1

Corpora: an introduction

Mark Davies

1 Introduction

Many introductions to English corpora attempt to provide a comprehensive list of the "most important" corpora currently available. While there are some advantages to such an approach, these lists are invariably outdated even before they are published, and hopelessly outdated after five to years.

In this introduction, I take a different approach. Rather than attempting to create a complete and exhaustive list, I focus on a handful of corpora (and related resources, such as text archives and the "Web as Corpus") that are *representative* of general classes of corpora. We will discuss the relative advantages and disadvantages of these *general classes* of resources, which will undoubtedly contain much better exemplars in years to come.

The types of corpora (and corpus-related resources) that we consider are the following:[1]

1. Small 1–5-million-word, first-generation corpora like the ***Brown Corpus*** (and others in the Brown "family," such as the LOB, Frown, and FLOB)
2. Moderately sized, second-generation, genre-balanced corpora, such as the 100-million-word ***British National Corpus***
3. Larger, more up-to-date (but still genre-balanced) corpora, such as the 450-million-word ***Corpus of Contemporary American English*** (COCA)[2]
4. Large text archives, such as **Lexis-Nexis**
5. Extremely large text archives, such as **Google Books**[3]
6. **The Web** as corpus, seen here through the lens of Google-based searches

[1] All of the corpora discussed in this chapter are "general" corpora, rather than corpora for a particular genre of English.

[2] See Davies 2008, 2011.

[3] See Michel *et al*. 2011.

Finally, we will consider very large "hybrid" corpora, which take data from text archives or the Web, but which then deliver this data through powerful architectures and interfaces. These include:

7. The web-based corpora available through **Sketch Engine**
8. An advanced interface to Google Books, available through **googlebooks.byu.edu**.

As we discuss these different types of corpora, we will first (in Sections 2–5) see how well they can provide data for a wide range of linguistic phenomena – lexical, morphological, syntactic, and semantic. As we do so, we will consider how the quantity and quality of the data are affected by the corpus size, as well as the corpus architecture and interface. Second, in Section 6 we will consider the issue of *variation* within English, by looking primarily at genre coverage and balance in the corpora. We will also briefly consider other types of variation, such as variation in time (i.e. historical corpora) and space (i.e. corpora that provide data on dialectal variation), as well as variation at the level of individual speakers and writers. In the concluding section, we will take an (admittedly risky) "flight of fancy" and imagine what type of corpora might be available in five, ten, or twenty years.

2 Providing data on a wide range of linguistic phenomena

A typical textbook on linguistics will contain chapters on phonology, lexis, morphology, syntax, and semantics, as well as variation by speaker, by time (language change), and in space (dialects). As a result, it is probably not unreasonable to expect modern corpora to provide useful data on these types of phenomena, as shown in Table 1.1. (Note that these searches are simply representative examples of different types of searches; i.e. this is obviously not an exhaustive list.)

Too often, linguists are artificially and needlessly limited by the size or the design or the architecture of the particular corpus that they have been using for years. As a result, they are in a sense "blind" to the full range of phenomena that can be studied with other robust, well-designed corpora. For this reason, we will consider in some detail in the following sections (with many concrete examples) how the quantity and quality of the data that we obtain from corpora (for the phenomena listed above) are a function of corpus size, architecture, and genre balance.

3 Corpus size (Brown, BNC, COCA)

In this section, we will consider the importance of corpus size, and we will attempt to answer two questions. First, how do the data from

Table 1.1 *Types of phenomena*

Lexical

1. Frequency and distribution of specific words and phrases
2. Lists of all common words in a language or genre

Morphology

3. Processes involving word formation (e.g., nouns formed with suffixes like **ism* or **ousness*)
4. Contrasts in the use of grammatical alternatives, such as *HAVE* + *proven/proved*, or *sincerest/most sincere*

Grammar/syntax

5. High-frequency grammatical features, like modals, passives, perfect or progressive aspect
6. Less frequent grammatical variation, such as choices with verb subcategorization (*John started to walk / walking; she'd like* (*for*) *him to stay overnight*)

Phraseological patterns

7. Collocational preferences for specific words (e.g., *true feelings* or *naked eye*)
8. Constructions, *e.g. [V NP into V-ing] (they talked him into staying*) or [V POSS *way* PREP] (*he elbowed his way through the crowd*)

Semantics

9. Collocates (generally) as a guide to meaning and usage (e.g. with *click* (n), *nibble* (v), or *serenely*)
10. Semantic prosody, e.g. the types of words preceding the verb *budge* or nouns following the verb *cause.*

"first-generation" corpora like the Brown Corpus (1 million words in size) compare to those from second-generation corpora (which have anywhere from 100 to 500 million words), in terms of providing enough occurrences of different linguistic phenomena? Second, is there much of a difference between a 100-million-word corpus (e.g. BNC) and a nearly 500-million-word corpus (COCA), or is 100 million words adequate?

We will examine these two questions empirically, by looking in turn at each of the ten phenomena presented in Table 1.1.[4] (Note that these numbers are probably a bit cryptic at this point, but they will be explained – phenomenon by phenomenon – in the discussion that follows.)

3.1 Lexical

Even for some moderately frequent words, a one-million-word corpus like Brown does not provide enough data for useful analyses. For example, 83 of the 1,000 most frequent adjectives in COCA occur five times or less in Brown, including such common words as *fun*, *offensive*, *medium*, *tender*, *teenage*, *coastal*, *scary*, *organizational*, *terrific*, *sexy*, *cute*, *innovative*, *risky*, *shiny*, *viable*, *hazardous*, *conceptual*, and *affordable* (all of which occur 5,000 times or more in COCA). Of the top 2,000 adjectives in COCA, 425 occur five times or less in Brown, and this rises to 2,053 of the top 5,000 and 5,106 of the top

[4] Note that the COCA data are based on the 450-million-word version from 2012, but counts for later years will be higher, since COCA grows by 20 million words a year. The BNC and Brown, on the other hand, are static.

Table 1.2 *Frequency of different phenomena in COCA, BNC, and Brown (numbers explained in detail in Sections 3.1–3.5)*

		COCA (450 m)	BNC (100 m)	Brown (1 m)
1	Lexical: individual	(See discussion in Section 3.1 above)		
2	Lexical: word lists	100,705	43,758	3,956
3	Morphology: substrings	*-ousness* 112 *-ism* 512	*-ousness* 25 *-ism* 278	*-ousness* 1 *-ism* 6
4	Morphology: compare	*prove{n/d}* 2,616 + 3,001 *sincere* 85 + 65	*prove{n/d}* 82 + 1,169 *sincere* 11 + 12	*prove{n/d}* 3 + 7 *sincere* 1 + 0
5	Syntax: high frequency	modals 5,794k perfects 1,837k *be* passives 2,900k	modals 1,421k perfects 446k *be* passives 890k	modals 14k perfects 4k *be* passives 10k
6	Syntax: low frequency	*love* 12,178 + 5,393 *hate* 3,968 + 1,773 *for* 931	*love* 1,192 + 351 *hate* 389 + 475 *for* 103	*love* 10 + 2 *hate* 8 + 2 *for* 0
7	Phraseology: words	*true feelings* 654 *naked eye* 175	*true feelings* 148 *naked eye* 53	*true feelings* 2 *naked eye* 0
8	Phraseology: constructions	*way* 251v : 15,868t *into* 275v : 2,160t	*way* 83v : 3,533t *into* 111v : 358t	*way* 15v : 44t *into* 6v : 6t
9	Semantics: collocates	*riddle* (n) 57 *nibble* (v) 96 *crumbled* (j) 33 *serenely* (r) 24	*riddle* (n) 0 *nibble* (v) 13 *crumbled* (j) 1 *serenely* (r) 4	*riddle* (n) 0 *nibble* (v) 0 *crumbled* (j) 0 *serenely* (r) 0
10	Semantics: prosody	*budge* (v) 1,427 *cause* (v) 1,344	*budge* (v) 164 *cause* (v) 358	*budge* (v) 3 *cause* (v) 0

10,000 (all of which occur 120 times or more in COCA). In addition, a Brown-based frequency list (for all words in the corpus) would be quite sparse. For example, only 3,956 lemmas occur 20 times or more in Brown, but this rises to more than 43,000 lemmas in the BNC and 100,000 lemmas in COCA. (Note that this is not due to norming, but rather it is the number of word types.)

