

Dispersions and adjusted frequencies in corpora: further explorations

Stefan Th. Gries

University of California, Santa Barbara

Abstract

In order to adjust observed frequencies of occurrence, previous studies have suggested a variety of measures of dispersion and adjusted frequencies. In a previous study, I reviewed many of these measures and suggested an alternative measure, *DP* (for deviation of proportions), which I argued to be conceptually simpler and more versatile than many competing measures. However, in spite of the relevance of dispersion for virtually all corpus-linguistic work, it is still a very much under-researched topic: to the best of my knowledge, there is not a single study investigating how different measures compare to each other when applied to large datasets, nor is there any work that attempts to determine how different measures match up with the kind of psycholinguistic data that dispersions and adjusted frequencies are supposed to represent. This article takes exploratory steps in both of these directions.

1. Introduction

Whether one likes it or not, corpus linguistics is all about distributional data and virtually every corpus-based paper reports how often some linguistic phenomenon occurred or how often it co-occurred with some other linguistic phenomenon or extralinguistic variable. Such frequency data are used for several different purposes: sometimes, they are just used descriptively, but outside of particular traditional schools of corpus linguistics, they are also often used to support particular points or applications in the domains of applied and theoretical linguistics as well as a tool for psycholinguists and psychologists. For example, in some theoretical approaches such as cognitive linguistics or usage-based grammar, frequency data are now regularly used in the domains of first and second/foreign language acquisition, the study of language and culture, grammaticalization, phonological reduction, morphological processing, syntactic alternations etc.

Interestingly enough, many of these approaches assume a connection between observed frequencies in a corpus and some mental correlate: in first language acquisition, input frequency is one of the most important determinants of word and construction learning; in cognitive-linguistic approaches frequency of encounter is one of the central determinants of degree of mental entrenchment / familiarity; for example, observed frequencies (or their logs) are good proxies towards the familiarity of words – cf. Howes and Solomon (1951) for recognition times, Oldfield and Wingfield (1965) as well as Forster and Chambers (1973) for

naming times, and Ellis (2002a, b) as well as Jurafsky (2003) and Gilquin and Gries (2009) for overviews. Thus, in probabilistic models of language production and comprehension, mental entrenchment in turn is correlated with accessibility (such that, for example, high frequencies of exposure make nodes more available for activation).

In spite of this central role of frequency in linguistics in general and in psycholinguistics in particular, it has become clear that frequencies of occurrence are not a perfect predictor of aspects of processing. This is for different kinds of reasons. First, because of the complexity of all aspects entering into processing effort: no one would deny that processing of words and concepts is determined by many more though highly intercorrelated aspects such as salience of words and concepts, recency of occurrence, concreteness/manipulability, to name but a few. Thus, any kind of frequency effect will be ridden with noise and, hence, necessarily indirect. Second, because of the fact that frequency of occurrence, however straightforward to define, does not enter into a straightforward one-to-one relationship with aspects of processing because any particular frequency of occurrence can arise from very different distributional patterns: a word *w* may occur 18-20 times in each of ten very different registers, or it may occur 190 times in only one of the 10 registers. While these two results look the same in a frequency list of the complete corpus of 10 registers, it is obvious that these results would not *be* the same: they would not be the same for the corpus linguist who may be interested in register-dependent vocabulary differences, and they would not be the same for the psycholinguist or language acquisition researcher, who knows that learning process in general exhibit a distributed learning or spacing effect:

given a certain number of exposures to a stimulus, or a certain amount of training, learning is always better when exposures or training trials are distributed over several sessions than when they are massed into one session. This finding is extremely robust in many domains of human cognition. (Ambridge et al. 2006: 175)

Surprisingly, there is not much corpus-linguistic work that deals with or let alone incorporates this potential bias, which in corpus linguistics is referred to as *dispersion*. I know of only a few studies that attempt to address this problem by developing measures of dispersion (i.e., measures that quantify the homogeneity of the distribution of a word in a corpus) or adjusted frequencies (i.e., frequencies that penalize words that are attested only in a small part of a corpus), and there is also only another handful of studies which actually use these measures or study them in more detail. In Gries (2008), I discuss all the measures proposed so far and illustrate that using frequencies alone runs the risk of yielding incorrect results. More specifically, I

- exemplified how frequencies of (co-)occurrence can be quite misleading;

- argued that measures of dispersion as well as adjusted frequencies may be needed to study and characterize our data more accurately;
- suggested a new and intuitively simple dispersion measure called *DP* to address several of the shortcomings that existing measures exhibit; and
- provided some resources for researchers: two small computer programs to compute dispersion measures and adjusted frequencies as well as dispersion measures and adjusted frequencies for thousands of words from four different corpora.

