This is a contribution from JSLS 5:2
© 2022. John Benjamins Publishing Company

What do (most of) our dispersion measures measure (most)? Dispersion?

Stefan Th. Gries
University of California, Santa Barbara, USA | JLU Giessen

This paper discusses the degree to which most of the most widely-used measures of dispersion in corpus linguistics are not particularly valid in the sense of actually measuring dispersion rather than some amalgam of a lot of frequency and a little dispersion. The paper demonstrates these issues on the basis of data from a variety of corpora. I then outline how to design a dispersion measure that only measures dispersion and show that (i) it indeed measures information that is different from frequency in an intuitive way and (ii) has a higher degree of predictive power of lexical decision times from the MALD database than nearly all other measures in nearly all corpora tested.

Keywords: dispersion, frequency, association, range, Juilland's *D*, Gries's *DP*, generalized additive modeling

1. Introduction

In some way,[1] just about any statistic in corpus linguistics is ultimately based on frequency of occurrence and co-occurrence: We report frequencies of tokens and/or types per se, we use frequencies to compute dispersion measures (DMs), or we use co-occurrence frequencies to compute association measures (AMs). For each of these three dimensions of statistical information, theoretical, cognitive, and psycholinguistic research has discussed cognitive/psycholinguistic mechanisms underlying these dimensions. For instance,

[1] The beginning of this paper is very similar to that of the previous/sister publication (Gries 2022). This is so that, for readers who only read one of them, each of the two papers is self-sufficient and, for readers who read both, that they can skim Sections 1 and 2 of this one without missing anything.

https://doi.org/10.1075/jsls.21029.gri | Published online: 30 November 2021
Journal of Second Language Studies 5:2 (2022), pp. 171–205. ISSN 2542-3835 | E-ISSN 2542-3843

- **token frequency** has been related to matters of (cognitive) entrenchment (Schmid 2010) and/or baseline activation levels in psycholinguistic models of the mental lexicon (see discussion by Baayen et al. 2016);
- **dispersion** has been considered as a proxy towards the commonness of a word (I am using *commonness* here as a 'technical term' that, while usually operationalized using frequency, is not the same as frequency, see Savický & Hlaváčová 2002) and has also been related to recency (e.g., Gries 2019a);
- **association** has been related to contingency and associative learning in, say, the Competition Model or in Ellis's CREED model (e.g. Ellis 2007a, b; Fu & Li, 2019), but has also played an important role in second language studies or learner corpus studies as in explorations of collocational knowledge (see Ellis et al. 2008; Durrant & Schmitt 2009, Bestgen & Granger 2014, or Siyanova-Chanturia 2015 for examples).

When corpus linguists want to quantify, say, the dispersion of an element in a corpus or the association of two elements in a corpus, they have to choose what types of dispersion measure (DM) or AM to use simply because for both dispersion and association many different measures have been proposed (Gries, 2021). For dispersion, Gries (2008, 2010) reviewed and compared about a dozen or so measures, for association, Evert (2009) and Pecina (2009) alone reviewed more than in 80 measures, and for both domains new measures have been proposed since, which of course raises the issue of which measure(s) to choose.

One of the most central aspects that should feature in any researcher's decision for a measure is of course **validity**, which can be approached from two important yet complementary perspectives. The first perspective is concerned with the desideratum that a measure *m* should really measure what it is intended to measure; that means a DM should be designed in such a way that it measures dispersion and an AM should be designed in such a way that it measures association. There are probably few who would disagree with this seemingly trivial statement, but there is another, complementary aspect to it which is less often considered: The values of a measure should measure, or 'react to', what they are intended to measure or 'react to', but also not measure or 'react to' much else, so that we can take/interpret the computed values at face value (no pun intended).

The second perspective is concerned with the fact that the results of some such measure should ideally be correlated (well (enough)) with the kind of external evidence that the measure is supposed to measure. For example, if a DM is truly a measure of the degree to which, say, a word in common or widespread use in a language *and* if one independently assumes that the dispersion of a word is related to how quickly one can recognize it in, say, a lexical decision task, it follows that a good DM should also correlate with such external data (e.g. experimentally-obtained reaction times).

