The Importance of Robust Corpora in Providing More Realistic Descriptions of Variation in English Grammar

Abstract: This paper provides many 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. We also show – in great detail – how several recent corpora – such as COCA (2008), COHA (2011), GloWbE (2013), and the BYU interface to the Google Book n-grams (2012) – allow us to accurately examine this full range of variation, in ways that are not available with smaller corpora. As a result, 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.

Keywords: corpora, syntax, variation, historical, dialectal, genres

1 Introduction

Too many grammars of English 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 (e.g. journalistic prose from the US in the 1990s–2000s). The use of corpus data might help to remedy this situation, but all too often even corpus linguists base their conclusions on corpora that fail to adequately take into account a full range of variation in the language.

In this paper, I will consider how syntactic phenomena can vary as a function of language change, genre-based differences, and dialectal differences. Equally as important, I 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, 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 1,100 + 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. As important as Biber’s work has been, there are two limitations. First, the data came from a proprietary corpus, which is not publicly available for use by others. Second, the 40 million word Longman corpus that they used – while quite robust for the mid-1990s – would now be seen as being rather small.

The Corpus of Contemporary American English (COCA; see Davies 2009, 2011), which was released in 2008, solves these two problems. First, it is 450 million words in size (and continues to grow by 20 million words each year), which makes it more than ten times as large as the Longman corpus. Second, it is publicly-accessible, which means that teachers and students can easily replicate the investigations in Biber et al. and also carry out new studies, which can be verified by others. In this section, I will provide examples of a handful of such phenomena.
To begin with, we might look at the relative frequency of all modals in all genres of COCA (spoken, fiction, popular magazines, newspapers, and academic). Researchers have often looked at modals in much smaller 2–4 million word corpora, since they are one of the few syntactic phenomena that can be studied successfully with corpora that size. Expanding this search, we can see 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. Notice how must is more common in the more “formal” genres (cf. Leech 2003), whereas have to is more common in the informal genres, such as spoken.

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; see Hundt 2001). Two simple searches in COCA show us that the be passive is much more common in the formal genres, especially in academic, where explicit agents (e.g. in describing a chemistry experiment) would sound strange. The get passive, on the other hand, is most frequent in the informal genres (such as spoken).

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 data for the discourse marker you know or the like quotative (and I’m like, “no way”) show that they are much more common in the spoken transcripts in COCA.

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 construction that combines the 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. Even in a 100 million word corpus like the British National Corpus, there are only two tokens, and probably even less in a much smaller 40 million word corpus like the Longman Corpus.

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, COCA is perhaps the only corpus that will show such genre-based differences.

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). Due to limitations of space, just a small handful of examples will be given in each section.

3 Researching recent and ongoing syntactic changes with COCA

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) is large enough in size to look at a wide range of phenomena, (2) continues to be updated, and (3) that has a genre composition that is essentially the same from year to year. All are crucial to mapping recent syntactic shifts.

As far as recent syntactic shifts, 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 (e.g. “and he’s like, ...”; cf. Barbieri 2009) has had a sustained increase since the early 1990s (see the right side of the chart). Consider also the so not construction (e.g. I’m so not interested in him). Although the tokens for this construction are relatively sparse, we can still see a clear increase in the construction over the past two decades.
We might also consider three other syntactic shifts in contemporary American English, where the changes are probably much less salient. First, there is a clear increase in the “end up V-ing” construction (e.g. he ended up paying too much) during the past two decades. Second, the get passive (Bill got hired last week) has steadily increased over the past two decades (and there has been a corresponding decrease in the be passive as well). Third, there has been a 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) (cf. Mair 2002), as is seen in Table 1.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>+ to [help] [p*] to [v*]</td>
<td>825</td>
<td>798</td>
<td>726</td>
<td>668</td>
<td>370</td>
</tr>
<tr>
<td>− to [help] [p*] [v*]</td>
<td>5,494</td>
<td>6,453</td>
<td>7,144</td>
<td>7,502</td>
<td>4,237</td>
</tr>
<tr>
<td>% − to</td>
<td>86.9%</td>
<td>89.0%</td>
<td>90.8%</td>
<td>91.8%</td>
<td>92.0%</td>
</tr>
</tbody>
</table>