(Cf. that study for definitions of, and references on, all the measures discussed here.) However, it is quite obvious that a variety of issues in this area remains to be explored in more detail, especially given that dispersion characteristics can influence any given corpus-based statistic.

First, we need much more information about the properties of the measures. Lyne's (1985) groundbreaking work is a laudable start: using scatterplots to compare a few dispersion measures, he was the first to try to come to grips with the various ways in which measures differ. However, his study was restricted to the few measures available at that time, and today's computational possibilities allow for much larger and/or more detailed investigations of the measures he used and later ones. For example, little is known about

- how the results of different dispersion measures or adjusted frequencies compare to each other (beyond Lyne's above study). This is problematic since there are now different kinds of measures.
 - First, parts-based measures, which take into consideration how often an element in question is attested in parts of the corpus, but disregard the order of corpus parts as well as where in these parts element in question occurs.
 - Second, distance-based measures, which take into consideration the distances between successive occurrences of the element in question in a corpus (and hence their order), but not its frequencies in different parts of a corpus.
 - Finally, hybrids, which take into consideration both the number of occurrences of the element in question in each corpus part and, within each part, distances between successive occurrences.
- to what degree, if any, these measures come in quantitatively definable, meaningful groups;
- which kinds of distributions (of authentic data) yield what kinds of results.

Such issues are relevant for, for example, being able to better compare dispersion measures from different studies, to be able to choose the best measure, or at least choose measures that are better suited to particular tasks at hand. In Section 2 of this paper, I will therefore compare the behaviour of all published

measures of dispersions and all adjusted frequencies I have come across on the basis of the 17,481 most frequent types (10,294,637 tokens) in the spoken component of the British National Corpus (XML version).

Second, we need to be more serious about validating our dispersion measures and adjusted frequencies. (Strictly speaking, this also applies to measures of co-occurrence strengths, but this is beyond the scope of the present paper.) Devising statistics that are theoretically motivated and make intuitive sense when applied to corpus data is a useful step, but also only the *first* step.

For example, given what we have seen above regarding the correlation of observed frequencies (or their logs) and the familiarity of words, psycholinguists or psychologists often use frequency information of words from corpora or databases to create experimental stimuli with an eye to controlling for frequency or familiarity. However, if dispersion plays the role some corpus linguists have argued, then controlling for frequency alone may turn out to be insufficient unless dispersion is considered at the same time. For corpus linguists, that means, though, that our measures must be validated against corpus-external evidence because, strictly speaking, as long as we corpus linguists do not show that our dispersions and adjusted frequencies correspond to something outside of our corpora, we have failed to provide the most elementary aspect of a new measure, its validation.¹

How could we come up with such evidence? First, we can perform experiments ourselves. For example, one could run experiments (i) on the fictitious corpus distributions discussed in Lyne (1985) and Gries (2008) to determine whether the measures are able to distinguish them or not and (ii) to determine which measures' results from large balanced corpora are most compatible with subjects' intuitions regarding the words' overall centrality in a language. Since dispersion and adjusted frequencies are used as proxies to familiarity, one could also check whether ways of presenting children with nonce words that differ in terms of the dispersion patterns of the word in question lead to different degrees of learning success (cf. studies on distributed learning such as Ambridge et al. 2006). Thankfully, the number of experimental validations of corpus-based studies is steadily increasing, and the field of dispersion should be no exception to this general trend.

Second, we can reanalyze published psycholinguistic data. In Section 3 below, I will correlate dispersion measures and adjusted frequencies with the response time latencies of Spieler and Balota (1997) as well as Balota and Spieler (1998), as well as lexical decision task data from Baayen (2008).

