FOCUS ARTICLE AND RESPONSES

Expanding horizons in the study of World Englishes with the 1.9 billion word Global Web-based English Corpus (GloWbE)

Mark Davies and Robert Fuchs
Brigham Young University and Münster University

In this paper, we provide an overview of the new GloWbE Corpus — the Corpus of **Glo**bal **W**eb-**b**ased **E**nglish. GloWbE is based on 1.9 billion words in 1.8 million web pages from 20 different English-speaking countries. Approximately 60 percent of the corpus comes from informal blogs, and the rest from a wide range of other genres and text types. Because of its large size, its architecture and interface, the corpus can be used to examine many types of variation among dialects, which might not be possible with other corpora — including variation in lexis, morphology, (medium- and low-frequency) syntactic constructions, variation in meaning, as well as discourse and its relationship to culture.

Keywords: English, Englishes, global, world, dialects, corpus, corpora, varieties of English

1. Introduction

One of the most important challenges facing researchers of World Englishes is the question of where to find raw data from speakers of these dialects. Possible data sources may include collections of newspapers, blogs, emails, SMS texts, transcripts from recorded conversations, or fictional literature. Studies based on each of these approaches are found in *English World-Wide* during the past four or five years.

Another possibility is to use "structured corpora". An important set of corpora for the study of World Englishes is the extended Brown family of corpora, which includes the Brown Corpus of 1960s American English (Francis 1964) and other parallel corpora of varieties and time points such as 1990s American English, 1960s and 1990s British English, as well as Australian English, New Zealand

English World-Wide 36:1 (2015), **1–27**. DOI 10.1075/eww.36.1.01dav
ISSN 0172–8865 / E-ISSN 1569–9730

English and Indian English (Bauer 1993; Hundt, Sand and Siemund 1998; Hundt, Sand and Skandera 1999; Johansson 1980; Peters 1987; Shastri 1988). Each of these individual corpora contains about one million words of text.

However, the most widely used corpus for research on World Englishes may be the International Corpus of English (ICE) (see Greenbaum 1996). The ICE components are composed of roughly one million words each (600,000 spoken and 400,000 written), and they currently provide data on 13 national varieties of English, including Great Britain, Ireland, Canada, New Zealand, Hong Kong, East Africa, India, Singapore, the Philippines, and Jamaica, as well as the USA, Nigeria, and Sri Lanka (just the written portion for these last three countries).

As noted, the ICE corpora have been very important for our understanding of World Englishes, as measured by the number of studies that have been based on these corpora. Nevertheless, as valuable as these corpora are, one important limitation is the size of the individual components in ICE. The majority of ICE is composed of transcripts of spoken language, which is extremely difficult and time-consuming to collect. Because the individual corpora have just one million words each, their primary usefulness is probably that they provide the possibility of looking at relatively high-frequency syntactic constructions, where even just a million words might yield enough data. On the other hand, they sometimes do not provide enough data for in-depth research on lexical variation, morphological variation, variation with medium- and lower-frequency syntactic constructions, or differences in word meaning between dialects.

Because of the limitations of smaller corpora, some researchers have created their own proprietary, ad-hoc corpora, in order to study phenomena that need to be based on much larger collections of data. For example, Hundt, Hoffmann and Mukherjee (2012) investigated the use of the hypothetical subjunctive (e.g. *as if he* {*was*/*were*} *king*), where a few million words of data would not have been nearly enough. To collect the needed data, they created a corpus of 146 million words of text from South Asian newspapers (and then compared this to newspaper data from the British National Corpus [BNC]), which provided extremely useful and insightful data for this construction. The downside of such proprietary corpora, however, is precisely the fact that they are proprietary. They are created by individual researchers for use on selected topics, but often are not available to a much wider range of researchers of World Englishes.

Recognizing the need to create a very large corpus of World Englishes, which would be available to a wide range of researchers, we recently collected the GloWbE Corpus (**Glo**bal **W**eb-**b**ased **E**nglish Corpus). This corpus is based on 1.9 billion words of text from 20 different countries, which includes six Inner Circle and 14 Outer Circle countries (on the distinction between Inner and Outer

Circle, see Kachru 1985).[1] The texts in the corpus consist of informal blogs (about 60 percent of the corpus) and other web-based materials, such as newspapers, magazines, company websites, and so on. As with the other corpora from corpus.byu.edu, GloWbE is freely available to all researchers at <http://corpus2.byu.edu/glowbe>.

In this paper we provide a number of concrete examples of how GloWbE allows researchers to carry out a wide range of studies on lexical, phraseological, morphological, syntactic, and semantic variation among dialects of English, many of which could probably not be studied with other, smaller corpora. Due to limitations of space, we will provide only a very brief discussion of many phenomena that have been studied in much more detail elsewhere. As a result, there are many aspects of these different phenomena that cannot and will not be exhaustively considered in this paper. But hopefully, this overview of the GloWbE Corpus will show how it can be an important part of researchers' "toolbox" of resources for studying World Englishes, along with other corpora such as ICE and the Brown-inspired corpora.

2. Designing and creating the GloWbE corpus

There were three goals in the creation of GloWbE: size, genre balance (including informal language), and accuracy in terms of identifying the dialect that it is representing. We will consider each of these goals as we discuss the design and creation of the corpus in this section.

In terms of size, the goal in creating GloWbE was to have a corpus that was large enough to permit research on a wide range of phenomena in World Englishes. To this end, there was really only one possible source for the texts, and that was web pages. Virtually all corpora that are larger than about 500 million words in size are based largely (or exclusively) on web pages. For example, this is the approach used for all of the large corpora from <www.sketchengine.co.uk>. However, as useful as the Sketch Engine corpora are, none of them allow for comparisons between different dialects of English.

But we also wanted to ensure that the web pages represented informal language fairly well. Recall that with the ICE corpora, about 60 percent of the total number of words for each country comes from transcriptions of spoken language,

1. In future updates to the corpus, other countries may be added as well. These may include countries like Malta, Cyprus, Cameroon, Burma, Trinidad and Tobago, and the Bahamas — all of which are former British colonies.

and the other 40 percent consists of more formal, written texts.[2] In the creation of GloWbE, we followed roughly the same approach. About 60 percent of the words for each country come from informal blogs, whereas the other 40 percent come from a wide variety of (often) more formal genres and text types.

The first task in creating the corpus was to get the URLs for millions of web pages from the 20 different countries. In order to do so, we ran hundreds of very high frequency 3-grams (three word strings) in the Corpus of Contemporary American English (COCA) against Google — phrases such as {*and from the*}, {*but if it*}, {*and they are*}, etc. Because of the high frequency of the search strings and because Google does not use search engine optimization criteria for phrases like *and from the*, it ends up listing essentially random URLs, which is precisely what we wanted. We stored these URLs in a database, along with all associated metadata (web site, country, page title, etc.).

In order to achieve a roughly 60/40 mix of informal and somewhat more formal language, we first collected one million URLs from a "general" search in Google, and then another million URLs from Google searches of just blogs. In the general search, however, about 20 percent of these were also blogs (there is no way to exclude them from "general" searches), which results in (roughly) a 60/40 mix overall.

The most challenging part of the corpus creation was ensuring that the web pages were correctly associated with each of the 20 countries in the corpus. To do so, we carried out each of these two sets of searches ("general" and blogs) for each country separately, using Google "Advanced Search", and limiting by "Region" (as Google calls it) — Canada, Ireland, India, Singapore, etc.

The question, of course, is how well Google has correctly identified web sites by country. For web sites with a top-level country domain (".LK" = Sri Lanka, ".SG" = Singapore, etc.) this is quite straightforward for Google. But in the case of ".com" and ".org" web sites, for example, it is much more difficult. In these cases, Google relies on several heuristics, including 1) the IP address for the web server, 2) who links to that website, and 3) who visits the website.[3]

For example, imagine a website <http://www.somesite.com>. The IP address suggests that the server is located in or near Singapore, Google has identified that 95 percent of the visitors to this website come from Singapore and also that 93 percent of the links to this website are from other known websites from Singapore (e.g. those ending in ".SG"). As a result, it is fairly safe for Google to assume that

2. The spoken part of ICE contains some formal (scripted) data and the written section includes informal writing (e.g. personal letters). In a sense, then, perhaps the best distinction is "medium", rather than "formality".

3. See <https://support.google.com/webmasters/answer/62399?hl=en>.

this website is from Singapore, in spite of the ".com" address. This approach may not be perfect, but it is very good. In the year since we created the corpus (for the first six months it was only available for internal testing), and after checking hundreds of websites to see where they are actually located (based for example on "info" pages on the sites), we have yet to find a single website whose country has not been correctly identified by Google.

After creating the list of URLs, we used HTTrack to download the two million web pages, and we then used JusText to remove "boilerplate" material from the web pages — recurring headers, footers, sidebars, and so on.[4] After this, we used the CLAWS 7 tagger[5] to tag the entire corpus. Finally, we imported the texts into a database, where they would use the same architecture and interface as the other corpora from <corpus.byu.edu>.

The end result was a 1.9 billion word corpus from about 1.8 million web pages in 20 different countries, as shown in Table 1:

Note that the United States (US) and Great Britain (GB) have the largest size (both about 386 million words), all six Inner Circle countries (as well as India) have at least 80 million words of text each, and nearly all of the 20 countries have at least 40 million words of text each (Ghana has 39 million and Tanzania 35 million). As a result of the sampling process, all subcorpora constitute representative samples of how these national varieties of English are used in web-based communication.

As has already been mentioned, GloWbE uses the same architecture and interface as the other corpora from <corpus.byu.edu>. One of the strengths of this architecture and interface is that it allows users to carry out useful comparisons of the different sections of the corpus. In other corpora such as the BNC or the COCA (see Davies 2009, 2011), we can compare and contrast different genres, or, as with the Corpus of Historical American English (COHA; see Davies 2012), we can compare different historical periods. In the case of GloWbE, of course, what we are comparing are the different national varieties. Although we will see many other examples throughout this paper, at this point we will give one quick example of such comparisons.