Finally, COCA can be used to look at recent changes with prescriptively-charged phenomena as well. For example, consider the split infinitive (to [verb] [Adv] > to [Adv] [Verb], e.g. to go boldly > 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 in Table 2, there is an increase in each five year block during the past two decades, for an increase of nearly 50% during this time.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>− split to [v*] <em>ly.[r</em>]</td>
<td>12,595</td>
<td>11,353</td>
<td>11,456</td>
<td>10,415</td>
<td>5,048</td>
</tr>
<tr>
<td>+ split to <em>ly.[r</em>] [v*]</td>
<td>7,681</td>
<td>8,907</td>
<td>9,955</td>
<td>10,819</td>
<td>6,353</td>
</tr>
<tr>
<td>% split</td>
<td>37.9%</td>
<td>44.0%</td>
<td>46.5%</td>
<td>51.0%</td>
<td>55.7%</td>
</tr>
</tbody>
</table>

These are just a handful of changes, and many more are discussed in Davies (2011). The important point is that COCA is probably the only corpus of English whose size and composition allow us to look at ongoing syntactic changes such as these.

4 Researching longer range syntactic changes with COHA

The Corpus of Historical American English (COHA) was released in 2010. It contains more than 400 million words of text 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 (see Davies 2012a, 2012b).

Carrying out research on diachronic syntax with COHA is both quick and easy. For example, we can easily see the increase in the need to V (I need to leave) or end up V-ing (I’ll end up getting there late) constructions. Notice the nice S-curve increase in both constructions in the last 40–50 years.

Even more complicated studies of diachronic syntax can be carried out quite easily with COHA. For example, Table 3 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: ([vm*] never|always [vv*]: he would never *answer his mail*). With the two quick and easy searches, we can clearly see in Table 3 and Figure 1 the sustained 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] in Table 4 (and Figure 2) represent 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).

<table>
<thead>
<tr>
<th>Table 4</th>
<th>Zero relative (the man – he saw)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A 1,666 1,680 1,756 2,025 2,044 2,032 1,734 1,462 1,513 1,389 1,278 1,122 904 968</td>
<td></td>
</tr>
<tr>
<td>B 4,929 5,117 6,126 7,778 8,453 8,921 9,681 10,931 9,937 9,088 9,079 9,069 8,239 8,664 7,696 7,617</td>
<td></td>
</tr>
<tr>
<td>% B 0.75 0.75 0.78 0.79 0.82 0.81 0.83 0.84 0.85 0.86 0.87 0.87 0.89 0.89 0.89</td>
<td></td>
</tr>
</tbody>
</table>

Of course we might want to change the relative clause subject, or experiment with different type of antecedents, or eliminate “false positives”, 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 above – in just a minute or so.

Small 2–4 million word historical corpora (like ARCHER or the Brown family of corpora) are limited primarily to looking at high frequency constructions, such as modals and auxiliaries. But because of its large size (100 times the size of these other corpora), COHA allows us to gain insight into a much wider range of syntactic changes in English.
5 Researching long-range syntactic changes with Google Books-Advanced

While COHA is composed of 400 million words of text, the Google Books n-grams are 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 et al. 2011) is extremely limited and simplistic, as far as syntactic searches go. 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 search for – one by one – the individual strings end up paying, ended up paying, ends up paying, and then start with tens of thousands of other embedded verbs – all of which would take weeks or months.

With the BYU/Advanced Google Books interface that we released in 2012, however, researchers can search by lemma and part of speech, and they could do a search like “end up V-ing” (which yields more than 400,000 tokens) in just 1–2 seconds. 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.

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 VERB (e.g. going to leave). We can easily search for “going to [v*]”, and we see the overall increase, as well as all of the matching strings. A further example is the “get passive” construction (e.g. got returned, get fired), which is definitely increasing over time (overall / individual forms). Again, with Google Books Standard, we would have to perform thousands of separate searches and perhaps spend days to get this data.