Interestingly enough, much work in corpus linguistics using DMs and AMs has not concerned itself enough with both of these two perspectives (maybe especially by neglecting the first – often for good reasons, see below – and I have done so myself too often), which can then also impact the second. Put differently and to say it out loud, if one's DM or AM is not *and* not only measuring what it is supposed to measure, then we are already beginning to fail the most basic test criterion, that of validity and then it's not a huge surprise that our measure might not correlate well with the kinds of external evidence we want to correlate it with or validate it against. As just mentioned, neglecting the first perspective – measuring what one wants to measure and nothing else – has often been done for probably just one single reason: simplicity and sortability along one dimension *d*: we all like to just click "Sort" and be done with it. If one computes DMs for how evenly words are distributed across a corpus and the DM conflates or, to put a more positive spin on it, 'conveniently integrates' information from various dimensions – hopefully with at least one of them being dispersion – then this might be (!) sufficient for a variety of lexicographic, applied, and maybe some descriptive purposes (and for many of those purposes the second perspective might not be relevant because, for instance, lexicographers don't need external psycholinguistic validation). In fact, sometimes the conflation of measures often returns 'intuitively satisfying' results precisely because the ranking one observes is actually not so much due to the dimension of information *d* one says one is using but (more) due to another dimension. This happens most often when, for example, frequency 'supports' the DM/AM *m* and, thus, makes *m* return results with a treacherously high(er) appeal. And that higher post hoc interpretive appeal has often made us ignore the fact that that appeal is not so much because *m* is so great and precise at capturing the dimension *d* we imply it captures (by its name) and the results are so great precisely because dimension *d* is exactly what matters, but because *m* actually reflects more than we say it does and it is actually everything that *m* uses above and beyond *d* that makes the results seem so great. More concretely, we might be calling something a *DM* and, correspondingly, interpret its results in terms of dispersion when, figuratively as well as statistically speaking, $^2/_3$ of what it returns is just re-packaged frequency information, same for AMs. This can even lead to the treacherous situation that corpus results based on some DM fit external evidence well mostly because the particular DM used is actually correlated more with frequency than with dispersion. In a way, in such a situation, it might be the fact that we are violating the first perspective (our DM is more determined by frequency than dispersion effects) that makes the measure seem to pass the second perspective (its result correlate well with external evidence).

Again, oftentimes this conflation is not necessarily a problem: Somewhat simplistically, the more descriptive the study, the less of a problem the conflation of different dimensions causes. However, as soon as the goal is more linguistic, theoretical, and/or psycholinguistic in nature than the simplest of descriptions, this kind of threat to validity becomes problematic and then addressing both perspectives is becoming more and more relevant: With interpretive goals, we need 'clean'/precise diagnostic tools (measures) – not ones tainted by conflation – that we can then maybe also relate to external evidence. And in fact, uncritically conflating frequency and dispersion can be problematic even in largely descriptive (or prescriptive) contexts such as lexicography (as when adjusted frequencies make words that are distributed completely differently seem distributed very similarly, see Gries 2020: 114 for one example).

In this paper, the second one of a 'two-paper paper', I want to discuss the notion of dispersion and how it is often computed and then used in corpus linguistic. I will focus on what I consider the most widely-used DMs (not that dispersion is widely used ...) and specifically on the question of how cleanly they actually measure dispersion and just dispersion. I will argue that nearly all of the most widely-used DMs are problematic precisely in the sense that they are not 'clean': They do not only measure dispersion but also frequency; in fact, they react more to frequency than they do to true dispersion, and in the sister publication to this paper, I argued that the same is true of the most widely-used AMs (such as G^2, t, or p_{FYE}).

This study, therefore, pursues the following goals. First and in Section 2, I will briefly recap an example to show how measures we use may be mislabeled a bit given that they reflect other things more; specifically, I will very briefly summarize how the 'association' measure G^2 seems to actually react more to frequency than to association or, minimally, conflates association information so much with frequency that the two are very hard to disentangle, which can lead to counterintuitive results. Then, Section 3 turns to the main point of the paper, DMs. Section 3.1 quickly surveys the main DMs that have been introduced and illustrate their computation from a term-document matrix. Section 3.2 motivates and defines a new DM that is by design untainted by frequency; this section uses the same logic that was employed in the sister publication to define an AM that is by design untainted by frequency. Section 3.3 then compares this new measure to a variety of traditional DMs in several corpora and shows that indeed it is much less correlated with frequency than all traditional DMs. Section 3.4 is a first attempt to validate the information that this new DM offers: using lexical decision times from the MALD database, I show that the new measure together with the (now largely orthogonal!) variable of frequency predicts reaction times better than all existing measures. Finally, Section 3.5 offers a very brief excursus on the collocations of *fast food* and *fast bowler* discussed in the previous paper on AMs to show how the new measure can also help in the case of co-occurrence/association data. Section 4 concludes. In order to make it easier for people to follow along or apply the logic of this paper to their own work, the exposition below will regularly provide R code; note, however, that understanding the R code is not required to understand the paper and readers unfamiliar with R can feel free to gloss over the code – the code is really only meant as help for readers who might want to write their own code for the measures discussed or proposed.

