QUANTIFYING THE SEMANTIC VALUE OF WORDS

DILLON NIEDERHUT
Department of Anthropology, University of California
Berkeley, USA
dillon.niederhut@berkeley.edu

Hypotheses about the evolution of human language often posit a role for the informativeness of speech acts. However, there has yet to be an accessible method for measuring the semantic value of a word or group of words. This paper outlines a novel test statistic for determining the relative semantic value of words in a language, given a corpus that approximates ecologically valid use of that language. Future work could use this test to disambiguate between social and cognitive factors in linguistic change.

1. Introduction

The extent to which human language systems evolved because it was beneficial for early hominins to communicate well has been a topic of heavy debate for some time. While this debate has advanced to the point of some researchers arguing that communicating poorly might be even more beneficial than doing it well, there has yet to be a method for measuring the communicative worth of a speech act (Pinker, Nowak, & Lee, 2008).[a] Were such a method to exist, hypotheses that posit a utility motive for the origin of language could be tested in natural experiments involving the acquisition or change of that language.

2. Natural word use follows a Zipfian distribution

In conversational English, a few words are used many times, and the rest are used with shocking rarity (Fig. 1). Even words that seem relatively common, like CARPET, SCALE, and WEIRD, appear fewer than once in every twenty thousand words. To put this in context, it has been estimated that the average person speaks 16,000 words per day (Mehl, Vazire, Ramirez-Esparza, Slatcher, & Pennebaker, 2007).

Intuitively, words that are used very frequently don't seem to carry much semantic value. The word MY, for example, really only tells you that something is

[a]One can measure the Shannon Information of human language, but high entropy words are not necessarily the same as semantically valuable words.

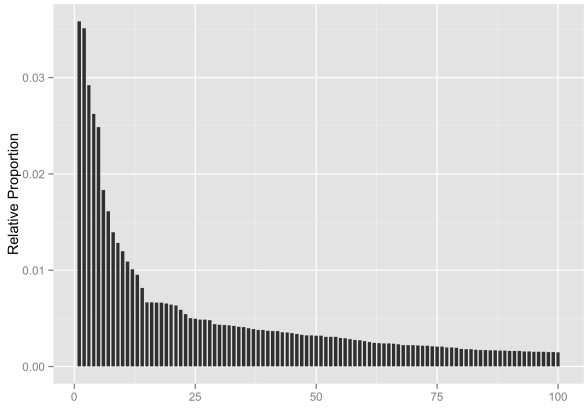

Figure 1. Relative proportion of the 100 most common words in English, from the sample described below.

being possessed by the speaker, but does not tell you anything about that something or about the speaker. To take another example, the word YEAR probably indicates a period of 365 days, but might not have the expected start date (i.e. a fiscal year) or might only refer to nine months out of those 365 days (i.e. an academic year).

On the other hand, uncommon words seem to carry a lot of semantic value. For example, were the speaker to use the word MULTICOLLINEAR, a native speaker of English familiar with statistics could assume the following statements (ordered by decreasing likelihood):

1. the speaker is a fairly educated person who does statistical analyses, speaking to one or more persons who are fairly educated and understand statistcs
2. the speaker, in particular, is well informed about the general linear model
3. the context under discussion is a model with two or more predictors that are highly correlated, and thus has unstable linear coefficients
4. the speaker will go on to mention ways to measure this, like the variance inflation factor, and ways to correct it, like principal component analysis
5. the speaker does not work with very large datasets or machine learning methods

In this particular case, a single word is giving us a rich set of inferences about who is involved in the conversation, their knowledge state, the topic under dis-

cussion, and what will happen next. It should be obvious that this is much more information than was provided by the words MY and YEAR.

However, *context* is difficult to define and even harder to measure. To continue the example above, measuring the education level of every person involved in a conversation, along with the major topic under discussion and the fields in which the speaker does not work in a dataset large enough to draw meaningful inferences about language is not currently feasible. However, the fourth point above – measuring the frequency of associated words like VARIANCE INFLATION FACTOR – is easily and frequently implemented.

3. Semantic value is a change in that distribution

We can imagine, then, that a rare word is one which refers to a rare context; and conversely, that rare contexts tend to be described with rare words. In that case, we may posit that the semantic value of a word is related to the distance between the words associated with it relative to their frequency of use in the language as a whole (Wittgenstein, 1953; Salton, Wong, & Yang, 1975; Deerwester, Dumais, Landauer, Furnas, & Harshman, 1990). To state this another way, an informative word is one which changes the relative proportions of the words surrounding it each time it appears.

This argument differs from prior work which assumes that the semantic content of a word is defined by the words that appear nearby (Gentner, 1983; Furnas, Landauer, Gomez, & Dumais, 1983). The model in this paper is the nearby words point to the same latent variable, which is the difficult-to-measure real-world context of the speech act. The implementation, however, is very similar to modern methods in computational semantics like pairwise mutual information or latent semantic analysis (LSA) (Lund & Burgess, 1996; Turney & Pantel, 2010). The difference here is that we are not trying to assign a similarity between the distributions of one single word versus another single word; we are attempting to assign a single value that demonstrates the relationship between a single word and the distribution of all words in the English language.

