 Collocations that the teachers did not approve**

<b>Teacher A</b>			
		<b>count</b>	<b>examples</b>
General text	uncommon	3	<i>adventure junkies</i>
	common	3	<i>small parachute, just like to, small bridge</i>
	free	16	<i>a bit of craziness</i>
IELTS	uncommon	17	<i>caning ban, low-angled sun, lower wattage bulbs</i>
	common	6	<i>gym session, increasingly rarer, job content</i>
	free	33	<i>keep rooms warmer, flexibility was listed as, choice of hours</i>
Abstracts	uncommon	12	<i>enact agency, complex articulation, marginalized situation</i>
	common	1	<i>rarely translated to</i>
	free	8	<i>highlight the multiple pathways, initiatives were afforded, activity were constructed</i>
total		89	
<b>Teacher B</b>			
General text	uncommon	7	<i>black-water rafting, rock faces, active holiday</i>
	common	29	<i>exciting adventure, young people, really good</i>
	free	37	<i>follow the river, want to stay, a lot of reading</i>
IELTS	uncommon	28	<i>lightweight materials, avid bushwalker, obesity summit</i>
	common	56	<i>school tradition, new homes, important thing</i>
	free	101	<i>disagree with the government, try to avoid, introduction of the ban</i>
Abstracts	uncommon	26	<i>school context, linguistic habitués, local level strategies</i>
	common	20	<i>small group, large city, present research</i>
	free	47	<i>the thesis begins in, promote the retention, the collection of documentation</i>
total		351	

The collocations that the teachers did not approve fall in three categories:

- uncommon combinations
- common combinations
- free combinations.

For each category, Table 7.6 gives the counts and some examples.

Of the total number of collocations not approved, 21% are uncommon combinations, most from IELTS and academic text. Some are particularly topic specific, which make them less useful in some sense—for example, *caning reform*, *control unruly students*, and *caning ban* from an article about “Caning in Thai Schools.” Others relate to the author’s choice of words and writing style—for example, *obese Australians*, *mail monopoly*, and *chubby stars*. The teachers did not recommend these combinations to their students, because they would not use

them themselves. Some combinations from academic text are extremely rare and may in fact be the creation of the author. Ten of them were tested using the WEB PHRASES collection, and their frequencies are given in Table 7.7. Among the first six, which do not occur in that collection, *multilingual social cohesive communication, interethnic and interracial family, intercultural identity formation* are rare technical terms; *enact agency, reframing of understandings* and *the life of class* are unusual combinations that few native speakers would use. The frequency of other four is low, despite the fact that they are made up of relatively common words.

Some uncommon collocations can be removed during the collection building process using the frequency cut-off value described in Section 6.2.1. However, we did not do so, because that value is adjustable according to the teacher or student's need, and we wanted to evaluate the performance of the collocation extraction algorithm without the interference of this variable.

Common combinations account for 26% of collocations not approved by the teachers. They are of the adjective + noun and adverb + verb type, and include extremely common words such as *new, good, really, things, small, young* and *large*. The teachers argued that combinations like *small parachute, young people, really good* should be treated as weak collocations, and therefore not worth deliberate learning.

Half the collocations not approved are free combinations and the majority are of the verb + noun and noun + verb types; the remainder are verb + *to* + verb and noun + *of* + noun. Sentences normally start with a noun phrase and then a verb + noun phrase, so verb + noun and noun + verb collocations constitute the core structure. Consequently, some system-identified verb + noun and noun + verb combinations are just free combinations of a verb and noun or vice versa. It is difficult for a computer program to judge the collocation strength of *pay the price* versus *pay the bills*: both are extremely common, with a frequency of 270,000 and 200,000 in WEB PHRASES respectively, but the first is idiomatic while the latter is a free combination. For the same reason, the teachers did not approve some verb + *to* + verb collocations that comprise common verbs (such as *try, want, need* and

**Table 7.7 Frequency of uncommon combinations in WEB PHRASES**

uncommon word combinations	frequency
<i>enact agency</i>	0
<i>multilingual social cohesive communication</i>	0
<i>interethnic and interracial family</i>	0
<i>intercultural identity formation</i>	0
<i>reframing of understandings</i>	0
<i>the life of class</i>	0
<i>forces mediate</i>	96
<i>local level strategies</i>	190
<i>cultural monitoring</i>	190
<i>marginalized situation</i>	210

**Table 7.8 Collocations not approved by teacher A**

*constant comparison, reframing of understandings, intercultural identity formation, lower-wattage bulbs, choice of study majors, starts with a string CV, work out, adventure junkies, a bit of craziness*

*help* as in *try to use, want to stay, need to help, and help to repair*), or noun + of + noun collocations that start with common of phrases (such as *percentage of, name of, position of, and bottom of*).

Of 89 collocations not approved by teacher A, teacher B shared 80. The nine collocations they did not agree on are given in Table 7.8. Both thought that adverb + verb collocations are not particularly useful. Teacher B tended to reject verb + noun and noun + verb collocations. The fact that teacher B's rate of rejection is four times higher than that of teacher A (351 vs. 89) indicates that they have a different view of what collocations are. As teacher B pointed out in her notes, she was trying to distinguish between idiomatic combinations and merely convenient grammatical constructions. On the other hand, teacher A focused on what her students would find useful to learn. She approved 90% of common collocations that teacher B did not. She pointed out that ones like *exciting adventure, similar uses, and minor mistakes* are useful for lower level students, and that noun + verb collocations in academic text are good for thesis writing.

Overall, the result of the evaluation was positive. CLS identified a significant number of collocations, ranging from 6% of the word count (24 collocations for a

370-word general article) to 12% of the word count (76 collocations for a 628-word abstract). Of 919 collocations identified by CLS, 6% were clearly incorrect. Of the remainder, one of the teachers (teacher A) approved 88%. The other teacher clearly had a different notion of what collocations are, which highlights the subjective nature of the task.

## 7.2 Evaluating the “cherry-picking” facility

This section describes an evaluation of the “cherry-picking” facility with eight students and their teacher. Recall that “cherry-picking” is the facility for students to collect collocations while reading the text. The goal of the evaluation was to obtain feedback from real users in an authentic context. The evaluation comprises three phases:

- testing collocation knowledge
- using CLS to collect collocations
- having students write a literature review with the collocations they collected.

They are introduced in the sections below.

### 7.2.1 Background

The evaluation was incorporated into a voluntary course in which the teacher provided students with support when writing a literature review for their Masters thesis proposal. Because the students were non-native speakers, the teacher focused not only on the content and organization of their proposals, but also on academic language aspects of their text. The teacher believed that studying collocations related to the student’s research topics helps him write more professionally. She intended to use this evaluation to introduce the concept of collocation, and provide a means by which students could improve their own writing in the future.

Eight students aged from 25–30, from Samoa, Cambodia, and the Solomon Islands, participated in the series of classes, once a week, across the semester. They could withdraw at any time if they decided not to proceed to their Masters by thesis. Their study topics were related to *educational leadership, curriculum*

**Table 7.9 Statistics of collocations used in the evaluation**

<b>topic</b>	<b>educational leadership</b>	<b>curriculum and change</b>	<b>literacy and bilingualism</b>
number of articles	10	12	14
running words	4215	5765	6877
noun + noun	156	215	296
noun + verb	67	89	116
noun + of + noun	108	115	119
adjective + noun	251	329	374
verb + noun	119	107	149
verb + verb	11	15	16
adverb + adjective	6	10	9
verb + adverb	12	9	10
verb + adjective	3	7	4
adverb + verb	10	11	13
<b>total</b>	<b>743</b>	<b>907</b>	<b>1106</b>

*and change*, and *literacy and bilingualism*. The teacher helped them formulate research questions and proofread what they had written. They were given two months to read related literature and one month to write a draft literature review.

In order to help students build up collocations related to their research area, the teacher collected 36 abstracts from PhD theses and used CLS to build three collections, each containing ten to fourteen articles. Table 7.9 shows the number of articles, the number of running words and extracted collocations. Noun + noun, noun + *of* + noun and adjective + noun are the dominant collocation types. On average, a 460-word abstract contains about 22 noun + noun, 11 noun + *of* + noun, and 31 adjective + noun collocations.

### 7.2.2 Testing collocation knowledge

At the beginning of the course, two tests were conducted to give an indication of the student's vocabulary size and collocation knowledge.

Vocabulary was tested using the Nation and Laufer Levels Test (1999) available on the *Compleat Lexical Tutor*.<sup>23</sup> The test comprises five levels, each with 18 fill-in-the-blanks questions in which students are asked to complete words with some characters missing. Words are chosen from the 2000, 3000, 5000, 10,000 and

<sup>23</sup> <http://www.lextutor.ca/tests/levels/productive>

**Table 7.10 Students' vocabulary test scores**

<b>student</b>	<b>2000</b>	<b>3000</b>	<b>5000</b>	<b>10,000</b>	<b>AWT</b>
A	77%	66%	55%	55%	66%
B	77%	61%	33%	11%	37%
C	<b>88%</b>	<b>88%</b>	66%	55%	61%
D	<b>88%</b>	66%	55%	35%	72%
E	<b>94%</b>	77%	<b>88%</b>	77%	<b>88%</b>
F	<b>94%</b>	66%	44%	33%	72%
G	72%	72%	55%	22%	66%
H	77%	44%	22%	16%	27%

academic wordlists. According to the specification, the pass threshold is 83% at each level; otherwise students are suggested to build up this level by working on the corresponding wordlist.

Table 7.10 shows the scores of each student on the five levels. All students did relatively well on the 2000 wordlist. Only one student passed the 83% threshold on 2000, 5000 and academic wordlists. No one reached 83% on the 10,000 wordlist. Student D and F, who performed poorly on the 5000 wordlist, achieve a better score on the academic wordlist, indicating that they have accumulated a certain academic vocabulary.

To examine the students' collocation knowledge related to their research area, they were given three minutes to brainstorm keywords and five minutes to brainstorm a list of collocations related to their study. Appendix E gives the full list of keywords and collocations. Of the total number of 111 collocations, only 10% constitute more than two words. Apart from five noun + *of* + noun collocations, the remainder are dominated by the noun + noun and adjective + noun types. Collocations produced by some students (for example, B and D) are diverse, covering a wide range of topics, while those of others are similar and narrow (for example, C and H).

**Table 7.11 Number of keyword and collocations produced by students**

student	keyword				collocations			
	total	2000	AWT	other	total	2000	AWT	other
A	16	44%	17%	39%	15	47%	40%	13%
B	27	52%	40%	8%	22	63%	32%	5%
C	13	61%	23%	16%	16	70%	22%	8%
D	15	20%	53%	27%	18	58%	25%	17%
E	-	-	-	-	13	63%	20%	17%
F	15	40%	33%	27%	17	64%	14%	22%
G	-	-	-	-	-	-	-	-
H	9	67%	11%	22%	10	70%	20%	10%
average	16	47%	30%	23%	16	64%	24%	12%
identified collocations	-	-	-	-	-	56%	33%	11%

(Note: Student G was absent that day, and student E did not produce keywords)

**Table 7.12 Results of Fill-in-Blanks tests**

student	receptive exercise	productive exercise
A	17 (85%)	4 (20%)
B	13 (65%)	5 (25%)
C	11 (55%)	7 (35%)
D	13 (65%)	7 (35%)
E	17 (85%)	3 (15%)
F	15 (75%)	5 (25%)
G	17 (85%)	5 (25%)
H	12 (60%)	5 (25%)
average	14 (72%)	5 (25%)

Table 7.11 gives the number of keywords and collocations produced by students, along with the vocabulary profile generated using the Vocabulary Profiles tool available on *Compleat Lexical Tutor*.<sup>24</sup> The first row shows that student A produced 16 keywords, 44% from the 2000 wordlist, 17% from the academic wordlist and 39% that do not appear on the lists. On average, each student brainstormed 16 keywords, and about half of which were on the 2000 wordlist. The same number of collocations was produced (16), about 60% from the 2000 wordlist. The last row gives the vocabulary profile of collocations from the three collections discussed above. Compared to the collocations produced by students,

<sup>24</sup> <http://www.lextutor.ca/vp/eng/>

ones from abstracts represent more academic words (33% vs. 24%) and fewer common words (56% vs. 64%).

In the three collections prepared for the students, the teacher created two Fill-in-Blanks exercises (Section 6.4.1), each containing 20 questions and focusing on noun + noun, noun + verb, adjective + noun, noun + *of* + noun collocations (see Appendix F). One tested receptive collocation knowledge: answers were available. The other tested productive knowledge: students had to provide their own answers. Table 7.12 shows the scores of each student on each exercise. They performed well on receptive exercises, with an average of 14 correct answers. In contrast, an average of 5 correct answers in the second exercise indicated that their productive knowledge in their research area was limited, which will inevitably hamper their writing ability.

