Mark Davies

Using (and Useful) Corpora for the Study of HEL

This short overview of publicly available online corpora may be of value to students of the history of the English language. It focuses on the historical corpora that are available from the Brigham Young University suite of corpora (corpus.byu.edu) and on the practical applications of them. But there are other valuable corpora, among them the *Helsinki Corpus* and other corpora from the University of Helsinki, the Brown family of corpora, the *Corpus of Late Modern English Texts*, and many smaller—one- to five-million-word—corpora. A good listing can be found at www.helsinki.fi/varieng/CoRD/.

There are also text archives, including *Early English Books Online (EEBO)*, *Eighteenth Century Collections Online (ECCO)*, ProQuest's *Literature Online*, and many historical newspapers. Although most of these lack sophisticated search engines, they can still be useful for the study of lexical and phraseological changes in English (see Davies, "Corpora").

Google Books (books.google.com) and *Google Books Ngram Viewer* (books.google.com/ngrams) contain data from hundreds of billions of words of text. The standard *Google Books* interface is limited mainly to looking at lexical frequency over time. A more powerful interface to the

same data can be found at googlebooks.byu.edu, which allows many of the types of searches discussed below, including syntactically oriented searches and the use of collocates (co-occurring words) to see changes in meaning (see Davies, "Making").

Language change is, of course, still ongoing, and many students find it interesting to look at change in progress today in order to understand similar mechanisms that may have been operative three or five hundred years ago. Perhaps the best corpus for this purpose is the *Corpus of Contemporary American English (COCA)*, which currently contains 520 million words of data (20 million words each year, 1990–2015) and is updated every year or two (see Davies, "Corpus"). In addition, the new *NOW* (News on the Web) *Corpus* grows by four to five million words each day (or about 130 million words each month, or 1.5 billion words each year) and thus allows students to look at changes essentially in real time.

The main historical corpora at Brigham Young University of value to HEL students, however, are the *Corpus of Historical American English* (*COHA*; 400 million words, 1810s–2000s), the *Time* corpus (*Time Magazine*; 100 million words, 1920s–2000s), and the British *Hansard* (speeches in British Parliament, 1.6 billion words, 1803–2005). The Brigham Young corpora include *Early English Books Online* (400 million words, 1470s–1690s), which will soon be replaced with a larger, newer version.[1]

This overview provides examples of the types of phenomena that students can research with *COHA*. Because all these corpora use the same architecture and interface, similar searches can be carried out with the other Brigham Young corpora. But most of these searches cannot be performed with the much smaller corpora—they require a larger historical corpus, like *COHA* (see Davies, "Expanding" and "Examining"). With *COHA*, students can find the frequency of words and phrases since the 1810s, including words and phrases that have decreased since the 1800s (*bosom, grieved, bestow, of no little*), that increased and then decreased (*anyhow, mustn't, naughty, as though to, far-out, swell* [adj.]), and that increased (*a lot of, guys, unleash, screw up, freak out*).

Because of the unique corpus architecture and interface, students can also find all words that were used more in one period than another, even when they do not have particular words in mind—the database is relational, so queries on it can ask all words in one historical period to be compared with all words in another. For example, verbs in the 1970s–2000s (*replicate, download*) can be compared with verbs in the 1930s–1960s (*grudge,* *effectuate*), and adjectives in the 1970s–2000s (*environmental, global*) can be compared with adjectives in the 1830s–1870s (*pecuniary, sagacious*).

Students can also find words and phrases that were related to changes in society and culture, or to historical events, such as *emancipation, steamship, telegraph, flapper*, fascis*, teenage**, and *communis**. In the course Corpus-Based Approach to the History of American English at Brigham Young, we examine words and phrases such as these to look at changes in American culture and society through the lens of corpus data (for the syllabus of this course, see davies-linguistics.byu.edu/elang495/).

Because the corpora are tagged for part of speech and are lemmatized, students can also search for changes with grammatical constructions—like *end up* V-ing, VERB PRON *into* VERB-ing (e.g., *talked them into going*), phrasal verbs (earlier = *gather up, take up*; later = *show up, wind up*), post-verbal negation with *need* (*needn't mention*), or sentence-initial *hopefully*. They can also look for stylistic constructions (part lexical, part syntactic), such as now "old-fashioned" [ADJ as to VERB] (*so good as to show me*), or [have quite VERB-ed] (*until she had quite finished*).

Students can also study morphological change and word formation by mass comparison of all matching forms in two time periods. Examples might be *heart* (earlier = *noble-hearted, heart-burnings*; later = *heartland, heartbeat*), *-able* adjectives (earlier = *supposable, exceptionable*; later = *affordable, predictable*) or the rise of *free (*tax-free, fat-free*) or *friendly words (*eco-friendly, kid-friendly*) in the past few decades.

The corpus also shows how the meaning or usage of words has changed over time, if students look at changes in collocates. Consider, for example, the collocates of *gay* in the 1800s (*laugh, colors*) and the 1970s–2000s (*lesbian, rights*). Collocate change can also signal cultural changes over time, such as adjectives used with *women* in the late 1800s (*tender, cultivated*) and the late 1900s (*sexual, divorced*).

These new corpora allow teachers and students of English to look at a wide range of changes in the language (lexical, morphological, syntactic, and semantic) in ways that would have been quite unthinkable even ten to fifteen years ago.

Note

1. *EEBO* has some drawbacks because of the types of texts it includes and how it was prepared. But it is still a useful source for changes in Early Modern English.

Works Cited

Davies, Mark. "Corpora: An Introduction." *The Cambridge Handbook of English Corpus Linguistics*, edited by Douglas Biber and Randi Reppen, Cambridge UP, 2015, pp. 11–31.

———. "The Corpus of Contemporary American English as the First Reliable Monitor Corpus of English." *Literary and Linguistic Computing*, vol. 25, no. 4, 2010, pp. 447–64.

———. "Examining Syntactic Variation in English: The Importance of Corpus Design and Corpus Size." *English Language and Linguistics*, vol. 19, no. 3, 2014, pp. 1–39.

———. "Expanding Horizons in Historical Linguistics with the 400-Million Word Corpus of Historical American English." *Corpora*, vol. 7, no. 2, 2012, pp. 121–57.

———. "Making Google Books N-Grams Useful for a Wide Range of Research on Language Change." *International Journal of Corpus Linguistics*, vol. 19, no. 3, 2014, pp. 401–16.