We can calculate the distance between the word distribution of all events that contain a single word of interest with the overall English distribution using the chi squared test:

$$\sum_{i}^{k} \frac{\left(p_{sample} * n_{sample} - p_{population} * n_{sample}\right)^2}{p_{population} * n_{sample}}$$

where k is the list of unique words in the population; i is one word in that list; n_{sample} is the total number of words used in communicative events that also include the word being measured; and, p_{sample} is the probability that any one in n words is word i.

To make this a little more concrete, imagine looking at the informativeness of

the word MULTICOLLINEAR. Communicative events that contain the word MULTICOLLINEAR also tend to contain words like VARIABLE, CLUSTER, and MODEL. These words have very high relative proportions in our sample, but low relative proportions in our population. So, for each of these, we might add something like:

$$\frac{(1E+01-1E-06)^2}{1E-06} = 1E+08$$

to the total sum. However, these events also contain many common words like TO, OF, and A, at close to their relative proportion in the total population of words. Each of these words adds a smaller value to the total sum.

This method of measuring distance is favorable in that the appearance of rare words is heavily weighted. It is disfavorable in that it is also sensitive to the total number of words, *n*. A very common word like MY has a very large *n*, so even small differences between the proportions of a word in the sample and the population produce values that are in the zeroth or first of magnitude. A very uncommon word with a small *n* will frequently produce values that are equal to the expected frequency of each word in the population, which is typically below 1E-05 – five orders of magnitude less. When summed across the number of unique words in the population (the length of *k*), the effect of *n* dominates the calculated value.

4. Correcting the magnitude of that change produces a test statistic

To produce a test statistic from the chi squared values, one first needs to correct the bias produced by the size of the sample. Then, the distribution of the test statistic must be characterized to produce population parameters for the expected mean value and variance. Both of these steps require real word linguistic data that has been decomposed into a distribution of frequency counts.

An English word distribution was created using a Python library written by the author, available at `https://github.com/deniederhut/redicorpus`, by randomly sampling comments from the discussion board at reddit.com/r/AskReddit over a period of 22 weeks. This particular discussion board was chosen both for its high traffic rate and the broad topics and conversational nature of the discussions there. This resulted in a total sample size of 4.29E+06 communicative events, with 1.08E+08 total unigrams and 2.93E+05 unique unigrams appearing more than once.

To describe the sampling distribution of chi square values, comments were randomly sampled from the total corpus twenty times each at the probabilities 1E-02, 1E-03, 1E-04, and 1E-05. In each of the 100 samples, term frequency was set to equal the number of comments, and the chi squared value of the term frequencies in the sample was calculated. As predicted above, the magnitude of the chi squared statistic is dependent on the size of the sample used to calculate the statistic (Fig. 2).

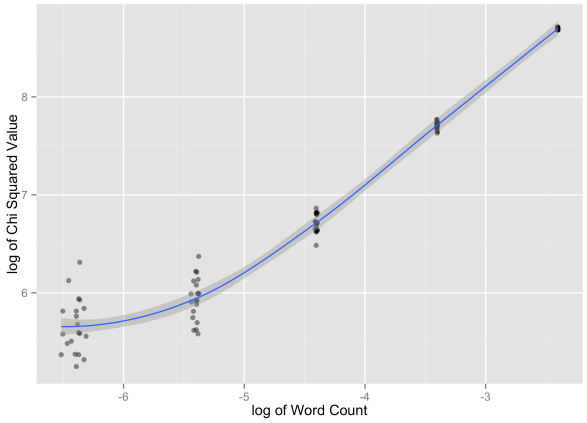

Figure 2. Chi squared values produced by comparing word distributions are convolved with the size of the sample.

The relationship between chi squared values and sample size linearizes when each variable is square root transformed. Happily, this transformation also causes the variability in chi squared values to become constant. After linearization, the chi squared values from each random sample are almost perfectly predicted by the total number of words in all the comments used to compute the value (Fig. 3).

The stable variability means that a simple test value can be created using the chi square value, corrected for word count, and divided by the constant standard deviation. For the sake of brevity, we'll call this the Zipf test.

$$z_{correct} = \frac{(-274.85 + \frac{\sqrt{n}}{6.37}}{288.33}$$

5. The test statistic conforms to expected behavior

As a proof of concept, the Zipf statistic was calculated for several words from the corpus (Table 1). Generally speaking, it produces values concurrent with our intuition. Very common words, like MY, DAY, and FEEL, have negative Zipf statistics that increase in linearly in magnitude with the square root of their relative proportion in the population. Words with a relative proportion around 1E-05 have Zipf statistics that are close to zero, or close to the corrected values derived from randomly sampling the entire population. Words that appear less frequently than this have an increasing change of producing positive Zipf statistics, indicating words with high semantic content.