Suppose that we are looking at the construction "[freak] [p*] out" (forms of *freak* + pronoun + *out*, e.g. *freaked me out*). First, by choosing the "Chart" option under "Display" and entering the search query, we can visualize the data in terms of overall frequency (Figure 1). This shows that the construction is much less com-

4. See <https://code.google.com/p/justext/>.

5. See <http://ucrel.lancs.ac.uk/claws/>.

Table 1. Size of GloWbE Corpus by country

Country	Code	Web sites	Web pages	Words
United States	US	82,260	275,156	386,809,355
Canada	CA	33,776	135,692	134,765,381
Great Britain	GB	64,351	381,841	387,615,074
Ireland	IE	15,840	102,147	101,029,231
Australia	AU	28,881	129,244	148,208,169
New Zealand	NZ	14,053	82,679	81,390,476
India	IN	18,618	113,765	96,430,888
Sri Lanka	LK	4,208	38,389	46,583,115
Pakistan	PK	4,955	42,769	51,367,152
Bangladesh	BD	5,712	45,059	39,658,255
Singapore	SG	8,339	45,459	42,974,705
Malaysia	MY	8,966	45,601	42,420,168
Philippines	PH	10,224	46,342	43,250,093
Hong Kong	HK	8,740	43,936	40,450,291
South Africa	ZA	10,308	45,264	45,364,498
Nigeria	NG	4,516	37,285	42,646,098
Ghana	GH	3,616	47,351	38,768,231
Kenya	KE	5,193	45,962	41,069,085
Tanzania	TZ	4,575	41,356	35,169,042
Jamaica	JM	3,488	46,748	39,663,666
TOTAL		340,619	1,792,045	1,885,632,973

mon in the South Asian varieties (IN, LK, PK, BD: India, Sri Lanka, Pakistan, Bangladesh) than in the Inner Circle countries.

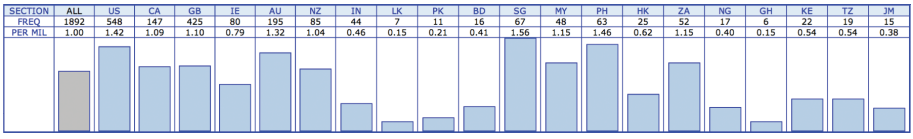


Figure 1. Overall frequency of *freak out*, by country

In addition to seeing overall frequency, it is also always possible to see the frequency of each individual matching form in each country, as shown in Figure 2. This table view again shows the relatively low frequency of this phrasal verb in the South Asian dialects.

	CONTEXT	ALL	US	CA	GB	IE	AU	NZ	IN	LK	PK	BD	SG	MY	PH	HK	ZA	NG	GH	KE	TZ	JM
1	FREAKED ME OUT	499	132	33	122	29	45	20	6	3	4	4	26	16	19	4	16	4	2	5	4	5
2	FREAKS ME OUT	351	113	30	57	11	35	18	12		3	5	10	13	16	7	7	1	2	3	6	2
3	FREAK ME OUT	266	78	19	73	11	26	9	3	2	1	1	8	6	8	1	10	3		5	2	
4	FREAKING ME OUT	154	39	10	34	5	16	10	3			1	9		7	3	6	1		2	2	6
5	FREAK YOU OUT	119	27	16	25	2	17	5	6		1	1	3	1	2	5	4	2	1	1		
6	FREAK THEM OUT	48	8	8	14	5	8		2						1		1	1				
7	FREAKED HIM OUT	44	18	7	7		4	2	1				2		1						2	
8	FREAKS YOU OUT	43	20	5	7	3	2	1	1	1				1	1					1		
9	FREAK HIM OUT	31	12	1	7	1	5	1						1	1		2					

Figure 2. Frequency of different forms of *freak out*, by country

In addition to these two main types of visualization, we can also directly compare two different countries or sets of countries — for example, what words occur much more in Ireland than in other countries, or in the South Asian dialects compared to Great Britain. We will see several examples of this in the sections that follow.

Finally, note that for any frequency chart (Figure 1) or table (Figure 2) display, users can click on the desired word and/or country to see the "Keyword in Context" (KWIC) entries (Figure 3).

CLICK FOR MORE CONTEXT	[?]	SAVE LIST	CHOOSE LIST	----------	CREATE NEW LIST	[?]
1	CA B	...ocalledmommylife.com	A	B	C	having someone (even though a trained specialist) manipulate my newborns body. Totally **freaked me out**. # That night after all of our visitors left around 8 so
2	IN B	...sandme.wordpress.com	A	B	C	, he even asked me to put the handle lock on just incase. that kinda **freaked me out**, and i woke up at 4am to check if my bike was
3	US G	jillwillrun.com	A	B	C	I forgot to warn you about the swelling. My doctor never mentioned and it **freaked me out**. Did they show you how to hold a pillow over your incision
4	GB G	helengrantbooks.com	A	B	C	# Well, I didn't stick around after that. The whole thing had **freaked me out** far too much. Somewhat later in the evening it occurred to me
5	PK G	tokyo.metblogs.com	A	B	C	's a need for them. In fact, the hushed silence in the trains **freaked me out** a little but I come from Southeast Asia and we've got a
6	GB G	anxietynomore.co.uk	A	B	C	thoughts be there about the hair loss. The first time it happened it really **freaked me out** but I am now more accepting. I a currently undergoing a thinning
7	NZ B	...reatfun4kidsblog.com	A	B	C	. I think Someone was looking after me, knowing how that would have totally **freaked me out**, don't you?
8	GB G	guardian.co.uk	A	B	C	, " Wow, this band are the greatest new band ever. " It **freaked me out** so much that I put it aside for a long time; it
9	GH G	ghanacelebrities.com	A	B	C	thinking of going for a refund of her booked ticket... lol! I kinda **freaked me out** when I saw people giving out flyers and advertisements all over the place
10	CA G	...banlegends.about.com	A	B	C	said to be a 5 year old boy in Surrey ND that got taken kinda **freaked me out** but thought i would look it up and see if its true or
11	IE B	weirdism.com	A	B	C	horrible, utter pain, and a moan... it was just disgusting. It **freaked me out**, so I simply dropped the drawer. My plan was to then
12	PH B	the.rainbowholic.me	A	B	C	, we had my tummy check and it was because of gallstones! It really **freaked me out** because I'm actually a healthy eater (now).: O

Figure 3. Keyword in Context (KWIC) display

They can also click on a KWIC line to get even more context — up to about a paragraph — as well as a link to the original web page, to see the complete context. In this paper none of the other 50+ figures show the KWIC lines, but it should be understood that they are fully accessible to users of the GloWbE corpus.

In the following sections, then, we will consider many different phenomena that show how GloWbE can be used to look at variation between dialects for a wide range of phenomena — lexis and phraseology (Section 3), morphology (Section 4), syntax (Section 5), semantics (Section 6), and discourse and culture (Section 7). As we mentioned previously, due to limitations of space, we can only dedicate a paragraph or two to each phenomenon. Because of this, the focus in these sections is on the overall range of possibilities for research with the corpus, and not an in-depth focus on new insights for any of the phenomena themselves.

3. Lexical variation

Previous researchers have noted that lexical frequency is very sensitive to corpus size. For example, Baker (2011: 70) notes that in the Brown family of corpora (one million words each), there might be a few hundred very high frequency words

with enough tokens to compare across dialects (e.g. *class*, *miss*, *black*, *true*, etc.), but such comparisons would be impossible for the vast majority of words. Other comparatively small corpora like ICE would also be similarly limited.

Because GloWbE is nearly two billion words in size, it is large enough to provide comparisons for many more words. For example, there are more than 100,000 distinct lemmas (101,559) that occur more than 100 times each in GloWbE, and nearly 200,000 (197,270) that occur 25 times each. In this section we will provide a few examples of the rich data on lexical variation in varieties of English which GloWbE provides.

We will begin with a fairly trivial example, simply to show that the GloWbE data does match what our intuition tells us should be happening, and then move on to some more interesting data. First, consider *fortnight* (Figure 4), which is much more frequent in British English (BrE) than in American English (AmE).

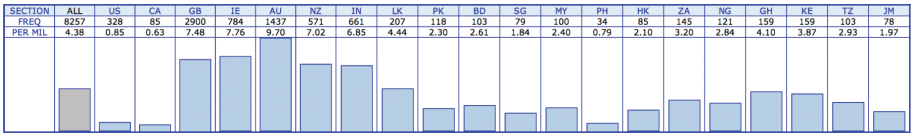


Figure 4. *Fortnight*

Of course, with GloWbE we can compare the frequency of any word in any country, compared to the other nineteen countries in the corpus. For example, *banjaxed* ('ruined, screwed up') is by far the most common in Irish English (Figure 5).

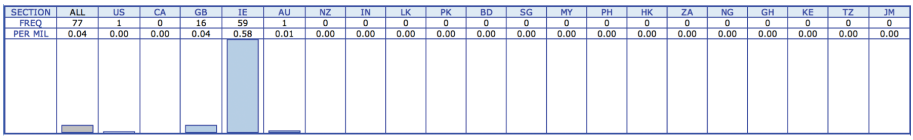


Figure 5. *banjaxed*

Turning to Outer Circle countries, we can see the frequency of *hand phone* ('mobile / cell phone') in Malaysian English (Figure 6).

