More than a peephole

Using large and diverse online corpora

Mark Davies
Brigham Young University, Utah

My two posts to the “Boot Camp” on the Corpora List thread were in response to two other posts by the computational linguist Linas Vepstas. In his first post, he responded to Bill Louw’s comments about large corpora like the Bank of English by saying:

> Except that the damned thing is not actually publicly available! It consists of copyrighted text. Damn! Those of us who actually need to do things outside of its limited, proscribed “allowed” usage are shit-outa-luck, and are reduced to analyzing Wikipedia, Project Gutenberg, and arxiv.org. I’d like to have access to a larger body of text that is more varied and diverse.[1]

I then sent a very short message to the List suggesting that Linas Vepstas take a look at the Corpus of Contemporary American English (COCA), which had been released a few months before in early 2008, and which contained (at that time) nearly 400 million words from a wide range of genres. He then replied:

> I need access to the original source, so that I can run it through my tools. I suppose that exploring the landscape by looking through a peephole in a wall is appropriate for some approaches, but clearly fails if what you need is not visible in the peephole.[2]

I then responded:

> I’m sure you’re aware of copyright issues. That’s the main reason that there are restrictions on access. Sure, we could give everyone full-text access, but then we’d probably end up in prison. But these interfaces certainly aren’t “peepholes” — far from it. Tens of thousands of people use them each month for an extremely wide range of research topics.[3]

In this short paper, I would like to expand on two issues raised in this thread. The first is a short discussion of copyright, and why it creates problems in terms of what can be made available with online corpora. The second and perhaps more

International Journal of Corpus Linguistics 15:3 (2010), **412–418**. DOI 10.1075/ijcl.15.3.13dav
ISSN 1384–6655 / E-ISSN 1569–9811 © John Benjamins Publishing Company

important and interesting issue is to go straight at the point that online corpora — such as those that I and others have created — are just a "peephole" into the language. I will argue that these corpora are of immense value for a wide range of other researchers, such as linguists who are interested in looking at language variation and change, as well as language teachers and learners. No resource is going to meet the needs of all possible users, but the point is to create as useful of tools as is possible for the largest number of researchers. Some online corpora do this quite well, we would argue, and these are resources that really are much more than "peepholes" into a language.

1. Copyright issues or "the travails of online corpora"

Linas Vepstas is correct in stating that there seems to be a choice between two types of texts. On the one hand, users can have full-text access to "public domain" materials like Project Gutenberg, Wikipedia, or even "Web as Corpus". Yet these are problematic in that either there are not a wide range of genres (Gutenberg, Wikipedia), or else it is difficult or impossible to separate the genres (Web as Corpus). As a result, these texts do a very poor job of representing the full range of what is contained in the language. On the other hand, text corpora like the British National Corpus, the Bank of English, or the Corpus of Contemporary American English do provide a much more balanced corpus, and therefore a much more realistic and also nuanced view of the language. However, these texts often do not come from public domain sources and are therefore problematic in terms of copyright and access.

For those who believe in creating corpora that are as representative of the "real world" language as possible, what are the options in terms of copyright and making the texts available to end users? In the examples that follow (which are based on material in Davies, in press b), I provide examples of three different ways in which things can play out. I do so with the goal of showing real-world approaches to creating public domain resources. This will enable us to move beyond the call to "make it all available, and completely for free" (and make it really large and quite balanced too, please), which seems to be the overly-simplistic mantra of some in the NLP or corpus linguistic community.

The first scenario in corpus creation is that there is generous "seed money" from a sponsoring institution. As a real-world example, we might consider the British National Corpus (which received millions of dollars in seed money from the Oxford University Press) or the Bank of English (which received millions of dollars in support from Collins-Cobuild). Publishers are enticed to be part of a consortium (in the case of the BNC), or the consortium is one large publisher.

These publishers participate because of the commercial value that they see in such an arrangement, and they provide the texts and the copyright permission to use the texts in the corpus.

The second way things can play out is what happened with the American National Corpus. There was no rich institution to provide seed money. Attempts were made to form a consortium, but this failed. Very few texts became available, what is there is not balanced in terms of sources or genres, and the corpus has not been completed.

The third approach is the one that we have taken in the creation of several other large corpora (see http://corpus.byu.edu). Because we knew that we did not have the millions of dollars that were used to create the BNC and the Bank of English, and wanting to avoid the fate of the ANC, we opted for a different plan. Copyrighted texts would be used as the basis of the corpus, but no copyright permission was requested or obtained from the copyright holders.

Without copyright permission we cannot ever release the full-text version of the corpus into the public domain. It is available, however, to search and access the corpus via a web-based interface that allows for an extremely wide range of queries. KWIC displays are limited to more or less what one sees on Google or in Google Books — the node word(s) surrounded by 40–60 words (180–200 words in expanded view). U.S. copyright law does allow for the use of copyrighted material, as long as (among other conditions) the end user does not have access to more than a certain percentage of the original text, and cannot "re-create" the original text by stringing together different pieces of the text. We have been criticized by some for not allowing full-text access to these corpora, but there simply is no legal alternative — none whatsoever.