3.2 Morphology

Morphologists are interested in morpheme ordering in English (see Hay and Baayen 2005), and it is therefore useful to look for the frequency of words with multiple suffixes, such as **ous+ness*. In COCA, there are 112 different forms that end in **ousness* and that have more than 10 tokens (e.g. *consciousness*, *seriousness*, *nervousness*, *righteousness*, *graciousness*, *dangerousness*), and this decreases to 25 in the BNC and just one in Brown (*consciousness*). In COCA, there are 512 words ending in **ism* with more than 20 tokens, 278 in the BNC, and only 6 in Brown (*communism*, *criticism*, *nationalism*, *mechanism*, *realism*, *anti-Semitism*). Morphologists are also interested in variation in competing word forms, such as *have* + *proven* or *proved*, because of insights that these give into how we process language (e.g. single or dual-route model). In COCA, there are 2,616 tokens of *have proven* and 3,001 for *have* *proved* in COCA. The BNC has 82 *have proven* and 1,169 *have proved*, and Brown has only 3 *have proven* and 7 *have proved*. Comparing adjectival forms (*sincerest* vs. *most sincere*), there are 85 and 65 tokens (respectively) in COCA, 11 and 12 in the BNC, and only 1 and 0 in Brown.

3.3 Syntax

High-frequency syntactic constructions are perhaps the one type of phenomenon where Brown provides sufficient data.[5] For example, there are 14,080 tokens of modals, 4,288 perfect constructions, and 9,985 *be* passives. In the BNC this increases to approximately 1,421,000 modals, 446,000 perfect constructions, and 890,000 *be* passives. And in COCA it is of course even more: 5,794,000, 1,837,000, and 2,900,000 tokens, respectively. But even for something as frequent as the *get* passive (*John got fired last week*) there are only 58 tokens in Brown, whereas there are about 9,000 in the BNC and in 70,000 in COCA. There are very few tokens of less common syntactic constructions (such as verbal subcategorization) in Brown. For [*to* V / V-*ing*] (*John hated* [*to buy*/ *buying*]), COCA has 12,178 tokens of [*love* to VERB] and 5,393 tokens of [*love* V-ing], and 3,968 + 1,773 tokens with *hate*. The BNC has 1,162+351 with *love* and 389+475 with *hate*. Brown, on the other hand, has only 10+2 with *love* and 8+2 with *hate* – too few to say much about this construction. With the ±*for* construction (when it is "optional," e.g. *I want* (*for*) *you to leave*) there are 931 tokens in COCA, 103 in the BNC, and 0 in Brown.

3.4 Phraseology

Specific words and phrases: Sinclair (2004a: 30–36) discusses the patterning of two different phrases: *naked eye* and *true feelings*. COCA has a robust 654 and 175 tokens (respectively), while the BNC has 148 and 53 tokens. Such an investigation of phraseology would be quite impossible in Brown, however, where there are only 2 and 0 tokens, respectively. Constructions: in COCA, there are 251 distinct verbs and 15,868 tokens for the "*way* construction," e.g. *make his way to*, *find their way into*, *push his way through*, *bluster their way out of*. This decreases to 83 verbs and 3,533 tokens in the BNC and only 15 verbs and 44 tokens in Brown – probably too few for an insightful analysis. With the "*into* V-*ing*" construction (e.g. *talk him into going*, *bribe her into getting*, *lure me into buying*), there are 275 distinct matrix verbs and 2,160 tokens in COCA, which decreases to 111 verbs and 358 tokens in the BNC, and only 6 verbs and 6 tokens in Brown – again, too few for any insightful analyses.

[5] For this reason, it is perhaps no surprise that the Brown family of corpora has been used for a number of insightful analyses of high-frequency grammatical phenomena in English, e.g. Leech *et al*. (2009).

3.5 Semantics

Collocates can provide useful insight into meaning and usage, following Firth's insight that "you shall know a word by the company it keeps" (1957: 11). But collocates are very sensitive to corpus size. For example, there are 15 distinct ADJ collocate lemmas of *riddle* (NOUN) that occur three times or more in COCA (span = 1L/0R), e.g. *great*, *ancient*, *cosmic*; 96 distinct NOUN collocate lemmas of *nibble* (VERB) occurring three times or more (span = 0L/4R), e.g. *edges*, *grass*, *ear*; 33 distinct NOUN collocate lemmas of *crumbled* (ADJ) occurring three times or more (span = 0L/2R), e.g. *cheese*, *bacon*, *bread*; and 24 distinct VERB collocate lemmas of *serenely* occurring three times or more (span = 3L/3R), e.g. *smile*, *float*, *gaze*. Because collocates are so sensitive to size, we find that these numbers decrease dramatically from 15, 96, 33, and 24 in COCA to 0, 13, 1, and 4 (respectively) in the BNC, and a dismal 0, 0, 0, 0 in Brown. An interesting use of collocates is their role in signaling "semantic prosody" (see Louw 1993), in which a word occurs primarily in a negative or positive context. For example, *budge* is nearly always preceded by negation (*it wouldn't budge*, *they couldn't budge it*), and *cause* takes primarily negative objects (e.g. *death*, *disease*, *pain*, *cancer*, *problems*). In order to see such patterns, however, we need large corpora. In COCA, there are 1,427 tokens of *budge* and 1,344 different object noun collocates of *cause* that occur at least 10 times each (span = 0L/4R). This decreases to 164 tokens of *budge* and 358 noun collocates of *cause* in the BNC, and just 3 tokens of *budge* and 0 noun collocates of *cause* (occurring ten times or more) – again, simply not enough for insightful analyses.

3.6 Accuracy of annotation in small and large corpora

As we have seen, large corpora have certain advantages in terms of providing data on a wide range of phenomena. But it is also true that there are some challenges associated with large corpora. This is particularly true in terms of the accuracy of annotation – both at the word level (e.g. accurate part of speech tagging) and the document level (e.g. accurate metadata for all of the texts in the corpus). And this is especially true when the corpus is created by a small team and with limited resources.

Consider first the issue of accuracy in document-level metadata. The Brown Corpus is composed of just 500 texts, and it is very easy to achieve 100 percent accuracy in terms of metadata. COCA, on the other hand, currently has more than 180,000 texts and the 400-million-word COHA historical corpus has more than 100,000 texts.[6] Even if COHA is 99.9 percent accurate in terms of metadata, there are potentially 50 or 100 texts that might have the wrong date, title, author, or genre classification. If one is researching the very first occurrence of a word or phrase or a

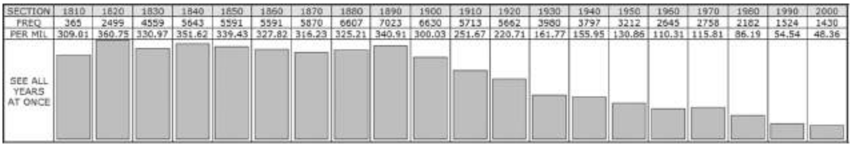
[6] COHA = *Corpus of Historical American English* – a historical, "companion" corpus for COCA; see Davies 2012a, 2012b, forthcoming.

construction with just 5–10 tokens, then even one text with the wrong metadata can cause serious problems. But in the case of a construction with 700 or 1,000 tokens, then 99.9 percent accuracy in metadata (with perhaps one errant token) should be sufficient.

