Examining Syntactic Variation in English: The Importance of Corpus Design and Corpus Size

Mark Davies (Brigham Young University)

Davies, Mark. 2013. Examining Syntactic Variation in English: The Importance of Corpus Design and Corpus Size. ***English Language and Linguistics*** **19.3, 1-39.** This paper examines a number of cases of syntactic variation in English – change over time, variation between genres, and variation between dialects. All of the data comes from large, structured corpora of English, including COCA, COHA, GloWbE, the BNC, and Google Books (Advanced). For many different types of syntactic constructions, only very large corpora, with the right type of architecture and interface, can provide the needed data to accurately describe these types of syntactic variation.

Key words: corpus, syntax, variation, dialect, historical, genre

1. Introduction

Too many grammars of English (or any language, for that matter) make overly-general statements about the grammaticality or acceptability of certain syntactic phenomena, without taking into account the fact that those judgments might apply to just one genre of one dialect at one particular point in time. As a result, their descriptions of English end up being quite artificial, and are therefore not nearly as insightful as they could otherwise be. The use of corpus data might help to remedy this situation, but all too often even corpus linguists base their conclusions on corpora that fails to adequately take into account a full range of variation in the language.

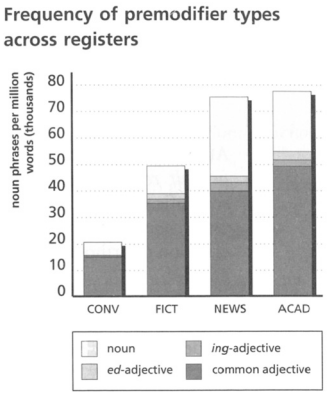
In this paper, we will consider how syntactic phenomena can vary as a function of language change, genre-based differences, and dialectal differences. Equally as important, we will consider how several recent corpora allow us to examine these three types of variation in ways that were quite impossible even four or five years. The overall message of this study, then, is that with the right type of corpora, we can account for variation in a much more reliable way, and thus provide much more insightful investigations into

English grammar.

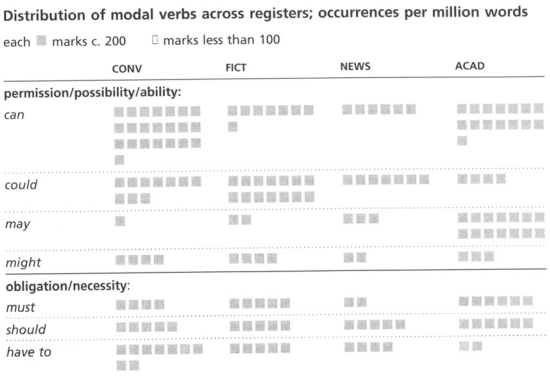
2. Researching genre-based variation with COCA

One of the most important types of syntactic variation is that which results from differences between genres. Biber et al (1999) was a groundbreaking 1100+ page book that provided hundreds of examples of significant differences in the frequency of syntactic constructions and features between spoken, fiction, newspaper, and academic texts. To take just two examples from among hundreds, they show the frequency of adjectival types by genre (Figure 1) and the frequency of modals by genre (Figure 2):

<Figure 1>



<Figure 2>



Due to its careful attention to detail and because of the wide range of phenomena that were covered, the Longman grammar will probably never be equaled as a guide to genre-based variation in English syntax. Nevertheless, while the corpus that they used was nearly "state of the art" in the 1990s, corpora have improved since that time, and therefore it might be interesting to revisit some of their data with larger and more recent corpora.

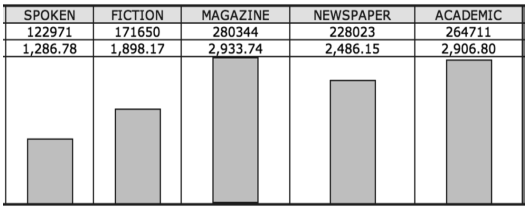
The Longman Corpus that Biber and his colleagues used was 40 million words in size (20 million words from the US and 20 million words from the UK), with 10 million words in each of the four genres of spoken, fiction,

newspaper, and academic. A corpus with 40 million words was very large in the 1990s, but it is somewhat on the "small" size in the 2010's. Perhaps the most serious limitation of their corpus, however, was that it was proprietary and not publicly-accessible. While researchers, teachers, and students could look at the hundreds of charts to examine genre-based variation, there was little possibility of ever replicating these investigations themselves.

In 2008 the Corpus of Contemporary American English (COCA) was released (see Davies 2009). As of the present time (2013) it is 450 million words in size – more than ten times as large as the Longman Corpus that was used by Biber and his colleagues. Best of all, it is publicly-accessible. As a result, researchers, teachers and students can easily replicate many of the investigations in Biber et al, and do so with a much larger and more recent corpus. In this section, we will provide examples of just a handful of such investigations.

Figure 1 above shows the frequency of different types of adjectives in the Longman corpus. With the simple search "*ing.[j*] [nn*]" we can find the frequency of –ING adjectives in COCA, and see the frequency by genre. (In this and the following charts, the first row of numbers shows the raw number of tokens, and the second row of numbers shows the normalized frequency – per million words – in the different genres.)

<Figure 3> Participial adjectival modification: ing.[j] [nn*]: overall frequency



In addition to seeing the overall frequency by genre, we can also see the individual matching strings in each genre, as in Figure 4.

<Figure 4> Participial adjectival modification: ing.[j] [nn*]: individual forms

	CONTEXT	ALL	SPOKEN	FICTION	MAGAZINE	NEWSPAPER	ACADEMIC
1	DEVELOPING COUNTRIES	3607	179		566	421	2441
2	GROWING NUMBER	2600	336		742	924	598
3	FOLLOWING YEAR	1783	73	104	605	444	557
4	WORKING CLASS	1568	348	58	187	186	789
5	BOILING WATER	1382	15	141	834	365	27
6	DEVELOPING WORLD	1356	104		345	204	703
7	MANAGING EDITOR	1302	266	8	226	782	20
8	MANAGING DIRECTOR	1279	115		288	805	71
9	INCREASING NUMBER	1254	76		332	309	537
10	FOLLOWING DAY	1179	87	369	312	165	246
11	NURSING HOMES	1147	155	4	177	323	488
12	INTERESTING THING	1120	988	33	62	33	4
13	TURNING POINT	1093	315	29	252	318	179

Figure 2 above shows the frequency of different modals in the Longman corpus by genre, and we can easily replicate this in COCA as well. For example, we see that *may* and *must* are the most frequent in academic texts, that *may* is quite uncommon in fiction, and that (not surprisingly) the contracted forms *'ll* (from *will*) and *'d* (from *would*) are least common in academic.

<Figure 5> Modal frequency by genre

	CONTEXT	ALL	SPOKEN	FICTION	MAGAZINE	NEWSPAPER	ACADEMIC
1	WOULD	1053578	266727	272262	173064	188137	153388
2	CAN	984914	238902	131811	248664	160535	205002
3	WILL	858879	213890	95522	186126	224317	139024
4	COULD	707864	137737	246489	121030	113880	88728
5	MAY	363146	63930	20835	90449	54627	133305
6	SHOULD	354185	88819	55889	67954	59606	81917
7	'LL	324740	121992	108116	55846	34665	4121
8	MIGHT	238757	45338	58794	48152	39336	47137
9	'D	190557	45615	88269	30422	23233	3018
10	MUST	189889	21813	44548	35231	28612	59685

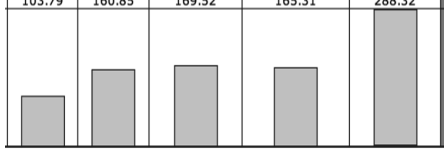
This genre-based variation is perhaps seen more clearly in the following two charts, which show the frequency of *must* and *have to* (a semi-modal) followed by a lexical verb (e.g. *must recognize, has to know*) in the five main genres of COCA (spoken, fiction, popular magazines, newspapers, and academic texts). Notice how *must* is more common in the more "formal" genres, whereas *have to* is more common in the informal genres, such as

spoken.

<Figure 6> Frequency by modals by genre: *must* as *have to*

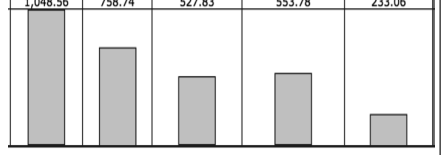
must [vv*]

SPOKEN	FICTION	MAGAZINE	NEWSPAPER	ACADEMIC
9920	14546	16199	15161	26256
103.79	160.85	169.52	165.31	288.32



have to

SPOKEN	FICTION	MAGAZINE	NEWSPAPER	ACADEMIC
100206	68612	50439	50791	21224
1,048.56	758.74	527.83	553.78	233.06

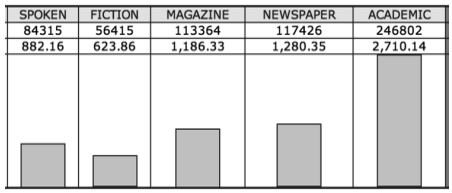


Another example of clear genre-based variation in COCA is related to the "*be*" and "*get*" passives (e.g. *John was / got fired from his job)*. Two simple searches in COCA show us that the *be* passive is much more common in the formal genres (especially academic), whereas the *get* passive is most frequent in the informal genres (such as spoken). (For background information on this construction, see Hundt 2001, Mair 2006, and Ruhlemann 2007).

<Figure 7> Frequency of *be* and *get* passives by genre

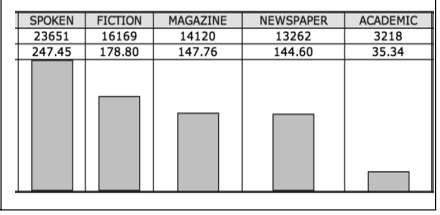
be passive

SPOKEN	FICTION	MAGAZINE	NEWSPAPER	ACADEMIC
84315	56415	113364	117426	246802
882.16	623.86	1,186.33	1,280.35	2,710.14



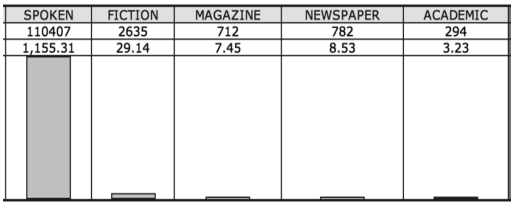
get passive

SPOKEN	FICTION	MAGAZINE	NEWSPAPER	ACADEMIC
23651	16169	14120	13262	3218
247.45	178.80	147.76	144.60	35.34



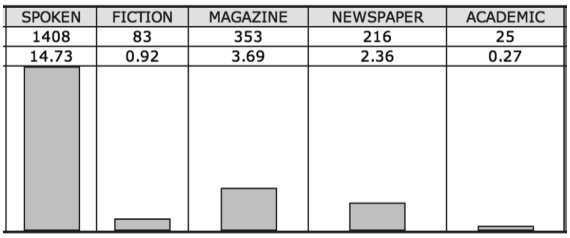
While the spoken transcripts in the Longman Corpus are from common, everyday conversation, the spoken transcripts in COCA come from *unscripted* conversation on national TV and radio broadcasts. As a result, some might think that this conversation in COCA would be too formal and stilted, but this is not the case. For example, the Figure 8 shows the frequency of the simple discourse markers like *you know*, which is of course much more common in the spoken transcripts:

<Figure 8> Frequency by the discourse marker [, *you know* ,] by genre



As we have mentioned, one of the real advantages of COCA (over the Longman Corpus used by Biber et al) is that it is much more recent. For certain types of constructions, this is an important advantage. For example, consider the data for the "quotative *like*" construction (*and* I'm like, "*I'm not going with you"*), shown in Figure 9 (for background information on this construction, see Tagliamonte and D'Arcy 2004, Buchstaller and D'Arcy 2009, and Barbieri 2009). As we will see in Section 3, this construction is quite recent, and is clearly increasing over time; hence a more recent corpus will provide many more tokens. But for our present purposes, we can see that there are significant differences in the frequency of the construction in the fives genres, with the construction being the most common (by far) in spoken, and virtually non-existent in academic. This also shows again that the spoken texts in COCA do reflect informal conversation quite well.