2. A brief recap: G^2 reacts more to frequency than to association

To illustrate the general nature of the problem – how measures that supposedly measure dimension *d* (e.g., dispersion/commonness or association/contingency) can actually return values that, to large extent, reflect something else – I will very briefly recap one example from the sister publication to this paper on AMs to show how the AM called the loglikelihood value or G^2 is not really a measure that delivers a clean association score but rather a measure that reflects mostly frequency and also some association.

How is G^2 computed? Most people do so from co-occurrence tables, which can be schematically represented as in Table 1.

Table 1. Schematic co-occurrence table of a word *w* and a construction *c*

	Construction: *c*	**Construction: other**	**Sum**
Word: *w*	*a*	*b*	*a*+*b*
Word: other	*c*	*d*	*c*+*d*
Sum	*a*+*c*	*b*+*d*	*a*+*b*+*c*+*d*

Let's define a hypothetical table of observed corpus results like `table.01.obs` as follows:

```
addmargins(table.01.obs <- matrix(c(50, 950, 350, 9998650), ncol=2,
   dimnames=list(WORD=c("w", "other"), CONSTRUCTION=c("c", "other"))))
##        CONSTRUCTION
## WORD       c   other      Sum
##   w       50     350      400
##   other  950 9998650  9999600
##   Sum   1000 9999000 10000000
```

For such a 2×2 table, G^2 can be computed from the observed and the expected frequencies, as represented in the usual formula here; we will use a small function

G2 (which also permits 0-frequencies that might otherwise cause problems for the log):

$$G^2 = 2\sum_{a}^{d} observed \times log\,\frac{observed}{expected}$$

```
c(“G2”=G2(table.01.obs))
##       G2
## 622.2269
```

However, the problem with G^2 is that it increases quite a bit when all frequencies of `table.01.obs` increase even though the ratios of the values in the table do not change (which of course entails that the actual association between *w* and *c* is no different from before):

```
addmargins(table.02.obs <- table.01.obs * 10)
##        CONSTRUCTION
## WORD        c     other       Sum
##   w       500      3500      4000
##   other  9500 99986500  99996000
##   Sum   10000 99990000 100000000
c(“G2”=G2(table.02.obs))
##       G2
## 6222.269
```

Similarly, G^2 also increases if only the overall frequency of *w* or *c* increases even if *w*'s or *c*'s distribution relative to *c* and *w* stay the same:

- in `table.03.obs`, *w* is twice as frequent as before, but still distributed with a 1-to-7 ratio over *c* vs. *other*;
- in `table.04.obs`, c is twice as frequent as before, but still distributed with a 1-to-19 ratio over *w* vs. *other*;
- yet in both tables, G^2 nearly doubles:

```
addmargins(table.03.obs <- matrix(c(100, 950, 700, 9998250), ncol=2,
   dimnames=list(WORD=c(“w”, “other”), CONSTRUCTION=c(“c”, “other”))))
##        CONSTRUCTION
## WORD       c   other      Sum
##   w      100     700      800
##   other  950 9998250  9999200
##   Sum   1050 9998950 10000000
c(“G2”=G2(table.03.obs))
##       G2
## 1239.451
addmargins(table.04.obs <- matrix(c(100, 1900, 300, 9997700), ncol=2,
   dimnames=list(WORD=c(“w”, “other”), CONSTRUCTION=c(“c”, “other”))))
##        CONSTRUCTION
## WORD       c   other      Sum
##   w      100     300      400
##   other 1900 9997700  9999600
##   Sum   2000 9998000 10000000
c(“G2”=G2(table.04.obs))
##       G2
## 1258.769
```

Importantly, a 'true' association-only measure such as the (log) odds ratio does not behave the same way because it recognizes that (i) what changed is mostly just the marginal frequencies of *w* and *c* and that (ii) the actual association of *w* and *c* is virtually the same in all four tables (note: I am computing the discounted odds ratio here, i.e. I add 0.5 to each cell before computing the odds ratio in case there is one or more cells with a frequency of 0):

```
log(odds.ratio(table.01.obs))
## [1] 7.323585
log(odds.ratio(table.02.obs))
## [1] 7.316393
log(odds.ratio(table.03.obs))
## [1] 7.319296
log(odds.ratio(table.04.obs))
## [1] 7.472703
```

In addition and more tellingly, the previous study then used a collocational case study – nouns after the adjective *fast* in the BNC to show that

- G^2-values are very much predictable from logged co-occurrence frequency (R^2 from a generalized additive model was 0.945);
- G^2-values are hardly at all predictable from a proper association measure such as the log odds ratios (R^2_{GAM} was 0.055);
- the association-only score of the log odds ratio was hardly predictable at all from logged co-occurrence frequency (R^2_{GAM} was 0.0241).

The focus of this paper is twofold: (i) to show that the situation is just as bad for most DMs, which mostly reflect frequency and not actually dispersion and (ii) to develop a gold-standard measure for dispersion that measures dispersion and just dispersion and is not also correlated with frequency already by its very design.

