Belles lettres Meets Big Data

Quantitative analysis of poetry and prose has roots deep in the 19th century.

Brian Hayes

The literary scholar needs a quiet room, a reading lamp, a notebook, a receptive mind—and algorithms for *n*-gram analysis, part-of-speech tagging, word-sense disambiguation, and sentence parsing. "Digital humanities" is all the rage these days in English departments. Recent meetings of the Modern Language Association have had dozens of sessions on the theme. Franco Moretti, a professor of English at Stanford University, insists that the tradition of "close reading"—giving careful attention to every word of a few canonical texts—must give way to "distant reading," where whole genres are subjected to quantitative analysis in bulk. According to the publisher of his books, "Moretti argues that literature scholars should stop reading books and start counting, graphing, and mapping them instead."

The idea of applying mathematical and computational tools to literature is hardly new. The first conference on "literary data processing" was held in 1964; it attracted 150 participants. The topics discussed included "computational stylistics" and a computer-aided assessment of John Milton's influence on Percy Bysshe Shelley. These were not the first such projects. Frederick Mosteller and David L. Wallace had already applied statistical methods to a case of disputed authorship in American history. They tabulated the frequencies of common words (*also, an, by, of,* etc.) in the Federalist Papers, seeking to determine which of those essays were written by Alexander Hamilton and which by James Madison. Earlier still—and without computer assistance—the British statisticians G. Udny Yule and C. B. Williams had studied variations in sentence length as a way of characterizing literary style and identifying authors.

I have lately become intrigued by two more pioneers of the digital humanities, who worked in an even earlier era, when *digital* could only refer to fingers, not computer chips. Both of these scholars were Americans born in the middle of the 19th century. One was a man of science who made a few brief forays into statistical language studies. The other was a professor of English literature who yearned to import scientific methods into his field.

Literary Spectroscopy

Thomas Corwin Mendenhall (1841–1924) grew up in rural eastern Ohio with little schooling, but he went on to quite a cosmopolitan career in science and pedagogy. In the 1870s he was one of the first faculty members of the new Ohio Agricultural and Mechanical College in Columbus. Then he went off to teach in Tokyo for a few years; by the time he came back, the Agricultural and Mechanical College had become the Ohio State University. Later, Mendenhall became president of the Rose Polytechnic Institute (now Rose-Hulman) in Terre Haute, Indiana; he was also president of Worcester Polytechnic Institute in Massachusetts and superintendent of the U.S. Coast and Geodetic Survey. He was a member of Sigma Xi, was elected to the National Academy of Sciences, and served a term as president of the American Association for the Advancement of Science. After retiring at age 60 he spent 10 years roaming Europe and Asia, followed by a return to small-town Ohio.

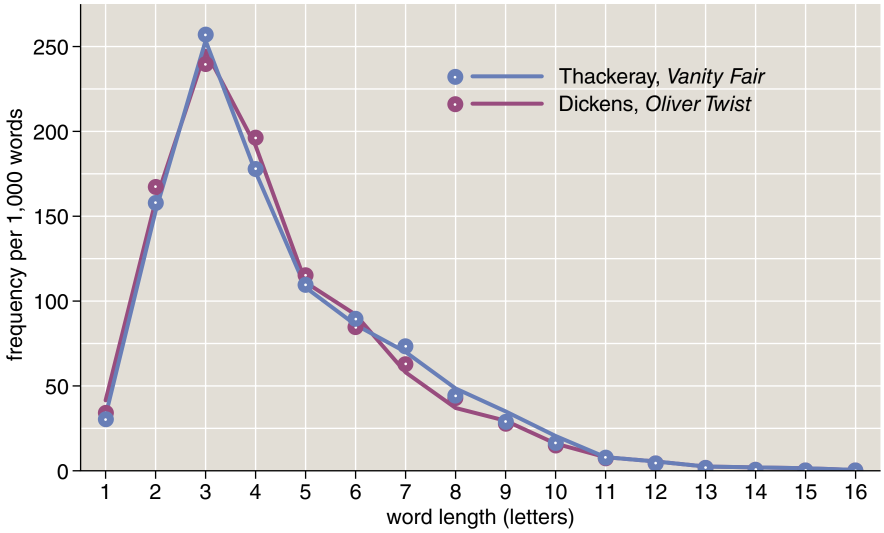
Mendenhall's scientific interests ranged from electrical machinery to geodesy to spectroscopy. The last of these topics provided a metaphorical context for his literary ventures. In a paper titled "The Characteristic Curves of Composition," published in *Science* in 1887, he remarked that the pattern of lines in a spectrum offers "indisputable evidence" for the presence of a chemical element.