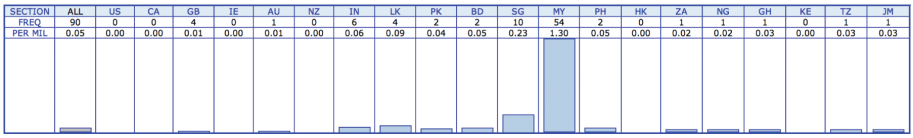
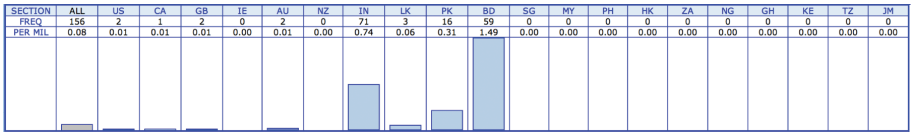


Figure 6. *hand phone*

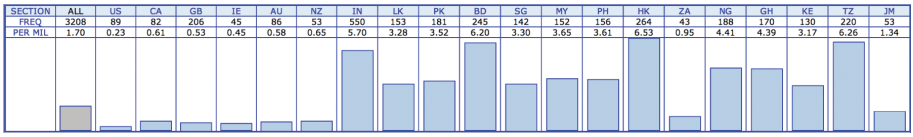
There are of course also lexical items that are limited not just to one country, but rather to a particular region. For example, *Eve teas** ('public sexual harassment') is limited primarily to South Asia (IN, LK, PK, BD) (Figure 7).



SECTION	ALL	US	CA	GB	IE	AU	NZ	IN	LK	PK	BD	SG	MY	PH	HK	ZA	NG	GH	KE	TZ	JM
FREQ	156	2	1	2	0	2	0	71	3	16	59	0	0	0	0	0	0	0	0	0	0
PER MIL	0.08	0.01	0.01	0.01	0.00	0.01	0.00	0.74	0.06	0.31	1.49	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Figure 7. *Eve teas**

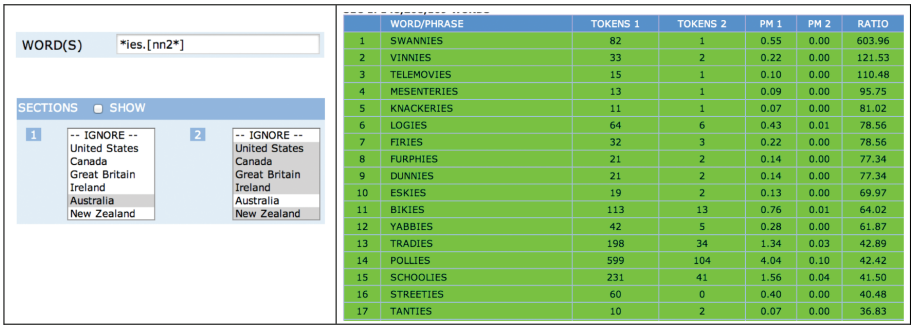
Perhaps even more interesting are those words that are limited mainly to non-Inner Circle dialects, such as *equipments* (Figure 8), which is extremely rare in the Inner Circle countries (Figure 8).



SECTION	ALL	US	CA	GB	IE	AU	NZ	IN	LK	PK	BD	SG	MY	PH	HK	ZA	NG	GH	KE	TZ	JM
FREQ	3208	89	82	206	45	86	53	550	153	181	245	142	152	156	264	43	188	170	130	220	53
PER MIL	1.70	0.23	0.61	0.53	0.45	0.58	0.65	5.70	3.28	3.52	6.20	3.30	3.65	3.61	6.53	0.95	4.41	4.39	3.17	6.26	1.34

Figure 8. *equipments*

In Figures 4–8, we searched for a particular word that we expected to be more frequent in a particular country or group of countries. But one of the real strengths of GloWbE is that it can quickly search through the database to find all words that are more frequent in one country compared with another. For example, via the web-based interface, we can search for **ies* plural nouns (*ies.[nn2*]) in Australian English (AusE; AU in the examples provided) that are not common in other Inner Circle dialects (US, CA, GB, IE, and NZ): United States, Canada, Great Britain, Ireland, and New Zealand, and the resulting list of words would be the one shown in Figure 9.



	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	SWANNIES	82	1	0.55	0.00	603.96
2	VINNIES	33	2	0.22	0.00	121.53
3	TELEMOVIES	15	1	0.10	0.00	110.48
4	MESENTERIES	13	1	0.09	0.00	95.75
5	KNACKERIES	11	1	0.07	0.00	81.02
6	LOGIES	64	6	0.43	0.01	78.56
7	FIRIES	32	3	0.22	0.00	78.56
8	FURPHIES	21	2	0.14	0.00	77.34
9	DUNNIES	21	2	0.14	0.00	77.34
10	ESKIES	19	2	0.13	0.00	69.97
11	BIKIES	113	13	0.76	0.01	64.02
12	YABBIES	42	5	0.28	0.00	61.87
13	TRADIES	198	34	1.34	0.03	42.89
14	POLLIES	599	104	4.04	0.10	42.42
15	SCHOOLIES	231	41	1.56	0.04	41.50
16	STREETIES	60	0	0.40	0.00	40.48
17	TANTIES	10	2	0.07	0.00	36.83

Figure 9. **ies* nouns in Australian English

Note that not all of the results are examples of the AusE *-ies* diminutive (e.g. *swannies*, *telemovies*, *mesenteries*), but the majority are: *vinnies* ('wine stores'), *firies* ('fire fighters'), *furphies* ('rumors'), *dunnies* ('toilets'), *eskies* ('coolers'), *bikies* ('bikers'), *tradies* ('tradesmen'), *pollies* ('politicians'), *schoolies* ('schoolchildren' or 'breaks from school'), *streeties* ('homeless people'), and *tanties* ('tantrums').

Of course we can also look for phrases, and not just individual words. For example, we can compare phrases like [*be*] *different to* (much more common in GB, IE, AU, and NZ than in US or CA; Figure 10) and [*be*] *different than* (more common in US and CA; Figure 11).

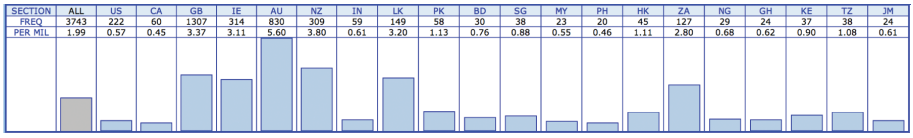


Figure 10. *be different to*

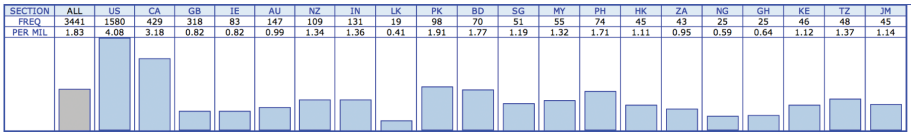


Figure 11. *be different than*

Other examples of differences in phraseology are [*keep*] *in view* in South Asian English (especially in the varieties found in India and Pakistan) (Figure 12), and [*discuss*] *about* (Figure 13) in the Outer Circle dialects.[6]

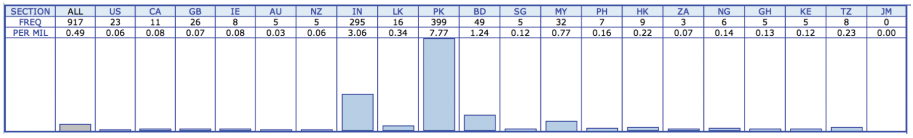


Figure 12. [*keep*] *in view*

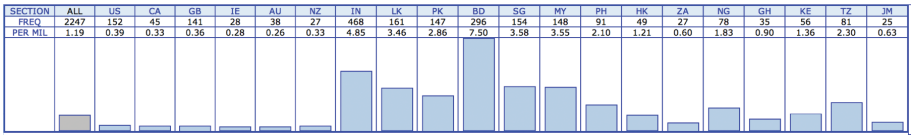


Figure 13. [*discuss*] *about*

4. Morphological variation

GloWbE can also be used to examine morphological variation among the different dialects of English. To take a fairly obvious and perhaps trivial example, we can search for the frequency of *had* + { *gotten* / *got* }, as shown in Figure 14, which is

6. Note that in the GloWbE search syntax, [j*] = [ADJ], [nn*] = (common) [NOUN], [p*] = [PRON], and [vv*] = (lexical) VERB.

based on 13,273 tokens. As we can see, the percentage of all tokens that use *gotten* (as opposed to *got*) is three to four times as high in the United States and Canada as it is in Great Britain, which has the lowest percentage of *gotten* of all of the Inner Circle dialects.

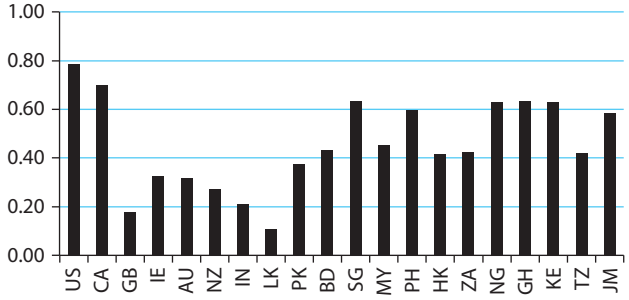


Figure 14. *had* + {***gotten*** / *got*}

Another case of competing past participles is [*have*] + {*proved*/*proven*}. There are 29,683 tokens in GloWbE, and they show that *proven* is much higher in the United States and Canada than in the other Inner Circle dialects (Figure 15). GB, on the other hand, is the Inner Circle dialect that most strongly prefers *proved*.

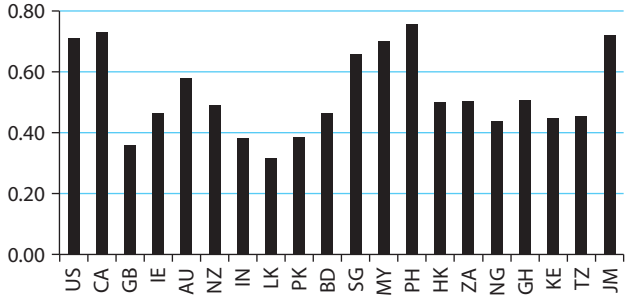


Figure 15. [*have*] + {***proven*** / *proved*]

GloWbE also shows a huge contrast between the US and CA and most of the other dialects with regards to *dove* versus *dived* (see Chambers 1998). There are 1,124 tokens of [pronoun] + {*dove*/*dived*} (e.g. *he dove into the pool*, *they dived into their homework*). Both US and CA use *dove* in about 77 percent of all cases, whereas GB is the dialect that uses *dove* the least, at only 18 percent (Figure 16).

