Ngrams Against Agnotology: Combatting Tobacco Industry Narratives About Addiction Through A Quantitative Analysis Of 14 Million Documents

Stephan Risi
risi@stanford.edu
Stanford University, United States of America

What happens when researchers have access to more documents than they could read in a lifetime? As a result of litigation, historians of tobacco have access to over 14 million formerly secret tobacco industry documents, containing incriminating internal memos and research reports but also newspaper clippings and consumer letters (UCSF Library and Center for Knowledge Management). Historians have used this treasure trove to document widespread fraud and systematic deception of smokers by the tobacco industry

(Brandt, 2007; Proctor, 2011). However, industry-friendly historians and tobacco lawyers have started to rewrite some this history by claiming that smokers always knew that smoking is addictive and causes cancer (Brandt, 2007; Proctor, 2011). Usually, these claims rest on a few, well selected documents that support a particular industry claim. Robert Proctor has called processes like this "agnotology," the cultural production of ignorance (Proctor, 2008). Indeed, the arguments of both pro- and anti-tobacco industry historians rely on the same corpus of data: an immense amount of publicly available and full-text searchable documents. Given 14 million documents, there will be some supporting almost any claim.

In this paper, I present one way to counter such agnotological assertions by studying broad trends across millions of documents with frequency analyses. In particular, I counter the claim that smokers always knew that smoking was addictive, an argument often made by tobacco lawyers in court to assign full responsibility to the smoker (Henningfield, Rose, & Zeller, 2006) To refute this assertion, I use frequency analyses with a validation measure to show that smoking only became widely understood as an addiction in the late 1980s and early 1990s, when scientists recognized that the same neural pathways were involved in dependence to both nicotine and harder drugs like heroin and cocaine. This inscription of addiction into the brain replaced older explanations of why people smoke, like personality traits or an oral fixation. Ultimately, I trace how the neurological understanding of nicotine addiction moved from research laboratories to the public: it led to the Surgeon General's warning labels; it enabled smokers to seek out new nicotine replacement therapies; and it made it possible for smokers to successfully sue the tobacco industry for the first time.

Frequency analysis, popularized by the Google Ngram Viewer, is one of the simplest mathematical tools in the arsenal of the digital humanities (Michel et al, 2011). By calculating the usage frequency of terms and expressions over time, it enables users to get a sense of when a term became more or less important. The mathematical simplicity confers it an important advantage: it scales very well not just to thousands but to millions of documents. For this study, I used all 14 million documents (about 10 billion tokens) dated between 1940 and 1998 to create a publicly available website (www.tobacco-analytics.org), where users can create their run their own frequency analyses, akin to Google ngrams.

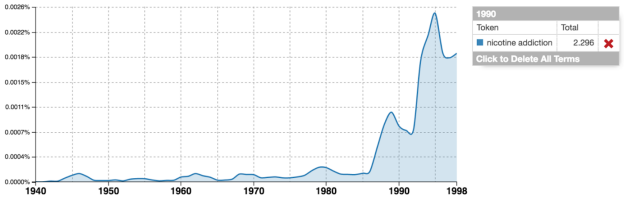

Figure 1: Screenshot of the relative frequencies of "nicotine addiction" in the tobacco documents from www.tobacco-analytics.org. The presentation will use a number of these graphs to show that smoking only became understood as an addiction in the late 1980s and early '90s.

The main drawback of this method is that the patterns found in the graphs of the Google Ngram Viewer are hard to validate: Does a spike in a particular year represent a statistically significant event or is it just a fluke? Is it caused by 10 or 1000 documents? I address this problem in two ways: First, I allow users to display the absolute number of appearances of a term by year to give them a sense of the number of documents that cause a spike. Second, I am developing a comparison statistic to calculate z-scores using the Corpus of Historical American English (COHA) (Davies, 2010). By comparing frequencies between the tobacco documents and the reference corpus (COHA), it allows me

to calculate when frequencies in the tobacco corpus deviate in a statistically significant way (Darwin, 2008, p. 208-222). Given, for example, the above graph of the relative frequencies of the term "nicotine addiction," z-scores can be used to show that the relative frequencies only started to deviate significantly from the comparison corpus in the 1980s.

The tobacco documents provide us with an opportunity to think through the problems that come with access to millions of secret documents. What if millions of dollars in settlements hinge on historical arguments? What if there are immense financial incentives to make false historical claims: to present narratives that are borne out in a few well selected documents, but which misrepresent the corpus as a whole? In the realm of tobacco, historical arguments and knowledge circulate far outside of academia in courtrooms to sway juries or in policy documents to change legislation. The immense size of the tobacco documents archive makes it possible to find a few documents supporting almost any claim. Findings from one group of documents can cancel out the findings from other documents; statements by one expert discredit those of another one. In these cases, quantitative analyses using the whole corpus can be an arbiter of these claims. They are not sufficient to advance historical arguments in themselves, but they can be used to test and disprove hypotheses made on the basis of a smaller set of documents. The Tobacco Analytics project makes powerful digital humanities tools available to tobacco researchers who may not have a technical background, and it allows historians to trace developments within the tobacco industry by examining the whole corpus with the click of a mouse.

Bibliography

Brandt, A. (2007). *The Cigarette Century: The Rise, Fall, and Deadly Persistence of the Product that Defined America*. New York: Basic Books.

Darwin, C. M. (2008). *Construction and Analysis of the University of Georgia Tobacco Documents Corpus.* PhD Dissertation, The University of Georgia, Athens, GA.

Davies, M. (2010). The Corpus of Historical American English: 400 million words, 1810-2009.

Henningfield, J., Rose, C., & Zeller, M. (2006). Tobacco Industry Litigation Position on Addiction: Continued Dependence on Past Views. *Tobacco Control, 15*(Suppl. 4), 27-36.

Kyriakoudes, L. M. (2006). Historians' Testimony on "Common Knowledge" of the Risks of Tobacco Use: A Review and Analysis of Experts Testifying on Behalf of Cigarette Manufacturers in Civil Litigation. *Tobacco Control, 15*(suppl 4), iv107-iv116.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., . . . Orwant, J. (2011). Quantitative Analysis of Culture Using Millions of Digitized Books. *Science, 331*(6014), 176-182.

Proctor, R. N. (2008). Agnotology. A Missing Term to Describe the Cultural Production of Ignorance (and Its Study). In R. N. Proctor & L. Schiebinger (Eds.), *Agnotology. The Making and Unmaking of Knowledge*. Stanford, CA: Stanford University Press.

Proctor, R. N. (2011). *Golden Holocaust: Origins of the Cigarette Catastrophe and the Case for Abolition*. Berkeley: University of California Press.