The 385+ million word *Corpus of Contemporary American English* (1990–2008+)

Design, architecture, and linguistic insights

Mark Davies
Brigham Young University

The *Corpus of Contemporary American English* (*COCA*), which was released online in early 2008, is the first large and diverse corpus of American English. In this paper, we first discuss the design of the corpus — which contains more than 385 million words from 1990–2008 (20 million words each year), balanced between spoken, fiction, popular magazines, newspapers, and academic journals. We also discuss the unique relational databases architecture, which allows for a wide range of queries that are not available (or are quite difficult) with other architectures and interfaces. To conclude, we consider insights from the corpus on a number of cases of genre-based variation and recent linguistic variation, including an extended analysis of phrasal verbs in contemporary American English.

Keywords: corpus, American English, relational databases, diachronic, genres, phrasal verbs

1. Introduction

The British National Corpus has been the source for many corpus-related studies since its release in the early 1990s — perhaps more than any other corpus of English during this same period. As valuable as it is, however, the BNC is beginning to show its age in some respects. First, there have not been any substantive additions to the textual corpus since it was released in 1993, although some texts have been corrected during this time. Second, there is no planned expansion to the BNC in the future, which means that it will unfortunately become increasingly out of date with regards to recent changes in English. (To be fair, though, the BNC was never conceived of as, or promised to be, a monitor corpus.) Finally, although the 100 million word corpus was extremely large for its time (the early 1990s), with

International Journal of Corpus Linguistics 14:2 (2009), **159–190**. DOI 10.1075/ijcl.14.2.02dav
ISSN 1384–6655 / E-ISSN 1569–9811 © John Benjamins Publishing Company

current technology it is possible to create much larger corpora, which would possibly be of even more value to researchers.

Since the BNC was released more than fifteen years ago, researchers have envisioned something similar for American English, as well as other varieties of English. In the late 1990s, work began on the American National Corpus, which was projected to have 100 million words and which would have a textual composition comparable to that of the BNC. However, ten years later, only a small portion of the ANC has been completed (approximately 22 million words), and substantive work on the corpus appears to have ended. In addition to the small size, the ANC also has a very limited set of genres, compared to the BNC. For example, there are only two magazines in the corpus, one newspaper, two academic journals, and the fiction texts represent only about half a million words of text (compared to 16 million words for the BNC). On the other hand, certain genres seem to be overrepresented. For example, nearly fifteen percent of the corpus comes from one single blog, which deals primarily with the teen movie *Buffy the Vampire Slayer*.

In this paper, we will discuss the Corpus of Contemporary American English (hereafter COCA) (http://www.americancorpus.org), a new corpus that has recently been placed online, and which is intended to compensate for the limitations of the two corpora just mentioned. The corpus contains more than 385 million words of American English from 1990 to 2008. There are 20 million words for each of these nineteen years, and 20 million words will be added to the corpus each year from this point forward. In addition, for each year the corpus is evenly divided between spoken, fiction, popular magazines, newspapers, and academic journals. In terms of the textual corpus, the corpus contrasts with the ANC in that it is the first large corpus of American English that contains texts from a wide range of genres. On the other hand, the corpus contrasts with the BNC in the sense that it is the only large publicly-available corpus of English that contains texts from the past fifteen years.

COCA uses the same architecture and web interface that we have created for other large corpora that we have placed online. Therefore, although we will refer to the "COCA architecture and interface" throughout this paper, this is simply a shorthand abbreviation for the architecture and interface used by COCA, yet which are also used for a wide range of corpora that we have created and placed online (see http://corpus.byu.edu). These corpora include BYU-BNC (our interface for the 100 million word British National Corpus), the TIME Corpus (100 million words of American English, 1920s–2000s), the Corpus del Español (100 million words, 1200s–1900s), and the Corpus do Português (45 million words, 1300s–1900s). By late 2009, we will have also placed online a genre-balanced, 300 million word corpus of American English from the early 1800s to the current time.

In this paper, we will first focus on the design and construction of the corpus, and show how with a relational database design we can acquire, store, and organize large amounts of texts with relative ease. We will then discuss the corpus architecture, and how the relational database architecture allows for an essentially unlimited number of levels of annotation, while still providing for very good performance on the large corpus. Finally, we will discuss ways in which the corpus architecture and interface work together to provide users with access to a number of types of queries which are not possible (or quite difficult) with most other architectures. To conclude, we will provide a somewhat extended analysis of phrasal verbs in contemporary American English, to show how the corpus can be used to carry out detailed analyses of current genre-based variation and recent linguistic shifts.

2. The composition of the corpus

The corpus was designed to be roughly comparable to the BNC in terms of text types. In the BNC, approximately 10% of the texts come from spoken, 16% from fiction, 15% from (popular) magazines, 10% from newspapers, and 15% from academic, with the balance coming from other genres. In the COCA, texts are evenly divided between spoken (20%), fiction (20%), popular magazines (20%), newspapers (20%) and academic journals (20%). This composition holds for the corpus overall, as well as for each year in the corpus. This means that researchers can compare data diachronically across the corpus, and be reasonably sure that the equivalent text composition from year to year will accurately show changes in the language.

As of October 2008, there are more than 150,000 texts in the corpus, and they come from a variety of sources:

Spoken (79+ million words): Transcripts of unscripted conversation from more than 150 different TV and radio programs (examples: *All Things Considered* (NPR), *Newshour* (PBS), *Good Morning America* (ABC), *Today Show* (NBC), *60 Minutes* (CBS), *Hannity and Colmes* (Fox), *Jerry Springer*, *Oprah,* etc).
Fiction (76+ million words): Short stories and plays from literary magazines, children's magazines, popular magazines, first chapters of first edition books 1990–present, and movie scripts.
Popular Magazines (81+ million words): Nearly 100 different magazines, with a good mix (overall, and by year) between specific domains (news, health, home and gardening, women, financial, religion, sports, etc). A few examples are *Time, Men's Health, Good Housekeeping, Cosmopolitan, Fortune, Christian Century, Sports Illustrated*, etc.

Newspapers (76+ million words): Ten newspapers from across the US, including: *USA Today, New York Times, Atlanta Journal Constitution, San Francisco Chronicle*, etc. There is also a good mix between different sections of the newspaper, such as local news, opinion, sports, financial, etc.
Academic Journals (76+ million words): Nearly 100 different peer-reviewed journals. These were selected to cover the entire range of the Library of Congress classification system (e.g. a certain percentage from B (philosophy, psychology, religion), D (world history), K (education), T (technology), etc.), again with a good mix both overall and by number of words per year.

For each year, the texts within each of the five genres are balanced between the sub-genres or domains just mentioned. For example, each year the newspapers are evenly divided between the ten newspapers (approximately 400,000 words each); approximately 10% of the fiction texts (400,000–500,000 words each year) come from movie scripts; the popular magazines maintain roughly the same composition from year to year (African-American, current events, sports, science, religion, health, etc.), and the same is true for the academic journals (science, history, religion and philosophy, technology, education, etc). For users who wish to obtain more detailed information on the composition of the corpus, it is possible to download files from the corpus website that show detailed information (source, title, author, pages, etc) for each of the 150,000+ texts in the corpus.

Some might wonder about the spoken texts, since these are based almost entirely on transcripts of *unscripted* conversation on television and radio programs. First, are they accurate — do they accurately reflect what is found on the original audio or video recording? (This is a serious problem with transcripts of the British parliament, as discussed in Mollin 2007). Second, are they really spontaneous (as would be hoped), or is there too much scripted material? Third and most importantly, do they represent well what we would find in "non-media" conversations, such as the type of conversations found in the BNC? These are all important questions, and we feel very confident that the nearly 80 million words of spoken text in COCA are indeed 1) very accurate 2) almost completely spontaneous, and 3) they do represent well non-media English. It is impossible to address each of these points at length in this article. However, a full discussion of these points, with all of the relevant links to samples of the audio and video files and unedited transcripts, as well as sample queries of the spoken corpus that show how well it does represent informal spoken English ("… you know …", "… well…", "…I mean…") can be found at the corpus website, via the link [More information / Texts / Spoken transcripts].

There might also be questions about the lack of Internet-based sources like emails, listservs, and blogs. There were two main reasons for not including these.

First, to facilitate diachronically-focused studies, we wanted to make sure that we had the same composition in the corpus from year to year, as was just discussed. While we might have been able to get listservs from 1990 to the present time, it would have been very difficult to get the same amount of emails for each year since 1990, and this would have of course been completely impossible for blogs (which didn't exist until the early 2000s). In any case, we likely could not have acquired 20 million words of "Internet"-genre texts for most of the years in the corpus, to match the size of the other five genres. Second, because this was the Corpus of Contemporary *American* English, we needed to limit the corpus to material produced in the United States, and with blogs, listservs, and emails, this is difficult (if not impossible) to control.

3. Creating the textual corpus

Obviously, one of the advantages of creating a corpus in 2008 — as opposed to 15–20 years ago, when the BNC was being constructed — is the amount of material that is already in electronic format and accessible via the Web. Two or three examples of materials from COCA may suffice. First, for the spoken material, sites like CNN have essentially all of their transcripts available back to at least 2000 (see http://transcripts.cnn.com). In the case of these CNN transcripts, there are more than 170 million words of text for the period 2000–2008, but only about 6 million words thereof were used for COCA. Second, many magazines and newspapers are now placing large archives of past issues online. To give just three examples, it is possible to access online all articles from Sports Illustrated (back to the 1950s; http://vault.sportsillustrated.cnn.com/), all articles from TIME (back to the 1920s; http://www.time.com/time/archive/), and all newspaper articles from the New York Times (http://query.nytimes.com/search/archive.html). In the case of fiction, it is possible to retrieve the first chapters from thousands of novels via several different websites (such as Barnes and Noble), and one can also download thousands of studio-version movie scripts from sites like Simply Scripts (http://www.simplyscripts.com/movie.html).

