

Zipf’s law is a consequence of coherent language production

Jake Ryland Williams,^{1,*} James P. Bagrow,^{2,†} Andrew J. Reagan,^{2,‡} Sharon E. Alajajian,^{1,§} Christopher M. Danforth,^{2,¶} and Peter Sheridan Dodds^{2,**}

¹*School of Information, University of California, Berkeley 102 South Hall #4600 Berkeley, CA 94720-4600.*

²*Department of Mathematics & Statistics, Vermont Complex Systems Center, Computational Story Lab, & the Vermont Advanced Computing Core, The University of Vermont, Burlington, VT 05401.*

(Dated: August 9, 2016)

The task of text segmentation may be undertaken at many levels in text analysis—paragraphs, sentences, words, or even letters. Here, we focus on a relatively fine scale of segmentation, hypothesizing it to be in accord with a stochastic model of language generation, as the smallest scale where independent units of meaning are produced. Our goals in this letter include the development of methods for the segmentation of these minimal independent units, which produce feature-representations of texts that align with the independence assumption of the bag-of-terms model, commonly used for prediction and classification in computational text analysis. We also propose the measurement of texts’ association (with respect to realized segmentations) to the model of language generation. We find (1) that our segmentations of phrases exhibit much better associations to the generation model than words and (2), that texts which are well fit are generally topically homogeneous. Because our generative model produces Zipf’s law, our study further suggests that Zipf’s law may be a consequence of homogeneity in language production.

PACS numbers: 89.65.-s, 89.75.-k, 89.70.-a

In our previous work on text partitioning [1], and again in our study on the effects of text mixing [2], we have observed the relevance of the selection model proposed long ago by Simon [3]. We will build on this model again, together with Zipf’s law [4, 5], as the basis for a model of frequency data of text.

The defining aspects of Simon’s model for language generation hold that as the terms of a text appear, they do so independently, and in proportion to their historic frequencies of occurrence. The first assumption (of memoryless term-term independence) is likewise the basis for the bag-of-terms model that is pervasive in the computational text analysis community. It is evident that this assumption fails for strictly defined words when one considers the dependence of terms that are bound, like *New York*, and *City*. As we have argued in [2], dependence can also be caused by mixing, where the rate of a term’s occurrence varies in a composite corpus of many texts as one transitions from document to document.

The two types of dependence indicate that for the Simon model to apply, and consequently, for the bag-of-terms model to apply, texts should be analyzed on the homogeneous (unmixed) scales of production (i.e., by writer, or even topic), and that the terms of texts should be segmented according to their local dependence (i.e., grouped together into irreducible expressions), rendering distributions of terms of mixed sizes. The first criterion can be relatively easy to satisfy, and in our analyses will come for free by the curation of the Project Gutenberg eBooks [6] database. The second, however, poses a much greater challenge, as the meanings that bind words into meaning-irreducible forms are only known a priori to the

writers who generate text for their meanings.

In our work on stochastic text partitioning [1], we proposed a simple mechanism for the fine-grained chunking of text into phrases. Under this framework, sentences were broken down stochastically, splitting on whitespace with a uniform bond-breakage probability, q . While this (uniform) framework is clearly not the ideal mechanism to isolate phrases for, say, dictionary lookups, we did find that random partitioning of phrases ($q = \frac{1}{2}$) produced Zipf laws extending many orders of magnitude in rank beyond the standard version for words only.

Here, we wish to explore informed partitions, taking into account empirical segmentation information. Some interesting options for estimating bond-breakage frequencies exist—in language learning environments, similar data are generated through exercises referred to as phrase-cued text, which have been shown to help learners develop prosody [7] (the patterns of stress and intonation, considered to be one of the essential features of reading fluency [8]). This suggests value in the phrase-scale of segmentation.

To obtain phrases, surveys (essentially classroom exercises) could be run that would benefit educators and students, while developing a base of training data for informed partitions. Ideally, such surveys would have students learning a language asked to segment their *own* writing. However, this is unconventional in the classroom, and poses a more difficult task to request. So, for now we will turn to expertly-segmented data.

In the computational linguistics (CL) community there have been recent efforts to establish benchmarks and testing sets for multiword expressions (MWEs) (loosely

defined as groups of “tokens in a sentence that cohere more strongly than ordinary syntactic combinations”) extraction [9]. While these data can be used to inform a partitioner, they are alone insufficient (narrowly focused on business reviews), and are better used for testing and benchmarking. Furthermore, it is not clear if MWEs and their extraction (used for work in information retrieval and other CL tasks) coincide with language generation and distributional representations, which are the focus, here. We have also executed a companion piece to this study [10], focused on the task of MWE extraction, where the text partitioning framework was tuned to exhibit its ability to perform as strongly, if not better than, the state-of-the-art [11], with very substantial increases in speed, while being trained on less data.