To take a pair of somewhat more complex constructions, consider first the “way construction”, which has been the focus of a great deal of research in construction grammar (see e.g. Goldberg 1997 for an introduction). In Google Books Advanced, we can simply search for “[vvd*] [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. Or consider the “causative into” construction – talked him into going, coerced them into buying, terrify me into doing, etc. (see Rudanko 2006; Wulff et al. 2007). The one simple search “[vv*] [p*] into [vvg*]” yields 30,200 tokens for 234 different strings.

As we have seen above, we can also carry out searches on two competing constructions to find evidence for syntactic change. For example, Table 5 (and Figure 3) 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; cf. Gonzalez-Alvarez 2003), and is based on 6,153,000 tokens from the 1810s–2000s. As can be seen in Table 5 and Figure 3, there is an increase in the use of the indicative since about the 1950s.

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). Two simple queries yield 3,812,000 tokens, which show a clear increase in the omission of the complementizer to (see Table 6 and Figure 4).

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] × [thousands of V₂] × [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 more than 6 million tokens of the “help (to) VERB” construction (Table 5, Figure 3), and the number of tokens for this one minor construction is more than three times as large as the total number of words in other often-used historical corpora such as ARCHER and the Brown family of corpora.

### 6 Researching dialectal variation in syntax in other World Englishes

Until recently, the largest corpus of English from several different dialects was the International Corpus of English (ICE), which is composed of one million words each from about thirteen English-speaking countries.
The corpus of **Global Web-based English** (GloWbE), which was released in 2013, is about 150 times as large as ICE, and it 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.

Note that nearly 60% of all of the texts in GloWbE (more than a billion words of text) come from blogs, which represent informal language quite nicely. Nevertheless, these are probably not as informal as the conversations in ICE, which model informal, spontaneous English extremely well.

To see how large corpora such as GloWbE can be used to look at dialectal variation in English, consider again the **“like construction”**, which has been discussed above. The 3,114 tokens in GloWbE show 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.

Another example deals with verbal subcategorization, with the “stop NP (from) V-ing” construction (e.g. stop them saying that, stop them from doing that). While the variant with *from* has roughly the same frequency in the different dialects, 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).

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, we can search for the **“way construction”** (e.g. made his way through the crowd, fought her way through the pitfalls), or the **“causative into”** construction (e.g. he talked her into staying, they forced me into admitting it).

Finally, with GloWbE we can also easily look at phenomena that straddle syntax and discourse. For example, we find that **“having said that...”** is relatively less common in American and Canadian English than the other Inner Circle dialects, whereas **“that said...”** is the most common in American and Canadian English.

In all of the cases discussed in this section, we have robust data from GloWbE – typically hundreds or even thousands of tokens. If we were using the ICE corpus, on the other hand, we would have many fewer tokens. The International Corpus of English is only about 1/150th the size of COCA (~13 million words in ICE, but nearly two billion words in COCA). So where we might have 2000 tokens in GloWbE, we might have just ten or fifteen total (for all countries) in ICE. Virtually none of the examples discussed in this section would be possible with any other corpus besides GloWbE.

### 7 Conclusion

In this paper, I 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.\(^1\) 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.

---

\(^1\) One type of variation that we have not discussed in this paper is variation at the level of the individual speaker, e.g. as a function of ethnicity, gender, or age. To do so, we would need a corpus that has coded all of this information for each individual speaker, in thousands or even millions of texts. The creation of such a corpus would obviously be extremely time intensive and expensive. The only large corpus that has done this (even partially) is the British National Corpus, which was generously funded by Oxford University Press, and which was created more than 20 years ago. Due to the costs involved, nothing similar has been done since, nor is it likely to be.
As we have seen, several recent corpora – such as COCA (2008), COHA (2011), GloWbE (2013), and the BYU interface to the Google Book n-grams (2012) – allow us to accurately examine this full range of variation. As we have seen, 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 often provides 100–200 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