Consider also word-level annotation, such as part-of-speech tagging. Suppose that we want to study the use of *for* as a conjunction (*, for had we known that . . .*). As Hundt and Leech (2012) point out, there are few enough tokens in the Brown Corpus (just 121 in all) that researchers have been able to manually examine each one of these to check the PoS tagging, and we see a clear decrease in *for* as a conjunction over time. COCA, on the other hand, has about 16,500 tokens of *for* that have been tagged as a conjunction and COHA has another 80,000 tokens for the period from the 1810s to the 1980s, which is far too many to examine manually.

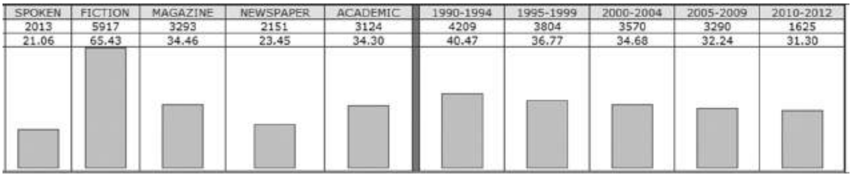
And yet because of the sheer number of tokens, I would argue, in this particular case we can still have confidence in the data. Consider Figures 1.1 and 1.2 which show nicely the decrease in *for* as a conjunction from the 1810s to the 2000s (COHA)[7] and then in the 1990s to the 2000s (COCA).[8]

In nearly every decade in COHA since the 1890s, *for* as a conjunction is less frequent than the preceding decade. And in COCA, it has decreased in every five-year period since 1990, and the decrease is still ongoing (as of 2012). So in this particular case, where Hundt and Leech (2012) suggested that there might be a problem with large corpora, it looks as though the large corpus works quite well. Further discussion and examples of other phenomena are addressed in Davies (2012b).



SECTION	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
FREQ	365	2499	4559	5643	5591	5591	5870	6607	7023	6630	5713	5662	3980	3797	3212	2645	2758	2182	1524	1430
PER MIL	309.01	360.75	330.97	351.62	339.43	327.82	316.23	325.21	340.91	300.03	251.67	220.71	161.77	155.95	130.86	110.31	115.81	86.19	54.54	48.36

Figure 1.1 Decrease in *for* (as conjunction in COHA), 1810s–2000s



SPOKEN	FICTION	MAGAZINE	NEWSPAPER	ACADEMIC	1990-1994	1995-1999	2000-2004	2005-2009	2010-2012
2013	5917	3293	2151	3124	4209	3804	3570	3290	1625
21.06	65.43	34.46	23.45	34.30	40.47	36.77	34.68	32.24	31.30

Figure 1.2 Decrease in *for* (as conjunction in COCA), 1990s–2012

[7] See http://corpus.byu.edu/coha/?c=coha&q=24684721

[8] See http://corpus.byu.edu/coca/?c=coca&q=24684733

There are of course additional questions, such as whether recall is as good as precision – since, by definition, we would have to manually look through large masses of (initially) incorrectly tagged tokens to improve precision. In addition, although precision may be 99 percent or higher in the case of *for* as a conjunction, there are undoubtedly other cases where POS tagging is more problematic. What is probably needed is a systematic study of accuracy of POS tagging for a wide range of phenomena in a number of (otherwise similar) small and large corpora, to determine just how much of an issue this might be.

4 Text archives and the Web[9]

Based on our analysis in the preceding section, it might appear that "bigger is always better." Of course, this is not the case. We have already discussed some of the challenges inherent in the creation of large corpora, in terms of accurate metadata and word-level annotation. In this section, we will consider several types of resources that dwarf traditional corpora in terms of their size, but which have limited use in researching many of the phenomena presented above in Table 1.1.

As an introduction to this section, we should remember that the usefulness of a corpus for end users is a function of at least

text (e.g. sentences and paragraphs in the corpus) +
annotation (e.g. document and word-level) +
architecture and interface

Just because a corpus is larger (number of words of text), it may have limited use if it is not annotated for parts of speech, or if the architecture is weak, or if the interface does not allow a wide range of linguistically relevant queries.

In this section, we will consider three examples of text archives – Lexis-Nexis (representing a wide range of similar text archives, such as ProQuest or EBSCO archives, other newspaper archives, or archives like Literature Online or Project Gutenberg), the Web (via Google), and Google Books. Table 1.3 shows how well these different resources do as far as allowing for the different types of research. Note that the checkmark in parentheses means that the search is probably only possible with (often significant) post-processing; i.e. it is not possible via the standard Web interface.

[9] In this section, we will discuss Web *as* Corpus, rather than Web *for* Corpus – an important discussion that is considered in some detail in Fletcher (2011). In Section 5, we will discuss Web *for* corpus.

Table 1.3 *Phenomena that can be researched with three text archives / Web*

		Lexis-Nexis	Web (via Google)	Google Books
1	Lexical: individual	(✓)	(✓)	(✓)
2	Lexical: word lists			
3	Morphology: substrings			
4	Morphology: compare forms	(✓)	(✓)	(✓)
5	Syntax: high frequency	(✓)	(✓)	(✓)
6	Syntax: low frequency	(✓)	(✓)	(✓)
7	Phraseology: words	(✓)		
8	Phraseology: constructions	(✓)	(✓)	(✓)
9	Semantics: collocates			
10	Semantics: prosody			

4.1 Lexical

There is no way to create frequency listing from text archives, at least via the standard interfaces for these resources. Nevertheless, with all three types of resources, it is certainly possible to see the frequency of an exact word or phrase, and of course the number of tokens will typically be much larger than with a 100- or 500-million-word corpus. For example, the adjectives in Table 1.4 – which shows their frequency in these text archives[10] – occur 20 times in COCA.

For some lexically oriented searches, there is really no alternative to an extremely large text archives, because of their sheer size. As can be seen, even COCA provides only relatively meager data for the words shown in Table 1.4. And this is even more pronounced for still-infrequent neologisms, where there may not be any tokens at all in a well-structured, half-billion-word corpus.

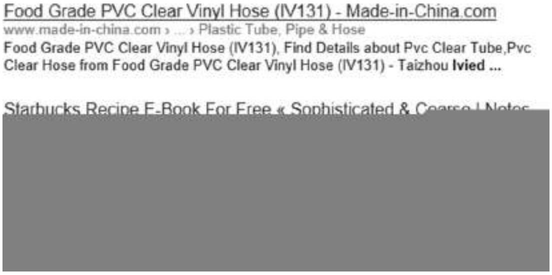
Nevertheless, there are a number of problems with the figures from text archives, shown in Table 1.4. First, in some cases the interface blocks access to more than a certain number of hits and it will not show the total number, as in the case of words with a frequency of 990–999 in Lexis-Nexis. Second, in text archives the numbers typically refer to the number of texts containing the word, rather than the total number of tokens. Third, as Kilgarriff (2007) notes, we need to be very skeptical of the numbers from (at least) Google. In the case of phrases, particularly, the numbers can be off by several orders of magnitude.[11] Fourth, in the case of Google Books, at present it is difficult to see (or extract) the number of tokens for a given word or phrase (see Figure 1.4). The results are displayed primarily as a "picture" of the frequency over time, and the actual raw number of tokens is deeply embedded in the HTML code for the web page.