2. Dispersions and adjusted frequencies: intercorrelations

To explore how dispersion measures and adjusted frequencies are intercorrelated with each other, I used data from the spoken component of the British National Corpus World Edition (XML). More specifically, I wrote an R script that

- loaded each corpus file of the BNC World Edition (XML) that contained spoken data, converted it to lower case, and retained only the lines that contained sentence numbers (regular expression: "<code>$\langle s \cdot n \rangle$");
- deleted all sequences of non-word tags and the material they refer to (regular expression: "<code>$\langle !w \cdot *? \rangle \cdot *? \rangle [^\langle * \rangle]$");
- split up the remaining data at sequences of optional spaces and word tags (regular expression: "<code>$\langle w \cdot *? \rangle$");
- printed the resulting word list into an output file such that the file contained all the word tokens from the corpus in the order of the files followed by one tab stop followed by the name of the file in which the word occurred.²

Then, for all word types that occurred ten or more times in the corpus, I used another R script to compute all 29 dispersion measures and all adjusted frequencies discussed in Gries (2008); the corpus parts I assumed were the individual files, which were processed in alphabetical order of their filenames. As a result, I obtained a table with these dispersion measures and adjusted frequencies in the columns for the 17,481 word forms in the rows.^{3,4}

Intercorrelations between these measures were explored using hierarchical agglomerative cluster analyses and, for additional graphical exploration, principal component analysis. Hierarchical agglomerative cluster analysis is a statistical tool well-suited to the task at hand. Such cluster analyses try to find structures in the data by successively amalgamating individual measures into larger groups such that the within-groups similarities are as large as possible compared to the between-groups similarities. While such a clustering approach is bottom-up and data-driven, the researcher has to make at least two important decisions. First, one has to decide on a measure of pairwise similarity on the basis of which the different elements to be compared – the dispersion measures and adjusted frequencies – are compared to each other. Second, one has to decide on an amalgamation rule, an algorithm that determines how groups of elements are merged. As to the former, I use a fairly standard measure, namely $1-r$, where r is the Pearson product-moment correlation between the vectors of any two measures that are being compared.⁵ As to the latter, I use Ward's method because it has been shown to be a reliable measure, good at identifying small clusters, and because of its affinity to the logic underlying ANOVAs. To avoid distortions by the different scales of the measures, the values were z -standardized by column,⁶ and I did separate analyses for the dispersion measures and the adjusted frequencies.

2.1 Dispersion measures

The result of the cluster analysis on dispersion measures is shown in Figure 1 (the abbreviations of the measures are listed in appendix 1).

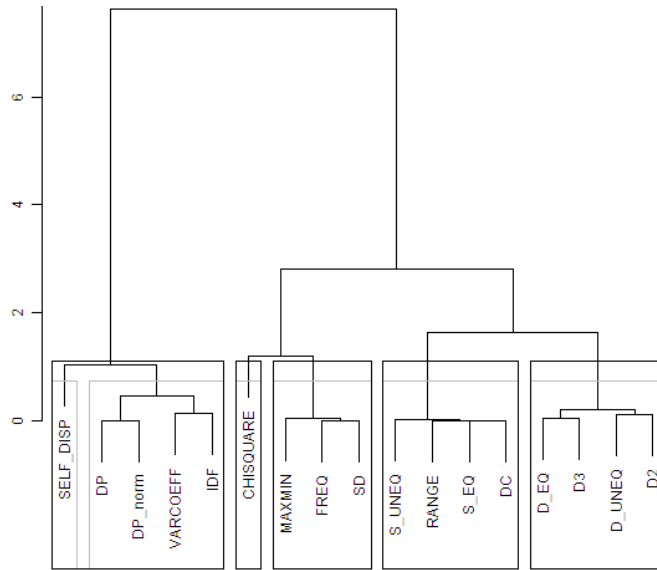


Figure 1: Dendrogram of dispersion measures⁷

The results suggest five different clusters:

- the maximal average silhouette width for a cluster solution (0.73) was obtained for six clusters (cf. the grey boxes), but this comes at the cost of assuming two clusters that consist of only one measure (assuming one-measure clusters is undesirable because such clusters mean that the one measure is in fact unique and cannot be merged with another one, which is after all the whole point of clustering);
- the second highest average silhouette width is practically the same (0.729), but has only one single-element cluster and a much higher average silhouette width than the next solution with fewer clusters (4 clusters: 0.675);
- the principal component analysis returned only four principal components with eigenvalues larger than 1; the loadings of the first two principal components (the first on the x -axis, the second on the y -axis), which together account for 77.2% of the variance in the data, are plotted in Figure 2 (the polygons represent the clusters from the HCA).