Due to limitations of space in this paper, we have provided just a few examples of how GloWbE can be used to look at morphological differences between varieties of English, and of course many other analyses of morphological variation can be carried out using the corpus.

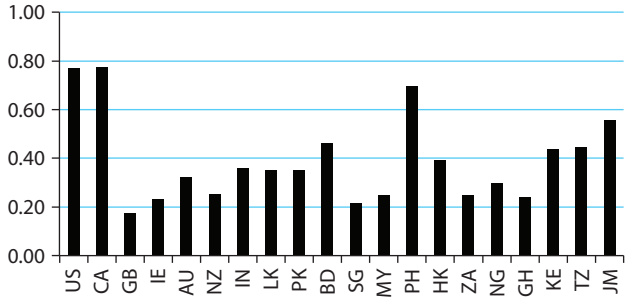


Figure 16. [PRON] + {***dove*** / *dived*}

5. Syntactic variation

In this section we will consider a number of different examples of how GloWbE can be used to carry out investigations of dialectal variation in syntax. We will start with some fairly simple examples, and then progress to more detailed phenomena, which relate to previous research on syntactic variation in English.

To begin with, consider the use of *likely* occurring between verbs (e.g. *they will **likely** have better careers*), which is discussed in Lindquist (2009: 209–271). Based on data from COCA and the BNC, Lindquist (2009) suggests that the construction is more frequent in AmE than in BrE, and this is clearly supported by the 36,703 tokens of the construction in GloWbE. The data can be visualized either by overall frequency (Figure 17) or by individual string (Figure 18).

SECTION	ALL	US	CA	GB	IE	AU	NZ	IN	LK	PK	BD	SG	MY	PH	HK	ZA	NG	GH	KE	TZ	JM
FREQ	36703	12265	5473	5112	1106	2129	1371	1273	416	611	661	798	713	726	804	552	622	442	597	508	524
PER MIL	19.48	31.71	40.61	13.19	10.95	14.36	16.84	13.20	8.93	11.89	16.74	18.57	17.12	16.79	19.88	12.17	14.59	11.40	14.54	14.45	13.24

Figure 17. [VERB] *likely* [VERB], overall by country

	CONTEXT	ALL	US	CA	GB	IE	AU	NZ	IN	LK	PK	BD	SG	MY	PH	HK	ZA	NG	GH	KE	TZ	JM
1	WILL LIKELY BE	6395	1861	837	913	171	334	228	316	87	139	173	147	138	170	205	112	83	81	118	157	125
2	WOULD LIKELY BE	2142	723	329	328	75	135	94	51	23	33	24	46	49	39	33	22	33	17	33	15	40
3	WILL LIKELY HAVE	1124	377	213	143	37	58	45	48	6	10	20	34	17	12	20	19	14	10	17	13	11
4	WOULD LIKELY HAVE	1069	368	168	178	44	77	58	25	12	16	13	14	16	7	8	10	9	9	11	7	19
5	'LL LIKELY BE	431	162	64	91	7	24	18	6	4	4	5	8	7	5	5	2	5	5	2	4	3
6	WILL LIKELY CONTINUE	396	142	53	51	11	20	11	13	5	7	4	11	5	2	16	10	6	4	13	7	5
7	WILL LIKELY GET	378	128	70	45	15	13	11	12	3	6	6	10	4	11	8	10	8	4	7	5	2

Figure 18. [VERB] *likely* [VERB], by form and by country

A somewhat "broader" construction is the "*way* construction", which has been widely studied from within the Construction Grammar model (see Goldberg 1997). The simple search for "[vv*]" followed within one or two words by "[ap*] way [i*] the [nn*]" will find all cases of strings like *made his way down the corridor*, *worked her way into the conversation*, *fought his way through the crowd*, and

so on. Within just a few seconds, users can search the 1.9 billion words to find the frequency of all 18,525 matching strings, and see all matching verbs and their frequency in each dialect (Figure 19).

		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

Figure 19. "*way* construction", by verb and by country

A quick glance at Figure 20 suggests that the "*way* construction" may be less frequent in the four South Asian varieties (IN, LK, PK, and BD), and this is confirmed by looking at the overall frequency in these dialects.

SECTION	ALL	US	CA	GB	IE	AU	NZ	IN	LK	PK	BD	SG	MY	PH	HK	ZA	NG	GH	KE	TZ	JM
FREQ	18525	3422	1481	4486	1130	1532	864	733	350	366	244	425	367	396	329	467	288	368	389	414	474
PER MIL	9.83	8.85	10.99	11.57	11.18	10.34	10.62	7.60	7.51	7.13	6.18	9.89	8.81	9.16	8.13	10.29	6.75	9.49	9.47	11.77	11.98

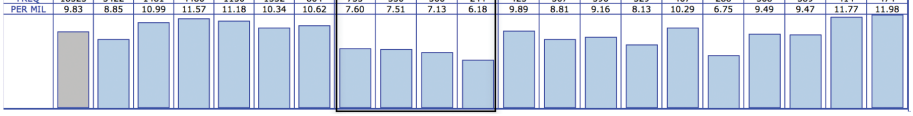


Figure 20. "*way* construction", overall by country

One of the most widely studied constructions during the past two decades or so has been the "quotative *like*" construction. A consistent theme in most of the recent research is that the construction has now clearly spread beyond being just an American phenomenon, and that it is now found in many other dialects of English (e.g. Tagliamonte and D'Arcy 2004 for Canadian English [CA]; Buchstaller 2006 for BrE; Rodríguez-Louro 2013 for AusE; and D'Arcy 2012 for New Zealand English).

The GloWbE corpus provides very interesting data on the distribution of the construction in blogs and other web pages from the different dialects. The 3,114 tokens of the construction in GloWbE ([c*] [p*] [be] like,|'; e.g. *and I was like*,) show that while the construction is still the most common in AmE, it is (in stair-step fashion) progressively less frequent in Canada, Great Britain, Ireland, Australia, and New Zealand, and that it is the least frequent in the South Asian dialects (Figure 21).

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	1.65	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

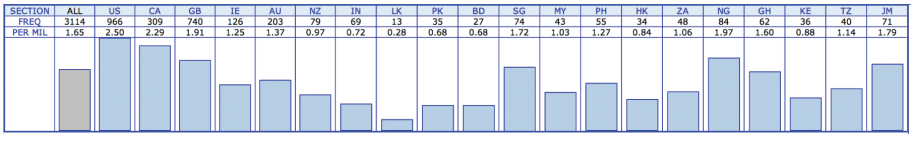


Figure 21. "quotative *like*" construction

Of course we can also use the corpus to look at syntactic phenomena where there is a strong prescriptive norm, and see how this plays out in the different dialects. For example, in AmE, *try and VERB* is quite stigmatized (e.g. *he'll try and talk to her tomorrow*) (see Hommerberg and Tottie 2007). GloWbE has 68,557 tokens of *try and*, and it clearly is less frequent (Figure 22) in the US (and Canada) than in the other Inner Circle dialects.

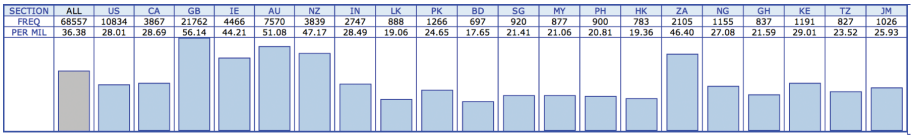


Figure 22. *try and* [VERB]

In all of the cases of syntactic variation discussed to this point, one simple search in GloWbE was able to provide the needed data. In many cases, however, we will want to combine two successive searches in GloWbE to see the relative frequency of two constructions.

For example, consider verbal complementation with *stop* (see Rudanko 2002: Chapter 4). Figure 23 shows the frequency without *from* (*they stopped him Ø leaving*) and a similar chart can be produced for the construction with *from* as well (*they stopped him* ***from*** *leaving*). Here the 21,455 tokens show that the construction without *from* is much less common in American and Canadian English.

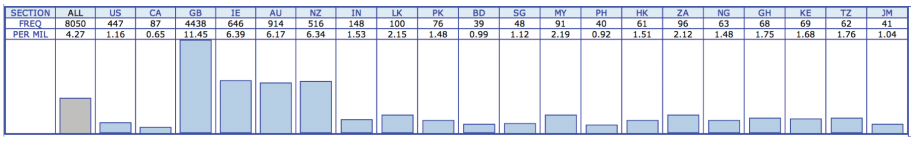


Figure 23. [*stop*] + [PRON] + [v-ing]

But we can also input the data from these two charts into a spreadsheet to see the percentage of all complements that lack *from*. Here we see perhaps even more clearly the relative absence of the *–from* construction in American and Canadian English (Figure 24).

Another example of dialectal variation is the (prescriptively) non-standard singular (SG) instead of the plural (PL) in cases like *there{'s/ are} some people next door*. Collins (2012) looks at this construction in eight different ICE corpora and suggests that speakers of Outer Circle varieties are much more reluctant to use the non-standard singular form with plural nominal subjects (e.g. *there's* (SG) *some people* (PL) *next door*), whereas this is not as much of an issue for speakers of the Inner Circle dialects. The data from the 153,916 tokens in GloWbE support this claim. Table 2 shows the number of tokens with the non-standard singular (e.g.

