N-gram

John Cayley

When applied to linguistic analysis, natural language processing, and text generation, an *n*-gram is, typically, a phrase composed of "*n*" that is "one or more" "grams" that are "tokens" (in the parlance of algorithmic parsing) or "words" (for our purposes). Linguistic *n*-grams are harvested from sequences of words that have, traditionally, been composed by human beings. Clearly, *n*-grams have historically established relative frequencies of occurrence within the corpora where they are found. These frequencies can be used to build a statistical model—most often a Markov model—for a corpus, and the model can be used to generate statistically probable sequences of words. This is the main engine of combinatory and automatic text generation (see COMBINATORY AND AUTOMATIC TEXT GENERATION). Some of the probable sequences generated from a model will, of course, already exist in the corpus, but many of them will not occur, either because these sequences have not yet been composed by human authors or because they would be considered "malformed" for reasons that are beyond the domain of statistical modeling. What, precisely, we can safely deem to be "beyond the domain of statistical modeling" is something of an issue, especially now, although it has been since the early days of the mathematical analysis of language use. Is language choice or chance (Herdan 1966)?

In December 2010, Google made its Ngram Viewer public (http://books.google.com/ngrams/). Intimately allied with this release was the publication of a major multiauthored paper in *Science* (Michel et al. 2011). This was a signal event that allowed us to see that, for some indeterminate amount of time, Google had been taking very seriously the statistical analysis of the corpora it has been harvesting from the Internet and elsewhere. In the case of the Ngram Viewer itself, the corpus is confined to millions of items from the Google Books digitization project. This corpus has also been normalized to a certain extent, as attested in the *Science* article, if not to the degree of thoroughness that is the (always unattainable) ideal of scholarly textual criticism. But there is no question about the "power" of the Ngram Viewer and what it represents for linguistic practice, including aesthetic literary practice. Set out in the *Science* article, there are enough fascinating examples of graphically represented "statements" emerging from the Ngram Viewer as a device of so-called quantitative cultural analysis to establish many major projects of research and, hopefully, language-driven aesthetically motivated data visualization.

Meanwhile, however, there are other service providers, such as Microsoft, also making their *n*-grams available (http://web-ngram.research.microsoft.com/info/), and thus it is becoming clear that this is the tip of a statistical analytic universe that is expanding around us, as language makers, at an explosive rate (Gleick 2011). The *n*-gram model that Google is building—from everything it can crawl from what we inscribe on the digital network—is as close as we may get to a model of "all" inscribed language. Access to this model is now tantalizingly on tap, literally at our finger tips. However, despite all blandishments to the contrary (such as Google's twin mottos "Don't be evil" and "Organize the world's information and make it universally accessible and useful"), access to these vital and potentially productive cultural vectors into and through what should be the inalienable commons of languages is mediated and controlled by the nonreciprocal application of proprietary algorithms; by terms of use or service; by outmoded legal considerations (because whole texts might be reconstituted from 5-gram data sets that include low-frequency *n*-grams, those with less than forty occurrences are not provided within data sets now "freely downloadable"; http://books.google.com/ngrams/datasets); and by the fact that, currently, the provision of these cultural vectors is funded and thus necessarily redirected by the vectors of commerce, via advertising, rather than by the needs and desires of the sciences, humanities, and arts. These data are constructed from language, the very medium of any practice of digital textuality, and so artists and critics of this medium—a commons within which all of us dwell—are increasingly engaging with the *n*-gram.

■ See also COMBINATORY AND AUTOMATIC TEXT GENERATION, COMPUTATIONAL LINGUISTICS, DATA, FLARF, SEARCH

References and Further Reading

Gleick, James. 2011. "How Google Dominates Us." *New York Review of Books* 58 (13). www.nybooks.com/articles/archives/2011/aug/18/how-google-dominates-us/.

Herdan, Gustav. 1966. *The Advanced Theory of Language as Choice and Chance*. Berlin: Springer.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, et al. 2011. "Quantitative Analysis of Culture Using Millions of Digitized Books." *Science* 331:176–182.

Shannon, Claude E., and Warren Weaver. (1949) 1998. *The Mathematical Theory of Communication*. Urbana: University of Illinois Press.