Many of the 150,000+ texts for the corpus, however, were downloaded from text archives that have full text of TV and radio transcripts, short stories, magazines, newspapers, and academic articles from thousands of different sources. While some of the materials were retrieved manually, others were retrieved automatically. Using VB.NET (a programming interface and language), we created a script that would check our database to see what sources to query (a particular magazine, academic journal, newspaper, TV transcript, etc) and how many words we needed from that source for a given year. The script then sent this information to Internet Explorer, which would enter that information into the search form at the text archive, check to see if we already had the articles that would be retrieved by the query, and (if not) then retrieve the new article(s). In so doing, it would store all of the relevant bibliographic information (publication data, title, author, number of words, etc.) in the database. It would continue this process until it reached the desired number of words for a particular source in a particular year.

One of the advantages of this approach is that the text from the articles was stored in the database alongside the metadata, and this facilitated the processing of a large amount of texts from many different sources. Had these 150,000 or so texts been stored as distinct files on the computer, this would have been much more difficult. The other advantage of this database-driven approach is that it will be quite easy in the future to add to COCA. We plan to add 20 million words of text each year from this point on. We have already created the databases that show the lists of sources and number of words from each source, as well as the scripts to obtain these texts from the data sources. Adding another 10 million words every six months is just a matter of running the scripts overnight, cleaning the texts (beyond what our script does automatically), tagging them, and importing them into the corpus. As a sidelight, we might mention that we use CLAWS-7 to tag the texts (see http://ucrel.lancs.ac.uk/claws/). Because the hardware for the corpus server is quite robust, we were able to tag approximately 25 million words per hour, so tagging another 10 million words every six months would require less than half an hour.

With the creation of any large, web-accessible corpus based on contemporary materials, there is obviously a question about copyright. We have followed the same essential approach that we have used for other large online corpora that we have placed online since 2001 (see http://corpus.byu.edu). While we could allow users full text access to the corpus, we purposely chose to limit KWIC displays to a limited number of words. Because the end users do not have access to the full text, and because usage of the corpus is logged, it would be very difficult for an end user to re-create a page of text, much less the entire article or book. In terms of US Fair Use Law, then, there is essentially no competition with and no adverse economic impact on the copyright holder. This 'snippet defense' is similar to the one used by Google and Google Books, and it has worked well for us as well since 2001.

4. Corpus architecture

The architecture for these corpora is based on extensive use of relational databases, and is an updated version of the architecture described in Davies (2009), and of a much earlier version described in Davies (2005). The main [seqWords] database contains a table with one row for each token in the corpus in sequential order (i.e. 385+ million rows for a 385+ million word corpus, such as COCA). The table (see Table 1) contains an [ID] column that shows the sequential position of each word in the corpus (1, 2, 3, … 385,000,000), a [wordID] column with the integer value for each unique type in the corpus (wordID), and a [textID] number that refers to one of the 150,000+ texts in the corpus:

Table 1. Main [seqWords] table

ID	textID	wordID
359653867	1034159	539
359653868	1034159	305
359653869	1034159	12799
359653870	**1034159**	**58779**
359653871	1034159	3
359653872	1034159	2636

The 'dictionary' table (see Table 2) contains part of speech, lemma, and frequency information for each of the 2.3 million types in the corpus, and the [wordID] value in this table relates to the [wordID] value in the table above. An example is the following:

Table 2. [Dictionary] table

freq	wordID	Form	lemma	POS
1752	14892	Claws	claw	nn2
601	31607	claw	claw	nn1
258	55107	clawing	claw	vvg
231	**58779**	**clawed**	**claw**	**vvd**
143	78859	claw	claw	vvi
130	82796	claw	claw	nn1_vv0

The 'sources' table (see Table 3) contains metadata on each of the 150,000+ texts in the corpus, and contains information on such things as genre, sub-genre, title, author, source information (e.g. magazine, issue, and pages), e.g.:

Table 3. [Sources] table

textID	Year	genre	sub-genre	Title	Author
1030037	2005	FIC	Novel	Mary, Mary	Patterson, James
1031736	2000	FIC	Novel	Joe College	Perrotta, Tom
1032934	2003	FIC	Novel	The Orion Protocol	Tigerman, Gary
1034159	**2001**	**FIC**	**Novel**	**Deep South**	**Barr, Nevada**
1031737	2000	FIC	Novel	The accidental bride	Harayda, Janice
1031741	2000	FIC	Novel	Timbuktu : a novel	Auster, Paul

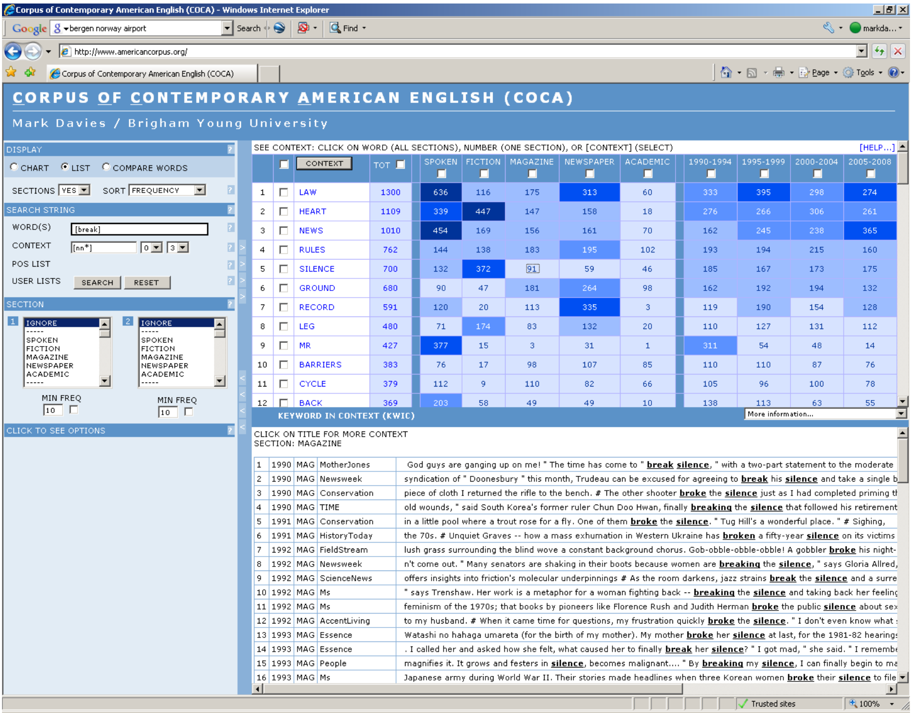
There are also other tables that contain supplementary word-level information. These include a ‘synonyms’ table with more than 370,000 entries (synonyms for more than 30,000 words) and tables for customized wordlists created by the corpus users (to be discussed in more detail below).

In our estimation, the relational database architecture allows a number of significant advantages over competing architectures. The first is speed and size. Because each of the tables is indexed (including the use of clustered indexes), queries of even large corpora are very fast. For example, it takes just about 1.3 seconds to find the top 100 noun collocates after the 23,000 tokens of *white* in the 100 million word BNC (*paper*, *house*, *wine*), and this increases to just 2.1 seconds for the 168,000 tokens of *white* in the 385+ million word American Corpus. Another example is that it takes about 1.2 seconds to find the 100 most frequent strings for *[end] up [vvg]* in the BYU-BNC corpus (*end up paying*, *ended up going*), and this is the same amount of time that it takes in the 385 million word American Corpus as well. In other words, the architecture is very scalable, with little or no decrease in speed, even as we move from a 100 million word corpus to a 385+ million word corpus. Even more complicated queries are quite fast. For example, *[[=clean]].[v*] the [n*]* searches for any form of any synonym of *clean* as a verb + *the* + a noun (*clean the house, scrubbing the sink, mopped the floor*), and it produces the 100 most frequent strings from COCA in less than two seconds.

The other main advantage of relational databases is that they allow for a ‘modular’ structure in which any number of additional feature sets can be incorporated into the architecture, with essentially no decrease in speed. One example of this is the synonyms-based query shown in the preceding paragraph. The [synonyms] table is linked to the [dictionary] table (see Table 2 above), and this table is in turn linked to the main [seqWords] table (Table 1 above), which contains each of the 385+ million words in context and in order. It would likewise be possible to add WordNet as another table (as we have already done for BYU-BNC), or WMatrix (see http://ucrel.lancs.ac.uk/wmatrix/), or CELEX (see http://www.ru.nl/celex/). Each table has its own index and there is very little “cost” in terms of the JOIN operations across tables. Imagine the complexity of such corpora with an XML format, where each feature (word form, POS, lemma, multiple synonyms, pronunciation, etc) is marked within the text itself. With relational databases, essentially any number of additional annotation features can be added via JOINed tables, and yet the architecture is very scalable, with very little performance hit for even very large corpora.

5. Corpus interface

COCA uses the same architecture and web interface as several other large corpora that we have placed online, which are listed above in Section 1. The following figure is a screenshot of the web interface, and it shows the main parts of this interface:


Corpus of Contemporary American English (COCA) - Windows Internet Explorer
Google
bergen norway airport
Search
Find
markda...
http://www.americancorpus.org/
Corpus of Contemporary American English (COCA)
Page
Tools
Trusted sites
100%

CORPUS OF CONTEMPORARY AMERICAN ENGLISH (COCA)

Mark Davies / Brigham Young University

DISPLAY

○ CHART ● LIST ○ COMPARE WORDS

SECTIONS YES SORT FREQUENCY

SEARCH STRING

WORD(S) [break]

CONTEXT [nn*] 0 3

POS LIST

USER LISTS SEARCH RESET

SECTION

1 IGNORE ----- SPOKEN FICTION MAGAZINE NEWSPAPER ACADEMIC -----

2 IGNORE ----- SPOKEN FICTION MAGAZINE NEWSPAPER ACADEMIC -----

MIN FREQ 10 □ MIN FREQ 10 □

CLICK TO SEE OPTIONS

SEE CONTEXT: CLICK ON WORD (ALL SECTIONS), NUMBER (ONE SECTION), OR [CONTEXT] (SELECT) [HELP...]