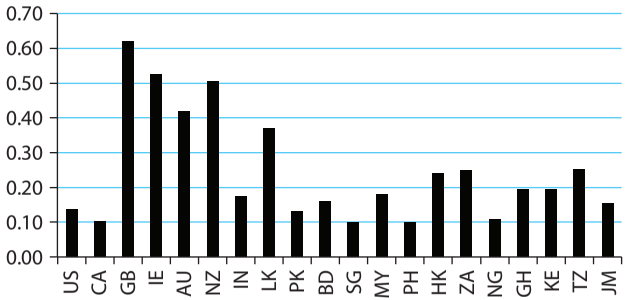


Figure 24. [*stop*] + [PRON] + [v-ing] (percentage without *from*)

Table 2. SG/PL agreement with existentials (*there is/are*)

	US	CA	GB	IE	AU	NZ	IN	LK	PK	BD	SG	MY	PH	HK	ZA	NG	GH	KE	TZ	JM
SG	1018	377	1230	383	582	237	89	23	39	32	93	58	66	55	58	28	37	35	25	43
PL	29401	10514	31220	7558	12432	6515	8499	4100	3970	2996	3405	3268	3757	3052	3257	3567	2866	3293	2583	3155
%SG	3.3	3.5	3.8	4.8	4.5	3.5	1.0	0.6	1.0	1.1	2.7	1.7	1.7	1.8	1.7	0.8	1.3	1.1	1.0	1.3

there's some people next door) and the plural form (e.g. *there **are** some people next door*), and the percentage use of the singular by country is shown in Figure 25.

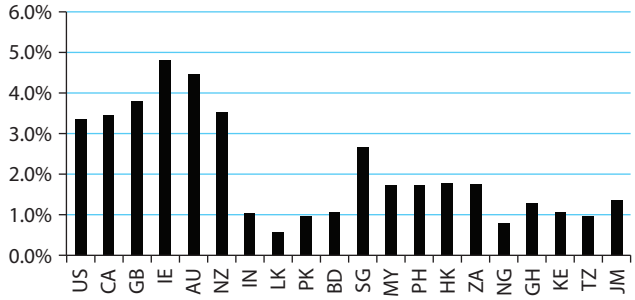


Figure 25. Percentage SG with existentials (e.g. *there's some people*)

Table 3 shows the data from another perspective. It depicts the overall percentage of the "non-standard" plural form in all Inner Circle and Outer Circle dialects. These data from GloWbE show that the Inner Circle dialects use the plural form about 2.9 times as frequently as the Outer Circle dialects (3.8/1.3), which agrees very nicely with Figure 2.7 in Collins (2012: 67).

Table 3. SG/PL agreement with existentials: Inner versus Outer Circle

	All Inner Circle	**All Outer Circle**
"Non-standard" PL	3,827	681
"Standard" SG	97,640	51,768
% PL	3.8%	1.3%

A similar case of linguistic conservatism on the part of speakers of Outer Circle varieties deals with another case of verbal agreement — this time with constructions like *each* (SG) *of them* (PL) {*is*|*are*} — in which the verb can agree with the formal head of the noun phrase (*each*/SG) or the notional head (*them*/PL). Prescriptively, agreement should be with the formal head (*each of them is*), but many speakers prefer agreement with the notional head (*each of them are*). To study this construction, we searched for *each*|*none of them*|*those*|*these are*|*were*|*have* for the non-standard plural form and *each*|*none of them*|*those*|*these is*|*was*|*has* for the standard singular form, as shown in Table 4. (Note that these two searches do not find all relevant forms, but those that they do find act as a good "proxy" for other forms, such as *each of my friends {is|are}*.)

Again, when we compare the Inner and Outer Circle dialects as a group, we see a striking difference. The 7,947 tokens in GloWbE show that speakers of the Inner Circle dialects are more likely to use the innovative, "incorrect" plural than are speakers of the Outer Circle dialects, who are linguistically more conservative.

Table 4. SG/PL agreement with e.g. *each of them* {*is*/*are*}

	US	CA	GB	IE	AU	NZ	IN	LK	PK	BD	SG	MY	PH	HK	ZA	NG	GH	KE	TZ	JM
PL	1369	334	1046	237	433	196	208	136	134	93	110	82	68	67	84	62	46	70	47	67
SG	509	188	540	157	215	102	197	81	160	93	62	105	83	74	78	126	94	101	48	45
%PL	73%	64%	66%	60%	67%	66%	51%	63%	46%	50%	64%	44%	45%	48%	52%	33%	33%	41%	49%	60%

Table 6. Subjunctive and indicative with hypotheticals, e.g. *as if he* {*was*/*were*}

	US	CA	GB	IE	AU	NZ	IN	LK	PK	BD	SG	MY	PH	HK	ZA	NG	GH	KE	TZ	JM
SUBJ	**16707**	3681	11983	2410	4111	1798	1654	911	1247	567	1254	1120	1301	835	835	1117	683	714	529	788
INDIC	**20969**	5963	24581	5522	8474	3937	2859	1337	2172	898	1972	1660	1692	1118	1939	2081	1328	1406	938	1798
%SUBJ	**44%**	38%	33%	30%	33%	31%	37%	41%	36%	39%	39%	40%	43%	43%	30%	35%	34%	34%	36%	30%

(And note that the chi square calculation shows that this difference is significant at $p < .0001$, as it is in Table 3 above.)

Table 5. SG/PL agreement with e.g. *each of them* {*is*/*are*}: Inner versus Outer Circle

	Inner Circle	**Outer Circle**
"Non-standard" PL	3,615	1,274
"Standard" SG	1,711	1,347
% PL	68%	49%

A final case of linguistic conservatism on the part of speakers of Outer Circle dialects deals with the use of the subjunctive in hypotheticals, which was the focus of Hundt, Hoffmann and Mukherjee (2012) for South Asian dialects. In their study, they looked at the frequency of the subjunctive (*were*) and the indicative (*was*) after *as if*, *as though*, *even if*, and other cases of *if*, e.g. *he acts as if he* {***was***/***were***} *king*. They found that speakers of the South Asian dialects used the older and more conservative subjunctive *were* more than speakers of BrE.

The data from the 146,889 tokens in GloWbE (Table 6) shows the same — to a point. All four South Asian dialects (shaded in Table 6) use the conservative *were* more than BrE (33 percent), as well as more than the other Inner Circle dialects of IE, AU, and NZ. Interestingly, AmE (shaded and bolded in Table 6) has the highest degree of the conservative *were*. The 44 percent figure in AmE is higher than any of the South Asian dialects studied in Hundt, Hoffmann and Mukherjee (2012), and the high degree of the subjunctive in American English confirms what Leech et al (2009) say about diachronic developments in the use of hypothetical subjunctives.

This may be due to strong prescriptive pressure for the use of the subjunctive in AmE (see Auer 2006), just as with the prescriptive pressure to avoid *try and*, discussed above. The data from GloWbE also show how data from a wide range of dialects of English can provide insights that otherwise might not be available with just a handful of dialects.

To finish this section on syntactic variation, we should note that GloWbE can also be used to look at discourse phenomena, such as discourse markers. We will briefly provide evidence for just one such construction. As Brinton (2009) notes, there is both historical and dialectal variation in the use of the two related discourse markers *having said that* and *that said*, which are used to refer back to something that has just been said, and then to provide an alternative point of view. The 8,529 tokens of *having said that*, show that this variant is much more common in BrE than in AmE (Figure 26), whereas the 11,208 tokens of *that said* show that this is much more common in AmE, and then progressively less frequent in CA, UK, IE, AU, and NZ (Figure 27).

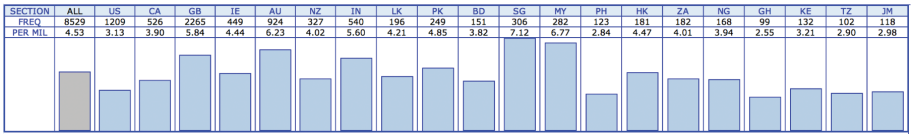


Figure 26. *having said that*

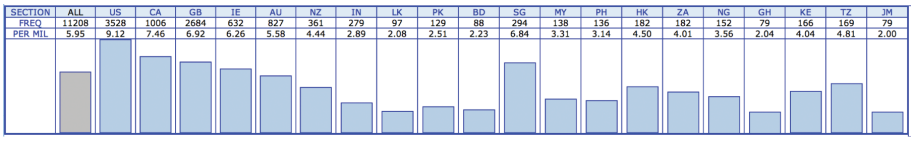


Figure 27. *that said*,

The strong preference for *that said* in AmE in GloWbE ties in nicely with historical data, which shows a large increase in *that said*, over the past few decades. For example, data from 81 tokens in the 400 million word COHA (left side of Figure 28) and 954 tokens in the COCA (right side of Figure 28) shows a clear increase over time. The important number is the frequency per million words, which has risen for example in COCA from 0.39 to 3.83 from the early 1990s to the 2010s.

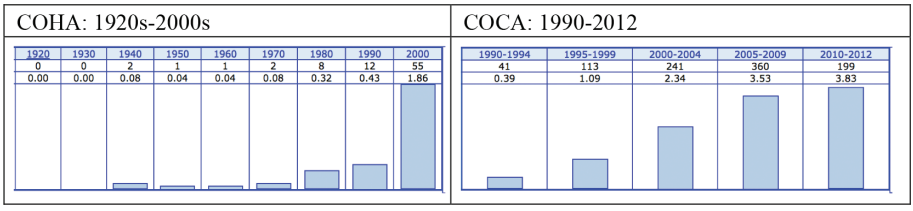


Figure 28. Historical increase in *that said*,

The ability to relate historical data to questions of dialectal variation is also enhanced in the BYU corpus interface. With just one click, users can seamlessly move from the results of one corpus to another (e.g. COCA or BNC for genre variation, and then GloWbE for dialectal variation, and then COHA for historical variation), and thus easily and quickly explore a wide range of variation in English.

6. Variation in meaning

What corpus-based data could provide evidence for differences in meaning between two or more dialects? For example, how would we know that in AmE *cupboard* is restricted primarily to storing items in a kitchen or pantry, whereas in BrE it can also be used for a storage place in other rooms in the house (cf. American *closet*)? Or how would we know that *scheme* is typically used in a negative sense in AmE, but that this is not the case in other varieties of English?

One approach would be to look at concordance lines for the word or phrase in different dialects, and to see whether the surrounding context might indicate differences in meaning. For example, Figure 29 shows a few of the concordance lines from the GB section of the GloWbE corpus.