> In a manner very similar, it is proposed to analyze a composition by forming what may be called a "word-spectrum," or "characteristic curve," which shall be a graphic representation of an arrangement of words according to their length and to the relative frequency of their occurrence.

Mendenhall's method was to select blocks of 1,000 words from a text, then record how many words in each block are of length one letter, two letters, three letters, and so on. He hoped to show that the resulting "spectrum" could serve as a reliable marker of authorial identity: The curve would be similar across all works by the same author, he thought, and different in works by different authors. He tested this hypothesis on novels by Charles Dickens (*Oliver Twist*) and William Makepeace Thackeray (*Vanity Fair*). Results based on 10,000 words from each novel are shown in the illustration at the top of the opposite page. Are the curves distinctive enough to serve as author fingerprints? Mendenhall concedes that the outcome is inconclusive, suggesting the need for more data.

Brian Hayes is senior writer for American Scientist. *Additional material related to the* Computing Science *column can be found online at http://bit-player.org. E-mail: brian@bit-player.org*

In preparing his 1887 paper, Mendenhall tallied the lengths of at least 30,000 words, and he recruited friends to count more. Out of curiosity, I tried hand-tabulating the lengths of the first 1,000 words of *Oliver Twist*. The task took more than an hour, and I made a dozen mistakes. How different the process with a computer—and with access to an archive of digitized texts, such as Project Gutenberg. In milliseconds, all the words in a huge, sternum-crushing Victorian novel are rendered into a table of a dozen or so numbers. Even with a complete inventory of word lengths, however, it's not clear that the spectral curves reliably discriminate between Dickens and Thackeray.



The "spectra" of word lengths in two Victorian novels was plotted in 1887 by Thomas Mendenhall, who hoped that such curves would serve as a fingerprinting device for identifying authors. The continuous curves are redrawn from Mendenhall's published graph, based on a sample of 10,000 words each selected from William Makepeace Thackeray's *Vanity Fair* and Charles Dickens's *Oliver Twist*. The dots show the word counts for the complete novels.

Shakespeare's Spectrum

Mendenhall published nothing more on word-length studies until 1901. Writing then in *Popular Science*, he sheepishly admitted he'd had an ulterior motive all along: To show that Francis Bacon wrote the plays of Shakespeare. For work on this famously vexed and vexing question, Mendenhall was able to secure funding. Augustus Hemenway, a Boston philanthropist (and Baconian partisan) agreed to pay the salaries of two Worcester women hired as letter counters, as well as the cost of building a special tabulating machine. The nature of the machine is not explained in detail, but it had a button for each possible number of letters in a word.

> One of the counters, with book in hand, called off "five," "two," "three," etc., as rapidly as possible, counting the letters in each word carefully and taking the words in their consecutive order, the other registering, as called, by pressing the proper buttons.

The two Worcester counters tallied 400,000 words of Shakespeare, 200,000 words of Bacon, and works of various other Elizabethan authors. They soon made a curious discovery: Whereas the word spectrum of most authors writing in English has its highest peak at words of three letters, Shakespeare exhibits an exceptional fondness for four-letter words. A computer survey of the complete works of Shakespeare confirms this observation. *(See illustration at right.)*

A second discovery must have been disappointing to Hemenway and Mendenhall: Bacon's word-length spectrum looks quite different from Shakespeare's, with a more conventional peak at three letters. Perhaps there was some consolation in seeing the spectral method itself vindicated, in that the two authors are clearly distinguished by their word-length curves.