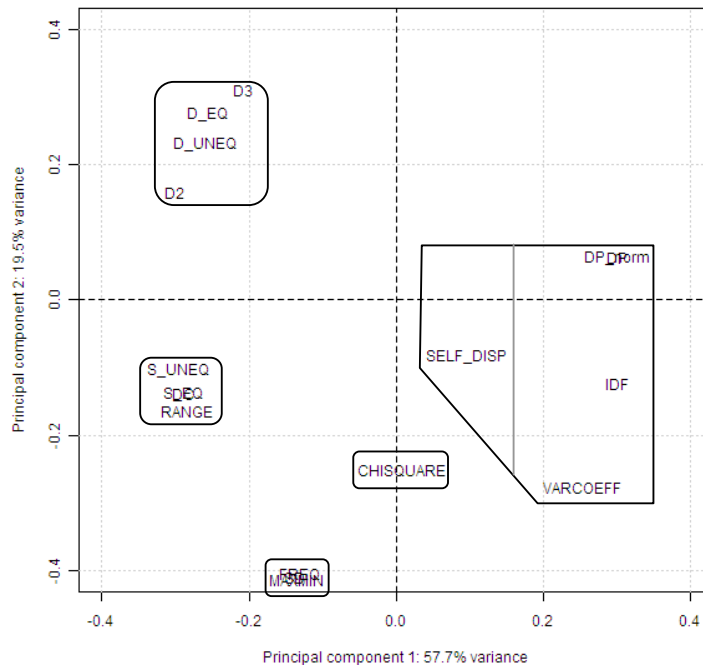


Figure 2: Loadings of the first two principal components

For reasons of space, I cannot discuss the results here in great detail. It is obvious, however, that the proposed 16 dispersion measures constitute five or six different kinds of measures that differ mainly along two dimensions and that exhibit varying degrees of homogeneity: both versions of Rosengren's *S*, *DC*, and the *range* are very similar to each other whereas the cluster containing *idf* and the variation coefficient is rather heterogeneous instead. Interestingly, a measure such as Carroll's D_2 , whose creator harshly attacked Juilland et al.'s *D* for a variety of perceived shortcomings, is actually very similar to it in terms of its overall behavior – in fact, much more similar than to most other measures. In addition, I checked how each cluster relates to raw frequency by inspecting one central member of each cluster. In order to force all values into a comparable range and given them the same orientation (high values indicate high clumpyness and low values indicate more even distributions), I *z*-standardized

- the vectors of $1-DC$ value and $1-D_{equal}$;
- the vectors of *idf*, chi-square and *DP* values.

These values were then plotted against log frequency and summarized with smoothers.⁸ The result shows that the different groups of values behave very differently with respect to frequency; cf. link 2 in Appendix 2. *DC*, D_{equal} , *idf* and

DP become smaller with larger observed word frequencies, but, on the whole, chi-square does not! In addition, the measures exhibit quite different ranges: while DC and DP are fairly similar, but the D_{equal} as more larger values than idf , and chi-square has extremely many large values. Finally, the curvature of the smoothers sometimes differs considerably: DC and DP again behave similarly, but different from both D_{equal} and idf .

What does all this show? In the present form, the data do not show much in terms of specific content. What they do show, however, is that different measures of dispersion will yield *very* different (ranges of) values when applied to actual data. Researchers must exercise caution in their choice of a measure of dispersion for their data: not only should they make sure that they choose a measure that exhibits all of the theoretically desirable characteristics,⁹ but they might also want to consider reporting, or basing their subsequent analysis on, the results of more than one measure, ideally from measures from the different clusters represented in Figure 1 and Figure 2. Interestingly, the one dispersion measure that is conceptually very different from all others does not exhibit a particularly special status in the evaluation: Washtell's self-dispersion is the only measure that does not only take into consideration the number of times an element is observed in a corpus part, but also the distances between the occurrences. On the one hand, this may seem like a theoretically very attractive feature, but it can also only be applied when a word occurs more than once in a corpus part. However, while this measure is different enough to constitute a cluster on its own when the less parsimonious 6-cluster solution is adopted, the principal component analysis shows that it is located in a relatively populated area of principal component space. More and maybe more diverse data are required to shed light on whether or not the additional computational effort of this theoretically attractive feature of self-dispersion is justified.

2.2 Adjusted frequencies

Interestingly, the result of the cluster analysis on adjusted frequencies does not merit a figure. Apart from Kromer's U_R , all other measures are grouped together; the only one that may be a little bit different from the rest is f_{AWT} : the average silhouette width for the two-cluster and three-cluster solutions are 0.89 and 0.65 respectively. What is more interesting to note is that the two different kinds of adjusted frequencies are not distinguished very much: the distance-based measures proposed by Savický and Hlaváčová are grouped together with several parts-based measures, which disregard distance information. Also, when one looks at how much in percent each adjusted frequency reduces the actually observed frequency, then the three measures which are furthest away from each other in the dendrogram behave *completely* differently; cf. Figure 3, which shows the non-parametric smoothers for Rosengren's AF (for equal corpus parts), Kromer's U_R , and Savický and Hlaváčová's f_{AWT} , and link 3 for the complete plot.

