Correction to Stefan Th. Gries' "Dispersions and adjusted frequencies in corpora" *International Journal of Corpus Linguistics* 13:4 (2008), 403-437

Jefrey Lijffijt and Stefan Th. Gries

Aalto University School of Science / University of California, Santa Barbara

Gries (2008) in this journal reviewed a variety of dispersion measures as well as adjusted frequencies and also proposed a measure for dispersion of elements in a corpus: *DP* (for *deviation of proportions*). This measure is computed as described in (i) to (iii) for an element *a* in *n* corpus parts.

(i) Determine the sizes s_{1-n} of each of the *n* corpus parts, which are normalized against the overall corpus size and correspond to expected percentages which take differently-sized corpus parts into consideration.

(ii) Determine the frequencies v_{1-n} with which *a* occurs in the *n* corpus parts, which are normalized against the overall number of occurrences of *a* and correspond to observed percentages.

(iii) Compute all *n* pairwise absolute differences of observed and expected percentages, sum them up, and divide the result by two.

The result is *DP*, which can theoretically range from approximately 0 to 1, where values close to 0 indicate that *a* is distributed across the *n* corpus parts as one would expect given the sizes of the *n* corpus parts. By contrast, values close to 1 indicate that *a* is distributed across the *n* corpus parts exactly the opposite way one would expect given the sizes of the *n* corpus parts. Table 1 is an example of how to compute *DP* if there are three equally large corpus parts, and one of these corpus parts contains $^2/_3$ of all occurrences of *a*, and another part contains the remaining $^1/_3$.

Table 1. Computation of *DP*, example 1

Step 1	Step 2	Step 3		
Expected %	Observed %	Abs. differences	Sum of abs. diff.	Divide by 2
0.333	0.667	0.334		
0.333	0.333	0	0.667	0.334
0.333	0	0.333		

While Gries' (2008) main point was to argue in favour of *DP*, some reviewers were concerned with the range of values that *DP* typically takes on. To address these concerns, Gries (2008) therefore also proposed a normalization: DP_{norm}, computed as shown in (1):

(1) $DP_{norm} = \frac{DP}{1-1/n}$

This erratum is concerned with this normalization of the measure. DP_{norm}, as defined in (1) only normalizes as intended when the n corpus parts over which a's dispersion is computed are equally large (just like several other dispersion measures previously discussed). In cases where the corpus parts are not equally large, DP_{norm} as defined in (1) can fall outside of the range between (and including) 0 and 1.

However, the proposed normalization can be easily fixed: instead of arriving at DP_{norm} by dividing DP by $1-1/n$, one should divide by the maximal possible value DP can become given the corpus in question. Specifically, DP is maximal when all occurrences of the word of interest are in the smallest part of the corpus. Hence the normalization factor should be $1-min_i(s_i)$, that is, one minus the size of the smallest corpus part. Thus we should redefine as in (2):

(2) $DP_{norm} = \frac{DP}{1-\min(s)}$

For the data in Table 1, this would mean that DP_{norm} is computed as in (3):

(3) $DP_{norm} = \frac{DP}{1-\min(s)} \cong \frac{0.334}{1-0.333} \cong 0.5$

Consider Table 2 and (4) for a case where the corpus sizes are not equal:

Table 2. Computation of *DP*, example 2

Step 1	Step 2	Step 3		
Expected %	Observed %	Abs. differences	Sum of abs. diff.	Divide by 2
0.4	0.6	0.2		
0.3	0.2	0.1	0.4	0.2
0.3	0.2	0.1		

(4) $DP_{norm} = \frac{DP}{1-\min(s)} \cong \frac{0.2}{1-0.3} \cong 0.2857$

For most of the standard corpora for which dispersion measures have been provided, the results change very little, given that these corpora involve many corpus parts (i.e., large n's) and, connected to that, very comparable corpus sizes (i.e., very similar s's) – usually, the results differ by less than $1/100$. However, we felt the correction was necessary and the dispersion data published on the companion website to Gries (2008, 2010) have been corrected in the meantime.

References

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

Gries, St. Th. 2010. "Dispersions and adjusted frequencies in corpora: Further explorations". In St. Th. Gries, S. Wulff & M. Davies (Eds.), *Corpus Linguistic Applications: Current Studies, New Directions*. Amsterdam: Rodopi, 197–212.

Authors' addresses