3. Dispersion measure: What do they measure and how?

3.1 Existing measures

Let's now apply the above logic to DMs and ask what DMs should do and then see whether that is what they actually do. They should quantify dispersion, the degree to which an element – usually, a word, but it could of course be any linguistic element – is distributed evenly in a corpus. Simply put, a word *W* can be distributed relatively evenly/regularly across the parts of a corpus, or relatively unevenly/clumpily. Most DMs fall into the interval [0,1] but those DMs form two classes differing in their orientation: For some, high values mean that words are distributed evenly, for others, high values mean words are distributed clumpily (which of course means each DM in the [0,1] interval can easily be transformed into the other class, if necessary). For a completely hypothetical, but still instructive example, consider a corpus with 500 pretty equally-sized parts (such as the Brown or LOB corpora or the ICE-GB) and a word *W* with a frequency of 1000 in the corpus as a whole. In such a case, if every corpus file contained 2 instances of *W*, this would constitute a completely even distribution, if all 1000 instances of *W* occurred in just 1 of the 500 corpus files (maybe even the smallest one), this would constitute a completely clumpy distribution, and any distribution of *W* in-between those two extremes should return a dispersion value somewhere between 0 and 1.

Gries (2008) surveyed a large number of dispersion statistics and adjusted frequencies and Gries (2010) showed that they fall into four groups (five, if chi-squared is recognized as a separate group):

- one group that includes Juilland's *D*, the probably most widely-used measure;
- one that includes *range* and Rosengrens's *S*;
- one that includes frequency, maxmin, and the standard deviation;
- one that includes Gries's own DP/DP_{norm}.

In what follows, I briefly discuss how some of these values are computed on the basis of a made-up toy example, namely a corpus represented here as a term-document matrix `tdm` (because that is the easiest and fastest way to compute the most widely-used DMs); consider the following `tdm` summarizing a toy 'corpus' with 16 different words (in the rows) and their frequencies in the 5 parts of the corpus (in the columns) in the cells:

```
##     p1 p2 p3 p4 p5
##   a  1  2  3  4  5
##   b  2  2  2  2  2
##   c  0  0  1  0  0
##   e  1  1  1  1  1
##   g  0  0  1  1  0
##   h  0  0  0  1  1
##   i  1  0  0  0  0
##   m  1  0  0  0  0
##   n  1  1  0  0  0
##   p  1  0  0  0  0
##   q  0  1  0  0  0
##   s  0  1  1  0  0
##   t  0  1  1  1  1
##   u  1  0  0  0  0
##   w  0  1  0  0  0
##   x  0  0  0  0  1
```

From this `tdm`, we can easily compute the absolute and relative frequencies of each word in each part of the corpus and in the complete corpus. For instance, *a* occurs 15 times in the corpus and 20% of its 15 occurrences are in p3:

```
rowSums(tdm.abs) # word frequencies
##  a  b  c  e  g  h  i  m  n  p  q  s  t  u  w  x
## 15 10  1  5  2  2  1  1  2  1  1  2  4  1  1  1
```

```
round(tdm.rel <- prop.table(tdm.abs, 1), 4)
##
##        p1     p2     p3     p4     p5
##   a 0.0667 0.1333 0.2000 0.2667 0.3333
##   b 0.2000 0.2000 0.2000 0.2000 0.2000
##   c 0.0000 0.0000 1.0000 0.0000 0.0000
##   e 0.2000 0.2000 0.2000 0.2000 0.2000
##   g 0.0000 0.0000 0.5000 0.5000 0.0000
##   h 0.0000 0.0000 0.0000 0.5000 0.5000
##   i 1.0000 0.0000 0.0000 0.0000 0.0000
##   m 1.0000 0.0000 0.0000 0.0000 0.0000
##   n 0.5000 0.5000 0.0000 0.0000 0.0000
##   p 1.0000 0.0000 0.0000 0.0000 0.0000
##   q 0.0000 1.0000 0.0000 0.0000 0.0000
##   s 0.0000 0.5000 0.5000 0.0000 0.0000
##   t 0.0000 0.2500 0.2500 0.2500 0.2500
##   u 1.0000 0.0000 0.0000 0.0000 0.0000
##   w 0.0000 1.0000 0.0000 0.0000 0.0000
##   x 0.0000 0.0000 0.0000 0.0000 1.0000
```

We can also compute the (absolute and relative) sizes of the corpus parts: p1 has 9 words, which corresponds to 18% of the 50-word corpus:

```
(corpus.part.sizes.abs <- colSums(tdm.abs)) # absolute
## p1 p2 p3 p4 p5
##  9 10 10 10 11
(corpus.part.sizes.rel <- colSums(tdm.abs) / sum(tdm.abs)) # relative
##   p1   p2   p3   p4   p5
## 0.18 0.20 0.20 0.20 0.22
```

