Published online 16 December 2010 | Nature | doi:10.1038/news.2010.677

News

Cultural goldmine lurks in digitized books

'Culturomics' uncovers fame, fortune and censorship from more than a century of words.

Philip Ball

Analysing decades of books can reveal important cultural trends.

FRANCK CAMHI / Alamy

The digitization of books by Google Books has sparked controversy over issues of copyright and book sales, but for linguists and cultural historians this vast project could offer an unprecedented treasure trove. In a paper published today in *Science*[1], researchers at Harvard University in Cambridge, Massachusetts, and the Google Books team in Mountain View, California, herald a new discipline called culturomics, which sifts through this literary bounty for insights into trends in what cultures can and will talk about through the written word.

Among the findings described by the collaboration, led by Jean-Baptiste Michel, a Harvard biologist, are the size of the English language (around a million words in 2000), the typical 'fame trajectories' of well-known people, and the literary signatures of censorship such as that imposed by Germany's Nazi government.

"The possibilities with such a new database, and the ability to analyse it in real time, are really exciting," says linguist Sheila Embleton of York University in Toronto, Canada.

"Quantitative analysis of this kind can reveal patterns of language usage and of the salience of a subject matter to a degree that would be impossible by other means," agrees historian Patricia Hudson of Cardiff University, UK.

"The really great aspect of all this is using huge databases, but they will have to be used in careful ways, especially considering alternative explanations and teasing out the differences in alternatives from the database," adds Royal Skousen, a linguist at Brigham Young University in Provo, Utah. "I do not like the term 'culturomics'," he adds. "It smacks too much of 'freakonomics', and both terms smack of amateur sociology."

Half a trillion words

Using statistical and computational techniques to analyse vast quantities of data in historical and linguistic research is nothing new — the fields known as quantitative history and quantitative linguistics already do this. But it is the sheer volume of the database created by Google Books that sets the new work apart.

"The possibilities with such a new database are really exciting."

So far, Google has digitized more than 15 million books, representing about 12% of all those ever published in all languages. Michel and his colleagues performed their analyses on just a third of this sample, selected for the quality of the optical character recognition in the digitization and the

reliability of information about a book's provenance, such as the date and place of publication.

The resulting data set contained over 500 billion words. This is far more than any single person could read: a fast reader would, without breaks for food and sleep, need 80 years to finish the books for the year 2000 alone.

Not all isolated strings of characters in texts are real words. Some are numbers, abbreviations or typos. In fact, 51% of the character strings in 1900, and 31% in 2000, were 'non-words'. "I really have trouble believing that," admits Embleton. "If it's true, it would really shake some of my foundational thoughts about English."

According to this account, the English language has grown by more than 70% during the past 50 years, and around 8,500 new words are being added each year. Moreover, only about half of the words currently in use are apparently documented in standard dictionaries. "That high amount of lexical 'dark matter' is also very hard to believe, and would also shake some foundations," says Embleton. "I'd love to see the data."

In principle she already can, because the researchers have made their database public at **www.culturomics.org**. This will allow others to explore the huge number of potential questions it suggests, not just about word use but about cultural history. Michel and colleagues offer two such examples, concerned with fame and censorship.

They say that actors reach their peak of fame, as recorded in references to names, around the age of 30, while writers take a decade longer but achieve a higher peak. "Science is a poor route to fame," they add. Physicists and biologists who achieve fame do so only late in life, and "even at their peak, mathematicians tend not to be appreciated by the public".

Big Brother's fingerprints

Nation-specific subsets of the data can show how references to ideas, events or people can drop out of sight because of state suppression. For example, the Jewish artist Marc Chagall virtually disappears from German writings in 1936-1944 (while remaining prominent in English-language books), and 'Trotsky' and 'Tiananmen Square' similarly vanish at certain sensitive points in time from Russian and Chinese works respectively. The authors also look at trends in references to feminism, God, diet and evolution.

"The ability, via modern technology, to look at just so much at once really opens horizons," says Embleton. However, Hudson cautions that making effective use of such a resource will require skill and judgement, not just number-crunching.

"How this quantitative evidence is generated and how it is interpreted are the most important factors in forming conclusions," she says. "Quantitative evidence of this kind must always address suitably framed general questions, and employed alongside qualitative evidence and reasoning, or it will not be worth a great deal."

References

1. Michel, J.-B. *et al. Science* advance online publication. **doi:10.1126/science.1199644** (2010).

Comments

*If you find something abusive or inappropriate or which does not otherwise comply with our **Terms** or **Community Guidelines**, please select the relevant 'Report this comment' link.*

Comments on this thread are vetted after posting.

Now **I** finally get to see what this app is all about! **#60768**

Report this comment Posted by: **Jackson Mary** | 2013-11-04 04:14:06 AM

Michel and his colleagues performed their analyses on just a third of this sample, selected for the good quality of the optical character recognition in the digitization and the reliability of information about a book's provenance, such as the date and place of publication **Fix Them or Fire Them** **#61020**

Report this comment Posted by: **Steven Shaer** | 2013-11-11 03:46:45 AM

Commenting is now closed.

Nature ISSN 0028-0836 EISSN 1476-4687