### 7.2.3 Collecting collocations

Students were given a one hour tutorial on how to cherry-pick collocations from the three collections built for them. Then they spent one hour identifying collocations that might be of use for writing a literature review. They were asked to collect at least 100 collocations and print them out. Then they wrote a literature review and were asked to highlight any uses of the collocations.

A cherry basket comprises a list of collocations and illustrative text that constitutes the sentence containing the collocation, and the preceding and following sentence (Appendix G gives an example). Table 7.13 summarizes the number of collocations selected by the students (grouped by type), their average length (in words), and their frequency in the WEB PHRASES collection.

Each student collected around 100 collocations, 92% of which are of verb + noun, noun + noun, adjective + noun and noun + *of* + noun types. They collected a few more noun + verb collocations writing, but showed little interest in the other five collocation types. It is understandable that verb + adverb, adverb + verb, adverb + adjective, verb + *to* + verb and verb + adjective are not common in academic text. Most of the collocations picked constitute more than two words (2.79 words per collocation). On average, the collocations occur 113,000 times in the WEB PHRASES collection. Together, this indicates that the students tended to pick long

**Table 7.13 Statistics of collocations collected by the students**

student	total	vn	nn	an	non	nv	vr	rv	ra	vv	va	length	frequency
A	94	17	23	33	14	4	2	1	0	0	0	2.77	55,000
B	103	18	27	44	12	1	0	0	0	1	0	2.52	146,000
C	106	26	20	36	15	4	1	2	0	2	0	2.89	91,000
D	95	21	14	38	11	8	3	0	0	0	0	2.89	128,000
E	123	32	35	34	12	4	0	0	2	0	0	2.92	55,000
F	110	25	26	33	14	4	3	1	1	1	0	2.96	75,000
G	112	16	15	39	27	4	1	1	1	0	0	2.90	95,000
H	114	24	26	42	14	2	1	1	1	0	0	2.52	260,000
average	105	22	23	37	15	4	1	1	1	1	0	2.79	113,000

vn: verb + noun; nn: noun + noun; an: adjective + noun; non: noun + *of* + noun;  
 nv: noun + verb; vr: verb + adverb; rv: adverb + verb; ra: adverb + adjective;  
 vv: verb + to + verb; va: verb + adjective

and common collocations. Vocabulary size seems to play a part in what kinds of collocation students prefer. For example, the weaker student H's collocations are particularly frequent compared to those of other students.

When cherry-picking collocations, only one student used the Category feature (Section 6.3.3). He grouped his collocations into *teaching and learning*, *bilingualism*, *culture value*, *literacy* and *other* categories. In general, collocations collected by students can be divided into two groups: topic specific and academic. Table 7.14 shows some examples for the most popular five collocation types. Apart from collocations related to their research topic, students also collected ones that are useful for any academic writing, such as *study concluded that*, *findings indicated that*, *nature of the research*.

#### 7.2.4 Results

In the end, only two students submitted the literature review; one did not use any collocations he collected; the other six failed to complete this aspect of the workshop series. As a result, the only text used for analysis was from one student, who wrote about a 700-word literature review with five collocations highlighted. Table 7.15 shows excerpts from the text with collocations highlighted in bold. They are:

- two verb + noun: *engage in innovative practices* and *improve the quality of education*

**Table 7.14 Collocations collected by students**

	<b>topic specific</b>	<b>academic and research specific</b>
adjective + noun	<i>environmental education curriculum sustainable development inadequate reading skills</i>	<i>informal observations research literature epistemological framework</i>
noun + noun	<i>management approaches school culture language shift</i>	<i>key assumption research literature focus groups</i>
noun + of + noun	<i>lack of policy direction effectiveness of the curriculum complexities of teaching</i>	<i>reliability of the research analysis of these data nature of the research</i>
verb + noun	<i>achieve educational change overcome the entrenched culture of translate into practice</i>	<i>look at local perspectives underpinned the study investigate ways</i>
noun + verb	<i>curriculum was developed with reform was implemented through skills were transferred to</i>	<i>study concluded that findings indicated that study is embedded in</i>

- two adjective + noun: *professional development programs* and *educational reforms*
- one noun + noun: *teacher self-efficacy*.

The teacher thought using these five collocations did make this student's writing more fluent and native-like.

### 7.2.5 Questionnaire

Two students and the teacher independently filled out a questionnaire with eleven questions (see Appendix H and I). The questionnaire aims to

- evaluate the student's understanding of the concept of collocation after using CLS,
- understand why students used fewer collocations than expected, and
- gather comments and suggestions for improving CLS.

After using the system, both students realized the importance of collocation knowledge in academic writing. In their words, collocation knowledge is helpful in that “text will make sense and [be] grammatically correct” and “provide clarity and more meaning to the pieces of writing.”

**Table 7.15 Use of collected collocations**

*However, there was evidence of some teachers who engage in innovative practices in science teaching.*

*What this implies is that teacher self-efficacy in teaching in rural schools like being innovative can be improved...*

*One of the ways to address problems ...is to involve them in professional development programs that will build up their capacity of pedagogical content knowledge to improve the quality of education.*

*... to meet the requirements of the new educational reforms.*

They explained why they did not make full use of the collocations they collected. First, their collocations were not particularly useful because they both changed research topic after collecting them. In their comments on how to improve the system, both students suggested that it should provide materials that are sufficiently close to their study. Second, when collecting collocations they did not know what they really needed for writing a literature review. They stated that they would do things differently if they were to do it again. Third, they were under time pressure to finish their writing: going through 100 collocations to find appropriate ones is not an easy job.

Finally, the students made positive comments about the system: it is easy to use and could help them improve their collocation knowledge if used regularly.

From her own observations during the evaluation and while helping students review drafts of literature reviews, the teacher was convinced that using CLS could help students identify collocations that they may have difficulty discovering on their own. Most importantly, this evaluation introduced the concept of collocation to the students and raised their awareness of this language phenomenon. She pointed out that there was evidence that the students either used the collocations they collected directly, or made acceptable changes before incorporating them into their literature review.

The teacher believed that students can benefit from using the system regularly. She proposed ways to prompt usage and increase motivation: (1) give positive feedback if they use a collocation a certain number of times; (2) compare the text they write with and without using the system, which will make them realize that by using collocations their text can appear more native-like.

The teacher thought that two factors contributed to the lack of the use of the collected collocations: unfamiliarity with the system, which was confirmed by feedback that the students would like to redo the “cherry-picking,” and the lack of feedback and monitoring.

With respect to how to improve the system, the teacher suggested linking collocations to other resources, such as domain specific glossaries.

Finally, the teacher remarked that (1) the human element is another factor that could make the system more or less useful, (2) less technical or difficult material can be built into the system so that it can be used by other learners, not just those doing academic study, and (3) some collocations identified by the system were not recognized by her instinctively.

### 7.2.6 Discussion

We recognize that this evaluation was conducted with a small group of students. It lasted four months and the students’ progress was not closely monitored because the course was not compulsory. In the end, two students submitted literature reviews, and only one used collocations he collected. However, both students and teachers confirmed that they have seen the value of collocation knowledge in academic writing and will continue to use CLS in the future.

## 7.3 *Theoretical evaluation with language teachers*

The effectiveness of the individual collocation activities was deemed to lie beyond the scope of this thesis, emphasis of which is the evaluation of users in particular contexts. Instead, a “theoretical” evaluation was carried out, in which a group of teachers examined the strengths and limitations of CLS, judged its compliance with language theory, and explored its use in the classroom.

### 7.3.1 Background

The evaluation was embedded into an online postgraduate study course *Second Language Learners and Learning in Mainstream Classrooms*, offered by the School of Education at the University of Waikato and available on the university's Moodle website. It covers second language acquisition theory and practice, and includes an examination of the nature, demands, and outcomes of language learning, different approaches to language learning, and introduces language learning theories.

Four former teachers, two Chinese and two New Zealanders, enrolled in this course. They had experience in second language teaching of Chinese, English and Te Reo Māori.

CLS was introduced to the teachers under the heading of language learning approaches. It was associated with the *content based and data driven learning* approach. Their lecturer used the system to build a collection that contains school journal materials from an adult literacy project. She prepared a user manual that contains step-by-step instructions, and corresponding screenshots, that introduces the “cherry-picking” facility, the Fill-in-Blanks exercise design interface for teachers, and an exercise created using the interface (see Appendix J).

Teachers were given an assignment titled “reflecting on the data driven language learning tool” that asked them to try out the system at their own time and pace, and post discussions of four questions (see below) on the course forum.

### 7.3.2 Results

Here is a summary of what the teachers thought: Appendix K gives the full discussion.

**Question 1: Do you understand how this tool works? What do you think the language learning principles are that the tool exploits?**

On the whole, the teachers were impressed by CLS and thought it was fun to use and should be attractive to students. The Māori language teacher thought that the system provided great input and feedback, and was keen on using it to teach Māori.

One teacher related the system to the repetition principle—repeated reading of the same text. She liked the links to other examples of the same collocation in different contexts, and suggested that the system could provide opportunities for students to actually use those examples, for example, by incorporating them into manipulative exercises.

One teacher connected the system with rote learning theory. She compared it with another language tool she had encountered before, and thought that CLS was more fun to work on.

**Question 2: What do you think the potential usefulness of the tool is for learners and teachers?**

The usefulness of CLS is summarized as following. For students, they can

1. do self-study at home,
2. create exercises for themselves for self-monitoring and self-evaluation, and
3. spend more time on learning because the system is fun to work on and easy to operate.

Teachers can

1. target the specific language and relate it to the context being studied,
2. save time when creating appropriate tasks to support learning,
3. reduce the complications that learners are exposed to by using the exercise parameters, and
4. spend more time with students who need extra help, while giving advanced students more freedom to study on their own—which is particularly useful in classrooms with large student numbers.

**Question 3: What do you think the potential limitations of the tool are for learners and teachers?**

The teachers focused on the limitations from their own perspective. First, the system should be used in conjunction with other activities such as listening and speaking, so that students can develop balanced language skills. Second, the teachers were concerned that the system may be limited by the range of text it provides and thought it could provide facilities to build their own bank of graded texts (note: the collection building facility described in Section 6.2 was not

finished at the time of evaluation). Third, one teacher pointed out that the available activity types were limited and students might guess the correct answer, which reduces the usefulness of exercises. Fourth, there is a potential mismatch in level between text and collocation examples. Fifth, although a printed summary of progress can be obtained, the system does not provide sufficient facilities for teachers to monitor a whole classroom of students within limited class time.

#### **Question 4: Could you think of using it? How and/or why?**

The teachers expressed their desire to use CLS in their classrooms. They commented that using input text to create exercises would challenge their students to process what they have learnt. They felt that the system provides “useful additional mileage and interaction with text.”

They thought that the system could be used in four ways:

1. practise target vocabulary,
2. revisit what students have learnt,
3. promote self-study, and
4. create catch-up work for students who miss classes, or students with low language proficiency.

One teacher proposed an interesting usage scenario: students collecting collocations and creating exercises for themselves, while the teacher adopts the role of guide and facilitator.

Finally, a few concerns were raised by teachers. First, the texts need to be appropriate for their students. Second, using the system might be a challenge for beginners. Third, some teachers and students may face a “computer literacy” problem. Fourth, some schools may not have facilities to run the system.

#### **7.3.3 Discussion**

This evaluation is anecdotal rather than quantitative. Four trainee teachers provided insightful comments and proposed many interesting ways to use CLS inside and outside the classroom. CLS has great potential to help teachers construct focused learning activities. It also can be a useful self-study tool for students to study language at their own pace. However, many issues need to be

addressed before such a system can be successfully incorporated into classrooms—for example, providing sufficient monitoring and feedback.

## 8. Conclusion

---

Collocations are one of the greatest challenges in second language learning. They are difficult to acquire because they are numerous and arbitrary. Printed dictionaries and online concordancers are useful resources, but the former are limited by physical size and the latter are not tailored to meet the needs of learners (Section 2.5). There is a wealth of language learning activities on the Web, but those specific to collocations are rare (Section 2.9.1). Despite widespread recognition of the importance of collocation learning, and the growing use of computers in second language learning, little research has been reported on computer-assisted collocation learning.

The goals of this thesis are to examine ways of presenting corpus data for effective collocation learning, and investigate how to construct a learning environment that helps learners systematically acquire collocation knowledge. To formalize these goals, two hypotheses were formulated:

- 1) Corpus data can be processed and organized in different ways to help learners expand collocation knowledge.
- 2) For a given collection of language learning text, pedagogically valuable collocations can be automatically identified and incorporated into an environment that facilitates the key learning activities of noticing, retrieval and generation.

In order to investigate these hypotheses in a concrete and constructive way, a computer-based collocation learning system called CLS has been constructed during the course of the investigation. It comprises two components: collocation resources and a learning platform. These substantiate the two hypotheses above, as reviewed in Sections 8.1 and 8.2.