<Figure 9> Frequency of the "quotative *like*" construction by genre

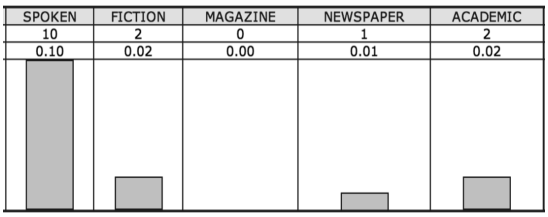


<Figure 10> Examples of the "quotative *like*" construction (KWIC)

CLICK FOR MORE CONTEXT							[?] SAVE LIST CHOOSE LIST ---------- CREATE NEW LIST [?]
1	2009	SPOK	NPR_TellMore	A	B	C	someone put chemicals in their hair? Then I just started laughing **and he 's like ,** do you do that mom? And I'm like absolutely I do.
2	2006	SPOK	NBC_Today	A	B	C	fun colors. CURRY: That's a good price. Ms-GOODMAN: **And it 's like ,** you know, tomboy chic. TEXT: Valentine's Day Gifts American Eagle
3	2009	SPOK	NPR_TalkNation	A	B	C	So that was different. (Soundbite-of-laugh Mr. MEDINA: **And I was like ,** you could see the - the video on YouTube right now.
4	2007	SPOK	ABC_GMA	A	B	C	lot of guys will wear like the Leatherman thing on the outside **and they 're like ,** in the office, you know, they're like a management consultant but
5	2008	SPOK	CBS_48Hours	A	B	C	he say, Do you want to be in a band? **And I was like ,** Yeah. I'll be in a band. I was like, Who
6	2011	SPOK	ABC_20/20	A	B	C	, is what I heard, what I think I heard. **And I was like ,** well, that's really strange. CHRIS-CUOMO-1-ABC# (Voiceover) She doesn't recognize the
7	2011	SPOK	ABC_Primetime	A	B	C	I got naked and they were, like, taking pictures. **And they were like ,** oh yeah. Then he got naked. COMMERCIAL-BREAK-# ANNOUNCER-# Are young models being
8	2010	SPOK	ABC_20/20	A	B	C	, it's 12:00 in - it's not even noon. **And she 's like ,** oh, that's okay. By the time we get all the stuff
9	2008	SPOK	ABC_20/20	A	B	C	is dead. " And my mom is, like screaming. **And I 'm like ,**' Who's the king? Who's the king? " DOCTOR-LARRY-CAHIL# People
10	1997	SPOK	NBC_Today	A	B	C	. You know, I'm sort of cynical and jaded, **and I 'm like ,** Oh, brother,' but that was really, really sweet. Dr-WEIL
11	2004	SPOK	PBS_Tavis	A	B	C	Michael, emote pain! I need pain, Michael! " **And I was like ,** " If I had like some words to go with this I could like
12	2011	SPOK	NPR_FreshAir	A	B	C	when I'm looking at him, I'm having mixed feelings **because I 'm like ,** well, if my dad is telling me this, and he's doing

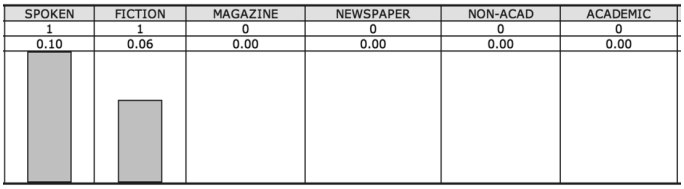
The other significant advantage of COCA over the Longman Corpus (in addition to being much more recent) is that it is much larger. For some low-frequency constructions, this is of crucial importance. For example, consider the following chart, which shows the frequency of the construction that combines passive, perfect, and progressive (e.g. *he had been being watched by the FBI*). We see clear effects of genre with the construction, in that it occurs much more in spoken than in the other genres. But note that there are only fifteen tokens in COCA, which contains 450 million words. In a much smaller 40 million word corpus like the Longman Corpus, there might only be one or two tokens.

<Figure 11> Frequency of [have been being V-ed] by genre in COCA



The importance of size in looking at genre-based differences is confirmed when we look at the frequency (by genre) of the "HAVE been being V-ed" construction in the British National Corpus, which contains only 100 million words. As Figure 12 shows, the construction occurs only two times, and it is therefore difficult to see how genre comes into play in this case.

<Figure 12> Frequency of [have been being V-ed] by genre in the BNC



In summary, then, we can use COCA to quickly and easily search for and document important genre-based variation in English syntax, to confirm the detailed genre-based data in Biber et al (1999). And for certain low-frequency constructions and for very recent syntactic constructions, COCA is perhaps the only corpus that will show such genre-based differences.

3. Researching recent and ongoing syntactic changes with COCA

In addition to genre-based variation, with the right kind of corpora we can also map out historical changes in syntax. In the following three sections, we will see how this can e done for very recent and ongoing changes in English with COCA (Section 3), over the past 200 years with the 400 million word Corpus of Historical American English [COHA] (in Section 4), and over the past 200 years with the 155 billion word Google Books (Advanced) n-grams databases (Section 5).

Turning first to recent, ongoing changes in English, I have argued elsewhere (see Davies 2011) that COCA is perhaps the only large corpus of English that allows us to look such changes. This is due to the fact that COCA is the only large corpus that 1) continues to be updated and 2) that has a genre composition that is essentially the same from year to year.

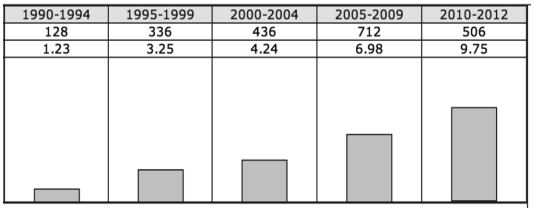
In this section, we will provide a handful of examples of how COCA data can provide insight into recent and ongoing syntactic shifts in English. Virtually none of these investigations would be possible with other corpora of contemporary English, either because 1) they are too small or 2) because – as with the Oxford English Corpus and the Bank of English – they do not have the same genre balance from year to year (again, see Davies 2011 for full

details).

We will first consider two very salient recent changes ("quotative *like*" and "*so not* ADJ"), followed by two changes in two prescriptively-focused constructions (*can*/*may* for permission, and split infinitives) and then finally three much less salient constructions: [*end up* V-ing], the "*get* passive", and [*help* (to) V].

First, let us consider the rise in two fairly salient grammatical constructions that have increased in frequency during the past two decades: the "quotative like" construction (*and he's <u>like</u>, "I'm not going with her"*) and the "*so not*" construction (*I'm <u>so not</u> interested in him*). Turning first to the "quotative *like*", recall that as Figure 9 indicated, this construction is much more common in the spoken genre than in the other genres. In addition to genre-based variation, there is also clear evidence for change over time, as is shown in the following chart from COCA.

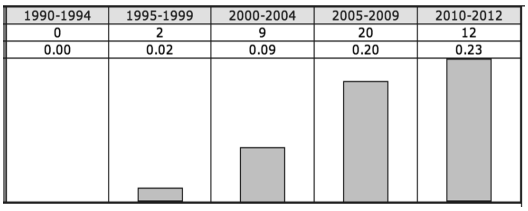
<Figure 13> Frequency of "quotative *like*" over time, 1990s-2000s



As the chart indicates, the frequency of this construction has steadily increased in each five-year period since the early 1990s. Via the corpus interface, it is also possible to see the normalized frequency in each individual year, and this shows that for nearly every year during the past decade, the frequency is higher than the year before.

Consider now the "so not" construction (*I'm so not interested in him*). As shown in the chart below, although the tokens for this construction are relatively sparse, but we still see a clear increase in the construction over time

<Figure 14> Frequency of [*so not* ADJ], 1990s-2000s



Let us now briefly consider two "prescriptive" issues –*can/may* for permission, and the split infinitive. First, consider the data for *can* vs. *may* (cf. Facchinetti 2000, Leech 2003, Millar 2009), as measured by the frequency of the two strings *can I* and *may I*. As the data show, there is a steady shift away from the prescriptive rule (i.e. from *may I* to *can I*) during the past two decades.

<Table 1> Frequency of *may / can (I)*, 1990s-2000s

	1990-94	1995-99	2000-04	2005-09	2010-12
may I	1223	855	768	722	328
can I	Top of Form 2976	3541	3027	3055	2024
% can I	70.9%	80.6%	79.8%	81.9%	86.5

Consider as another prescriptive rule the split infinitive (to [verb] [Adv] > to [Adv] [Verb], e.g. *to go boldy* > *to boldly go*) (cf. Close 1987). This is measured by the percentage of –ly adverbs (e.g. *boldly, quickly*) either before or after the infinitive following *to*. As can be seen, there is an increase in each five year block during the past two decades.

<Table 2> Frequency of split infinitive (e.g. *to go boldly* > *to boldly go*), 1990s-2000s

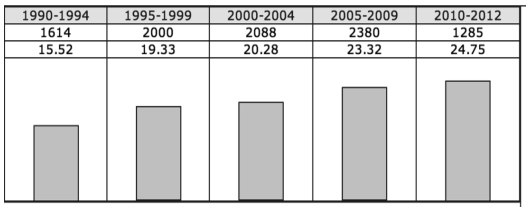
	search string	1990-94	1995-99	2000-04	2005-09	2010-12
- split	to [v*] *ly.[r*]	17675	15981	16124	14999	7164
+ split	to *ly.[r*] [v*]	8068	9349	10419	11368	6641
% split		31.3%	36.9%	39.3%	43.1%	48.1%

To this point, we have looked at two salient, recent grammatical constructions and two fairly salient prescriptive rules. For these phenomena,

however, sociolinguistic surveys or other means of gathering data might also be sufficient, since the speakers are quite aware of the phenomena. Where corpora really shine, however, is for the "lower level" constructions where speakers themselves seem quite unaware of what is going on. To conclude this section, consider three more syntactic shifts in contemporary American English (from among many that we could choose): the rise in the "end up V-ing" construction (*we'll end up paying too much*), the increase in the "*get* passive", and the shift from [help to V] to [help V].

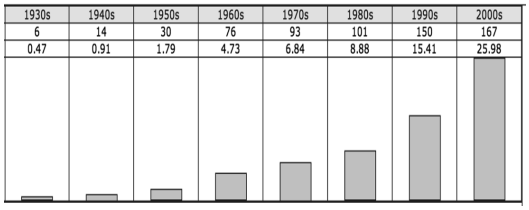
First, Figure 15 shows the increase in the "end up V-ing" construction over the past two decades.

<Figure 15> Frequency of [end up V-ing], 1990s-2000s



Notice that the normalized frequency increases in each five year period since the early 1990s. In fact, this continues a trend that has been in progress for the last 80-90 years, as shown in data from the 100 million word TIME Corpus of Historical American English (http://corpus.byu.edu/time):

<Figure 16> Frequency of [end up V-ing] by decade, 1930s-2000s



The second low-level shift is the rise in the "*get* passive" (*Bill got hired last week*, vs. *Bill was hired last week*), who genre distribution is discussed in Figure 7 above. The following table was not produced directly by the COCA

interface, but it is based on two searches in COCA (the *be* passive: [be] [vvn*] and the *get* passive: [get] [vvn*]). It shows the percentage of all passives (*be* or *get*) that occur with *get*.