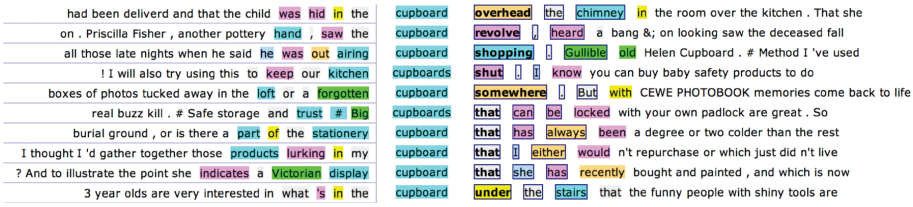


Figure 29. Concordance lines for [*cupboard*]

Notice that in these sentences, the cupboard is over the chimney (#1) or under the stairs (#10), that boxes of photos (#5) or stationary (#7) are stored there, and that it is possible to purchase a stand-alone cupboard (#9) — all of which would seem strange in AmE.

However, given a large enough corpus, we can use another approach. Rather than looking at all 8,726 tokens of the lemma *cupboard* in GloWbE, for example, we can simply have the corpus interface look for all collocates of *cupboard*. We can then compare the collocates to see which ones occur in one dialect but not another, and which may therefore signal differences in meaning and usage.

For example, Figure 30 shows a comparison of the collocates of *cupboard* in 386 million words of AmE (left) and 387 million words from GB (right) in GloWbE.

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	REFRIGERATOR	9	1	0.02	0.00	9.02
2	CLOSETS	12	2	0.03	0.01	6.01
3	ACTIVITY	6	1	0.02	0.00	6.01
4	CLOSET	9	2	0.02	0.01	4.51
5	PANTRY	8	3	0.02	0.01	2.67
6	PLATE	5	2	0.01	0.01	2.51
7	ITEMS	9	6	0.02	0.02	1.50
8	MOTHER	6	4	0.02	0.01	1.50
9	STUFF	6	6	0.02	0.02	1.00
10	WAY	5	5	0.01	0.01	1.00
11	GLASS	6	7	0.02	0.02	0.86
12	YEAR	6	8	0.02	0.02	0.75
13	YEARS	6	10	0.02	0.03	0.60
14	BATHROOM	7	15	0.02	0.04	0.47

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	AIRING	131	3	0.34	0.01	43.58
2	DAYS	15	1	0.04	0.00	14.97
3	SIDE	14	1	0.04	0.00	13.97
4	STORAGE	41	3	0.11	0.01	13.64
5	BROOM	53	4	0.14	0.01	13.22
6	SKELETONS	39	3	0.10	0.01	12.97
7	DUST	13	1	0.03	0.00	12.97
8	SKELETON	13	1	0.03	0.00	12.97
9	WARDROBES	13	1	0.03	0.00	12.97
10	FUME	25	2	0.06	0.01	12.47
11	STORE	72	7	0.19	0.02	10.26
12	BACK	92	10	0.24	0.03	9.18
13	CEREAL	9	1	0.02	0.00	8.98
14	WARDROBE	9	1	0.02	0.00	8.98

Figure 30. Collocates of [*cupboard*]: US (left), GB (right)

While not all of the collocates are of course relevant, many are. For example, *refrigerator* and *pantry* are more frequent (per million words) in AmE, probably because there are more references to *cupboard* in the context of a kitchen. In BrE, on the other hand, there are references to *brooms* and *wardrobes*, as well as to *skeletons in the cupboard*, all of which would be used with *closet* in AmE. (It is also important to recall that we can click on any collocate to see it in context with *cupboard* in the corpus, as shown in Figure 3 above.)

Further, let us consider the collocates of *scheme* in AmE and BrE, as shown in Figure 31. In AmE (left), there are references to *alleged*, *evil*, *fraudulent*, *Ponzi*, *(get) rich quick*, and *illegal* schemes, whereas in BrE (right) the collocates are much more prosaic and neutral in tone (or even positive: note *generous*, *innovative*, *competent*, and *qualified*). In corpus linguistic terms, we could say that *scheme* has "negative prosody" in AmE (cf. Louw 1993), whereas this is not the case for BrE.

6	ALLEGED	26	5	0.07	0.01	5.21	6	OVERSEAS	31	1	0.08	0.00	30.94
7	EVIL	48	10	0.12	0.03	4.81	7	DEFINED	127	5	0.33	0.01	25.35
8	FRAUDULENT	62	18	0.16	0.05	3.45	8	GENEROUS	50	2	0.13	0.01	24.95
9	NEFARIOUS	27	9	0.07	0.02	3.01	9	LABOUR	25	1	0.06	0.00	24.95
10	PONZI	617	255	1.60	0.66	2.42	10	TAX-AVOIDANCE	25	1	0.06	0.00	24.95
11	FEDERAL	30	13	0.08	0.03	2.31	11	SCOTTISH	24	1	0.06	0.00	23.95
12	REGULATORY	50	22	0.13	0.06	2.28	12	INNOVATIVE	70	3	0.18	0.01	23.28
13	AMERICAN	22	13	0.06	0.03	1.70	13	AUTOMATIC	23	1	0.06	0.00	22.95
14	ELABORATE	71	43	0.18	0.11	1.65	14	COMPETENT	23	1	0.06	0.00	22.95
15	RICH	86	55	0.22	0.14	1.57	15	QUALIFIED	22	1	0.06	0.00	21.95
16	QUICK	57	38	0.15	0.10	1.50	16	JOINT	21	1	0.05	0.00	20.96
17	ILLEGAL	53	36	0.14	0.09	1.48	17	VULNERABLE	21	1	0.05	0.00	20.96

Figure 31. Collocates of [*scheme*]: US (left), GB (right)

In these three cases, we compared BrE and AmE. This was done for two reasons. First, these are the two varieties with a global reach, and many speakers of other varieties are familiar with them. Second, these are the two largest segments of GloWbE, at about 385 million words each. Such comparisons may still be possible with smaller segments, perhaps even with countries like Tanzania (35 million words), Ghana (39 million words), or Bangladesh (40 million words), which are among the smallest in the corpus. This is especially the case if regional dialects are compared (e.g. Africa = 203 million words, or South Asia = 234 million words).

7. Discourse and culture

In this section, we will consider "discourse", in the sense of "which topics of discussion are more common" in one dialect (or groups of dialects) than another, and "what is being said" about particular concepts in different dialects.

At the most basic level, we can use GloWbE to compare the frequency of particular words in different dialects (cf. Section 3) and then consider how this may relate to the culture of those speakers. For example, words with *Buddh** (e.g. *Buddhist*, *Buddhism*, *Buddha*) are the most common in Sri Lanka (LK) — the one country in the corpus that is predominantly Buddhist (Figure 32).

SECTION	ALL	US	CA	GB	IE	AU	NZ	IN	LK	PK	BD	SG	MY	PH	HK	ZA	NG	GH	KE	TZ	JM
FREQ	120303	7870	1395	6193	1326	3287	1581	9001	55912	1840	5244	5134	7762	1280	10485	390	268	275	326	393	341
PER MIL	63.84	20.35	10.35	15.98	13.12	22.18	19.42	93.34	1,200.26	35.82	132.81	119.47	186.39	29.60	259.21	8.60	6.28	7.09	7.94	11.18	8.62

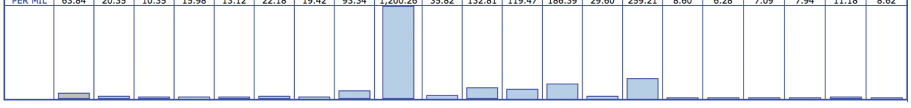


Figure 32. *Buddh**

Not surprisingly, *Quran* and *Allah* are likewise the most common in Pakistan, Bangladesh, and Malaysia — the three countries in the corpus with the greatest proportion of Muslims. Or consider *feminism*, which is the most common overall in the Inner Circle countries (Figure 33), although the frequency in Ireland (perhaps the most culturally conservative of these countries) is the lowest of these Inner Circle countries.

SECTION	ALL	US	CA	GB	IE	AU	NZ	IN	LK	PK	BD	SG	MY	PH	HK	ZA	NG	GH	KE	TZ	JM
FREQ	12235	4159	887	2932	557	1491	484	249	124	126	96	61	92	78	54	152	257	139	166	84	47
PER MIL	6.49	10.75	6.58	7.56	5.51	10.06	5.95	2.58	2.66	2.45	2.43	1.42	2.21	1.80	1.33	3.35	6.03	3.59	4.04	2.39	1.19

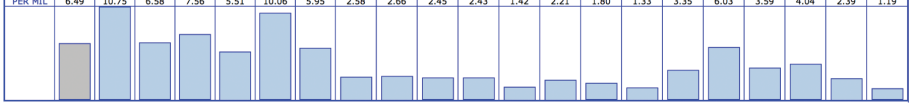


Figure 33. *feminism*

In addition to simply looking at word frequency, we can also compare the collocates of a given word, to see "what is being said" about particular concepts in different countries. For example, Figure 34 shows the most frequent adjectival collocates of *belief* in South Asia (left) and the six Inner Circle countries (right). Notice the use of *Hindu*, *Muslim*, *Islamic*, *polytheistic*, *monotheistic*, *sectarian*, and *heretical* in South Asia (all of which are probably related primarily to religion), compared to *liberal*, *deepest*, *positive*, *economic*, *confident*, *causal*, and *non-religious* in the Inner Circle countries (more secular).