	□	CONTEXT	TOT □	SPOKEN □	FICTION □	MAGAZINE □	NEWSPAPER □	ACADEMIC □	1990-1994 □	1995-1999 □	2000-2004 □	2005-2008 □
1	□	LAW	1300	636	116	175	313	60	333	395	298	274
2	□	HEART	1109	339	447	147	158	18	276	266	306	261
3	□	NEWS	1010	454	169	156	161	70	162	245	238	365
4	□	RULES	762	144	138	183	195	102	193	194	215	160
5	□	SILENCE	700	132	372	91	59	46	185	167	173	175
6	□	GROUND	680	90	47	181	264	98	162	192	194	132
7	□	RECORD	591	120	20	113	335	3	119	190	154	128
8	□	LEG	480	71	174	83	132	20	110	127	131	112
9	□	MR	427	377	15	3	31	1	311	54	48	14
10	□	BARRIERS	383	76	17	98	107	85	110	110	87	76
11	□	CYCLE	379	112	9	110	82	66	105	96	100	78
12	□	BACK	369	203	58	49	49	10	138	113	63	55

KEYWORD IN CONTEXT (KWIC) More information...

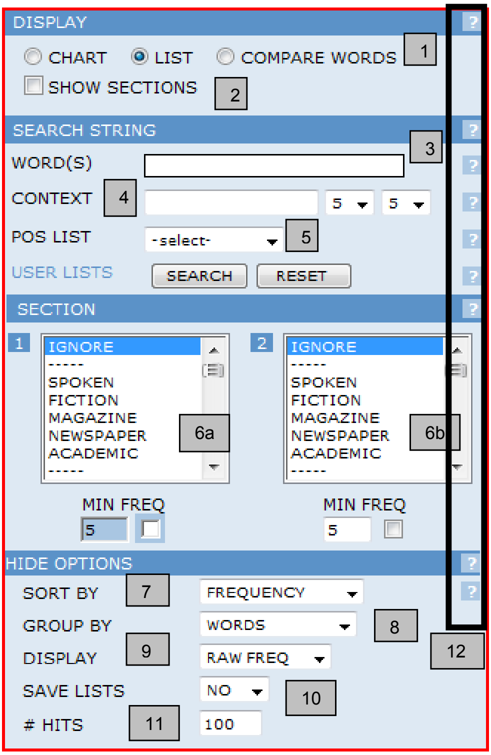
CLICK ON TITLE FOR MORE CONTEXT
SECTION: MAGAZINE

1	1990	MAG	MotherJones	God guys are ganging up on me! " The time has come to " **break** **silence**, " with a two-part statement to the moderate
2	1990	MAG	Newsweek	syndication of " Doonesbury " this month, Trudeau can be excused for agreeing to **break** his **silence** and take a single b
3	1990	MAG	Conservation	piece of cloth I returned the rifle to the bench. # The other shooter **broke** the **silence** just as I had completed priming th
4	1990	MAG	TIME	old wounds, " said South Korea's former ruler Chun Doo Hwan, finally **breaking** the **silence** that followed his retirement
5	1991	MAG	Conservation	in a little pool where a trout rose for a fly. One of them **broke** the **silence**. " Tug Hill's a wonderful place. " # Sighing,
6	1991	MAG	HistoryToday	the 70s. # Unquiet Graves -- how a mass exhumation in Western Ukraine has **broken** a fifty-year **silence** on its victims
7	1992	MAG	FieldStream	lush grass surrounding the blind wove a constant background chorus. Gob-obble-obble-obble! A gobbler **broke** his night-
8	1992	MAG	Newsweek	n't come out. " Many senators are shaking in their boots because women are **breaking** the **silence**, " says Gloria Allred,
9	1992	MAG	ScienceNews	offers insights into friction's molecular underpinnings # As the room darkens, jazz strains **break** the **silence** and a surre
10	1992	MAG	Ms	" says Trenshaw. Her work is a metaphor for a woman fighting back -- **breaking** the **silence** and taking back her feeling
11	1992	MAG	Ms	feminism of the 1970s; that books by pioneers like Florence Rush and Judith Herman **broke** the public **silence** about sex
12	1992	MAG	AccentLiving	to my husband. # When it came time for questions, my frustration quickly **broke** the **silence**. " I don't even know what :
13	1993	MAG	Essence	Watashi no hahaga umareta (for the birth of my mother). My mother **broke** her **silence** at last, for the 1981-82 hearings
14	1993	MAG	Essence	. I called her and asked how she felt, what caused her to finally **break** her **silence**? " I got mad, " she said. " I remembe
15	1993	MAG	People	magnifies it. It grows and festers in **silence**, becomes malignant.... " By **breaking** my **silence**, I can finally begin to ma
16	1993	MAG	Ms	Japanese army during World War II. Their stories made headlines when three Korean women **broke** their **silence** to file

Figure 1. Corpus interface

Users fill out the search form in the left frame of the window, they see the frequency listings or charts in the upper right-hand frame, and they can then click on any of the entries from the frequency lists to see the Keyword in Context (KWIC) display in the lower right-hand frame. They can also click on any of the KWIC entries to see even more context (approximately one paragraph) in this lower right-hand frame. Note also that there is a drop-down box (between the upper and lower right-hand frames) which provides help for many different topics.

As seen in Figure 2 below, users specify in the [DISPLAY] section (1) of the search form what type of display they want to see in the frequency listing: charts (showing the total for all matching strings), lists (individual entry for each matching string), or word comparisons (discussed below). In (2), they can choose to see the frequency of each matching word, phrase, or collocate in each of the five genres (spoken, fiction, popular magazines, newspapers, and academic journals) as well as the four time periods (1990–94, 1995–99, 2000–04, and 2005–07) for the [LIST] display.


DISPLAY

○ CHART ● LIST ○ COMPARE WORDS

□ SHOW SECTIONS

SEARCH STRING

WORD(S)

CONTEXT 5 5

POS LIST -select-

SECTION

1 IGNORE ----- SPOKEN FICTION MAGAZINE NEWSPAPER ACADEMIC -----

MIN FREQ 5 □

2 IGNORE ----- SPOKEN FICTION MAGAZINE NEWSPAPER ACADEMIC -----

MIN FREQ 5 □

HIDE OPTIONS

SORT BY FREQUENCY

GROUP BY WORDS

DISPLAY RAW FREQ

SAVE LISTS NO

HITS 100

Figure 2. Search form

The actual query (word, phrase, synonym set, grammatical construction, etc.) is entered into the [WORD] section (3). In Section (4), users can indicate collocates/contextual information, which is the main way of searching for collocates, and they can indicate the size of the collocation span. The [POS LIST] (5) gives a drop-down list that allows users to select from among more than fifty part of speech tags, which are then input into the search form (3 or 4).

In the [SECTIONS] part of the search form (6a, 6b), users can limit the query to a particular part of the corpus (e.g. FICTION, MAG:Sports, 1990–94, 2008, or any combination of sections), and via the [MIN FREQ] section below that, they can indicate the minimum number of times that it needs to occur in that section of the corpus. Via the second [SECTION] part of the search form (6b), users can select a second group of sections against which to compare the results from the first group (6a) (e.g. FIC:Movies vs. FICTION, MAG:Sports vs. MAGAZINES, or 1990–1999 vs. 2000–2008). Detailed examples of these types of queries will be provided in Section 12 of this paper.

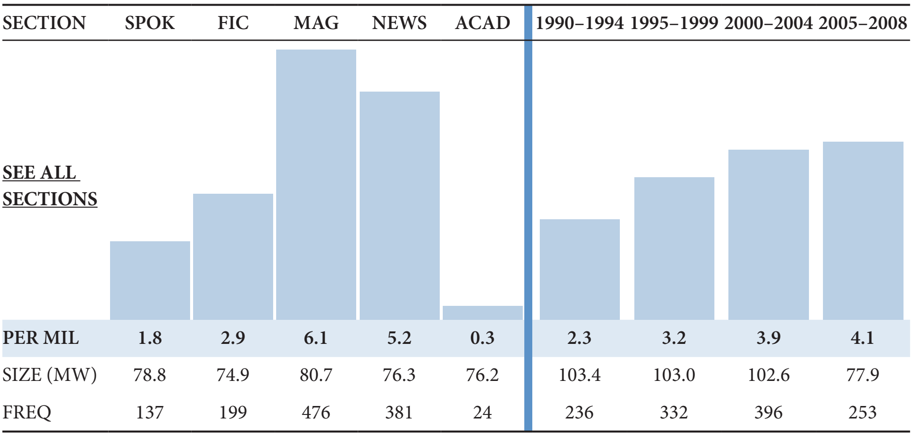
There are more options via the [OPTIONS] section of the search form. First, users can select how to sort the results (7) — by raw frequency or by 'relevance', which is Mutual Information score (for collocates) or A/B contrasts (e.g. words in one section of the corpus but not in the other, or collocates that occur with one word but not with a second one). Users can also choose how to group the results (8) — by lemma (e.g. *watch*), words (e.g. *watch, watched, watches* as separate entries), or no grouping (e.g. two entries for *watches*, as noun and as verb). In (9) they can choose whether to see raw frequencies, tokens per million, or a combination of these. In (10) they can choose to save their results to the database and re-use them in later queries, and in (11) they can choose how many entries to see (1–1000). Note also that all of the question marks (12) take the user to context sensitive help pages that explain that particular option in the search form.

6. Basic query syntax

The query syntax allows for a wide range of searches, including words, phrases, substrings, parts of speech, lemma, collocates, synonyms, customized word lists, limits by genre and by time period, or any combination of these. Appendix 1 gives an overview of the possible types of searches, and more detailed discussion will be found in the sections that follow.