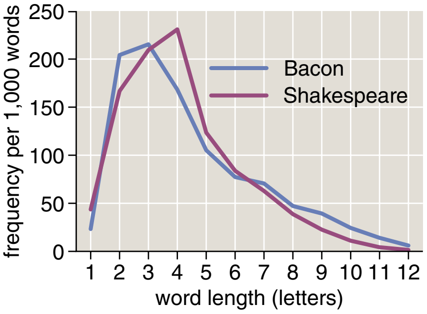
The Shrinking English Sentence

My second precocious digital humanist is Lucius Adelno Sherman (1847–1933), who was born and educated in New England. After graduating from Yale in 1871, he stayed in New Haven to teach Greek, Latin, and English at the Hopkins Grammar School. He also published a translation of the Swedish poem *Frithiof's Saga*, and worked toward a Yale Ph.D., which was awarded in 1875. His doctoral dissertation, "A Grammatical Analysis of the Old English Poem, 'The Owl and the Nightingale,'" already suggests a quantitative bent to his literary work. For example, he lists the frequencies of various prepositions, conjunctions, and forms of negation in the poem.

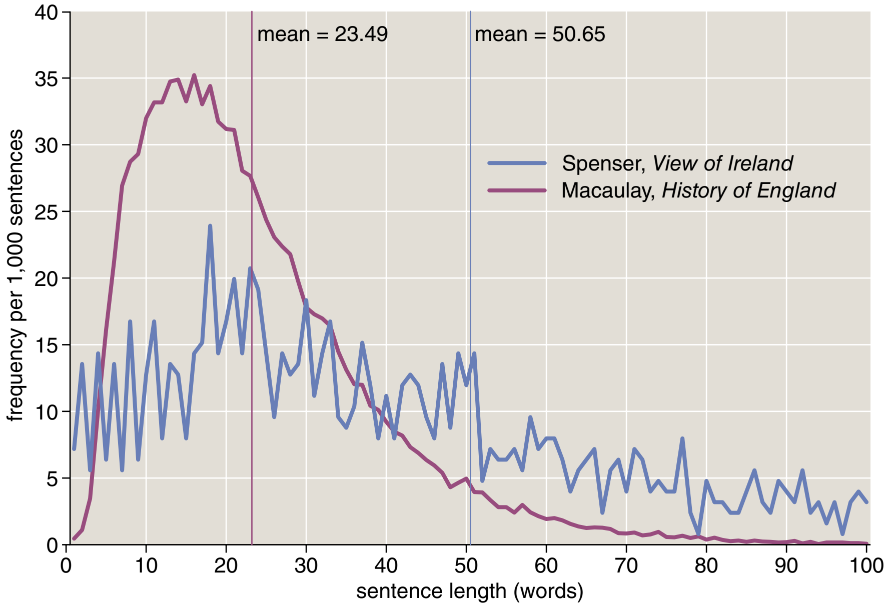
In 1882 Sherman accepted an appointment in the English department at the University of Nebraska in Lincoln. At the time, the Lincoln campus consisted of a single building, but both the city and the university were growing at a frenzied pace and would soon become a cultural capital of the prairies. Sherman remained in Lincoln through the rest of his life and career, serving in due course as department chair and dean. He published extensively on Shakespeare and other Elizabethan dramatists, and wrote plays of his own. And then there was the peculiar book that concerns me here: *Analytics of Literature: A Manual for the Objective Study of English Prose and Poetry*, published in 1893.

The first half of *Analytics* is a fairly conventional introduction to rhetoric and poetics, with chapters on meter and rhyme, figures of speech, the emotional force of words—a lot of close reading. Then Sherman suddenly goes all quantitative, launching into a discussion of sentence length.

Sherman was motivated by broader questions than the authorship puzzles that concerned Mendenhall. While teaching the historical development of English literature, Sherman took note of pervasive changes in sentence structure. In the chronological progression from Chaucer in the 14th century to



Word-length spectra for Francis Bacon and William Shakespeare fail to support the notion that Bacon wrote Shakespeare's plays—to the disappointment of Mendenhall.



