

# Towards a More Useful Measure of Dispersion in Corpus Data

---

Stefan Th. Gries<sup>1</sup>

## Abstract

The most frequently used statistic in corpus linguistics is the frequency of occurrence of some linguistic variable or the frequency of co-occurrence of two or more linguistic variables. However, it is widely known that frequencies of (co-)occurrence may sometimes be quite misleading. For example, Leech, Rayson, and Wilson (2001) show that while the words *HIV*, *keeper*, and *lively* are about equally frequent in the British National Corpus (16 p.m.), *HIV* is much less specialized such that it occurs in a much smaller number of files than *keeper* and *lively*. Similarly, Gries (2006) shows how investigating the association of verbs to the imperative in the ICE-GB can be severely distorted by words that are highly frequent in just one out of 500 files.

In order to handle such problems, several scholars suggested a variety of dispersion measures; these include the range, the standard deviation or the variation coefficient, Juilland's *D*, Carroll's *D2*, Rosengren's *S*, the usage coefficient, and inverse document frequency (cf. Oakes 1998 or Rayson 2003 for overviews). However, these coefficients suffer from the problem that they require the corpus parts for which a dispersion measure is computed to be identically large (Oakes 1998: 191), which is usually not true. Likewise, chi-square does not rely on this assumption, but can take on widely varying values depending on whether expected frequencies become very small.

In this paper, I will propose for discussion a very simple alternative measure, which (i) allows to quantify dispersion just like the above, (ii) does not rely on the unwarranted assumption of equally-sized corpus parts, and (iii) is not affected by small expected frequencies. I will exemplify this measure using both words and word-construction pairings from different frequency bands and with different degrees of dispersion in the British National Corpus Sampler.

## References

- Gries, Stefan Th. 2006. Some proposals towards more rigorous corpus linguistics. *Zeitschrift für Anglistik und Amerikanistik* 54.2:191–202.
- Leech, Geoffrey N., Rayson, Paul, and Andrew Wilson. 2001. *Word frequencies in written and spoken English: based on the British National Corpus*. Longman: London.
- Oakes, Michael. 1998. *Statistics for corpus linguistics*. Edinburgh: Edinburgh University Press.
- Rayson, Paul. 2003. *Matrix: a statistical method and software tool for linguistic analysis through corpus comparison*. Unpublished Ph.D. dissertation, Lancaster University.

---

<sup>1</sup> e-mail: stgries@linguistics.ucsb.edu