The Use of Corpora in Second Language Learning

by

Michael Colley

(Under the direction of William A. Kretzschmar, Jr.)

Abstract

Corpora can be a valuable resource for students learning a foreign language. Unlike most language resources, corpora show language as it is actually used in real situations. However, attempts to use corpora in second language learning have had limited success due to the fact that they use software designed for linguists rather than language learners. This thesis discusses the development and use of a program designed with the needs of language learners in mind. Upon entering a word, the program extracts phrases from the corpus based on words that most frequently collocate with the given word. This is useful for language learners because it allows them to see how the meanings of words are shaped by the surrounding context. Compared to programs used by linguists, it does a better job of hiding the subtleties of word usage and emphasizes the most frequent and therefore most useful characteristics of the language.

Index words: Corpus studies, second language learning, programming, applied corpus linguistics

Chapter 1

Introduction

Students who are learning a widely-spoken second language generally have a lot of tools at their disposal: dictionaries, grammar books, phrase books, cassettes and CDs, reading materials, classes, and even computer software. One tool that has not yet made it into the mainstream, despite some experimental trials by researchers, is the use of corpora. There is no substitute for experience; in order to learn a foreign language, you have to be exposed to it, either through writing, spoken words, or (preferably) both. A corpus is like a large database of language experience. Unlike dictionaries, grammar books, and phrase books, corpora contain only attested language, language that was actually produced by somebody for some real purpose, not just for the purpose of illustrating a grammatical point or the use of a word.

The most useful feature of a corpus is that it is searchable electronically. By itself, a corpus is no more useful than a large collection of reading material. With computer software, however, a corpus becomes a valuable tool that linguists are increasingly using to do language research. The goal of this thesis is to design a program that will help make corpora an equally valuable tool for second language learners. Many researchers, such as Wooldridge (1991), Bernardini (2000, 2002), and Mair (2002) have already experimented with using corpora to help teach a foreign language. But there has not yet been any initiative to modify the software that linguists use to fit the needs of second language learners.

There are two major approaches to the use of corpora in language pedagogy. Traditionally, the most common approach has been to use corpora in the preparation of classroom materials such as readings or exercises. With rapid advances in computer memory and speed however,

it has also become popular for teachers to give students direct access to corpora so they can investigate the language themselves. Silvia Bernardini (2000) has called this second approach the 'non-mediated' use of corpora, since the students interact directly with the corpus, rather than being presented with materials that the teacher has prepared using a corpus. The use of the term 'non-mediated' is a little misleading however. While it is accurate in the sense that the students are interacting directly with the corpus data, the corpus is still mediated by the software used to visualize it, in most cases a concordance. A corpus is simply a collection of texts, usually selected with some criteria in mind. What makes a corpus valuable is the ability to search through it and view its contents in different ways, such as via a word list or concordance. A truly non-mediated use of a corpus, that is reading it from start to finish, would not be a particularly useful or enlightening experience.

Programs used in conjunction with corpora are like a new pair of eyes that allow users to see things in a text that they would not have otherwise noticed. The text does not change, only the way it is presented to the user. In this sense, the presentation of data may be more important than the data itself. Texts have been around for centuries, but it is only with the ability to rapidly change the way they are presented that linguists have begun to see them in a different light. One implication of this is that there is no single best way to present the data from a corpus. The most common types of programs used with corpora have been word lists and concordances. The most popular type of concordance that linguists use is the key-word-in context (KWIC) concordance, which shows the word in interest (the keyword) lined up and centered, along with a certain span of words to the left and right of the keyword. An example is shown in Figure 5.1 on page 55.

Concordances that can be sorted by the words to the left or right of the keyword are particularly useful in noticing patterns of word usage such as frequent collocates. Software packages like WordSmith Tools offer a collection of simple tools to make a word list and a KWIC concordance along with numerous additional, often more complex tools for analyzing texts. WordSmith and similar programs were designed primarily for linguists, with features

that are best suited to linguistic research. Although researchers such as Bernardini (2000, 2002) have been successful in using WordSmith and other programs in the classroom, the tools themselves were not designed for second language learning. It is not surprising then that such approaches often put the language learner in the role of a language researcher. While this may be a useful experience for some, it is not the only way that corpora can be used. The program created for this thesis will be designed specifically for second language learners, and thus will cater more to the goal of language learning than language research.

The program will focus on phrases rather than individual words. Corpus researchers like Michael Stubbs (2001) have argued that phrases, as opposed to words, are the main unit in which language is understood. In natural conversation and writing, words are seldom used in isolation. Thus the dictionary model of language, in which individual words have a list of definitions from which language users can choose, is misleading. Instead, corpus research has shown that most common words tend to collocate around a relatively small number of other words. This suggests that word choice is not completely open or random but is constrained by the context of adjacent words. This fact is reflected in collocation dictionaries like The BBI Combinatory Dictionary of English (Benson et al. 1997). The BBI dictionary differs from most phrase dictionaries in that it does not simply list common idioms in English like 'hit the bull's eye.' Instead, it lists phrases formed by frequent collocates, such as frequent verb-noun combinations (e.g. 'make a bet') and preposition-noun combinations (e.g. 'in advance'). These types of expressions are more important for second language learners than colloquial idioms since they tend to occur with much greater frequency. These are the types of expressions that this program will focus on.

Although most concordance programs allow users to search for phrases, the user must know in advance which phrases are important. For a second language learner, however, this is not a straightforward decision. For example, there is no logical reason we *take* advantage of things or *make* use of something, rather than 'make advantage' or 'take use'. The use of one word over another depends not only on meaning but also on the surrounding context.

For this reason, the program created for this thesis will be designed to work with phrases. It will assume no prior knowledge of what the phrases are, but based on their frequency of occurrence in the corpus being examined, the program will extract them and summarize the results for the user. This will be a major step towards making corpora more user-friendly for non-linguists and also making them useful for language learners who may not be proficient in the language they are investigating.

Although most of the algorithms used by this program to search through corpora exist in other programs such as WordSmith, no program has combined these algorithms in order to automate the task of finding phrases within the mass of data that simple searches return. To do this, a simple yet effective algorithm is used in which the results from one search through the corpus are then searched a second and third time to find phrases of multiple word lengths, as opposed to just two-word expressions. As explained in more detail in Section 4.3.2, this gives the program the ability to know that, for example, 'behalf of' is not in itself an expression, but rather a part of the larger phrase 'on behalf of.'

Screen shots of the program are available in Appendix A. When the program is first started, the user is shown a brief tutorial explaining what the program does, what a corpus is, which corpora are available, and other useful information. This tutorial is also available from the Help menu. Figure A.1 shows the Main Form[1] of the program with each of the components labeled. The names of the components will be used frequently throughout this thesis, particular in Chapter 4 'Creating the Program.' Figure A.2 shows an example of a search for the word 'hold' in the Brown and Frown corpora. The user can select a corpus by clicking on the File Button or by selecting File → Load Text... from the main menu. This opens the Directory dialog box shown in Figure A.3. If more than one text is selected, the user will be prompted to provide a name for the corpus. This way the selected texts will be treated together as single corpus, and the user can select the corpus again by using

[1]For consistency I have used capital letters in the names of all forms, classes and visual components of the program

the Select Text drop-down box on the main form. When the program starts, any previously selected corpora will be loaded into the drop-down box. Users can use the drop-down box to switch between corpora at any time, allowing a very easy means of comparing the results of a single search in more than one corpus.

I have called the program Phraser (at least until I can think of a better name). The user can enter words via the Enter Word edit box on the Main Form. Phraser then displays a list of common expressions containing that word, along with their frequency in the corpus. The user can then click any expression to see examples of how it is used. The expressions are listed in the scrollable List Box in the top right corner of the Main Form. The last item in the List Box is always 'View all occurrences of...', followed by the word the user entered. If no expressions are found, this will be the only item in the List Box. Example sentences containing the expression are shown in a large memo component called Display that takes up most of the Main Form. It is like the blank area of a word processor, but only used to display text. The Display uses rich text format (RTF), rather than plain ASCII text, so that the expressions can be highlighted in each sentence.

I decided to show the expressions in the context of complete sentences rather than the key-word-in-context (KWIC) concordance used by other programs because I have found that it is much easier to see how words or expressions are being used with complete sentences. The KWIC format is useful for noticing collocates of the keyword, but since this program automatically does this for the user, there is no need to show the results in a KWIC concordance.

Sometimes however, a single sentence is not enough to fully understand how an expression is being used. For this reason, users can double click on any sentence to see more of the context. This brings up a separate form called Viewer showing the sentence that the user clicked on (highlighted) along with 2400 bytes of surrounding text, which is enough to fill up the screen. Figure A.4 in Appendix A shows an example from the phrase 'hold on.'