7. Simple frequency queries (charts)

Perhaps the most basic type of search is one that finds the overall frequency of a word, phrase, substring, or grammatical construction in the five main genres (spoken, fiction, popular magazines, newspapers, and academic journals) and the four time periods represented in the corpus (1990–94, 1995–99, 2000–04, and 2005–08). To see this frequency data, users simply select [CHART] in the search form and then input the word, phrase, or grammatical construction. For example, if users search for the word *funky*, they will see:



SECTION	SPOK	FIC	MAG	NEWS	ACAD	1990–1994	1995–1999	2000–2004	2005–2008
PER MIL	1.8	2.9	6.1	5.2	0.3	2.3	3.2	3.9	4.1
SIZE (MW)	78.8	74.9	80.7	76.3	76.2	103.4	103.0	102.6	77.9
FREQ	137	199	476	381	24	236	332	396	253

Figure 3. Frequency by genre and year ("*funky*")

This indicates that the word nearly doubled in frequency between 1990–94 and 2000–04, but that the increase has slowed since that time. In addition, we see that the word is most frequent in magazines, followed by newspapers, fiction, spoken, and (at a very low frequency) academic.

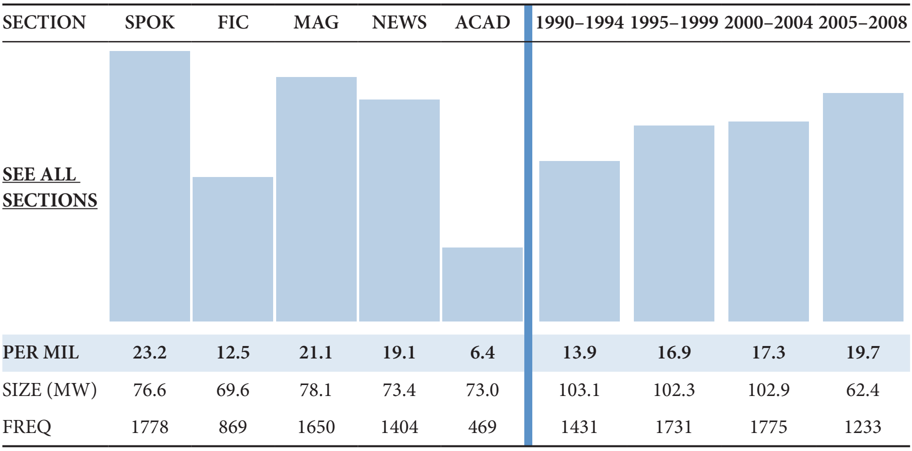
Users can click on [SEE ALL SECTIONS] to find the frequency in nearly fifty different genres. For *funky*, this table would show the following (just the first few rows are shown in Table 4):

Table 4. Frequency by sub-genre ("*funky*")

	SECTION NAME	# PER MILLION	# TOKENS	# WORDS
1	MAG:Entertain	27.0	94	3,479,537
2	MAG:Afric-Amer	16.7	56	3,357,201
3	NEWS:Life	11.8	152	12,883,819
4	FIC:Movies	7.7	71	9,208,594
5	MAG:Sports	7.4	70	9,440,867
6	NEWS:Misc	6.7	165	24,691,471
7	MAG:Home/Health	6.3	118	18,822,071

This shows that the word is most frequent in magazines (especially entertainment, African-American, sports, and home and health magazines), newspapers ('lifestyle' sections), and movies.

The chart displays are useful for much more than just isolated words and phrases. They can be used to see the frequency of virtually anything that can be entered into the search form, including syntactic constructions. Users could enter *[vv*]* to see the overall frequency of lexical verbs in each section, or *[vv*] about ./,* to see the frequency of preposition stranding with the preposition *about* (before a full stop or a comma), or virtually any other construction. With access to frequency charts by genre, even beginning users can easily replicate the types of searches used for corpus-based books like the *Longman Grammar of Spoken and Written English* (Biber et al. 1999). To show just one example, the following is the chart for *[end] up [vvg]* (all forms of *end* + *up* + VVG verb form: *ended up watching, ends up paying*, etc):



SECTION	SPOK	FIC	MAG	NEWS	ACAD	1990–1994	1995–1999	2000–2004	2005–2008
PER MIL	23.2	12.5	21.1	19.1	6.4	13.9	16.9	17.3	19.7
SIZE (MW)	76.6	69.6	78.1	73.4	73.0	103.1	102.3	102.9	62.4
FREQ	1778	869	1650	1404	469	1431	1731	1775	1233

Figure 4. Frequency by genre and year ([end] up [VVG])

This shows, among other things, that the construction has increased in frequency about 50% during the past two decades. This fits in well with the data from our 100 million word TIME Corpus (http://corpus.byu.edu/time/), which shows a steady increase in the construction since it arose in the early 1900s.

8. More advanced frequency queries

Once users see the overall frequency of a construction, as in Figure 4 above, they can click on any of the bars to see the most frequent matching strings in that genre or time period. For example, by clicking on the [SPOKEN] bar in the figure above, one would see the following table, which shows the most frequent strings for *[end] up [vvg]* in the spoken texts (just the first few entries are shown in Table 5):

Table 5. List of all matching forms ([end] *up* [VVG])

	TOTAL	SPOK	FIC	MAG	NEWS	ACAD	1990–	1995–	2000–	2005–
							94	99	04	07
1 END UP GETTING	151	**76**	17	24	30	4	39	37	46	29
2 END UP PAYING	187	**68**	7	48	54	10	58	53	47	29
3 ENDED UP GETTING	153	**61**	21	23	39	9	31	35	46	41
4 END UP GOING	93	**37**	7	23	20	6	18	35	25	15
5 ENDED UP GOING	102	**33**	19	24	25	1	16	33	30	23

Rather than going through the chart displays, however, it is possible to go directly to the table or list display, and by so doing there are other options available to the user as well. To do this, users simply select [LIST] rather than [CHART] in the search form. For example, users could look for *[j*] smile*, and they would see results like the following (note that these are color-coded on the web interface to show relative normalized frequency, and that it is also possible to see the figures for tokens per million words):

Table 6. List of all matching forms ([J*] *smile*)

		TOTAL	SPOK	FIC	MAG	NEWS	ACAD	1990–	1995–	2000–	2005–
								94	99	04	07
1	BIG SMILE	441	85	226	57	64	9	109	125	118	89
2	LITTLE SMILE	268	24	208	27	6	3	96	64	72	36
3	SLIGHT SMILE	184	4	149	14	12	5	58	39	48	39
4	WRY SMILE	183	2	118	26	35	2	59	44	45	35
5	BROAD SMILE	167	5	86	34	41	1	54	36	38	39

Users can also choose to have the results sorted by Mutual Information score, and to set a lower bound on the number of tokens. For the preceding query, the following would be the MI-ranked results with a lower bound of ten tokens (note that users can also see the frequency for each section of the corpus, as shown in the preceding table, but not shown in Table 7):

Table 7. List of all matching forms, sorted by Mutual Information score ([J*] *smile*)

		TOTAL	ALL	%	MI
1	WRY SMILE	183	903	20.3	8.97
2	RUEFUL SMILE	56	278	20.1	8.97
3	GAP-TOOTHED SMILE	19	95	20.1	8.96
4	BEATIFIC SMILE	34	189	18.0	8.85
5	TOOTHY SMILE	38	237	16.0	8.74
6	IMPISH SMILE	27	200	13.5	8.57

In the case of *wry smile*, for example, there are 183 tokens in the corpus. There are 903 tokens with *wry*, so that 20.3% of all of the occurrences of *wry* are with *smile*, which yields a Mutual Information score of 8.97.

It is also possible to group the results by lemma or by word and part of speech. For example, the query *[vv*] * course* would produce results like *veer off course, veered off course, altered the course, altering the course*, etc. When the results are grouped by lemma, however, the results would be the following (where the brackets indicate lemma, and where the results are again sorted by Mutual Information score):

Table 8. Grouping by lemma ([VV*] * *course*)

		TOTAL	ALL	%	MI
1	[VEER] [OFF] [COURSE]	29	1704	1.7	6.16
2	[ALTER] [THE] [COURSE]	105	10749	0.98	5.61
3	[STAY] [THE] [COURSE]	460	93544	0.49	4.92
4	[STEER] [A] [COURSE]	25	5330	0.47	4.87
5	[CHART] [A] [COURSE]	58	12742	0.46	4.84

Clicking on the first entry would show KWIC entries like the following (a partial list here), which show the different forms for *veer*:

Table 9. Keyword in Context (KWIC) display

1	1990	ACAD	RehabResrch	Typically, when a vehicle becomes laterally unstable, it will rapidly **veer** **off** **course** if there is no steering. If someone is steering, the vehicle (if
7	1994	FIC	HarpersMag	happening. # It was barely detectable at first. He was **veering** **off** **course**, but just slightly, and it could have been no more than a trick
8	1994	MAG	Prevention	flag every intersection or fork in the road so we won't **veer** **off** **course**. # I glide through a shady grove. Then, as I emerge into
13	1999	SPOK	CBS_48Hours	not been a scenic one. It's a life that first **veered** **off** **course** years ago when her own mother walked out on her. I'm sure
18	2003	FIC	Today's Parent	Talk about moral issues with your child, especially if he's **veering** **off** **course**. And talk about your mistakes honestly, if he points those out. There
25	2006	SPOK	Fox_Cavuto	Right. Full disclosure before the fact. If somebody **veers** **off** **course**, tell them there. Don't wait a few years. Yes.

9. Collocates

The queries that we have considered to this point have involved single words (including part of speech or lemma, as well as substring), or phrases with a certain number of words. For example, the query that we have just discussed is a three word string composed of *[vv*] * course*. It is also possible to search for collocates anywhere within a ten word span to either side of the node word. To do a collocates-based search, users simply enter the node word or phrase in the WORD(S) part of the search form (see [3] in Figure 2 above) and the desired collocates expression in the CONTEXT part of the form (see [4] in Figure 2).

