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
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):

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

| SPOKEN | FICTION | MAGAZINE | NEWSPAPER | ACADEMIC |
|--------|---------|----------|-----------|----------|
| 1,286.78 | 1,898.17 | 2,933.75 | 2,486.13 | 2,956.80 |

In addition to seeing the overall frequency by genre, we can also see the
individual matching strings in each genre, as in Figure 4.
Participial adjectival modification: ing.[j] [nn*]:
individual forms

| CONTEXT                  | ALL | SPOKEN | FICTION | MAGAZINE | NEWSPAPER | ACADEMIC |
|--------------------------|-----|--------|---------|----------|-----------|----------|
| DEVELOPING COUNTRIES     | 3607| 179    | 566     | 421      | 344       | 3442     |
| GROWING NUMBER           | 2600| 236    | 742     | 634      | 398       | 398      |
| FOLLOWING YEAR           | 1783| 73     | 104     | 605      | 444       | 357      |
| WORKING CLASS            | 1508| 248    | 58      | 187      | 106       | 789      |
| BOILING WATER            | 1382| 15     | 141     | 834      | 365       | 27       |
| DEVELOPING WORLD         | 1356| 104    | 845     | 178      | 783       | 783      |
| MANAGING DIRECTOR        | 1302| 266    | 216     | 392      | 220       | 220      |
| MANAGING EDITOR          | 1279| 115    | 268     | 805      | 71        | 71       |
| INCREASING NUMBER        | 1254| 76     | 332     | 309      | 537       | 537      |
| FOLLOWING DAY            | 1179| 87     | 309     | 312      | 185       | 246      |
| NURSING HOMES            | 1147| 105    | 4       | 377      | 323       | 488      |
| INTERESTING THINGS       | 1130| 93     | 33      | 62       | 37        | 6        |
| TURNING POINT            | 1093| 315    | 29      | 282      | 318       | 379      |

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.

Modal frequency by genre

| CONTEXT | ALL | SPOKEN | FICTION | MAGAZINE | NEWSPAPER | ACADEMIC |
|---------|-----|--------|---------|----------|-----------|----------|
| WOULD   | 105570| 296727 | 272262  | 173864   | 188137    | 153388   |
| CAN     | 984914| 239952 | 131811  | 248664   | 160535    | 205002   |
| WILL    | 858979| 213890 | 95222   | 186126   | 224317    | 179824   |
| COULD   | 707874| 137737 | 248489  | 121100   | 113880    | 88708    |
| MAY     | 663146| 69330  | 20835   | 90449    | 54627     | 133005   |
| SHOULD  | 394185| 88819  | 55889   | 67954    | 59066     | 81917    |
| LL      | 368470| 122092 | 108116  | 55984    | 36665     | 41211    |
| RIGHT   | 238737| 45338  | 18794   | 48152    | 39336     | 47137    |
| ‘D      | 190557| 45655  | 88269   | 30422    | 32323     | 3518     |
| MUST    | 189885| 21811  | 44546   | 35231    | 26832     | 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
Examining syntactic variation in English: The importance of corpus design …… 5
spoken.

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

| must [vv*]                        | have to |
|-----------------------------------|--------|
| SPOKEN FICTION MAGAZINE NEWSPAPER ACADEMIC | SPOKEN FICTION MAGAZINE NEWSPAPER ACADEMIC |
| 9920 14546 16199 15161 26256       | 10020 69612 50439 50791 21224 |
| 103.79 160.85 169.52 165.31 288.32 | 1,048.50 798.74 527.83 355.78 233.06 |

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 | get passive |
|------------|-------------|
| SPOKEN FICTION MAGAZINE NEWSPAPER ACADEMIC | SPOKEN FICTION MAGAZINE NEWSPAPER ACADEMIC |
| 84315 56415 113364 117426 246802       | 23651 16169 14120 13262 3218 |
| 882.16 623.86 1186.33 1280.35 2270.14 | 247.45 179.80 147.76 144.60 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:
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 five 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.
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.

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.
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 be 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
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 like, “I’m not going with her”) and the “so not” construction (I’m so not 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.

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.
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.

| search string | 1990-94 | 1995-99 | 2000-04 | 2005-09 | 2010-12 |
|---------------|---------|---------|---------|---------|---------|
| - split       | 17675   | 15981   | 16124   | 14999   | 7164    |
| + split       | 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](image1.png)

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](image2.png)

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.
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.*
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 | 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 | 0.79 | 0.79 | 0.82 | 0.81 | 0.83 | 0.84 | 0.85 | 0.86 | 0.86 | 0.87 | 0.87 | 0.89 | 0.89 |
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.

|      | 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  | 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 |
Examining syntactic variation in English: The importance of corpus design …… 17

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.

|      | 1870 | 1880 | 1890 | 1900 | 1910 | 1920 | 1930 | 1940 | 1950 | 1960 | 1970 | 1980 | 1990 | 2000 |
|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| from | 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  |
| %    | 0.00 | 0.00 | 0.00 | 0.01 | 0.01 | 0.02 | 0.03 | 0.06 | 0.06 | 0.08 | 0.10 | 0.17 | 0.20 |

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.
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
Examining syntactic variation in English: The importance of corpus design

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”**

![Overall frequency chart](image)

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**

![Forms chart](image)

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”

![Graph showing overall frequency of the construction “going to VERB”.](image)