[10] The average for these ten adjectives will of course be different from that of another set of adjectives, but it nonetheless lets us get a general sense of these resources. The size (lower row) in each case is calculated by finding the ratio with COCA, where the word occurs 20 times in 450 million words. We know that Google Books (English) is actually about 250 billion words, so the 232-billion-word estimate there is fairly accurate.

[11] For example, a search of the phrase *would be taken for a* shows 1,610,000 hits in Google, but after paging through the results, one finds that there are actually only 528 hits (accessed February 2013).

Table 1.4 *Frequency of very infrequent words in BNC, COCA, and three text archives / Web*

	BNC	COCA	Lexis-Nexis	Google Books	Web (Google)
accentual	19	20	168	26,155	244,000
biggish	40	20	999	4,577	504,000
coloristic	0	20	992	6,853	141,000
consummatory	1	20	71	25,710	109,000
disassociative	0	20	580	1,108	178,000
folkloristic	0	20	542	11,209	195,000
freckly	6	20	999	1,178	505,000
ivied	1	20	776	13,166	187,000
Kennedyesque	2	20	987	512	62,100
unbruised	3	20	995	3,240	86,500
Average	7.2	20	711	10,355	221,160
Size (??)	100 m	450 m	15 billion	232 billion	5 trillion



Food Grade PVC Clear Vinyl Hose (IV131) - Made-in-China.com
www.made-in-china.com › ... › Plastic Tube, Pipe & Hose
Food Grade PVC Clear Vinyl Hose (IV131), Find Details about Pvc Clear Tube,Pvc Clear Hose from Food Grade PVC Clear Vinyl Hose (IV131) - Taizhou **Ivied** ...

[redacted]

Figure 1.3 "Snippet" view in Google (Web)

Another challenge with text archives is working with the often limited interface. For example, consider the output from Google in Figure 1.3 (Lexis-Nexis and Google Books have similar displays). Unlike the interfaces for many structured corpora, where it is possible to display nicely sorted KWIC (Keyword in Context) lines, in the case of text archives one would have to write a script to extract the data in a usable format.

4.2 Morphology

Via the standard web interfaces, it is typically not possible to search by substrings. It is however, possible to compare competing word forms, such as *HAVE* + *proven* / *proved*. In Lexis-Nexis, this is straightforward. In Google (Web), one must remember that the frequency of strings can be wildly inaccurate (see note 11), so the comparison of these two numbers can also be very inaccurate. Finally, in Google Books it is possible to compare alternate forms, as in Figure 1.4. But again the actual raw numbers are very hard to extract (they are displayed primarily as cryptic "percentage" figures, as shown in Figure 1.4).

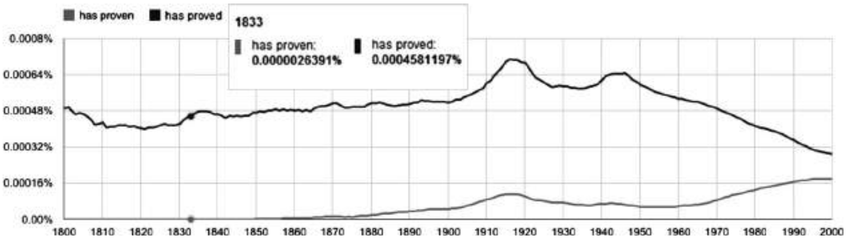


Figure 1.4 Frequency chart in Google Books

4.3 Syntax

Syntactically oriented queries present a real challenge for the often simple interfaces of text archives. Take for example the high-frequency *be* passive or the less frequent case of verbal subcategorization [*to* V / V-*ing*]. Text archives typically do not allow searching by part of speech (or even by lemma), so we would need to search for hundreds or thousands of matching strings one by one, e.g. *start to see*, *started to see*, *started to notice*, etc. One option would be to write a script to serially carry out many searches and then store the results for each search. But such queries are certainly not possible for most users, natively via the interface.

4.4 Phraseology

In cases like *true feelings* or *naked eye*, again one could write a script to parse through hundreds of "snippet" entries (as in Figure 1.3 above), and then store the snippets (and the accompanying metadata) in a KWIC format that can be analyzed with another piece of software. But even in Lexis-Nexis, which has the best KWIC-oriented display, it is cumbersome at best to try to use the output from the web interface, without additional processing. In cases like the "*way* construction" (e.g. *pushed his way through the crowd*) or the "*into* V-*ing*" construction (e.g. *he talked her into staying*), one is faced with the same problem as with the syntactically oriented searches in Section 4.3 above. Users would have to write a program to serially input thousands or tens of thousands of exact strings (e.g. *talk her into going*, *talk him into staying*, *coax Fred into doing*, etc.), and then store the results.

4.5 Semantics

In order to extract the collocates using the native interface for these text archives, one would (again) have to write a program to parse through the simple "snippet" output and save each snippet, and then post-process these data to extract collocates. However, even this would probably not be possible, since most text archives severely limit the number of "snippets" for a given search (e.g. 1,000 in Lexis-Nexis, Google (Web), and Google Books). With only 1,000 tokens per word or phrase, it is impossible to create a robust dataset to extract collocates.

4.6 Summary

Text archives are initially quite appealing, because of their sheer size. For certain types of lexically oriented queries (e.g. very low-frequency words or neologisms), they may be the only option, and they may also be sufficient for comparisons of alternating word forms (e.g. *have + proven/proved*, or *he snuck/sneaked*). But for virtually all other types of searches, the simplistic interface simply cannot generate the desired data, without significant post-processing.

5 Hybrid "corpora": text archives + full-featured architecture and interface

We saw in Section 3 that size is crucial: small 2–4-million-word corpora are at times limited in terms of the range of data that they can provide. But as we have seen in Section 4, size is not everything – most text archives have such a simplistic interface that they also are very limited in the range of queries that they offer. As we will see in this section, the best solution may be to take the texts from a text archive or the Web (containing billions of words of data), and then combine this with a robust corpus architecture.

As examples of this "hybrid" approach, in this section we will consider two sets of corpora. First, we will look at the corpora from Sketch Engine (www.sketchengine.co.uk). All of the corpora in Sketch Engine that are publicly accessible and that are more than a billion words in size are based on web pages, and there are currently three corpora of English that contain more than a billion words of text. Second, we will consider the different "corpora" that are available from googlebooks.byu.edu, which are based on the n-grams from books.google.com/ngrams/, and which range in size from 45 to 155 billion words.

As Table 1.5 shows, both of these hybrid corpora offer a wide range of searches.

Table 1.5 *Phenomena that can be researched with two "hybrid" corpora*

		Sketch Engine	googlebooks.byu.edu
1	Lexical: individual	✓	✓
2	Lexical: word lists	✓	✓
3	Morphology: substrings	✓	✓
4	Morphology: compare forms	✓	✓
5	Syntax: high frequency	✓	(✓)
6	Syntax: low frequency	✓	(✓)
7	Phraseology: words	✓	
8	Phraseology: constructions	✓	(✓)
9	Semantics: collocates	✓	(✓)
10	Semantics: prosody	✓	(✓)

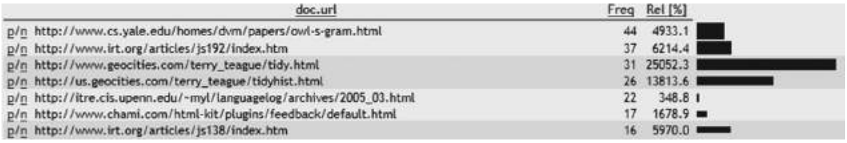
5.1 Lexical

Both corpora allow users to search for a particular word or phrase, and to see the frequency of the word in the different sections of the corpus. For example, Figures 1.5 and 1.6 show the frequency of *tidy* in different web pages in the 3.2-billion-word enTenTen08 corpus from Sketch Engine (Figure 1.5) and in each decade from the 1810s to the 2000s in the 155-billion-word American English dataset from Google Books (BYU) (Figure 1.6).