### 8.1 Presenting corpus data for collocation learning

CLS's collocation resources contain rich material generated from a trillion word tokens in the form of *n*-grams. They were filtered by wordlists and syntactic constraints, and then organized into three digital library collections (Chapter 3):

- WEB PRONOUN PHRASES
- WEB PHRASES
- WEB COLLOCATIONS.

The first collection helps learners to locate pronoun phrases that contain a particular term in three ways: phrases that contain it, patterns that precede it and patterns that follow it. The second collection allows learners to explore free word combinations, unconstrained by grammatical class. The third collection houses a large volume of naturally occurring collocations, organized by syntactic pattern and ranked by frequency.

Chapter 4 evaluated the quality and quantity of WEB COLLOCATIONS. A comparison of five statistical measures on Web and BNC bigrams supported our decision to present the collocations in descending frequency order. The part-of-speech tagger achieved 80% accuracy on five-grams, which indicates that the impact of the restricted context that five-grams provide is mild and that collocations extracted from them are, in general, acceptable. WEB COLLOCATIONS contains far more collocations than those in the *Oxford Collocation Dictionary for Students of English*. However, it does contain mis-categorized collocations caused by tagging errors and inconsistency of word class assignment. Users need to be advised of this by the interface, and through training.

To examine the effectiveness of these resources in term of helping students expand collocation knowledge, we invited language learners to use CLS while writing (Chapter 5). For the WEB PRONOUN PHRASES collection, students were asked to write short descriptions about themselves to elicit pronoun use. The results demonstrate that the system helped students check grammatical errors, and generate, expand and confirm the text they wrote.

This collection only contains pronoun phrases, but self-expressions do not necessarily begin with pronoun words. It would be easy to add *my-*, *his-*, and *her-*

grams. Including *n*-grams that contain words that are commonly used to express interests, feelings, and emotions—such as *happy*, *sad*, *like*, *dislike*, and *angry*—is possible, but this wordlist would need to be manually selected and categorized by language instructors.

The WEB PRONOUN PHRASES collection interface provides access to other lexical resources—WordNet, Roget’s thesaurus, the Edinburgh Word Association thesaurus and Yasumasa Someya’s lemma list (Section 3.3.2). Unfortunately, their volume is rather small, and students found them unsatisfactory. WordNet does provide a few examples of how the words are used in context, but the other three only offer a list of single words. These resources were chosen for demonstration purposes, and can be easily expanded by allowing teachers to add more items, and including other resources.

Students also complained that pronoun phrases on the first result page were similar. It is true that particular language structures sometimes dominate the search results. For example, of the top 30 *I*-phrases that contain the word *like*, 23 relate to the *I would like to* pattern. One remedy is to remove phrases with similar structures, but the extent to which this should be done needs further investigation.

For the WEB COLLOCATIONS and WEB PHRASES collections, nine students from an IELTS writing preparation class were recruited. Each wrote an essay, in which teachers and we examined each language error and determined whether, in principle, CLS could help resolve it. Then we marked the position of the errors and asked the students to use the system to correct them. The results were extremely encouraging. Of a total of 108 errors, CLS could help resolve 95, and the students actually resolved 73 without any human assistance. In a majority of cases, the result was a clear improvement in their writing.

WEB COLLOCATIONS seems less successful than WEB PHRASES. One teacher complained that organizing collocations according to syntactic patterns confuses users because this made it difficult to identify useful collocations—some lower level students may not be familiar with the concept of word class. One solution would be to provide an interface that presents the most common collocations first, regardless of syntactic patterns. Taking the word *cause* as an example, the verb

*cause + noun* (*cause problems*, *cause damage*, *cause harm*), this would give high priority to the noun *cause + of + noun* (*cause of death*, *cause of action*, *cause of the problem*), and adjective + the noun *cause* (*the leading cause*, *a common cause*, *a major cause*).

In conclusion, to support the first hypothesis three digital library collections were built from Web *n*-grams to demonstrate how to process and organize corpus data to help learners expand their collocation knowledge. They proved to be useful and effective in helping students improve writing.

There are, of course, some limitations. All three collections are based on a historical dump of the Web, and have been further filtered: this falsely rejects some acceptable phrases—such as ones containing neologisms like *google*. To counter this, new words could be manually added into the wordlist used to filter *n*-grams. Furthermore, grammatical errors in Web text may confuse less advanced learners, and the situation is exacerbated when they occur frequently—for example, *may not suitable* occurs 602 times in the WEB PHRASES collection. Here, user training is needed.

Although Web text was used for the investigation, our work is not restricted to it—the same technologies apply to other corpora. In addition, other collections or sub-collections could be made. For example, one could focus on learning epistemic adverbs such as *certainly* or *probably*, identified by Biber (2006) as occurring frequently in university spoken and written language. This could be useful for students in English Study for Academic Purposes courses, and for those preparing for university study. Another might contain particular sentence heads—for example, sentences starting with the words *As*, *Despite*, or *With*—to help learners construct sentences. Last but not least, collections could focus on particular domains, such as quantification, to support theme or function-oriented vocabulary learning.

## 8.2 Constructing a collocation learning system

In the CLS collocation learning platform (Chapter 6), teachers build digital library collections from articles they have prepared for their students. Collocations are

automatically identified, and organized by syntactic pattern. Once the collection is constructed, learners interact through an interface specially designed for them to seek, study, and collect collocations. While reading the articles, their attention is drawn to highlighted examples. They recycle and consolidate what they have learnt through exercises that are generated from the content of the WEB COLLOCATIONS collection combined with the collections built by teachers. Students expand and enrich their knowledge by examining related collocations retrieved from WEB COLLOCATIONS, and by studying exemplary text in the BNC and live samples from today's Web.

CLS has undergone three evaluations (Chapter 7). The first compared collocations automatically identified with those manually selected by two teachers, of which one teacher approved 88% and the other 58%. Both teachers selected a large number of preposition collocations and free form phrases. However, they did not always agree with each other as to what a collocation is. The evaluation confirmed the subjective nature of collocation identification. In this sense, CLS can help teachers reflect on what they have noticed and what they have unthinkingly ignored.

This evaluation indicates that the ten collocation types that CLS covers (Table 3.6) are not equally effective: some are more useful for particular types of text, and for particular student groups. Noun + verb collocations from academic text—for example, *the research reveals, the finding indicated that, the study examines*—are useful patterns that students can apply in their academic writing. However, ones from general text may differ because most are free combinations. On the other hand, verb + noun collocations from general text help lower level students construct sentences. CLS does allow teachers to switch on/off certain collocation types (Section 6.2.1), but they need to be advised of this.

Although CLS focuses on lexical collocations, preposition-related collocations—such as preposition + noun, noun + preposition and verb + preposition—can easily be added, and presented in a separate interface to draw attention to prepositions.

There are other categories that are of great pedagogical value, but have been excluded by this thesis. Some contain non-adjacent items like *make up one's mind*,

*serve somebody well, chase somebody up.* Others are categories of lexical phrase as defined by Nattinger and Decarrico (1992). Lexical phrases include chunks of language of varying length: both short, relatively fixed phrases like *as it were, on the other hand*; and phrases with a fixed, basic frame with slots for various fillers, like *a \_\_\_\_ ago (a year ago, a month ago)*, and *the \_\_\_\_er X, the \_\_\_\_er Y (the higher X, the higher Y, the longer you wait, the sleepier you get)*. They differ from the collocations targeted by this thesis in that they are used to perform defined functions. For example, *I'll say* indicates agreement; *see you later* indicates parting; *as far as I know* is a qualifier.

Nattinger and Decarrico (1992) group them into four categories:

- polywords: short phrases that function much like individual lexical items, for example *by the way*.
- phrasal constraints: short- to medium-length phrases, for example *as I was saying*.
- institutionalized expressions: lexical phrases of sentence length, usually functioning as separate utterances, for example *how do you do?*
- sentence builders: lexical phrases that provide the framework for an entire sentence, for example *my point is that\_\_\_\_.*

*Polywords* and *phrasal constraints* are word level phrases. They differ in that the former are fixed and the latter somehow variable. For example, *by the way* is fixed because *way* and *by* cannot be substituted by other words without compromising the functional meaning, whereas *as I was saying* is variable because *saying* can safely be replaced by *mentioning*. *Institutionalized expressions* and *sentence builders* are sentence level counterparts of polywords and phrasal constraints.

There are two approaches to identifying non-adjacent collocations. First, an interface could be provided for teachers to mark them manually. Second, patterns that could be recognized by computers can be defined in advance, for example, *make up \*{1-2} mind*, where \*{1-2} means that one or two words can be inserted between *up* and *mind*. The identification of lexical phrases will utilize pre-compiled phrase lists because they are grammatically or lexically variable, and their associated functions depend largely on human judgment. Extracting fixed

phrases is straightforward, but for variable ones, slots and their associated fillers can only be identified if pre-defined patterns are available.

The quality of automatically identified collocations is affected by the underlying natural language processing tools and the collocation identification algorithm. Section 7.1.2 discussed three causes of problematic collocations: tagging errors, partial collocations and incorrect chunking. It is impossible to eliminate tagging errors without human intervention. Partial collocations can be alleviated by defining more sophisticated patterns, such as extending noun + noun or noun + of + noun to include more nouns and the conjunction words *or* and *and*. However, they will never be complete because of the complexity of human language. Parsing sentences at the phrase level (Section 4.2.1) may help reduce incorrect chunking. For example, the phrase level parsing of example 2 shown in Table 7.2 is

```
[NP some/DT people/NNS] [VP make/VBP] [NP a/DT living/NN] [VP  

researching/NN] [NP the/DT family/NN histories/NNS] [PP of/IN] [NP  

others/NNS]
```

Here, *a living researching* would not be identified as a noun + noun collocation because the second noun *researching* belongs to the following verb phrase (indicated by VP). However, this solution is not perfect. First, the syntactic patterns defined for each collocation type would become more complex, which may introduce more partial collocations. Second, errors in phrase level parsing are inevitable, as evidenced in example 1 of Table 7.2:

```
[NP The/DT present/JJ research/NN] [VP focused/VBD] [PP on/IN] [NP  

the/DT strategies/NNS families/NNS] [VP used/VBD to/TO promote/VB]  

[NP home/NN language/NN] [VP learning/VBG] [PP in/IN] [NP oral/JJ  

and/CC] [VP written/VBN] [PP from/IN]
```

Here, *the strategies families* is incorrectly identified as a single noun phrase (indicated by NP). Further investigations are needed to find a balanced solution.

For all these reasons, an interface for manually adding and removing automatically identified collocations in the text is necessary so that teachers can eliminate inappropriate ones and pick up free form phrases.

The second evaluation tested the “cherry-picking” facility with students doing university study. They were asked to select collocations related to their study topic and then use them when writing a literature review. The evaluation plan was interrupted because some students changed their topic, which made the collections built for them obsolete, and some did not submit the literature review in the end. Another problem was that it is time-consuming to go through the printed cherry-basket to locate useful collocations while writing. CLS did provide a full-text search facility on collocations (Section 6.3.1), but it was not exploited by the students because of time constraints. As a result, only one student’s data was obtained. Although, this is rather disappointing, I nevertheless believe that the evaluation procedure was valid and the results were positive. Both students and teachers confirmed that using the system helped them achieve a better understanding of the concept of collocation, which will benefit students in the long run.

The third evaluation invited four trainee teachers to examine CLS and discuss its strengths and limitations. On the whole, they were impressed by the system and provided useful feedback and suggestions.

In conclusion, to support the second hypothesis a collocation learning platform was built based on the three well recognized processes that lead to lexical acquisition. To my knowledge, it is the first computer system that aims to help students systematically learn collocations. It is intended as a research prototype rather than a production system, which limits the ability to conduct fully satisfactory evaluations. As one teacher pointed out, human factors play a role in making full use of CLS. How to build collections that suits student needs, and how to keep students motivated and provide useful feedback and monitoring, both require further investigation.

CLS is still in its initial form. Studying how teachers actually use it in the classroom, and how students react to it, will yield more conclusive results and constructive suggestions for further development.

### 8.3 Into the future

CLS is a successful system that has been demonstrated in many workshops and conferences. In fact, Pathways College at the University of Waikato is using the WEB PHRASES and WEB COLLOCATIONS collections as language support tools. They have been introduced to PhD and Masters students in workshops that run regularly on campus. They are also helping me write this thesis! There will be many improvements in the future. Here are a few.

Some students could not make full use of CLS's collocation resources because they did not have sufficient vocabulary to formulate search terms. The choice of word form (singular or plural, verb base form or past participle, noun or corresponding adjective) and the presence or absence of articles may yield substantially different results. For example, in the WEB PHRASES collection, 20 different words follow *make difference* and all occur less than 1000 times. There are more than 100 for *make a difference* and the top ten occur more than 10,000 times. I plan to investigate ways to help students choose the right search terms by checking terms entered by students before submitting them to CLS. Given *make difference*, the system could suggest *make a difference*, *make any difference*, or *make no difference*, because they are far more frequent than *make difference*.

