DIGITAL DATA

Google Opens Books to New Cultural Studies

In March 2007, a young man with dark, curly hair and a Brooklyn accent knocked on the door of Peter Norvig, the head of research at Google in Mountain View, California. It was Erez Lieberman Aiden, a mathematician doing a Ph.D. in genomics at Harvard University, and he wanted some data. Specifically, Lieberman Aiden wanted access to Google Books, the company's ambitious—and controversial—project to digitally scan every page of every book ever published.

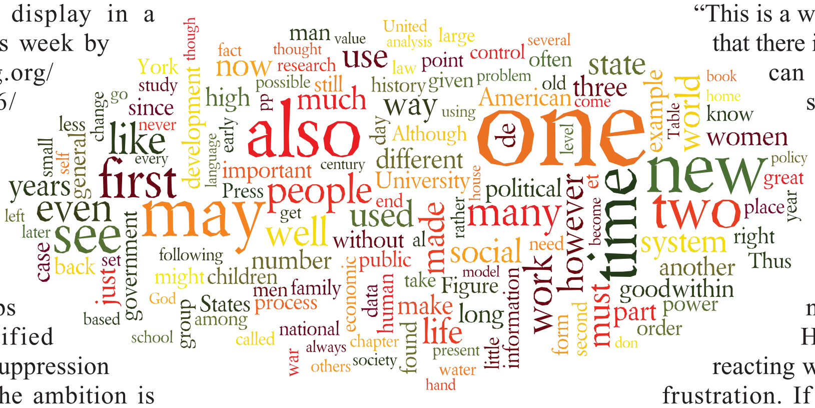
By analyzing the growth, change, and decline of published words over the centuries, the mathematician argued, it should be possible to rigorously study the evolution of culture on a grand scale. "I didn't think the idea was crazy," recalls Norvig. "We were doing the scanning anyway, so we would have the data."

The first explorations of the Google Books data are now on display in a study published online this week by *Science* (www.sciencemag.org/content/early/2010/12/16/science.1199644.abstract). The researchers have revealed 500,000 English words missed by all dictionaries, tracked the rise and fall of ideologies and famous people, and, perhaps most provocatively, identified possible cases of political suppression unknown to historians. "The ambition is enormous," says Nicholas Dames, a literary scholar at Columbia University.

The project almost didn't get off the ground because of the legal uncertainty surrounding Google Books. Most of its content is protected by copyright, and the entire project is currently under attack by a class action lawsuit from book publishers and authors. Norvig admits he had concerns about the legality of sharing the digital books, which cannot be distributed without compensating the authors. But Lieberman Aiden had an idea. By converting the text of the scanned books into a single, massive "n-gram" database—a map of the context and frequency of words across history—scholars could do quantitative research on the tomes without actually reading them. That was enough to persuade Norvig.

Lieberman Aiden teamed up with fellow Harvard Ph.D. student Jean-Baptiste Michel. The pair were already exploring ways to study written language with mathematical techniques borrowed from evolutionary biology. Their 2007 study of the evolution of English verbs, for example, made the cover of *Nature*. But they had never contended with the amount of data that Google Books offered. It currently includes 2 trillion words from 15 million books, about 12% of every book in every language published since the Gutenberg Bible in 1450. By comparison, the human genome is a mere 3-billion-letter poem.

Michel took on the task of creating the software tools to explore the data. For the analysis, they pulled in a dozen more researchers, including Harvard linguist Steven Pinker. The first surprise, says Pinker, is that books contain "a huge amount of lexical dark matter." Even after excluding proper nouns, more than 50% of the words in the n-gram database do not appear in any published dictionary. Widely used words such as "deletable" and obscure ones like "slenthem" (a type of musical instrument) slipped below the radar of standard references. By the research team's estimate, the size of the English language has nearly doubled over the past century, to more than 1 million words. And vocabulary seems to be growing faster now than ever before.



It was also possible to measure the cultural influence of individual people across the centuries. For example, notes Pinker, tracking the ebb and flow of "Sigmund Freud" and "Charles Darwin" reveals an ongoing intellectual shift: Freud has been losing ground, and Darwin finally overtook him in 2005.

Analysis of the n-gram database can also reveal patterns that have escaped the attention of historians. Aviva Presser Aiden led an analysis of the names of people that appear in German books in the first half of the 20th century. (She is a medical student at Harvard and the wife of Erez Lieberman Aiden.) A large number of artists and academics of this era are known to have been censored during the Nazi period, for being either Jewish or "degenerate," such as the painter Pablo Picasso. Indeed, the n-gram trace of their names in the German corpus plummets during that period, while it remains steady in the English corpus.

Once the researchers had identified this signature of political suppression, they analyzed the "fame trace" of all people mentioned in German books across the same period, ranking them with a "suppression index." They sent a sample of those names to a historian in Israel for validation. Over 80% of the people identified by the suppression index are known to have been censored—for example, because their names were on blacklists—proving that the technique works. But more intriguing, there is now a list of people who may have been victims of suppression unknown to history.

"This is a wake-up call to the humanities that there is a new style of research that can complement the traditional styles," says Jon Orwant, a computer scientist and director of digital humanities initiatives at Google. In a nod to data-intensive genomics, Michel and Lieberman Aiden call this nascent field "culturomics."

Humanities scholars are reacting with a mix of excitement and frustration. If the available tools can be expanded beyond word frequency, "it could become extremely useful," says Geoffrey Nunberg, a linguist at the University of California, Berkeley. "But calling it 'culturomics' is arrogant." Nunberg dismisses most of the study's analyses as "almost embarrassingly crude."

Although he applauds the current study, Dames has a score of other analyses he would like to perform on the Google Books corpus that are not yet possible with the n-gram database. For example, a search of the words in the vicinity of "God" could reveal "semantic shifts" over history, Dames says. But the current database only reveals the five-word neighborhood around any given term.

Orwant says that both the available data and analytical tools will expand: "We're going to make this as open-source as possible." With the study's publication, Google is releasing the n-gram database for public use. The current version is available at www.culturomics.org.

–JOHN BOHANNON

CREDITS: J. B. MICHEL *ET AL.*; WORDLE.COM