5.2 Morphology

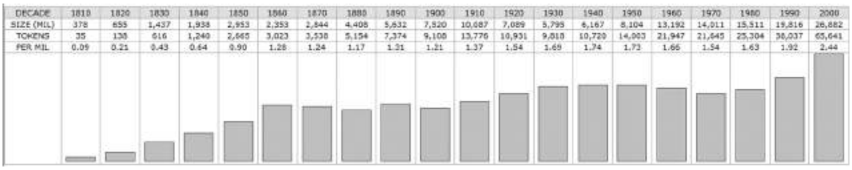
With both corpora, it is possible to generate word lists, including words that contain particular substrings. For example, the common words ending in **ism* in the Sketch Engine enTenTen08 corpus (about 3 billion words, from web pages) are *terrorism* (126,534 tokens), *mechanism* (95,034), *criticism* (92,190), *capitalism* (47,624), *journalism* (44,863), *racism* (43,451), *tourism* (37,552), *baptism* (33,774), *socialism* (28,717), and *organism* (23,629). In the Google Books (BYU) corpora, we can also see the distribution by decade (Figure 1.7).

Since we can easily search for strings in these corpora (see Section 5.1 above), we can also easily compare word forms, e.g. *he sneaked/snuck*. Figure 1.8 shows the frequency by decade in Google Books (BYU) (note the increasing use of *snuck* over time).



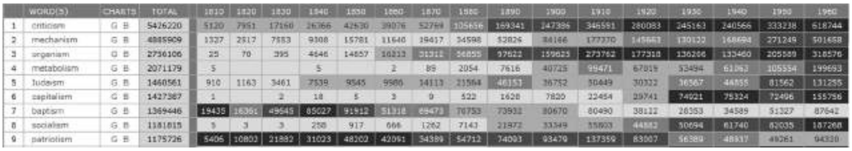
	doc.url	Freq	Rel [%]
p/n	http://www.cs.yale.edu/homes/dvm/papers/owl-s-gram.html	44	4933.1
p/n	http://www.irt.org/articles/js192/index.htm	37	6214.4
p/n	http://www.geocities.com/terry_teague/tidy.html	31	25052.3
p/n	http://us.geocities.com/terry_teague/tidyhist.html	26	13813.6
p/n	http://itre.cis.upenn.edu/~myl/languagelog/archives/2005_03.html	22	348.8
p/n	http://www.chami.com/html-kit/plugins/feedback/default.html	17	1678.9
p/n	http://www.irt.org/articles/js138/index.htm	16	5970.0

Figure 1.5 Lexical frequency: Sketch Engine



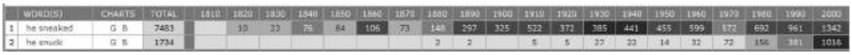
DECADE	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
SIZE (MIL)	378	655	1,437	1,938	2,953	2,353	2,844	4,408	5,632	7,520	10,087	7,089	5,795	6,167	8,104	13,192	14,011	15,511	19,816	26,882
TOKENS	35	138	616	1,240	2,665	3,023	3,538	5,154	7,374	9,108	13,776	10,931	9,818	10,720	14,003	21,947	21,645	25,304	38,037	65,641
PER MIL	0.09	0.21	0.43	0.64	0.90	1.28	1.24	1.17	1.31	1.21	1.37	1.54	1.69	1.74	1.73	1.66	1.54	1.63	1.92	2.44

Figure 1.6 Lexical frequency: Google Books (BYU)



	WORD(S)	CHARTS	TOTAL	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960
1	criticism	G B	5426220	5120	7951	17160	26766	42630	39076	52769	105656	169341	247386	346591	280083	245163	240566	333238	618744
2	mechanism	G B	4885909	1327	2517	7583	9308	15781	11640	19417	34598	52826	84166	177370	145661	130122	168694	271249	501658
3	organism	G B	2736106	25	70	395	4646	14857	16213	31311	56855	97622	159825	273762	177318	136206	133460	205589	318576
4	metabolism	G B	2071179	5			5		2	89	2054	7616	40725	99471	67015	53494	61062	105554	199693
5	Judaism	G B	1460561	910	1163	3461	7539	9545	9586	14113	21564	46153	36752	30449	30322	36567	44855	81562	131255
6	capitalism	G B	1427387	1		2	18	5	3	9	522	1628	7820	22454	29741	74921	75324	72496	155756
7	baptism	G B	1369446	19435	16361	49545	85027	91912	51318	69473	70753	73932	80670	80490	38112	28353	34589	51327	87642
8	socialism	G B	1181815	5	3	3	258	917	666	1262	7143	21972	33349	55803	44882	50694	61740	82035	187268
9	patriotism	G B	1175726	5406	10802	21882	31023	48202	42091	34389	54712	74093	93479	137359	83007	56389	48937	49261	94320

Figure 1.7 Word forms (**ism*) in Google Books (BYU)



	WORD(S)	CHARTS	TOTAL	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
1	he sneaked	G B	7483		10	23	76	84	106	73	148	297	325	522	372	385	441	455	599	572	692	961	1342
2	he snuck	G B	1734								2	2		5	5	27	22	14	32	72	156	381	1016

Figure 1.8 Morphological variation in Google Books (BYU)

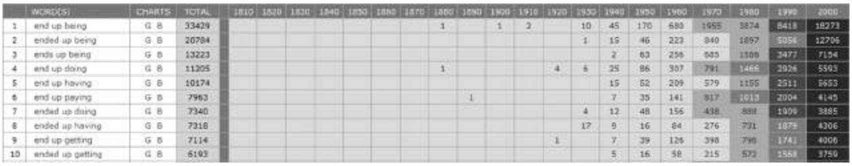
5.3 Syntax

The Sketch Engine corpora are tagged with TreeTagger and they are searched via Corpus Query Language (CQL), a widely used corpus architecture and search engine. This allows us to find constructions like the [*END up* V-*ing*] construction (e.g. *you'll end up paying too much*).[12]

Google Books (BYU) is a bit more problematic, in terms of syntactic searches. The version of the Google Books n-grams that it uses does not include part of speech or lemma. As a result, in a search like "[*end*] *up* [v?g*]," it creates the search "on the fly," based on COCA data. It finds all forms of the lemma *end* from COCA, followed by the word *up*, followed by any word in COCA that is tagged [v?g*] (e.g. *watching*, *knowing*) at least 50 percent of the time (50 percent is the default value, and it can be any number 1–100). Nevertheless, for most queries this works quite well. For example, Figure 1.9 shows the first entries for "[*end*] *up* [v?g*]," and these 400,000+ tokens are retrieved from the 155 billion words (*n*-grams) in about two seconds.

