
THE WALL STREET JOURNAL

WSJ.com

BOOKS | March 16, 2012, 6:43 p.m. ET

The New Science of the Birth and Death of Words

Have physicists discovered the evolutionary laws of language in Google's library?

By CHRISTOPHER SHEA

Can physicists produce insights about language that have eluded linguists and English professors? That possibility was put to the test this week when a team of physicists published a paper drawing on Google's massive collection of scanned books. They claim to have identified universal laws governing the birth, life course and death of words.

The paper marks an advance in a new field dubbed "Culturomics": the application of data-crunching to subjects typically considered part of the humanities. Last year a group of social scientists and evolutionary theorists, plus the Google Books team, showed off the kinds of things that could be done with Google's data, which include the contents of five-million-plus books, dating back to 1800.



Getty Images

Higher death rates for words, the authors of a new study say, are largely a matter of homogenization.

Published in *Science*, that paper gave the best-yet estimate of the true number of words in English—a million, far more than any dictionary has recorded (the 2002 Webster's Third New International Dictionary has 348,000). More than half of the language, the authors wrote, is "dark matter" that has evaded standard dictionaries.

The paper also tracked word usage through time (each year, for instance, 1% of the world's English-speaking population switches from "sneaked" to "snuck"). It also showed that we seem to be putting history behind us more quickly, judging by the speed with which terms fall out of use. References to the year "1880" dropped by half in the 32 years after that date, while the half-life of "1973" was a mere decade.

In the new paper, Alexander Petersen, Joel Tenenbaum and their co-authors looked at the ebb and flow of word usage across various fields. "All these different words are battling it out against synonyms, variant spellings and related words," says Mr. Tenenbaum. "It's an inherently competitive, evolutionary environment."

When the scientists analyzed the data, they found striking patterns not just in English but also in Spanish and Hebrew. There has been, the authors say, a "dramatic shift in the birth rate and death rates of words": Deaths have increased and births have slowed.

English continues to grow—the 2011 Culturomics paper suggested a rate of 8,500 new words a year. The new paper, however, says that the growth rate is slowing. Partly because the language is already so rich, the

"marginal utility" of new words is declining: Existing things are already well described. This led them to a related finding: The words that manage to be born now become more popular than new words used to get, possibly because they describe something genuinely new (think "iPod," "Internet," "Twitter").

Higher death rates for words, the authors say, are largely a matter of homogenization. The explorer William Clark (of Lewis & Clark) spelled "Sioux" 27 different ways in his journals ("Sieoux," "Seaux," "Souixx," etc.), and several of those variants would have made it into 19th-century books. Today spell-checking programs and vigilant copy editors choke off such chaotic variety much more quickly, in effect speeding up the natural selection of words. (The database does not include the world of text- and Twitter-speak, so some of the verbal chaos may just have shifted online.)

Synonyms also fight Darwinian battles. In one chart, the authors document that "Roentgenogram" was by far the most popular term for "X-ray" (or "radiogram," another contender) for much of the 20th century, but it began a steep decline in 1960 and is now dead. ("Death," in language, is not as final as with humans: It refers to extreme rarity.) "Loanmoneys" died circa 1950, killed off by "loans." "Persistency" today is breathing its last, defeated in the race for survival by "persistence."

The authors even identified a universal "tipping point" in the life cycle of new words: Roughly 30 to 50 years after their birth, they either enter the long-term lexicon or tumble off a cliff into disuse. The authors suggest that this may be because that stretch of decades marks the point when dictionary makers approve or disapprove new candidates for inclusion. Or perhaps it's generational turnover: Children accept or reject their parents' coinages.

The similar trajectory of word birth and death across time in three languages is "striking," though the research is still too new to evaluate fully, says Mark Liberman, a professor of linguistics at the University of Pennsylvania. Among the questions raised by critics: Since older books are harder to scan, how much of the word "death" is simply the disappearance of words garbled by the Google process itself?

In the end, words and sentences aren't atoms and molecules, even if they can be fodder for the same formulas.

A version of this article appeared March 17, 2012, on page C3 in the U.S. edition of The Wall Street Journal, with the headline: The New Science of the Birth and Death of Words.

Copyright 2012 Dow Jones & Company, Inc. All Rights Reserved

This copy is for your personal, non-commercial use only. Distribution and use of this material are governed by our [Subscriber Agreement](#) and by copyright law. For non-personal use or to order multiple copies, please contact Dow Jones Reprints at 1-800-843-0008 or visit www.djreprints.com

Don't Miss...

Powered by Taboola



**Art Linkletter
Estate Sells for
\$10.3M**
00:37



**Power Suits Get a
More Feminine
Look**
02:23