Historical changes in the distribution of sentence lengths were noted in the 1890s by Lucius A. Sherman. Edmund Spenser, writing at the end of the 16th century, averaged 50 words a sentence, and the distribution of lengths was very broad. Thomas Macaulay, in the middle of the 19th century, had a narrower distribution and an average of just 23 words per sentence. The blue curve is more jagged because it is based on a smaller sample: 1,254 sentences versus 41,371.

Shakespeare in the 17th to Emerson in the 19th, sentences seemed to grow simpler, to lose much of their "heaviness" and intricacy. Of course Sherman was hardly the first to notice that modern English syntax differs from medieval and Elizabethan practice, but his approach was novel: He believed the nature of the change should be susceptible to scientific inquiry. "The right way and the only way to learn the facts and principles of English prose development was plainly to study the literature objectively, with scalpel and microscope in hand."

An obvious way to begin this inquiry was simply to measure the lengths of sentences, and thus Sherman undertook a great counting project. By experiment he found that a sample of 500 sentences was enough to characterize an author's habits, and so he tallied such samples for a dozen writers. Some basic facts quickly emerged. Robert Fabyan, writing circa 1500, produced sentences with an average length of 63 words. Edmund Spenser, a century later, wrote 50-word sentences. By the time we come to Ralph Waldo Emerson in the middle of the 19th century, the average sentence length has dropped to 20.5 words. Comparing the overall averages for early and modern writers "furnish[es] evidence that the English prose sentence had dropped something like half its weight since Shakespeare's times."

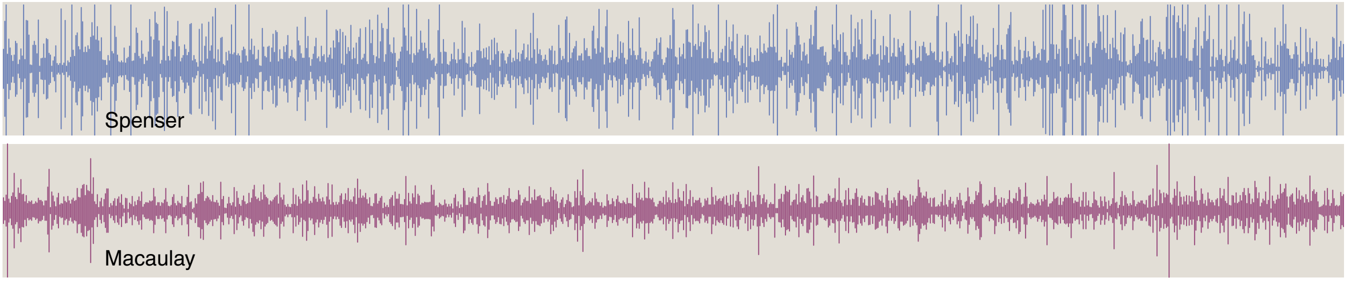
As his sentence-length data accumulated, Sherman began noticing other patterns. Some of his observations are mere curiosities; for example, he remarked on an excess of odd numbers in the sentence lengths of Thomas Babington Macaulay and an excess of prime numbers in those of Thomas De Quincy. But Sherman also explored the distribution of sentence lengths—what Mendenhall might have called the *sentence spectrum*—which conveys much more information than a simple average. *(See illustration at left.)* He also remarked on rhythms created by variations in length from one sentence to the next. *(See illustration below.)* And he went on to examine the evolution of subtler linguistic properties such as the number of verbs per sentence and the linkage between clauses in complex sentences.

The Lit Lab

Sherman performed some prodigious feats of counting. One summer he set aside three weeks to tally the words in all 40,000+ sentences of Macaulay's five-volume *History of England*. But he didn't do all the counting by himself. After all, he was a professor. He had students!