Collocation learning is a daunting task. Learning is likely to be most effective and sustainable if the learning environment puts learners in situations that make them want to use language, and presents them with challenges that they feel motivated to meet. Future developments will provide learners with opportunities to interact with their peers or teachers through computer-mediated communication tools such as text-based *chat*. Students will participate in activities in different ways—as individuals, pairs or groups that work in competition or collaboration.

For example, in the Collocation Guessing exercise (Section 6.4.2)

*plain, dark, white, bitter, milk, bar of—chocolate*

in single player mode, the computer presents the collocates of the word *chocolate* one by one until the learner guesses the word. Learners can compete to see who needs the fewest collocates to make the right guess. In paired mode, one learner is

assigned the list of collocates and presents them one by one to the other learner who does the guessing. In order to make their work efficient, the presenting party needs to choose the order carefully, moving from general words to stronger collocates. In this mode, pairs of learners work collaboratively with each other, and can also compete with other pairs.

The only monitoring and feedback that CLS currently provides takes the form of a summary report of an exercise. More is needed. How to monitor the development of the student's collocation knowledge, provide appropriate feedback, and generate exercises tailored to individual needs is a challenging research project. The trainee teachers suggested that CLS has the potential to become a self-study and self-evaluation tool for students. For that, a comprehensive monitoring facility is needed. One option is to incorporate CLS into existing course management systems like Moodle so that teachers can use the facilities they provide to assess student progress.

Last but not least, it would be valuable to conduct longitudinal research to assess the impact of using CLS on the development of collocation knowledge of EFL students. I hope to deploy CLS in a primary school in China, where young children are learning English for the first time, build collections using their textbook, ask students to do cherry-picking activities and collocation exercises regularly, and evaluate their collocation knowledge after one or more years.

CLS will be a useful tool for supporting collocation learning, especially in an EFL environment where teaching is grammar-oriented and exam-driven. I was taught to learn English by studying grammar rules, and still remember how hard it was to differentiate the words *look*, *see*, and *watch* by studying their definitions in a dictionary. I understand the difficulties inherent in changing the way that English is taught, but I hope that this thesis will be a small step in that direction. Evaluation of CLS will be ongoing, and will lead to refinements in both the material it provides and the interfaces through which teachers and learners use it. I believe that the deployment of computer-based collocation learning systems is an exciting development that will transform language learning.

## References

---

- Anthony, L. (2006). "Developing a Freeware, Multiplatform Corpus Analysis Toolkit for the Technical Writing Classroom." *IEEE Transactions on Professional Communication*, 49(3), 275–286.
- Arabski, J. (1979). *Errors as indicators of the development of interlanguage*. Katowice: Uniwersytet Śląski.
- Bahns, J. and Eldaw, M. (1993). "Should we teach EFL students collocations?" *System*, 21(1), 101–114.
- Bates, M.J. (1989). "The design of browsing and berrypicking techniques for the online search interface." *Online Review*, 13, 407–424.
- Benson, M., Benson, E., and Ilson, R. (1986). *The BBI combinatory dictionary of English: A guide to word combinations*. Amsterdam/Philadelphia: John Benjamins.
- Benson, M., Benson, E., and Ilson, R. (1997). *The BBI combinatory dictionary of English: A guide to word combinations*. Amsterdam/Philadelphia: John Benjamins.
- Biber, D. (2006). "Stance in spoken and written university registers." *Journal of English for Academic Purposes*, 5(2), 97–116.
- Biber, D. and Kurjian, J. (2007). "Towards a taxonomy of web registers and text types: A multi dimensional analysis." *Language and Computers*, 59(1), 109–130.
- Bishop, H. (2004). "The effect of typographic salience on the look up and comprehension of unknown formulaic sequences." In N. Schmidt (Ed.), *Formulaic sequences: Acquisition, processing, and use*, 227–244. Philadelphia, PA, USA: John Benjamins Publishing Company.
- Biskup, D. (1992). "L1 influence on learner's renderings of English collocations: A polish/German empirical study." In P.J.L. Arnaud and H. Béjoint (Eds.), *Vocabulary and applied linguistics*, 85–93. London: Macmillan.
- Bowerman, C. (1993). *Intelligent computer-aided language learning. LICE: a system to support undergraduates writing in German*. Unpublished doctoral dissertation, UMIST, Manchester.
- Braun, S. (2005). "From pedagogically relevant corpora to authentic language learning contents." *ReCALL*, 17(1), 47–64, Cambridge University Press.
- Brown, D.F. (1974). "Advanced vocabulary teaching: The problem of collocation." *RELC Journal*, 5(2), 1–11.
- Biber, D., Johansson, S., Leech, G., Conrad, S., and Finegan, E. (1999). *The Longman grammar of spoken and written English*. London: Longman.

- Chambers, A. and O'Sullivan, I. (2004). "Corpus consultation and advanced learners' writing skills in French." *ReCALL*, 16(1), 158–172, Cambridge University Press.
- Chambers, A. (2005). "Integrating corpus consultation in language studies." *Language Learning and Technology*, 9(2), 111–125. Retrieved November 28, 2008, from <http://llt.msu.edu/vol9num2/chambers/default.html>
- Chambers, A. and O'Riordan, S. (2006). "Integrating a corpus of classroom discourse in language teacher education: the case of discourse markers." *ReCALL*, 18 (1), 83–104, Cambridge University Press.
- Channell, J. (1981). "Applying semantic theory to vocabulary teaching." *English Language Teaching Journal*, 35, 115–122.
- Chinnery, G. M. (2008). "You've got some GALL: Google-assisted language learning." *Language Learning and Technology*, 12(1), 3–11.
- Church, K. and Hanks, P. (1989). "Word association norms, mutual information, and lexicography". *Proceedings of the 27th Annual Meeting of the Association for computational Linguistics*. Association for Computational Linguistics: 76–83
- Church, K., W. Gale, P. Hanks, and D. Hindle. (1991). "Using Statistics in Lexical Analysis." in Zernik (Ed.) *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, 115–164. Lawrence Erlbaum Associates Publishers.
- Cobb, T. (n.d.). *Compleat Lexical Tutor*. Retrieved July 7, 2008, from <http://www.lextutor.ca/>
- Conzett, J. (2000). "Integrating collocation into a reading and writing course." In M. Lewis (Ed.), *Teaching Collocation*, 70–87. LTP, England.
- Cowie, A.P. (1981). "The treatment of collocations and idioms in learners' dictionaries." *Applied Linguistics*, 2(3), 223–235.
- Coxhead, A. (1998). *An academic word list*. Occasional Publication Number 18, LALS, Victoria University of Wellington, New Zealand.
- Debski, R. (2003). "Analysis of research in CALL (1998-2000) with a reflection on CALL as an academic discipline." *RECALL*, 15(2), 177–188. Cambridge University Press.
- Dodigovic, M. (2005). *Artificial intelligence in second language learning: Raising Error Awareness*. Multilingual Matters, New York.
- Ellis, N. C. (2001) "Memory for language." in P. Robinson (Ed.), *Cognition and Second Language instruction*. Cambridge: Cambridge University Press.
- Fano, R.M. (1961). *Transmission of Information: A Statistical theory of Communications*. Wiley, New York.
- Farghal, M., and Obeidat, H. (1995). "Collocations: A neglected variable in EFL." *IRAL*, 33(4), 315–331.

- Firth, J.R. (1957). "Modes of Meaning." Papers in *Linguistics 1934-51*, 190–215. Oxford University Press.
- Fletcher W.H. (2005). "Concordancing the Web: promise and problems, tools and techniques." Retrieved September 21, 2009, from <http://www.kwicfinder.com/FletcherConcordancingWeb2005.pdf>
- Fuentes, C.A. (2003). "The use of corpora and IT in a comparative evaluation approach to oral business English." *ReCALL*, 15(2), 189–201.
- Gabrielatos, C. (2005). "Corpora and language teaching: Just a fling or wedding bells?" *Teaching English as a second or foreign language*, 8(4). Retrieved March 12, 2009, from <http://tesl-ej.org.ezproxy.waikato.ac.nz/ej32/a1.html>
- Goulden, R., Nation, P. and Read, J. (1990). "How large can a receptive vocabulary be?" *Applied Linguistics*, 11, 341–363.
- Hill, J. and Lewis, M. Eds. (1997). *LTP Dictionary of Selected Collocations*, LTP.
- Hill, J. (1999). "Collocational competence." *ETP*, 11.
- Hill, J. (2000) "Revising priorities: form grammatical failure to collocational success." In M. Lewis (Ed.), *Teaching collocation*, 70–87. LTP, England.
- Hill, J., Lewis M., and Lewis M. (2000) "Classroom strategies, activities and exercises." In M. Lewis (Ed.), *Teaching collocation*, 88–117. LTP, England.
- Hoey, M. (2000). "A world beyond collocation: new perspectives on vocabulary teaching." In M. Lewis (Ed.), *Teaching Collocation*, 224–243. LTP, England.
- Howarth, P. (1998). "The phraseology of learners' academic writing" In A. Cowie (Ed.), *Phraseology: Theory, Analysis and Applications*, 161–86. Oxford: Oxford University Press.
- Hulstijn, J. H. and Laufer, B. (2001). "Some empirical evidence for the involvement load hypothesis in vocabulary learning." *Language Learning*, 51, 539–558.
- Hwang, K. and Nation, I.S.P. (1995). "Where would general service vocabulary stop and special purposes vocabulary begin?" *System*, 23(1), 35–41.
- International English Language Testing System (IELTS), (1997). *Specimen materials handbook*. Retrieved January 21, 2009, from <http://www.scribd.com/doc/13570277>.
- Justeson J., and Katz, S. (1995). "Principled disambiguation: Discriminating adjective senses with modified nouns." *Computational Linguistics*, 21(1), 1–27.
- Kaltenböck, G. and Larcher, B. (2005). "Computer corpora and the language classroom: on the potential and limitations of computer corpora in language teaching." *ReCALL*, 17(1), 65–84. Cambridge University Press.

- Kessler, B., Nunberg, G., and Schütze, H. (1997). "Automatic detection of text genre." In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, 32–38. Madrid.
- Kilgariff, A., and Grefenstette, G. (2003). "Introduction to the social issue on the web as corpus." *Computational Linguistics*, 29(3), 333–347.
- Krenn B., and Evert, S. (2001). "Can we do better than frequency? A case study on extracting PP-verb collocations." In *Proceeding of the ACL Workshop on Collocations*. Toulouse, France.
- Laufer, B. and Nation, P. (1999). "A vocabulary size test of controlled productive ability." *Language Testing*, 16(1), 33–51.
- Leech, G., Garside, R., and Bryant, M. (1994). CLAWS4: The tagging of the British National Corpus. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING 94)*, 622–628. Kyoto, Japan,.
- Leech, G. and Smith, N. (2000). "The British National Corpus (version 2) with Improved Word-class Tagging." Retrieved November 3, 2009, from [http://www.natcorp.ox.ac.uk/docs/bnc2postag\\_manual.htm](http://www.natcorp.ox.ac.uk/docs/bnc2postag_manual.htm).
- Leech, G., and Svartvik, J. (1975). *A Communicative Grammar of English*. Harlow, Longman.
- Lewis, M. (1993). *The lexical approach*. Language Teaching Publication, England.
- Lewis, M. (1997). *Implementing the lexical approach: putting theory into practice*. Hove: Language Teaching Publications.
- Lewis, M. (Ed.) (2000). *Teaching Collocations*. Hove: Language Teaching Publications.
- Lewis, M. (2000). "Learning in the lexical approach" In M. Lewis (Ed.), *Teaching Collocation*, 155–184. LTP, England.
- Manning, C. and Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press.
- Marton, W. (1977). "Foreign vocabulary learning as problem no. 1 of language teaching at the advanced level." *Interlanguage Studies Bulletin*, 2(1), 33–57.
- Meyer, C. F. (2002). *English corpus linguistics. An introduction*. Cambridge: Cambridge University Press.
- Mitchell, T. F. (1971). Linguistic "going on": collocations and other lexical matters arising on the syntagmatic record. *Archivum Linguisticum*, 2, 35–69.
- Moskowitz, G. (1978). *Caring and sharing in the Foreign Language class*. Heinle & Heinle.
- Nagy, W. E. (1997). "On the role of context in first- and second-language vocabulary learning." In N. Schmitt, and M. McCarthy (Eds.), *Vocabulary*

