FREDERIC KAPLAN

Linguistic Capitalism and Algorithmic Mediation

GOOGLE MADE 50 BILLION DOLLARS in revenue in 2012, an impressive financial result for a company created less than fifteen year ago.[1] That figure represents about 140 millions dollars per day, 5 million dollars per hour. By the time you have finished reading this article (about six minutes), Google will have made about 500,000 dollars. What does Google actually sell to get such astonishing results? Words. Millions of words.

The success of Google's highly original business model is the story of two algorithms. The first—pioneering a new way of associating web pages to queries based on keywords—has made Google popular. The second—assigning a commercial value to those keywords—has made Google rich.

In 1998, search engines could be used to search for web pages containing certain keywords, but they used inefficient and easily hackable ranking methods, such as the number of occurrences of a search keyword within a page. Most of those methods were not scalable as the number of web pages grew.[2] Larry Page, Google's cofounder, designed an alterative computation of the relevance of search results by adapting a ranking principle that is well established in the academic world: the most important documents are the most cited. He invented a recursive formulation of this principle by computing the value of a page based on the sum of the values of documents citing it.[3] Each citation behaved like a vote whose weight was proportional to the number of citations of the citing document. With this voting principle, classification and search results kept improving as the World Wide Web continued to extend: the more documents, the finer the ranking. The relevance of the results provided rapidly outperformed the other major search

ABSTRACT Google's highly successful business model is based on selling words that appear in search queries. Organizing several million of auctions per minute, the company has created the first global linguistic market and demonstrated that *linguistic capitalism* is a lucrative business domain, one in which billions of dollars can be realized per year. Google's services need to be interpreted form this perspective. This article argues that linguistic capitalism implies not an *economy of attention* but an *economy of expression*. As several million users worldwide daily express themselves through one of Google's interfaces, the texts they produce are systematically mediated by algorithms. In this new context, natural languages could progressively evolve to seamlessly integrate the linguistic biases of algorithms and the economical constraints of the global linguistic economy. REPRESENTATIONS 127. Summer 2014 ISSN 0734-6018, electronic ISSN 1533-855X, pages 57–63. DOI: 10.1525/rep.2014.127.4.57.

engines and, by the beginning of the year 2000, Google was the most popular portal for accessing information on the World Wide Web.

This algorithm relies heavily on the blind mechanisms of so-called collective intelligence. It functions well if document creators ignore the existence of the ranking and if actors do not deliberately try to create content to enhance their scoring artificially.[4] As expected, in the last ten years, many algorithms have been developed to deceive Google's ranking criteria. Such algorithms optimize textual and intertextual content to push content further in the search results. Google has kept updating its methods for detecting those algorithmically produced fake contents, but text-producing algorithms have continued to improve their ability to outwit Google's countermeasures. This is how the first "linguistic war" on the Internet, characterized by the first massive production of algorithmic texts on the Word Wide Web, started. I will discuss some of its multiple implications later in this article.

In March 2000, the "Internet bubble" collapsed, and many "start-ups" offering good use value but no exchange value went bankrupt. Most of their business models were based on selling advertising space on high-traffic web pages, hoping that the popularity of their services would motivate high prices. Again, Google's founders had a clever intuition. They realized that they were accumulating a form of *linguistic capital* as the number of Google's users continued to grow, and to enter ever-larger numbers of search queries. Google managed to transform this linguistic capital into actual money by organizing an algorithmic auction model for selling keywords.

The principle of this system is well known. Every time a user views search results on Google, several sponsored links with a short text are presented. Advertisers pay only if Google displays their ad and users click on the link. In order to choose what ad to display, an algorithm organizes a bidding process in three steps.

First, advertisers select a keyword—for instance "vacation"—and define the maximum price they would be ready to pay if a user arrives on their site by clicking on the link of the ad. To help advertisers, Google gives an estimate of the amount one needs to offer to have a reasonable chance of being among the selected ads. However, a high bid does not automatically guarantee selection.

Second, Google associates a quality score with the ad. This figure, ranging from 1 to 10, evaluates the global "quality" of the ad, which is computed through a complex combination of various factors, including the relevance of the text ad regarding the keyword, the average number of clicks on the ad, and the performance and quality of the linked website.[5] This score measures how well the ad is working (remember that Google is only making money if users actually click on the advertiser link). The exact computation method is kept secret and can be changed at any time by Google.

Third, the rank of an ad is calculated by multiplying the bid times the quality score and sorting the results from the highest to the lowest result." An ad with a good score and medium bid can overcome a less efficient ad with a higher bid. Eventually, the price paid by the advertisers is not their maximum auction offer but a slightly lower price, one computed on a second-price auction model.[6]

Such auctions happen every time a user enters a search query—about three billion times per day in 2012—millions of times per minute.[7] Google has created the first global, real-time, and multilingual linguistic market. As a consequence, the fluctuation of the price of keywords indirectly reflects global linguistic movements. The value of some keywords like "snowboarding" or "bikini" varies seasonally. The increase and decrease of the word "gold" is linked with the perceived state of financial crisis. Google makes a lot of money on some very competitive keywords like "flowers," "hotels," "vacation," and "love." It also organizes bids for buying the names of famous people ("Picasso," "Freud"). Bidding strategies vary.[8] Anything that can be named can be associated with a bid.

