LETTER

Uncontrolled corpus composition drives an apparent surge in cognitive distortions

Benjamin Schmidt[a,1], Steven T. Piantadosi[b], and Kyle Mahowald[c]

Bollen et al. (1) present an exciting interdisciplinary combination of clinical psychology and corpus linguistics. They find that phrases chosen by cognitive-behavioral experts to reflect negative thoughts—"cognitive distortions"—have sharply increased in frequency since the 1980s in three Google Books datasets.

Unfortunately, their work faces a foundational limitation: For the reported patterns to be meaningful, corpus frequencies must actually reflect cognitive

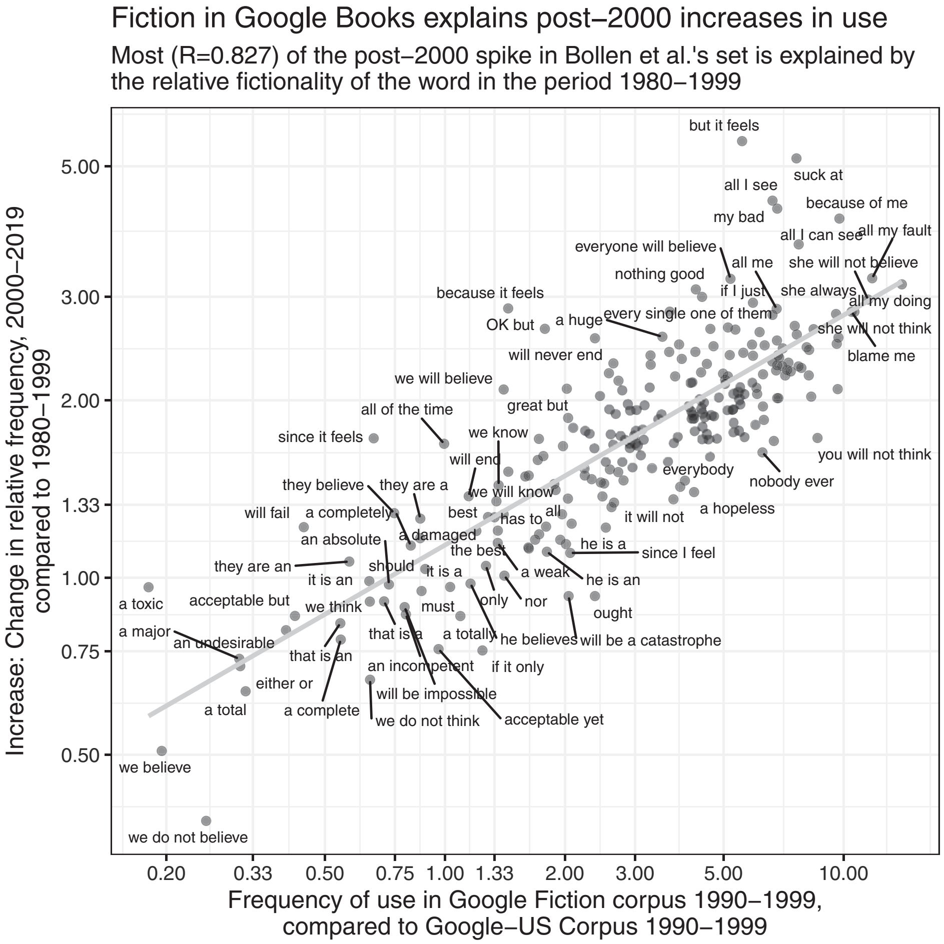

Fig. 1. The *x* axis is a measure of a word's frequency in Google's Fiction corpus relative to the general Google US corpus, between 1990 and 1999. The *y* axis shows the change in frequency between 1990–1999 and 2000–2019. The positive correlation shows that words which are more likely to appear in fiction between 1990 and 1999 were more likely to increase in overall frequency after 2000. This is consistent with a shift toward including more fiction in the corpus post-2000.

[a]Department of History, New York University, New York, NY 10012; [b]Department of Psychology, University of California, Berkeley, CA 94720; and [c]Department of Linguistics, University of Texas at Austin, Austin, TX 78712

Author contributions: B.S., S.T.P., and K.M. designed research, performed research, analyzed data, and wrote the paper.

The authors declare no competing interest.

[1]To whom correspondence may be addressed. Email: bs145@nyu.edu.

Published November 1, 2021.