<Table 3> Frequency of "*get* passive" vs. "*be* passive", 1990s-2000s

	1990-94	1995-99	2000-04	2005-09	2010-12
be	672188	625102	609466	570799	282262
get	14129	15888	15959	16867	9241
% get	2.1%	2.5%	2.6%	3.0%	3.3%

As one can see, the *get* passive steadily increases from one time period to the next, and the overall effect since the early 1990s is that the *get* passive has increased (compared to the *be* passive) more than 50% during this time.

The final low-level syntactic change is the slow but consistent shift from [help to V] to [help V] (*I'll help Mary to clean the room* > *I'll help Mary clean the room*), which is a change that has been commented on from a corpus -based approach by for previous studies on changes and variation with complements of *help*, see Kjellmer (1985); Mair (1995, 2002); and Rohdenburg (2009), among others.

<Table 4> Frequency of [help to V / help V], 1990s-2000s

	search string	1990-94	1995-99	2000-04	2005-09	2010-12
+ to	[help] [p*] to [v*]	825	798	726	668	370
- to	[help] [p*] [v*]	5494	6453	7144	7502	4237
% -to		86.9%	89.0%	90.8%	91.8%	92.0%

This data from COCA complements the data from the TIME Corpus, which also shows a slow but steady evolution towards the bare infinitive (*help him clean the room*) from the 1920s to the 2000s.

<Table 5> Frequency of [help to V / help V] by decade, 1920s-2000s

	1920s	1930s	1940s	1950s	1960s	1970s	1980s	1990s	2000s
+ to	15	33	47	54	53	54	24	11	8
- to	73	214	316	369	287	303	270	391	363
% - to	83%	87%	87%	87%	84%	85%	92%	97%	98%

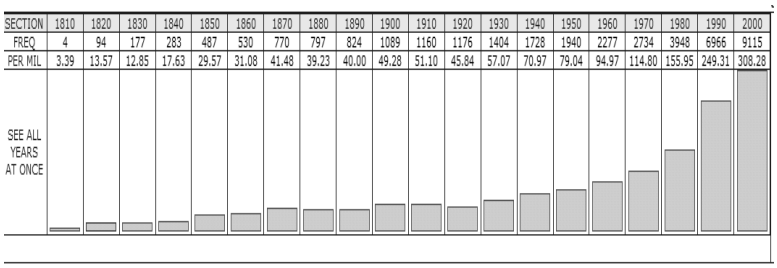
To summarize this section, we have seen that COCA – perhaps uniquely – can quickly and easily provide data on a wide range of ongoing syntactic shifts in contemporary English. Other large corpora such as the Bank of English and the Oxford English Corpus do provide data from different years in the 1990s and 2000s, but they crucially do not have the same genre balance from year to year, which cripples their use as monitor corpora (see Davies 2011 for a more complete discussion). On the other hand, small corpora like the Brown family corpora – which do have texts from the 1960s and 1990s and which have been used to compare high frequency syntactic constructions in the two periods – are just too small to look at a wide range of syntactic shifts. COCA alone seems to have the right balance to look at such changes.

4. Researching longer range syntactic changes with COHA

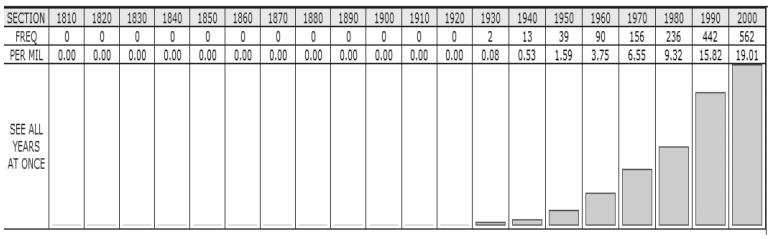
The Corpus of Historical American English (COHA) was released in 2010, and it contains more than 400 million words from a wide range of genres, and it maintains roughly the same genre balance from decade to decade. At 400 million words, it is about 100 times as large as any other genre-balanced historical corpus of English, and as a result it allows us to gain much more insight into syntactic changes in English than is possible with any other corpus. The majority of the phenomena shown in this section could not be studied successfully with small 2-4 million words corpora.

Carrying out research on diachronic syntax with COHA is both quick and easy. For example, the following two charts show the increase in the *need to V* (*we need to leave*) and the *end up V-ing* (*we'll end up getting there late*) constructions. Notice the nice S-curve increase in both constructions in the last 40-50 years. In terms of extracting the data, it is just a matter of inputting the correct search string (*[need] to [v*]* and *[end] up [v?g*]*) and COHA will find all of the tokens (1827 tokens for *end up V-ing* and 37,503 tokens for *need to V*) and create the chart in less than two seconds.

<Figure 17> Frequency of "*need to* [v*]", 1810s-2000s



<Figure 18> Frequency of "[end] up V-ing", 1810s-2000s



Even more complicated studies of diachronic syntax can be carried out quite easily with COHA. For example, Table 6 considers adverb placement with modals. [A] represents pre-verbal placement (never|always [vm*] [vv*] : *he* never would *answer his mail*) while [B] is post-verbal placement: (he [vm*] never|always [vv*] : *he* would never *answer his mail*). In this case we just carry out both searches (49,311 tokens total), copy the data from the two charts into Excel, and create a ratio of B/(A+B). In less than one minute total, we can clearly see the shift towards post-verbal placement: *he would* never *answer his mail*.

<Table 6> Frequency of post-verbal negation, 1810s-2000s

	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
A	490	523	437	389	435	423	405	281	280	241	157	147	122	135	82
B	2301	2547	2772	2608	2864	3128	3180	3051	2922	3143	2815	2755	3137	3665	3876
% B	0.82	0.83	0.86	0.87	0.87	0.88	0.89	0.92	0.91	0.93	0.95	0.95	0.96	0.96	0.98

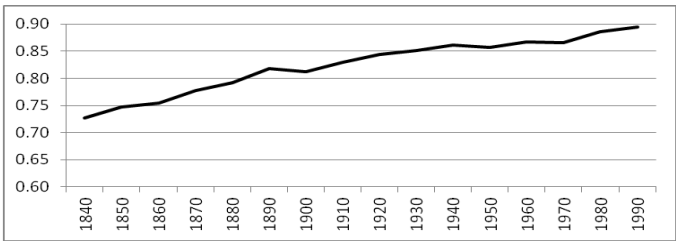
<Figure 19> Frequency of post-verbal negation, 1810s-2000s

Consider now a syntactic search that would likely be quite complex with other corpora, but which can be done quite easily with COHA. This deals with the increase in null relative pronouns at the expense of overt relative pronouns . [A] below represents overt relative pronouns with *he* as relative clause subject ([nn*] that|which|who|whom he [vv*]: *the woman that he married*) while [B] is zero relative pronoun: ([nn*] – he [vv*]: *the woman that he married*). As before, we simply copy the data from the two charts and do a simple ration in Excel. Of course we might want to change the relative clause subject, experiment with different type of antecedents, and so on. But the point is that with COHA, we can do even relatively complex searches such as this – resulting in clear and unambiguous data like that shown below – in just a minute or so.

<Table 7> Zero relative (*the man – he saw*), 1810s-2000s

	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990
A	1835	1668	1683	1758	2052	1911	2067	1995	2039	1740	1463	1516	1392	1291	1124	910
B	4871	4939	5155	6139	7841	8586	8972	9693	10983	9964	9098	9106	9089	8273	8697	7739
% B	0.73	0.75	0.75	0.78	0.79	0.82	0.81	0.83	0.84	0.85	0.86	0.86	0.87	0.87	0.89	0.89

<Figure 20> Zero relative (*the man – he saw*), 1810s-2000s



The previous investigations related to descriptive-oriented phenomena, but we can also use COHA to look at more prescriptively-oriented phenomena, as is shown with the following two prescriptive rules. The first is the shift from *may* to *can* for permission (as measured by the ratio of the two phrases *may I* and *can I*). Table 8 contains the data from 13,346 tokens from 1990 to 2009, and the following chart shows perhaps more clearly the shift from *may* to *can* during this time. Notice that although there are some increases and decreases in terms of the percentage of *can* (perhaps due to the varying effect of the prescriptive rule at times), the gray trendline shows the overall increase in *can,* and we see that it is now 50% more common than it was 100 years ago.

<Table 8> *Can I* vs *may I*, 1810s-2000s

	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
may I	488	485	498	460	451	550	456	390	473	327	348
can I	559	577	543	572	731	675	833	813	887	1135	1095
% can I	0.53	0.54	0.52	0.55	0.62	0.55	0.65	0.68	0.65	0.78	0.76

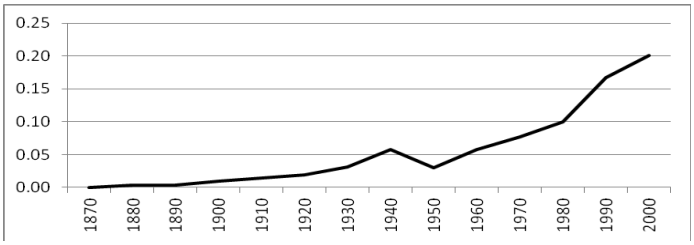
<Figure 21> *Can I* vs *may I*, 1810s-2000s

The second prescriptive rule shows the shift from *different from* to *different than* from the 1870s to the current time (*Bill is quite different from/than the others*), and is based on 9,636 tokens (there are virtually no cases of *different than* before the 1870s, and so the chart starts at that point). The increase in *different than* is perhaps more noticeable in the following chart, where we see that although there was still some tentativeness in the 1940s-1950s, the increase in *different than* has been quite pronounced since that time.

<Table 9> *different* + *from* / *than*, 1810s-2000s

different +	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
from	537	535	513	627	683	663	631	641	668	686	664	692	796	747
than	0	2	2	6	10	13	20	37	20	40	51	69	133	150
% than	0.00	0.00	0.00	0.01	0.01	0.02	0.03	0.06	0.03	0.06	0.08	0.10	0.17	0.20

<Figure 22> *different* + *than* (vs *from*), 1810s-2000s

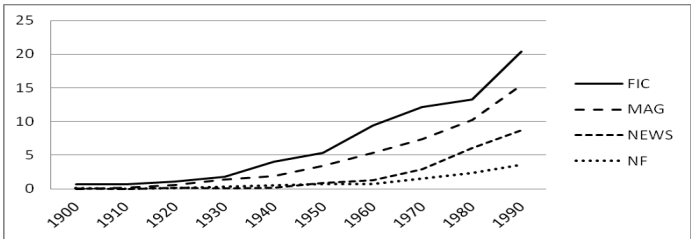


Finally, note that all of the data above is drawn from the complete corpus. As we know, however, language change often spreads through genres, perhaps starting in the more informal genres and then spreading to the more formal genres over time. We can easily map this out with COHA as well. For example, Table 10 shows the frequency per million words for the *end up* constructions (+ADJ: *he ended up dead*, and also +V-ing: *he ended up doing more than he wanted*). We run the query four times, selecting each of the different genres. We then copy the data into Excel (as in Table 10) and we can then see (as in the chart below) how in every decade since the early 1900s, the construction has been most common in the more informal genres.