Some words and expressions have therefore become commodities with different monetary values that can be "bought" from Google. In some sense, Google has extended capitalism to language, transforming linguistic capital into money. The company has demonstrated that *linguistic capitalism* is a lucrative business domain, one in which billions of dollars of revenue can be realized per year. Understanding the rules of this new economical game is of crucial importance.

It is important to understand that, although in principle every word can become the subject of bidding, in practice only some words do. Anna Jobin and I, in a 2013 study, suggest naming the lexicon of the commodified derivate of the English language *Google-ese*, by analogy Google's commodified derivate of the French language *Googlais*, its German equivalent *Googlisch*, and so on, while other ad-selling search engines with potentially different algorithms and economic markets could be associated with different lexica, leading to, for example, *Bingese*, *Bingais*, and *Bingisch* for Bing.[9] The very existence of *Google-ese*, *Googlais*, *Googlisch*, and the like—that is, specific keywords bought by advertisers and marketers—accounts for the company's financial success, and many of the services it provides free of charge must be studied from this perspective.

When Google's autocompletion service transforms on the fly a misspelled word, it does more than offer a service. It transforms linguistic material without value (not much bidding on misspelled words) into a potentially profitable economic resource. When Google automatically extends a sentence you have started to type, it does more than save you some time, it transforms your expression into one that is statistically more regular based on the linguistic

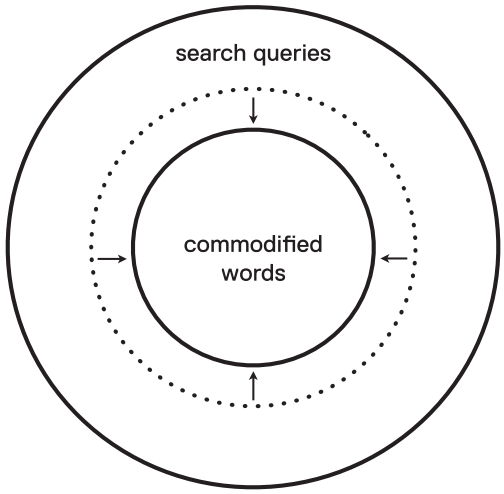

FIGURE 1. Autocompletion services can transform linguistic material without value (not much bidding on misspelled words) into a potentially profitable economic resource.

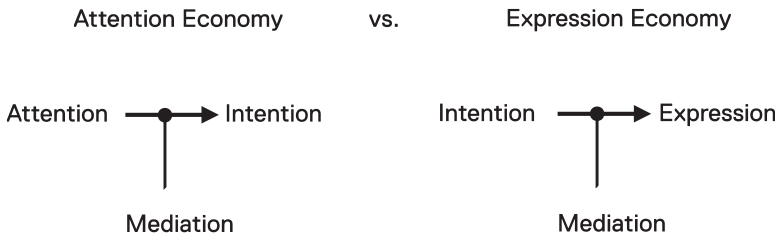

FIGURE 2. Attention economy vs. expression economy.

data it daily gathers. Even if Google's autocompletion may not be explicitly biased toward more economically valuable expressions, it nevertheless tends to transform natural language into more regular, economically exploitable linguistic subsets. The more we use Google's linguistic prosthesis, the more this transformation is likely to happen (fig. 1).

We are now several million users worldwide who daily express ourselves through one of Google's interfaces (Google Docs, Gmail, Google+, and so on). Google is certainly the first economic actor to have understood that the logic of linguistic capitalism implies not an *economy of attention* but an *economy of expression* (fig. 2). The goal in this new economic game is not to catch the users' gaze but to develop intimate and sustainable linguistic relationships with the largest possible number of users in order to model linguistic change

accurately and mediate linguistic expression systematically. The discovery of this previously unknown territory of capitalism announces new economic wars. Google arrived first and can now benefit from an advance in terms of linguistic capital, that is, the amount of linguistic data it daily mediates, but other players will learn to master the new and relatively simple rules of this game.

Can we anticipate the global linguistic effects of the commodification of words and general algorithmic mediation of textual expression? Should we observe a general tendency toward more regular and less idiomatic forms of linguistic expression, more suitable to for purchase? This might be a reasonable hypothesis if we consider the effect of autocompletion algorithms in isolation and the growing tendency to use them as mediators in our textual expression. Unfortunately, these text-transforming algorithms are often based on statistical models of the texts encountered on the Internet, and these texts are no longer "pure" natural language resources, but have themselves already been altered by various algorithmic mediations.