We have dwelt on this final approach at some length because this is probably the most realistic scenario of the three for others contemplating the creation of corpora. Well-funded corpora like the BNC or the Bank of English are the exception, not the rule. Some other lucky individuals may be part of some large government-funded plan to create a "national corpus" of a given language. But by and large, there is little if any money available to obtain copyright permissions, or to entice publishers to be part of a consortium. In this case, the only option is to use copyrighted materials, and then restrict access to the end users in some way. So while this does not provide the complete access to the texts in the way that some in the NLP community might want, it is in fact the only realistic option in many cases.

2. More than a peephole

Now that we have briefly considered *why* there may be some limits in terms of end-user access to online corpora, let us now discuss whether this handicaps the corpora to such a point that they end up being nothing more than "peepholes" — forcing us to miss out on wide, expansive vistas into the full landscape of what is happening in the language.

In the discussion that follows, we will focus on the types of data that one can obtain from the architecture and interface that are used for the corpora from corpus.byu.edu — the Corpus of Contemporary American English (400 million words, 1990–2009+), the Corpus of Historical American English (400 million words, 1810's-2000's), the TIME Corpus (100 million words, 1920's-2000's), BYU-BNC (our interface to the British National Corpus), and corpora of Spanish and Portuguese. Some of these types of searches are also available from other online corpus interfaces, such as BNCweb, Sketch Engine, and other approaches that are based on the IMS Corpus Workbench, as well as alternate approaches such as VISL and PIE — although only the full range of searches discussed below can be found via our interface (see Davies 2009 for more details).

At the most basic level, users of these corpora can search by word (*mysterious*), phrase *(nooks and crannies* or *faint* + noun), lemmas (all forms of words, like *sing* or *tall*), wildcards (un*ly or r?n*), and more complex searches such as un-X-ed adjectives (*unexpected*, *unlimited*) or verb + any word + a form of *ground* (*hit the ground*, *breaking new ground*). Besides seeing the "frequency results" window, users can of course see the word or phrase in context — up to about 90–100 words to the left and 90–100 words to the right. This is of course a limitation based on the fact that these are copyrighted texts, as discussed above.

Besides choosing to see a list of all the individual matching strings, users can also see a chart display that shows the total frequency of all strings in the five "macro" registers (spoken, fiction, popular magazines, newspapers, and academic journals), or (in historical corpora) across the different centuries or decades in the corpus. Via the chart display, users can also see the frequency of the word or phrase in sub-registers as well, such as movie scripts, children's fiction, women's magazines, or medical journals (or even year by year, such as each of the 200 years in the Corpus of Historical American English, 1810–2009). Via the web interface, users can search for collocates, which of course provide insight into the meaning and usage of words and phrases. For example, users can search for the most common nouns near *thick* (*hair*, *glasses*, *smoke*), adjectives near *smile* (sorted by Mutual Information score; *wry*, *rueful*, *toothy*), nouns after *look into* (*eyes*, *future*, *matter*), or words starting with *clos** near *eyes* (*closing my eyes*, *their eyes were closed*).

Users can also include information about genre or a specific time period directly as part of the query. This allows them to see how words and phrases vary across speech and many different types of written texts. For example, they can easily find which words and phrases occur much more frequently in one register than another, such as *faint* + [noun] in fiction (*smile*, *light*, *sound*), or verbs in the slot [*we * that*] in academic writing (*hypothesize*, *assume*, *suggest*). They can also apply this to collocates, such as nouns with the verb *break* in newspapers (*record*, *law*, *ground*) or adjectives with *woman* in fiction texts (*young*, *beautiful*). Finally, they can compare one section to another, such as nouns near *chair* in academic (*department*, *committee*) vs fiction (*back*, *table*), nouns with *passionate* in fiction (*kiss*, *lovemaking*) vs newspaper (*fans*, *game*), adjectives in medical journals compared to other journals (*preoperative*, *histopathologic*), or verbs in sports magazines compared to other magazines (*ski*, *blitz*, *punt*).

Finally, users can easily carry out semantically-oriented searches. For example, they can compare nouns that appear with *small* (*percentage*, *fraction*) and *little* (*brother*, *attention*), verbs with *he/she* (*bombed*, *campaigned* vs *sniffled*, *giggled*), adjectives with *Democrats* (*electable*, *fun*) and *Republicans* (*extremist*, *mean-spirited*), or verbs with *Bush* (*drill*, *deploy*) and *Clinton* (*grope*, *inhale*). They can also find the frequency and distribution of synonyms of a given word, such as *beautiful* or the verb *clean*, see which synonyms are more frequent in competing registers (such as synonyms of *strong* in fiction (*beefy*, *burly*) and academic (*effective*, *deep-seated*)), and they can use synonyms as part of a more complex query (such as all synonyms of *clean* with any noun; *clean the house*, *wipe the sweat*, *mopping the floor*). Finally, they can create "customized lists" for any category that interests them, and then re-use these in subsequent queries (such as colors + clothes, or words related to *beautiful* + synonyms of *woman*).