1.1 Thesis organization

Chapter Two of this thesis talks in more detail about some of the previous attempts to use corpora to help second language learners. It is not intended as an exhaustive account of every work done in this area, as the pedagogical use of corpora is a very large field, and much of it has little to do with this thesis. It does, however, discuss some of the major studies done in this area. It also talks about some of the other programs used to search corpora.

Chapter Three talks about the corpora used in this thesis to test the program. There are four in all[2]: The Brown and Frown corpora, the Switchboard corpus, a Wall Street Journal corpus, and an academic corpus.

Chapter Four shows how the program was created. Flow charts of various functions and event handlers are available in appendix B, while the programming code itself is available in appendix C.

Chapter Five is devoted to using the program. Four example searches are given and their results are discussed. In addition, the program was tested using two ESL students: one beginner and one advanced student. Chapter Five also discusses some improvements and further developments that could be made.

[2]Actually there are five, though Brown and Frown are used together as a single corpus.

Chapter 2

Review of literature and other programs used with corpora

One of the most important early projects that used corpus data to help second language learners was the construction of the Cobuild dictionary, published in 1987. The dictionary was based on a 20 million-word corpus (200 million for the second edition, published in 1995), and its most distinctive feature was the use of clear, plain-language definitions, using complete sentences and, when possible, using only words that are more common than the word being defined. A typical definition is the one for 'motive': "Your motive for doing something is your reason for doing it"(Sinclair 1995). The second edition even gives an indication of how frequently each word occurs in the corpus so that learners (or teachers) can decide how important it is for their vocabulary.

Cobuild is a big project, and the corpus used in creating it, called the Bank of English, is now one of the largest corpora in existence. Most studies dealing with the pedagogical use of corpora are much more small-scale, however, often confined to a single classroom situation. Until recently, the most common approach to using corpora in second language learning has been in the creation of teaching materials for the students. Wooldridge (1991) for example made KWIC concordances from Georges Simenon's *Le Chien Jaune* to find sentences that illustrate the use of certain groups of vocabulary items, such as words related to poison, as well as features of French vocabulary that are often difficult for speakers of English, like the difference between *savoir* and *connaître* 'to know.' The students themselves do not sit down at the computer with access to the corpus, which in this case is just a single novel. From the students' point of view, then, this experience may not be much different than other types of activities they do in class. Although using a corpus in this way may be valuable for

creating better teaching materials and finding real-word examples for particular grammatical or lexical features, the students are not necessarily aware that anything has changed in the way they learn.

A different but related use of corpora in second language learning is the creation and analysis of so-called learner corpora, that is, corpora of language produced by people learning English. Again, this is mainly used to help teachers and designers of teaching materials, as well as linguists studying second language acquisition. By systematically analyzing the language of non-native speakers learning English, for example, researchers can refine their approach to EFL error analysis. Sylviane Granger (1998) notes that with learner corpora, researchers can investigate the overuse and/or underuse of certain features of the target language, where as before most work focused only on errors. Geoffrey Leech (1998) lists several other uses of learner corpora, including comparison of linguistic performance by speakers of different linguistic backgrounds (i.e. native language transfer), and helping learners make use of the full range of expressive possibilities in a language.

The analysis of learner corpora is certainly an important field, especially when the results are compared to other corpora. A corpus of native-speaker language, like the ones used in this thesis, provides valuable information on what is common and normal in a language, but it does not give any information about the kinds of words or grammatical constructions that are difficult for non-native speakers. However, this type of research is also distinctly different from the use of corpora that I am proposing in this thesis. The use of learner corpora has a lot in common with the use of native-speaker corpora by teachers and designers of teaching materials in that both approaches add a new tool to a traditional style of teaching. Neither approach puts the learner in charge of analyzing the data, an approach that is quite different from any previous method of language learning.

Johns (1991) was one of the earliest studies arguing for a radically different method of learning known as data-driven learning. In traditional methods, Johns argues, the teacher's role is to impart his or her knowledge of the rules of the target language, then evaluate

the students to make sure that they have successfully mastered these rules. This approach assumes that the teacher is already aware of all of the grammatical rules of the language, something which linguists know cannot be true. In data-driven learning however, the teacher guides the students in a process of discovery, using corpus data to answer the students' questions about the language. In this approach, both the teacher and the students learn together. Johns gives an example in which one of his students provided an explanation for when to use 'convince' vs. 'persuade' that turned out to be more useful for the other students than his own explanation.

Data-driven learning was made possible by the rapid advances in computer speed and memory that began in the early nineteen nineties. Still, when compared to today, the resources available for most of the early studies were quite limited. For this reason, most early attempts at data-driven learning focused on smaller corpora, sometimes consisting of just one or two texts. Mparutsa et al. (1991) for example, used very small corpora, ranging from 6,000 to about 30,000 words, to help students at the University of Zimbabwe, where all instruction is done in English. The case studies focus on academic English, noting that most of the students come to the university having learned English from literary sources and often find it difficult to understand the language used in textbooks. Through concordances, they note that many words have a more narrow set of meanings when used in a particular academic field, as with 'demand' and 'equilibrium' in economics.

Hunston (2002) summarizes some of the benefits of data-driven learning in general. Since students are encouraged to use corpus data to answer their own questions, they are frequently more motivated to learn about the language and usually remember what they have learned better than students who learn via grammar books. In has also been hypothesized that data-driven learning improves students' ability to learn from context, something which becomes increasingly important as students move away from controlled classroom situations to real world situation in which they must communicate the target language.

Bernardini (2000) was one of the first 'classroom concordancing' studies to make use of large corpora, rather than the small, usually specialized corpora used in previous studies. The study is based on a seminar in which Italian students of English translation used an online version of the British National Corpus to solve a particular research question. Bernardini seems to advocate putting language learners in the role of corpus linguists. In the seminar, students were asked to come up with their own research question and use the corpus to answer it. In a more recent study, Bernardini (2002) attempts to find ways of holding the students' interest in the corpus activity, since she noticed that the initial enthusiasm in some students working with corpora eventually diminished. She suggests the use of more than one corpus as well as a variety of tools (i.e. software) for analyzing them. In both of these studies, the students seemed to enjoy the activity and could see its usefulness for learning English. But it also made them aware of the fact that learning a language is a life-long activity, due to the subtleties of the language that they wound up investigating in their projects. This may be a good thing in that it encourages them to move beyond oversimplified grammar book explanations, but at the same time it can also be quite intimidating, giving them the false impression that they need to know every detail of the language to be able to use it proficiently.

Along these same lines, Mair (2002) focuses on the use of corpora to help non-native speakers gain native or native-like competence in English. He uses the example of the relatively rare construction 'possibility to do something' as opposed to the more typical 'possibility of doing something', of which the former occurs 29 times in the British National Corpus, a 100-million-word corpus.

Few studies of this type question the value of the software that is being used to make concordances from the corpus. One exception is Wible et al. (2002), who create a 'lexical difficulty filter' for the concordance. One major problem of using concordances in the classroom is that there is no control over the level of the difficulty of the sentences taken from the corpus. The filter they created thus ranks the sentences in order of difficulty by flagging

low-frequency words in the context of the keyword. The result is that the sentences near the top of the list tend to be the best sentences for illustrating the use of the keyword.

Most of the software used in these studies have been fairly simple concordancing programs. None of the studies talk in depth about the software used, nor do they usually justify its use over another program. One of the most common programs used in recent years is WordSmith, and in terms of the number of features that it offers, it is one of the most useful for corpus researchers. In fact, many of its features provided the inspiration for my program. Its creator, Mike Scott, provides a useful summary of many of its other features in Scott (2001).

WordSmith creates word lists and concordances as well as any program. It also calculates collocates, a useful feature that is discussed further in Chapter 4. It can show the spatial distribution of the keyword in the corpus and can statistically compare more than one corpus, just to name a few of its features. Whether or not any of the additional features would be useful for second language learners is another question, since the program was designed for linguistic analysis, not language learning. The answer is most likely that they are not, or at least that they would require substantial training before they could be useful.

Unlike most early concordance programs, WordSmith includes a Windows-style GUI (graphical user interface), a feature that increases its ease of use among students who are familiar with using Windows programs. Despite the user interface, it is not always straightforward to use the program due to the large number of buttons and menus the user must click through to make a concordance. Unlike my program, it does not put its most important features in a single window. One way to evaluate the ease of use of a Windows program is to count the minimum number of clicks on either the mouse or the keyboard that are required to perform a particular task. Using WordSmith version 4, it requires a minimum of twelve clicks to create a concordance starting from the program's opening screen. This count does not include the entry of the keyword itself. With my program the minimum number of clicks is one, the clicking of the OK Button. To be fair, seven of the clicks in WordSmith were for

selecting a corpus. If the user wants to use a corpus other than the default corpus on my program, an additional two clicks are required if the corpus has been previously loaded, and four clicks if it has not. Either way, this shows that in terms of its ease of use, WordSmith has a lot of room for improvement.