For training, we turn to the bond-breakage information held latently inside of the Wiktionary [12] with its trove of larger-than-word entries and their example usages. Noting that example usages of phrases in the Wiktionary are consistently coded in boldfaced text, we are able to resolve isolated empirical bond-breakage frequency data. For example, at the time of writing this letter, the entry for the phrase “have a ball” was represented by the usage:

The kids **had a ball** playing in the fountain.

which informed us that (kids, had) is a pair more likely split, and that (had, a) is a pair more likely bound. The Wiktionary examples have the added bonus of being trustworthy, on account of their being archetypal-usage examples, and will serve to inform our partitions for all of the experiments in this letter. However, to make informed text partitions available in many other languages (where examples may be less prolific and/or coded differently), we also construct bond-breakage frequencies in a similar manner, but from the embedded hyperlinks on Wikipedia, which are ubiquitous and coded-for consistently across languages. From either source of data, we define the partitioning probability, $q(w_\ell, w_r)$, for an ordered pair of words, (w_ℓ, w_r) , according to the total number of known bond breaking appearances $f_b(w_\ell, w_r)$ and the total number of bond preserving appearances $f_p(w_\ell, w_r)$:

$$q(w_\ell, w_r) = \frac{f_b(w_\ell, w_r)}{f_p(w_\ell, w_r) + f_b(w_\ell, w_r)},$$

with a default value of $q(w_\ell, w_r) = 1$ for all pairs unobserved in a training set. Note that this model is still quite naïve, and may be improved by allowing values of q to vary, depending on terms existing farther away in a text. However, training a more sophisticated partitioner like this would be done best if the training data were fully segmented (which might be accomplished with phrase-cued text).

We note that both sources (examples and hyperlinks) have a reasonable degree of accord (considering English),

with the resulting q -values falling on the same side of $q = 0.5$ for $\sim 65\%$ of all unique word pairs (which later will be all that is necessary to ensure equivalent partitions), and as much as 75% for those most frequently appearing. While the two sources share coverage on about 20,000 word pairs (and disagree on a substantial number), we note that the Wikipedia hyperlinks source is much larger, with coverage on many proper nouns (names of places, people, etc.), which were found to be important for MWE identification in [10].

In our original work on text partitioning [1], we considered stochastic text partitions, where bonds between pairs words would either be preserved or broken at random (uniformly). Despite our solving for the analytic expectation across all possible partitions of phrase-partition frequencies, it became necessary to execute one-off random partitions for the study of true rank-frequency distributions. However for CL tasks, these stochastic partitions have little value on account of their variation, and while the variation in the case of informed stochastic partitions is reduced, it is still not ideal, as CL tasks require consistency.

A perhaps unexpected bonus to the naïveté of our assumptions (specifically, defining the q ’s to be independent of one another) is that the maximum likelihood partition (MLP), or most frequently occurring partition of a text, is easily computed. Accepting this partition (for a text) is deterministic, and accomplished simply by breaking a bond whenever its bounding pair of words have a breakage probability of at least 0.5—note that this conservatively sets the ill-determined case of $q = 0.5$ to bond-breaking. Using this scheme, a text can be segmented very rapidly based on the Wiktionary (with the same computational complexity as simple word count), and requires no global information. For all of these reasons—speed, determinism, and maximum likelihood—we continue with the MLP in this work.

We now return to Simon’s model of language generation, which analytically produces single-parameter power-law rank-frequency distributions, and is our hypothesis as the dominant productive mechanism for Zipf’s law. Zipf’s law succinctly formulates a relationship between the frequencies of occurrence of terms, and their ranks by frequency (descending):

$$f = C \cdot r^{-\theta}, \quad (1)$$

for a scaling constant, C . Simon’s model has one tunable parameter—the rate of term introduction, α . When language is generated by Simon’s model, Zipf’s scaling exponent is given by $\theta = 1 - \alpha$, and in the case of an empirical text, the term introduction rate can be inferred as $\alpha = N/M$, where N is the number of unique term types in the text, and M is the total number of all terms occurring in a text.

