"Speaking about Seeing": Supplementary material

Zekun Sun and Chaz Firestone

Across several different types of stimuli (shapes, matrices, and motion paths), subjects eventually use fewer words to describe more complex stimuli. Why? And in particular, what *about* their descriptions is changing such that the descriptions become shorter? The main text of our paper not only reports this phenomenon itself, but also shows that the frequency of "random" words ("random", "irregular", "odd", etc.) increases as stimulus complexity increases. Here, we report another content analysis investigating the relationship between the complexity of a stimulus and the rarity of the words used to describe it.

Before continuing, a few caveats are in order. (Indeed, some of these caveats are why we report this analysis here in this supplement rather than in the main text.) First, this is a purely exploratory approach: Most of the specific analyses reported below occurred to us only after collecting (and in some cases examining) our data, and so they are not "a priori", or pre-registered, etc. Second, there is no one agreed-upon way to analyze word frequency, and as a result there are many analytical degrees of freedom available to an experimenter; as can be seen below, in many cases we present multiple different analyses for the "same" question, often with different results. Third, the necessarily imperfect nature of automated speech transcription (which was approximately 90% accurate across our experiments) is uniquely challenging for this type of analysis (compared to the analyses reported in the main text of our paper). In particular, occasionally "mishearing" one word as another (e.g., transcribing "shape" as "shave") has little or no consequence for the primary analyses in our paper, which simply concerns the number of words used in a description. It even has little consequence for our "random words" analysis, since a good-but-imperfect transcription is still acceptable for seeking out specific keywords. By contrast, looking for rare words, as we do below, could be very sensitive to transcription errors, since the mistranscribed word might be much less frequent than the word the subject actually said (e.g., "shave" being much less frequent than "shape").

For these reasons (and still others we mention below), we think that the results reported below are far from conclusive, but still interesting enough to report here in this supplement. And to avoid biases arising from our analytical choices, we report here every analysis we have conducted, which includes variations on the corpus used, the inclusion of "stop words", and whether the calculation is made over the number of unique words vs. the average frequency of the words in a given description. We hope that these exploratory analyses are interesting and informative for readers, and we also invite readers to explore these data on their own.

Do more complex stimuli get more interesting / rarer words?

One reason that more complex stimuli might receive shorter verbal descriptions is that the descriptions pack richer detail into fewer words — e.g., by using rare or infrequent vocabulary that carries with it more information. For example, if a subject were to describe a shape as "a blob with big round corners" vs. "a fractal snowflake atop wilting flowers", both descriptions technically have the same length (6 words), but the second description uses words that are much more specific and rare, and so in information-theoretic terms carry "more" information. Did that happen here?

To find out, we computed the average frequency of the words spoken in a given shape's descriptions. Our first analytical choice was with regard to the corpus used: Either (1) the frequency of words in a pre-established corpus of English (what we'll below call "Global Frequency"); or 2) the frequency of words from a given description with respect to all the descriptions collected across the experiment (what we'll below call "Local Frequency"). For example, using Global Frequency, the word "triangle" is fairly infrequent (since it doesn't appear too often in English corpora); but using Local Frequency, the word "triangle" is quite frequent (since it appears quite a bit in subjects' descriptions of shapes).

We then ran three analyses within each of these approaches. We either (a) examined the frequency of each word in a given description, then averaged those frequencies together (as long a given word was more than 2 letters long) to give that description a frequency score; (b) ran the same analysis but without including "stop words" (i.e., words such as "who", "that", "with", "the", and so on, as specified by the NLTK standard; Bird et al., 2009); and (c) counted up all the unique words generated for each stimulus, and measured the frequency of those (so that, e.g., repeatedly saying "square" counts as only one instance of "square"). As can be seen below, these choices often produce different patterns of results, and so we report them all in full.

1 Global frequency

Our first approach computed word frequency with respect to a widely used corpus of English: *wordfreq* 2.2.1 (`https://pypi.org/project/wordfreq/`; Speer et al., 2018).