Another useful program is the Cobuild Concordance, a demonstration of which is available for free on the internet (http://titania.cobuild.collins.co.uk). The demonstration searches the Bank of English, the same corpus used to make the second edition of the Cobuild dictionary (Sinclair 1995). One of the nice features of this program is that it takes advantage of part-of-speech tagging, allowing users to search for keywords in a specified context. For example, if a user was interested in phrasal verbs, he or she could enter a verb and specify that it must be followed by a preposition. The code for a preposition in this program is 'IN', so a search for 'put IN' would return a concordance of 'put' followed immediately by a preposition. A plus sign plus a number indicates the span of words to be considered. For example, 'put+3IN' would find the word 'put' followed by a preposition within a span of three words to the right, as in '**put** an end **to** these things.'

This is a very useful tool for finding examples of specific types of phrases. It does, however, require users to know in advance what type of phrase they are looking for. There are eighteen codes for different parts of speech, which fortunately is not overwhelming in comparison to the level of detail most part-of-speech taggers include, but it is enough so that it may never occur to many learners to try certain combinations. The Cobuild Concordance thus falls into the same category as WordSmith: a very powerful tool, but ultimately better suited to people who already know a lot about English and particularly to linguists.

At this stage in its development my program does not offer as many features as WordSmith or the Cobuild Concordance. Its purpose for now is to illustrate that second language learners do not have to become experts in corpus linguistics, as studies by Johns (1991), Bernardini (2000, 2002) and others imply. Like the 'lexical difficulty filter' described in Wible

et al. (2002), my program attempts to find in the corpus what is useful for the learner, rather than overwhelming the user with all of the subtleties and minute details of the language.

Chapter 5

Using and testing the program

Because Phraser finds phrases based on frequency, its results are not limited to any particular type of phrase. This is an advantage over programs like the Cobuild Concordance, which requires users to have an idea of what they are looking for in advance. A few examples of the types of phrases the program is likely to find are: verb-preposition pairs or phrasal verbs ('look at'), adjective-preposition pairs ('similar to'), verb-noun pairs ('take place'), verb-adjective pairs ('make sure'), adverb-adjective pairs ('highly unlikely'), as well as longer phrases like 'get rid of' and 'keep in mind.' Idiomatic expressions like the ones often found in phrase books ('under the weather') do not occur very frequently, at least not in the corpora I have used. While such expressions may add a certain color or charm to one's speech, they are not essential for most forms of interaction. The program thus focuses on what is probably the most important part of the language—the part that is neither too common, like the phrases filtered out by the stop list, nor too odd or obscure, like the phrases that do not show up frequently in a corpus.

Since Phraser was designed to be a flexible tool for all types of phrases, I will show how it can be used by focusing on the results of searches for four different words. An alternative approach would be to focus on certain types of phrases and show how the program can be used to find them. This approach would be well suited for a program like the Cobuild Concordance, but for this program it does not demonstrate its full capacity. The other problem with focusing on types of phrases is that learners do not always categorize phrases like linguists do. Most learners do not stop and wonder "what types of prepositions can I use

with the word 'give'?" A more likely question would simply be "how do you use 'give'?", a question that can be more easily answered with a program like this one.

5.1 Example Searches

To show some examples of the types of phrases this program finds, I will discuss the results of four different searches. The words used are 'hold' (Section 5.1.1), 'mind' (Section 5.1.2), 'real' (Section 5.1.3), and 'hand' (Section 5.1.4). These words were selected because they result in a variety of different types of phrases and because they have a number of different uses depending on the context in which they are used.

5.1.1 'Hold'

The resulting phrase list for a search for the word 'hold' in the Brown and Frown corpora is listed in Tables 5.1 and 5.2. Table 5.1 shows the search without any stop words in effect; Table 5.2 shows the same search with all of the possible types of stop words included. These are the same lists that are displayed in the Main Form's List Box. To save space, however, I have listed the phrases in several columns. The user can click on each of these phrases to see a concordance. The user can control the contents of the stop list by selecting Advanced → StopWords... from the main menu as described in Section 4.3.5.

Looking at Table 5.1, the need for a stop list is immediately apparent. Many of the phrases included are due to the fact that extremely high-frequency words like articles and conjunctions tend to collocate with almost every word. Thus since 'hold' is a verb, it is not surprising that the top phrase is the verb with the infinitive marker 'to.' The next two are 'hold' followed by the articles 'the' and 'a.' The same is likely to be true of any transitive verb, since articles frequently occur before the verb's direct object, as do possessive adjectives as in 'hold my.'

Table 5.1: Resulting phrases from a search for 'hold', with no stop words in effect (Brown and Frown Corpora)

to hold (102)	would hold (6)	who hold (3)
hold the (35)	hold him (6)	they hold (3)
hold a (22)	hold of the (6)	his hold (3)
hold on (19)	took hold (5)	on hold (3)
to hold the (17)	not hold (5)	still hold (3)
hold of (14)	a hold (4)	hold its (3)
hold it (13)	can hold (4)	hold still (3)
hold them (11)	hold out (4)	enough to hold (3)
and hold (10)	hold onto (4)	seemed to hold (3)
hold his (10)	hold this (4)	trying to hold (3)
will hold (9)	hold their (4)	to hold down (3)
i hold (9)	hold her (4)	to hold him (3)
to hold a (9)	hold and (4)	to hold back (3)
you hold (8)	to hold up (4)	i hold my (3)
hold up (8)	to hold his (4)	get hold of (3)
hold your (8)	to hold your (4)	don't hold with (3)
we hold (7)	hold on the (4)	got hold of (3)
hold that (7)	to hold on to (4)	take hold of (3)
hold down (7)	taken hold (3)	hold your fire (3)
hold in (7)	which hold (3)	hold in the (3)
to hold them (7)	that hold (3)	View all occurrences of "hold" (303)
could hold (6)		

Table 5.2: Resulting phrases from a search for 'hold', with all stop words in effect (Brown and Frown Corpora)

hold on (19)	hold out (4)	seemed to hold (3)
hold of (14)	hold onto (4)	trying to hold (3)
hold up (8)	taken hold (3)	get hold of (3)
hold down (7)	on hold (3)	got hold of (3)
hold in (7)	still hold (3)	take hold of (3)
took hold (5)	hold still (3)	hold your fire (3)
hold back (5)	hold with (3)	View all occurrences of "hold" (303)
hold on to (5)	enough to hold (3)	

Table 5.2 shows the effectiveness of the stop list in eliminating such phrases. In this case, the list includes phrases that second language learners are likely to find more useful. These include phrasal verbs, which are made up of a verb plus a preposition like 'on' or an adverb like 'down.' Examples from the list are 'hold on', 'hold up' and 'hold down.' The list also includes frequent verb-noun combinations, with 'hold' being used as a noun in this case. Examples include 'took hold' and 'get hold of.' There is also a verb-adjective combination ('hold still') as well as others ('still hold', 'trying to hold', 'hold your fire'). Thus the program does not focus on any one type of phrase, but instead emphasizes frequency of occurrence in the corpus. The user does not need to be aware of the fact that 'hold' can be used as both a verb (as in 'hold out') and a noun (as in 'get hold of') since the program will find examples of both if they exist in the corpus. This is an advantage, especially for English, since words in English are seldom confined to a single grammatical category. Even new words have a tendency to expand their grammatical categories from their origin, as in 'she I.M.'d me.' Unlike the Cobuild Concordance, this program allows the user to focus on word usage rather than grammatical function.