The easiest-to-compute DM is *range*, which answers the question 'what is the number/proportion of corpus parts in which a/each word is attested in at least once?' or 'how much of the corpus in parts do you have to look at to see all instances of the word?'. For instance,

- *a* is in every corpus part so *range* is 1: to see all instances of *a*, you need to look at 100% of the corpus parts;
- *x* is just in one corpus part: to see all instances of it, you need to look at only 20% of the corpus parts:

```
(ranges <- apply(tdm.abs, 1, \(af) sum(af > 0))) / ncol(tdm.abs)
# af = anonymous function
##   a   b   c   e   g   h   i   m   n   p   q   s   t   u   w   x
## 1.0 1.0 0.2 1.0 0.4 0.4 0.2 0.2 0.4 0.2 0.2 0.4 0.8 0.2 0.2 0.2
```

This measure is very crude because it considers neither the number of times a word occurs in each corpus part nor, even more importantly, the sizes of the corpus parts. Just on the side, therefore, I would actually like to propose a slightly modified version of *range*, namely one that incorporates the sizes of the corpus parts in which a word is attested to at least some extent: $range_{\text{withsize}}$ is the sum of the sizes of the corpus parts in which the word is attested. This answers the question 'how much of the corpus have you seen *maximally* when you saw all instances of the word?':

```
(ranges.withsize <- apply(tdm.rel, 1,
   \(af) sum(corpus.part.sizes.rel[af > 0]) ))
##    a    b    c    e    g    h    i    m    n    p    q    s    t
## 1.00 1.00 0.20 1.00 0.40 0.42 0.18 0.18 0.38 0.18 0.20 0.40 0.82
##    u    w    x
## 0.18 0.20 0.22
```

This measure is already more discriminatory than *range* because it can distinguish the dispersion of *m* and *x*: each occurs in only one part of the corpus (so their traditional *range* values are 0.2) but *x*'s one occurrence is in a bigger corpus part than that of *m*, which is why its $range_{\text{withsize}}$ value is higher – a simple step but it already affords the widely-used measure of *range* more discriminatory power.

Two other simple measures are (i) sd_{pop}, the standard deviation (for the population) of the frequencies of each word in each file, and (ii) *varcoeff*, the variation coefficient (i.e. the standard deviation normalized by the mean):

```
sd.pop <- function (values) { sd(values)*sqrt((length(values)- 1) /
   length(values)) }
round(sds <- apply(tdm.abs, 1, sd.pop), 3)
##     a     b     c     e     g     h     i     m     n     p     q
## 1.414 0.000 0.400 0.000 0.490 0.490 0.400 0.400 0.490 0.400 0.400
##     s     t     u     w     x
## 0.490 0.400 0.400 0.400 0.400

round(varcoefs <- sds / apply(tdm.abs, 1, mean), 3)
##     a     b     c     e     g     h     i     m     n     p
## 0.471 0.000 2.000 0.000 1.225 1.225 2.000 2.000 1.225 2.000
##     q     s     t     u     w     x
## 2.000 1.225 0.500 2.000 2.000 2.000
```

The final more general measure is *idf* (inverse document frequency, see Spärck Jones 1972, Robertson 2004), the (here, binary) log of the number of corpus parts divided by the range:

```
round(idfs <- log2(ncol(tdm.abs)/ranges), 3)
##     a     b     c     e     g     h     i     m     n     p     q
## 0.000 0.000 2.322 0.000 1.322 1.322 2.322 2.322 1.322 2.322 2.322
##     s     t     u     w     x
## 1.322 0.322 2.322 2.322 2.322
```

Then, there is a variety of dedicated dispersion measures, and Juilland's *D* is the most widely known one (see Juilland et al. 1970). For the version we consider here, the one that can handle differently large corpus parts, we first need for each word how much in percent it makes up of each corpus part. The following shows that for the first three words; *a*, for instance, makes up 11.11% of the first corpus part (it occurs in there once and p1 has 9 words):

```
head(perc.of.corpus.part.that.is.element <- t(apply(tdm.abs, 1,
   \(af) af/corpus.part.sizes.abs)), 3)
##
##            p1  p2  p3  p4        p5
##   a 0.1111111 0.2 0.3 0.4 0.4545455
##   b 0.2222222 0.2 0.2 0.2 0.1818182
##   c 0.0000000 0.0 0.1 0.0 0.0000000
```