For example, to find the most frequent nouns near *thick*, users would enter *thick* into WORD(S) and *[nn*]* into CONTEXT, and would then see results like the following (note also that as with the 'slot-based' queries shown in the previous section, these results can also be sorted by Mutual Information score):

Table 10. Collocates display (Noun collocates of *thick*)

		TOTAL	SPOK	FIC	MAG	NEWS	ACAD	1990–94	1995–99	2000–04	2005–07
1	HAIR	935	15	747	101	60	12	242	241	280	172
2	INCH	440	34	33	225	136	12	81	137	142	80
3	AIR	382	17	258	57	39	11	113	113	99	57
4	SMOKE	291	50	150	43	42	6	85	65	90	51
5	GLASSES	285	14	200	33	28	10	78	83	74	50
6	INCHES	284	16	42	123	75	28	70	86	84	44
7	LAYER	248	7	66	103	33	39	62	66	71	49

The preceding example — a particular part of speech near a given word — is only the simplest of collocate-based queries. The query syntax and corpus architecture allow for a much wider range of collocate-based searches, both in terms of the node word as well as the collocates, and also as to how the results are sorted (raw frequency or MI score) and how they are grouped (by collocates alone or by the node word and the collocates, and by either word or lemma in either case). Examples of these are given in Appendix 2. In essence, the corpus architecture allows users to search for virtually anything "near" anything else in the 385+ million word corpus.

10. Word comparisons

Researchers have recognized the value of corpora in using collocates to tease apart slight differences between near-synonyms (e.g. *small* and *little*), or to provide insight into culturally-defined differences between two terms (e.g. *girls* and *boys*) (see, for example, Sinclair 1991 or Stubbs 1996). The architecture of COCA allows users to carry out searches like this quickly and easily, by comparing the collocates of two contrasting words or lemmas. For example, to compare the collocates of *small* and *little*, a user would simply select COMPARE WORDS, then enter *small* in one search field and *little* in the other, and then select [nn*] as for CONTEXT. Finally, s/he might specify that the first word (*small* or *little*) should occur at least 20 times with the given noun, while the opposing adjective occurs at least two times. The user would then see the following:

Table 11. Word comparisons (small / little [NN*])

WORD 1 (W1): SMALL (0.55)					
	WORD	**W1**	**W2**	**W1/W2**	**SCORE**
1	AMOUNTS	859	3	286.3	521.6
2	PERCENTAGE	927	4	231.8	422.2
3	FRACTION	532	5	106.4	193.8
4	FIRMS	390	4	97.5	177.6
5	BUSINESSES	2210	24	92.1	167.7
6	FARMERS	492	6	82.0	149.4
7	SCALE	400	7	57.1	104.1
WORD 2 (W2): LITTLE (1.82)					
	WORD	**W2**	**W1**	**W2/W1**	**SCORE**
1	WHILE	2208	4	552.0	303.0
2	BIT	16425	90	182.5	100.2
3	BROTHER	1073	6	178.8	98.2
4	TROUBLE	496	4	124.0	68.1
5	SISTER	929	8	116.1	63.7
6	ATTENTION	1377	13	105.9	58.1
7	FUN	292	3	97.3	53.4

Table 11 shows that there are .55 tokens of small for every token of little in the corpus, and 1.82 tokens of little for every token of small. Therefore, all other things being equal, any noun collocate would occur about half as much (.55) with small as with little. In the case of firms, however, (#4 on the left side of the table), there are about 98 tokens with small for every token with little (W1/W2), which is about 178 times the expected rate (SCORE). In the case of sister, on the other hand, this collocate occurs about 64 times as frequently with little (vs. small) than the overall frequency of these two words would predict.

The following table provides a few additional examples of word comparisons that can be done with the corpus. [A] and [B] refer to the two words being compared, collocate POS shows the part of speech of the collocates, and the rightmost two columns show the collocates that occur with either [A] or [B] much more than the overall frequency of either of these two words would suggest.

Table 12. Examples of word comparisons

	[A]	[B]	**Collocate POS**	**Collocates with [A]**	**Collocates with [B]**
1	[boy]	[girl]	[j*]	growing, rude	sexy, working
2	Democrats	Republicans	[j*]	open-minded, fun	mean-spirited, greedy
3	Clinton	Bush	[v*]	confessed, groped, inhale	assure, deploying, stumbles
4	utter.[j*]	sheer	[nn*]	silence, despair	beauty, joy
5	ground.[n*]	floor.[n*]	[j*]	common, solid	concrete, dirty
6	[rob].[v*]	[steal].[v*]	[nn*]	bank, store	cars, money

In summary, the simple yet quick word comparisons that are possible with this corpus would be of value to many different types of users. Linguists can quickly contrast synonyms, language learners can move beyond simple thesauruses to see differences between words, and even those in cultural studies, political science, and other social sciences can quickly and easily compare how contrasting words are used in contemporary American English.

11. Integrated thesaurus and customized wordlists

One of the advantages of using a relational database architecture is that it allows us to integrate into the architecture any number of new databases with additional annotation features. Perhaps the best example of this is the thesaurus that we have integrated into the architecture — a thesaurus that contains more than 370,000 synonyms for more than 30,000 different words. As we will see, the information from this thesaurus database can be used seamlessly as part of the query syntax.

In a typical thesaurus, users would see that the following are synonyms for *beautiful*: *wonderful, attractive, striking, lovely, handsome, charming, stunning, magnificent, gorgeous, superb, scenic, exquisite, delightful, pleasing, good-looking, picturesque,* and *fine-looking*. Obviously, however, some of these words are more frequent than others, and they would have a different distribution in different genres. An inexperienced language learner might end up sounding strange if s/he uses *exquisite* or *picturesque* much more than *beautiful* or *wonderful*.

COCA allows users to enter a simple query like [=beautiful], and then see the following (Table 13 shows just a partial listing of all of the synonyms):

Table 13. Synonyms list (partial listing for *beautiful*)

WORD(S)	TOTAL	SPOK	FIC	MAG	NEWS	ACAD	1990–	1995–	2000–	2005–
							94	99	04	07
BEAUTIFUL	36404	8297	13004	7809	4776	2518	8588	10552	10734	6530
		108.36	186.78	100.03	65.04	34.47	83.28	103.13	104.36	104.58
WONDERFUL	25523	10688	4848	4601	4160	1226	6821	8193	6934	3575
		139.59	69.63	58.93	56.65	16.78	66.14	80.07	67.42	57.25
ATTRACTIVE	10231	1308	1838	2922	1939	2224	3083	2849	2697	1602
		17.08	26.4	37.43	26.4	30.45	29.9	27.84	26.22	25.66
STRIKING	8956	1071	1128	2274	1883	2600	2645	2557	2425	1329
		13.99	16.2	29.13	25.64	35.59	25.65	24.99	23.58	21.28
LOVELY	8132	1584	3629	1642	999	278	2119	2450	2267	1296
		20.69	52.12	21.03	13.6	3.81	20.55	23.94	22.04	20.76
HANDSOME	6640	436	3452	1464	975	313	1830	1815	1853	1142
		5.69	49.58	18.75	13.28	4.28	17.75	17.74	18.02	18.29
CHARMING	4395	622	1619	1028	911	215	1158	1132	1199	906
		8.12	23.25	13.17	12.41	2.94	11.23	11.06	11.66	14.51
GORGEOUS	3718	947	1075	1028	598	70	738	952	1132	896
		12.37	15.44	13.17	8.14	0.96	7.16	9.3	11.01	14.35
SUPERB	2622	332	197	1089	715	289	871	687	703	361
		4.34	2.83	13.95	9.74	3.96	8.45	6.71	6.84	5.78
SCENIC	2034	71	123	941	661	238	596	587	544	307
		0.93	1.77	12.05	9	3.26	5.78	5.74	5.29	4.92
EXQUISITE	1971	106	564	708	352	241	527	553	550	341
		1.38	8.1	9.07	4.79	3.3	5.11	5.4	5.35	5.46

This table (which is about the most complex one that the user might see — most would be much simpler) contains a wealth of information. It shows all of the matching synonyms for *beautiful* in the thesaurus, along with their overall frequency and the frequency in each of the five main genres and the four time periods, as well as the tokens per million words (located below the raw frequency count for each word). A quick look at the color coding in the table shows, for example, that most of the synonyms are much less frequent in ACADEMIC than in the other registers. On the other hand, many of the synonyms are more frequent in FICTION, especially ones like *lovely, handsome,* and *charming*. One can also see some tentative trends regarding recent shifts with these words. *Gorgeous*, for example, has doubled in usage over the past eighteen years (7.16 tokens per million words in 1990–94 to 14.35 in 2005–07), while *superb* has decreased more than 30% during this same time (8.45 tokens per million words in 1990–94 to 5.78 in 2005–07).

In addition to seeing the frequency of the synonyms of a given word, as in the table above, it is also possible to include synonyms as part of more complex queries. For example, the simple query *[=clean].[v*]* would show that synonyms of *clean* as a verb are words like *wipe, dust, scrub, polish, cleanse, scour, mop, vacuum, launder,* and the users would see the distribution and frequency of each of these synonyms. However, this synonym information can be used as part of a more complex query, such as *[=clean].[v*] * [nn*]*, which would yield *clean the house, wiped the sweat, mopping the floors, dusts the shelves*, etc, along with the frequency and distribution of each string. In this way, users can move past simple form-based searches to ones that include a fairly robust semantic component.

One last feature of note is that it is possible for users to create their own customized wordlists, which they can again integrate seamlessly into the query syntax. There are two ways of creating these lists. First, they can save a subset of the words or phrases from an existing search. For example, they could search for the synonyms of *beautiful,* or *crash,* or *money*, and then save just the synonyms that are of interest to them. Similarly, they could find the collocates of a given word, and then save some of these collocates in their own wordlist. They could create from scratch a wordlist, such as emotions (*sad, happy, worried, ecstatic*, etc), colors (*blue, green, red*, etc), or parts of clothing (*shirt, blouse, suspenders, hat*, etc). In any of these cases, they simply create a name for the list and store it via the web interface under their chosen username.

