O fractions, my fractions!

Machines are getting better at literary analysis

Algorithms that identify the voices of authors and characters should be celebrated, not scorned

Alamy

Prospero Mar 8th 2018 | by J.T.

IN "Dead Poets Society" (1989), John Keating, a teacher at a 1950s American boarding school, played by Robin Williams, draws a chart, its shape dictated by a fictional essay called "Understanding Poetry". The horizontal axis measures a poem's technical quality, the vertical axis shows its importance, and the combination of the two determines its greatness. After allowing his pupils to draw such a chart for Lord Byron and William Shakespeare, Mr Keating declares the essay "excrement", and orders them to rip it out of their poetry anthologies. "This is a battle, a war, and the casualties could be your hearts and souls," he rumbles. There are "armies of academics going forward measuring poetry", with little regard for passion, beauty or romance.

Doubtless Mr Keating would have been dismayed to read "The Transformation of Gender in English-Language Fiction (http://culturalanalytics.org/2018/02/the-transformation-of-gender-in-english-language-fiction/) ", a paper published last month in the Journal of Cultural Analytics. The authors—Ted Underwood and Sabrina Lee of the University of Illinois, and David Bamman of the University of California, Berkeley—have trained a series of machine-learning models on a broad corpus of 104,000 works of fiction written between 1700 and 2010. The database, which the academics compiled from the HathiTrust Digital Library and the Chicago Novel Corpus, is enormous but not exhaustive. It contains almost all classic novels, but only about half of the books that have been listed in *Publishers Weekly*, an American trade magazine. Nonetheless, the authors believe that it is a reasonable representation of the overall market for fiction, since the historical share of female authors is similar to that in *Publishers Weekly*. The algorithms they have trained on the data have allowed them to explore a range of gendered issues (see chart).

Latest stories

Can Donald Trump sack Robert Mueller?
THE ECONOMIST EXPLAINS

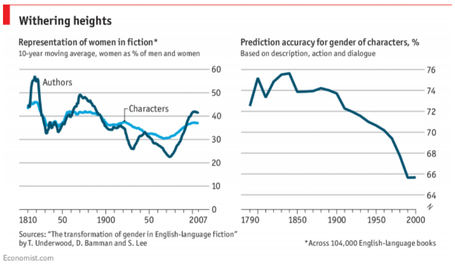
American corporations come out against Donald Trump's proposed tariffs
GRAPHIC DETAIL

Dogs and cats may have been involved in Maya rituals
SCIENCE AND TECHNOLOGY

See more

One model identifies an author's gender, and finds that the share of books written by women fell from about half at the start of the 19th century to less than a quarter in the 1960s, followed by a rebound to roughly 40% today. A second model identifies the gender of characters via their names and pronouns, with more than 90% accuracy, and shows a similar trend: the share of the narrative given to fictional women declined over 150 years, before recovering slightly. A third model tries to determine a character's gender based only on the language used in descriptions, actions and dialogue. Such predictions were right 75% of the time in 1800 but just 65% of the time in 2000, suggesting that the fictional women and men are behaving in less stereotypable ways.

Mr Keating would have called such research piffle. He taught that the purpose of reading is to feel, "to savour words and language": medicine, law, business and engineering are noble pursuits that keep us alive, but literature stirs the emotions that make life worth living. Yet in a 3,500-word essay


Withering heights

Sources: "The transformation of gender in English-language fiction" by T. Underwood, D. Bamman and S. Lee

*Across 104,000 English-language books

Economist.com

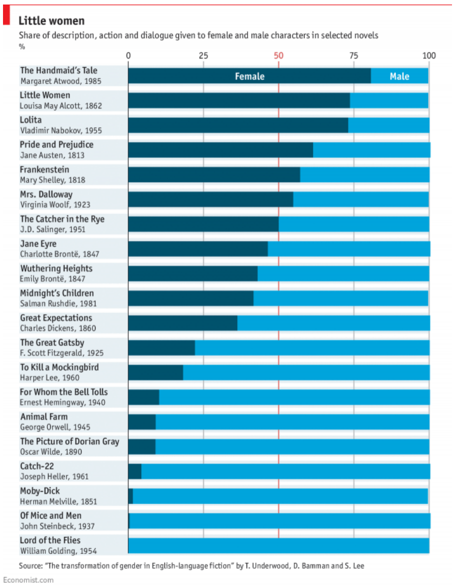
(https://www.theatlantic.com/education/archive/2014/02/-em-dead-poets-society-em-is-a-terrible-defense-of-the-humanities/283853/) for *The Atlantic* in 2014, Kevin Dettmar, a professor of English at Pomona College, criticised the film's anti-intellectualism. He argued that defending literature purely for its sentimental value encourages the belief that "the humanities is easy, a soft option; that the humanities doesn't train thinkers".