1.1 All words

Our first analysis computed the frequency of each word in a given description, and then averaged those frequencies together such that the description would have a mean word frequency that (indirectly) reflects the "information density" of that description. For example, if a subject said "almost a perfect square with one corner chopped off", every word (except "a") received a frequency score, and then those scores are averaged together. We then averaged together the scores of each of these descriptions for a given stimulus (e.g., a single shape), such that every stimulus has a frequency score that is the average of the average frequency scores across each of its descriptions. Finally, we correlated these scores (one for each stimulus) against that stimulus' "objective" complexity for each experiment (i.e., skeletal entropy, organization level, or # of direction changes).

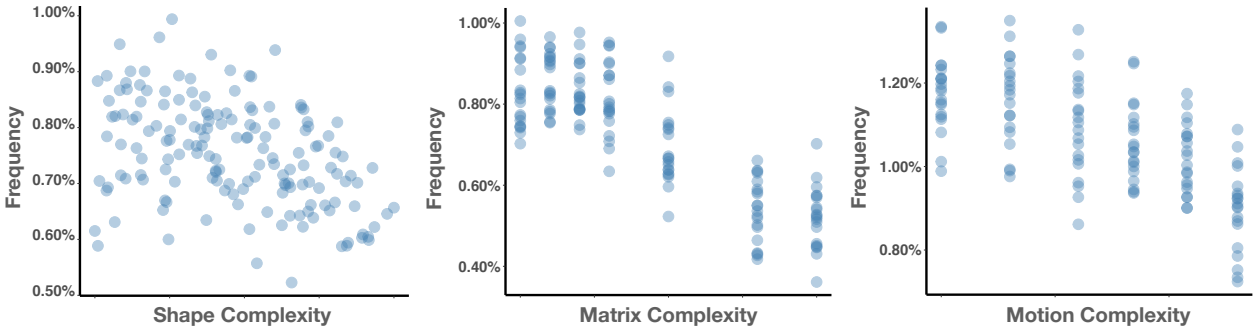

Figure 1. The relationship between the objective complexity of a stimulus (i.e., shapes, matrices and motion-paths; x-axis), and the averaged word frequency of descriptions of that stimulus (y-axis). For each word, its frequency value comes from a commonly used English corpus; for each description, the averaged word frequency is calculated from all of its words, including stop words.

Exp.1 (Shapes): A linear model with skeletal surprisal as a predictor explained a significant proportion of variance in average word frequency of descriptions, with the two quantities being negatively related($R^2 = 0.17$,$F(1, 158) = 33.01$, $p = 4.59 \times 10^{-8}$). However, a quadratic model was also significant ($R^2 = 0.22$, $F(2, 157) = 21.97$, $p = 3.86 \times 10^{-9}$), including a significant quadratic term ($b = -2.95 \times 10^{-3}$, $95\%CI = [-4.86 \times 10^{-3}, -1.03 \times 10^{-3}]$, $t = -3.04$, $p < 0.01$).

Exp.2 (Matrices): A linear model with matrix randomness as a predictor explained a significant proportion of variance in average word frequency of descriptions, with the two quantities being negatively related. ($R^2 = 0.70$ $F(1, 138) = 323.9$, $p < 2.20 \times 10^{-16}$.) Here, the quadratic model explained a similar proportion of variance ($R^2 = 0.71$, $F(2, 137) = 165.0$, $p = 2.20 \times 10^{-16}$), and the quadratic term was *not* significant ($b = -1.26 \times 10^{-3}$, $95\%CI = [-2.83 \times 10^{-3}, 3.06 \times 10^{-4}]$, $t = -1.59$, $p = 0.11$).

Exp.3 (Motion): A linear model with # of direction changes as a predictor explained a significant proportion of variance in average word frequency of descriptions, with the two quantities being negatively related. ($R^2 = 0.44$ $F(1, 118) = 91.84$, $p < 2.2 \times 10^{-16}$.). Again, the quadratic model explained a similar

proportion of variance ($R^2 = 0.47$, $F(2, 117) = 51.81$, $p = 2.2 \times 10^{-16}$), and the quadratic term was significant ($b = -1.42 \times 10^{-3}$, $95\%CI = [-2.48 \times 10^{-3}, -3.62 \times 10^{-4}]$, $t = -2.66$, $p < 0.01$).