These customized wordlists are saved in a database on the server, and can then be used a day, week, or year later as part of another query. For example, if a user *lingprof* creates a list for words related to emotions, s/he can then use these words as part of the query: *[r*] [lingprof:emotions] that*, to retrieve strings like *pretty worried that*, *quite sad that, extremely perturbed that,* etc. Likewise, these customized lists can be used as part of a collocates search. For example, the user *lingprof* might create a second list named *familyMember* (with *mother, mom, brother, uncle*, etc), and then search for any *familyMember* within six words of one of the *emotions* words, e.g. *her aunt was quite happy to see that*, *when Dad is as angry as that*, *they were excited that Mom could be there*, etc. Again, the ability to incorporate user-defined lists as part of the query, as well as the basic corpus architecture, allows users to carry out quite complex semantically-oriented queries on the corpus.

12. Searching by and contrasting sections of the corpus

One final feature of the corpus is the ability to sort by and compare by frequency in different sections of the corpus, and this is an outgrowth of the basic relational database architecture of the corpus. Each word in the corpus is clearly identified with one of the genres (and sub-genres), and the clustered indexes allow fast retrieval of data from the relevant section of the corpus. This allows users to very quickly compute the frequency of words and strings of words in different sections of the corpus and to compare these section frequencies to each other (for an overview of genre/register-based differences, see Biber et al. 1998: 32–51, Biber et al. 1999).

For example, users could find the most frequent words ending in *-ment* in ACADEMIC (*development, environment, government*), strings with *hard* + *[nn*]* in MAGAZINES (*hard work, hard drive, hard time, hard disk*), nouns near *chair* in FICTION (*back, table, desk, room*), (potential) synonyms of *smart* in FICTION (*bright, cool, quick, sharp*), adjectives in ACAD:Medicine (*other, significant, clinical, medical*), or nouns after forms of *get* in 2005–07 (*people, job, lot, way, money*).

The real power of section-based searching, however, is the ability to see what occurs in one section of the corpus, as opposed to another. For example, the following table compares the collocates of *chair* in FICTION and ACADEMIC, and clearly shows the very different word senses in the two sections:

Table 14. Comparison of collocates by section (noun collocates of *chair* in ACAD and FIC)

SEC 1: ACADEMIC

	WORD	SEC1	SEC2	PM1	PM2	RATIO
1	DEAN	25	2	0.34	0.03	11.91
2	BOARD	76	8	1.04	0.11	9.05
3	COLLEGE	25	3	0.34	0.04	7.94
4	SECTION	39	5	0.53	0.07	7.43
5	COUNCIL	14	2	0.19	0.03	6.67
6	MUSIC	27	4	0.37	0.06	6.43
7	CONFERENCE	19	3	0.26	0.04	6.04
8	COMMITTEE	145	23	1.99	0.33	6.01

SEC 2: FICTION

	WORD	SEC2	SEC1	PM2	PM1	RATIO
1	KITCHEN	197	2	2.83	0.03	103.35
2	LEATHER	209	3	3.00	0.04	73.10
3	LAWN	185	3	2.66	0.04	64.70
4	EYES	107	2	1.54	0.03	56.13
5	WINDOW	156	4	2.24	0.05	40.92
6	FATHER	78	2	1.12	0.03	40.92
7	SWIVEL	137	4	1.97	0.05	35.94
8	ARMS	170	5	2.44	0.07	35.67

As can be seen, the collocates of *chair* that occur much more in ACADEMIC than FICTION are *dean, board, college*, etc, while those in FICTION but not ACADEMIC are *kitchen, leather, lawn*, etc. The tables show the frequency of each collocate with chair in the two sections (e.g. 197 tokens of *kitchen* near *chair* in fiction but only 2 tokens of *kitchen* near *chair* in academic). These then are converted to tokens per million words in the two sections (2.83 in FICTION, .03 in ACADEMIC), and the ratio figure (103.35) is the ratio of the normalized tokens per million figures for the two sections. As can be seen in this table, the data clearly show that in academic texts, *chair* refers to the position on a committee, whereas in fiction texts it refers to the piece of furniture. All of this is accomplished via one simple query, with just a few clicks of the mouse.

Other examples of comparisons between two 'macro-genres' (spoken, fiction, popular magazines, newspapers, and academic) might be: *hard [nn*]* in MAGAZINES vs. SPOKEN (*hard disk / frost / snow / use / edges*), *-ment* words in ACADEMIC vs. FICTION (*underachievement, debridement, self-assessment, apportionment*), or synonyms of *smart* in NEWSPAPERS vs. ACADEMIC (*ritzy, nifty, brainy, stylish, glitzy, chic, trendy*). It is also possible to compare across time periods. For example, one could find synonyms of *smart* that have increased in usage from 1990–1999 to 2000–07 (*swanky, chic, stylish, ritzy*) or which have decreased in usage during that time (*impertinent, vigorous, shrewd, well-groomed, dashing*). Another example would be phrases with *green [nn*]* that have increased from the 1990s to the 2000s (*green zone / building / home / power*), which show the influence of the Iraq War or the environmental movement during this time.

These section comparisons can also be useful to find which words occur in a sub-section, compared to the larger section of which it is a part. Examples of this would be verbs in NEWS:Money compared to NEWSPAPERS overall (*restate, telecommute, bundle, hedge, digitize, liquidate*), adjectives in ACAD:Medicine vs ACADEMIC (*preoperative, parotid, endoscopic, laryngeal, histiopathologic*), or verbs in MAG:Sports compared to MAGAZINES overall (*re-sign, ski, blitz, grunt, punt, hunt, fish, pedal*). These easy to set up searches can produce lists that could be used by someone who is interested in ESP — English for Specific Purposes (cf. Gavioli 2005).

As one final note to this section, we should mention that because of the unique relational database architecture that we use, the ability to search in, limit by, and compare across sections of the corpus is perhaps more powerful than that of any other existing architecture for large corpora. There are five different architectures for large publicly-accessible corpora (of English). Two of them have just the British National Corpus — Just the Word (http://193.133.140.102/JustTheWord/), and Phrases in English (http://pie.usna.edu/). Three others have the BNC and other corpora as well — Sketch Engine (http://www.sketchengine.co.uk/), VISL (http://corp.hum.sdu.dk/), and BNCweb (http://bncweb.lancs.ac.uk/; where the IMS Corpus Workbench architecture has been used for other corpora as well). Three of these architectures — *Phrases in English*, *Just the Word*, and *VISL* — do not have any ability to limit searches by section of the corpus (e.g. genre or time period). With both BNCweb and Sketch Engine, users can limit searches by section, and the interface for Sketch Engine is somewhat less cumbersome in terms of identifying these sections. With both BNCweb and Sketch Engine, however, it is impossible to compare across registers. In other words, users can carry out a search on one section (call it Section 1) of the corpus, carry out a search on a second section (Section 2), and then load the results sets into some other program to compare the two sections. The users would then use that other program (probably a relational database) to see what is unique or much more common in Section 1. With our architecture, however, all of this can be done via the corpus interface very quickly, with just a couple clicks of the mouse.

13. A concrete example: Phrasal verbs in English

13.1 Phrasal verbs in contemporary American English

To conclude, let us now present a concrete example that shows in a somewhat more integrated way the power and functionality of the corpus and the corpus architecture. In the previous sections, we discussed several different phenomena that show how the corpus can be used to see the frequency of words, phrases, substrings, parts of speech, or lemmas (or any combination of these), as well as collocates and synonyms, and how these can be used to compare words, genres, and time periods. In this section, we will look at one single phenomenon, the use of phrasal verbs in contemporary American English, to see how all of this functionality can be joined together to look at the phenomenon from several different points of view.

Phrasal verbs are of course of interest for a number of reasons. First, as 'multiword expressions', they are on the interface of syntax and semantics, and these two domains interact in interesting ways. Second, from a pedagogical standpoint, phrasal verbs in English have long been recognized as a real area of difficulty for language learners. While there are many different learner-oriented dictionaries of phrasal verbs in English (e.g. *Longman Dictionary of Phrasal Verbs*, *Collins Cobuild Dictionary of Phrasal Verbs*, *NTC's Dictionary of Phrasal Verbs and Other Idiomatic Verbal Phrases*), almost none of them are based on the type of rich corpus data that one can obtain from a robust corpus like COCA (but see McCarthy & O'Dell 2004, as well as Gardner & Davies 2007). Third, there are clear differences between genres in terms of phrasal verbs, and there are interesting changes with these verbs over time as well (see Hiltunen 1994; Claridge 1997; Claridge 2000).

13.2 Basic frequency listings

In terms of obtaining data on phrasal verbs, users can enter the search string *[vvi] [rp*]* (infinitival form of lexical verbs + adverbial particle). They would then see the most frequent phrasal verbs across all registers — *find out*, *go back*, *come back*, *figure out*, *go out*, *pick up*, *come up*, etc. They can also use *[vv*] [rp*]* to search for all forms of a given verb (not just infinitival forms), and then have these grouped by lemma, yielding results like [*go*] *on* (*goes on, going on, went on*, etc), [*come*] *back*, [*come*] *up*, [*go*] *back*, [*pick*] *up*, [*find*] *out*, etc. And of course, they can limit the search to a particular adverbial particle, such as *up* (*grow up, set up, end up*), *down* (*sit down, break down, shut down*), *out* (*find out, point out, turn out*), or *over* (*take over, go over, look over*). It is also possible to search for separable phrasal verbs (*find it out*, *look the words up*) by using the [Context] searches. Users would input *[vv*]* as [Word] and *[rp*]* as [Context], anywhere within 3–4 words to the right of *[vv*]*.

One of the advantages of our corpus architecture is that it allows users to find *all* matching strings — not just those that occur above a certain frequency. For example, there are more than 3,100 distinct phrasal verbs (grouped by lemma) with the single particle *up*. But only about 1,460 of these 3,100 verbs occur three times or more, which is the threshold for inclusion in corpus architectures like the *Phrases in English* (PIE) interface to the BNC (see http://pie.usna.edu). More than half are very low frequency items like *gulp up* (2 tokens), *grovel up* (2), *cringe up* (2), *barf up* (2), *splotch up* (1), *sputter up* (1), *squeak up* (1), and *pumple up* (1). With a limited architecture like PIE, then, more than 50% of all phrasal verb types will be lost. And yet these very low frequency forms are often interesting in terms of new forms that are just barely entering into the language, or which are almost completely on the way out. Certainly we would want any architecture to find all of the relevant forms. Another important point is that any corpus that is much smaller than the 385+ million word COCA (such as the 100 million word BNC) is certainly going to miss many of these very low frequency items.