- description, acquisition and pedagogy*, 64–83. Cambridge: Cambridge University Press.
- Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge University Press.
- Nattinger, J. R. and DeCarrico, J. S. (1992). *Lexical phrases and language teaching*. Oxford: Oxford University Press.
- Nesselhauf, N. (2003). “The use of collocations by advanced learners of English and some implications for teaching.” *Applied Linguistics*, 24(2), 223–242.
- Oxford Advanced Learners’ Dictionary* (6th Edition) (2000), Oxford University Press.
- Oxford Collocation Dictionary for Students of English* (2nd Edition) (2009), Oxford University Press.
- Palmer, H.E. (1933). *Second interim report on English Collocations*. Tokyo: Kaitakusha.
- Pawley, A. and Syder, F.H. (1983). “Two puzzles for linguistic theory: nativelike selection and nativelike fluency”, in J.C. Richards and R.W. Schmidt (Eds.), *Language and Communication*, 191–225. London: Longman.
- Peachey, N. (2005). “Concordancers in ELT.” In *British Council teaching English*. Retrieved October 28, 2008, from <http://www.teachingenglish.org.uk/think/articles/concordancers-elt>.
- Renouf, A., Kehoe, A., and Banergee, W. (2007). “WebCorp: An integrated system for web text search.” In C. Nesselhauf, M. Hundt, and C. Biewer (Eds.), *Corpus linguistics and the web*, 47–68. Amsterdam: Rodopi.
- Renouf, A. and J. M. Sinclair. (1991). “Collocational frameworks in English.” In K. Aijmer and B. Altenberg (Eds.), *English Corpus Linguistics. Studies in Honor of Jan Svartvik*, 128–143. London: Longman.
- Robb, T. (2003). “Google as a quick ’n dirty corpus tool.” *TESL-EJ*, 7(2). Retrieved November 28, 2008, from <http://www-writing.berkeley.edu/TESL-EJ/ej26/int.html>
- Salaberry, R. (1996). “A theoretical foundation for the development of pedagogical tasks in computer mediated communication.” *CALICO*, 14(1), 5–34.
- Seretan, V., Nerima, L. and Wehrli, E. (2004). “Using the Web as a Corpus for the Syntactic-Based Collocation Identification.” In *Proceedings of International Conference on Language Resources and Evaluation (LREC)*, 1871–1874, Lisbon, Portugal.
- Shei, C. C. (2008). “Discovering the hidden treasure on the Internet: using Google to uncover the veil of phraseology.” *Computer Assisted Language Learning*, 21(1), 67–85.
- Sinclair, J. McH. (1966). “Beginning the study of lexis.” In C.E. Basell, J. C. Cateford, M. A.K. Halliday, and R. H. Robins (Eds.), *In memory of J.R. Firth*, 410–430.

- Sinclair, J. McH. (2004a). *Trust text: language, corpus and discourse*. Routledge, London.
- Sinclair, J. McH. (2004b). *How to use corpora in second language teaching*. John Benjamins Publishing Company, Philadelphia, PA, USA.
- Stubbs, M., and Barth, I. (2003). "Using recurrent phrases as text-type discriminators: A quantitative method and some findings." *Functions of Language*, 10(1), 61–104.
- Swan, M. (1996). *Language teaching is teaching language*. Plenary IATEFL.
- Wermter, J. and Hahn, U. (2006). "You Can't Beat Frequency (Unless You Use Linguistic Knowledge): A Qualitative Evaluation of Association Measures for Collocation and Term Extraction." *Association for Computational Linguistics (ACL)*, 785–792.
- West, M. (1953). *A general service list of English words*. Longman, Green and Co., London.
- Witten, I.H., Bainbridge, D. and Nichols, D.M. (2010). *How to Build a Digital Library*. Morgan Kaufmann, Burlington, MA (second edition).
- Woolard, G. (2000). "Collocation—encouraging learner independence." In M. Lewis (Ed.) *Teaching Collocation*, 28–46. LTP, England.
- Wood, M. (1981). *A definition of Idiom*. Manchester, England: Center for Computational Linguistics, University of Manchester.
- Wray, A. (2002). *Formulaic Language and the Lexicon*. New York: Oxford University Press.
- Yorio, C.A. (1980). "Conventionalized language forms and the development of communicative competence." *TESOL Quarterly*, 14(4), 433–442.
- Yoon, H., and Hirvela, A. (2004). "ESL student attitudes toward corpus use in L2 writing." *Journal of Second Language Writing*, 13, 257–283.
- Yoon, H. (2008). "More than a linguistic reference: The influence of corpus technology on L2 academic writing." *Language and Learning Technology*, 12(2), 31–48.

## Appendix A Function word list

---

This is a list of the function words of English. It is used in this project to differentiate between content words and function words in a text.

### **Adverbial particles**

again ago almost already also always anywhere back else even ever everywhere far hence here hither how however near nearby nearly never not now nowhere often only quite rather sometimes somewhere soon still then thence there therefore thither thus today tomorrow too underneath very when whence where whither why yes yesterday yet

### **Auxiliary verbs (including contractions)**

am are aren't be been being can can't could couldn't did didn't do does doesn't doing done don't get gets getting got had hadn't has hasn't have haven't having he'd he'll he's I'd I'll I'm is I've isn't it's may might must mustn't ought oughtn't shall shan't she'd she'll she's should shouldn't that's they'd they'll they're was wasn't we'd we'll were we're weren't we've will won't would wouldn't you'd you'll you're you've

### **Prepositions and conjunctions** (one category since there is some overlap)

about above after along although among and around as at before below beneath beside between beyond but by down during except for from if in into near nor of off on or out over round since so than that though through till to towards under unless until up whereas while with within without

### **Determiners and pronouns** (omitting archaic thou, thee, etc.)

a all an another any anybody anything both each either enough every everybody everyone everything few fewer he her hers herself him himself his I it its itself less many me mine more most much my myself neither no nobody none no-one nothing other others our ours ourselves she some somebody someone something such that the their theirs them themselves these they this those us we what which who whom whose you your yours yourself yourselves

### **Numbers**

billion billionth eight eighteen eighteenth eighth eightieth eighty eleven eleventh fifteen fifteenth fifth fiftieth fifty first five fortieth forth four fourteen fourteenth fourth hundred hundredth last million millionth next nine nineteen nineteenth ninetieth ninety ninth once one second seven seventeen seventeenth seventh seventieth seventy six sixteen sixteenth sixth sixtieth sixty ten tenth third thirteen thirteenth thirtieth thirty thousand thousandth three thrice twelfth twelve twentieth twenty twice two



## Appendix B Penn Treebank tags

---

### Phrase Level

ADJP	Adjective Phrase.
ADVP	Adverb Phrase.
CONJP	Conjunction Phrase.
FRAG	Fragment.
INTJ	Interjection. Corresponds approximately to the part-of-speech tag UH.
LST	List marker. Includes surrounding punctuation.
NAC	Not a Constituent; used to show the scope of certain prenominal modifiers within an NP.
NP	Noun Phrase.
NX	Used within certain complex NPs to mark the head of the NP. Corresponds very roughly to N-bar level but used quite differently.
PP	Prepositional Phrase.
PRN	Parenthetical.
PRT	Particle. Category for words that should be tagged RP.
QP	Quantifier Phrase (i.e. complex measure/amount phrase); used within NP.
RRC	Reduced Relative Clause.
UCP	Unlike Coordinated Phrase.
VP	Verb Phrase.
WHADJP	<i>Wh</i> -adjective Phrase. Adjectival phrase containing a <i>wh</i> -adverb, as in <i>how hot</i> .
WHAVP	<i>Wh</i> -adverb Phrase. Introduces a clause with an NP gap. May be null (containing the 0 complementizer) or lexical, containing a <i>wh</i> -adverb such as <i>how</i> or <i>why</i> .
WHNP	<i>Wh</i> -noun Phrase. Introduces a clause with an NP gap. May be null (containing the 0 complementizer) or lexical, containing some <i>wh</i> -word, e.g., <i>who</i> , <i>which book</i> , <i>whose daughter</i> , <i>none of which</i> , or <i>how many leopards</i> .
WHPP	<i>Wh</i> -prepositional Phrase. Prepositional phrase containing a <i>wh</i> -noun phrase (such as <i>of which</i> or <i>by whose authority</i> ) that either introduces a PP gap or is contained by a WHNP.
X	Unknown, uncertain, or unbracketable. X is often used for bracketing typos and in bracketing <i>the...the</i> .

### Word level

CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential <i>there</i>
FW	Foreign word

IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PRP	Personal pronoun
PRP\$	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	<i>to</i>
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Wh-determiner
WP	Wh-pronoun
WP\$	Possessive wh-pronoun
WRB	Wh-adverb

## Appendix C User guide

---

This is the user guide for the WEB COLLOCATIONS and WEB PHRASES collections.

### **Exercise one: collect useful phrases and structures using WEB COLLOCATIONS and PHRASES**

Topic: *The threat of nuclear weapons maintains world peace.  
Nuclear power provides cheap and clean energy.*

*The benefits of nuclear technology far outweigh the disadvantages.*

Gather useful phrases related to the topic using WEB COLLOCATIONS

1. pick some keywords, for example *threat, nuclear, weapons, energy, benefits, disadvantages*
2. search for collocations using those words
  - go to WEB COLLOCATIONS
  - click **search**
  - type in *nuclear*
  - click **show collocations**
  - click *nuclear weapons*
  - click *nuclear weapons* again to see Web samples
  - click the icon that follows *nuclear weapons* to see BNC samples
  - type in *benefits*
  - click **show collocations**
  - look up the collocations that of your interest
3. try out other words yourself
4. write down those you think might be useful

Gather useful sentence structures related to the topic using WEB PHRASES

5. pick some phrases, for example *nuclear power, nuclear weapons, world peace*
6. search for phrases that contain those phrases
  - go to Web Phrases
  - click **search**
  - type in *nuclear power*
  - click **show phrases**
  - click *nuclear power is -> nuclear power is the -> nuclear power is the best -> Web* for web samples
  - click *nuclear power is -> nuclear power is a -> nuclear power is a good -> Web* for Web samples

- try out other phrases
- type in *world peace*
- select **phrase preceding** for search for
- click **show phrases**
- click *to world peace* -> *contribute to world peace* -> *can contribute to world peace* -> **Web** for web samples

### **Exercise two: find related words, synonyms and antonyms**

Sample 1: *investigate*

- go to WEB COLLOCATIONS
- type in *investigate*
- click **show collocations**
- click *investigate the effect of*
- type in *effect*
- click **show collocations**
- click *take the effect*

Sample 2: *rise*

- go to WEB COLLOCATIONS
- type in *rise*
- click **show collocations**
- click *rise up*
- click **show collocations**
- click *rates rise*
- write down **prices, shares ...**
- type in *prices*
- click **show collocations**
- click *prices include*
- type in *shares*
- click **show collocations**
- click *shares will delete*

### **Exercise three: using WEB PHRASES to check grammar errors**

Sample 1: *government must be responsible of their welfare*

- go to WEB PHRASES
- click **Search**
- type in *be responsible*
- select **phrase following** for search for
- click **show phrases**
- click *be responsible for*

Sample 2: *we can do something to make easier their life*

- go to WEB PHRASES
- type in *make easier their life*
- click **show phrases**

- try *make \* their life*
- select **phrase following** for search for
- try *make their life*

Sample 3: *They have increased day to day*

- go to WEB PHRASES
- type in *increased day \* day*
- click **show phrases**

Sample 4: *This problem would resolve a little*

- go to WEB PHRASES
- type in *problem would*
- click **show phrases**
- click *problem would be*

#### **Exercise four: check choice of words**

Sample 1: *It is difficult to think about nuclear power as a good source*

- go to WEB PHRASES
- click **Search**
- type in *think \* nuclear power as*
- click **show phrases**
- type in *nuclear power as*
- select **phrase preceding** for search for
- click **show phrases**
- click *see nuclear power as -> some see nuclear power as -> Web* for web samples
- try out *consider nuclear power as*

Sample 2: *nuclear power is limited to few hands*

- go to WEB PHRASES
- click **Search**
- type in *is \* \* few hands*
- click **show phrase**
- click *is in a few hands* to see web samples
- type in *in a few hands*
- select **phrase preceding** for search for
- click *aggregated in a few hands -> Web*
- click *concentrated in a few hands -> Web*

Sample 3: *Each country does not give threat to other country*

- go to WEB COLLOCATIONS
- click **Search**
- type in *threat*
- click **show collocations**
- click *pose a threat*
- click *pose a threat* again to see web samples

Sample 4: *The problem began with the development ....*

- go to WEB COLLOCATIONS
- click **Search**
- type in *problem*
- click **show collocations**
- click *problem seems to*
- click *the problem lies in*

#### **Exercise five: expand your text**

Sample 1: *We will all benefit from it*

- go to WEB COLLOCATIONS
- click **Search**
- type in *benefit*
- click **show collocations**
- click *benefit greatly*
- click *benefit greatly* again to see web samples
- try out other collocations

#### **Exercise six: use more precise words or different structures**

Sample 1: *It will be very important as the energy crisis is ....*

- go to WEB PHRASES
- type in *It will be \* important*
- click **show phrases**

Sample 2: *as the energy crisis is not far ahead*

- go to WEB PHRASES
- type in *energy crisis is*
- select **phrase following** for search for
- click **show phrases**

Sample 3: *There is really no danger for the public*

- go to WEB PHRASES
- type in *no danger*
- select **phrase preceding** for search for
- click **show phrases**
- click *poses no danger* -> *It poses no danger* -> **Web** for web samples
- type in *no danger* again
- select **phrase following** for search for
- click **show phrases**

Sample 4: *The cities must have another solutions*

- go to WEB PHRASES
- type in *must \* solutions*

- click **show phrases**
- type in *must \* \* solutions*
- click **show phrases**

Sample 5: *nuclear power can give us more benefits than ....*

- go to WEB PHRASES
- type in *more benefits than*
- select **phrase preceding** for search for
- click **show phrases**

#### **Exercise six: learn how to use a verb**

Sample 1: *If we outweigh the advantages and disadvantages of....*

- go to WEB PHRASES
- type in *outweigh*
- select **phrase following** for search for
- click **show phrases**
- type in *outweigh benefits*
- select **phrase preceding** for search for
- type in *outweigh the advantages*
- select **phrase preceding** for search for
- click **show phrases**

Sample 2: I intend to *examine about the solutions of these problems*

- go to WEB COLLOCATIONS
- click **Search**
- type in *examine*
- click **show collocations**
- click *examine the effects of*
- type in *solutions*
- click **show collocations**
- click *find solutions*



## Appendix D Evaluation of collocation resources

---

This is the full result of the evaluation of the WEB PHRASES and WEB COLLOCATIONS collections.