5.4 Phraseology

In just a couple of seconds, Sketch Engine can provide users with concordance lines for words, phrases, or even syntactic strings, which can be re-sorted, thinned, and so on. Google Books (BYU) cannot generate these concordance lines, because it is based just on n-grams. The actual text is in the Google Books "snippets" (e.g. Figure 1.3), and users would have the same problems extracting data for concordance lines from these snippets as they would in using Web data generally, as was discussed above in Section 4.1. Sketch Engine can search quite nicely for the patterns in which constructions occur – the same as it does for advanced syntactic searches generally, as seen in Section 5.3 above. For example, for the [VERB NP *into* V-*ing*] construction, Sketch Engine finds about 3,900 tokens (Figure 1.10),[13] and Google Books (BYU) finds about 30,200 tokens[14] (Figure 1.11).[15]



	WORD(S)	CHARTS	TOTAL	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
1	end up being	G B	33429								1		1	2		10	45	170	680	1955	3874	8418	18273
2	ended up being	G B	20784													1	15	46	223	840	1897	5056	12706
3	ends up being	G B	13223														2	63	256	685	1586	3477	7154
4	end up doing	G B	11205								1				4	6	25	86	307	791	1466	2926	5593
5	end up having	G B	10174														15	52	209	579	1155	2511	5653
6	end up paying	G B	7963									1					7	35	141	617	1013	2004	4145
7	ended up doing	G B	7340													4	12	48	156	438	888	1909	3885
8	ended up having	G B	7318													17	9	16	84	276	731	1879	4306
9	end up getting	G B	7114												1		7	39	126	398	796	1741	4006
10	ended up getting	G B	6193														5	16	58	215	573	1568	3758

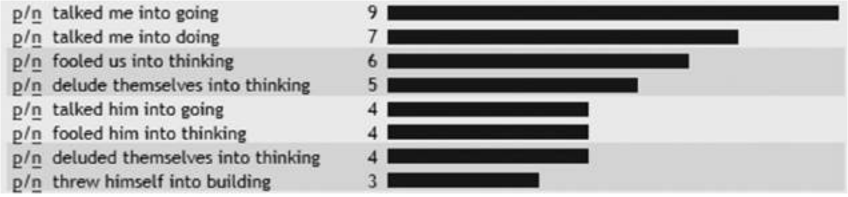
Figure 1.9 Probabilistic POS tagging in Google Books (BYU)

[12] [lemma = "end"] [word = "up"] [tag = "VVG"], which yields about 52,000 tokens.

[13] [tag = "VV."] [tag = "PP"] [word = "into"] [tag = "VVG"]

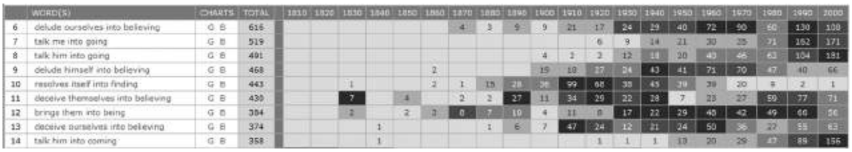
[14] Google Books (BYU) is based on the Google Books n-grams, which only include n-grams that occur 40 times or more in the corpus. This creates complications for a search like *[vv*] [p*] into [v?g*]*, where because of all of the different verbs, the vast majority of unique strings (e.g. *they bamboozled us into recapping*) will not occur 40 times or more, and will therefore not appear in the "corpus."

[15] [vv*] [p*] into [v?g*]



p/n	talked me into going	9
p/n	talked me into doing	7
p/n	fooled us into thinking	6
p/n	delude themselves into thinking	5
p/n	talked him into going	4
p/n	fooled him into thinking	4
p/n	deluded themselves into thinking	4
p/n	threw himself into building	3

Figure 1.10 Syntactic searches in Sketch Engine



	WORD(S)	CHARTS	TOTAL	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
6	delude ourselves into believing	G B	616							4	3	9	9	21	17	24	29	40	72	90	60	130	100
7	talk me into going	G B	519												6	9	14	21	30	25	71	162	171
8	talk him into going	G B	491										4	1	2	12	18	20	40	46	63	104	181
9	delude himself into believing	G B	468						2				15	18	27	24	43	41	71	70	47	40	66
10	resolves itself into finding	G B	443			1			2	1	15	28	36	99	68	38	45	39	39	20	8	2	1
11	deceive themselves into believing	G B	430			7		4		2	2	27	11	34	29	22	28	7	23	27	59	77	71
12	brings them into being	G B	384			2		2	3	0	7	10	4	11	8	17	22	29	48	42	49	66	56
13	deceive ourselves into believing	G B	374				1				1	6	7	47	24	12	21	24	50	36	27	55	63
14	talk him into coming	G B	358				1								1	1	1	13	20	29	47	89	156

Figure 1.11 Syntactic searches in Google Books (BYU)

5.5 Semantics

As we saw above in Section 3.5, collocates are extremely sensitive to corpus size. Table 1.6 shows the number of collocates (with a minimum lemma frequency of at least three tokens) for four different searches.

It is interesting, however, that for some words at least, there appears to be a "diminishing return" with corpus size. Although the number of collocates between the BNC and COCA (4–5 times as large as the BNC) is striking, in a corpus like enTenTen12 from Sketch Engine (which is 25 times as large as COCA), for some words (e.g. *nibble* or *serenely*) there is not nearly as significant a yield in collocates. We will discuss some possible explanations for this in the following section, as we discuss the composition of the corpora.

Table 1.6 *Number of collocates in different corpora*

		coll:span	Brown	BNC	COCA	SketchEng[b]	GB: BYU
Genre[a]	node word	Size	1 m	100 m	450 m	11.2 b	155 b
FIC, ACAD	*riddle* (N)	J: 1L/0R	0	0	15	228	188
FIC, MAG	*nibble* (V)	N: 0L/4R	0	13	99	~90[c]	~58[d]
MAG	*crumbled* (J)	N: 0L/1R	0	1	28	115	92
FIC	*serenely* (R)	V: 3L/3R	0	4	28	36	54

[a] This is the genre of COCA in which the word is the most frequent, which ends of being important as we talk about genre balance in Section 6.

[b] The Sketch Engine *enTenTen12* corpus – currently 11,192,000,000 words in size

[c] Sketch Engine groups collocates by grammatical relation, so it separates for example direct object (*nibble the cheese*) from object of preposition (*nibble on the cheese*). We have done our best to group the collocates from different relations and calculate their total frequency, but 90 is an approximate number.

[d] Collocates in Google Books (BYU) work differently than the other BYU corpora, like COCA or COHA, since Google Books is based on *n*-grams. As a result, these numbers are an approximate. Remember also the issue with the 40 token threshold, explained in note 14.

6 Accounting for and describing variation: genre, historical, dialect, and demographic

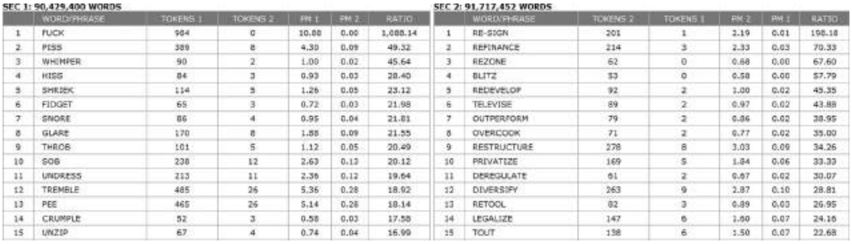
6.1 The importance of genre variation

Hundreds of studies over the past decade or two have shown the crucial importance of genre in describing language. Perhaps the best example of this is Biber *et al.* (1998), which shows how very different the language is in different genres – in terms of syntax and morphology (with somewhat less attention given in the book to lexis and semantics).