SEC 1: 233,866,709 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	CHIEF BELIEF	10	1	0.04	0.00	53.01
2	HINDU BELIEFS	47	6	0.20	0.00	41.53
3	SECTARIAN BELIEFS	13	2	0.06	0.00	34.46
4	CORRUPT BELIEFS	12	2	0.05	0.00	31.81
5	AGE-OLD BELIEF	14	5	0.06	0.00	14.84
6	BLIND BELIEFS	11	4	0.05	0.00	14.58
7	POLYTHEISTIC BELIEFS	18	7	0.08	0.01	13.63
8	ESSENTIAL BELIEFS	15	6	0.06	0.00	13.25
9	HINDU BELIEF	42	22	0.18	0.02	10.12
10	WRONG BELIEF	43	25	0.18	0.02	9.12
11	WRONG BELIEFS	51	31	0.22	0.03	8.72
12	ISLAMIC BELIEF	113	79	0.48	0.06	7.58
13	HERETICAL BELIEFS	10	7	0.04	0.01	7.57
14	BUDDHIST BELIEF	27	19	0.12	0.02	7.53
15	MONOTHEISTIC BELIEF	15	12	0.06	0.01	6.63
16	ISLAMIC BELIEFS	111	94	0.47	0.08	6.26
17	MUSLIM BELIEF	39	39	0.17	0.03	5.30

SEC 2: 1,239,817,686 WORDS

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	SILLY BELIEFS	186	1	0.15	0.00	35.09
2	THEISTIC BELIEF	77	1	0.06	0.00	14.52
3	CONTRADICTORY BELIEFS	53	1	0.04	0.00	10.00
4	LIBERAL BELIEFS	42	1	0.03	0.00	7.92
5	APPARENT BELIEF	79	2	0.06	0.01	7.45
6	DEEPEST BELIEFS	39	1	0.03	0.00	7.36
7	SILLY BELIEF	34	1	0.03	0.00	6.41
8	POSITIVE BELIEF	58	2	0.05	0.01	5.47
9	SIMPLE BELIEF	58	2	0.05	0.01	5.47
10	CATHOLIC BELIEF	83	3	0.07	0.01	5.22
11	DIFFERING BELIEFS	55	2	0.04	0.01	5.19
12	DEEP BELIEFS	27	1	0.02	0.00	5.09
13	ECONOMIC BELIEFS	27	1	0.02	0.00	5.09
14	CONFIDENT BELIEF	25	1	0.02	0.00	4.72
15	NON-RELIGIOUS BELIEFS	24	1	0.02	0.00	4.53
16	CAUSAL BELIEFS	24	1	0.02	0.00	4.53
17	PRE-EXISTING BELIEFS	24	1	0.02	0.00	4.53

Figure 34. Collocates of [*belief*]: South Asia (left), Inner Circle (right)

Another example of the ability to gain cultural insight from the comparison of collocates are the adjectival collocates of the lemma *marriage* in the Outer Circle countries (left) and the Inner Circle countries (right) in Figure 35. In the Outer Circle countries, there is concern about *inter-caste*, *fixed*, and *forceful* marriages, as well as *permanent* versus *temporary* marriages (perhaps as a husband is forced to look for work outside of his home country). In the Inner Circle countries, on the other hand, people are apparently more concerned with the "hot button" topic of same-sex marriage, with adjectives like *opposite-sex* and *same-sex*, and related words like *anti-gay*, *supporting* and *preserving* (i.e. traditional heterosexual mar-

riage), as well as *pro-abortion* and *unborn* — apparently referring to "conservatives" and "liberals", in the context of their views on same-sex marriages.

SEC 1: 644,753,594 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	INTER-CASTE	91	2	0.14	0.00	87.49
2	FIXED	45	1	0.07	0.00	86.53
3	PHILIPPINE	35	1	0.05	0.00	67.30
4	FORCEFUL	26	1	0.04	0.00	50.00
5	NIGERIAN	51	2	0.08	0.00	49.03
6	CUSTOMARY	359	23	0.56	0.02	30.01
7	FIXED-TIME	164	0	0.25	0.00	25.44
8	HALAL	22	2	0.03	0.00	21.15
9	HINDU	271	26	0.42	0.02	20.04
10	PERMANENT	427	54	0.66	0.04	15.21
11	TEMPORARY	678	103	1.05	0.08	12.66
12	BLISSFUL	72	12	0.11	0.01	11.54
13	IRREPARABLE	23	4	0.04	0.00	11.06
14	ISLAMIC	322	65	0.50	0.05	9.53

SEC 2: 1,239,817,686 WORDS

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	IRISH	121	1	0.10	0.00	62.92
2	OPPOSITE-SEX	102	1	0.08	0.00	53.04
3	AUSTRALIAN	136	3	0.11	0.00	23.58
4	PRO	78	2	0.06	0.00	20.28
5	CONSISTENT	30	1	0.02	0.00	15.60
6	SAME-GENDER	79	3	0.06	0.00	13.69
7	IDENTICAL	48	2	0.04	0.00	12.48
8	NARROW	24	1	0.02	0.00	12.48
9	FACTO	47	2	0.04	0.00	12.22
10	PRO-ABORTION	23	1	0.02	0.00	11.96
11	UNBORN	20	1	0.02	0.00	10.40
12	ANTI-GAY	175	9	0.14	0.01	10.11
13	SUPPORTING	109	7	0.09	0.01	8.10
14	PRESERVING	28	2	0.02	0.00	7.28

Figure 35. Collocates of *families*: Outer Core (left), Inner Circle (right)

Before leaving this section, which deals with how the GloWbE data can provide cultural insights, we want to come back to something that we discussed in Section 3, where we dealt with lexical differences between the dialects. Recall that in that section, we looked for **ies* words in Australia, compared to the other Inner Circle countries, and found examples like *firies*, *bikies*, and *tradies*. These are interesting from a lexical point of view, but they provided some insight into cultural differences between the different countries.

We can do similar searches, however, which provide more culturally interesting data. For example, if we compare all **ism* words in Great Britain and South Asia, we find the following (Figure 36). In Great Britain (left), people are writing about *Euroscepticism*, *Labourism*, *presenteeism*, *nimbyism* (*nimby* = 'not in my backyard'), *monetarism*, *Thatcherism*, and *Blairism* — with most of these being political in nature. In South Asia, on the other hand, the **ism* words are much more related to religion — *Qadianism*, *castism*, *Talibanism*, *Vaisnaism*, *Shivaism*, *Shiaism*, and so on (with the exception of *Naxalism*). Thus there seems to be a real difference in terms of what people in these two regions are writing about on the Web.

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	EUROSCEPTICISM	137	1	0.35	0.00	82.66
2	LABOURISM	124	1	0.32	0.00	74.82
3	ANTI-FASCISM	86	1	0.22	0.00	51.89
4	PRESENTEEISM	81	1	0.21	0.00	48.87
5	THROMBOEMBOLISM	432	6	1.11	0.03	43.44
6	NIMBYISM	72	1	0.19	0.00	43.44
7	PRESBYTERIANISM	44	1	0.11	0.00	26.55
8	ISOMERISM	86	2	0.22	0.01	25.94
9	MONETARISM	81	2	0.21	0.01	24.44
10	BLAIRISM	93	0	0.24	0.00	23.99
11	THATCHERISM	382	10	0.99	0.04	23.05
12	LOCALISM	765	22	1.97	0.09	20.98
13	ANGLICANISM	259	8	0.67	0.03	19.53
14	BONAPARTISM	63	2	0.16	0.01	19.01
15	LUDDISM	31	1	0.08	0.00	18.70

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	QADIANISM	100	1	0.43	0.00	165.74
2	NAXALISM	145	2	0.62	0.01	120.16
3	ISMAILISM	68	1	0.29	0.00	112.70
4	SHAIVISM	56	1	0.24	0.00	92.82
5	MELIORISM	43	1	0.18	0.00	71.27
6	BRAHMANISM	114	3	0.49	0.01	62.98
7	ETERNALISM	72	2	0.31	0.01	59.67
8	CASTEISM	249	7	1.06	0.02	58.96
9	MARXISM-LENINISM-MAOISM	127	0	0.54	0.00	54.30
10	BUDHISM	107	4	0.46	0.01	44.34
11	JISM	404	16	1.73	0.04	41.85
12	MUHAMMADANISM	23	1	0.10	0.00	38.12
13	BISM	44	2	0.19	0.01	36.46
14	VAISHNAVISM	83	0	0.35	0.00	35.49
15	JAINISM	551	32	2.36	0.08	28.54

Figure 36. **ism* words: Great Britain (left), South Asia (right)

Finally, rather than comparing two countries or sets of countries, we can simply leave the query "open" as far as country goes, and look at all **ism* words in all countries. Here we find that discussions of *tourism* are more common in Africa and Jamaica, people in South Asia are writing more about *terrorism*, *autism* is

	CONTEXT	ALL	US	CA	GB	IE	AU	NZ	IN	LK	PK	BD	SG	MY	PH	HK	ZA	NG	GH	KE	TZ	JM
1	TOURISM	66204	2859	3177	7376	3290	4234	3871	3563	3716	922	1703	2135	2451	2312	2945	2635	1092	2838	3746	6370	4969
2	CRITICISM	62731	14465	3644	15808	3164	4983	2298	3017	1839	2200	1148	811	1022	813	1123	1450	1314	1036	967	717	912
3	MECHANISM	44334	8850	2552	8021	2291	3576	1793	3274	1736	1105	1178	885	920	705	1634	1063	758	830	1344	1067	752
4	TERRORISM	42198	8783	1909	6845	732	2101	881	2940	5427	5529	1570	317	471	317	417	395	1277	461	1023	544	259
5	JOURNALISM	41459	10280	2878	10441	1591	3953	1090	1693	997	743	927	522	332	613	647	840	784	908	895	863	462
6	CAPITALISM	37319	9466	2266	10261	1942	2835	1549	1356	681	602	871	461	218	367	875	850	516	394	370	817	622
7	RACISM	36539	11535	1894	8545	1859	2987	1051	797	1082	579	332	501	831	198	327	1185	586	675	505	367	703
8	BUDDHISM	21793	1829	309	1434	351	757	390	1790	9061	324	828	845	1200	314	1954	74	66	68	58	86	55
9	AUTISM	20293	7240	1508	5277	1585	2207	260	715	73	56	273	73	98	159	104	65	38	76	36	70	380
10	SOCIALISM	19837	6423	792	4291	1020	1732	733	746	292	284	535	191	114	223	534	412	173	202	156	690	294
11	OPTIMISM	15125	2950	1249	3764	765	988	532	677	265	375	324	346	244	327	303	295	363	379	483	242	254
12	NATIONALISM	14389	1521	880	3052	1020	851	268	1032	1473	886	772	141	184	285	310	368	347	277	212	229	281
13	COMMUNISM	14198	4465	630	3284	632	1249	401	501	190	317	377	202	227	234	330	395	159	118	132	207	148
14	BAPTISM	12367	2696	1506	1314	965	917	814	193	178	82	793	129	88	695	252	284	223	571	166	300	201
15	FEMINISM	12215	4157	886	2928	556	1491	482	248	124	125	95	61	92	77	52	152	255	138	165	84	47
16	ATHEISM	10705	4644	354	1763	404	1565	552	238	70	241	101	37	74	139	66	131	114	16	94	51	51

Figure 37. **ism* words; all dialects

discussed most in the Inner Circle countries, and the topic of *feminism* is more common in the Inner Circle countries as well, as we discussed earlier in this section.