<Table 10> [end] up ADJ, 1900s-2000s

GENRE	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990
fiction	0.63	0.66	1.09	1.75	4.02	5.34	9.38	12.18	13.27	20.36
magazine	0	0.13	0.55	1.38	1.93	3.38	5.34	7.35	10.19	15.46
newspaper	0.05	0	0.12	0.08	0.21	0.86	1.33	2.86	6.04	8.66
non-fic book	0.09	0.04	0.12	0.28	0.53	0.73	0.71	1.51	2.37	3.61

<Figure 23> [end] up ADJ, 1900s-2000s



In summary, we can easily and quickly study a wide range of syntactic phenomena with the 400 million word COHA corpus, which was released in 2010. But the majority of these constructions occur too infrequently to be studied with a small 2-4 million word corpus like ARCHER or the Brown family of corpora.

5. Researching long-range syntactic changes with Google Books-Advanced

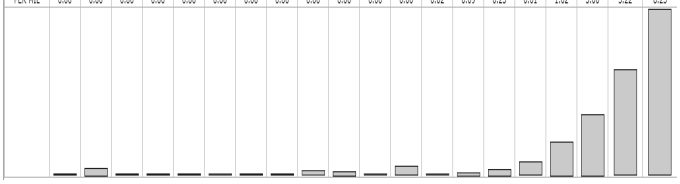
While COHA is composed of 400 million words of text, the Google Books dataset (see http://books.google.com/ngrams) is based on 155 billion words of data from millions of books, and this is just the data from the American English dataset.

Unfortunately, the "standard" Google Books interface (see Michel, Lieberman, et al 2011) is extremely limited and simplistic, as far as syntactic searches. It is difficult or impossible to search by either lemma or part of speech. For example, to search for the construction "end up V-ing" (*ended up paying, ends up looking*, etc), one would have to look – in sequence – for the individual strings *end up paying, ended up paying, ends up paying*, and then

start with tens of thousands of other verbs – all of which would take weeks or months. With the Advanced Google Books interface that we have developed, however (see http://googlebooks.byu.edu), researchers can search by lemma and part of speech, and they could do a search like this in just 1-2 seconds. For example, the following is the data for the construction; note that there are more than 400,000 tokens of this construction.

<Figure 24> Overall frequency of the construction "*end up* V-ing"

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	0	2	0	0	0	0	0	3	1	11	10	29	143	566	2,041	8,085	22,681	46,556	103,440	221,225
PER MIL	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.09	0.25	0.61	1.62	3.00	5.22	8.23



In addition to seeing the overall frequency, researchers can also see the frequency of each matching string in each decade, and then click on any of these to see the book excerpts at books.google.com. (Note: to emphasize the range of verbs following *end*, here we show just the forms with *ended*, but we could see examples with all forms of *end* just as easily).

<Figure 25> Forms of the construction "*end up* V-ing" by decade

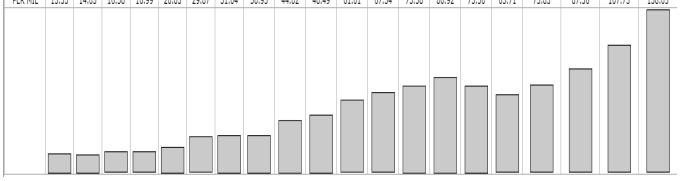
	WORD(S)	CHARTS	TOTAL	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
1	ended up being	G	20784													1	15	46	223	840	1897	5056	12706
2	ended up doing	G	7340													4	12	48	156	438	888	1909	3885
3	ended up having	G	7318													17	9	16	84	276	731	1879	4306
4	ended up getting	G	6193														5	16	58	215	572	1568	3759
5	ended up going	G	5387														4	24	41	201	601	1438	3078
6	ended up taking	G	4550												1		1	11	50	210	446	1227	2604
7	ended up working	G	4230									1				2	3	9	54	205	450	1114	2392
8	ended up making	G	3766										1			1	4	14	52	180	445	912	2157
9	ended up staying	G	3250										1					5	29	126	356	816	1917
10	ended up paying	G	2537														3	10	82	203	296	666	1277
11	ended up losing	G	2305										1					5	41	132	226	562	1338

Another example of a syntactic search that is quite easy and fast in Google Books Advanced (but quite impossible in Google Books Standard) is the increase in the periphrastic future with *going to* (e.g. *going to leave*). We can easily search for "going to [v*]", and we see the overall increase (Figure 26),

as well as all of the matching strings (Figure 27).

<Figure 26> Overall frequency of the construction "*going to* VERB"

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	5,863	9,585	23,839	32,927	61,494	70,264	88,283	136,410	247,945	364,677	615,380	477,373	426,357	499,052	593,971	866,843	1,034,367	1,358,101	2,135,178	3,711,104
PER MIL	15.53	14.63	16.58	16.99	20.83	29.87	31.04	30.95	44.02	48.49	61.01	67.34	73.58	80.92	73.30	65.71	73.83	87.56	107.75	138.05



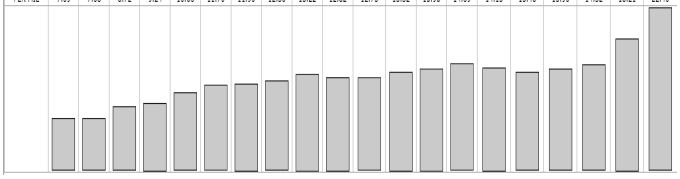
<Figure 27> Forms of the construction "*going to* VERB" by decade

	WORD(S)	CHARTS	TOTAL	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980
1	going to be	G	2366706	604	970	2236	2978	5839	7611	10186	16211	30445	46324	92438	75506	71743	91050	106789	161164	202366	271048
2	going to do	G	900246	152	244	817	1340	3580	4908	5962	9718	18749	28909	51524	38702	33786	37899	43517	59171	71534	94439
3	going to have	G	736856	53	105	391	745	2028	2780	3834	5600	9984	15722	27953	23455	23016	27685	33843	47976	65093	86758
4	going to get	G	561063	15	22	191	295	835	998	1558	2482	5335	9855	20287	16413	17549	23127	26042	37116	46897	64749
5	going to take	G	369650	185	302	860	1115	2258	2454	3335	5137	8797	13174	21271	15538	13173	14789	17421	24423	28248	36740
6	going to make	G	359228	166	306	634	1018	1916	2341	2805	4461	7939	11735	20918	14913	12169	14103	16447	23324	27832	36851
7	going to happen	G	284049	8	30	80	170	403	564	753	1470	3254	4896	10981	8835	9337	11094	13222	19923	24163	31492
8	going to say	G	259531	277	696	1757	1999	2990	3838	4050	5912	10435	12041	16281	12414	10153	10096	13478	17745	17809	23093
9	going to tell	G	238483	80	156	567	747	1449	1823	2273	3156	5896	9159	14067	10427	8704	9080	11511	14758	16867	23749
10	going to give	G	210366	166	191	597	819	1517	1727	2282	3248	6076	9070	13711	10440	8491	8754	10583	15740	16490	20168

Another example is the "*get* passive" construction (e.g. *got returned, get fired*), which is definitely increasing over time. (Again, with Google Books Standard, we would have to perform thousands of separate searches to get this data.)

<Figure 28> Overall frequency of the construction "*get* V-ed"

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	2,677	4,639	12,540	17,907	31,461	27,519	33,852	54,212	74,440	96,402	128,377	95,866	81,022	90,575	114,677	177,796	194,742	225,196	360,866	602,053
PER MIL	7.09	7.08	8.72	9.24	10.66	11.70	11.90	12.30	13.22	12.82	12.73	13.52	13.98	14.69	14.15	13.48	13.90	14.52	18.21	22.40



<Figure 29> Forms of the construction "*get* V-ed" by decade

	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	get rid	G	931233	1621	2793	7502	10322	17251	14507	17624	28349	37775	48944	62219	46060	36883	40321	49451	78282	76018	76696	109779	168636
2	get dressed	G	68451	1	2	14	10	37	48	102	98	217	417	871	991	1331	2055	3149	4457	5652	8939	14840	25220
3	get acquainted	G	66889	64	133	325	422	659	560	736	1235	1644	3178	6028	4957	4136	4980	6213	6529	5412	5299	6440	7939
4	get killed	G	52499	5	1	20	37	120	176	186	402	854	1080	2088	1507	1807	2520	2675	4129	4905	5971	9130	14886
5	get discouraged	G	18099	1	1	20	25	88	108	161	249	318	764	944	759	573	642	761	989	1520	1912	3161	5103
6	get arrested	G	13003				1	1	11	13	19	45	93	272	225	322	210	314	770	1256	1448	2809	5194
7	get bogged	G	12710	1	1		1	2	18	4	10	14	9	40	20	67	126	356	897	1315	1841	3025	4963
8	get published	G	10862	1	1	3	12	20	10	26	37	54	89	110	125	123	150	305	552	871	1409	2435	4529
9	get promoted	G	9642	1	11	5	11	21	21	19	35	44	50	114	129	122	162	347	515	756	1493	2040	3746
10	get thrown	G	9159			2	12	19	14	31	45	61	134	189	131	215	266	342	545	733	995	1911	3514
11	get taken	G	9144		5	6	14	40	68	64	61	194	171	390	239	215	239	318	496	695	1076	1685	3168
12	get hooked	G	8634	1		1	2	4	9	15		13	24	38	53	93	141	152	417	901	1155	2224	3391

To take a somewhat more complex construction, consider the "*way* construction", which has been the focus of a great deal of research in construction grammar (see Israel 1996, Goldberg 1995, and Goldberg 1997 for an introduction). In Google Books Advanced, we can simply search for "[vv*] [ap*] way [i*]" to find more than 1,083,000 tokens for 3000 unique strings like *find their way into, make his way through, groping their way into*, and so on. If desired, we could also compare the verbs (*feel, shove, grope, elbow*, etc) that are used in different periods, to see the influence of semantic factors over time.

<Figure 30> Forms of the construction "V-ed *his way* PREP" by decade

	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	elbowed his way through	G	2449			3	10	32	47	41	39	109	185	136	231	111	172	129	152	200	124	169	239	320
2	wormed his way into	G	2120				1	1	8	12	4	8	23	21	91	118	164	152	208	290	222	183	230	384
3	shouldered his way through	G	1590			1	2		6	7	1	4	31	70	93	81	98	88	148	162	105	135	219	339
4	felt his way along	G	1483			3	6	7	17	20	14	52	57	96	97	107	71	100	81	117	93	100	197	248
5	groped his way through	G	1095	3	9	16	19	39	40	22	42	76	57	90	80	57	81	74	92	62	53	86	97	
6	took his way towards	G	1029	4	12	22	71	103	84	77	115	214	124	37	35	25	4	20	24	24	7	17	10	
7	felt his way through	G	800		1	2	2	12	6	8	10	34	47	68	48	38	63	60	56	59	65	85	136	
8	took his way through	G	793	5	21	45	43	66	51	39	88	79	97	69	45	33	9	15	31	27	6	15	9	
9	shoved his way through	G	785									9	4	9	17	36	34	38	60	44	69	168	297	
10	groped his way into	G	670		6	9	12	35	9	16	38	64	52	66	51	42	45	41	56	28	32	27	41	
11	groped his way along	G	658		2	2	8	38	23	32	25	29	53	61	54	49	34	34	46	28	34	53	53	
12	felt his way into	G	584				1	5	1	10	7	12	29	28	50	25	46	39	62	50	46	44	57	72
13	elbowed his way into	G	581					2	15	8	21	24	34	44	33	30	35	27	36	48	34	34	57	99
14	wended his way through	G	563				2	11	20	16	18	18	21	49	35	22	10	15	14	22	20	25	67	178

Consider one other construction – the "causative V-ing" construction: *talked him into going, coerced them into buying, terrify me into doing*, etc. (For previous discussion of the historical development of and variation with this construction – based on much smaller corpora – see Rudanko 2000 (chapter 5), 2003, 2005, 2006; Rudanko and Luodes 2005 (chapter 2); Gries and Stefanowitsch 2003; and Wulff, Stefanowitsch, and Gries 2007). The one simple search "[vv*] [p*] into [vvg*]" yields 30,200 tokens for 234 different strings.