We will here provide just a few pieces of data from COCA – a robust, well-balanced corpus – to show the importance of genre. First, consider Figure 1.12, which shows verbs that are much more common in fiction (left) than newspapers (right). Imagine that we had a corpus composed only of newspapers (which are very easy to obtain). In this case, words like those on the left would be almost completely absent in the corpus, while those on the right would be massively over-represented.

Figures 1.13–1.17 show extreme variation between genres in COCA for other phenomena as well. Figure 1.13 shows how much more common *-al* adjectives are in academic (adjectives that are at least ten letters in length). Figures 1.14–1.17 show a number of grammatical phenomena where there are significant variations between genres: preposition stranding with *to* (e.g. *the man I was talking to*), the *get* passive (e.g. *John got fired from his job*), the quotative *like* (e.g. *and I'm like, what's the problem?*), and *real* instead of *really* before adjectives (e.g. *he was real sick*).

The important differences between genres extend to meaning as well. For example, Figure 1.18 shows collocates that are much more common



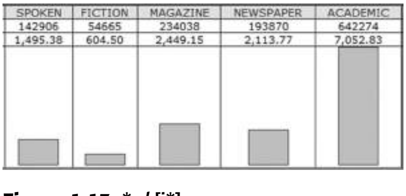
SEC 1: 90,429,400 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	FUCK	904	0	10.00	0.00	1,088.14
2	PISS	389	8	4.30	0.09	49.32
3	WHIMPER	90	2	1.00	0.02	45.64
4	HISS	84	3	0.93	0.03	28.40
5	SHRIEK	114	5	1.26	0.05	23.12
6	FIDGET	65	3	0.72	0.03	21.98
7	SNORE	86	4	0.95	0.04	21.81
8	GLARE	170	8	1.88	0.09	21.55
9	THROB	101	5	1.12	0.05	20.49
10	SOB	238	12	2.63	0.13	20.12
11	UNDRESS	213	11	2.36	0.12	19.64
12	TREMBLE	485	26	5.36	0.28	18.92
13	PEE	465	26	5.14	0.28	18.14
14	CRUMPLE	52	3	0.58	0.03	17.58
15	UNZIP	67	4	0.74	0.04	16.99

SEC 2: 91,717,452 WORDS

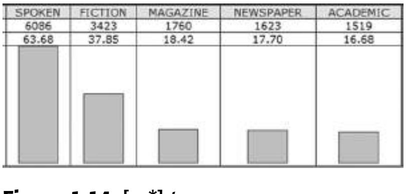
	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	RE-SIGN	201	1	2.19	0.01	198.18
2	REFINANCE	214	3	2.33	0.03	70.33
3	REZONE	62	0	0.68	0.00	67.60
4	BLITZ	53	0	0.58	0.00	57.79
5	REDEVELOP	92	2	1.00	0.02	45.35
6	TELEVISE	89	2	0.97	0.02	43.88
7	OUTPERFORM	79	2	0.86	0.02	38.95
8	OVERCOOK	71	2	0.77	0.02	35.00
9	RESTRUCTURE	278	8	3.03	0.09	34.26
10	PRIVATIZE	169	5	1.84	0.06	33.33
11	DEREGULATE	61	2	0.67	0.02	30.07
12	DIVERSIFY	263	9	2.87	0.10	28.81
13	RETOOL	82	3	0.89	0.03	26.95
14	LEGALIZE	147	6	1.60	0.07	24.16
15	TOUT	138	6	1.50	0.07	22.68

Figure 1.12 Lexis of fiction and newspapers in COCA



SPOKEN	FICTION	MAGAZINE	NEWSPAPER	ACADEMIC
142906	54665	234038	193870	642274
1,495.38	604.50	2,449.15	2,113.77	7,052.83

Figure 1.13 **al*.[j*]



SPOKEN	FICTION	MAGAZINE	NEWSPAPER	ACADEMIC
6086	3423	1760	1623	1519
63.68	37.85	18.42	17.70	16.68

Figure 1.14 [vv*] *to*

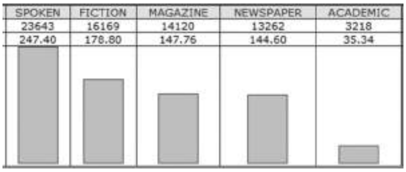


Figure 1.15 *get* passive

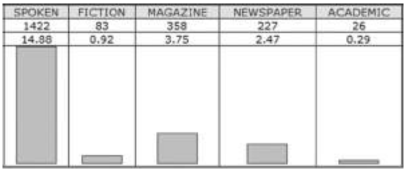


Figure 1.16 quotative *like*: [c*] [p*] [*be*] *like*,|

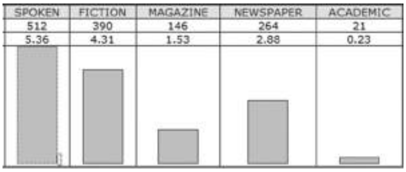
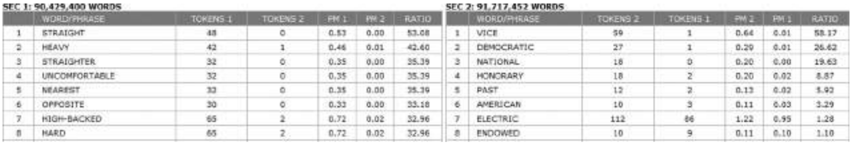


Figure 1.17 [*be*] *real* [j*] [y*]


SEC 1: 90,429,400 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	STRAIGHT	48	0	0.53	0.00	53.08
2	HEAVY	42	1	0.46	0.01	42.60
3	STRAIGHTER	32	0	0.35	0.00	35.39
4	UNCOMFORTABLE	32	0	0.35	0.00	35.39
5	NEAREST	32	0	0.35	0.00	35.39
6	OPPOSITE	30	0	0.33	0.00	33.18
7	HIGH-BACKED	65	2	0.72	0.02	32.96
8	HARD	65	2	0.72	0.02	32.96

SEC 2: 91,717,452 WORDS

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	VICE	59	1	0.64	0.01	58.17
2	DEMOCRATIC	27	1	0.29	0.01	26.62
3	NATIONAL	18	0	0.20	0.00	19.63
4	HONORARY	18	2	0.20	0.02	8.87
5	PAST	12	2	0.13	0.02	5.92
6	AMERICAN	10	3	0.11	0.03	3.29
7	ELECTRIC	112	86	1.22	0.95	1.28
8	ENDOWED	10	9	0.11	0.10	1.10

Figure 1.18 Collocates of *chair* in fiction and newspapers in COCA

with *chair* in fiction than in newspapers (left) and those that are much more common in newspapers than in fiction (right). Again, if our corpus is composed of one easily obtainable genre (like newspapers), we see only one very narrow "slice" of the language.

Even more than with newspapers, it is very easy to create extremely large corpora that are based solely on web pages. It is no surprise that virtually all of the corpora over 100 million words in size in Sketch Engine, for example, are based exclusively on web pages. But the question is – how representative are web pages of the full range of variation in the language?

Consider Table 1.7, which compares the lexis of the "traditional" genres to the web-only corpora. In this case, we compare word frequency in COCA and the BNC to the 2-billion-word *Corpus of Global Web-based English* (GloWbE),[16] which is just based (like Sketch Engine) on web pages. Here we see how many words in a 100,000 word list of English[17] have roughly the same normalized frequency in the (entire) GloWbE as in different genres of COCA and the BNC. For example, there are 13,386 words (from among the 100,000 total in the list) whose normalized frequency in COCA newspapers is roughly the same as that of GloWbE – i.e. the ratio is between 0.8 and 1.2.