In other words, across these three experiments, more complex stimuli tended to receive less frequent words, as can be seen in the corresponding graphs. This might suggest that, even with a plateau or reduction in verbal description length, descriptions pack "more information" into fewer words.

1.2 Stop words excluded

Intriguingly, however, we found a different (and in some cases opposite) pattern when stop words (such as "who", "that", "with", "the", and so on) were excluded.

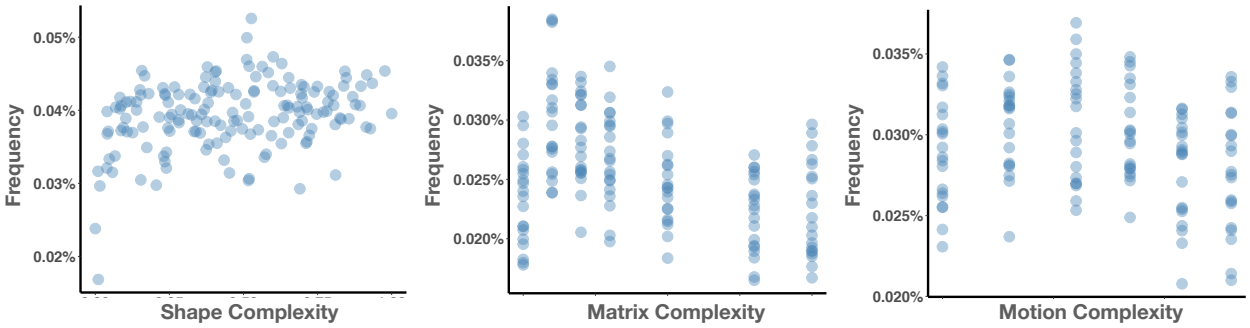

Figure 2. The same analysis as in Figure 1, but with stop words (such as "who", "that", "with", "the", and so on) excluded.

Exp.1 (Shapes): A linear model with skeletal surprisal as a predictor explained a significant but very small proportion of variance in average word frequency of descriptions, with the two quantities being *positively* related($R^2 = 0.093$, $F(1, 158) = 16.22$, $p = 8.75 \times 10^{-5}$). However, a quadratic model explained significantly greater variance ($R^2 = 0.13$, $F(2, 157) = 12.01$, $p = 1.40 \times 10^{-5}$), with a significant quadratic term, $b = -1.40 \times 10^{-4}$, $95\%CI = [-2.43 \times 10^{-4}, -3.67 \times 10^{-5}]$, $t = -2.68$, $p < 0.01$).

Exp.2 (Matrices): A linear model with matrix randomness as a predictor explained a significant proportion of variance in average word frequency of descriptions, with the two quantities being negatively related($R^2 = 0.16$, $F(1, 138) = 26.54$, $p = 8.73 \times 10^{-7}$). Here, the quadratic model explained a greater proportion of variance ($R^2 = 0.18$, $F(2, 137) = 15.5$, $p = 8.51 \times 10^{-7}$), and the quadratic term was marginally significant ($b = -7.94 \times 10^{-5}$, $95\%CI = [-1.59 \times 10^{-4}, 1.41 \times 10^{-8}]$, $t = -1.98$, $p = 0.050$).

Exp.3 (Motion): A linear model with # of direction changes as a predictor explained a significant but very small proportion of variance in average word frequency of descriptions, with the two quantities being negatively related ($R^2 = 0.039$,$F(1, 118) = 4.74$, $p < 0.05$). Again, the quadratic model explained a greater proportion of variance ($R^2 = 0.11$, $F(2, 117) = 7.16$, $p < 0.01$), and the quadratic term was significant ($b = -5.35 \times 10^{-5}$, $95\%CI = [-8.83 \times 10^{-5}, -1.86 \times 10^{-5}]$, $t = -3.04$, $p < 0.01$).