Sherman presents his analytic method as a pedagogic tool as much as a research program. Looking around at other university departments, he applauds the transformation then underway in the teaching of physics, chemistry, and biology, where memorization and classroom recitation gave way first to lab-bench demonstrations by the lecturer and then to hands-on experiments by the students. In a similar way he aimed to make English a laboratory course, in which students would dissect poetry and prose to identify the vital organs.

Not all of his students greeted this innovation with enthusiasm. One of the skeptics was Willa Cather, a novelist-to-be who was already writing professionally when she was an undergraduate. She and Sherman dueled as rival columnists and critics for the Lincoln newspapers, and she wrote satiric verses about the analytics course for the campus literary magazine. Years later, Cather mockingly recalled her time in Sherman's class as "trying to



Lengths of the first 1,000 sentences in Spenser and Macaulay are represented by lengths of vertical strokes. The resulting graphs look much like the waveforms of recorded sounds, and they suggest quite different rhythms of prose composition.

find the least common multiple of Hamlet and the greatest common divisor of Macbeth."

Beyond Nebraska, *Analytics of Literature* did not go entirely unnoticed—a review in *Science* called it "epoch-making"—and yet it clearly failed in its mission to transform literary criticism into a laboratory science.

Today the book seems almost entirely forgotten. (I learned of it from Mark Liberman of the University of Pennsylvania, writing on the Language Log website.) Whereas Mendenhall's work is still cited by statisticians, Sherman is seldom mentioned outside of a few specialized realms: Nebraska history, biographies of Willa Cather, and studies of "readability."

Books to Not Read

There is something undeniably risible about the earnest savant, engrossed in the counting of letters, words, and sentences, while the ordinary reader finds a better use for books. Willa Cather was not alone in poking fun at this figure. Jonathan Swift, in *Gulliver's Travels*, wrote of the professor who "made the strictest Computation of the general Proportion there is in the Book between the Numbers of Particles, Nouns, and Verbs, and other Parts of Speech."

To some extent Sherman deserves his obscurity; there is much in *Analytics of Literature* that now seems muddleheaded or misguided. And yet his basic observations about changes in sentence structure might yet reward further investigation—most likely as a linguistic rather than a literary phenomenon.

Sherman did the hard work of counting and compiling data, but he wasn't able to formulate much of an explanation of why and how the changes in syntax came about. His interpretation was vaguely evolutionary, framed in terms of the progressivist and teleological ideas that then ruled evolutionary thought. According to this view, written English prose had been steadily gaining in "fitness," and reached its culmination in Sherman's own time. The sentences of Elizabethan writers "are prevailingly either crabbed or heavy," he wrote. "Ordinary modern prose, on the other hand, is clear, and almost as effective to the understanding as oral speech." There's no sign that Sherman gave any thought to how this evolutionary process might continue into the future. What would he make of the English sentence in the age of texting? LOL.

Perhaps the digital humanists of the 21st century will rediscover and extend Sherman's work on sentence structure. They are already re-implementing his idea of the literature lab. I recently came upon the syllabus for an English course at Northeastern University where the class assignment begins: "Choose a big Victorian novel to not read."

Bibliography

Funda, E. I. 2005. "With scalpel and microscope in hand": The influence of Professor Lucius Sherman's 19th-century literary pedagogy on Willa Cather's developing aesthetic. *Prospects* 29:289–324.

Mendenhall, T. C. 1887. The characteristic curves of composition. *Science* 9:237–249.

Mendenhall, T. C. 1901. A mechanical solution of a literary problem. *Popular Science Monthly* 60:97–105.

Moretti, F. 2013. *Distant Reading*. London: Verso.

Mosteller, F., and D. L. Wallace. 1964. *Inference and Disputed Authorship: The Federalist*. Reading, MA: Addison-Wesley.

Sherman, L. A. 1892. On certain facts and principles in the development of form in literature. *University Studies of the University of Nebraska*, Vol. 1, No. 4.

Sherman, L. A. 1893. *Analytics of Literature: A Manual for the Objective Study of English Prose and Poetry*. Boston: Ginn & Company.