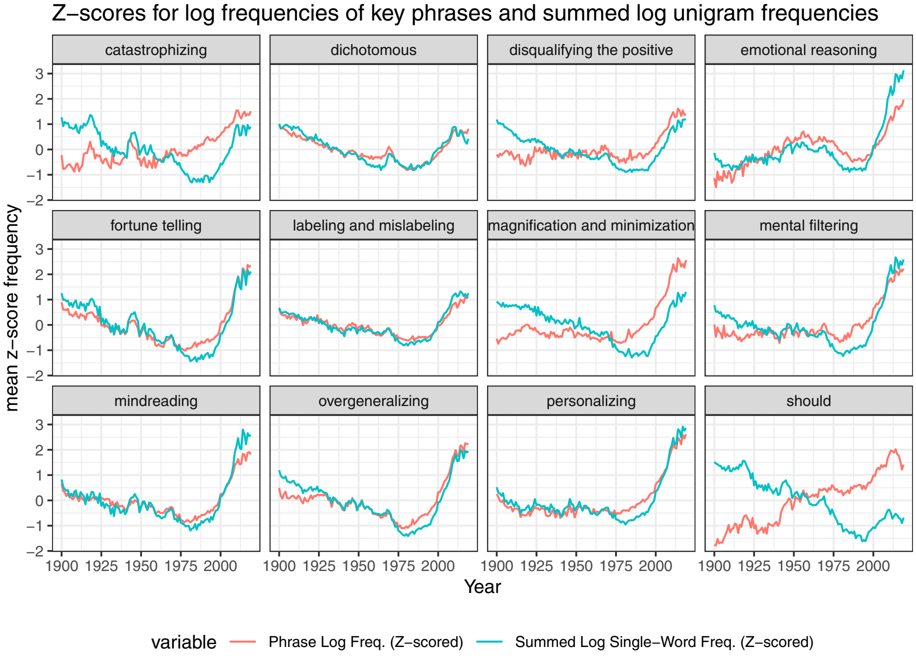

Fig. 2. **Split by the type of cognitive distortion, the mean *z*-scored frequency as a function of year for each phrase (e.g., logP(*but it feels*)) compared to the summed log unigram frequencies of the phrase's constituent words (e.g., logP(*but*) + logP(*it*) + logP(*feels*)) in the 2019 Google Books corpus. Unigram phrases are excluded from the analysis since the two values would not differ for unigrams. Most categories show very little difference between the trajectories of the phrases and their constituent words.**

distortions in the real world. But words in books are not clinical interviews, and word frequencies are not psychiatric assessments. Extrapolating to "entire societies" from phrases in library books is also problematic: English-language authors in Google Books talk about "Derrida" 3 times as much as "The Beatles," and talk about "the Federal Reserve" 30 times as much as "the grocery store."

Studies using Google Ngrams must properly negotiate changes in corpus composition (2). Many of the clinically relevant phrases ("my bad," "I will not," "but I feel") contain features like personal pronouns and verbs of thought that are more common in fiction than nonfiction. Although the authors state that Pechenik et al. (3) documents a recent decrease in fiction (1), it, in fact, describes a significantly different period in the much smaller 2012 Google corpus. Fig. 1 explores the effect of "fictionality" by comparing word frequencies in Google's 2019 "fiction" corpus to the 2019 US corpus Bollen et al. use. It shows that over two-thirds ($R^2 = 0.68$) of the post-2000 "hockey stick" effect can be explained solely by a word's fictionality in the period 1980–2000. We suggest the proportion of fiction in the corpus increased greatly as Google began including books directly from publishers, rather than academic libraries, after 2000. At least some of the remaining effect seems to be due to the fact that their experts provided contemporary phrases rather than historical ones (e.g., "huge" instead of "immense," "terrible" instead of "wretched"), which prevents fair longitudinal comparison.

Moreover, ordinary phrases like "it is an," "I will not," or "we know" could plausibly represent many things besides "internalizing disorders." The word "you," for example, tripled in usage rate from 1989 to 2019. Fig. 2 adjusts for this by comparing the joint log probability of a phrase to the summed log probability of its individual words. These are highly correlated, suggesting much of the increase in the frequency of these n-grams can be explained not by movement in the phrases but by changes in corpus frequency of the phrases' component words. For several categories, the phrases appear to increase in frequency less than would be expected based on their constituent words. The component words that increased most from 1970 to 2019 ("suck[s]," "still," "she," "not," "I") likely did so not in response to a cognitive shift but because of changes in corpus composition or because they are modern slang.

These considerations seriously undermine the approach of analyzing broad-scale textual patterns for clinical significance.

1 J. Bollen *et al.*, Historical language records reveal a surge of cognitive distortions in recent decades. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2102061118 (2021).
2 J. B. Michel *et al.*; Google Books Team, Quantitative analysis of culture using millions of digitized books. *Science* **331**, 176–182 (2011).
3 E. A. Pechenick, C. M. Danforth, P. S. Dodds, Characterizing the Google Books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PLoS One* **10**, e0137041 (2015).