With that, we can compute Juilland's *D*, which is based on the variation coefficient of these percentages normalizing for the number of corpus parts:

```
round(juillandsd <- apply(perc.of.corpus.part.that.is.element, 1,
   \(af) 1-((sd.pop(af)/mean(af))/sqrt(ncol(tdm.abs)-1))), 3)
##     a     b     c     e     g     h     i     m     n     p     q
## 0.785 0.968 0.000 0.968 0.388 0.386 0.000 0.000 0.386 0.000 0.000
##     s     t     u     w     x
## 0.388 0.749 0.000 0.000 0.000
```

Carroll's (1970) D_2 is the normalized entropy of these percentages:

```
   entropy4d2 <- function (distr) {
   -sum((temp <- distr[distr > 0]/sum(distr)) * log2(temp)) /
   log2(length(distr))
}
round(carrollsd2 <- apply(perc.of.corpus.part.that.is.element, 1,
entropy4d2), 3)
##     a     b     c     e     g     h     i     m     n     p     q
## 0.938 0.999 0.000 0.999 0.431 0.430 0.000 0.000 0.430 0.000 0.000
##     s     t     u     w     x
## 0.431 0.861 0.000 0.000 0.000
```

Carroll's (1970) D_2 is virtually perfectly (negatively) correlated with the Kullback-Leibler divergence, which is the relative entropy of how much the distribution of the occurrences of a word across the files differs from the corpus part sizes; those values can be normalized to the [0,1] interval as well:

```
KLD <- function (post.true, prior.theory) {
   logs <- log2(post.true/prior.theory); logs[logs==-Inf] <- 0;
   return(sum(post.true*logs))
}
round(klds <- apply(tdm.rel, 1, KLD, corpus.part.sizes.rel), 3)
##     a     b     c     e     g     h     i     m     n     p     q
## 0.137 0.003 2.322 0.003 1.322 1.253 2.474 2.474 1.398 2.474 2.322
##     s     t     u     w     x
## 1.322 0.288 2.474 2.322 2.184

round(kldsnorm <- 1-2^(-klds), 3)
##     a     b     c     e     g     h     i     m     n     p     q
## 0.091 0.002 0.800 0.002 0.600 0.580 0.820 0.820 0.621 0.820 0.800
##     s     t     u     w     x
## 0.600 0.181 0.820 0.800 0.780
```

Rosengren's (1971) *S* is based on the relative sizes of the corpus parts and computed as follows:

```
round(rosengrenss <- apply(tdm.abs, 1, \(af)
   (sum(sqrt(af * corpus.part.sizes.rel))^2) / sum(af)), 3)
##     a     b     c     e     g     h     i     m     n     p     q
## 0.950 0.999 0.200 0.999 0.400 0.420 0.180 0.180 0.380 0.180 0.200
##     s     t     u     w     x
## 0.400 0.820 0.180 0.200 0.220
```

Finally, Gries's *DP* (for deviation of proportions) is computed from the difference between the distribution of the occurrences of a word across the files differs and the corpus part sizes, with a possible normalization added to make DP_{norm} fully exhaust the interval [0,1] even for corpora with very few parts (for corpora with many parts, the difference is negligible anyway):

```
(griessdp <- apply(tdm.rel, 1, \(af) sum(abs(af-corpus.part.sizes.rel))/2))
##    a    b    c    e    g    h    i    m    n    p    q
## 0.18 0.02 0.80 0.02 0.60 0.58 0.82 0.82 0.62 0.82 0.80
##    s    t    u    w    x
## 0.60 0.18 0.82 0.80 0.78

round(griessdpnorm <- griessdp/(1-min(corpus.part.sizes.rel)), 3)
##     a     b     c     e     g     h     i     m     n     p     q
## 0.220 0.024 0.976 0.024 0.732 0.707 1.000 1.000 0.756 1.000 0.976
##     s     t     u     w     x
## 0.732 0.220 1.000 0.976 0.951
```

If one computes all these measures for all word types of the Brown corpus, one finds that every one of these DMs is very highly correlated with frequency. The following are R^2-values that quantify how much of the dispersion values of all words is predictable from just the logged frequency of the words and we can see that none of the values is < 0.8:

```
##         VC     JUILLD         SD        KLD     CARRD2        IDF
##     0.8073     0.8074     0.8524     0.8727     0.8729     0.9035
##    KLDNORM   ROSGRENS         DP     DPNORM      RANGE RANGEWSIZE
##     0.9235     0.9440     0.9555     0.9555     0.9619     0.9619
```