Both are partly right. Great literature can move readers in a way that few other academic subjects can. It can also, when perused critically, stimulate reasoning, empathy and debate. Neurologists have struggled to prove (https://www.thecut.com/2015/09/does-literature-really-beef-up-your-brain.html) that reading fiction actually improves those functions, but they have shown that interrogating a text activates the relevant parts of the brain. For those who believe that studying literature critically is worthwhile, the lessons that can be gleaned from big data and machine learning are valuable.

Take the gender of authors. This ought to be one of the most basic questions for literary scholars to answer: has fiction become more or less dominated by men? Before the advent of digital humanities, a field which applies computer science to the arts, the response could only be subjective or based on small samples. "The Transformation of Gender" provides an objective answer that would surprise many people and should provoke more research. The post-1960s rebound in female authorship, for instance, could have many causes. Ms Lee notes that it follows the rise of the paperback novel and coincides with a proliferation in romance imprints.

Readers will also be intrigued by charts showing how the language used to depict men and women has changed. "Heart", "mind" and "spirits" were once strongly feminine but have now become neutral, while "house" has switched from landed male owners to domestic female occupants. Yet Mr Bamman argues that the most promising product of such research is an elementary one: a machine's ability to identify literary characters. E. M. Forster, a British novelist, described the people in a story as "word-masses", made up simply of description, action and dialogue. It is now possible for an algorithm to ingest a text, identify the subjects of each word using context, and split them into these masses. Indeed, one of the techniques used in the paper is known as the "bag-of-words model".

Mr Underwood notes that the algorithms are far from perfect. Though they can be used to examine individual books (see chart), they also make mistakes, especially when a first-person narrator is framing the story. Across a wider sample, however, they can be deployed more confidently. A paper published by Mr Bamman in 2013 was able to identify character stereotypes from 42,000 Wikipedia film summaries, which clustered Batman with Jason Bourne and the Joker with Dracula. A follow-up in 2014 confirmed various literary theories about the similarities between characters in the novels of Charles Dickens and Jane Austen, among other writers.



Little women

Share of description, action and dialogue given to female and male characters in selected novels

Source: "The transformation of gender in English-language fiction" by T. Underwood, D. Bamman and S. Lee

Economist.com

The latter study was also able to separate the author's voice—that is, the mannerisms that make each writer unique—from those of the characters, who have their own quirks. Mr Bamman explains that identifying individual people might also help algorithms to understand plot, since a sudden change in personnel usually indicates a change of scene. The ability to isolate these formal elements of writing and compare them across a vast body of work is being harnessed by other scholars, too. The latest edition of the "New Oxford Shakespeare" has claimed that 17 of the bard's 44 plays were produced collaboratively (https://www.economist.com/blogs/prospero/2017/03/revenge-maths-mob) , based on an analysis of how his contemporaries used "function words" like "and" or "with".

Such author-attribution has been used since the 1950s, when two statisticians (with no background in history) proved that 12 essays from "The Federalist Papers", claimed by both Alexander Hamilton and James Madison, were far more like Madison's in style. Looking at those function words (like "while" versus "whilst", or "among" versus "between") was more definitive than examining the ideas in the essays. But computers and digital corpora make this far faster today: Ben Blatt adopted these techniques for many clever experiments in "Nabokov's Favorite Word is Mauve", his book from 2017.

Artificial intelligence is still a long way from being able to write new arguments coherently, as we discovered when we recently attempted to automate an article for our Science and Technology section (https://www.economist.com/news/science-and-technology/21732805-weve-got-few-years-left-least-how-soon-will-computers-replace-economists) . When it comes to metaphor and allusion, humans will always believe that they have the upper hand. But it would be folly to ignore the help that machine learning can offer to those seeking empirical answers to literary questions. These techniques can enrich readers' understanding of the books they love, without quelling their enthusiasm. To borrow another line of Mr Keating's, as he encourages his students to stand on their desks: "We must constantly look at things in a different way."