In summary, across these three experiments, moderately complex stimuli tended to receive more frequent words compared to more complex or simple stimuli. Evidently, excluding stop words significantly changed the relationship that we found in previous section.[1]

1.3 Unique words

Another way to analyze these data is not to compute the average information density of each token word present in a single description of a stimulus, but rather to ask *what kinds of words* are used to describe that stimulus across all descriptions (with each word counting only once). In this analysis, we extracted all the unique words that a stimulus received from all descriptions (stop words excluded), and then asked whether the averaged frequency of these unique words varied along with the complexity of the stimulus.

[1]One relevant observation to make here is that the proportion of stop words varies as a function of description length: At least in our dataset, longer descriptions tended to have a great proportion of stop words. This, in turn, means that excluding stop words decreases the average word frequency for longer descriptions more than it does for shorter descriptions, and could be the reason why there is such a drastic reversal in the observed pattern once these words are removed.

Exp.1 (Shapes): A linear model with skeletal surprisal as a predictor explained a significant proportion of variance in average word frequency of descriptions, with the two quantities being negatively related ($R^2 = 0.11$, $F(1, 158) = 20.12$, $p = 1.39 \times 10^{-5}$). However, a quadratic model explained a greater proportion of variance ($R^2 = 0.15$, $F(2, 157) = 13.41$, $p = 4.19 \times 10^{-6}$), including a significant quadratic term, $b = 1.13 \times 10^{-5}$, $95\%CI = [1.04 \times 10^{-4}, 1.32 \times 10^{-3}]$, $t = 2.46$, $p < 0.05$).

Exp.2 (Matrices): Neither a linear nor a quadratic model with randomness as a predictor reached significance, and so stimulus complexity explained almost none of the variance in word frequency (linear model: $R^2 = 0.020$ $F(1, 138) = 2.83$, $p = 0.09$; quadratic model: $R^2 = 0.025$; $F(2, 137) = 1.77$, $p = 0.17$).

Exp.3 (Motion): A linear model with # of direction changes as a predictor explained a significant but small proportion of variance in average word frequency of descriptions, with the two quantities being *negatively* related($R^2 = 0.053$, $F(1, 118) = 6.64$, $p = 0.011$). However, a quadratic model explained a similarly small portion of variance ($R^2 = 0.059$, $F(2, 117) = 3.67$, $p < 0.05$), and the quadratic term was not significant, $b = 1.24 \times 10^{-5}$, $95\%CI = [-1.67 \times 10^{-5}, -4.15 \times 10^{-5}]$, $t = 0.84$, $p = 0.4$).

Here, Experiment 1 and 3 showed a negative relationship relationship between these two variables. Neither a linear nor a quadratic model relating stimulus complexity to word frequency reached statistical significance for matrices.

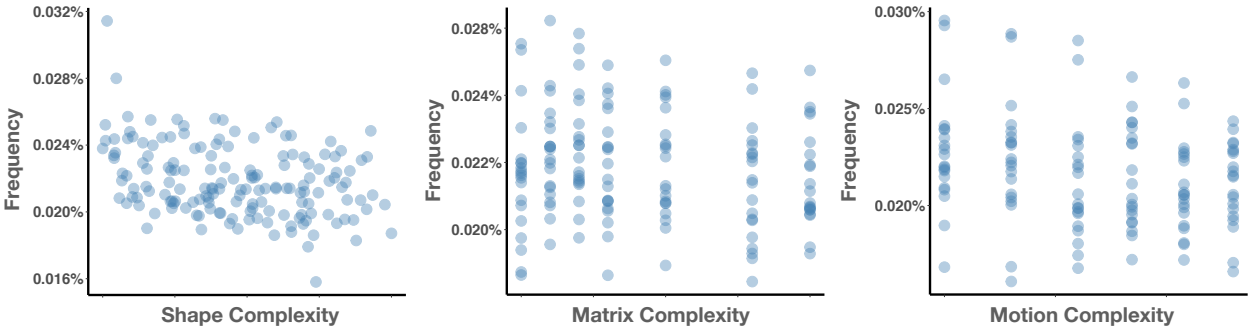

Figure 3. In the "unique words" analysis, we averaged together the frequencies of all the unique words that a given stimulus received, and plotted those average frequencies against the complexity of the stimulus.