(Incidentally, *range* and *range*$_{\text{withsize}}$ are so highly correlated here because the parts of Brown are nearly all the same size – in corpora with more varied sizes such as the BNC that would be different. Also, less well-known or differently-designed measures like the adjusted frequency measures proposed by Savický & Hlaváčová (2002), which are not based on the division of the corpus into parts, are also highly correlated with frequency.) This is clearly reminiscent of how, in the previous paper on AMs, logged co-occurrence frequency on its own was so predictive of AMs such as G^2 or t that one wonders whether these should even be called AMs anymore. And the above is not an isolated finding just for the Brown corpus. We find the same in the ICE-GB:

```
##         SD         VC     JUILLD        KLD     CARRD2        IDF
##     0.8062     0.8177     0.8184     0.8771     0.8853     0.9075
##    KLDNORM   ROSGRENS RANGEWSIZE      RANGE         DP     DPNORM
##     0.9159     0.9318     0.9471     0.9475     0.9482     0.9482
```

In other corpora, frequency accounts a little less for the dispersion measures, but still very well. For the complete BNC, for instance, nearly all R^2-values are around 0.8 and higher – only *KLD* fares at least a bit better (in how it nearly always has a by far lower correlation with frequency than all other measures):

```
##        KLD         VC     JUILLD    KLDNORM     CARRD2   ROSGRENS
##     0.6612     0.7884     0.7948     0.8497     0.8767     0.9015
##         SD        IDF         DP     DPNORM RANGEWSIZE      RANGE
##     0.9056     0.9127     0.9243     0.9243     0.9437     0.9507
```

Similarly for the BNC Baby, …

```
##        KLD         VC     JUILLD    KLDNORM     CARRD2   ROSGRENS
##     0.6051     0.7460     0.7533     0.8011     0.8219     0.8443
##         DP     DPNORM        IDF RANGEWSIZE      RANGE         SD
##     0.8571     0.8571     0.8679     0.8911     0.9104     0.9275
```

and the BNC sampler, …

```
##        KLD         VC     JUILLD    KLDNORM     CARRD2   ROSGRENS
##     0.6406     0.7428     0.7607     0.8222     0.8248     0.8647
##        IDF         DP     DPNORM RANGEWSIZE         SD      RANGE
##     0.8701     0.8802     0.8802     0.9067     0.9107     0.9210
```

and the spoken part of the BNC.

```
##        KLD         VC     JUILLD    KLDNORM     CARRD2   ROSGRENS
##     0.6243     0.8203     0.8262     0.8927     0.8959     0.9198
##         DP     DPNORM        IDF RANGEWSIZE      RANGE         SD
##     0.9236     0.9236     0.9281     0.9465     0.9607     0.9688
```

There is one additional comment on these observations that merits brief mention, especially for readers who have seen me discuss dispersion before. It has sometimes been argued that frequency and dispersion measures are highly correlated – the discussions I have witnessed involved *range*, Juilland's *D*, and Gries's *DP* – and that, therefore, it might suffice to use frequency only (i.e. to not bother with dispersion at all) or just a simple dispersion measure like *range* (in spite of the huge information loss of traditional *range*). To that, I usually responded by stating that (i) yes, frequency and dispersion are highly correlated (in Gries 2020, I myself report an R^2 from a GAM of regressing *DP* on logged frequency of > 0.8), but also that (ii) this correlation between frequency and dispersion is actually not strong at all when one considers the limited range of frequencies from which words for psycholinguistic experimentation might be recruited. For example, in the spoken BNC, the R^2 for the correlation between frequency and dispersion for words with frequencies in the interval [2036, 5838] (i.e. decent pmw frequencies of 195.6 and 560.8) is a mere 0.086. Thus, my main point in those discussions was not to argue against a generally high correlation of dispersion and frequency – that correlation indubitably exists and has been documented in my writing and elsewhere. Instead, my main point was always that, *for a certain range of words*, namely exactly the range that is interesting in much psycholinguistic work, the two are not so correlated so, if one is interested in controlling for words' commonness, one needs to consider frequency *and* dispersion because frequency is not the best proxy for commonness (see, e.g., Adelman et al. 2006; Brysbaert & New 2009; Baayen 2010; Gries 2010; Brysbaert et al. 2019) and, in that range, frequency does not also cover dispersion well enough.

That being said, the fact remains that, in general, i.e. over all words in a corpus, dispersion values are so highly correlated with (logged) frequency of occurrence that, by analogy to the previous discussion of AMs, one cannot help but wonder to what degree dispersion values make an *additional, useful, yet unique* contribution to frequency values when it comes to operationalizing words' 'commonness'. And the fact that dispersion is overall so highly related to frequency of course means that one might also doubt what DMs such as Juilland's *D* or Gries's *DP* contribute even within a frequency range where they are less correlated with frequency: Does the way these DMs are computed make a principally different contribution only in a certain frequency range? And if the answer was yes, would it then even make sense to compute dispersion values for all words, even those outside of that range? And how would one identify the range where frequency should be augmented with dispersion in the first place? It seems to me what is needed is something for dispersion that functions like the (logged) odds ratio for AMs, a gold-standard DM whose computation is *by definition* unrelated to frequency (perspective 1 from above), and then we need to check what, if anything, that DM contributes when it is correlated with external data (perspective 2 from above). The next subsection is devoted to discussing such a new DM and builds on the logic with which such a measure was developed for AMs in the sister publication to this one.