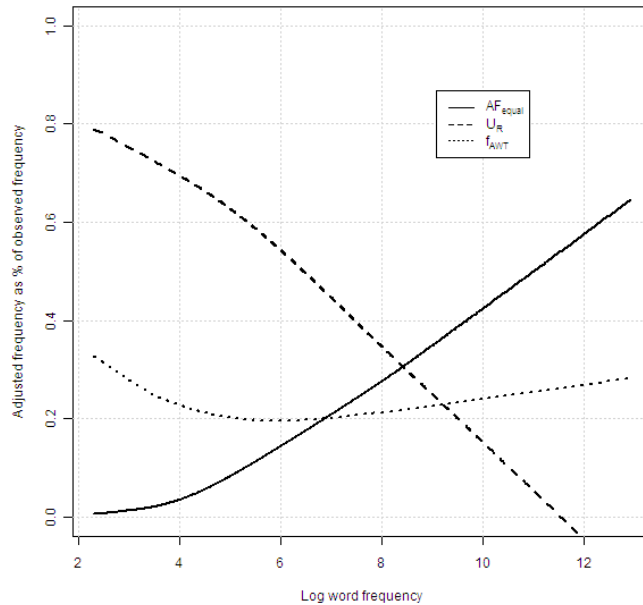


Figure 3: Non-parametric smoothers summarizing three adjusted frequencies

It is hard to imagine a more diverse result. The more frequent words are, the less Kromer's U_R reduces their frequency, but at the same time the more Rosengren's AF does, and f_{AWT} is different from both. I have little to say about this particular result other than that it clearly emphasizes that we know next to nothing about how different adjusted frequencies behave and what they actually mean or do. More exploration is necessary but even more important is that we begin to validate the two dozen or so dispersion measures and adjusted frequencies we have at our disposal. A first step in this direction will be taken in the next section.

3. Dispersions and adjusted frequencies: validation against psycholinguistic data

While dispersion measures and adjusted frequencies were developed with rather practical motivations in mind (e.g., to provide lexicographers with more reliable statistics than raw frequencies), it is probably fair to say that our knowledge of dispersion measures and adjusted frequencies is approximately inversely proportional to what we know about their accuracy, reliability, and predictive power. In this section, I want to briefly explore how the measures we have relate to the kind of psycholinguistic data they are presumably supposed to relate to. If dispersion measures are really better indicators of, for example, the familiarity of words (and, hence, somewhat indirectly to maybe even to the concepts these

words evoke), if adjusted frequencies are truly more appropriate indicators of cognitive entrenchment, then we should find robust correlations between our measures and psycholinguistic results such as response time latencies. Unfortunately, it will become obvious that the data raise more questions than they answer.

As a first example, I correlated the response time latencies of young and old native speakers of English to monosyllabic words from Spieler and Balota (1997) and Balota and Spieler (1998). All dispersion measures and adjusted frequencies were centered and the correlation coefficient used was Kendall's τ . Given that not all dispersion measures have the same orientation (cf. above Section 2.1), the correlations between the measures and the response time latencies can be both positive and negative: the larger an effect, the more Kendall's τ will deviate from zero; cf. Figure 4 for the results for young speakers and Figure 5 for the results for old speakers; the x -axis labels ("d" and "f" indicate whether the plotted measure is a measure of dispersion or an (adjusted) frequency and given the large n , all correlations are highly significant.

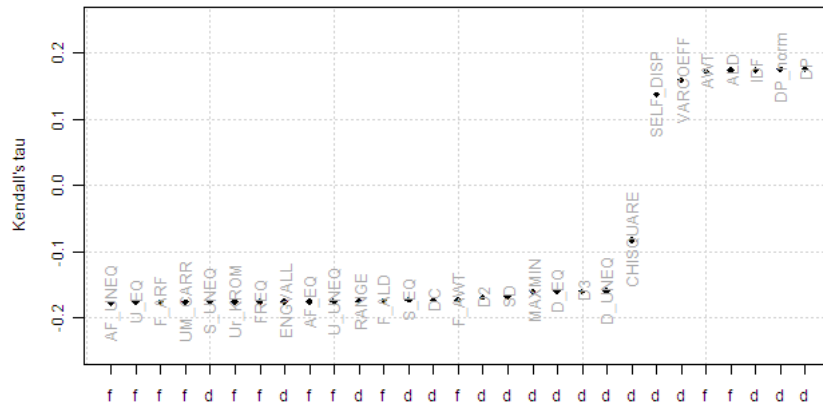


Figure 4: Correlations between Balota and Spieler's (1998) response time latencies (young speakers) and the dispersion measures and adjusted frequencies surveyed in Gries (2008)