2 Local frequency

All of the above analyses compute word frequency by cross-referencing the relevant words in a corpus of English. However, it is not entirely clear that this is the best approach for our research question, since some words that are relatively infrequent in English as a whole might actually be quite frequent in the context of our experiment (e.g., words like "triangle"). So, in addition to the word frequencies computed from an English corpus, we also computed those words' "local frequencies" with respect to the "corpus" comprising all the words spoken in our experiments (a total of 374,157 words across all spoken descriptions in all three experiments). Other than this difference, all of the analyses below were carried out in the same way as with the "global" corpus above

2.1 All words

Exp.1 (Shapes): A linear model with skeletal surprisal as a predictor explained a significant proportion of variance in average word frequency of descriptions, with the two quantities being negatively related ($R^2 = 0.27$, $F(1, 158) = 58.96$, $p = 1.58 \times 10^{-12}$). However, a quadratic model was also significant ($R^2 = 0.31$, $F(2, 157) = 35.05$, $p = 2.59 \times 10^{-13}$), including a significant quadratic term, $b = -5.11 \times 10^{-3}$, $95\%CI = [-8.58 \times 10^{-3}, -1.62 \times 10^{-3}]$, $t = -2.90$, $p < 0.01$).

Exp.2 (Matrices): A linear model with matrix randomness as a predictor explained a significant proportion of variance in average word frequency of descriptions, with the two quantities being negatively related ($R^2 = 0.69$, $F(1, 138) = 313.1$, $p = 2.20 \times 10^{-16}$). Here, the quadratic model explained a similar proportion of variance ($R^2 = 0.70$, $F(2, 137) = 156.5$, $p = 2.20 \times 10^{-16}$), and the quadratic term was not significant ($b = -1.11 \times 10^{-3}$, $95\%CI = [-3.89 \times 10^{-3}, 1.67 \times 10^{-3}]$, $t = -0.79$, $p = 0.43$).

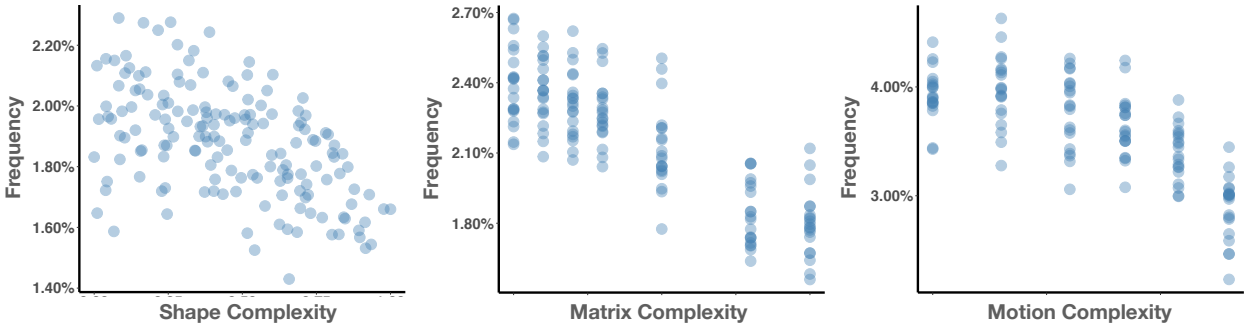

Figure 4. For each word, its frequency is how often it appears in our database, and for each description, the averaged word frequency is calculated from all its words, including stop words.

Exp.3 (Motion): A linear model with # of direction changes as a predictor explained a significant proportion of variance in average word frequency of descriptions, with the two quantities being negatively related ($R^2 = 0.52$,$F(1, 118) = 127.5$, $p = 2.20 \times 10^{-16}$). However, the quadratic model explained a greater proportion of variance ($R^2 = 0.62$, $F(2, 117) = 97.06$, $p < 2.20 \times 10^{-16}$), and the quadratic term was significant ($b = -8.93 \times 10^{-3}$, $95\%CI = [-0.012, -0.0058]$, $t = -5.70$, $p = 8.96 \times 10^{-8}$).