<b>student text</b>	<b>search terms</b>	<b>system suggestion</b>	<b>student's change</b>
<b>noun phrase</b>			
1 the important improvement	the important *	the important contribution	the important contribution
2 the most famous period of world	the most * period	active/successful/exciting period	the most successful
3 the main cultural value	* culture value	traditional culture value	traditional culture value
4 industrial lake	lake	artificial lake	artificial lake
5 a fancy and good position	position	a unique position	fairly unique
6 a country's cultures and histories	a country's *	a country's culture and history	culture and history
7 the most powerful attractions	the most * attractions	the most popular attractions	famous attractions
8 has implemented noticeable projects	projects	exciting projects	new projects
9 generation of youth	generation	younger generation	younger generation
10 contemporary arts building among our society	contemporary *	contemporary art gallery/museums in	contemporary art gallery
11 historical arts	compare historical arts and historical art	historical art	a historical art
12 classical artifacts	artifacts	historical/ancient art	ancient art
13 has deep interests in	has * interest in	has special/strong/particular interests in	particular interests in
14 a fast-paced development of knowledge	* development of knowledge	progressive development of knowledge	the progressive development
15 modern art's appearing	modern art	the development of modern art	the development of modern art
16 numerous of countries	countries	a number of countries	the number of countries
17 two types art of	two types *	two types of	two types of arts
18 the tendency about the society	the society	the development of society	the value of society
19 a great deal of arts	* of arts	a wide range of	a large number of arts
20 old artifacts are related to people	old artifacts *	old artifacts of people	old artifacts that prefer to collect

21	the most important steps of our evolution	* of evolution	stages of evolution	stages of evolution
22	a great deal of museum	museums	a number of museums	a large number of museums
23	numerous of museums	museums	a number of museums	a number of museums
24	important events in their times	events * * time	events of that time	at that time
25	the popularity of modern technology	* of modern technology	advent/development of modern technology	the development of modern technology
26	the behavior of ancient people	* of ancient people	lives of ancient people	the lives of ancient people
27	one of this evidence	one of * evidence	one of the evidence	one evidence is that
28	a element of a national spirit	of a national spirit	expression of a national spirit	a element of a national spirit
29	technology and economical development	technology and * development	technology and economic development	
30	traditional and historical art	traditional and * art	traditional art	
31	they have established specialization centres	* centres	cultural centres	
32	new-age artists	artists	contemporary artists	
33	these sorts of art	* of art	various forms of art	
34	lots of arts aspects	* of art	all aspects of art	
35	a currently representation of a society			
36	social experts			

**verb + noun**

37	spend their tour	tour	take a tour	spend their holiday
38	deserve attention of public	* public attention	attract public attention	attract public attention
39	take an important role in	role	play an important in	play an important role
40	cultivate their children with a good art understanding	* their children with art *	provide their children with art education	encourage their children to develop ...
41	afford citizens more entertainments	entertainment	offer citizen more entertainment	provide citizen more entertainment
42	reinforce the income for a particular country	income	generate/increase the income	increase the income
43	know clearly about their culture	* their culture	understand their culture	know about
44	make the country become more famous	make the country *	make the country attractive	make the country more attractive

45	save the history	history	preserve the history	preserve the history
46	let the society to become more	* the society	make society more	
47	balance their consciousness about culture	consciousness	raise cultural consciousness	raise the consciousness
48	forms of art ... have paid attention to some people	attention	draw attention	draw attention to
49	have an assumption about	assumption	make an assumption	make an assumption
50	preserve their artifacts to be shown as	preserve artifacts *	preserve artifacts for study and display	to gain attention
51	looking for learn about history	looking for * about	looking for information about	looking for information
52	acquire their history	history	learn about the history	know about their history
53	keep the culture and tradition value	the cultural value	increase the cultural value	increase the culture value
54	people have seen an increasing emphasis on	have * emphasis have placed * emphasis	have placed great emphasis	observed a particular emphasis on
55	ignore the significance of	significance	not recognize the significance	
56	show the cultures	culture	promote a culture	
57	art seals off the greatest events	* the events	record the events	
58	if the public only emphasizes the traditional one	* the tradition art	if the public only preserve the traditional art	
59	arts ... have produced the historical value	* the historical value	preserve the historical value	
60	construct numerous museums	museum	build museums	
61	lead to help common people	people * help ordinary people	young people can help ordinary people	
62	reserve mental values			
63	giving a better value to historical art			

**noun + verb**

64	the essay favour	the essay * * favor	I favour	I favour
65	are aware of ... a lot	are * aware of	are fully aware of	are fully aware of
66	the profound influence created by it	the profound influence * by	the profound influence exerted by	brought

**preposition phrase**

67	in the current society	in * society	in modern society	in the morden
68	during period	during * period	during the period of	during the period
69	there are some opinions towards	opinion * opinion in *	opinion in favor of	different opinions of
70	in the other hand	* the other hand	on the other hand	on the other hand
71	over the times	over *	over time	over time
72	in the field of culture	* of culture	in terms of culture	
73	in some stages	* some stages	at some stages	
74	are careless for the modern art			

**phrasal verbs /verb + preposition**

75	play an important role on	play an important role *	play an important role in	play an import role in
76	other critics point to the new ...	critics point *	critics point out that	point out
77	many countries today famous with	* famous famous *	is famous for	are famous for
78	Italy very famous with	Italy * famous	Italy is famous for	Italy is very famous for
79	give priority for encouraging ...	give priority *	give priority to	give priority to
80	different from cultural to the other	different from * to	different from culture to culture	different culture to culture
81	represents the culture on a whole			

**grammatical**

82	record that what happened	record * happened	record what happened	record what happens
83	is equal important to	is * important	is equally important	is equally important
84	more likely to be preserve	to be *	to be preserved	to be preserved
85	people who interested in	people who * interested	people who are interested	who are interested
86	become one the most	become one * the most	become one of the most	one of the most
87	It is a great art	* great art	the great art	the great art
88	try hardly to preserve	try * to	try hard to	try to
89	compare between ... and	compare * and	compare ... and ...	compare ... and ...
90	many century ago	many * ago	many centuries ago	many centuries ago
91	contain wide range of	contain * wide range	contain a wide range	a wide range of
92	we as next generation	as * next generation	as the next generation	the next generation
93	to people who were lived	people who * lived	people who had lived here	

94	is influence by	is * by	is influenced by	
95	devolving of interest			
96	be comparison with wine			
97	in the main while			
98	now a day			
99	the different socialization being change			
100	arts has represented			
101	keep going with the society			

**verb + complement**

102	are more worth and value	art is more	important	beneficial and valuable
103	society has become increasingly fascinating	society has become more *	society has become more open and tolerant	accepting
104	argument may be true	argument may be *	argument may be valid	
105	the society to become more valuable	make the society *	make the society more open	society become more open and liberal

**adverb use**

106	I almost totally agree	I * agree	I generally agree	I mostly agree
107	are aware of a lot	are * aware of	are fully aware of	are fully aware of
108	modern people strongly claim that	* claim that	proudly claim that	



## Appendix E Keywords and collocations produced by students

---

These are the keywords and collocations produced by the students to test their vocabulary related to their research topics.

<b>Student</b>	<b>Keywords</b>	<b>Collocations</b>
A	<i>leadership, management, administration, manage, leader, leading, policy, policies, managing, administering, collaboration, collaborating, democratic, sharing, control, bureaucratic, mentoring, mentors</i>	<i>leadership practice management practice leading role management paradigms collaborative leadership democratic leadership administering roles management control cultural leadership management policies mentoring policies management issues leadership issues administrative issues administrative policies</i>
B	<i>leadership, innovation, creativity, principle, change, improvement, enhance, education, teachers, student, learning, teaching, capacity, performance, development, standard, style, idea, risk, adaptation, involvement, engagement, commitment, vision, goal, empowerment, decision</i>	<i>educational leadership leadership approach leadership style curriculum development teaching capacity established vision teacher empowerment student performance teacher performance principle commitment standard teaching taking risk change creation performance enhancement teacher involvement school goal shared leadership decision making innovation activities leadership capacity educational change educational innovation</i>
C	<i>transition, teaching, teachers,</i>	<i>teaching methods</i>

	<i>students, perceptions, classroom, explore, materials, methods, difficulties, schools, learning, principals</i>	<i>teaching materials learning difficulties teachers' transitions teachers' difficulties students' learning teachers' perceptions transition difficulties teachers learning students' perception principal's perception students' difficulties classroom learning classroom difficulties school materials explore difficulties</i>
D	<i>science, innovation, perception, practical, rural, literacy, education, technology, innovative, relevancy, concepts, contexts, conceptualize, contextualize, culture</i>	<i>science innovations technology education science literacy innovative ideas science education conceptualizing science science perception constructivist approach scientific concepts distance mode inquiry learning collaborative learning science education scientific applications indigenous science environmental science formative assessments practical exercises</i>
E		<i>bilingual education language acquisition second language learning code switching standard variety non standard variety classroom practice contrastive analysis transfer of language language interference academic writing writing skills critical literacy</i>
F	<i>critical, thinking, effective, reading, skills, impact, bilingual, literacy, strategy, approach, theory, conversational, academic, ability,</i>	<i>critical thinking effective reading skills effective speakers English as a second language</i>

	<i>classrooms</i>	<i>critical literacy as a teaching method</i> <i>disadvantage of critical literacy</i> <i>theory of critical literacy</i> <i>impact of bilingual education</i> <i>immersion bilingual education</i> <i>transitional bilingual education</i> <i>reading ability</i> <i>teaching strategy</i> <i>code switching</i> <i>students literacy</i> <i>conversational language</i> <i>academic language</i> <i>home literacy</i>
H	<i>curriculum, reform, designer, planner, teaching, change, learning, development learners</i>	<i>political curriculum</i> <i>cultural curriculum</i> <i>old curriculum</i> <i>new curriculum</i> <i>economic curriculum</i> <i>national curriculum</i> <i>worldwide curriculum</i> <i>local curriculum</i> <i>English curriculum</i> <i>joint curriculum</i>

(Note: Student G was absent that day and student E didn't produce keywords)



## Appendix F Fill-in-Blanks exercises

---

Example Fill-in-Blanks exercises used to evaluate the “cherry-picking” facility

**Collocation Fill-in-Blanks**

Score: 0 out of 5

How to play | Summary report

levels (1) approach (1) observations (1) planning (1) design (1)

- This thesis is a work-in-progress that articulates my research journey based on the development of a curriculum innovation in environmental education. This journey had two distinct, but intertwined phases: action research based fieldwork, conducted collaboratively, to create a whole school approach to environmental *education curriculum* \_\_\_\_\_; and a phase of analysis and reflection based on the emerging findings, as I sought to create personal "living educational theory" about change and innovation.
- A qualitative research approach has been used to investigate the experiences, opinions, presuppositions, and interpretations of stakeholders in this complex context. The multiple case studies conducted via fifty-seven informants represent the whole education system namely at the higher, middle and *classroom management* \_\_\_\_\_. The methods of data collection encompass focus groups, semi structured interviews and analysis of documents.
- It views the organizational behaviour and its respective environment as a complex whole of interrelating, interdependent parts rather than fragmented entities, which go beyond simple cause and effect relationships. A qualitative *research* \_\_\_\_\_ has been used to investigate the experiences, opinions, presuppositions, and interpretations of stakeholders in this complex context. The multiple case studies conducted via fifty-seven informants represent the whole education system namely at the higher, middle and classroom management levels.
- I also look at local perspectives on the place of exams and physical discipline. Fieldwork included *classroom* \_\_\_\_\_ in rural and urban settings. The thesis documents how children approach learning at school, how teachers go about their work, and how teachers and students interact.