Aside from the phrase list, the essential part of the program is the concordance. Figure A.2 in Appendix A shows the concordance for the phrase 'hold on' from a screen shot of the program. Compare this to the KWIC concordance of the same phrase, as adapted from WordSmith in Figure 5.1. The whole-sentence format used in Phraser is clearly better suited to second language learners than the KWIC format used in WordSmith. The KWIC format focuses on the immediate context, that is on the words surrounding the keyword. The whole-sentence format focuses on a larger context and makes it easier to see how the word is being used in the sentence. The whole-sentence format is more natural, since rarely in readings are people interested in only a particular word or phrase. The KWIC format, although invaluable for linguistic and other types of research, is ultimately not a natural way to read. It was designed to be able to notice linguistic patterns, and it is best suited to what it was designed for.

```
       wages. The Generals Continue To Hold On Thailand's military loses a battle
   by his grandmother. She struggled to hold on to her favorite grandson, even in
      Collor de Mello faces a battle to hold on to his job after a congressional
  to get out a cool, poised, "Won't you hold on a second, please", I covered up
   meted out in one analogous instance. Hold on tight. First of all, the six figures
     Accordingly, aristocrats tended to hold on to their land. It has often been
     "For the murder of Jack Wiggins." "Hold on there, Marshal," Fred said.
British to so disruptively retain their hold on these vital western posts in
  a passion for the polka and wanted to hold on to that. It could not be done
        idealism, despite its powerful hold on the political traditions of our
  knew, so that he quickly released his hold on the goat and pretended to be
       constantly to give the British a hold on this region, from whence they
       The dancer who never loosens her hold on a parasol, begins to feel that it is
 denominations are rapidly losing their hold on the central city. The key to
the Italian tradition of letting singers hold on to their notes, but to restrain them
   pieces. As soon as the fox has taken hold on most of the populace he imports
       gun on the desk, Marshal". "Now, hold on, damn it; I won't"- Red Hogan's
  out he moves, the thinner will be his hold on conclusive evidence, and the
      At least I had been unable to lay hold on the experience of conversion. Try
 only expressing our present emotion. I hold, on the contrary, that we mean to
```

Figure 5.1: Example of a KWIC display of the phrase 'hold on' from the Frown and Brown corpora (adapted from WordSmith).

5.1.2 'Mind'

For this example, I have included the following types of words in the stop list: articles, conjunctions, subject pronouns, object pronouns, relative pronouns, and possessive adjectives. This is the default content of the stop list unless the user changes the settings. The stop list essentially works as a filter to remove phrases that may not be of use to the user. With these settings, the filter is neither too strict nor too passive. However, since it is not the job of the programmer to decide what the user wants, the contents of the stop list can easily be set as the user wishes.

The result for a search for 'mind' with the default settings is shown in Table 5.3. 'Mind' is another word like 'hold' that can used in more than one part of speech and has several different meanings depending on its context. One can quickly see the variety of uses for the word in phrases like 'in mind', 'never mind', 'don't mind', and 'change her mind.' Many

Table 5.3: Resulting phrases from a search for 'mind' (Brown and Frown Corpora)

in mind (72)	mind in (6)	american mind (3)
of mind (41)	mind for (6)	wouldn't mind (3)
in his mind (31)	frame of mind (6)	jack's mind (3)
mind that (26)	would you mind (6)	mind on (3)
of the mind (21)	public mind (5)	mind does (3)
never mind (17)	own mind (5)	mind until (3)
don't mind (17)	mind from (5)	mind about (3)
to mind (16)	did not mind (5)	mind had (3)
mind of (15)	changed his mind (5)	mind are (3)
mind was (14)	of her mind (5)	goal in mind (3)
mind to (12)	of the human mind (5)	that in mind (3)
mind is (12)	machine in the mind (5)	change his mind (3)
of his mind (12)	made up his mind (5)	before his mind (3)
mind as (11)	scientific mind (4)	came to mind (3)
have in mind (11)	didn't mind (4)	on my mind (3)
in mind that (11)	won't mind (4)	up your mind (3)
state of mind (10)	mind were (4)	through her mind (3)
keep in mind (10)	independence of mind (4)	changed her mind (3)
of my mind (10)	of mind that (4)	mind and heart (3)
had in mind (9)	through his mind (4)	to bear in mind (3)
in the mind (9)	change her mind (4)	to keep in mind (3)
in my mind (9)	up her mind (4)	science of the mind (3)
up his mind (8)	bear in mind that (4)	up his mind that (3)
on his mind (8)	keep in mind that (4)	corner of his mind (3)
in her mind (8)	in the public mind (4)	doubt in my mind (3)
bear in mind (7)	out of my mind (4)	in his own mind (3)
human mind (6)	whose mind (3)	if you don't mind (3)
mind at (6)	york mind (3)	View all occurrences of "mind" (578)

of the phrases take the form of preposition + 'mind', though even these show a variety of senses, as can be seen from example sentences. 'In mind' tends to be used for what people are thinking, often showing the thoughts that lead up to or could lead up to a particular decision or action. Examples include:

> Rhode Island is going to examine its Sunday sales law with possible revisions **in mind**.
>
> Without any definite plan **in mind**, she went to a judge to see what could be done.
>
> In recognition of the growing trend for second homes, or vacation cottages, we have designed this one specifically with the family handyman **in mind**.

'In the mind', unlike 'in mind' refers more literally to the mind itself. Examples include;

> Scott Fitzgerald said it is a sign of genius to be able to entertain **in the mind** two mutually contradictory ideas without going insane.
>
> The machine **in the mind** offers a more muscular approach.
>
> The town itself, whatever the source of the individual characters, is, as a fictional setting, a metaphor for the female place **in the mind**.

'To mind' may also involve thoughts, but more often it is used with images, remembrances, associations, or sudden ideas. Examples include:

> The words of Cardinal Newman come forcibly **to mind**: "Oh how we hate one another for the love of God"!
>
> The subject of immortality brings **to mind** a vivid incident which took place in 1929 at Montreux in Switzerland.
>
> And when I think about it, the words from a song in a minor Broadway musical, 'Salvation,' come **to mind**.

'To mind' may also occur as verb phrase, as in "Anta, his wife, never seemed **to mind**."

The phrases found with 'on' always occur with a possessive adjective, as in 'on his mind' and 'on my mind.' These are often used for things that are worrying people, for example:

> Winston was relieved; those presents had been **on his mind**.
>
> In spite of the hundred things he had **on his mind**, Winston went and put his arm around her waist.
>
> "I got a lot **on my mind** right now and there are things I can't carry to the field," Davis said.

The example sentences show enough of the context that the non-native speaker of English can use them to help understand the meaning of these phrases. For example 'revisions' and 'definite plan' suggest a certain sense of 'in mind', while 'relieved' and 'in spite of the hundred things' suggest a completely difference sense for 'on. . . mind'.

5.1.3 'Real'

Part of what makes this program a valuable tool for all types of learners is the ability to use it with specialized corpora. This way, learners can focus on a particular style or register of English in which they would like to become more proficient. Certainly each register of a language has its own vocabulary that a learner must master. But many common words take on different meanings and may be used differently depending on the formality of the situation, the register, and so forth. This is something that shows up very well in this program. The word 'real' is a good example. It is a word that occurs frequently in all four of the corpora used—the Brown and Frown corpora (joined), the Switchboard corpus, the Wall Street Journal corpus, and the academic corpus—but it used differently in each one.

In the Brown and Frown corpora, common phrases with 'real' include 'real estate', '(in) real life', 'very real', 'real world', 'real problem', and '(there was) no real.' In these phrases and in the example sentences, 'real' is typically used to make things concrete, to specify

that one is talking about reality and not something fanciful, imaginary, delusional, etc. For example:

> Everyone knows that private detectives **in real life** are not like Sam Spade and Pat Novak, but the real and the imaginary musician are closely linked.
>
> Our problem, therefore, is to devise processes more modest in their aspirations, adjusted to the **real world** of sovereign nation states and diverse and hostile communities.
>
> American Catholic colleges and universities are, in a **very real** sense, the product of "private enterprise"- the "private enterprise" of religious communities.

'Real estate' of course has a separate meaning of its own. The second language learner is likely to conclude from these examples that 'real' is the opposite of 'imaginary' or 'false', which is probably the prototypical definition for most people. The Switchboard corpus, however, gives a very different picture of 'real.' Here, 'real' is more commonly used as an intensifier, as shown in phrases like 'real good', 'real nice', and 'real hard.' Some examples include:

> So that's still a **real good** show too I that one tends to come on earlier in the day than I want to turn the TV on.
>
> So I think that's **real nice** too to come up with different options do you like the job sharing.
>
> Just brownies or French doughnuts would have been good but it's **real hard** to make them they don't really come out like they do in New Orleans up here I don't know why.

A learner wishing to become more fluent in informal, conversational English will quickly notice this additional use of 'real' in the spoken corpus. The phrases 'real estate' and 'real world' also occur, showing that the other sense of real has not become obsolete in the spoken

data. Sometimes, however, the two senses seems to blur, as in these examples for the phrase 'have a real problem':

> Yeah and and and I **have a real problem** with the government you know giving away things to the the other countries that have as as much ability to to do harm to us as anyone.
>
> And I **have a real problem** with anything pesticide like pesticides or anything like that so.

In these sentences, 'real' is not so much the opposite of 'imaginary' as it is just an intensifier, similar to saying 'big problem.'