**Figure 27** Forms of the construction “going to VERB” by decade

| DECADE | going to be | going to do | going to have | going to get | going to have | going to have | going to have | going to have | going to have | going to have | going to have | going to have |
|--------|-------------|-------------|--------------|-------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| 1970s  | 90,123      | 30,894      | 13,943       | 64,682      | 10,214       | 2,823        | 22,346       | 9,700        | 2,865        | 5,450        | 2,823        | 1,569        |
| 1980s  | 78,156      | 27,984      | 12,182       | 62,856      | 10,034       | 2,635        | 21,622       | 9,250        | 2,680        | 5,220        | 2,642        | 1,480        |
| 1990s  | 68,688      | 26,734      | 11,324       | 57,528      | 9,778        | 2,416        | 20,224       | 8,830        | 2,490        | 4,960        | 2,388        | 1,229        |
| 2000s  | 59,220      | 25,474      | 10,486       | 52,152      | 9,332        | 2,207        | 18,944       | 8,378        | 2,330        | 4,572        | 2,175        | 1,035        |

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”

![Graph showing overall frequency of the construction “get V-ed”.](image)
Examining syntactic variation in English: The importance of corpus design ……  21

<Figure 29> Forms of the construction “get V-ed” by decade

| DECADE | TOTAL |
|--------|-------|
| 1900  | 0     |
| 1920  | 0     |
| 1940  | 0     |
| 1960  | 0     |
| 1980  | 0     |

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

| DECADE | TOTAL |
|--------|-------|
| 1900  | 0     |
| 1920  | 0     |
| 1940  | 0     |
| 1960  | 0     |
| 1980  | 0     |

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*] [vvg*]” yields 30,200 tokens for 234 different strings.
<Figure 31> Forms of the construction “VERB NP into V-ing” by decade

| DECADES | CHARMS | TOTAL |
|---------|--------|-------|
| 1910    | 258    | 356   |
| 1920    |        | 1     |
| 1930    |        | 1     |
| 1940    |        | 1     |
| 1950    |        | 1     |
| 1960    |        | 1     |
| 1970    |        | 1     |
| 1980    |        | 1     |
| 1990    |        | 1     |

In the examples above, we searched for just one particular string (such as “[end] up [vv*]” 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 | 76,633 | 71,461 | 142,617 | 147,712 | 287,666 | 510,332 |
| % indic | 0.29 | 0.27 | 0.26 | 0.27 | 0.27 | 0.28 | 0.31 | 0.34 | 0.38 | 0.45 |
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.

|     | 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|
| %   | 0.26 | 0.29 | 0.36 | 0.39  | 0.39  | 0.44  | 0.59  | 0.64  | 0.77  | 0.85  |

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.

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](https://example.com/figure35.png)

Turning to an even less frequent construction, we find that there are only 31 tokens of the \[\text{end up } \text{V-ing}\] 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 \[\text{end up } \text{V-ing}\] 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).
Examining syntactic variation in English: The importance of corpus design

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 36: [end up V-ing] in COCA/BNC

| SECTION | ALL | FICTION | NON-FICTION | NEWSPAPER | NO-ACADEMIC | ACADEMIC | BNC | FICTION | NON-FICTION | NEWSPAPER | NO-ACADEMIC | ACADEMIC | 1980-1984 | 1985-1989 | 1990-1994 | 1995-1999 | 2000-2004 | 2005-2009 | 2010-2012 |
|----------|-----|---------|-------------|-----------|--------------|-----------|-----|---------|-------------|-----------|--------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| PRE-PRE | 52.54 | 48.79 | 34.17 | 27.95 | 9.35 | 14.77 | 26.66 | 26.34 | 16.85 | 37.00 |

| SECTION | ALL | FICTION | NON-FICTION | NEWSPAPER | NO-ACADEMIC | ACADEMIC | BNC | FICTION | NON-FICTION | NEWSPAPER | NO-ACADEMIC | ACADEMIC | 1980-1984 | 1985-1989 | 1990-1994 | 1995-1999 | 2000-2004 | 2005-2009 | 2010-2012 |
|----------|-----|---------|-------------|-----------|--------------|-----------|-----|---------|-------------|-----------|--------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| PRE-PRE | 10.98 | 37.55 | 0.74 | 23.82 | 17.05 | 7.95 | 3.72 | 9.31 |

### 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](image1.png) Quotative like construction in GloWbE

| Time   | US | GB | NL | FR | DE | IT | ES | PL | RU | JP |
|--------|----|----|----|----|----|----|----|----|----|----|
| 2012   | 0.9 | 0.8 | 0.7 | 0.6 | 0.5 | 0.4 | 0.3 | 0.2 | 0.1 | 0.0 |
| 2013   | 1.0 | 0.9 | 0.8 | 0.7 | 0.6 | 0.5 | 0.4 | 0.3 | 0.2 | 0.1 |

![Figure 39](image2.png) Concordance lines for “quotative like” in British portion of GloWbE

| Line | Source | Text |
|------|--------|------|
| 1    | GB     | I can make a lot of music on this keyboard. |
| 2    | GB     | And I can make a lot of music on this keyboard. |
| 3    | GB     | I can make a lot of music on this keyboard. |
| 4    | GB     | I can make a lot of music on this keyboard. |
| 5    | GB     | I can make a lot of music on this keyboard. |
| 6    | GB     | I can make a lot of music on this keyboard. |
| 7    | GB     | I can make a lot of music on this keyboard. |
| 8    | GB     | I can make a lot of music on this keyboard. |
| 9    | GB     | I can make a lot of music on this keyboard. |
| 10   | GB     | I can make a lot of music on this keyboard. |

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
Examining syntactic variation in English: The importance of corpus design ……  29

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.
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:
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
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.
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.

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:
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).

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.

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.
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,
Examining syntactic variation in English: The importance of corpus design ……

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 Intocausative”. 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, Geoff, 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.
Examining syntactic variation in English: The importance of corpus design …… 39

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. Stefanovitsch, 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
