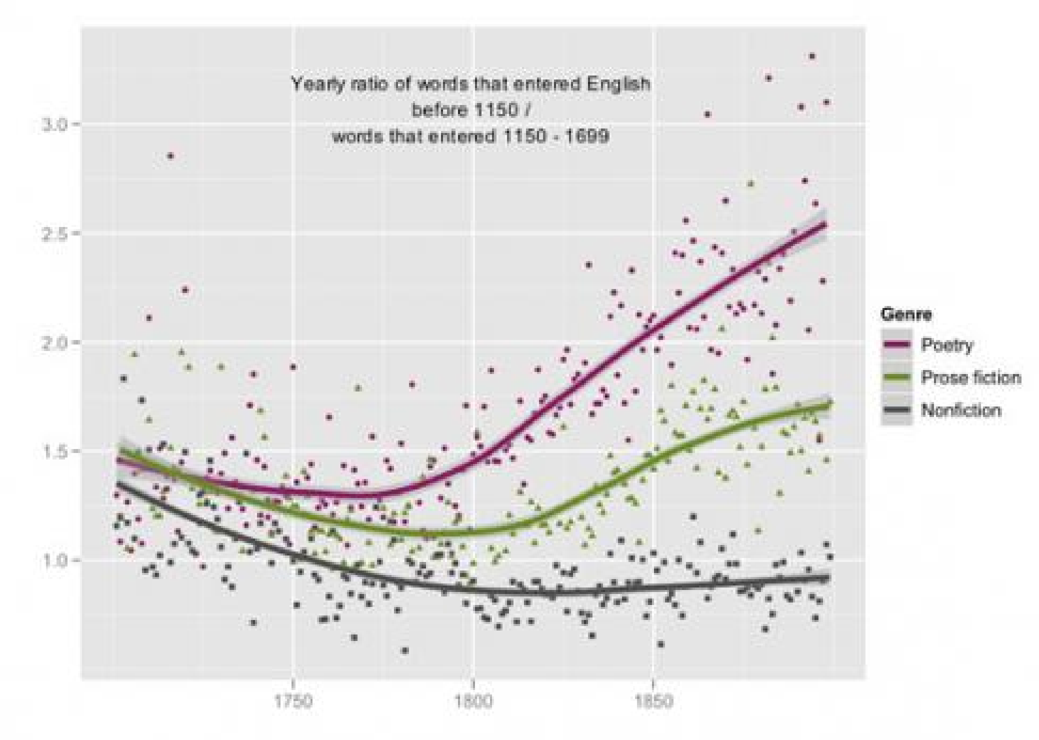
PHYS.ORG

Exhaustive computer research project shows shift in English language

May 16 2013, by Dusty Rhodes



This graph shows a sharp increase in the use of "informal" Old English words in literary writing after 1800. Credit: Ted Underwood

University of Illinois English professor Ted Underwood recently wrapped up a research project involving more than 4,200 books. Since that work revealed dramatic shifts in the English language between the 18th and 19th centuries, he's now expanding his research to include more than 470,000 books – almost every English language book written during that era and preserved in a university library.

How did he find time to read 4,000 books, let alone 400,000? He didn't, of course. Underwood, who teaches 18th- and 19th-century literature, worked with the U. of I.'s Institute for Computing in Humanities, Arts and Social Science (I-CHASS) and the HathiTrust Research Center (a collaboration of the U. of I. and Indiana University) to develop computer programs to crawl through digitized copies of the books, counting words and sorting genres.

Underwood's data mining venture has already yielded some gems. "Increasingly, I'm finding that there are big patterns to be discovered in literary history at that scale," he said. "We just hadn't been able to back up far enough to see it."

He initially set out simply to confirm his hunch that the English language acquired a bit of starchiness around 1800. "There's a very Latinate diction that sets in around that time," he said. "I had a vague sense that written English became more formal. For example, you no longer 'need' something; you 'require' it."