Across these three experiments, more complex stimuli tended to receive less frequent words, as can be seen in the graphs above. This might suggest that, even with a plateau or reduction in verbal description length, descriptions of complex stimuli pack "more information" into fewer words.

2.2 Stop words excluded

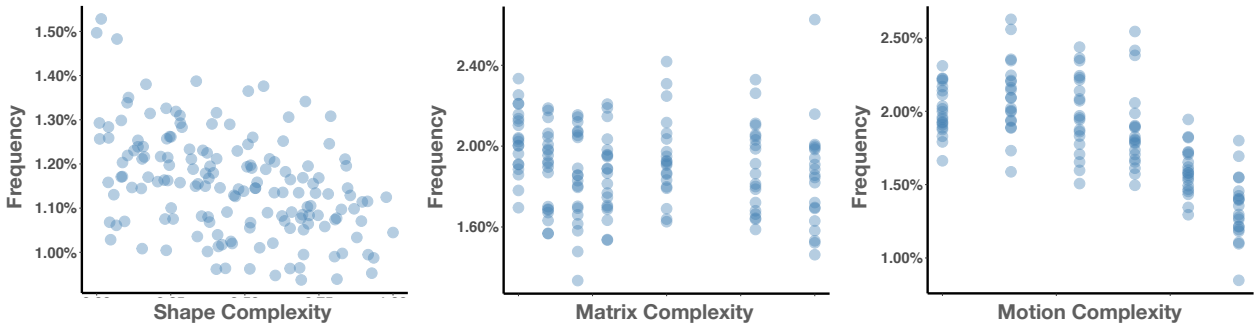

Figure 5. For each word, its frequency is how often it appears in our database, and for each description, the averaged word frequency is calculated from all its words, excluding stop words.

Exp.1 (Shapes): A linear model with skeletal surprisal as a predictor explained a significant proportion of variance in average word frequency of descriptions, with the two quantities being negatively related ($R^2 = 0.19$, $F(1, 158) = 36.73$, $p = 9.61 \times 10^{-9}$). A quadratic model was also significant ($R^2 = 0.19$, $F(2, 157) = 18.74$, $p = 5.01 \times 10^{-8}$), but the quadratic term itself did not reach significance, $b = -1.11 \times 10^{-3}$, $95\%CI = [-1.33 \times 10^{-3}, 3.54 \times 10^{-3}]$, $t = 0.90$, $p = 0.37$).

Exp.2 (Matrices): Neither a linear nor a quadratic model reached significance, and so matrix randomness explained almost none of the variance in word frequency (linear model: $R^2 = 0.013$, $F(1, 138) = 1.81$, $p = 0.18$; quadratic model: $R^2 = 0.017$, $F(2, 137) = 1.19$, $p = 0.31$).

Exp.3 (Motion): A linear model with # of direction changes as a predictor explained a significant proportion of variance in average word frequency of descriptions, with the two quantities being negatively related ($R^2 = 0.46$,$F(1, 118) = 98.96$, $p = 2.20 \times 10^{-16}$). However, the quadratic model explained a greater proportion of variance ($R^2 = 0.59$, $F(2, 117) = 83.95$, $p = 2.20 \times 10^{-16}$), and the quadratic term was significant ($b = -7.79 \times 10^{-3}$, $95\%CI = [-0.010, -0.0053]$, $t = -6.16$, $p = 1.06 \times 10^{-8}$).