From the academic corpus, the top phrase is 'real world', while other phrases show 'real' being used in a way that is more peculiar to certain academic fields: 'real time', 'real numbers', and 'real wages.' From the Wall Street Journal corpus, it is not surprising that 'real estate' dominates the phrase list. Aside from 'real estate' itself there are 'real estate investment' and 'of real estate mortgage.' The phrase 'real time' also occurs here, as in:

> The chief financial officer of the firm was in continuous contact with all of the credit officers providing them with **real time** cash balances information.

These examples show the usefulness of the academic corpus and the Wall Street Journal corpus as sources for information on a specialized form of language. Such corpora are a very valuable resource for virtually all non-native speakers of English and, in some cases, even for native speakers. No matter what one's proficiency in a language, there are always areas for which the usage of certain words can be quite different from what one is used to. Specialized corpora are a valuable tool for people wanting to "learn the lingo" so to speak.

5.1.4 'Hand'

One thing that will become apparent to users of Phraser is that the meanings of words are largely determined by their context. Many times, what a word means in isolation is very

different from the way it is normally used in context. This is an important concept because words are seldom used in isolation; they are nearly always used with other words. A good example of this is 'hand.' By itself, the word seems very straightforward—it is just the name of a body part. Phraser, however, emphasizes another view of 'hand' that is more important for learners of English. The top phrase in all four corpora is 'on the other hand.' This phrase almost never refers to the hand itself, but uses 'hand' metaphorically. This is also one of the rare examples where a four-word phrase occurs at the top of the list. Phraser is able to do this because the shorter phrase 'other hand' (or 'the other hand') almost always occurs in the expression 'on the other hand.' Therefore, it eliminated the shorter phrase in favor of the complete one.

The phrase's high occurrence in all four corpora will emphasize for users its importance in a variety of different styles of English, both conversational and formal. Some examples from each corpus are:

> The hall, **on the other hand**, appeared lifeless and deserted on these long waterfront afternoons. (Brown and Frown)
>
> Yeah now my roomie **on the other hand** he is a power user. (Switchboard)
>
> Living organisms, **on the other hand**, cannot stay the same without changing constantly, and they use their environment to their advantage. (Academic)
>
> **On the other hand**, a small decline in gold prices can wipe out profits. (Wall Street Journal)

Other phrases show that the metaphoric use of 'hand' is not restricted to this one phrase. Another common phrase that occurs in all four corpora is 'right hand' and/or 'left hand.' For these it is important to look at the example sentences, since these phrases appear to use hand in its literal body-part sense. Occasionally they are used in this way as in "When her **right hand** was incapacitated by the rheumatism, Sadie learned to write with her **left**

hand." More often, however, the phrases are synonymous with 'right' and 'left' respectively, often used to specify the side of an object, even though the object probably does not have hands. An example is "The **left-hand** numbers in each cell represent utility indicators or net payoff values for A, the **right-hand** numbers those for B." With examples like these, users can note that in the metaphoric use of the phrase, the words are usually hyphenated. But more importantly, these and other phrases show that 'hand' is a very common word in English and occurs in a wide variety of expressions.

There are also some differences in the phrases found in each corpus. The phrases 'hand in' and 'hand over' occur frequently in the spoken corpus and in the Brown and Frown corpora, but do not show up in the Wall Street Journal or academic corpus. 'Out of hand' is one of the top phrases in the spoken corpus, but occurs toward the bottom of the list in the other corpora. Since users can easily switch from one corpus to another using the SelectText box, this program is particularly good at noticing subtle differences in word usage among various corpora. This type of data is merely observational of course; for linguistic research, much more carefully designed and selected corpora would be necessary, and a program like WordSmith would be better for doing statistical analyses on the results. Users of Phraser are more likely to focus on a single corpus that they are interested in. I point out these differences to show how each corpus can emphasize a different style or register of English, and how these differences show up in the search results from this program.

5.2 Testing the program

To test the usefulness of Phraser, I selected two ESL students to answer a short quiz using the program. One is a Brazilian student who has been studying English for less than a year. Thus she does not speak English fluently, even though she is able to read and understand a little. The other is a Korean student who has studied English for many years. She speaks English well, but still makes occasional errors, especially in writing.

The quiz consists of thirteen questions, with the easiest questions toward the beginning. The complete quiz in giving in Appendix D. The questions on the quiz are intended to evaluate three different things: the usefulness of the program's phrase list, the usefulness of the sentences, and the user's understanding of how to use to program.

Questions 1, 2, 3, and 5 are intended to evaluate the usefulness of the phrase list. These questions do not require the user to look at the example sentences. For example, question 1 asks the user to select the appropriate word for the following sentence:

> I was looking for something different in my career, so I decided to _____ advantage of the new opportunities.

The choices given are 'make', 'take', 'get', and 'find.' To find the answer, the user only needs to consult the list of phrases for 'advantage' and notice that 'take advantage' is one of the phrases. Questions like this one test the program's ability to help users select the right word when the surrounding words more or less dictate which word should be used.

Questions 4, 6, 8, 9 11, 12, and 13 require the user to try to infer the meaning of the phrases from the sentences, though 6 can be reasonably guessed by only looking at the phrases. For example, question 8 asks the user to try to determine when to use the expressions 'in mind', 'to mind' and 'on my mind.' These questions test the program's ability to help users learn how to use particular words and assume that they are already able to read and understand the example sentences.

Questions 7 and 10 are related to the actual use of the program. Question 7 asks the user to search for the word 'apart' in the Switchboard corpus, from which the phrases 'falling apart', 'fell apart' and 'is falling apart' are found. The question is how the user can find out if the word 'apart' is used in other expressions. The answer is by either clicking on 'View all occurrences of "apart"' or by searching for 'apart' in another corpus. Question 10 asks the user to search for the word 'wrong' and notice the phrases 'what's wrong' and 'what's wrong

with'. The question is how to find out if the phrase 'what's right (with)' is also common. This requires the user to search for the word 'right'.

5.2.1 Results

The beginning ESL student did not understand English well enough to read the quiz, so I had to help her in Portuguese. I did not translate any of the words she looked up, however, nor any word or phrase that she found in the program. If the question was a fill-in-the-blank, I did not help her with any words in the sentence. Despite her limited use of English, she was able to answer questions 1, 2, 3, and 5 successfully. The questions which required her to use the example sentences were more difficult however. The only ones she got right were 9b and 13. Question 9b was to find an expression with the word 'keep' that fits in the sentence:

I had to run in order to _____ them.

She was able to infer from the example sentences what the expression 'keep up with' meant and saw that it fit in the sentence. Question 13 was to compare the usage of the word 'quarter' in the academic and the Wall Street Journal corpora. She was able to see that 'quarter' in business English typically refers to a fiscal quarter, as in 'fourth quarter earnings.' For the rest of the questions however, she said that there were too many words she did not understand to be able to tell what the sentences meant.

Questions 7 and 10 were somewhat confusing for her. For question 7, she did not realize that 'falling' and 'fell' were related, so her answer was "by clicking on the other expressions." For question 10, to find out if 'what's right (with)' is a common expression, she tried entering the entire expression rather than just the word 'right.'

Phraser turned out to be a much more useful program for the more advanced ESL student. Unlike the beginning student, she knew enough English to be able to read the example sentences from the program and interpret how the phrases were being used. Some of the questions on the quiz were too easy for her, namely 1, 2, 3, and 5. I encouraged her, however,

to try to use to program to find the answer, even though for these questions she already knew what the answer was. For the rest of the questions however, the program appeared to be very helpful to her.

She answered all but one the questions correctly. The one she got wrong was number 4, which was a fill-in-the blank question:

> I can't find my camera. Can you help me look _____ it?
>
> A. at B. like C. for D. up

Her initial response was A 'at.' I asked her to look more closely at the example sentences in the program. Her confusion was between the choices 'look at' and 'look for.' Eventually she concluded that 'look at' is more of a perception, and 'look for' is more of an action. She was then able to choose C 'for' as the correct answer.

Questions 8 and 9 were the most difficult for her. Question 8 asks about the difference among the expressions 'in mind', 'to mind', and 'on... mind', as described in Section 5.1.2. She did not have much of a problem with 'to mind', since she noticed that it tends to occur with 'come' and 'bring', as in 'come to mind' and 'bring to mind'. The difference between 'in mind' and 'on... mind' was more difficult, though she eventually concluded that 'in mind' often occurs with plans and 'on... mind' often occurs with burdens. For question 9, she had some doubts about the following fill-in-the-blank question, which asks for an expression using 'keep':

> If you work at home, it's important to _____ your business expenses.

The answer I intended was 'keep track of', though she responded with 'keep a lid on.' Either answer seems appropriate however, depending on the meaning of the sentence, which admittedly is not clear as it is written.

For question 11, she had no difficulty realizing that 'went wrong' and 'went right' are not related. 'Went wrong' occurs in sentences such as:

> They attributed everything that **went wrong** in Russia to German influence and intrigue.