In some sense, the results are striking. On the one hand, both panels show the same measures as resulting in the strongest correlations: *ALD*, *DP/DPnorm*, and *idf* (as measures with positive correlation coefficients) and *AF_{uneq}* and *U_{uneq}* (as measures with negative correlation coefficients). On the other hand, it is equally obvious that with very few exceptions, it doesn't seem to matter which measure is chosen since most of the correlations are of the same strength (which also means that Kromer's U_R – in spite of the claim of it being more psycholinguistically grounded – does not result in a stronger correlation with the psycholinguistic data).

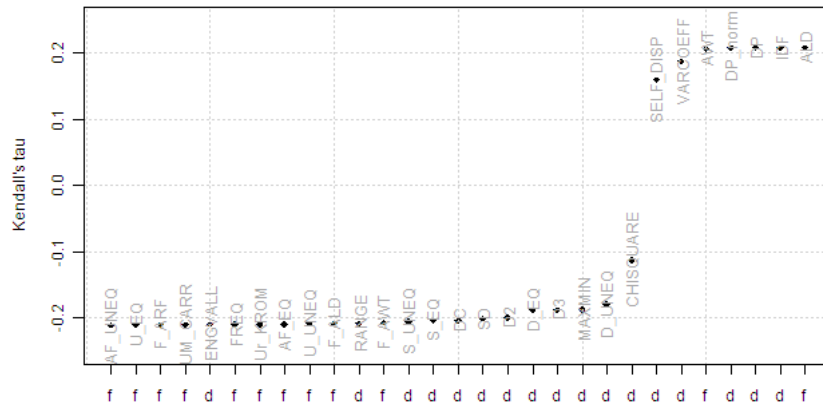


Figure 5: Correlations between Balota and Spieler's (1998) response time latencies (old speakers) and the dispersion measures and adjusted frequencies surveyed in Gries (2008)

Even this interim conclusion, however, is undermined once we do the same kind of computation for the lexical decision task data of Baayen (2008), which are represented in Figure 6.

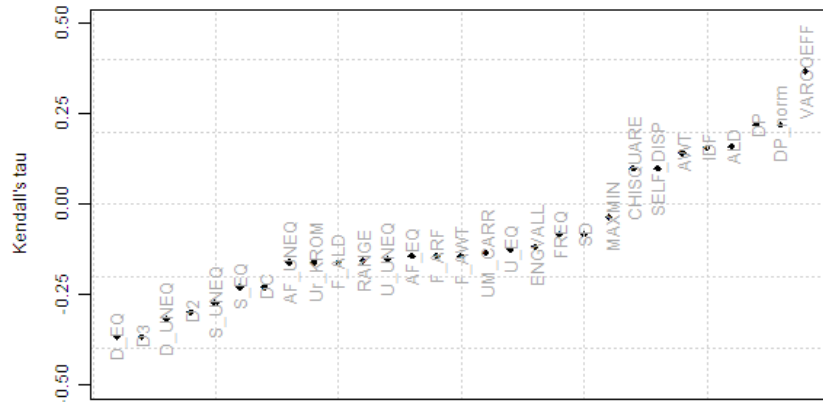


Figure 6: Correlations between Baayen's (2008) lexical decision task times (for native speakers) and the dispersion measures and adjusted frequencies surveyed in Gries (2008)

Again, *ALD* and DP/DP_{norm} are among the strongest correlations, only surpassed, perhaps surprisingly, by the variation coefficient, but D_{equal} and D_3 also exhibit strong correlations although their correlations with Balota and Spieler's data were only somewhat moderate. On the more positive side, compared to Figures 4 and 5, this time there is a clear cline with some measures clearly very

close to a null correlation, and as a matter of fact, only D_{equal} , D_3 , $D_{unequal}$, and the variation coefficient correlate significantly with the psycholinguistic measure.