<Figure 31> Forms of the construction "VERB NP into V-ing" by decade

	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	talk him into going	G	491										4	2	2	12	18	20	40	46	62	104	181
2	talk him into coming	G	358				1								1	1	1	13	20	29	47	89	156
…..																								
32	tricked him into giving	G	54											1		4	2	4	4	4	10	6	19
33	coaxed him into taking	G	54								1	6	1	1		2	5	4	4	7	4	6	13
34	forced him into making	G	52							1		1				4	4		7	6	10	7	12
35	frighten him into giving	G	52							1	1	7	4	1	4	5	3	2	4	4	6	3	7
36	provoke him into making	G	51							1	1	1					1	7	12	6	5	8	9
37	talk him into moving	G	51													1	1	2	1	3	11	10	22
38	talk him into selling	G	49												1	2	5	1	4	6	9	7	14
39	leads him into making	G	49							1	1				4	4	5	3	10	10	1	6	4
40	talk him into helping	G	49														2		3	1	3	16	24
41	coaxed him into going	G	48							2	5	2	3			3	5	3	6	2	2	8	7
42	forced him into taking	G	45			1					2	1	2	1					2	8	9	9	10
43	fooled him into believing	G	44											3	3	3	4	1	4	1	3	6	16
44	coax him into taking	G	43										1	2	2	3	1		3	4	5	8	14
45	mislead him into believing	G	41												5	2	2	7	2	4	3	8	8

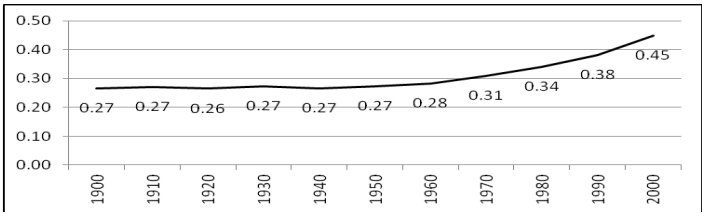
In the examples above, we searched for just one particular string (such as "[end] up [vvg*]" or "[vv*] [ap*] way [i*]") and then retrieved the frequency of each matching string (e.g. Figure 31). But it is also possible to carry out more advanced research as well. For example, we could compare the frequency of two competing constructions to see how one construction is increasing at the expense of the other.

For example, the following table and chart provide data for the use of the subjunctive and indicative in the context "if I/he/she/it was/were" (e.g. *if I was/were*), and is based on 6,153,000 tokens from the 1810s-2000s. (For an introduction to recent changes with the subjunctive in English, see Gonzalez-Alvarez 2003; Peters 1998; and Rohdenburg 2009.) We did one simple search for the subjunctive and then another for the indicative, and then compare the frequencies in a spreadsheet. As can be seen, there is an increase in the use of the indicative since about the 1950s.

<Table 11> Subjunctive vs indicate with *if*

	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
Subj	271,173	314,531	228,277	185,032	189,935	239,876	361,927	331,667	345,971	465,334	629,778
Indic	98,417	116,839	82,125	69,514	68,948	89,690	142,617	147,712	178,489	287,666	510,332
% indic	0.27	0.27	0.26	0.27	0.27	0.27	0.28	0.31	0.34	0.38	0.45

<Figure 32> Percentage of *if* clauses with indicative (vs. subjunctive)

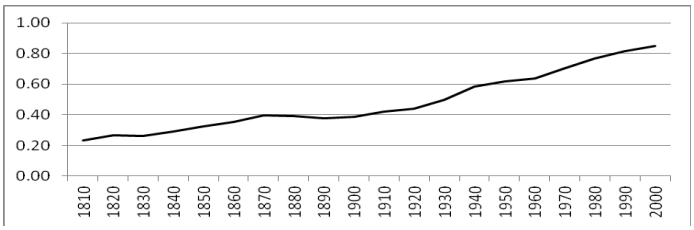


Another example of syntactic variation over time deals with verbal subcategorization, in this case whether or not *to* is used in complements of *help* (*help him* to *do it* vs *help him* − *do it*) (see Table 4 above for data from the 1990s-2000s). Two simple queries yield 3,812,000 tokens, which show a clear increase in the omission of the complementizer *to*.

<Table 12> *help* NP (*to*) VERB

	1820	1840	1860	1880	1900	1920	1940	1960	1980	2000
to	1,481	5,058	8,242	17,563	40,179	46,904	42,336	89,654	90,702	209,335
zero	533	2,077	4,541	11,285	25,600	37,196	59,769	157,044	302,561	1,186,332
% zero	0.26	0.29	0.36	0.39	0.39	0.44	0.59	0.64	0.77	0.85

<Figure 33> Percentage of *help* complements without *to*



The contrast between Google Books Standard and Google Books Advanced − in terms of how they can be used to look at syntactic change − is quite striking. For example, in the case of the "causative V-ing" construction discussed above ("V1 NP into V2-ing"), we would have to search for [thousands of V1] x [thousands of V2] x [all possible pronouns] (e.g. *forced*

him into accepting, coax us into returning). There would be hundreds of thousands or even millions of unique strings, and it would take months or perhaps years to carry out this research in GB-S. In GB-Adv, on the other hand, we have all of the data in just 2-3 seconds.

Finally, notice the incredibly large number of tokens for these constructions. For example, there are nearly 4 million tokens of the "help (to) VERB" construction (Table 12, Figure 33), and this number of tokens for this one minor construction is almost twice as large as the total number of words in some corpora such as ARCHER and the Brown family of corpora.

6. Researching dialectal variation in syntax with COCA and the BNC

In addition to genre-based variation in syntax and historical change in syntax, with the right type of corpus we can look at dialectal variation in syntax. In this section, we will consider differences between British and American English, which are the two dialects that have been compared in most detail.

The most typical route to studying syntactic differences between British and American English has been to use the four million words of text in the Brown family of corpora. These corpora are comprised of the Brown corpus (1 million words, US, 1960s), LOB (1 million words, UK, 1960s), FROWN (1 million words, US, 1990s), and FLOB (1 million words, UK, 1990s). Unfortunately, with this approach, only very high frequency constructions such as modals and auxiliaries can be studied.

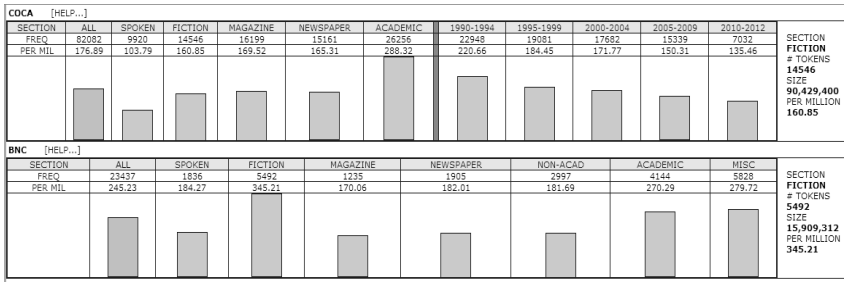
For example, Leech et al (2009) is a collection of papers that look at syntactic differences between British and American English, and they are based primarily on the four corpora in the Brown family. An investigation of the chapters in this book show that more than half deal with just very high frequency phenomena like modals, progressives, passives, and high-frequency phenomena related to the noun phrase. So as insightful as these studies might be for high frequency syntactic studies (and these corpora have been of great value for studying certain types of syntactic change, during the past few decades), these corpora do not have enough data to be used for many

medium- and low-frequency syntactic constructions (see Davies 2012a, 2012b, and 2012c for a more complete discussion of this issue).

Fortunately, with the release of COCA in 2008, we now have a large corpus of American English (450 million words, 1990-2012) and with the British National Corpus (BNC) a large corpus of British English (100 million words, 1980s-1993), which can be compared against each other to look at a wide range of syntactic constructions in the two dialects, and not just high frequency constructions, as with the Brown family of corpora. These comparisons are also greatly facilitated by the fact that – with just one click – users can re-do a COCA search in the BNC (or a BNC search in COCA), to compare the two dialects.

As an initial example, consider the following data from the BNC and COCA, which shows that *must* + lexical verb (e.g. *they must admit that…*) is more common in British than American English (245 tokens per million in the BNC; 177 in COCA). Note that this has already been shown in previous research, but the fact that it shows up so nicely in the BNC and COCA as well should be reassuring to those whose research has been limited primarily to the Brown family of corpora. Note also that in COCA, [*must* + lexical verb] is least common in the most informal dialect (Spoken) and the most common in the most formal dialect (Academic), and that its frequency is decreasing in each five-year period since the early 1990s.

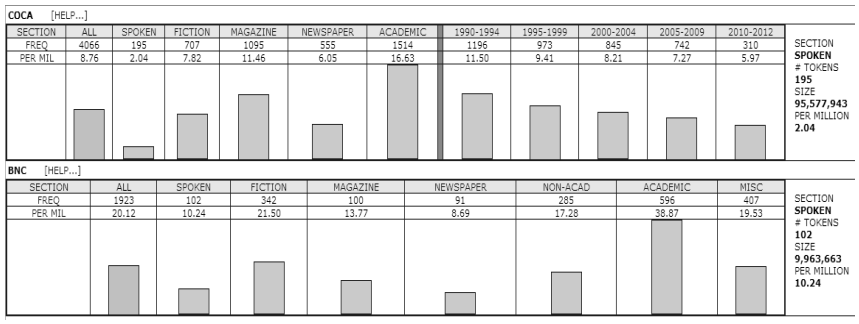
<Figure 34> *must* + lexical verb in COCA/BNC



Let us now turn to a somewhat less frequent construction – post-verbal negation with the verb *need* (e.g. *they need not concern you*). The Brown family of corpora have 45 tokens in the US corpora (Brown and Frown) and 69 in

the British corpora (LOB and FLOB). In COCA and the BNC there are nearly 6,000 tokens. In less than five seconds, we can see that the construction is more than twice as common in the BNC, and that in COCA, the construction is associated mainly with the more formal genres (e.g. eight times as common in Academic as Spoken), and that the construction is decreasing in frequency over time.

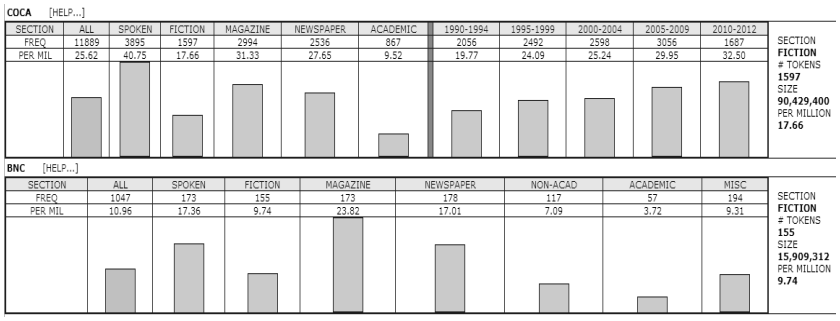
<Figure 35> *need* + NEG + VERB in COCA/BNC



Turning to an even less frequent construction, we find that there are only 31 tokens of the [*end up* V-ing] construction in the Brown corpora (e.g. *they ended up paying too much*). Even with this small amount of data, however, it looks like the construction is more common in the US (21 vs 10 tokens) and that it is increasing from the 1960s to the 1990s (3 vs 28 tokens).