'Went right' occurs in sentences such as:

> "And your golden god", said Samuel Burns, "probably **went right** home and poured himself into a boiling bath."

She concluded that 'went wrong' is used when something wrong happens, while 'went right' is most often used to mean 'move directly to somewhere.' She was also able to notice the difference in usage of the word 'real' in the academic and Switchboard corpora, as described in section 5.1.3, as well as the peculiar usage of 'quarter' in the Wall Street Journal corpus.

5.2.2 Discussion

Although these two tests are not intended as formal research into the usefulness of Phraser, they do show some promising indications that the program is worth-while. Both students admitted that they were not very good with computers, and the second was even a little apprehensive about whether she would be able to learn how to use the program. Once she saw it, however, she caught on very quickly and immediately realized it could be useful for her. She was particularly interested in the academic corpus, being a foreign student in the United States. For most students, it is likely to be in the academic world that their language is most often evaluated, so an academic corpus is clearly the most valuable for them. After the test, she even said that if I sell the program some day, she would definitely buy it. Her only complaint was that it was sometimes difficult to locate particular phrases if there was a long list. This may be due to the nature of the quiz, which requires the user to focus on particular phrases, rather than just the most frequent ones. I asked if it would be useful if there was an option to put the phrases in alphabetical order, and she said that would help a lot. That way, users can choose which order is most convenient for them.

The beginning student's reaction to the program was also positive, despite not being able to answer the more difficult questions. She said that she was surprised that she was able to answer as many questions as she did, having only been through beginning English classes for speakers of Portuguese. Rather than being frustrating or intimidating to her, the program actually appeared to make her more confident in her English, since she saw that some questions that appeared difficult were actually very easy using the program (namely the multiple choice fill-in-the-blank questions).

5.3 Further developments

There are many additional features that could be added to this program to make it more useful. The test with the beginning ESL student suggests that the program is of some use for beginning students, but they are unlikely to learn as much from it as a more advanced student. One thing that could make this program more useful for beginning students is the ability to use it with a parallel bilingual corpus. This way, students could compare sentences in English with translations in their own language. Such a feature would not translate words or phrases for the students, since the student could consult a foreign-language dictionary for this. Instead, it would allow students to see that a word in English, like 'mind' or 'hand' in the examples given earlier (Sections 5.1.2 and 5.1.4), is not equivalent to the dictionary definition in their own language, but rather takes on a life of its own depending on how it is used. The translations would merely serve to help the user understand the sentence, which was a problem for the beginning ESL student who tested this program. I suggested the idea of using a bilingual corpus to her, and she said it would help a lot, adding that she would easily be able to answer all of the questions if she knew what the sentences meant.

Many other useful features could be added if the corpora were tagged for parts of search. Only one of the corpora used in this thesis (the Switchboard corpus) has part-of-speech tags. To make full use of them, however, would require a more sophisticated indexing system, or

else a completely different method for locating words in the corpus. With the ability to know what part of speech a word belongs to however, the program could do more than just look for collocations; it could also look for colligations, words that are related by a particular grammatical feature. For example, one feature of English that is often difficult for speakers of other languages is knowing when to use the infinitive or the '-ing' form of the verb, as in "She likes to shop" vs. "She likes shopping." The current program would not be of much use for this, however with a tagged corpus it could group all infinitives together as well as all '-ing' forms and show which one is more frequent. For example, a search for 'avoid' would find many examples of 'avoid + -ING', but would probably not find 'avoid + INFINITIVE' (cf. 'avoid going to school' vs. *'avoid to go to school').

Along these same lines, other features could be added if the program grouped words together based on certain grammatical categories, like possessive adjectives, or on lemmas. With the current program for example, a search for the word 'shook' returns phrases like 'shook his head' and 'shook her head.' The program could recognize 'his' and 'her' as possessive adjectives and replace them with a general word like 'one's.' A similar type of thing could be done with lemmas, for example, a search for 'shook' could also search for 'shake', 'shakes', 'shaken', and 'shaking' and group the results accordingly, arriving at generalized expressions like 'shake one's head.' There is some value in leaving the program like it is, however. Phrases like 'shake one's head' are typical of dictionary and phrase book entries, but they rarely occur as such in real language. 'Shook his head' and 'shook her head' are far more typical, and arguably, some learners may benefit more from seeing the phrases as they actually occur. Secondly, many verb phrases are used more in one tense than another, while noun phrases may exist in the singular but not in the plural, or vice versa. For example, a search for 'time' finds phrases such as 'at the same time', 'for the first time', and 'long time', while a search for 'times' results in phrases like 'at times', 'three times' and 'at all times.' At least in this example, it does not appear that the behavior of 'time' is much like that of

'times'. Examples like these, though, may be more of an exception to the general tendency, and there may be some way for the program to deal with them.

As for the internal algorithms of the program, the multiple search algorithm described in Section 4.3.2 seems to work quite well for finding most types of expressions. Its only major disadvantage is that it does not work well for phrasal verbs in which the phrasal part has been separated from the verb, as in '**take** the trash **out**'. A possible solution to this would be to make a full table of collocates, like the one shown in table 4.1 for WordSmith, rather that restricting the view to the collocates immediately to the left and right. This would not be necessary in all of the searches, only the first one for the keyword itself. This way, the program could include sentences with phrases like 'take the trash out' in the examples for 'take out'.

The indexing system for this program also has room for improvement. Ideally, the best method for text retrieval would be one that can search the text reasonably quickly without using an index. Having to create an index is problematic because it may take ten to twenty minutes for even a small corpus, and several hours or more for a large corpus. However, file indexing is a complex and developing field of study in itself, so the best solution to this problem is to find a ready-made text retrieval system to use in this program. These are rarely if ever free for the Windows operating system, however, and I did not have the resources to get one for this project.

Chapter 6

Conclusions

There are no doubt many other features that could be added to Phraser, but the basic framework of the program that has been created already appears to be useful for students learning English. The program is flexible enough that it can be used in many different ways. With a sufficiently large corpus, it could be used as a reference guide to help students who are not sure of how to use a particular word. In this sense, it could be used as another type of dictionary, one that gives examples instead of definitions. It is important to remember, however, that a small or moderately sized corpus like that ones used in this thesis do not contain examples of all possible combinations of words, only the most probable ones. For this reason, I envision this program as being used primarily as a tool for learning. In a classroom, the program could be used in conjunction with other assignments, such as exercises, readings, and compositions. For example, rather than explaining the difference between 'much', 'many', and 'a lot', an instructor could let students use this program and have them "discover" the difference themselves. Since the program automates the task of finding phrases, this would not require a large amount of extra work for the student, nor does it require them to have any training in corpus research. Although the program does not provide the answers to students' questions, it does present the data in a way that makes the answers easy to find.

The tests with both a beginning and an advanced ESL student suggest that the program can be useful to both in some ways, but it is more likely to be useful for students who know enough English to be able to understand words and phrases from their context. Thus the program is not a substitute for a dictionary or phrase book; it is a tool that gives students

experience with real language, but in an organized way that allows them to focus on words with which they are having difficulty.

The strength of the program lies in its ability to look at words in the context of other words, not just as individual units with discrete and independent meanings. In this way, it forces the learner to think more along the lines of "how do I use the word X?" rather than "what does X mean?"

Bibliography

Benson, Morton, Evelyn Benson, and Robert Ilson 1997. *The BBI Combinatory Dictionary of English: A Guide to Word Combinations.* Revised Edition. Amsterdam: John Benjamins Publishing Company.

Bernadini, Silvia 2000. "Systematising serendipity: Proposals for concordancing large corpora with language learners." *Rethinking Language Pedagogy from a Corpus Perspective: Papers from the third international conference on Teaching and Language Corpora.* Lou Burnard and Tony McEnery (eds.) Frankfurt am Main: Peter Lang.

— 2002. "Exploring new directions for discovery learning." *Teaching and Learning by Doing Corpus Analysis: Proceedings of the fourth international conference on Teaching and Language Corpora.* Bernhard Kettemann and Georg Marko (eds.) Amsterdam: Rodopi.

Granger, Sylviane 1998. "The computer learner corpus: a versatile new source of data for SLA research." *Learner English on Computer.* Sylviane Granger, ed. London: Longman.

Hunston, Susan 2002. *Corpora in Applied Linguistics.* Cambridge: Cambridge University Press.

Johns, Tim 1991. "Should you be persuaded—Two samples of data-driven learning materials." *Classroom Concordancing.* Tim Johns and Phillip King, eds. ELR Journal, Vol. 4.

Leech, Geoffrey 1998. "Preface." *Learner English on Computer.* Sylviane Granger, ed. London: Longman.