4. Summary and some (preliminary) conclusions

While the results of Section 2 provide at least some clue(s) for future studies, the results of Section 3 do not yet inspire a lot of hope. Section 2 showed that when the proposed dispersion measures are applied to most of the words in the spoken component of the BNC, they fall into approximately five groups along two dimensions and take on a bewildering range of values. It is probably safe to say that chi-square is not a particularly useful measure since across the full range of observed frequencies, it exhibits an extremely high range of values, so chi-square does not appear to be particularly discriminatory. However, apart from that, the dispersion measures mainly differ in the degree to which they reach higher values with increasing frequency and none of them reaches really high levels of predictive power (which was to be expected, recall Section 1).

For the adjusted frequencies, the picture is more diverse: the measures do not fall into nicely distinct groups other than U_R vs. the rest, but if three core measures are explored, the ways in which an adjusted frequency reduces the observed frequency exhibit all possible directions of correlation with the actually observed frequency.

In spite of the diversity of these results, recommendations for future work are clear: avoid chi-square, use several different measures of dispersion from the identified groups, bear in mind the potentially confounding factor of corpus part sizes, and explore self-dispersion as well as the distance-based measures to determine whether or not they ultimately yield more revealing results.

Section 3 brings good news and bad news. The good news is that for all three psycholinguistic measures, there are significant correlations between at least some dispersion measures and adjusted frequencies, the highest absolute correlations are provided by a small set of measures (ALD , DP , and the variation coefficient are among them), and, crucially, these measures are more highly correlated with the psycholinguistic results than raw frequencies of occurrence. It is particularly interesting that a general measure of dispersion such as the variation coefficient, which has not been designed specifically to handle corpus data, scores so well. The bad news, however, is that the data are as yet too small and too heterogeneous to allow making more meaningful recommendations than that, (i) focusing on these three measures probably increases the likelihood of good results and (ii) we need to know more.

On a methodological level, it also emerges that even though Lyne's earlier work on comparing different measures of dispersion has been a major milestone, it is now time to include more measures and adopt a multivariate perspective. Lyne used a plot-based exploration on selected (fictitious) distributions, but the present approach shows that (i) looking at more than 17,000 word types and (ii) using more sophisticated methods – robust smoothing approaches, hierarchical cluster analysis and principal components analysis – have more to offer than was

available at the time of Lyne's work. While this paper could only take a small step towards answering all the questions that arise from the literature, I hope I could provide some initial and interesting results and some incentive to explore these issues further. After all, what are dispersions and adjusted frequencies good for when we don't know what they do and what exactly they measure ...

Notes

* I thank three anonymous reviewers for their comments and suggestions. The usual disclaimers apply.

¹ One laudable exception is the recent work by Ellis and colleagues, who show that range has significant predictive power above and beyond raw frequency of occurrence, and it is this kind of evidence we must provide in order to show our efforts are more than devising clever equations.

² All retrieval operations, statistics, and graphs were computed with R 2.8.0 (cf. R Development Core Team 2008).

³ Scripts to compute dispersion measures and adjusted frequencies as well as dispersion measures and adjusted frequencies for words from four different corpora are available from my website; cf. link 1 in Appendix 2.

⁴ Since it is as yet an unresolved question exactly how dispersions and adjusted frequencies react to different numbers of corpus parts (esp. in combination with differently sized corpus parts), it needs to be mentioned how similar the corpus parts are to each other. In this case, the file sizes (in words) were all rather similar to each other: the relative entropy of the file sizes is 0.914 and thus relatively close to the theoretical maximum.

⁵ I used 1-r as a measure to be able to better compare the results of the cluster analysis with the of the principal components analysis. A cluster analysis based in Kendall's τ as a similarity measure yielded a virtually identical dendrogram, the sole difference being that the two clusters on the right of Figure 1 were more similar to each other.

⁶ To z-standardize a value x from a vector/range of values, you subtract the mean of all the values from x and divide the result by the standard deviation of all the

values: $z = \frac{x - \mu}{\sigma}$.

⁷ Cf. the appendix for the meanings of the abbreviations.

⁸ The smoother I used are locally-weighted polynomial regressions, i.e. regression lines that try to summarize the cloud of points in a scatterplot without the restriction of typical linear regressions that the line must be straight; cf. `lowess` in R.