To further our understanding of this new linguistic environment we should distinguish between text on web pages that have been authored entirely by humans, which we call *primary resources*, and *secondary resources*, in which algorithms have played an important shaping role. Secondary resources include texts produced by spambots, bot-edited text such as Wikipedia articles (in which algorithms correct and structure texts essentially produced by human writers); texts produced through machine translation, through automatic summarizers (creating short texts out of long ones), and through text-spinning engines (created for the purpose of enlarging the "lexicon footprint" of a text in order to receive more visitors through search engines); and other forms of automatically generated articles, where textual content is produced out of a structured database.[10]

Distinguishing primary from secondary resources is not a trivial issue. In certain cases, human readers can tell the difference based on a feeling of oddity induced by texts produced by algorithms. These oddities are due to specific lexical or syntactic biases of these algorithms that often produce syntactically correct sentences—that no natural speaker would ever write. Unfortunately, it is much more difficult for an algorithm to detect such distinction automatically. As the proportion of these secondary resources compared to primary resource increases, computational linguistic statistical models may include some of the vocabulary and expression forms originating from these algorithms. As a consequence, these expressions may be suggested to us as statistically plausible forms by autocompletion algorithms.

If we follow this hypothesis, natural languages could progressively evolve to seamlessly integrate the linguistic biases of algorithms and the economical constraints of the global linguistic economy. Are we witnessing a new stage of

"grammatization" through yet another retroaction of technology and economy on natural languages?[11] Should we expect something like a pidgin or a creole to emerge, whose syntax and vocabulary would be influenced by the linguistic capacity of machines and economic value of words? We should definitely monitor the evolution of commodified lexicons. We should also conduct research on the new algorithmic dialects and work toward designing algorithms to recognize them automatically. Eventually, we should track and document creolization phenomena, if they occur. Through the commodification of words and the advent of algorithms as a new media, something is likely happen to language, and, although we are not yet sure what it will be, new tools must be built in order to understand this global linguistic evolution.

Notes

1. "Google Inc. Announces Fourth Quarter and Fiscal Year 2012 Results," Google, "Investor Relations," http://investor.google.com/earnings/2012/Q4_google_earnings.html.
2. John Battelle, *The Search: How Google and Its Rivals Rewrote the Rules of Business and Transformed Our Culture* (New York, 2005).
3. Lawrence Page, "Method for Node Ranking in a Linked Database," Google, "Patents," http://www.google.com/patents/US6285999.
4. Dominique Cardon, "Dans L'esprit Du PageRank," *Réseaux* 177, no. 1 (2013): 63, doi:10.3917/res.177.0063.
5. Micky Lee, "Google Ads and the Blindspot Debate," *Media, Culture & Society* 33, no. 3 (April 2011): 433–47, doi:10.1177/0163443710394902.
6. Benjamin Edelman, Michael Ostrovsky, and Michael Schwarz, "Internet Advertising and the Generalized Second-Price Auction: Selling Billions of Dollars Worth of Keywords," *American Economic Review* 97, no. 1 (March 2007): 242–59.
7. "2012 Search Trends: The World," Google, "Zeitgeist 2012," http://www.google.com/zeitgeist/2012/.
8. E. D'Avanzo, T. Kuflik, and A. Elia, "Online Advertising Using Linguistic Knowledge," in *Information Technology and Innovation Trends in Organizations*, ed. Alessandro D'Atri et al., (Heidelburg, 2011), 143–50, SpringerLink, http://www.springerlink.com/content/h730042483307369/abstract/.
9. Anna Jobin and Frédéric Kaplan,"Are Google's Linguistic Prosthesis Biased Towards Commercially More Interesting Expressions? A Preliminary Study on the Linguistic Effects of Autocompletion Algorithms," (paper presented at the *Digital Humanities* conference at the University of Nebraska-Lincoln, 16-19 July 2013, http://dh2013.unl.edu/abstracts/ab-223.html).
10. Zoltán Gyöngyi and Hector Garcia-Molina, "Web Spam Taxonomy," (paper presented at the *First International Workshop on Adversarial Information Retrieval on the Web [AIRWeb]*, 10–14 May 2005, Chiba, Japan), http://airweb.cse.lehigh.edu/2005/gyongyi.pdf. R. Stuart Geiger, "The Lives of Bots," in *Critical Point of View: A Wikipedia Reader*, ed. Geert Lovink and Nathaniel Tkacz (Amsterdam, 2011), 78–93, http://www.networkcultures.org/_uploads/%237reader_Wikipedia.pdf. Philipp Koehn et al., "Moses: Open Source Toolkit for Statistical

Machine Translation," (paper presented at the *45th Annual Meeting of the ACL* [Association for Computational Linguistics], 2007, Stroudsburg, PA), in *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions* (Stroudsburg, PA, 2007), 177–80; Harold Somers, "Review Article: Example-based Machine Translation," *Machine Translation* 14, no. 2 (June 1999): 113–57, doi:10.1023/A:1008109312730. Philip Parker, 2013. "Method and Apparatus for Automated Authoring and Marketing," Google, "Patents," http://www.google.com/patents/US7266767.

11. Sylvain Auroux, *La révolution technologique de la grammatisation: introduction à l'histoire des sciences du langage* (Liège, 1994).