8. Conclusion

In Sections 3–7, we have seen a number of examples showing how the data from GloWbE can be used to insightfully investigate a wide range of phenomena in different dialects of English. One aspect of this that we have alluded to throughout the paper, but which we might touch on in a somewhat more detailed fashion here, is the importance of corpus size.

Other than GloWbE, the only other corpus of English that contains data from a number of different dialects, and which is organized in a way that allows us to compare across these dialects, is ICE. As we have discussed previously, ICE contains one million words each for 14 different dialects (11 of which contain both spoken and written English), for a total of about 12,200,000 words of text. GloWbE, on the other hand, contains about 1.9 billion words of data. In other words, GloWbE is more than 150 times larger than ICE. Where ICE may yield 20–30 tokens of a given word, phrase, or construction, GloWbE will often yield 150 times as many, or in other words 3,000–4,000 tokens for the same phenomenon. Another advantage of GloWbE is that it provides data on a number of varieties so far not included in ICE (such as Pakistani and Malaysian English).

For high frequency syntactic constructions, ICE often has enough data, and this is why it is probably no surprise that so many ICE-based studies in fact deal with rather high frequency constructions. But for many of the phenomena discussed in this paper, ICE would probably not have enough tokens. For example, most of the words and phrases shown in Section 3 occur 500–2000 times in GloWbE, yet they would only occur between perhaps four and 15 times in ICE. In terms of morphological variation, contrasting forms like *dived*/*dove* occur

1,000–1,200 times in GloWbE, and they might therefore only occur six or seven times in ICE. In GloWbE there are about 8,000 tokens for a construction like *each of them* {*is*|*are*}, and in ICE there would be only about 50 tokens — probably too few to say much of interest. And things are even more problematic in terms of the number of tokens for collocates shown in Section 6 and Section 7. For a given collocate, there are often only 30–40 tokens in GloWbE, and with a corpus only 1/150th the size, we might be lucky to have a single token in ICE.

But of course size is not everything. The ICE corpora have been constructed very carefully, and for phenomena where "every token counts" and when there can be no "messiness" at all in the data, the carefully-curated, manually annotated ICE corpora may be more useful than GloWbE. Likewise, for phenomena where actual spoken material is needed, ICE will probably be better than GloWbE, where there is no spoken data (although the 60 percent or so of texts in GloWbE that come from blogs do provide fairly informal language). Finally, in GloWbE we only know that a website is from a particular country, but there might be speakers from other countries who have posted to that website. In ICE, on the other hand, care has been taken to ensure that all speakers are from the country in question.

In other words, it is probably not an "either/or" issue when it comes to the use of different corpora, in which the use of one corpus precludes the use of another. Researchers may want to use ICE for some studies, GloWbE for others, and perhaps proprietary corpora that they have created for yet other studies. All of these can be seen as useful "tools" in the researchers' "toolbox", and they complement each other nicely.

To the extent, though, that researchers do adopt GloWbE as part of their "toolbox" (along with ICE and other corpora), they will be able to expand their horizons in terms of the types of variation that they consider, as they carry out research on World Englishes.

References

Auer, Anita. 2006. "Precept and Practice: The Influence of Prescriptivism on the English Subjunctive". *Linguistic Insights – Studies in Language and Communication* 39: 33–53.

Baker, Paul. 2011. "Times May Change but We'll Always Have Money: A Corpus Driven Examination of Vocabulary Change in Four Diachronic Corpora". *Journal of English Linguistics* 39: 65–88. DOI: 10.1177/0075424210368368

Bauer, Laurie. 1993. *Manual of Information to Accompany the Wellington Corpus of Written New Zealand English*. Wellington: Victoria University of Wellington.

Brinton, Laurel. 2009. The Development of 'that said'. Paper presented at American Association of Corpus Linguistics, University of Alberta.

Buchstaller, Isabelle. 2006. "Social Stereotypes, Personality Traits and Regional Perception Displaced: Attitudes towards the 'New' Quotatives in the U.K.". *Journal of Sociolinguistics* 10: 362–381. DOI: 10.1111/j.1360-6441.2006.00332.x

Chambers, Jack K. 1998. "Social Embedding of Changes in Progress". *Journal of English Linguistics* 26: 5–36. DOI: 10.1177/007542429802600102

Collins, Peter. 2012. "Singular Agreement in There-Existentials: An Intervarietal Corpus-Based Study". *English World-Wide* 33: 53–68. DOI: 10.1075/eww.33.1.03col

Corpus of Global Web-based English. http://corpus2.byu.edu/glowbe.

D'Arcy, Alexandra. 2012. "The Diachrony of Quotation: Evidence from New Zealand English". *Language Variation and Change* 24: 343–369. DOI: 10.1017/S0954394512000166

Davies, Mark. 2009. "The 385 + Million Word Corpus of Contemporary American English (1990–2008 +): Design, Architecture, and Linguistic Insights". *International Journal of Corpus Linguistics*. 14: 159–190. DOI: 10.1075/ijcl.14.2.02dav

Davies, Mark. 2011. "The Corpus of Contemporary American English as the First Reliable Monitor Corpus of English". *Literary and Linguistic Computing* 25: 447–465. DOI: 10.1093/llc/fqq018

Davies, Mark. 2012. "Expanding Horizons in Historical Linguistics with the 400 Million Word Corpus of Historical American English". *Corpora* 7: 121–157. DOI: 10.3366/cor.2012.0024

Francis, W. Nelson. 1964. "A Standard Sample of Present-Day English for Use with Digital Computers. Report to the US Office of Education in Cooperative Research Project No. E-007".

Goldberg, Adele. 1997. "Making One's Way Through the Data". In Masayoshi Shibatani and Sandra Thompson, eds. *Grammatical Constructions: Their Form and Meaning*. Oxford: Clarendon Press, 29–53.

Greenbaum, Sidney, ed. 1996. *Comparing English Worldwide: The International Corpus of English*. Oxford: Oxford University Press.

Hommerberg, Charlotte, and Gunnel Tottie. 2007. "'Try to' or 'try and'? Verb Complementation in British and American English". *ICAME Journal* 31: 45–64.

Hundt, Marianne, Andrea Sand, and Rainer Siemund. 1998. *Manual of Information to Accompany the Freiburg-LOB Corpus of British English ('F-LOB')*. Freiburg: Department of English. Albert-Ludwigs-Universität Freiburg.

Hundt, Marianne, Andrea Sand, and Paul Skandera. 1999. *Manual of Information to Accompany the Freiburg-Brown Corpus of American English ('Frown')*. Freiburg: Department of English. Albert-Ludwigs-Universität Freiburg.

Hundt, Marianne, Sebastian Hoffmann, and Joybrato Mukherjee. 2012. "The Hypothetical Subjunctive in South Asian Englishes: Local Developments in the Use of a Global Construction". *English World-Wide* 33: 147–164. DOI: 10.1075/eww.33.2.02hun

Johansson, Stig. 1980. "The LOB Corpus of British English Texts: Presentation and Comments". *ALLC Journal* 1: 25–36.

Kachru, Braj B. 1985. "Standards, Codification and Sociolinguistic Realism: The English Language in the Outer Circle". In Randolph Quirk and Henry Widdowson, eds. *English in the World: Teaching and Learning the Language and Literatures*. Cambridge: Cambridge University Press, 11–30.

Kilgarriff, Adam et al. 2014. "The Sketch Engine: Ten Years On." *Lexicography* 1: 1–30. www.sketchengine.co.uk (accessed August 20, 2014). DOI: 10.1007/s40607-014-0009-9

Leech, Geoffrey, Marianne Hundt, Christian Mair, and Nicholas Smith. 2009. *Change in Contemporary English: A Grammatical Study*. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511642210

Lindquist, Hans. 2009. *Corpus Linguistics and the Description of English*. Edinburgh: Edinburgh University Press.

Louw, William. 1993. "Irony in the Text or Insincerity in the Writer? The Diagnostic Potential of Semantic Prosodies". In Mona Baker, Gill Francis, Elena Tognini-Bonelli, eds. *Text and Technology: In Honour of John Sinclair*. Amsterdam: John Benjamins, 157–176. DOI: 10.1075/z.64.11lou

Peters, Pam. 1987. "Towards a Corpus of Australian English". *ICAME Journal* 11: 27–38.

Rodríguez Louro, Celeste. 2013. "Quotatives Down Under: *Be like* in Cross-Generational Australian English Speech". *English World-Wide* 34: 48–76. DOI: 10.1075/eww.34.1.03rod

Rudanko, Juhani. 2002. *Complements and Constructions*. Lanham, MD: University Press of America.

Shastri, S. V. 1988. "The Kolhapur Corpus of Indian English and Work Done on Its Basis So Far." *ICAME Journal* 12: 15–26.

Tagliamonte, Sali, and Alexandra D'Arcy. 2004. "'He's like, she's like': The Quotative System in Canadian Youth". *Journal of Sociolinguistics* 8: 493–514. DOI: 10.1111/j.1467-9841.2004.00271.x

Author's address

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