The main caveat for the compatibility of Zipf’s empirical law with Simon’s theoretical model arises from the

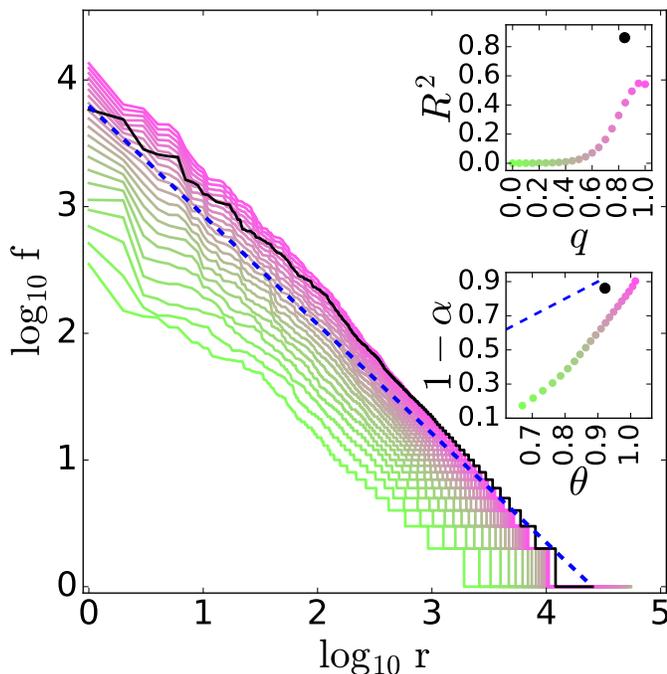


FIG. 1: The main axes show rank-frequency distributions from Herman Melville’s “Moby Dick” for one-off uniform stochastic partitions for values of q ranging from 0 (green) to 1 (pink), together with a one-off, dictionary-informed stochastic partition (black) and its Zipf/Simon model (blue, dashed). In the upper inset we present the Zipf/Simon goodness of fit as a function of q (with the black point for the informed partition positioned at the text’s average value of q), and in the lower inset we present regressed scalings (θ) against those determined by the Zipf/Simon model ($1 - \alpha$). In the lower inset, complete agreement occurs along the (blue dashed) line $1 - \alpha = \theta$, and for both insets, colors of points correspond to those in the main axes.

fact that Simon’s model is only able to generate single-scaling rank-frequency distributions with scaling parameters θ in $[0, 1)$. While many texts exhibit multiple scalings and values of θ larger than 1, we have seen that these anomalies are often the result of text mixing and term dependence [2]. Consequently, we view the degree to which a particular Zipf/Simon model aligns with empirical rank-frequency data as a measure of the goodness of fit for the ‘bag-of-terms’ representation (tokenization) of a text (regardless of how the terms are defined, i.e., words or phrases).

The Zipf/Simon fit for a text of N unique and M total words is defined as follows. Assuming a text was generated precisely according to Simon’s model, the constant word introduction rate is inferred by the text-wide average: $\alpha = N/M$, and an estimate of the scaling exponent is then $\theta_{\text{mod}} = 1 - \alpha$. The exact form of the model’s fit is then obtained by computing the constant of proportionality: $C_{\text{mod}} = N^{\theta_{\text{mod}}}$, whereupon we may take the coefficient of determination, or R^2 , as a measure of

goodness of fit for the Zipf/Simon model, and as a result of the Simon model’s independence assumption, a measure of the quality of a text’s bag-of-terms representation. Note that this will also give us an opportunity to evaluate the qualitative nature of language produced by Simon’s model.

Since the connection of Simon’s model to the parameter θ occurs naturally in the complimentary cumulative distribution function (CCDF) of term frequencies, we measure and regress all quantities along CCDFs, while we present all results in the intuitive and familiar rank-frequency (Zipf) representations.

In Fig. 1, we plot an example informed partition for Herman Melville’s “Moby Dick” (black line, main axes), and behind it, the spectrum of uniform stochastic partitions ranging from $q = 0$ (green) sentences/clauses, to $q = 1$ (pink) words. One can see the distributions steepen along the gradient as q is increased (which is generally the trend throughout the eBooks), and as much can be seen in the lower inset, where a regressed scaling parameter θ is compared with $\theta_{\text{mod}} = 1 - \alpha$. Also within the main axes lies the Zipf/Simon fit for the the informed partition (blue, dashed), which appears reasonable by inspection, and outperforms all uniform partitions (including the $q = 1$ word partition) by goodness of fit of the Zipf/Simon model (upper inset), where the informed partition is indicated by the black point (in both insets). Note that this increased quality of fit by the Zipf/Simon model is also observed readily in the lower inset, where θ and $1 - \alpha$ nearly align.