As a more focused example of the range of queries and how they can be used to look at language variation and change, consider the following questions dealing with current change in English (see Davies, in press a, for more details). All of these questions — as with all of the queries above — can be answered with just one click after filling out the simple web-based search form:

- Lexical change: What is the frequency of *morph*, *old-school*, *gift* (as a verb), *freak out*, *perfect storm*, *(think) outside the box*, or *political* correct** over time? What verbs or nouns or adjectives or phrasal verbs with *up* are used a lot more in 2005–09 than in 1990–94?
- Morphological change: Are words with the suffix *-gate* (indicating "scandal") and the suffix *-friendly* more frequent in the 1990's or the 2000's? What is the frequency of words ending in *-ism* (e.g. *communism*, *terrorism*) in each time period since the early 1990's, and which *-ism* words are more common in the 2000's than the 1990's (and vice versa)?

- Syntactic change: Are the following increasing or decreasing (and when): end up V-ing, *get* passive (*got hired*), "quotative *like*" (*he's like*, *I'm not going*), *so not* ADJ (*I'm so not interested in her*).
- Semantic change: Changes over time with collocates (nearby words) can often indicate changes in meaning or the usage of a given word. See if this is true for the following words: *green*, *web*, *engine*.
- Discourse analysis ("what are we saying about X?"): Compare the collocates for the given words in the 1990's and the 2000's: *crisis*, *terror*, *gay*, *religion*. Or look at the collocates of *nuclear* and *crisis* in each time period since the early 1990's.

Note that these queries deal with just the last twenty years of American English (1990–2009). When the 400-million-word Corpus of Historical American English (1810's-2000's) comes online in Summer 2010, it will be possible to carry out queries like this with the same degree of ease, to look at a wide range of changes during the past two hundred years, with much more precision than with any other resource — whether it be web-based corpora or text archives like Google Books or Project Gutenberg.

While we have focused here on using online corpora to look at language variation and change, they are of course also very useful for language teachers, language learners, and materials developers. For example, language learners can move far beyond a typical thesaurus, to compare the overall frequency and the usage of two competing synonyms, to see the full range of collocates that occur with all of the synonyms in a particular semantic field, or so see which synonyms are used in different genres — all with one click of the mouse. With the ability to have the corpus generate lists of words that are much more common in one genre than another (or sub-genre, like Academic-Technology or Newspaper-Sports), students and developers can quickly and easily create materials for English for Specific Purposes.

To conclude, however, we should recognize that even with all of the types of searches that are possible via the web-based interface, what many in the NLP community still need and want is access to the raw texts themselves. As we have discussed, this is not possible in many cases, because of copyright issues. Nevertheless, in many cases it is still possible to obtain large amounts of word frequency and n-gram information from these corpora, since that data would not violate copyright law (at least US Copyright Law).

For example, we recently released the *Frequency Dictionary of American English: Word Sketches, Collocates, and Thematic Lists* (Davies & Gardner 2010), and much more data is available in machine-readable format than what is available in the printed book (see http://www.wordfrequency.info). These data contain the top 200–300 collocates for each of the top 20,000 lemmas of English (for a total of more than 4,300,000 node/collocate pairs), with frequency and Mutual Information score, and one can optionally get frequency lists that show the frequency by genre as well. Finally, we have n-grams data on all 2-grams and 3-grams in the entire 400-million-word Corpus of Contemporary American English. While not as large as the Google n-grams database, it does contain data from a wide range of identifiable genres, which is of course not possible with the Google web data.

With all of the preceding as examples of what can be done with these web-based corpora, as well as the frequency and n-grams data that is available offline, is it fair to say that these corpora only provide a small "peephole" into the language? We believe that it provides much, much more than that.

Notes

[1] Vepstas, message # 12267, 18 August 2008.

[2] Vepstas, message # 12270, 18 August 2008.

[3] Davies, message # 12271, 18 August 2008.

References

Davies, M. 2009. "The 385+ million word *Corpus of Contemporary American English* (1990–2008+): Design, architecture, and linguistic insights". *International Journal of Corpus Linguistics*, 14 (2), 159–190.

Davies, M. & Gardner, D. 2010. *A Frequency Dictionary of American English: Word Sketches, Collocates, and Thematic Lists*. London: Routledge.

Davies, M. In press, a. "The *Corpus of Contemporary American English* as the first reliable monitor corpus of English". *Literary and Linguistic Computing*.

Davies, M. In press, b. "Corpus linguistics questions and answers". In G. Barnbrook & V. Viana (Eds.), *Perspectives on Corpus Linguistics*. Amsterdam/Philadelphia: John Benjamins.

Author's address

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

Mark_Davies@byu.edu