[16] http://corpus2.byu.edu/web/ [17] www.wordfrequency.info

Table 1.7 *Similarity of lexis in web-based GloWbE and genres in COCA and BNC*

COCA	No. words	BNC	No. words
NEWSPAPER	13,836	MAGAZINE	8,743
MAGAZINE	13,349	NEWSPAPER	8,677
ACADEMIC	11,828	ACADEMIC	7,032
SPOKEN	10,793	FICTION	6,335
FICTION	8,804	SPOKEN	4,667

As can be seen, at least in terms of lexis, the web-only corpus is most like newspapers and magazines, but "web" lexis does a much poorer job of representing the lexis of the academic genre, or especially fiction and spoken. This may be why at times even very large web-only corpora do not improve significantly on the data from a well-balanced corpus (like COCA or the BNC). As we saw in Table 1.6, even a corpus like the 11.2-billion-word Sketch Engine *enTenTen12* corpus (25 times as large as COCA) provides only minimally better data for words that are most common in genres like fiction (e.g. *nibble* or *serenely*).

If we compare certain morphological and syntactic phenomena in the web only corpora to more balanced corpora, the situation becomes even more confusing. For example, the normalized frequency of *-al* adjectives[18] is 2,244 per million words in GloWbE-US,[19] which places it between COCA magazines and newspapers (see Figure 1.13 above). But the normalized frequency of the *get* passive ([get] [vvn*]; *John got fired last week*) is (at 239 tokens per million words) the most similar to spoken (see Figure 1.15 above). And strangely enough, the normalized frequency of *real* + ADJ ([be] real [j*] [y*]; *he was real smart*) is 0.77 in GloWbE-US, which is most like COCA Academic (see Figure 1.17). As we see, depending on the particular phenomena that we are studying, the web corpora are "all over the map" in terms of which of the "traditional" genres they best represent. As a result, it would be difficult to know ahead of time – for any particular phenomena – how representative of other genres a web-only corpus would be.

In summary, virtually all corpora with more than 1 billion words are composed of just web pages. But these large web page-based corpora do not represent particularly well the full range of variation that we see in genre-balanced corpora like the 100-million-word BNC, the 440-million-word *Bank of English*, or the 450-million-word (and growing) *Corpus of Contemporary American English* – which is currently the largest publicly available, genre-balanced corpus of English.

[18] *-al* adjectives that are at least ten letters long, e.g. *environmental* or *educational*.

[19] The 400 million words from the United States in the 2-billion-word GloWbE corpus.

6.2 Other types of variation

Besides genre-based variation, other important types of variation are change over time, variation between dialects, and variation at the level of the individual speaker.

In terms of historical variation, I have suggested at some length in other studies that perhaps the only historical corpus of English that is currently available, which can account for a full range of lexical, morphological, phraseological, syntactic, and semantic variation over the past 200 years (e.g. items 1–10 of Table 1.1) is the 400-million-word *Corpus of Historical American English* (COHA; see Davies 2012a, 2012b, forthcoming). I have also suggested that for very recent changes in English, the only reliable monitor corpus – which maintains the same genre balance from year to year (a crucial factor, which virtually all previous studies seem to have overlooked) and which is large enough to study a wide range of phenomena – is the *Corpus of Contemporary American English* (COCA; see Davies 2011).

In terms of dialectal variation, the International Corpus of English (Greenbaum 1996; Hundt and Gut 2012) can describe the range of variation about as well as other 1-million-word corpora, as we discussed in Section 3. The 1.9-billion-word *Corpus of Global Web-based English* (GloWbE) can account for the full range of linguistic phenomena shown in Table 1.1, since it uses the same architecture and interface as COCA and COHA. This includes queries that show variation between dialects, and which allow us to compare one dialect (or set of dialects) to another. But we must remember that this corpus – as is the case with virtually all corpora of its size – is based solely on web pages from these twenty countries – with the accompanying limitations discussed in Section 6.1 above. One other option is to use a corpus interface like that of the BYU corpora, which allow side-by-side comparisons of a wide range of phenomena in corpora from different countries (e.g. BNC for British English, COCA for American English, and the Strathy Corpus for Canadian English).[20]

Finally, in terms of demographic variation – variation at the level of the speaker (e.g. gender or age) – the *British National Corpus* is currently the only corpus that was designed, constructed, and annotated in such as way that it is possible to compare at the level of the individual speaker – *and* which is large enough to enable research on the full range of linguistic variation (items 1–10 of Table 1.1). But the degree to which end users can use this information is dependent on the corpus interface for the BNC. BNCweb[21] is currently the best interface for the BNC, in terms of researching demographic variation.[22]

[20] See http://corpus.byu.edu/comparing-corpora.asp for a wide range of side-by-side comparisons of the BNC and COCA, including lexical and syntactic frequency, collocates (to examine semantic contrasts), and so on.

[21] http://bncweb.lancs.ac.uk/

[22] See www.natcorp.ox.ac.uk/docs/URG/BNCdes.html

7 Some concluding remarks (and a crystal ball)

In this introductory chapter on "corpora," rather than attempting to discuss all of the important English corpora that are currently available (an impossible task), we have focused instead on different *types* of corpora (with just a few examples of each), and we have paid particular attention to *general* issues of size (Section 4), architecture (Section 5), and variation (Section 6). We have seen that there are relatively few corpora (perhaps limited to just the BNC, the *Bank of English*, and COCA and COHA) that (1) are large, (2) allow a wide range of searches, and (3) provide data from a wide range of genres.

Even at this level of abstraction, some of what we have considered will still be outdated almost as soon as this chapter published, and much of this will be hopelessly outdated within five to ten years. This is due in part to dramatic changes that I believe are on the verge of taking place, particularly in terms of data from social media. For example, Twitter already provides real-time "fire hose" access to every single tweet[23] – hundreds of millions of words of data each day – and Facebook and other social media sites may soon do so as well.

Imagine the situation five, ten, or twenty years from now, when researchers will be able to download billions of words of data *every day* from Facebook or other social media sites. For each status update or post that comes through, they will have accompanying metadata that show the gender, general age range, and approximate geographical location of the author. Assume further that because of advances in technology, they are able to efficiently process hundreds of billions of words of data at a rate that is hundreds or thousands of times as fast as today. One can therefore imagine a scenario – in the not-too-distant future – in which a researcher can examine the use of a particular word, or phrase, or syntactic construction – virtually in real time, and with incredible detail (gender, age, and location).

For example, researchers could examine two competing syntactic constructions (e.g. +/– *to* with *help: help Mary clean the room*, *help Mary to clean the room*), and see which of the two is more common in the US or the UK, between men and women, between different age groups, as a function of the embedded verb, or in data from this year compared to data from last year. Even the largest "structured" corpora from the present time (e.g. Sketch Engine corpora, GloWbE, COCA, or the BNC) cannot provide this degree of granularity. And this one example from the domain of syntax can be multiplied endlessly for other variations in syntax, or in lexis, morphology, phraseology, or meaning. At this point, I suspect that many of us will look back with nostalgia on the "quaint" 100- or 500-million-word corpora

[23] See https://dev.twitter.com/docs/streaming-apis/streams/public

that we currently have available, and wonder how we were able do so much with so little.

While this is an admittedly hypothetical scenario, what is probably beyond dispute is that the corpora that will be available to us in a decade or two will be truly revolutionary, at least from our current vantage point. The only question, then, is whether we will take advantage of the new resources that are certain to come our way.