Here, Experiments 1 and 3 showed that the complexity of stimuli negatively correlated with word frequency. However, excluding stop words from the analysis appeared to change the pattern observed in

Experiment 2.[2]

2.3 Unique words

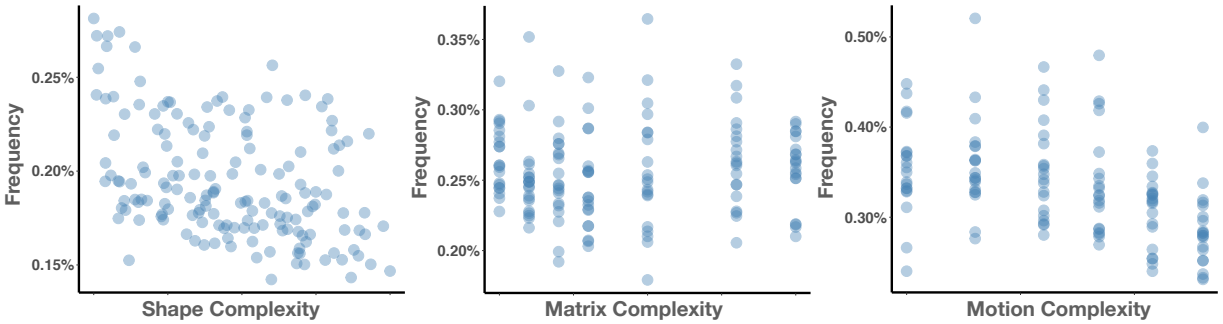

Figure 6. For each word, its frequency is how often it appears in our database, and for each stimulus, its frequency is calculated from all the unique words it received.

Exp.1 (Shapes): A linear model with skeletal surprisal as a predictor explained a significant proportion of variance in average word frequency of descriptions, with the two quantities being negatively related ($R^2 = 0.17$, $F(1, 158) = 32.6$, $p = 5.46 \times 10^{-8}$). However, a quadratic model explained a similar proportion of variance ($R^2 = 0.19$, $F(2, 157) = 18.65$, $p = 5.40 \times 10^{-8}$), including a significant quadratic term, $b = 6.69 \times 10^{-4}$, $95\%CI = [1.40 \times 10^{-5}, 1.32 \times 10^{-3}]$, $t = 2.02$, $p = 0.045$).

Exp.2 (Matrices): Neither a linear nor a quadratic model even reached significance, and so matrix randomness explained almost none of the variance in word frequency (linear model: $R^2 = 0.003$, $F(1, 138) = 0.47$, $p = 0.49$; quadratic model: $R^2 = 0.011$, $F(2, 137) = 0.77$, $p = 0.47$).

Exp.3 (Motion): A linear model with # of direction changes as a predictor explained a significant proportion of variance in average word frequency of descriptions, with the two quantities being negatively related ($R^2 = 0.19$,$F(1, 118) = 28.07$, $p = 5.51 \times 10^{-7}$). And the quadratic model explained a greater proportion of variance ($R^2 = 0.24$, $F(2, 117) = 18.42$, $p = 1.11 \times 10^{-7}$), including a significant quadratic term ($b = -1.24 \times 10^{-3}$, $95\%CI = [-1.90 \times 10^{-4}, 3.31 \times 10^{-4}]$, $t = -2.70$, $p < 0.01$).

3 Conclusion

Are shorter descriptions caused by rarer or more informative words? As shown above, it's hard to say! Choices about which words to include/exclude, or which corpus to use, can lead to uncertain or even contradictory findings. Moreover, the (un)reliability of automated speech transcription poses a unique challenge to this sort of analysis (in ways that apply less, or not at all, to the main analyses in our paper). These patterns prevent us from reaching a firm conclusion about the relationship between word frequency and complexity. Future experiments that are explicitly designed to address this question might enable further progress on this (tricky) problem.

References

Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

Speer, R., Chin, J., Lin, A., Jewett, S., and Nathan, L. (2018). LuminosoInsight/wordfreq: v2.2. *journal*.

[2]Another stop-words-related observation: Here (in the "local frequency" analysis), excluding stop words seemed to have a less dramatic effect on the overall pattern than it did earlier (in the "global frequency" analysis. We suspect this is because the difference in frequency between stop words and other content words (e.g., the difference in frequency between "with" and "triangle") is much greater in the context of a naturalistic corpus of English (where "with" appears much more often than "triangle") than here in our experiments (where "with" appears very frequently, but "triangle" does too). For this reason, excluding stop words may have been less consequential here than earlier.