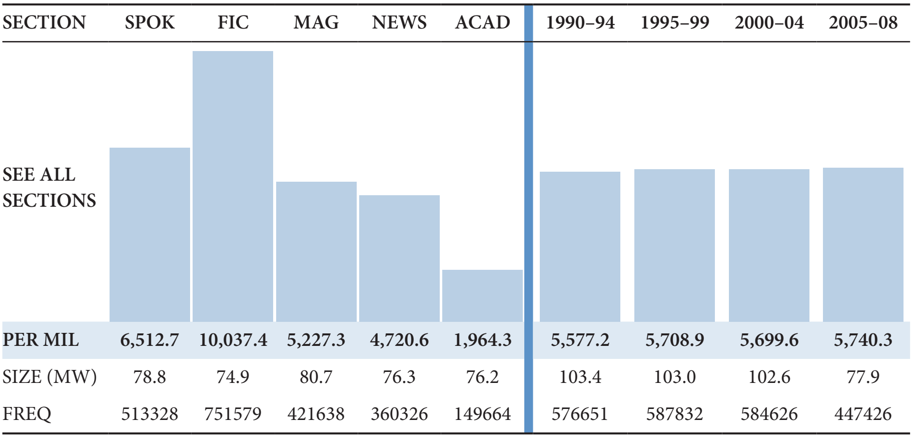
13.3 Semantically-oriented queries

The collocates feature of the corpus can also provide useful insight into the meaning and use of the different phrasal verbs. To find nominal collocates of *break down*, for example, users would simply enter *[break] down* (all forms of the lemma *break* + *down*) and then specify *[nn*]* in the [Context] field. They would then see the results list, with *barriers, car, system, door, time, tears, people, process*, etc, and these collocates could also be ranked by Mutual Information score. While it might seem trivial to find and sort collocates of phrasal verbs, this is actually something that is unique to the architecture used for COCA. All other architectures for large corpora either cannot do collocates (*VISL, Phrases in English*), or else they can find collocates only for single words (*Sketch Engine, BNCweb,* and *Just the Word*).

The ability to compare the collocates of different words and phrases can also provide useful insight into the meaning of competing phrasal verbs. For example, the difference between *burn up* and *burn down* might not be readily apparent (for example, a house can either "burn up" or "burn down"). With the [Compare Words] feature, however, we can easily compare the collocates to see the difference. *Burn up* takes the collocates *atmosphere, calories, fever, body, energy, fuel*, while *burn down* takes *houses, buildings, candles, barn, school*. Thus *burn up* tends to relate to noun phrases in nearby adverbial clauses or else to more figurative burning, while *burn down* tends to relate to more literal, direct objects of *burn*.

13.4 Recent linguistic shifts

As we have discussed above in Section 12, other architectures and interfaces for large corpora are fairly limited in their ability to view, limit and contrast the frequency across sections of the corpus (such as by genre or by time period), or to quickly and easily find the frequency in all sections. With the COCA architecture and interface, however, users can find the overall frequency of phrasal verbs by section (genre or historical period) by simply entering *[vv*] [rp*]* (any verb followed by an adverbial particle) and then select the [CHART] display. They would then see the total frequency of all phrasal verbs in each of the five main genres and in the four time periods:



SECTION	SPOK	FIC	MAG	NEWS	ACAD	1990–94	1995–99	2000–04	2005–08
PER MIL	6,512.7	10,037.4	5,227.3	4,720.6	1,964.3	5,577.2	5,708.9	5,699.6	5,740.3
SIZE (MW)	78.8	74.9	80.7	76.3	76.2	103.4	103.0	102.6	77.9
FREQ	513328	751579	421638	360326	149664	576651	587832	584626	447426

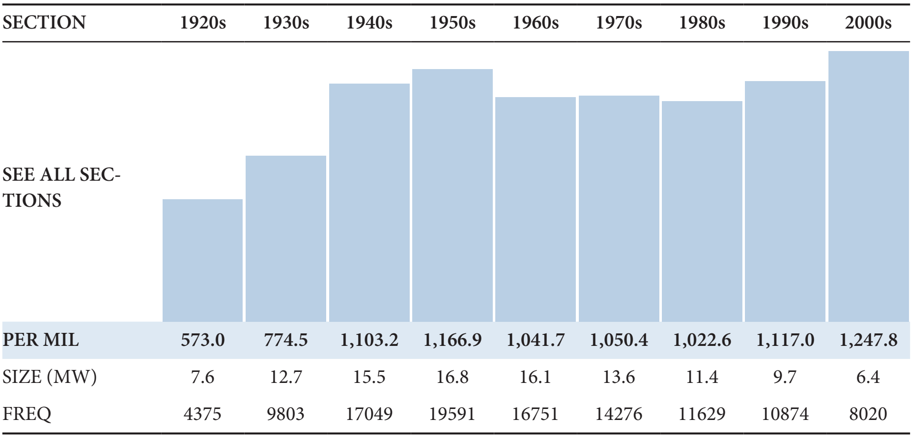
Figure 5. Frequency of phrasal verbs by genre and year

As can be seen from the chart above, COCA can provide us with interesting insight into recent historical shifts in English, which is something that is impossible (or very difficult) with any other existing corpus of English. The BNC lacks much of a historical timespan (it is mainly from the mid-1980s to 1993), and has of course not been updated since 1993. The Bank of English (http://www.titania.bham.ac.uk/) is — besides COCA — the only corpus of contemporary English with much of a diachronic nature to it, but it is not freely-available. In addition, its architecture is quite limited (much more so than BNCweb and Sketch Engine) in terms of searching by time period or even visualizing change across time.

With COCA, it is quite easy to map out recent changes in English. The chart above, for example, shows the frequency of phrasal verbs in each of the short 4–5 year periods since the early 1990s (the four columns to the right). The data from COCA indicate that the total frequency of phrasal verbs has continued to increase during the past two decades — by about 2–3 percent in each of the four time periods.

Another way of measuring this is to find the number of different phrasal verbs that have a given frequency in each of the different time periods, or in other words the 'productivity' of this construction. To find this, users can limit the search to a given time period, specify a minimum frequency (such as 100), and they will again see that the frequency is increasing slightly over time. There are 705 different phrasal verbs with a frequency of at least 100 in 1990–1994, 714 in 1995–1999, and 729 in 2000–2004. (It is not possible to know the total number of types with a frequency for 2005–2009, since there are not yet any texts from Oct-Dec 2008 or from 2009. But the totals for 2000–2002 (819 types) compared to 2005–2007 (829 types) show that the increase in number of types per time block is still increasing.)

While there are several corpus-based studies of phrasal verbs from Early Modern English (cf. Hiltunen 1994; Claridge 1997; Claridge 2000), there is virtually nothing for the last 100–200 years. Yet a search of the TIME Corpus (100 million words, American English, 1920s–2000s; see http://corpus.byu.edu/time/) shows that the total frequency of phrasal verbs is now more than twice what it was in the 1920s. After a period of relative stability from the 1940s–1980s, it has recently begun to increase again as can be seen in Figure 6.



SECTION	1920s	1930s	1940s	1950s	1960s	1970s	1980s	1990s	2000s
SEE ALL SECTIONS									
PER MIL	**573.0**	**774.5**	**1,103.2**	**1,166.9**	**1,041.7**	**1,050.4**	**1,022.6**	**1,117.0**	**1,247.8**
SIZE (MW)	7.6	12.7	15.5	16.8	16.1	13.6	11.4	9.7	6.4
FREQ	4375	9803	17049	19591	16751	14276	11629	10874	8020

Figure 6. Frequency of phrasal verbs in American English, 1920s–2000s (TIME Corpus)

In summary, then, the data from COCA (1990–present) support nicely and expand on the data from other historical corpora, such as the TIME Corpus.

In terms of the historical dimension, it is also interesting not only to see the overall increase in phrasal verbs, but to examine exactly which verbs are coming into (or perhaps dropping out of) the language. With the COCA interface, users simply select one time period (e.g. 2005–2008) for [Section 1], another time period (e.g. 1990–1994) for [Section 2], enter the search string (*[vv*] [rp*]* in the case of phrasal verbs), and the corpus architecture will compare the frequencies of all matching forms in the two time periods. Using this approach, we find that the following phrasal verbs occur much more in 2005–2008 than in 1990–1994: *hit up, drill down, queue up, feed off, rein in, dial down, price out, stress out*, and *log in*. On the other hand, the following phrasal verbs have experienced a significant decrease from 1990–94 to 2005–2008: *cool out, linger on, blank out, foul up*, and *ease off*. As can be seen, COCA can be a powerful tool to find neologisms and track their usage over time — perhaps more easily than with any other corpus of contemporary (American) English.

13.5 Genre-based variation

In addition to examining the historical dimension, COCA also allows users to quickly and easily compare the frequency across genres — again, perhaps more easily than with any other corpus architecture. For example, users can generate charts like Figure 5 above, which show the frequency of any given word, phrase, or construction in the five major genres. As we see in Figure 5, these data for phrasal verbs agree quite nicely with the data in Biber et al. (1999: 407–413). As in the

Longman corpus that was used by Biber et al., we find that phrasal verbs are the most frequent in fiction, followed by spoken, newspapers, and then academic.

With our architecture and interface, however, we can move far beyond just the 'macro-genres' to see the frequency of phrasal verbs in all 42 of the sub-genres of the corpus. By clicking on [See All Sections], users can see, for example, that movie scripts and juvenile fiction are the sub-genres of fiction that have the most phrasal verbs. For popular magazines, it is children's and women's magazines where phrasal verbs are the most common, with arts and religion the least. For newspapers, they are most common in sports and lifestyle, with general national news the least. And in the academic genre, phrasal verbs are the most common in the humanities and least common in medical journals.