Zooming out to the larger collection of (roughly 20,000) English eBooks, we focus on comparing two partitions for each text—the $q = 1$ (word) partitions, and the informed (phrase) partitions. In Fig. 2, we compare the goodness of fit by the Zipf/Simon model between the two partition types. This comparison divides the collection of books into two subsets that we will refer to as the word-based and phrase-based texts. Our demarcation is given by the line $R^2_{q=1} = R^2_{\text{MLP}}$ (red, dashed), and shows that most texts were better bag-models under the informed (phrase-based, $R^2_{q=1} < R^2_{\text{MLP}}$) partition framework. While a good number of texts ($\sim 20\%$) are word-based, we note that these texts tend to have weaker R^2 values in general, which can be seen in Fig. 3 where we apply an R^2 cutoff to the books and measure the portion of the data set removed. This quantifies the numbers of phrase-and word-based texts that are strong fits ($R^2 > 0.6$) to be in a proportion of more than 12:1, respectively, indicating that a strongly-fit bag-of-terms representation of a text is better (and generally quite significantly) achieved by MLP phrases more than 90% of the time.

From Fig. 3, we can see that $R^2 \sim 0.6$ appears near a sharp drop for the cutoff, which, if applied as a threshold for analysis, results in approximately a 20% loss (red line/stars) for the dataset (largely due to the low- R^2 ,

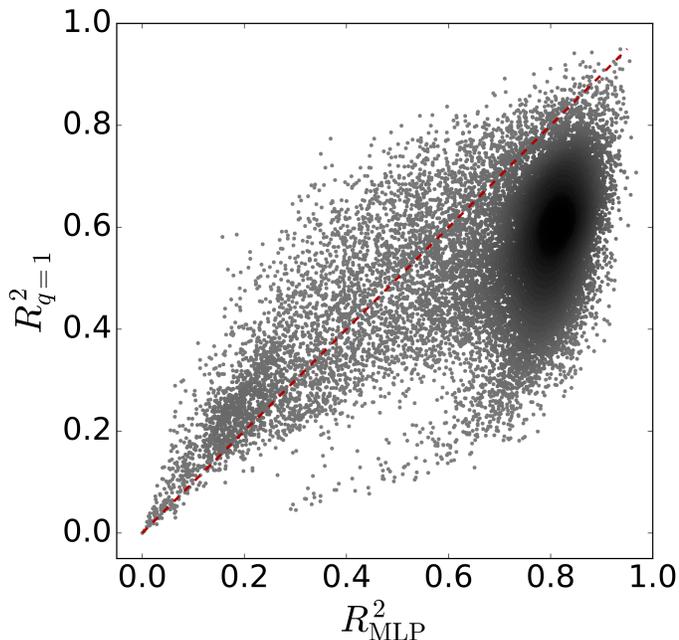


FIG. 2: We plot the goodness of fit (R^2) of the Zipf/Simon model, applied to the texts of the English Project Gutenberg eBooks database, where the dictionary-informed one-off partitions (horizontal axis) are plotted against the $q = 1$ word partitions (vertical axis). The discriminating line (red, dashed, $R^2_{MLP} = R^2_{q=1}$) helps divide the collection into texts that are word-based and phrase-based.

word-based texts). By allowing the books to be either phrase- or word-based (according to their R^2 values), we are then able to accommodate a strong ($R^2 \geq 0.6$) ‘bag’ analysis for over 80% of the dataset, which is a large improvement over the conventional, blanket usage of words (dashed blue line/open circles), which only accommodates strong fits for approximately 37.5% of the collection.

While one can view the $\sim 20\%$ of poorly-fit books as a loss to quality analysis, it is possible that refinement of the partition function (through better data or modeling) would allow one to obtain better tokenizations/fits for all books. However, one can also look at these 20% in a different way—as potentially being unfit for the bag-of-terms framework. Sorting the eBooks according to $\max\{R^2_{q=1}, R^2_{MLP}\}$, we find the ten poorest fits to include dictionaries, spelling books, and books of extremely small size (often as placeholders for other media), indicating that low R^2 values appropriately identify texts unfit for analyses.