Of course, the data from COCA and the BNC is much more robust. There are nearly 13,000 tokens, and they show that the [*end up* V-ing] construction is more than twice as common in the US as in the UK, that in the US (but not UK) it is the most common in the informal genres, and that it is increasing in frequency in each five-year period in the US (of course there is no such diachronic data for the BNC, since it is not designed to be used as a historical or monitor corpus).

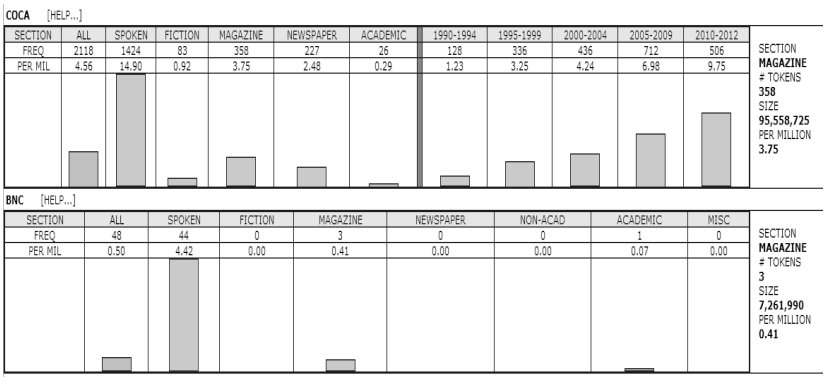
<Figure 36> [*end up* V-ing] in COCA/BNC



Remember, however, that the BNC is limited to texts from a generation ago (the 1980s and early 1990s), whereas COCA is added to year-by-year (and thus currently included texts through 2012). If the construction is increasing over time, then any more recent corpus (e.g. COCA, which alone includes texts from the last 20 years) will have more tokens.

Let us now examine an even more interesting and recent construction: the "quotative *like*" construction, e.g. "*and I'm like, I don't want it*"), which has been discussed in the sections above. The following data from COCA and the BNC show that it is nearly ten times as frequent in COCA (4.6 per million COCA and 0.5 in the BNC). In addition, it is most common in the more informal genres in COCA, and it is increasing in each five-year period in COCA.

<Figure 37> Quotative *like* construction in COCA/BNC

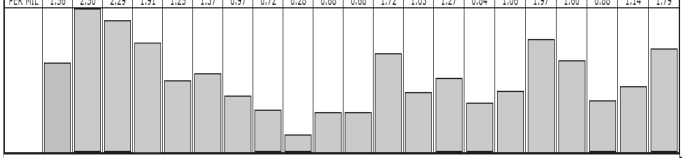


Again, however, we have to worry about the fact that we are "comparing apples and oranges" to some degree as we use COCA (continually updated; current as of 2012) and the BNC (now a generation old). Any construction that is increasing over time has the potential to appear more common in American English by the mere fact that COCA is a more modern corpus.

Interestingly, if we look at a corpus whose texts in British and American English are completely contemporaneous, this huge gap with the "quotative *like*" construction is much smaller. For example, the 1.9 billion word GloWbE corpus (web pages from 20 English-speaking countries, 2012-2013) shows that "quotative *like*" is only slightly more frequent in American than British English (2.5 per million in US and 1.9 per million in GB (Great Britain)), and the KWIC lines following that provide examples of the construction from the GB portion of the corpus.

<Figure 38> Quotative *like* construction in GloWbE

SECTION	ALL	US	CA	GB	IE	AU	NZ	IN	LK	PK	BD	SG	MY	PH	HK	ZA	NG	GH	KE	TZ	JM
FREQ	3114	966	309	740	126	203	79	69	13	35	27	74	43	55	34	48	84	62	36	40	71
PER MIL	1.56	2.50	2.29	1.91	1.25	1.37	0.97	0.72	0.28	0.68	0.68	1.72	1.03	1.27	0.84	1.06	1.97	1.60	0.88	1.14	1.79



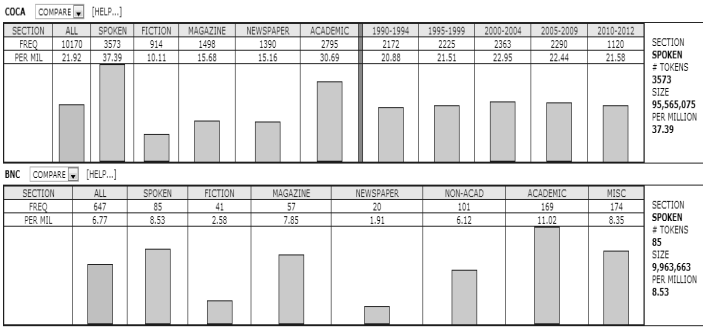
<Figure 39> Concordance lines for "quotative *like*" in Britsh portion of GloWbE

1	GB	G	guru.bafta.org	A	B	C	can attest to that. They were very excited about it, **so I was like** ,' Holy shit, now I have to do a play!' because
2	GB	G	guru.bafta.org	A	B	C	' Well sure, it's all downhill that way.' **And I was like** ,' Oh, okay. He's got a repertoire, and doesn't
3	GB	G	femalefirst.co.uk	A	B	C	who I love, and I was showing him my stomach, **and he was like** ,' Errr...' and I was like,' I'll do a
4	GB	G	femalefirst.co.uk	A	B	C	my stomach, and he was like,' Errr...' **and I was like** ,' I'll do a handstand for you!' And I kicked him
5	GB	G	eurogamer.net	A	B	C	, but over here is the area of effect fear. Things **where it 's like** , in some dungeons this might be really good, but in this encounter this
6	GB	G	blokely.com	A	B	C	' # " But touring is all I've ever done **since I was like** , 18, 20 years old. I couldn't do anything else I used
7	GB	G	eurogamer.net	A	B	C	light the lamps with whale-oil tanks, " says Smith. " **And we were like** , you know, what if this was a combatant who was made to burn
8	GB	G	dailymail.co.uk	A	B	C	money. # ' The public seem to be picking on her **and I 'm like** ," Carry on, keep on picking on Dorries, she deserves everything she
9	GB	G	uk.answers.yahoo.com	A	B	C	UK? I had my induction day at college on the 17th **and it was like** , meh. I didn't know anyone there and I thought everyone else was
10	GB	G	dailymail.co.uk	A	B	C	just said, do you actually wan na be my girlfriend? **And I was like** , yeah. I really liked him. I knew Gemma then for about two

Of course, not all of the dialectal differences in syntax are due to the fact that COCA is a generation more recent than the BNC. For example, consider the data with the two competing constructions [*all the* NOUNs] and [*all of the*

NOUNs] (e.g. *all* (*of*) *the reasons*). The following chart shows the frequency of [*all* ***of*** *the* NOUNs] in COCA and the BNC, and we see that it is much more common in COCA. Notice, however, the genre patterning in COCA, where the construction is not limited to primarily formal or informal genres, and note also that the frequency is fairly static over time. Nevertheless, the construction is more than three times as frequent in COCA as in the BNC (21.9 tokens per million in COCA, 6.8 in the BNC).

<Figure 40> [*all* ***of*** *the* NOUN] in COCA and BNC



If we compare the frequency of the two constructions in COCA and the BNC, we see that the construction with *of* is much more common in American English, and this difference is significant (using Chi square) at $p < .00001$.

<Table 13> [all (of) the NOUN] in COCA and BNC

	all the [nn2]	all **of** the [nn2]	% all **of**
COCA	58,345	10,170	14.8%
BNC	15,116	647	4.1%

Again, however, the data from the much smaller Brown family of corpora is much less helpful. In this case, the results from the two dialects are virtually the same, and (using Chi square) there is no significant difference between the two dialects.

<Table 14> [all (of) the NOUN] in the Brown family of corpora

	all the [nn2]	all of the [nn2]	% all **of**
Am: Brown/Frown	295	20	6.3%
Br: LOB/FLOB	293	22	6.9%

7. Researching dialectal variation in syntax in other World Englishes

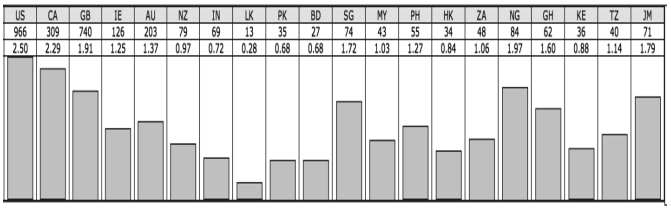
Obviously, there are more than just the British and American dialects of English, and it would be nice to be able to compare a wide range of dialects from the English-speaking world. Until recently, perhaps the most ambitious to do this from a corpus-based perspective has been the International Corpus of English (ICE), which is composed of one million words each from a number of English-speaking countries.

However, just as one or two million words was too small in terms of the historical corpora or for comparisons of just British and American English (with the Brown family of corpora), the same is true for ICE. At one million words each, these corpora are too small to look at anything but the most frequent syntactic constructions, such as modals and auxiliaries (which have already been extensively studied during the past two decades).

As a result, we have recently released the Global Web-based English (GloWbE) corpus, which contains nearly two billion words of English from twenty different countries. The countries with the largest corpora are the US and the UK (about 385 million words each), but there are also at least 40 million words each from the other countries as well (and in many cases many more than that): Canada, Ireland, Australia, New Zealand, India, Sri Lanka, Pakistan, Bangladesh, Singapore, Malaysia, Philippines, Hong Kong, South Africa, Nigeria, Ghana, Kenya, Tanzania, and Jamaica.

To see how large corpora such as GloWbE can be used to look at dialectal variation in English in ways that cannot be done with smaller corpora, consider again the "*like* construction", which has been discussed above. Figure 41 (copied from Figure 38 above) shows the frequency of the construction (3,114 tokens total) in each of the twenty dialects in GloWbe:

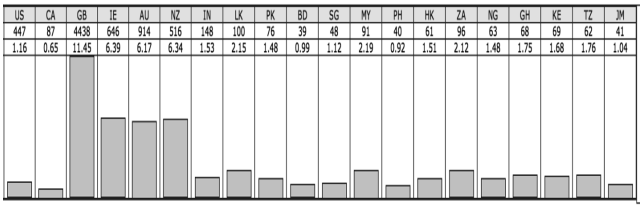
<Figure 41> "Quotative *like*" in twenty dialects of GloWbE



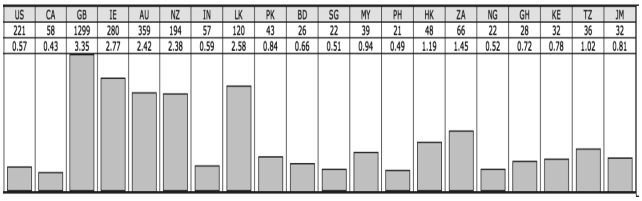
Note that the construction is the most frequent in the US and Canada, but that it is also relatively common in the other "core" countries of English as well, including Great Britain (GB), Ireland (IE), Australia (AU), and New Zealand (NZ) – in roughly descending order of frequency. But the important point is the raw frequency in these countries. In the US it occurs 966 times in 385 million words, or about 2.5 tokens per million words. In Great Britain it is even less common – 740 tokens in about 385 million words, or about 1.9 tokens per million words. If we had only one million words from each dialect (as in ICE), we would be comparing two or three tokens in one dialect and perhaps one token each in the other dialects, which would be quite meaningless. But with the two billion words in GloWbE, we have enough data to make interesting comparisons.

To take another concrete example, consider the construction "stop/prevent NP (from) V-ing" (e.g. *stop them saying that, prevent them from doing such things*). The following charts show the frequency of the construction without *from* with *stop* and *prevent*, followed by the construction with *from* in the different dialects.