Mair, Christian 2002. "Empowering non-native speakers: The hidden surplus value of corpora in continental English departments." *Teaching and Learning by Doing Corpus Analysis: Proceedings of the fourth international conference on Teaching and Language Corpora.* Bernhard Kettemann and Georg Marko (eds.) Amsterdam: Rodopi.

Mparutsa, Cynthia, Alison Love, and Andrew Morrison 1991. "Bringing concord to the ESP classroom." *Classroom Concordancing.* Tim Johns and Phillip King, eds. ELR Journal, Vol. 4.

Scott, Mike 2001. "Comparing corpora and identifying key words, collocations, frequency distributions through the WordSmith Tools suite of computer programs." *Small Corpus Studies and ELT: Theory and Practice* Mohsen Ghadessy, Alex Henry, and Robert L. Roseberry (eds.) Amsterdam: John Benjamins.

Stubbs, Michael 2001. *Words and Phrases: Corpus Studies of Lexical Semantics.* Malden: Blackwell Publishers.

Wible, David, et al. 2002. "Toward automating a personalized concordancer for data-driven learning: a lexical difficulty filter for language learners." *Teaching and Learning by Doing Corpus Analysis: Proceedings of the fourth international conference on Teaching and Language Corpora.* Bernhard Kettemann and Georg Marko (eds.) Amsterdam: Rodopi.

Witten, Ian H., Alistair Moffat and Timothy C. Bell 1999. *Managing Gigabytes: Compressing and Indexing Documents and Images.* Second Edition. San Francisco: Morgan Kaufmann Publishers, Inc.

Wooldridge, T. Russon 1991. "A CALL application in vocabulary and grammar". *CCH Working Papers, vol. 1.* http://www.chass.utoronto.ca/epc/chwp/wulfric1/index.html.

Bradon, Edwin Philip 1987. *Do Teachers Care About Truth? Epistemological Issues for Education.* London: Allen & Unwin.
http://www.uwichill.edu.bb/bnccde/epb/PREF.html

Buchanan, James M 1975. *The Limits of Liberty: Between Anarchy and Leviathan.* Chicago: University of Chicago Press.
http://www.econlib.org/library/Buchanan/buchCv7Contents.html

Cohen, Bernard Leonard 1990. *The Nuclear Energy Option: An Alternative for the 90s.* New York : Plenum Press. http://www.phyast.pitt.edu/∼blc/BOOK.htm

Committee on Issues in the Transborder Flow of Scientific Data et al. 1997. *Bits of Power: Issues in Global Access to Scientific Data.* Washington, D.C.: National Academy Press.
http://www.nap.edu/readingroom/books/BitsOfPower/

Darrigol, Olivier. 1992. *From c-Numbers to q-Numbers: The Classical Analogy in the History of Quantum Theory.* Berkeley: University of California Press.
http://ark.cdlib.org/ark:/13030/ft4t1nb2gv/

Francis, W.N., and H. Kucera. 1979. *Brown Corpus.* Providence, RI: Brown University. ICAME CD-Rom.

Goertzel, Ben 1993. *The Evolving Mind.* Langhorne, PA: Gordon and Breach.
http://www.goertzel.org/books/mind/contents.html

Greenwald, Lloyd N/A. Wall Street Journal Corpus.
http://plan.mcs.drexel.edu/courses/ml/software/

Herman, Ellen 1995. *The Romance of American Psychology: Political Culture in the Age of Experts.* Berkeley: University of California Press.
http://ark.cdlib.org/ark:/13030/ft696nb3n8/

Hundt, Marianne, Andrea Sand, and Paul Skandera 1999. *The Freiburg-Brown Corpus of American English* Freiburg: Albert-Ludwigs-Universität. ICAME CD-Rom.

Ide, Nancy, and Randi Reppen, Keith Suderman 2003. *Switchboard Corpus*, from *The American National Corpus.* American National Corpus Project, CD-Rom.

Krupat, Arnold 1992. *Ethnocriticism: Ethnography, History, Literature.* Berkeley: University of California Press. http://ark.cdlib.org/ark:/13030/ft9m3nb6fh/

Lodish, Harvey 2003. *Molecular Cell Biology.* New York: W.H. Freeman and Company. http://www.ncbi.nlm.nih.gov/books/

Sahtouris, Elisabet 1996. *Earthdance: Living Systems in Evolution.* Santa Barbara, CA: Metalog Books. http://www.ratical.com/LifeWeb/Erthdnce/erthdnce.html

Sinclair, John, ed. 1995. *Collins Cobuild English Dictionary.* London: Harper Collins.

Stephenson, Neal 1999. *In the beginning. . . was the command line* New York: Avon Books. http://www.cryptonomicon.com/beginning.html

Von Mises, Ludwig 1979. *Economic Policy: Thoughts for Today and Tomorrow.* South Bend, Ind.: Regnery/Gateway. http://www.mises.org/etexts/ecopol.asp

Appendix A

Screen Shots

Note: Screen shots of all the forms in Phraser are shown in this appendix. Figure A.1 shows the Main Form with the names of the visual components labeled. The other figures show the program exactly as it appears to the user. The Main Form is resizable, and users can maximize it if desired.

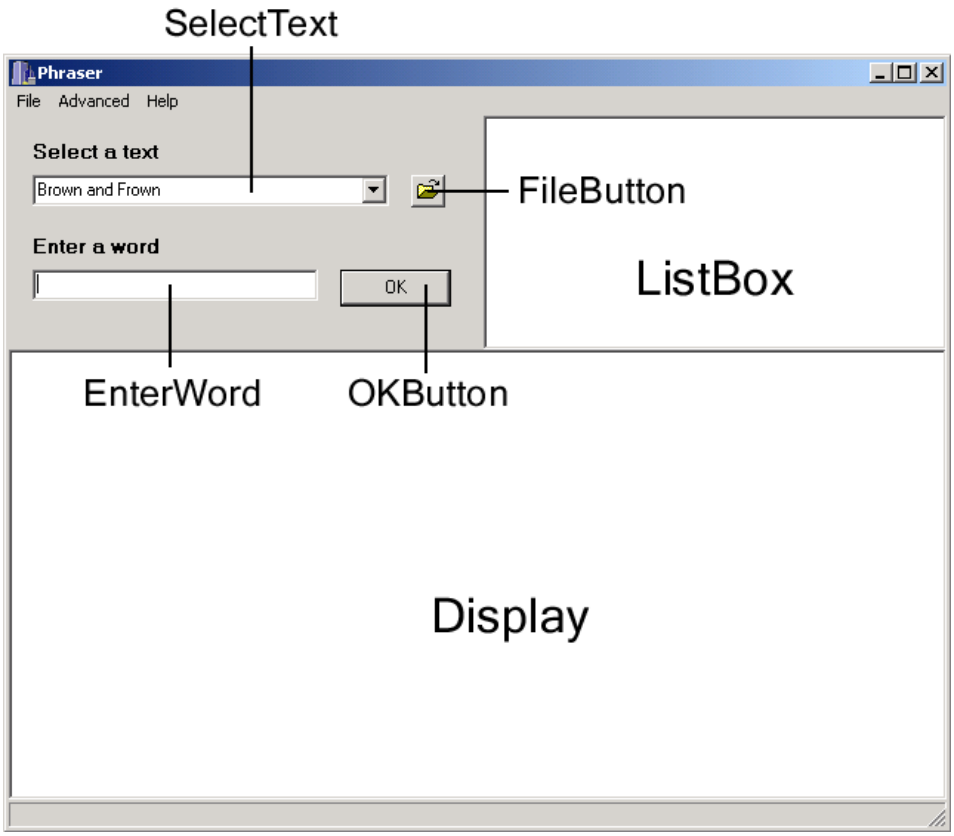

Figure A.1: Screen shot of the Main Form, with names of components labeled.