Check answers | Close window

(a) Receptive exercise

**Collocation Fill-in-Blanks**

Score: 0 out of 5

How to play | Summary report

levels (1) approach (1) observations (1) planning (1) design (1)

- Until the job satisfaction of principals becomes as explicit as their job dissatisfaction, few teachers will aspire to this role. It was also found that \_\_\_\_\_ of succession planning at an organisational and school level inhibits teachers' principal class leadership aspirations. The research identified the policy, planning and research implications that arise from the findings.
- Within the teacher and assistant principal groups, data were also sought from those who were aspiring to principal roles and those who were not aspiring to such roles. The \_\_\_\_\_ of the research was to understand the factors that influenced teachers (including those in leadership roles) in their decision making to apply, or not to apply for principal class leadership roles.
- Personal factors such as time required by the job, the perceived stress level of the job, and effect on family, are strong disincentives to promote, particularly for women. Many teachers believe that the current role expectations of principals would not allow them to balance the demands of their personal life and their work life; the administrative demands and community \_\_\_\_\_ of the role are, in particular, seen as too demanding. A major inhibiting factor for teachers' leadership aspirations is their lack of an understanding of the high levels of job satisfaction that balance the stresses of the principal role.
- The literature illuminated several key themes, which formed the conceptual framework underpinning the research. These included school improvement, \_\_\_\_\_ of learning communities, teacher commitment and motivation, changing roles of principals and promotion of teacher leadership. Given the purpose of this study it seemed fitting that the approach of the study should be predominantly interpretive and orchestrated through multiple site case study.

Check answers | Close window

(b) Productive exercise



## Appendix G Cherry basket

---

A cherry basket comprises a list of collocations and illustrative text that constitutes the sentence containing the collocation, and the preceding and following sentence.

### **qualitative study**

This thesis is a qualitative study into aspects of primary education in Samoa. Using student, parent and teacher interview material, I investigate local perspectives on why education is important, what children should learn, how children learn, and what constitutes 'good' teaching.

### **professional development programmes**

Curriculum and assessment change has been unrelenting and even the most conscientious teachers often feel overwhelmed. At national and local levels, professional development programmes have assisted teachers to address these changes and a number of approaches have been adopted. However, while teachers have engaged in professional development programmes, the actual benefits to classroom teaching and learning have been less certain.

### **teacher interview material**

This thesis is a qualitative study into aspects of primary education in Samoa. Using student, parent and teacher interview material, I investigate local perspectives on why education is important, what children should learn, how children learn, and what constitutes 'good' teaching. I also look at local perspectives on the place of exams and physical discipline.

### **look at local perspectives**

Using student, parent and teacher interview material, I investigate local perspectives on why education is important, what children should learn, how children learn, and what constitutes 'good' teaching. I also look at local perspectives on the place of exams and physical discipline. Fieldwork included classroom observations in rural and urban settings.

### **reflect fundamentally**

Education policies are profoundly influenced by Western ideologies and practices. These reflect fundamentally different ways of thinking about children, their relationships with adults, teaching, and learning. By contrast, teaching practices in Samoa are consistent with local beliefs, values and understandings, and the material realities of a small, fiscally constrained Pacific nation.

### **local beliefs**

These reflect fundamentally different ways of thinking about children, their relationships with adults, teaching, and learning. By contrast, teaching practices in Samoa are consistent with local beliefs, values and understandings, and the material realities of a small, fiscally constrained Pacific nation. Policy initiatives are often met with inertia and resistance.

### **teacher self-efficacy**

The current study attempted to identify conditions that affect the manner in which Western Australian primary school teachers perceive recent curriculum changes; the types of support they access; and the relative usefulness of this support. Based on preliminary findings in the first phase of this study and the research literature it was expected that teacher self-efficacy, teacher characteristics such as age and years of teaching, and school context such as the level of 'innovativeness' would prove to be influential in the process of implementing new

initiatives. A model expressing the relationships between these concepts was developed and evaluated in the second phase of this study.

### **informal observations**

By focusing on the attitudes and behaviours of teachers from 'innovative' schools it was thought more could be learned than in schools that maintain the status quo. Qualitative methods of semi-structured interviews, informal observations and the analysis of websites and school documents were utilised throughout this phase. The second phase of the study employed a quantitative approach, based on the findings of the first phase, specifically a process of questionnaire construction and distribution throughout the defined population.

### **process of questionnaire construction**

Qualitative methods of semi-structured interviews, informal observations and the analysis of websites and school documents were utilised throughout this phase. The second phase of the study employed a quantitative approach, based on the findings of the first phase, specifically a process of questionnaire construction and distribution throughout the defined population.

### **engage in innovative practices**

In addition, most teachers will modify initiatives to meet the needs of their students and to fit in with their existing orientations. Consequently, school structures need to become more flexible to encourage teachers to engage in innovative practices. Interestingly, the self-efficacy of a teacher influences the way they perceive and cope with curriculum change, however teacher characteristics, such as age and the number of years teaching, did not yield substantially different results when teachers were categorised along these dimensions.

### **individual knowledge**

In recent years, professional development programmes that have been made available to teachers in New Zealand and other western countries have not often achieved the desired outcomes of improved teacher practice and decision making, or increased student achievement. The professional development research literature implies that the reason for this situation, is the inadequacy of programmes that do not acknowledge the teacher as a learner with individual knowledge, experience and priorities for their learning. As resources and attention continue to be focused on improving curriculum policies and classroom decision making to enhance student literacy achievement and reduce disparities, it is important to continue the search for teacher learning opportunities that achieve the desired goals.

### **learn about the research process**

This alternative approach to professional development sought to investigate the outcomes of teacher researcher partnership projects, each designed by individual teachers who worked with the facilitator to address their self-identified 'questions about practice'. The facilitator and the teachers worked together over a fifteen month period during which time they had individual and group meetings to learn about the research process and to design and implement their individual projects. The facilitator as researcher, gathered data from the teachers using qualitative methods and the teachers in turn gathered their own data to inform the progress and outcomes of their projects.

### **research literature**

The current study attempted to identify conditions that affect the manner in which Western Australian primary school teachers perceive recent curriculum changes; the types of support they access; and the relative usefulness of this support. Based on preliminary findings in the first phase of this study and the research literature it was expected that teacher self-efficacy, teacher characteristics such as age and years of teaching, and school context such as the level of 'innovativeness' would prove to be influential in the process of implementing new initiatives. A model expressing the relationships between these concepts was developed and evaluated in the second phase of this study.

## Appendix H Cherry-picking questionnaire (student)

---

### **Student A**

1. Do you think you understand the concept of collocation, before or after using the system, and what is it?

*After using the system.*

2. How important do you think collocation knowledge is in writing an academic text? Is this view influenced by the use of the system?

*This is important as it provides clarity and more meaning to the piece of writing*

3. How confident are you in your collocation knowledge related to your study topic?

*I am quite confident now after the using the system.*

4. Do you think collecting a set of collocations related to your study topic was helpful in writing the literature review, and if so, in what way?

*Yes it was helpful although the topics were not very related to my study. What was important was the concept of collocations.*

5. You have collected about 100 collocations, but only used six of them in the literature review, why?

*The answer is related to (Q. 4) of not really related to my topic of study. That is although I selected 100 collocation they were not collocations that could used correctly in my literature review. One thing that I learned was how to use the system to access collocations so now I can use it more appropriately in my literature review.*

6. Do you think collocations you collected were useful? Would you do it differently if you were asked to redo it again?

*The collocations I collected were not very useful. If I redo it again I can get better collocations that can be used correctly and appropriately in my literature review and thesis writing.*

7. Other than the collocations highlighted by the system, what other phrases you would like to store as well?

*Not sure. May be phrases that will cover concerns of the purpose of practical science activities, comparison between urban and rural secondary schools, Nature of science, what is science, photographic study etc.*

8. Do you think you can improve your collocation knowledge by using the system regularly?

*Certainly yes, I am sure I can improve my collocations if I use the system regularly.*

9. What other things do you like to be included in the “Cherry basket”?

*Concerns as raised in answer to (Q.7).*

10. Do you think the system is easy to use? What can be improved?

*The system is easy to use if it is used regularly.*

11. What other facilities you would like the system to provide?

*Proofing reading may be.*

### **Student B**

1. Do you think you understand the concept of collocation, before or after using the system, and what is it?

*I think I only understand the concept of collocation after using the system. collocation is the grouping of words in a sentence.*

2. How important do you think collocation knowledge is in writing an academic text? Is this view influenced by the use of the system?

*I think collocation knowledge is very important in writing an academic text. Through using the correct collocation, a text will make sense and grammatically correct.*

*Yes, this view was influenced by the use of the system.*

3. How confident are you in your collocation knowledge related to your study topic?

*To a certain degree, I'm confident in my collocation knowledge, and in relation to my study topic, it has helped me a lot as I was doing my literature review for my Master thesis.*

4. Do you think collecting a set of collocations related to your study topic was helpful in writing the literature review, and if so, in what way?

*Yes, as I have expressed above, collecting a set of collocations related to my study topic was helpful in writing the literature review. It helped me to use the right and correct groups of words in their contexts.*

5. You have collected about 100 collocations, but only used six of them in the literature review, why?

*May be it was because of the relevancy of my topic to the 100 collocations I collected.*

6. Do you think collocations you collected were useful? Would you do it differently if you were asked to redo it again?

*Yes, I think the collocations I collected were useful. Yes, now that I have some experience in using the system, I would do it differently if I'm asked to redo it again.*

7. Other than the collocations highlighted by the system, what other phrases you would like to store as well?

*Apart from the collocations highlighted by the system, other phrases that I would like to store as well are the very relevant collocations to my current research topic.*

8. Do you think you can improve your collocation knowledge by using the system regularly?

*Definitely, if I use the system regularly, I can improve my collocation knowledge.*

9. What other things do you like to be included in the “Cherry basket”?

*For the moment, I’m satisfied with the things in the cherry basket. They are basically enough for now.*

10. Do you think the system is easy to use? What can be improved?

*Yes, the system is easy to use.*

11. What other facilities you would like the system to provide?

*As mentioned above (10), for now, I’m satisfied with the system.*



## Appendix I Cherry-picking questionnaire (teacher)

---

1. To what extent do you think collocation knowledge contributes to effective academic writing? Can you give me an example of your thinking?

*I think collocation knowledge is extremely helpful. It is an extension of vocabulary knowledge and reflects flexibility of use. A student really needs to know how the word is used and what its most likely collocates are. So for instance you might know the word 'epistemology' from a list, but it's crucial to know that its form and possibly most frequent collocate is represented in the word sequence 'epistemological beliefs'.*

2. Do you think there is value in students identifying/being made aware of collocations in a particular topic area related to students' work? If so, in what way?

*Yes absolutely. So the example above is a good one from the theory of education domain. Also if you work within domains in this way, they are clearly collocations that students are motivated to want to know and be able to use.*

*I think the first point though in your question relates to identifying and being made aware of collocations. This is important because students find it difficult to isolate and identify the boundary of collocations. Awareness activities are fundamental to learning.*

3. Do you think your students have a better understanding of the concept of collocations after using the system?

*Yes they clearly did and showed some evidence of flexibility and generative use. That is, they made a few minor changes - all acceptable - to some of the collocations.*

4. Do you think your students can improve their collocation knowledge by using the system regularly?

*Yes I do. I think if they were to use the system regularly they would have an inventory of useful, categorized collocations. Maybe by ticking them off when they used them in their writing - as I tried to get them to do - this would help them to monitor their learning and development in this area.*

5. How could you support/encourage them to do this?

*This is the big question, and one that I haven't resolved at all yet. Some goal setting would be good. Maybe students could get some sort of positive feedback after using their picked collocations a certain number of times.*

*Perhaps getting students to compare the texts they write without recourse to the system with a re-write using the system may help them to see how much more ‘native-like’ and academic their texts sound.*

6. The students only used a few of collected collocations in their writing, do you know why?

*I think that it was the newness of the system and the lack of feedback and monitoring. So that’s what I’ve tried to think about dressing in the comments above in 5.*

7. What are the limitations of the system?

*The system has the capacity to link collocations or words to other resources. So I think we need to tailor those resources to need. For example, it may be helpful to have a link to a domain specific glossary. The system can do this. So it’s not really a limitation – rather a limitation of the way the teacher set it up.*

8. What other facilities you would like the system to provide?

*As above – a link to a domain specific glossaries.*

9. Do you think you will use the system in your class in the future and how?

*Yes I want to trial it again and build it in more systematically to the programme so that there are incentives (feedback and monitoring) for students to use it.*

10. Apart from this evaluation, can you think of other uses of the system in a classroom?

*I think it has particular application for domain specific learning. I think that there are some fascinating domains or areas, where we might have texts that are more – or less – technical; more – or less – difficult.*

11. What have you learned from this evaluation?

*Mostly I’ve thought a lot about the issue of encouraging students to use the system regularly and independently. So the technology can be fabulous but there’s the human element – it has to appear immediately useful or interesting to students.*

*Then other thing is that the collocation identified by the system are not always ones that I would intuitively choose as a teacher.*

## Appendix J Instructions for teachers

---

These are the instructions given to the teachers who tested CLS.