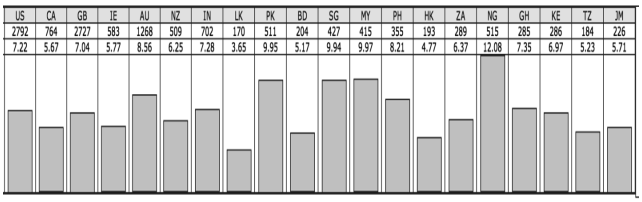
<Figure 42> [stop + PRON + V-ing] in GloWbE



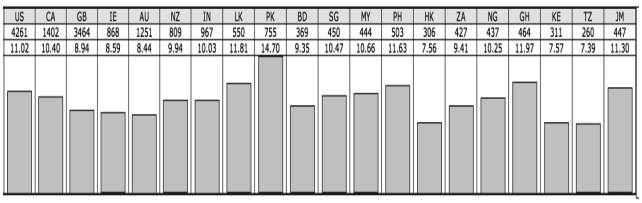
<Figure 43> [prevent + PRON + V-ing] in GloWbE



<Figure 44> [stop + PRON + from V-ing] in GloWbE



<Figure 45> [prevent + PRON + from V-ing] in GloWbE

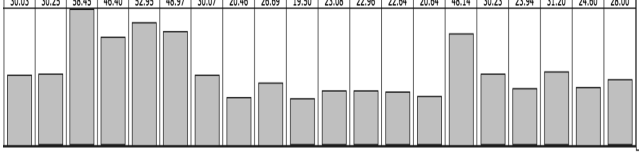


As can be clearly seen, the variant without *from* is much less frequent in the United States and Canada than in the other "core" dialects of English (GB, IE, AU, NZ). But again, the important point is the number of tokens in the different dialects – often just 2-5 tokens per million words, in many cases. With a corpus like ICE, which only has one million words per dialect, we could be quite unable to look at verbal subcategorization in cases like this.

Consider now the "try and VERB" construction (e.g. *they should try and do it tomorrow*). This is a construction that is proscribed in style guides in the United States and Canada (where only *try TO* is accepted). And the GloWbE data clearly shows the effect of this prescriptive rule – the construction is much less frequent in the US and Canada than in the other "core" dialects. So GloWbE can be used to look at prescriptive issues in interesting ways as well.

<Figure 46> [try and VERB] in GloWbE

US	CA	GB	IE	AU	NZ	IN	LK	PK	BD	SG	MY	PH	HK	ZA	NG	GH	KE	TZ	JM
11614	4077	22657	4688	7847	3986	2900	953	1371	770	992	956	979	835	2184	1289	928	1281	865	1108
30.03	30.25	58.45	46.40	52.95	48.97	30.07	20.46	26.69	19.50	23.08	22.96	22.64	20.64	48.14	30.23	23.94	31.20	24.60	28.00



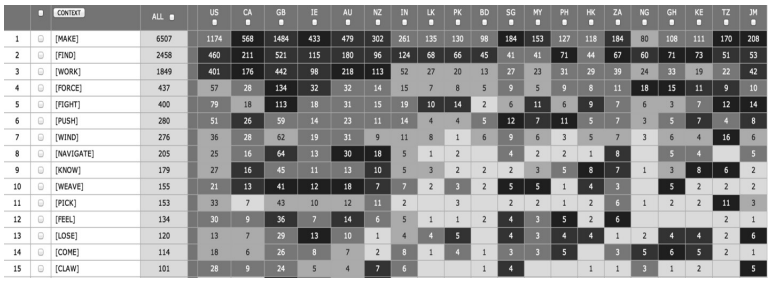
With GloWbE, it is also possible to see all of the matching strings for certain constructions, as well as the frequency of each of these strings in each of the twenty dialects. For example, Figure 47 shows the strings with the "go + ADJ" construction (*go crazy, go bankrupt*, etc.), where there is often strong negative semantic prosody.

<Figure 47> [go ADJ] in GloWbE

		CONTEXT	ALL	US	CA	GB	IE	AU	NZ	IN	LK	PK	BD	SG	MY	PH	HK	ZA	NG	GH	KE	TZ	JM
1		[GO] [CRAZY]	5680	1494	384	1089	215	418	184	281	70	122	73	316	208	163	102	89	130	84	66	54	138
2		[GO] [BANKRUPT]	4142	1654	328	784	160	274	102	168	53	42	60	90	119	42	70	45	39	38	32	13	29
3		[GO] [VIRAL]	3842	997	321	604	185	276	78	234	56	66	70	148	111	140	69	104	124	66	74	32	87
4		[GO] [LIVE]	3508	824	221	975	200	276	180	154	73	46	39	92	42	40	30	95	42	48	59	37	35
5		[GO] [MAD]	3398	415	94	1353	341	298	100	185	60	61	45	50	60	55	39	54	64	31	36	13	44
6		[GO] [WRONG]	3375	627	183	986	144	256	132	188	93	77	50	64	91	51	36	83	83	88	61	47	35
7		[GO] [UNNOTICED]	3320	639	273	733	182	217	122	206	107	89	52	54	68	72	35	93	68	88	84	57	81
8		[GO] [BAD]	2627	896	240	367	64	196	68	103	23	51	38	67	50	60	41	29	81	53	55	30	115
9		[GO] [MISSING]	2599	402	150	797	162	239	105	131	117	54	40	38	55	28	38	57	21	36	46	25	58
10		[GO] [STRONG]	2317	488	162	578	139	176	74	128	28	39	34	62	59	39	39	104	25	29	32	26	56
11		[GO] [HUNGRY]	2173	460	139	446	132	143	91	87	30	58	61	48	34	55	21	47	45	56	111	58	51
12		[GO] [BUST]	2033	255	54	1105	198	111	76	45	19	7	7	26	34	8	28	14	7	12	14	7	6
13		[GO] [WILD]	1860	454	122	405	109	151	64	57	25	29	15	41	49	46	19	43	44	39	39	30	79

Another example is the "way construction" (e.g. *made his way through the crowd, fought her way through the pitfalls*), which has been a favorite topic of research from within the Construction Grammar framework (see Figure 30 from COHA above, along with citations there). Using GloWbE, in just three or four seconds we can find all matching strings, and see their frequency in each dialect:

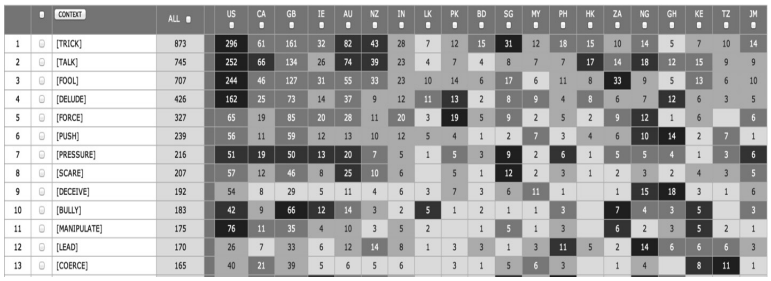
<Figure 48> The "*way* construction" in GloWbE



	CONTEXT	ALL	US	CA	GB	IE	AU	NZ	IN	LK	PK	BD	SG	MY	PH	HK	ZA	NG	GH	KE	TZ	JM
1	[MAKE]	6507	1174	568	1484	433	479	302	261	135	130	98	184	153	127	118	184	80	108	111	170	208
2	[FIND]	2458	460	211	521	115	180	96	124	68	66	45	41	41	71	44	67	60	71	73	51	53
3	[WORK]	1849	401	176	442	98	218	113	52	27	20	13	27	23	31	29	39	24	33	19	22	42
4	[FORCE]	437	57	28	134	32	32	14	15	7	8	5	9	5	9	8	11	18	15	11	9	10
5	[FIGHT]	400	79	18	113	18	31	15	19	10	14	2	6	11	6	9	7	6	3	7	12	14
6	[PUSH]	280	51	26	59	14	23	11	14	4	4	5	12	7	11	5	7	3	5	7	4	8
7	[WIND]	276	36	28	62	19	31	9	11	8	1	6	9	6	3	5	7	3	6	4	16	6
8	[NAVIGATE]	205	25	16	64	13	30	18	5	1	2		4	2	2	1	8		5	4		5
9	[KNOW]	179	27	16	45	11	13	10	5	3	2	2	2	3	5	8	7	1	3	8	6	2
10	[WEAVE]	155	21	13	41	12	18	7	7	2	3	2	5	5	1	4	3		5	2	2	2
11	[PICK]	153	33	7	43	10	12	11	2		3		2	2	1	2	6	1	2	2	11	3
12	[FEEL]	134	30	9	36	7	14	6	5	1	1	2	4	3	5	2	6				2	1
13	[LOSE]	120	13	7	29	13	10	1	4	4	5		4	3	4	4	1	2	4	4	2	6
14	[COME]	114	18	6	26	8	7	2	8	1	4	1	3	3	5		3	5	6	5	2	1
15	[CLAW]	101	28	9	24	5	4	7	6			1	4			1	1	3	1	2		5

One final example of a construction that has been a popular topic from within the Construction Grammar framework is the "into V-ing causative" (e.g. *he talked her into staying, they forced me into admitting it*) (see Figure 30 from COHA and related citations above). With GloWbE, we can again easily see the matching strings (here represented by just the matrix verb with these constructions: *trick, talk, fool*, etc).

<Figure 49> The "causative into V-ing" construction in GloWbE



	CONTEXT	ALL	US	CA	GB	IE	AU	NZ	IN	LK	PK	BD	SG	MY	PH	HK	ZA	NG	GH	KE	TZ	JM
1	[TRICK]	873	296	61	161	32	82	43	28	7	12	15	31	12	18	15	10	14	5	7	10	14
2	[TALK]	745	252	66	134	26	74	39	23	4	7	4	8	7	7	17	14	18	12	15	9	9
3	[FOOL]	707	244	46	127	31	55	33	23	10	14	6	17	6	11	8	33	9	5	13	6	10
4	[DELUDE]	426	162	25	73	14	37	9	12	11	13	2	8	9	4	8	6	7	12	6	3	5
5	[FORCE]	327	65	19	85	20	28	11	20	3	19	5	9	2	5	2	9	12	1	6		6
6	[PUSH]	239	56	11	59	12	13	10	12	5	4	1	2	7	3	4	6	10	14	2	7	1
7	[PRESSURE]	216	51	19	50	13	20	7	5	1	5	3	9	2	6	1	5	5	4	1	3	6
8	[SCARE]	207	57	12	46	8	25	10	6		5	1	12	2	3	1	2	3	2	4	3	5
9	[DECEIVE]	192	54	8	29	5	11	4	6	3	7	3	6	11	1		1	15	18	3	1	6
10	[BULLY]	183	42	9	66	12	14	3	2	5	1	2	1	1	3		7	4	3	5		3
11	[MANIPULATE]	175	76	11	35	4	10	3	5	2		1	5	1	3		6	2	3	5	2	1
12	[LEAD]	170	26	7	33	6	12	14	8	1	3	3	1	3	11	5	2	14	6	6	6	3
13	[COERCE]	165	40	21	39	5	6	5	6		3	1	5	6	3		1	4		8	11	1

Again, the important point for each of the three constructions just discussed is that the token count for each matching string in each dialect is quite small. If we had just a one million word corpus (as with ICE), for example, then rather than 244, 24, and 127 tokens with *fool* in the "into V-ing" construction (shown above in Figure 49), we could have just two or three tokens in each dialect – which would be far too few to say much about the construction.

While the previous examples dealt with constructions, we can also use GloWbE to look at many phenomena that straddle lexis and grammar. To

give just one example, Figure 50 shows the prepositions that are used with the word *integrated* in the different dialects.