To explore the qualitative nature of fit quality and the Zipf/Simon model, we also consider the variation of R^2_{MLP} (strictly, as $R^2_{q=1}$ is so frequently outperformed) across the different Library of Congress classifications (Fig. 4, top), and subjects (Fig. 4, bottom) provided in the eBooks collection. This variation can be seen in

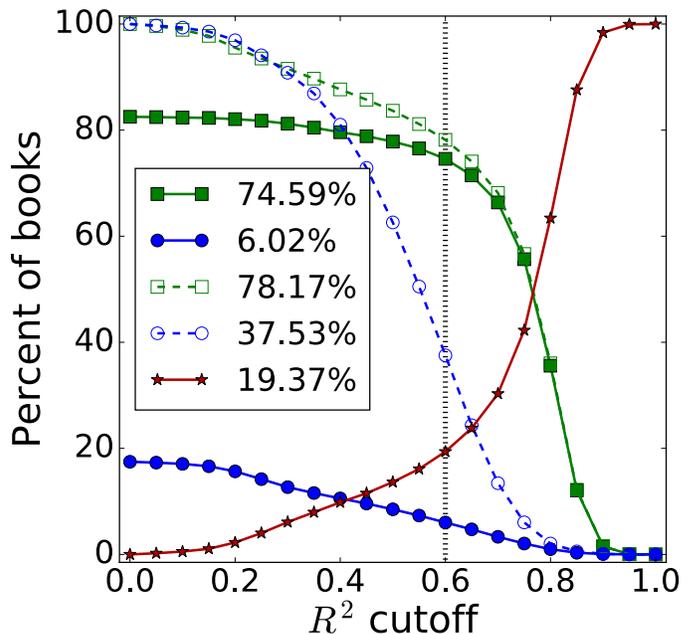


FIG. 3: The percent of books in the Project Gutenberg English eBooks database that are phrase-based (solid green line, filled squares), word-based (solid blue line, filled circles), and poorly fit (red line, stars), when a cutoff in R^2 is applied (right). A dropoff in R^2_{MLP} occurs near $R^2 \sim 0.6$, which is denoted by the vertical, black dotted line. We also present curves that indicate the percents of books remaining above the cutoff when blanketed usage of either words (blue dashed line, open circles) or informed phrases (green dashed line, open squares) are applied.

Fig. 4, where we observe stark differences in the ranges of box plots. As the majority of texts have an MLP representation that is relatively strong ($R^2_{MLP} > 0.6$), the medians for most classifications are high. However, it is clear that texts under classification A—general works, which includes dictionaries, encyclopedias, and periodicals—are more often poorly-fit. This is of note, as all other classifications, including textbook-reference classifications, such as medicine (R), law (K), and naval (V) and military (U) science, exhibit medians substantially above 0.6, though we note that many classifications exhibit outliers of low R^2_{MLP} , notably with language and literature (P), which contains long, brief, and mixed texts.

The main difference we note between classification A and the other reference classifications mentioned is that the general works of A are frequently not topically homogeneous. This observation is echoed in the bottom of Fig. 4, where we consider the eBooks’ subjects. Here, we again see low values for dictionaries and periodicals, but likewise for other heterogeneous subjects, such as questions and answers, travel, and social life. We also find that poetry, short stories, and science fiction all exhibit a large proportion of low R^2_{MLP} values. These subjects all

share the commonality of frequent brevity. This makes some sense in relation to the Simon model, as the generation of a text may not have an opportunity to stabilize (distributionally). We also note that the science fiction subject may frequently use phrases that are outside of the informed partitioner’s training.

Our results show that informed text partitioning is capable of improving the basic methodologies of feature selection in text analysis on a broad scale, allowing for better adherence to assumptions (specifically, for the bag-of-terms model) that underpin a vast collection of algorithms currently in practice throughout industry and academia. Additionally, phrase-based text analysis improves the independent interpretability of features—in the soft sense, at the level of user experience. With phrase-based text analysis, end-point users (e.g., policy makers or product users interpreting lists of phrase-feature topics from a topic model readout) who may not understand the workings of an algorithm will be better able to interpret results, as phrases provide critical context for interpretation. We have also proposed the measurement of association to a model of language generation, which we have found here to be the likely mechanism for the production of topically-homogeneous language, allowing us to assert new meaning for the presence of Zipf’s law in natural language—long observed, though largely a mystery. We also suggest the possibility that as phrase-segmentation methods improve, association to the Zipf/Simon model might also be used as a metric for determination of topical breaks in text, since R_{MLP}^2 does appear to increase in the presence of topically-homogeneous texts.