The Zipf test provides a method for quantifying the relative informativeness

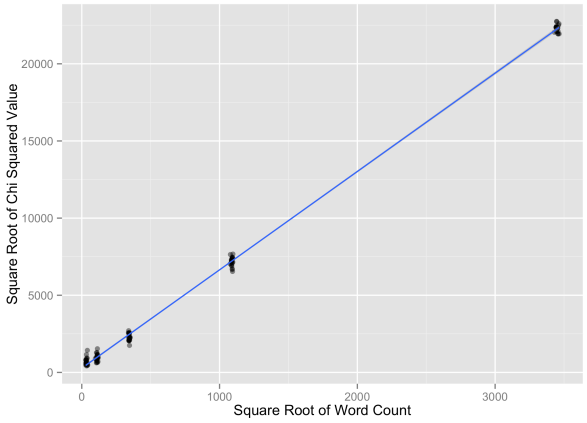

Figure 3. Square-root transforming the chi squared values and sample sizes linearizes both the relationship and the variability in y.

Term	Relative Proportion	Test Value
my	1.30E-02	-9.41E+01
day	2.12E-03	-5.82E+01
feel	1.71E-03	-5.04E+01
record	1.24E-04	-7.13E+00
unclear	1.11E-05	-8.12E-01
hither	6.09E-07	6.80E-01
dill	2.92E-06	8.25E-01
omlette	3.42E-07	1.78E+00
multicollinear	3.69E-08	3.28E+00

of any word, given a language corpus that is divided by communicative events. Additionally, this is an objective method that can be implemented largely without human intervention, and in any language that is easily tokenized or lemmatized. It measures the semantic value of a given word by comparing the frequencies of other words used in the same context with the distribution of all the words that appear in the corpus.

6. Hypotheses made tractable by this test

If the evolution of human language was driven by a need for effective communication, we would expect evolutionary pressures to produce cognitive systems that prioritize the acquisition of information-heavy terms. Specifically, we would

hypothesize that words with high semantic value would:

- be learned sooner in infancy; and,
- spread more quickly through a population; and,
- be preferentially adopted across languages.

More generally, the Zipf test should be useful in testing predictions of language change and use that include social and cognitive factors. For example, one could ask whether the diffusion of linguistic variants is better predicted by the utility of the word, or its use as a marker of social identity (Eisenstein, O'Connor, Smith, & Xing, 2014). It should also be possible to create historical data on the rate of semantic bleaching of words, and to investigate if a relationship exists between that rate and populations employing the word.

7. Supplementary information

Data were collected in Python 2.7.8 on Ubuntu Server 14.0.4, and were analyzed with Revolution R Open[b] based on CRAN release v. 3.2.1, "World-Famous Astronaut", (R Core Team, 2015). Tables were produced with xtable, and figures were produced with ggplot2 (Dahl, 2014; Wickham, 2009). The code and data necessary to reproduce this paper are available at `https://github.com/deniederhut/quantifying-semantic-value`.

Acknowledgements

The author would like to acknowledge Gabe Doyle and Dan Yurovksy for their methodological advice; and Terrence Deacon, Madza Virgens, and Drew Halley for their comments on this manuscript.

References

Dahl, D. (2014). *xtable: Export tables to latex or html.* (R package version 1.7-4)

Deerwester, S., Dumais, S., Landauer, T., Furnas, G., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*, 391–407.

Eisenstein, J., O'Connor, B., Smith, N., & Xing, E. (2014). Diffusion of lexical change in social media. *PLoS One*, *9*, e113114.

Furnas, G., Landauer, T., Gomez, L., & Dumais, S. (1983). Statistical se-mantics: Analysis of the potential performance of keyword information systems. *Bell System Technical Journal*, *62*, 1753–1806.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, *7*, 155–170.

[b] https://mran.revolutionanalytics.com/open/

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, and Computers*, *28*, 203–208.

Mehl, M., Vazire, S., Ramirez-Esparza, N., Slatcher, R., & Pennebaker, J. (2007). Are women really more talkative than men? *Science*, *317*, 82.

Pinker, S., Nowak, M., & Lee, J. (2008). The logic of indirect speech. *Proceedings of the National Academy of Sciences*, *105*, 833-838.

R Core Team. (2015). *R: A language and environment for statistical computing.* Vienna, Austria.

Salton, G., Wong, A., & Yang, C. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*, 613–620.

Turney, P., & Pantel, P. (2010). From frequency to meaning; vector space models of semantics. *Journal of Artificial Intelligence Research*, *37*, 141–188.

Wickham, H. (2009). *ggplot2: elegant graphics for data analysis.* Springer New York.

Wittgenstein, L. (1953). *Philosophical investigations.* Hoboken: Blackwell.