Of course it is also possible to see the frequency of each individual type in each of the five major genres, as shown in Table 15:

Table 15. Phrasal verbs in the five major genres (grouped by lemma)

	WORD(S)	TOTAL	SPOK	FIC	MAG	NEWS	ACAD
1	[GO] [ON]	51233	23212 294.50	13315 177.82	6012 74.53	6110 80.05	2584 33.92
2	[COME] [BACK]	39146	19119 242.57	11676 155.93	3422 42.42	4667 61.14	262 3.44
3	[COME] [UP]	36554	19080 242.07	7072 94.45	4843 60.04	4785 62.69	774 10.16
4	[GO] [BACK]	35165	14401 182.71	11281 150.66	4122 51.10	4625 60.59	736 9.66
5	[PICK] [UP]	31379	5844 74.14	14167 189.20	5715 70.85	5251 68.79	402 5.28
6	[FIND] [OUT]	26002	11630 147.55	6056 80.88	4405 54.61	2939 38.50	972 12.76
7	[COME] [OUT]	25723	12531 158.98	6334 84.59	3172 39.33	3431 44.95	255 3.35
8	[GO] [OUT]	25432	9941 126.12	7960 106.31	3476 43.09	3981 52.15	74 0.97
9	[GROW] [UP]	23250	6279 79.66	3923 52.39	5387 66.79	6715 87.97	946 12.42
10	[POINT] [OUT]	22123	5221 66.24	1914 25.56	5339 66.19	3672 48.11	5977 78.45

This table shows the raw frequency (above) and normalized frequency per million words (below) for each phrasal verb in each genre. We can see, for example, that *come back, pick up, come out*, and especially *go out* are not very frequent in academic, whereas *point out* is more frequent in academic than in any of the other genres.

The corpus architecture and interface also make it quite easy to find which phrasal verbs are frequent in one genre (or set of genres), compared to another. As discussed in Section 12 and as shown in Figure 2, the COCA architecture is unique in the way that it allows users to choose or create genres "on the fly" via the corpus interface, and then compare any linguistic features in these two sets of genres. For example, to find the most frequent phrasal verbs in fiction — compared to the other four genres — one simply selects [Fiction] for [Section 1] and the other four genres for [Section 2]. This would yield phrasal verbs for fiction like *glance around, squint up, slump back,* or *peel out* (with many of these being movement verbs). For magazines, they would be *trim off, scrape down, press out*, and *leaf out* ("*the first tree to leaf out in spring*") (with many of these relating to hobbies, in articles that would not be found in other genres). In newspapers many of the most frequent phrasal verbs deal with sports (*line out, ground out, foul out, fly out*), since there is more sports reporting in this genre than in the other four. And in academic they are verbs like *elaborate on, center around, refer back*, and *trace out*, which refer to the research process.

13.6 Conclusion

In Section 13, we have briefly used phrasal verbs as a test case to show how COCA can be used to quickly and easily retrieve data on lexical, historical, and genre-based variation, in a way that is probably not possible with any other corpus (or corpus architecture). Because the entire corpus architecture is based on relational databases, all of the frequency information for words, phrases, and constructions in different sections of the corpus is already stored in the database, or else can be quickly retrieved from the corpus. Even the most complex queries discussed in this section take only a few seconds to search for and display results from the 385+ million words of text.

We believe that this ability to quickly search, limit by, and compare frequencies across different sections of the corpus is not found at this level in any other corpus architecture. And of course, because of its textual composition, COCA is the only corpus that can even begin to provide robust data like this for American English, as well as the only publicly-available corpus of English to provide data from the last decade or two. Both of these features make it uniquely valuable as a tool to examine current genre-based variation and recent diachronic shifts in the language.

References

American National Corpus. http://www.americannationalcorpus (accessed February 2009).

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

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finnegan, E. 1999. *Longman Grammar of Spoken and Written English*. London: Longman.

British National Corpus. www.natcorp.ox.ac.uk (accessed February 2009).

Collins Cobuild Dictionary of Phrasal Verbs. 1989. London: Collins Publishers.

Corpus of Contemporary American English (COCA). http://www.americancorpus.org. Created 2008 (accessed February 2009).

Claridge, C. 1997. "A century in the life of multi-word verbs". In M. Ljung (Ed.), *Corpus-based Studies in English*. Amsterdam: Rodopi, 69–85.

Claridge, C. 2000. *Multi-Word Verbs in Early Modern English: A Corpus-Based Study*. Amsterdam: Rodopi.

Davies, M. 2005. "The advantage of using relational databases for large corpora: Speed, advanced queries, and unlimited annotation". *International Journal of Corpus Linguistics*, 10 (3), 307–334.

Davies, M. 2009. "Word frequency in context: Alternative architectures for examining related words, register variation and historical change". In D. Archer (Ed.), *What's in a Word-list? Investigating Word Frequency and Keyword Extraction*. London: Ashgate, 53–68.

Gardner, D. & Davies, M. 2007. "Pointing out frequent phrasal verbs: A corpus-based analysis". *TESOL Quarterly*, 41 (2), 339–359.

Gavioli, L. 2005. *Exploring Corpora for ESP Learning*. Amsterdam/Philadelphia: John Benjamins.

Hiltunen, R. 1994. "On phrasal verbs in Early Modern English: Notes on lexis and style". In D. Kastovsky (Ed.), *Studies in Early Modern English*. Berlin: Mouton de Gruyter, 129–140.

Longman Dictionary of Phrasal Verbs. 1983. Courtney, R. (Ed.). Harlow: Longman.

McCarthy, M. & O'Dell, F. 2004. *English Phrasal Verbs in Use*. Cambridge: Cambridge University Press.

Mollin, S. 2007. "The Hansard hazard: Gauging the accuracy of British parliamentary transcripts". *Corpora*, 2 (2), 187–210.

NTC's Dictionary of Phrasal Verbs and Other Idiomatic Verbal Phrases. 1993. Spears, R. A. (Ed.). Lincolnwood, Illinois: National Textbook Co.

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

Stubbs, M. 1996. *Text and Corpus Analysis*. Oxford: Blackwell.

TIME Corpus of American English. http://corpus.byu.edu/time (accessed February 2009).

Appendix 1. Query syntax (basic)

Syntax	Examples	Meaning	Sample matches
word1	mysterious	One exact word	*mysterious*
word1/word2	watching/looking	Either word (no space)	*watching, looking*
word1 word2	nooks and crannies	Multiple exact words	*nooks and crannies*
* word	fairly *	Words + undefined	*fairly evenly, fairly tame*
* word *	* terms *	“slots” “ “ “ “	*in terms of, to terms with*
*xx *xxx* x?xx	un*ly *heart* r?n*	Wildcard: * = any # letters Wildcard: ? = one letter	*unlikely, unusually* *hearts, sweetheart, heart-throb* *run, running, ran*
[pos] [pos*]	[vvg] [v?g] thick [nn*]	Part of speech (exact) Part of speech (wildcard)	*walking, talking* *having, being, talking* *thick glasses, thick accent*
word*.[pos] word [pos*] [pos*] * word	dis*.[vvd] haunting [nn*] [vvi] * sound	Wildcard with exact POS Wildcard POS + lemma Exact POS + any word + word	*discovered, disappeared, discussed* *haunting images, haunting images* *hear a sound, want to sound, muffle the sound*
[word]	[sing] [tall]	Lemmas	*sing, singing, sang* *tall, taller, tallest*
[=word]	[=clean] [=clean].[v*] [[=clean]].[v*] [[=clean]].[v*] the [*nn*]	Synonyms Limited by POS POS; all forms of lemma POS; lemma; with nouns	*clean, pure, fresh, mop,* *clean, mop, scrub, polish* *clean, mopping, scrubbed,* *mops the floor, scrubbed the pot*
[user:list]	[davies:clothes] [[davies:clothes]] [[davies:clothes]].[n*]	Customized lists All forms of lemma Lemma; POS = noun	*tie, shirt, blouse* *tie, tying, socks, socked, shirt,* *tie, ties, sock, socks* (i.e. just nouns)
Any other combination	[put] on * [lingprof:clothes]	Example: Form of *put* + *on* + any word + word in [clothes] list created by [lingprof]	*put on her skirt, putting on blue jeans*

Appendix 2. Types of context-based searches

NODE	COLLO-CATES	SPAN (L/R)	EXPLANATION	SORT BY GROUP BY	EXAMPLES
laugh.[n*]	*	5/5	Any words within five words of the noun *laugh*	Percentage Collocates	hearty, scornful
[thick]	[nn*]	0/4	A form of *thick* followed by a noun	Frequency Collocates	glasses, smoke
[look] into	[nn*]	0/6	Nouns after a form of *look + into*	Frequency Collocates	eyes, future
[eye]	clos*	5/5	Words starting with *clos** within five words of a form of *eye*	Frequency Both words	closed // eye, closing // eyes
[feel] like	[*vvg*]	0/4	A form of *feel* followed by a gerund	Frequency Collocates	crying, taking
find	time	0/4	*Find* followed by *time*	Frequency Collocates	time
work/job	hard/tough/difficult	4/0	*Work* or *job* preceded by *hard* or *tough* or *difficult*	Frequency Both words	hard // work, tough // job
[=publish]	[n*]	0/4	Nouns after a synonym of *publish*	Frequency Both words	publish // book, issue // statement, print // money
[=expensive]	[[jones:clothes]]	0/5	Synonym of *expensive* followed by a form of a word in the *clothes* list created by *jones*	Frequency Both words	expensive / shoes, pricey // shirt
[=boy]	[=happy]	5/5	Synonym of *happy* near a synonym of *boy*	Frequency Both words	happy // child, delighted // boy

Author's address

Mark Davies
Brigham Young University
4071 JFSB
Department of Linguistics and English Language
Brigham Young University
Provo, UT 84602 USA

mark_davies@byu.edu