To make these tools both explorable and available to the computational text analysis community, we have with this letter developed a Python package for text partitioning (for detailed information, see <https://github.com/jakerylandwilliams/partitioner>), which makes our tools available in 11 languages (English, German, Russian, Portuguese, Polish, Dutch, Italian, French, Finnish, Spanish and Greek), and in addition present an explorable online appendix (<http://jakerylandwilliams.github.io/partitioner/>).

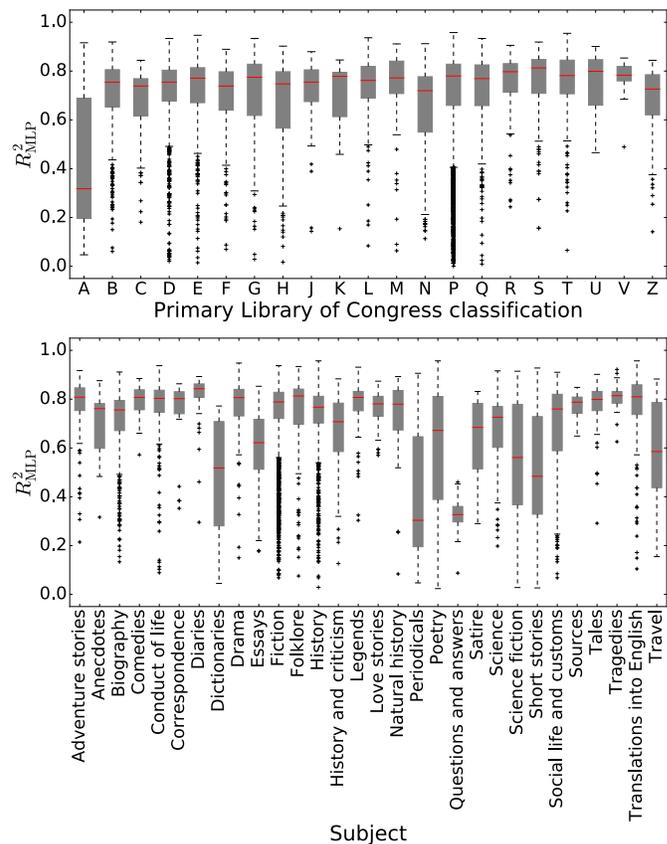


FIG. 4: Box plots, showing variation of R_{MLP}^2 across the primary Library of Congress classifications (top), and a selection of the Project Gutenberg eBooks subjects (bottom). All boxes represent at least 50 texts.

Dodds, CoRR (2015), <http://arxiv.org/abs/1409.3870>.

- [3] H. A. Simon, *Biometrika* **42**, 425 (1955).
- [4] G. K. Zipf, *The Psycho-Biology of Language* (Houghton-Mifflin, 1935).
- [5] G. K. Zipf, *Human Behaviour and the Principle of Least-Effort* (Addison-Wesley, 1949).
- [6] Project Gutenberg (2010), <http://www.gutenberg.org>.
- [7] M. J. Glavach, *The Practical Teacher* (2011).
- [8] J. Miller and P. J. Schwanenflugel, *Journal of Educational Psychology* **98**, 839 (2006).
- [9] N. Schneider, S. Onuffer, N. Kazour, E. Danchik, M. T. Mordowanec, H. Conrad, and N. A. Smith, in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)* (European Language Resources Association (ELRA), Reykjavik, Iceland, 2014), ISBN 978-2-9517408-8-4.
- [10] J. R. Williams, CoRR (2016), URL <http://people.ischool.berkeley.edu/~jakeryland/documents/williams2016b.pdf>.
- [11] N. Schneider and N. A. Smith, in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Association for Computational Linguistics, Denver, Colorado, 2015), pp. 1537–1547, URL <http://www.aclweb.org/anthology/N15-1177>.
- [12] Wiktionary (2014), <http://dumps.wikimedia.org/enwiktionary/2014>.

* Electronic address: jakerylandwilliams@berkeley.edu

† Electronic address: james.bagrow@uvm.edu

‡ Electronic address: Andrew.Reagan@uvm.edu

§ Electronic address: sharon@ischool.berkeley.edu

¶ Electronic address: chris.danforth@uvm.edu

** Electronic address: peter.dodds@uvm.edu

- [1] J. R. Williams, P. R. Lessard, S. Desu, E. M. Clark, J. P. Bagrow, C. M. Danforth, and P. S. Dodds, CoRR **abs/1406.5181** (2015), <http://arxiv.org/abs/1406.5181>.
- [2] J. R. Williams, J. P. Bagrow, C. M. Danforth, and P. S.