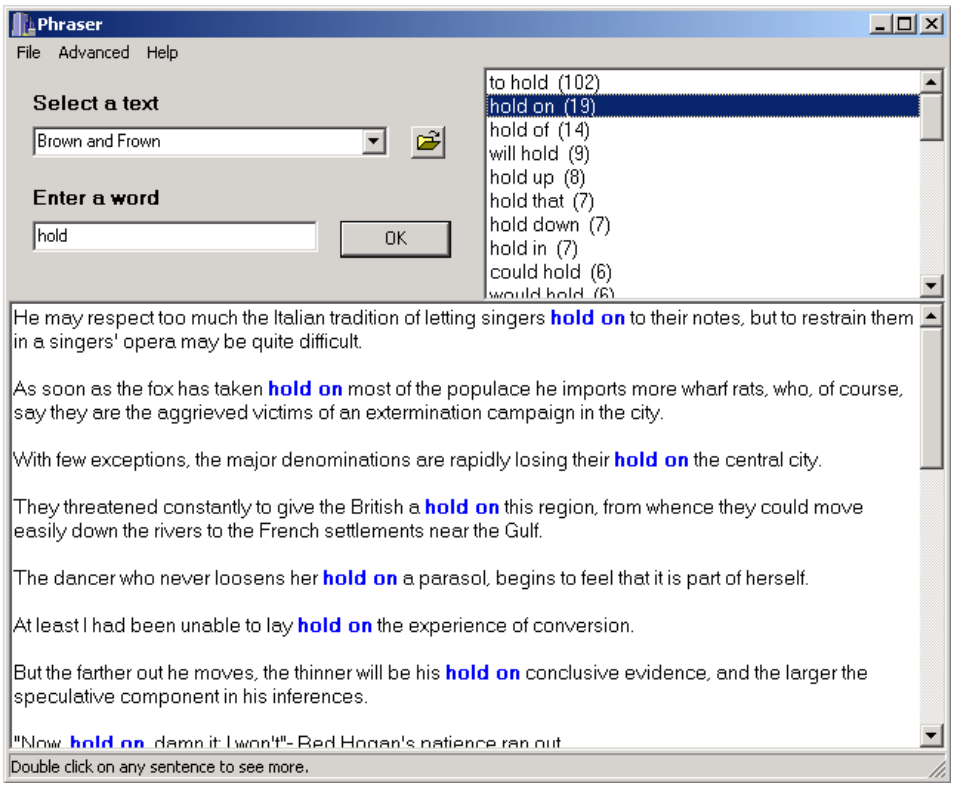

Figure A.2: Screen shot of the Main Form, with an example search for 'hold.'

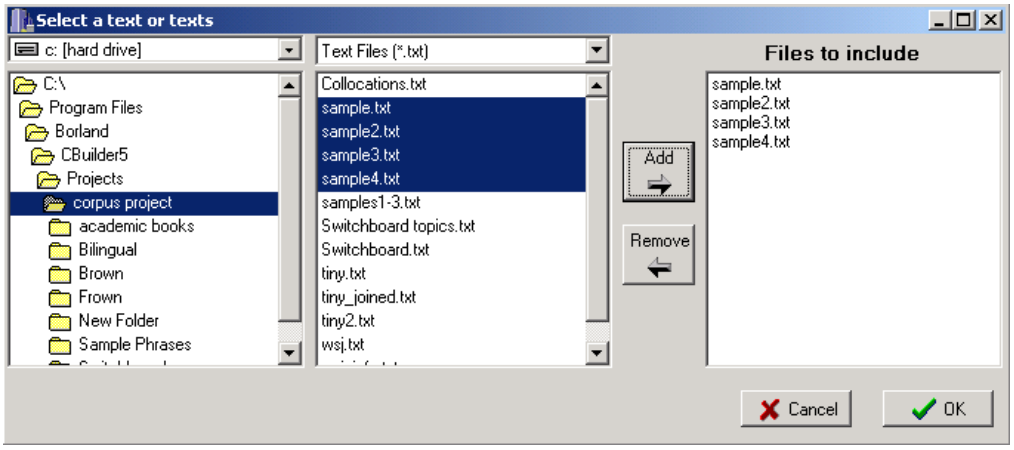

Figure A.3: Screen shot of the Directory form.

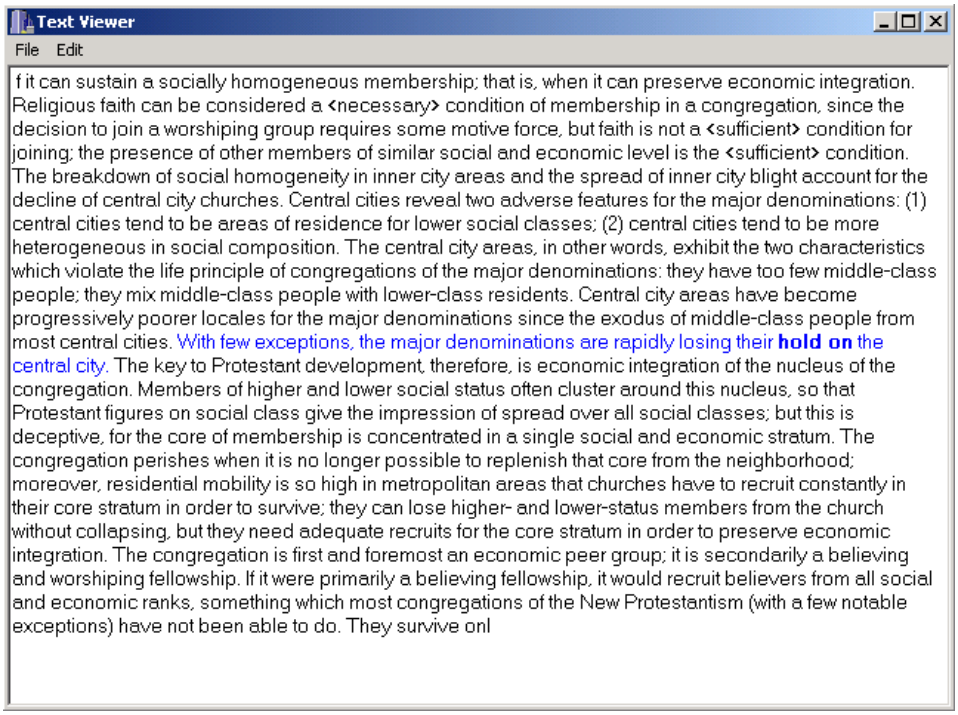

Figure A.4: Screen shot of the Viewer form, with an example from 'hold on.'

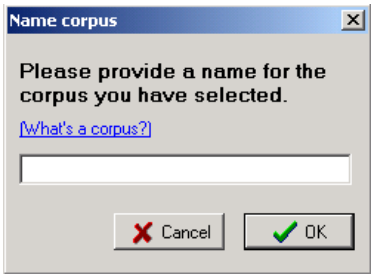

Figure A.5: Screen shot of the Name Corpus Form.

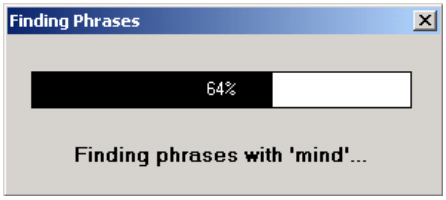

Figure A.6: Screen shot of the Progress Form, searching for 'mind'.

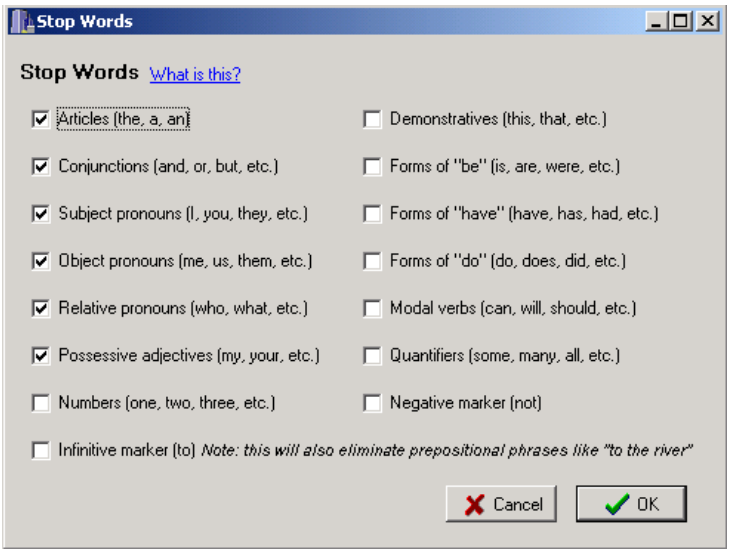

Figure A.7: Screen shot of the Stop Words Form.

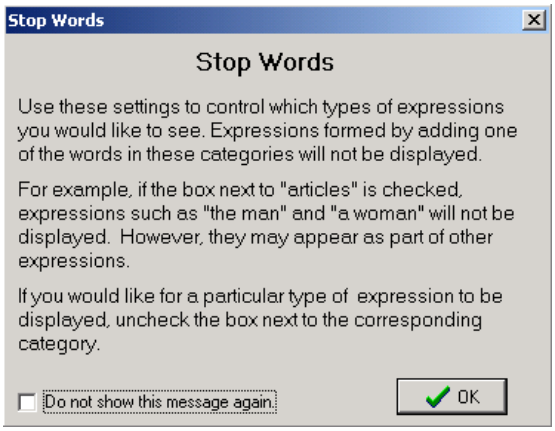

Figure A.8: Screen shot of the Stop Words Message.