Zimmermann, Richard

University of Manchester

Richard.zimmermann@manchester.ac.uk

Type of Contribution: Full Paper

Word growth dispersion

A single corpus part measure of lexical dispersion

Abstract

This paper introduces a measure of lexical dispersion across corpora, called *Dispersion of Word Growth* (D_{WG}). The following summarizes the essence of the calculation of the measure. First, determine the *Word Growth* (*WG*), an ordered set of the positions of all occurrences (*n*=the number of occurrences of the word) in a corpus (*N*=the size of the corpus).

(1) *WG* = {position(1), position(2), …(position(*n*)}

Second, get the distances between every occurrence starting from the second position.

(2) Distances = position(i+1) - position(i), $1 < i \leq n$

Third, calculate the deviations of the distances (*DD*) from a perfectly evenly distributed word, for which all occurrences would be equally spaced out at distance *N*/*n*.

(3) *DD* = Distances - *N*/*n*

Finally, put the sum of the absolute *DD*s in relation to the sum of deviations that would be obtained from the worst possible distribution of the word in the corpus (DD_{worst}). The worst possible distribution arises if all word occurrences were placed immediately adjacent to each other.[1]

(4) D_{WG} = sum(*DD*) / sum(DD_{worst})

[1] The actual calculation of DWG involves several additional steps and details not specified here.

If all word occurrences follow a perfectly even distribution, the sum of their distance deviations will be 0, and 0 divided by the worst distribution is 0. If a word distribution actually follows the worst distribution, then the worst distribution divided by the worst distribution will be 1. Hence, D_{WG} ranges between 0 and 1, with 0 indicating perfectly even dispersion and 1 maximally uneven distribution.

There is good reason to believe that D_{WG} does in fact map onto the concept of lexical dispersion. Applications of D_{WG} to natural language corpora return results that replicate findings from other dispersion measures. High frequency function words are well dispersed (*the, of*). High frequency items that are relatively badly dispersed are pronouns (*she, I*). Badly dispersed words are technical terms or names (*thyroid, Win*ston). Well dispersed words are those that straddle functional and lexical items (probably, make). However, the vast majority of words fall in between the latter two classes and show unpredictable dispersion behaviour (water, white, know). It is predominantly for those words that a dispersion measure is informative and relevant. Figure 1 illustrates the relation between DWG and frequency in the Brown corpus (Kučera & Francis, 1964).

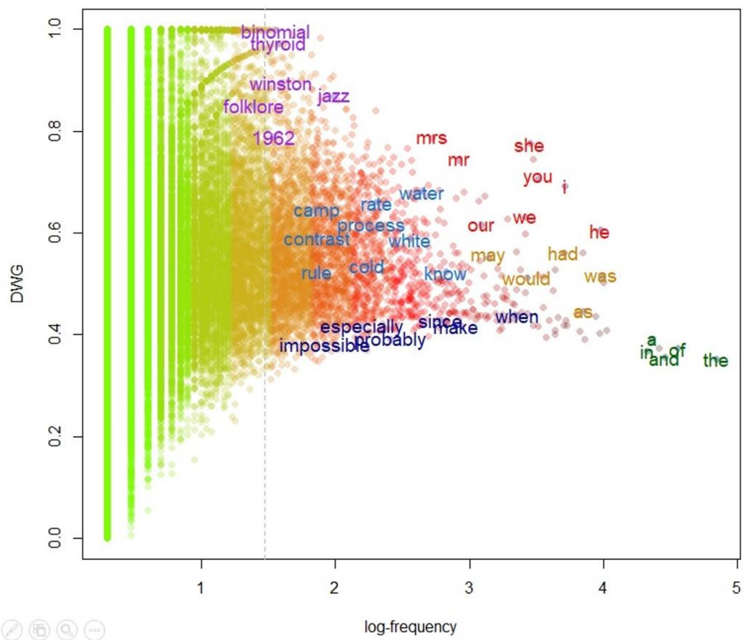

Figure 1: The distribution of DWG in Brown by word frequency (log scale).

Furthermore, D_{WG} is highly correlated with other dispersion measures. For example, it shows a high correlation of *r*=0.78 with *DP* (Griess, 2008) for the Brown corpus.

Based on investigations of several corpora, a five-class scale of D_{WG} seems reasonable. Values close to 0 indicating near-identical distances between occurrences do not normally exist in natural language. Experiments with artificial corpora show that randomly distributed words have a D_{WG} of c. 0.37, which constitutes a benchmark for highly uniform dispersion. Extremely skewed word growth distributions, however, do exist in natural language. These are cases in which a word's occurrences are placed in close proximity to each other so that its distribution actually approaches its worst possible distribution. For example, the word *tent* in Brown shows extremely small distances between its occurrences, as shown in (5), which results in a large D_{WG} value.

(5) There were umbrella tents, wall tents, cottage tents, station wagon tents, pup tents, Pop tents, Baker tents, tents with exterior frames (Brown, E-SkillsTradeHobbies, Sample E31 from Sports Age, 24:9 (1961))

The proposal is as follows: A D_{WG} value of less than 0.45 may indicate very even dispersion, 0.45 to 0.525 even dispersion, 0.525-0.6 average dispersion, 0.6 to 0.7 uneven dispersion and a value greater than 0.7 very uneven dispersion.

Following the literature on comparisons between dispersion measures (e.g., Burch et al., 2017), the performance of D_{WG} should be assessed for different criteria. Perhaps most importantly, the measure will be sensitive to the arrangement of the individual texts that make up a corpus. However, preliminary experiments on assessing the importance of this effect (with 15 corpus parts) suggest that variability caused by different corpus orders may be negligible, at least for large *n*. Next, dispersion measures of a word vary across different corpora. A preliminary experiment (with 10 natural 1m word corpora) shows that consistency is comparable for *DP* and D_{WG}. Finally, dispersion should be applicable for the correction of word frequencies, for instance, for the creation of learner vocabulary lists. Subtracting the D_{WG} percentage from a word type's frequency results in a high correlation with adjusted frequency.

D_{WG} has several additional advantages. The most important ones are the fact that it can be calculate easily from a single corpus part, relatively straight-forward interpretability, scale invariance with respect to *n* and *N* for randomly distributed words, and a low correlation with frequency so that dispersion and frequency are separated into two distinct dimensions.

References

Burch, B., Egbert, J., & Biber, D. (2017). Measuring and interpreting lexical dispersion in corpus linguistics. *Journal of Research Design and Statistics in Linguistics and Communication Science, 3*(2), 189-216.

Gries S. Th. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics, 13*(4), 403-437.

Kučera, H., & Francis, N. W. (1964). *A standard corpus of present-day edited American English, for use with digital computers (Brown)*. Providence, RI: Brown University.