LETTER

Reply to Garcia et al.: Common mistakes in measuring frequency-dependent word characteristics

Isabel M. Kloumann[f], James P. Bagrow[a,b], Karine Megerdoomian[g], Matthew T. McMahon[g], Brian F. Tivnan[b,g], and Christopher M. Danforth[a,b]

[a]Computational Story Lab, Vermont Advanced Computing Core, Department of Mathematics and Statistics, and [b]Vermont Complex Systems Center, University of Vermont, Burlington, VT 05401; [c]Center for Computational Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139; [d]School of Mathematical Sciences, North Terrace Campus, The University of Adelaide, Adelaide, SA 5005, Australia; [e]Applied Mathematics, University of Washington, Seattle, WA 98195; [f]Center for Applied Mathematics, Cornell University, Ithaca, NY 14853; and [g]The MITRE Corporation, McLean, VA 22102

The concerns expressed by Garcia et al. (1) are misplaced due to a range of misconceptions about word usage frequency, word rank, and expert-constructed word lists such as LIWC (Linguist Inquiry and Word Count) (2). We provide a complete response in our paper's online appendices (3). Garcia et al. (1) suggest that the set of function words in the LIWC dataset (2) show a wide spectrum of average happiness with positive skew (figure 1A in ref. 1) when, according to their interpretation, these words should exhibit a Dirac δ function located at neutral ($h_{avg} = 5$ on a 1–9 scale). However, many words tagged as function words in the LIWC dataset readily elicit an emotional response in raters as exemplified by "greatest" ($h_{avg} = 7.26$), "best" ($h_{avg} = 7.26$), "negative" ($h_{avg} = 2.42$), and "worst" ($h_{avg} = 2.10$). In our study (3), basic function words that are expected to be neutral, such as "the" ($h_{avg} = 4.98$) and "to" ($h_{avg} = 4.98$), were appropriately scored as such. Moreover, no meaningful statement about biases can be made for sets of words chosen without frequency of use properly incorporated.

Garcia et al. (1) compare our work on English with a similar sized survey by Warriner et al. (4). Warriner et al. generated a merged list of 13,915 English words, the bulk of which are a list of lemmas taken from movie subtitles, a mismatch with the corpora we used in creating our English word list labMT (language assessment by Mechanical Turk). In figure 1B of ref. 1, Garcia et al. make a flawed comparison between the two word lists because the words behind each histogram are not the same. For shared words, the minor difference in median h_{avg} of 0.07—much less than the observed positivity bias—cannot be because of our use of cartoon faces (emoticons). The earlier Affective Norms for English Words (ANEW) study upon which we modeled our work (5) also uses cartoons and yet found a lower median for words shared with Warriner et al. (5.29 versus 5.44) (4). All three datasets agree well in more general statistical comparisons (4).

In attempting to say anything about a given quality of words as it relates to use frequency within a specific corpora, a complete census of words by frequency must be on hand, otherwise uncontrolled sampling issues arise. In Fig. 1*A*, we plot average happiness as a function of frequency of use for the word list Garcia et al. (1) created from Google Books. The scatter plot is clearly unsuitable for linear regression. We show an estimate of cumulative coverage at the bottom, which crashes soon after reaching 5,000 words.

Sampling issues aside, Garcia et al. (1) state that regression against frequency f is a better choice than using rank r because information is lost in moving from f to r. However, the general adherence of natural language to Zipf's law, $f \sim r^{-1}$, provides an immediate counterargument, even acknowledging the possibility of a scaling break (6). Fig. 1*B* shows how use rank is well suited for regression, and is the basis for the "jellyfish" plots we presented in our work (3). In Fig. 1*C*, we present how h_{avg} behaves as a function of $1/f$, illustrating both the error in choosing $\log_{10} f$ and that our results will be essentially unchanged if we regress against $1/f$.

ACKNOWLEDGMENTS. This work was supported in part by National Science Foundation Grant DMS-0940271 (to C.M.D.) and National Science Foundation CAREER Award 0846668 (to P.S.D.).

Peter Sheridan Dodds[a,b,1], Eric M. Clark[a,b], Suma Desu[c], Morgan R. Frank[c], Andrew J. Reagan[a,b], Jake Ryland Williams[a,b], Lewis Mitchell[d], Kameron Decker Harris[e],

1 Garcia D, Garas A, Schweitzer F (2015) The language-dependent relationship between word happiness and frequency. *Proc Natl Acad Sci USA* 112:E2983.

2 Pennebaker JW, Booth RJ, Francis ME (2007) Linguistic Inquiry and Word Count: LIWC 2007. Available at homepage.psy.utexas.edu/HomePage/Faculty/Pennebaker/Reprints/LIWC2007_OperatorManual.pdf. Accessed May 15, 2014.