⁹ These "theoretically desirable characteristics" include the ability of dispersion measures (i) to handle differently-sizes corpus parts, (ii) to fall only into the range the dispersion measure is supposed to fall into, (iii) to exhaust that range (i.e., not cluster only in small range of the complete theoretical range), (iv) to not be overly

sensitive to the overall numbers of corpus parts, (v) to be sensitive enough, but not too sensitive, given extreme distributions and zero occurrences, and others; cf. Gries (2008: Section 2.4 and 5 for discussion and exemplification).

References

- Ambridge, B., A.L. Theakston, E.V.M. Lieven, and M. Tomasello. (2006), 'The distributed learning effect for children's acquisition of an abstract syntactic construction', *Cognitive Development*, 21: 174-93.
- Baayen, R.H. (2008), *Analyzing linguistic data: a practical introduction to statistics with R*. Cambridge: Cambridge University Press.
- Balota, D.A. and D.H. Spieler. (1998), 'The utility of item level analyses in model evaluation: a reply to Seidenberg and Plaut', *Psychological Science* 9.3:238-40.
- Ellis, N.C. (2002a), 'Frequency effects in language processing and acquisition', *Studies in Second Language Acquisition*, 24:143-88.
- Ellis, N.C. (2002b), 'Frequency effects in language processing and acquisition', *Studies in Second Language Acquisition*, 24: 297-339.
- Forster, K.I. & S.M. Chambers. (1973), 'Lexical access and naming time', *Journal of Verbal Learning and Verbal Behavior*, 12:627-35.
- Gilquin, G. & St.Th. Gries. (to appear) Corpora and experimental methods: a state-of-the-art review. *Corpus Linguistics and Linguistic Theory*, 5.
- Gries, St.Th. 2008. Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics*, 13: 403-37.
- Howes, D.H. & R.L. Solomon. (1951), 'Visual duration threshold as a function of word-probability', *Journal of Experimental Psychology*, 41: 401-10.
- Jurafsky, D. (2003), 'Probabilistic modeling in psycholinguistics', in R. Bod, J. Hay and S. Jannedy (eds.) *Probabilistic linguistics*. Cambridge, MA: The MIT Press, 39-96.
- Lyne, A.A. (1985), 'Dispersion', in *The vocabulary of French business correspondence*. Geneva, Paris: Slatkine-Champion, 101-24.
- Oldfield, R. and A. Wingfield. (1965), 'Response latencies in naming objects', *Quarterly Journal of Experimental Psychology*, A 17: 273-81.
- R Development Core Team. (2008), R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.
- Spieler D.H. and D.A. Balota. (1997), 'Bringing computational models of word naming down to the item level', *Psychological Science*, 8:411-6.

Appendix 1

Abbreviation	Measure
FREQ	observed frequency of word w
RANGE	number of parts with word w
MAXMIN	max. freq. of w /part - min. freq. of w /part
SD	standard deviation of frequencies
VARCOEFF	variation coefficient of frequencies
CHISQUARE	chi-square value of the frequency distribution
D_EQ	Juilland et al.'s D (assuming equal parts)
D_UNEQ	Juilland et al.'s D (not assuming equal parts)
D2	Carroll's D_2
S_EQ	Rosengren's S (assuming equal parts)
S_UNEQ	Rosengren's S (not assuming equal parts)
D3	Lyne's D_3
DC	Distributional Consistency
IDF	Inverse Document Frequency
ENGVALL	Engvall's measure
U_EQ	Juilland et al.'s usage coefficient U (assuming equal parts)
U_UNEQ	Juilland et al.'s usage coefficient U (not assuming equal parts)
UM_CARR	Carroll's U_m
AF_EQ	Rosengren's Adjusted Frequency AF (assuming equal parts)
AF_UNEQ	Rosengren's Adjusted Frequency AF (not assuming equal parts)
Ur_KROM	Kromer's U_R
F_ARF	Savický and Hlaváčová's f_{ARF}
AWT	Savický and Hlaváčová's AWT
F_AWT	Savický and Hlaváčová's f_{AWT}
ALD	Savický and Hlaváčová's ALD
F_ALD	Savický and Hlaváčová's f_{ALD}
SELF_DISP	Washtell's self-dispersion
DP	Gries's Deviation of Proportions
DP_norm	Gries's Deviation of Proportions (normalized)

Appendix 2

- link 1: <http://www.linguistics.ucsb.edu/faculty/stgries/research/dispersion/links.html>
- link 2: http://www.linguistics.ucsb.edu/faculty/stgries/research/dispersion/comparison_dispersion.png
- link 3: http://www.linguistics.ucsb.edu/faculty/stgries/research/dispersion/comparison_adjfreq.png