### 1. Trying out a data driven language learning tool

Go to the following site:

<http://flax.nzdl.org/greenstone3/flax?a=p&sa=about&s1.display=activities&c=adminc3>

This is a collection of texts that I have put together for you to explore.

A collection can be made from any texts – as long as they’re not too long, and any number of texts. They can be:

- domain and/or topic specific
- language item specific e.g., the personal pronoun sub-collection
- at a particular level e.g., using texts from IELTS
- from a particular source or of a particular genre e.g., newspaper reporting; Wikipedia

This collection contains a few texts from school journal materials, used for an adult literacy project.

### 2. As a language learner

You can **browse** the collection of texts for particular patterns of words and their contexts.

You can click on one text e.g., *The vege car* from the list below.

titles    Level

- ▶ More than a Mountaineer M ountaineer
- ▶ Playing with Words
- ▶ The vege car

You will get the text for *The vege car*; and by clicking on **Collocations** it will give you the frequent patterns (the frequent collocations) that appear in the text.

 The vege car

[Original](#) | [Collocations](#)

Who hasn't grumbled every time they [pull into a service station](#)? The Macdonald family of Palmerston North isn't worried about the [price of petrol](#).     When their tank is empty, they just [pour in vegetable oil](#) - the kind that [takeaway shops](#) use for frying their chips. And yes, the Macdonalds can even [drive their car](#) on [vegetable oil](#) after it's been used for cooking - thanks to Dad, James nui Macdonald.

Click the two different bunches of cherries and see what you get: one will allow you to 'pick' useful collocations for your own personal list; the other will give you more collocations for the first word – and the second word of the collocation.

If you click on the other two icons they give you extended contexts for the collocations. One context is the live web, and the other is a large database of British English called the British National Corpus (BNC). I could link this to other databases if I wanted to which would give other contexts.

Back to the 'picking' function. When I do this as a learner I am compiling a personal list of useful collocations and examples of those collocations in context. To view what I have picked I go to the following icon on the right side of the page (the cherries):


 The vege car

If I click that I will be able to see all the examples I have chosen. This will be displayed in the following way:

### Cherry Basket

[Add Category](#)
[Hide Samples](#)
[Print friendly](#)

 price of petrol 	   1
<ul style="list-style-type: none"> <li>○ Who hasn't grumbled every time they pull into a service station? The Macdonald family of Palmerston North isn't worried about the <a href="#">price of petrol</a>. When their tank is empty, they just pour in vegetable oil - the kind that <a href="#">takeaway shops</a> use for frying their chips.</li> </ul>	
 save money 	   1
<ul style="list-style-type: none"> <li>○ Around three years ago, James nui decided it was time to build a car that didn't need petrol. It wasn't just to <a href="#">save money</a>. James wanted to protest against the war in Iraq, which he believes is partly about powerful countries like the United States wanting oil from the Middle East.</li> </ul>	
 need petrol 	   1
<ul style="list-style-type: none"> <li>○ Around three years ago, James nui decided it was time to build a car that didn't <a href="#">need petrol</a>. It wasn't just to <a href="#">save money</a>.</li> </ul>	

I can categorise those useful collocations as well if I like. See the tab **Add category**.

### 3. As a teacher

I can automatically generate exercises. The examples I have set up is a Fill-in-Blanks.

Click on **Activities**, and **Create an exercise**. This panel will come up. It gives you options about what type of collocation you want to have in your exercise. In this example, I've selected the Verb + Noun. Other options exist such as Adjective + Noun. Pull down the menu to see them all.

**Exercise parameters**

Collocation Type: Verb+Noun  
Contains words (separated by comma):

Number of sentences to choose from: 59

**Activity parameters**

Number of sentences: 10  
Collocation position: first word  
Show alternative words: Yes

Review ? Print ? Display ?

This shows me that I have 59 sentences across all the texts which have this collocation type. However I've selected that I only want 10 examples and that I want the first word eliminated.

By clicking the **Review** tab, I get something like the panel below. This is actually an example working with Adjective + Noun combinations. If as a teacher I don't like any of the automatically generated options, I can discard them, by clicking the **Discard** tab.

same (1) cheaper (1) nitric (1) helpful (1) eldest (1) little (1) powerful (1)  
crude (1) great (2)

1. Why not drive from one end of the country to the other? That way, he could show off his car and teach people about cleaner, cheaper fuels at the \_\_\_\_\_ time. James also wanted to show that a car like his could be made in Aotearoa.
2. Why not drive from one end of the country to the other? That way, he could show off his car and teach people about cleaner, \_\_\_\_\_ fuels at the same time. James also wanted to show that a car like his could be made in Aotearoa.

Discard Show answer Discard Show answer

If I click the **Display** tab, this is what I get – a self-checking exercise all ready to go for learners.

**Collocation Fill-in-Blanks**

Score: 0 out of 5

[How to play](#)[Summary report](#)

1. Why not drive from one end of the country to the other? That way, he could show off his car and teach people about cleaner, [.....] *fuels* at the same time. James also wanted to show that a car like his could be made in Aotearoa.
2. Driving a car that runs on petrol puts harmful gases into the air. These include carbon monoxide, carbon dioxide, [.....] *oxide*, nitrogen dioxide, and hydrocarbons. Why are they harmful?
3. Somewhere between Christchurch and Blenheim, James nui got a speeding ticket! "I was a [.....] *bit* surprised and very embarrassed," says James nui.
4. It wasn't just to save money. James wanted to protest against the war in Iraq, which he believes is partly about [.....] *countries* like the United States wanting oil from the Middle East. James nui, who is Ngā Puhi and Ngāti Whātua, also likes the fact that his car is in keeping with traditional Māori values.
5. In 1893, a man named Rudolf Diesel designed an engine that could run on vegetable oil – peanut oil, to be exact. Diesel engines are still around, but most now run on [.....] *oil* ... the stuff that James nui wants to avoid.

[Check answers](#)

## Appendix K Teachers' discussions

---

Here are the four questions and teachers' discussions about CLS.

1. Do you understand how this tool works? What do you think are some of the language learning principles that the tool exploits?

**Student A:** *I think this is really a really cool resource. I can see that it is great input for students and really challenges their processing. It gives great feedback and helps to monitor progress. Can I use a template to create the same sorts of work in te reo Maori?*

**Student B:** *I had a quick look at this the other day and thought "wow" looks fantastic. I wouldn't profess to understand how this tool works. My command of IT is way below that. However it is fun to use and as a user I could work my way around finding how it is operated from that perspective. So being fun is one principle it exploits. Then repetition.... the exercises involve repeated reading of the same text. it would be good if there was a facility to manipulate other examples of the same collocation in different contexts. The links to the different corpus are useful for high level learners but are not manipulative.... no opportunity to actually use them.*

**Students C:** *After I have experienced it, I feel so mysterious, I have a question: who invent it? A talented person. I think teaching and learning are opposite.*

**Students D:** *If I understand properly, rote-learning is one of the main language learning principles exploited here.*

*In a way, it is quite similar to a English learning software me and my colleague used before. It is called **Issues in English**. I could see quite a lot of similarities between them including 'specific topic; specific language items' In Issues in English there are also vocabulary exercises for synonyms and antonyms etc. However the design of this tool looks much more fun, which should attract learners more!*

2. What do you think the potential usefulness of the tool is for: the learners? the teacher?

**Student A:** *I think the potential usefulness is very high. It's another tool to add to the bank of activities that can be used to support language learning. For the learners I can't see too many issues. It means that those with computers can work on exercises at home. As a teacher one of the things I like most is the specific nature of the programme. I can really target a particular aspect of language that is being learned and also ensure it relates to the context being studied. Very cool.*

**Student B:** *I think it looks to be hugely useful. Presumably as time goes on a range of texts at different levels and different exercises are to be included.*

*Is there the facility for teachers to post their own texts?*

*For the learner the usefulness is in more practice. For the teacher the ease of creating appropriate tasks to support learning.*

*Somewhere on it I saw the opportunity to limit the use of tense, sentence length (I think it was in relation to an activity not yet up there) numbers of modal verbs etc. That would be useful in narrowing the complications learners are exposed to.*

**Students C:** *I think it is of great use to both of them, because for students, it is funny, easy to operate. It can attract students. Besides, students can learn much language knowledge from it, such as collocation, words and so on. As for teachers, they can create exercises easily and quickly, it is time-saving. As for the process of teaching, teachers can show this system to students, then let students to do it by themselves, which gives students deep impression.*

**Student D:** *What I wish to add here apart from all of yours is that it should be really useful for language learners (regardless of the levels) to do the self-monitoring and self-evaluation to some extent.*

*For teachers, I personally feel that it is more feasible for them to offer help to those relatively slower learners while the others are all set up and 'picking cherries' happily by themselves. This is something that can hardly be realized in traditional classrooms, especially when the size of the class is large like in many Asian countries and regions.*

3. What do you think the potential limitations of the tool are for: the learners? the teacher?

**Student A:** *Initially I thought it might be difficult for a teacher to keep track of how students are going and give them feedback on that basis. But then I thought there looks like a feature that summarises progress. Students could print that off and give it to the teacher as activities are completed and it would be easy for a teacher to monitor progress. Another limitation(?) is that a teacher must remember that this approach should be used in conjunction with other activities. If a teacher did not ensure students had opportunities to practise skill development in the strands of listening and korero, then this approach would limit student progress.*

**Student B:** *As Robin points out, the tool cannot be regarded as a complete language learning tool. It could only ever be one of a kit. a useful addition to a reading programme and to vocabulary development and practice. As such it would be limited by the range of texts provided (unless it is easy for teachers to put up their own ..... but then that takes the labour saving advantages away) and a good gradation of levels. Ideally I think learners would be able to work through a graded bank of texts.*

*Cross referencing to the original source of the text would make it more useful for teachers to link to classroom programmes. The tasks on the Library site could then be advocated as follow up to class work.*

*One limitation I see is the mismatch in levels between the texts and the examples in the corpus.*

**Students C:** Well...I think both of them have limits, because this system emphasize grammar, but there is no other things, this point is for learners. As for teachers, they have few styles of teaching, only multiple choice, which has a luck factor. I think this is really good to Chinese education system- exams and grammar

**Student D:** I agree with you all that this tool should only be employed as a supplement to language learning and teaching.

*It is very likely that students become reliant on the 'luck factor' while carrying on with the exercises without thinking actively and independently; for teachers, they may find it hard to monitor the whole classroom of learners within the limited class time although they do get a printed summary progress later on (as Robin pointed out).*

#### 4. Could you think of using it? How and/or Why?

**Student A:** I would love to be able to use this tool. Does it work in te reo Maori too? I think these exercises are great tools to use as activities for input text and also to really challenge the learner to process what they have been learning. Working individually or in groups would work just fine.

**Student B:** I woudl use it if the texts were appropriate for my learners and if I had on-line computer facilities to work with them initially to see they were able to use it appropriately.

*I see it would provide useful additional mileage and interaction with text as well as practice with target vocab. As Robin says it coudl be used inividually or in groups. The latter would provide opporunity for negotiation of meaning.*

**Student C:** Yeap, I want to use. Because I consider it as a new model of teaching. I want to use this in my classroom. First, I let students collect "cherries" by themslves, and then they will become teachers to create exercises to themselves. Teachers only play a role of guidance. You know, computer is very popular now, but there is no this kind of language exercises system, and it is good to self-study. I think students will enjoy it very much.

**Student D:** Yes I certainly would like everybody else here, although it will be challenging to start with. I think an in-depth orientation is needed and then practising trying out some sets of exercises(if not all) by myself is essential before I put it into classroom with my students.

*No matter how it is to be carried out, students' individual levels and needs (eg. specific grammar items) should be taken into account.*

*Another advantage of this tool is that teachers can use it to help those who have missed some classes, or have low proficiency to do some catch -up work after class hours (if there is a language lab available of course.)*

*By the way, I just wish to point out another factor that may affect the effectiveness of this tool, that is 'computer-literacy' of both teachers and students, especially the former. What's more, the required facilities that each school can or can not provide accordingly is another issue.*