<Figure 50> *Integrated* + PREP in GloWbE

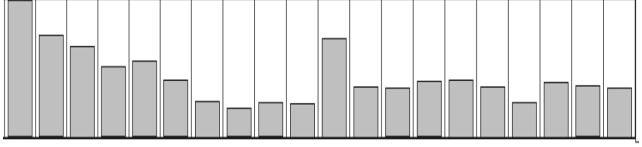
	CONTEXT	ALL	US	CA	GB	IE	AU	NZ	IN	LK	PK	BD	SG	MY	PH	HK	ZA	NG	GH	KE	TZ	JM
1	INTEGRATED INTO	5899	983	517	1063	323	419	233	309	156	123	150	147	98	119	195	221	138	173	245	166	121
2	INTEGRATED WITH	2968	485	224	478	142	191	135	323	97	67	118	104	81	46	122	84	36	57	81	59	38
3	INTEGRATED IN	1186	164	67	211	51	48	28	116	21	36	47	40	33	46	44	37	17	36	64	49	31
4	INTEGRATED WITHIN	260	38	22	56	10	24	11	17	8	6		3	4	5	11	10	1	4	17	7	6
5	INTEGRATED TO	187	23	12	20	4	13	6	28	13	3	11	5	6	11	5	6	5	6	6	1	3
6	INTEGRATED BY	125	11	23	18	3	9	2	15	6	3	4	1	1	4	7	1	2	4	3	7	1
7	INTEGRATED ACROSS	97	22	7	14	8	11	9	8		2		2		4	2	5			1	1	1
8	INTEGRATED AS	96	12	9	24	4	6	1	7	2	1	2	2	7	3	3		3	2	3	5	
9	INTEGRATED ON	75	10	3	16	2	5	5	10	3	4	2	2	2	2	1	3	3			1	1

While the "core" dialect prefer the prepositions *into* (and *with*), South Asian dialects such as India allow a number of otherwise "non-standard" prepositions, such as *in, within*, and *to* (e.g. *they adopted Hinduism and integrated it* in *the Indian caste system*).

Just as the preceding example straddles lexis and syntax, with GloWbE we can also easily look at phenomena that straddle syntax and discourse. For example, the following example shows the frequency of the phrase ". *that said* ,". As can be seen, it is the most common in US English, and then is progressively less common in Canada (CA), Great Britain (GB), Ireland (IE), Australia (AU), and New Zealand (NZ). And again, in one million word corpora like those in the International Corpus of English, the token counts would be quite small – typically just 5-10 tokens per dialect in ICE – whereas in GloWbE they occur a total of more than 24,000 times in the corpus.

<Figure 51> "*that said*" in GloWbE

US	CA	GB	IE	AU	NZ	IN	LK	PK	BD	SG	MY	PH	HK	ZA	NG	GH	KE	TZ	JM
8222	2131	5478	1097	1736	713	529	199	268	201	662	323	323	347	398	331	204	347	276	300
21.26	15.81	14.13	10.86	11.71	8.76	5.49	4.27	5.22	5.09	15.40	7.76	7.47	8.58	8.77	7.76	5.26	8.45	7.85	7.58



8. Conclusion

In this paper, we have provided many different concrete examples of how English grammar varies in important ways, as a function of differences between genres, as a function of language change, and as a function of differences between dialects. All of this data shows that it is far too simplistic to say that "Structure X is acceptable or common in English", when in fact it may be in one historical period but not in the corpus 50 or 100 years before, or in just American English but not British English, or in just academic English but not spoken English.

As we have seen, several recent corpora – such as COCA, COHA, GloWbE, and the BYU interface to the BNC – allow us to accurately examine this full range of variation. In addition, as we have seen, the reason that these corpora often provide data that is more insightful than other corpora is because these corpora are both more recent and much larger than previous corpora. Corpus size is often a crucial factor. New 400 million word corpora like COHA can provide 100 times as much data as previous small corpora like the Brown family of corpora or ARCHER, and a corpus like GloWbE provides 100 times as much data as the combined ICE corpora. These new corpora are also very user friendly, especially in the sense that it is possible – via the corpus interface at corpus.byu.edu – to seamlessly move between these different corpora (with just one click) to compare phenomena over time, between dialects, and between genres. The end result is that these new corpora allow us to provide a much more reliable and insightful view into English syntax than was possible even four or five years ago.

References

Barbieri, F. 2009. Quotative 'Be Like' in American English: Ephemeral or Here to Stay?. *English World-Wide* 30, 68-90.

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

Buchstaller, I. and D'Arcy, A. 2009. Localized Globalization: A Multi-Local,

Multivariate Investigation Of Quotative 'Be Like'. *Journal of Sociolinguistics* 13, 291-331.

Close, R. 1987. Notes on the Split Infinitive. *Journal of English Linguistics* 20, 217-229.

Davies, M. 2009. The 385+ Million Word Corpus of Contemporary American English (1990-2008+): Design, Architecture, and Linguistic Insights. *International Journal of Corpus Linguistics* 14, 159-90.

Davies, Mark. 2011. The Corpus of Contemporary American English as the First Reliable Monitor Corpus of English. *Literary and Linguistic Computing* 25, 447-65.

Davies, Mark. 2012a. Examining Recent Changes in English: Some Methodological Issues. In Terttu Nevalainen and Elizabeth Closs Traugott (eds.). *Handbook on the History of English: Rethinking Approaches to the History of English*, 263-87. Oxford: Oxford Univ. Press.

Davies, Mark. 2012b. Expanding Horizons in Historical Linguistics with the 400 million word Corpus of Historical American English. *Corpora* 7, 121-57.

Davies, Mark. 2012c. Recent shifts with three nonfinite verbal complements in English: Data from the 100 million word TIME Corpus (1920s-2000s). In Bas Aarts, Joanne Close, Geoffrey Leech, Sean Wallis (eds.). *Current Change in the English Verb Phrase*, 46-67. Cambridge: Cambridge Univ. Press.

Facchinetti, R. 2000. Be Able to in Present-Day British English. In Mair, C. and Hundt, M. (eds), *Corpus Linguistics and Linguistic Theory*. Amsterdam: Rodopi. 117-130.

Francis, Gill, Susan Hunston and Elizabeth Manning, eds. 1996. *Collins Cobuild Grammar Patterns 1: Verbs*. London, HarperCollins

Goldberg, A. 1997 "Making One's Way Through the Data". In M. Shibatani and S. Thompson (eds.), *Grammatical Constructions: Their Form and Meaning*. Oxford: Clarendon Press, 29-53.

Goldberg, A. 1995. Constructions: A construction grammar approach to argument structure. Chicago: University of Chicago Press.

Gonzalez-Alvarez, D. 2003. "If he come vs. if he comes, if he shall come: Some Remarks on the Subjunctive in Conditional Protases in Early and Late Modern English". *Neuphilologische Mitteilungen* 104.3, 303-313.

Gries, S., and A. Stefanowitsch. 2003. "Co-Varying Collexemes in the Into-Causative". In M. Achard and S. Kemmer (eds.), *Language, Culture, and Mind*. Stanford, CA: CSLI Publications, 225-36.

Hundt, M. 2001. "What Corpora Tell Us about the Grammaticalisation of Voice in get-Constructions". *Studies in Language* 25.1, 49-87.

Hunston, Susan, and Gill Francis. 2000. *Pattern Grammar: A Corpus-Driven Approach to the Lexical Grammar of English*. Amsterdam: John Benjamins.

Israel, M. 1996. "The Way Constructions Grow". In A. Goldberg (ed.), *Conceptual Structure, Discourse and Language*. Stanford: CSLI, 217-230.

Kjellmer, G. 1985. Help to / Help Ø Revisited, *English Studies* 66, 156-161.

Leech, G. 2003. Modality on the move: the English modal auxiliaries 1961-1992. In Facchinetti, R., Krug, M., and Palmer, F. (eds.) *Modality in Contemporary English*. Berlin: Mouton de Gruyter. 224-40.

Leech, Geoffrey, Marianne Hundt, Christian Mair and Nick Smith. 2009. Change in Contemporary English: A Grammatical Study. Cambridge: Cambridge University Press

Mair, C. 2006 Tracking Ongoing Grammatical Change and Recent Diversification in Present-Day Standard English: The Complementary Role of Small and Large Corpora. In Renouf, A. and Kehoe, A. (eds). *The Changing Face Of Corpus Linguistics*. Amsterdam: Rodopi. 355-76

Mair, C. 1995. "Changing patterns of complementation, and concomitant grammaticalisation, of the verb help in present-day British English". In B. Aarts and C. Meyer (eds.), The verb in contemporary English: theory and description. Cambridge: Cambridge University Press, 258-72.

Mair, C. 2002. "Three Changing Patterns of Verb Complementation in Late Modern English: A Real-Time Study Based on Matching Text Corpora". *English Language and Linguistics* 6, 105-131.

McEnery, T., Xiao, Z. (2005). Help or Help to: What Do Corpora Have to Say?, *English Studies* 86, 161-187.

Michel, J. B., Y. Kui Shen, A. Presser Aiden, A. Veres, M. Gray, The Google Books Team, J. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. Nowak, and E. Lieberman Aiden. 2011. "Quantitative Analysis of Culture Using Millions of Digitized Books". *Science* 331, 176-182.

Millar, N. 2009. Modal verbs in TIME: Frequency changes 1923-2006, *International Journal of Corpus Linguistics* 14, 191-220.

Peters, P. 1998. "The Survival of the Subjunctive: Evidence of Its Use in Australia and Elsewhere". *English World-Wide* 19.1, 87-103.

Rohdenburg, G. 2009. "Grammatical Divergence between British and American English in the Nineteenth and Early Twentieth Centuries". *Linguistic Insights – Studies in Language and Communication* 77, 301-329.

Rudanko, J, and Lea Luodes. 2005. Complementation in British and American English. Lanham, MD: University Press of America.

Rudanko, J. 2000. Corpora and Complementation. Lanham, MD: University Press of America.

Rudanko, J. 2003. "Comparing Alternate Complements of Object Control Verbs: Evidence from the Bank of English Corpus". In P. Leistyna and C. Meyer (eds.), *Corpus Analysis: Language Structure and Use*. Amsterdam: Rodopi, 273-83.

Rudanko, J. 2005. "Lexico-Grammatical Innovation in Current British and American English: A Case Study on the Transitive into -ing Pattern with Evidence from the Bank of English Corpus". *Studia Neophilologica* 77, 171-87.

Rudanko, J. 2006a. "Watching English grammar change: A case study on complement selection in British and American English". *English Language and Linguistics* 10, 31-48.

Rudanko, Juhani. 2006b. Emergent Alternation in Complement Selection: The spread of the Transitive *into -ing* Construction in British and American English. *English Linguistics* 34.4, 312-331.

Ruhlemann, C. 2007. Lexical Grammar: The GET-Passive As A Case In Point. *ICAME Journal* 31, 111-127.

Tagliamonte, S. and D'Arcy, A. 2004. He's like, She's like: The Quotative System in Canadian Youth. *Journal of Sociolinguistics* 8, 493-514.

Wulff, S., A. Stefanowitsch, S. Gries. 2007. "Brutal Brits and Persuasive Americans: Variety Specific Meaning Construction in the into-Causative". In G. Radden, et al (eds.), *Aspects Of Meaning Construction*. Amsterdam: John Benjamins, 265-281.

Department of Linguistics and English Language
Brigham Young University
Mark_Davies@byu.edu

Received: October 31, 2013
Reviewed: December 11, 2013
Accepted: December 12, 2013