He used data from Google Books to find the 10,000 words used most frequently in 18th- and 19th-century books (not counting determiners, prepositions, conjunctions and pronouns). By using a "Web-scraper" and dictionary.com, he traced the etymology of each of those 10,000 words, sorted them by the date of entry into the English language, and divided them into two groups – pre- and post-Norman Conquest. The 1066 invasion made French the official language of Britain, used (along with Latin) for all written business and spoken by the ruling class. The Old English vernacular spoken by the lower class survived, but by the time English regained its status as the official language, some 200 years later, the vocabulary of business and government had been lost, replaced by French and Latin words. "So that's sort of a rough-and-ready way of assessing formality," Underwood said.

Using that 12th-century dividing line, Underwood plotted the ratio of formal to informal words used in books each year from 1700 to 1899. The resulting graph confirmed his assumption, but just barely. "You did get more Latinate words around 1800," he said, "but it was a pretty slight effect."

When he sorted the books by genre, however, the result was surprisingly profound. While the language of nonfiction works indeed became more formal, the language of fiction, drama and especially poetry became more formal until about 1775, and then reversed course and increasingly relied on less-formal, pre-Norman Conquest words. By 1899, Old English words occurred in drama and fiction at a rate more than 1 ½ times higher than in nonfiction, according to Underwood's data. In poetry, the rate was almost three times higher.

This large shift coincides with the advent of the style of writing we now think of as "literature" – writing that's set apart by its imaginative intention. "If you go back to 1700, literature basically means anything written – just literacy, actually. Our concept of literature as fictive and aesthetic really emerges in the late 18th century," Underwood said. "I think the most important result I stumbled upon was how literary genres changed from non-literary prose."

In an article titled "The Emergence of Literary Diction," published in the *Journal of Digital Humanities*, Underwood and co-author Jordan Sellers, a graduate student at the U. of I., explained their discovery and theorized about why literature – considered to be the most artful form of writing – adopted the simplest vocabulary.

"In a sense, poetry became more specialized than it had been before: Its diction became more remote from prose. But it specialized in the direction of old words that would appear plain, common and universal," they wrote.

Underwood suspects that "literature" as we know it evolved into a vehicle for expressing individual experience, leaving nonfiction writers to analyze abstract ideas and social structures. "In my view, the decreasing formality of literary language was a side effect of this emphasis on the elementary and the personal," he said.

This research will be included in his book, "Why Literary Periods Mattered: Historical Contrast and the Prestige of English Studies," to be published this summer by Stanford University Press. This research recently won him an $85,000 Digital Innovation fellowship from the American Council of Learned Societies and a $57,000 digital humanities startup grant from the National Endowment for the Humanities, which he is using to expand his research to 470,000 volumes.

One of the first steps involves categorizing each of those volumes by genre – a task that's trickier than it sounds. "Like Aristotle's 'Poetics' – it's about poetry, but it's not poetry," Underwood said. He has already crafted a program to automatically classify texts based on word patterns (similar to how a spam filter works on email), but he is constantly tweaking and refining the process.

"Genres change across time, and genre boundaries are fuzzy. So humanists are rightly wary of crisp computational solutions to problems that are really fuzzy," he said. "We can build in and acknowledge some of that fuzziness."

Using computer programs to analyze creative writing is a branch of digital humanities – a field so new that it's still considered controversial in certain circles. "In the academy, I think it's viewed with a mixture of excitement and apprehension," Underwood said. However, Underwood's father is in computer science, and Underwood spent summers during his undergraduate years working for his dad, writing computer programs. Combining that skill with his passion for literature comes naturally to the English professor, but he realizes it may not come so easily to his colleagues.

"What I'm trying to do now is create tools that will make it easier for other researchers to use this bigger collection," he said.

He shares his data, his processes and his programs through his blog, The Stone and the Shell (tedunderwood.com), which takes its name from Wordworth's epic poem, "The Prelude," in which a shell seems to represent poetry, and a stone represents mathematics.

More information: journalofdigitalhumanities.org … -and-jordan-sellers/

Provided by University of Illinois at Urbana-Champaign

Citation: Exhaustive computer research project shows shift in English language (2013, May 16) retrieved 11 May 2024 from https://phys.org/news/2013-05-exhaustive-shift-english-language.html