3 Dodds PS, et al. (2015) Human language reveals a universal positivity bias. *Proc Natl Acad Sci USA* 112(8):2389–2394.

4 Warriner AB, Kuperman V, Brysbaert M (2013) Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behav Res Methods* 45(4):1191–1207.

5 Bradley MM, Lang PJ (1999) *Affective Norms for English Words (anew): Stimuli, Instruction Manual and Affective Ratings. Technical report c-1* (Univ of Florida, Gainesville, FL).

6 Williams JR, Bagrow JP, Danforth CM, Dodds PS (2015) Text mixing shapes the anatomy of rank-frequency distributions: A modern Zipfian mechanics for natural language. arXiv:1409.3870.

Author contributions: P.S.D., E.M.C., S.D., M.R.F., A.J.R., J.R.W., L.M., K.D.H., I.M.K., J.P.B., K.M., M.T.M., B.F.T., and C.M.D. analyzed data; and P.S.D. and C.M.D. wrote the paper.

The authors declare no conflict of interest.

[1]To whom correspondence should be addressed. Email: pdodds@uvm.edu.

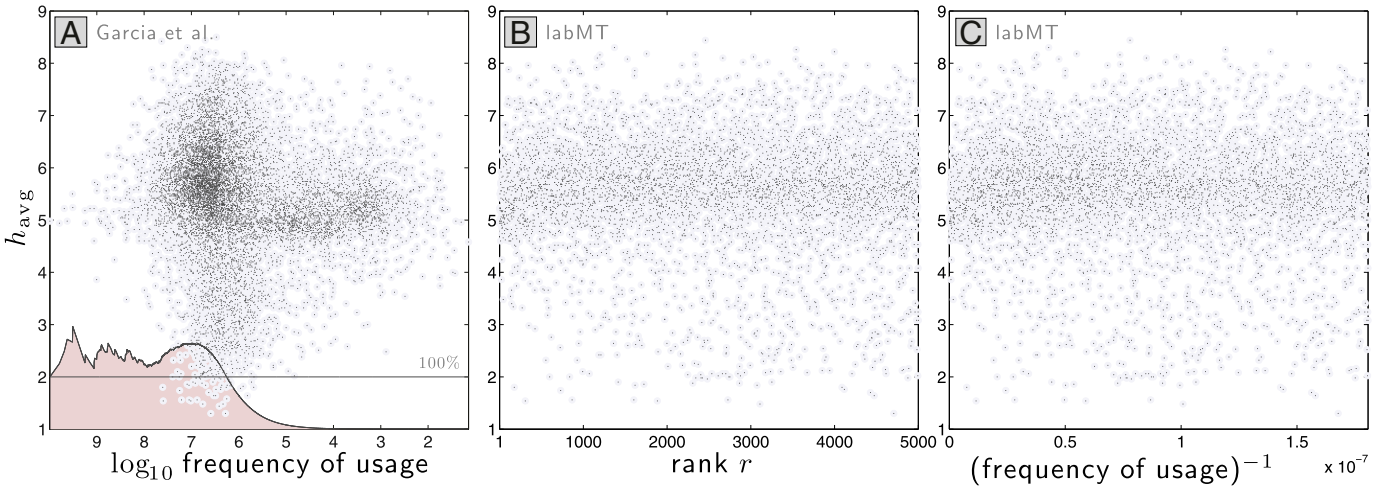

Fig. 1. (*A*) Scatterplot of h_{avg} as a function of word usage frequency for the English Google Books word list generated by Garcia et al. (1). Uncontrolled subsampling of lower frequency words yields a poor candidate for linear regression. The lower curve provides a coarse estimate of cumulative lexicon coverage as a function of usage frequency f using Zipf's law $f_r \sim f_1 r^{-1}$ inverted as $r \sim f_1/f_r$. The rapid drop off begins at around rank 5,000, the involved lexicon size for Google Books in labMT (2). (*B*) Scatterplot of h_{avg} as a function of rank r for the 5,000 words for Google Books contributing to labMT, the basis of our "jellyfish" plots (2). (*C*) The same data as in *B* but now plotted against the inverse of usage frequency. The approximate adherence to Zipf's law $f \sim r^{-1}$ means there is no substantive loss of information if regression is performed on the correct transformation of frequency. Linear regression fits for the first two scatterplots are $h_{avg} \simeq 0.089 \log_{10} f + 4.85$ and $h_{avg} \simeq -3.04 \times 10^{-5} r + 5.62$ (as reported in ref. 3). Note difference in signs, and the far weaker trend for the statistically appropriate regression against rank in *B*. Pearson correlation coefficients: +0.105, −0.042, and −0.043 with P values 6.15×10^{-26}, 3.03×10^{-3}, and 2.57×10^{-3}. Spearman correlation coefficients: +0.201, −0.013, and −0.013 with P values 6.37×10^{-92}, 0.350, and 0.350 (*B* and *C* must match). The Spearman analysis indicates that an assumption of a nonmonotonic relationship between h_{avg} and rank r is well supported.