3.2 A new measure: Motivation and development

The main insight that triggered a certain unease with existing DMs (including my own) came from two sources. The first of these is the above-discussed fact that all main DMs right now are so strongly correlated with frequency that they do not seem to be 'clean' DMs that measure dispersion and not much else. This is a threat to validity: how can we interpret something in terms of dispersion if our measure is actually 0.9-correlated with something else (frequency) and we do not even have a gold standard measure for dispersion in the first place? The analogy I drew in the first paper on AMs to drive home this point is the following. Readers need to ask themselves the following question: Would they be happy if they went to a lab to have their blood checked for their cholesterol level, give a sample, pay the lab, and then be sent an email with the sentence "Your HDL level is *x*" but then they find out that the value *x* they are given is only correlated with their HDL value with an R^2_{GAM} of 0.1 but it reflects their blood glucose level really well with an R^2_{GAM} of 0.9? I doubt they would. Yes, that value is also interesting from a general health perspective – just as frequency is generally interesting for many (corpus-) linguistic applications – but it's not quite the same now, is it? So why do we do that in our own research?

The second impetus to address this actually arose for me from an example that I have often used to explain the need to not just consider the frequency of words as a proxy for their commonness but also their dispersion. In the Brown corpus, the two words *enormous* and *staining* have the same frequency of occurrence of 37 instances, but they have very different ranges: the 37 instances of *enormous* are in 36 different parts whereas the 37 instances of *staining* are in 1 part only. [2] Intuitively, this would mean one should expect massive differences in DMs because *enormous* is nearly as perfectly evenly distributed as a word with a frequency of 37 can be while *staining* is nearly as perfectly clumpily distributed as a word with that frequency can be – but many measures do not reflect that at all, as is shown in Table 2 (with the words *croaked* and *the* added for comparison)

Table 2. Some dispersion values for four words in the Brown corpus

				Rosengren's			
	range	*range*$_{\text{withsize}}$	*sd*	*S*	*KLD*$_{\text{norm}}$	*DP*	*DP*$_{\text{norm}}$
enormous	0.072	0.07201	0.27	0.072	0.93	0.92799	0.92977
staining	0.002	0.002	1.65	0.002	0.99	0.99800	0.99992
croaked	0.002	0.0019	0.04	0.002	0.99	0.99	1
the	1	1	34.4	0.985	0.03	0.094	0.095

For instance, given that *range*/*range*$_{\text{withsize}}$ can range from 0 to 1, from a strictly dispersion-based point of view (!), it seems 'weird' that a word like *enormous*, which is about as evenly dispersed as a word with that frequency can be, scores values that are close to the theoretical minimum; the same is true for Rosengren's *S*, the *DP*-measures, and *KLD*$_{\text{norm}}$ (although these measures have the reverse orientation). If *enormous* could talk, it would say to a researcher that quantifies its dispersion with, say, any of these measures, "You say you're looking to quantify dispersion, but you're giving me a low value mostly because of my low frequency because, let's face it, given my 37 occurrences, how could I possibly be more evenly dispersed?". For *sd*, it's nearly worse: The value for *the*, easily one of the most evenly distributed words in the English language, is very high, ok, but *staining* scores a little bit *higher* than *enormous* as if it was *more* evenly distributed than *enormous*, which it, clearly, is not. Now, one might say that *enormous* is evenly distributed, it's just not frequent enough (!) to show up in many corpus parts and that's why it is ranked as it is, with a dispersion value in the clumpiest decile. But *that* is the problem, because now we are again using the notion of

[2] Brysbaert & New 2009: 985 discuss the similar case of *creasy* and *measly*, both of which occur 63 times in SUBTLEX$_{\text{US}}$, but *creasy*'s occurrence are all in one corpus part whereas *measly*'s are spread out over 59 parts; similarly, see Oakes & Farrow 2007: 91 for discussion of an example similar to *staining* in Brown, namely *thalidomide* in FLOB.

frequency to 'excuse'/motivate a somewhat counterintuitive dispersion result. We would in essence be saying "*enormous* is not *frequent* (!) enough to score a dispersion value that recognizes its nearly maximally even dispersion". It is of course every researcher's right to say, "I want my measure *m* to reflect a little bit of dispersion and a lot of frequency", but then that researcher – most of us at this point – must answer the criticism that, if that is what their *m* does, calling it a DM is less than intuitive because it amounts to sneaking in a variable (frequency) into our analysis/discussion of a measure that, if it was 'clean', would actually measure something else (here, dispersion – in the previous paper, association). And in corpus linguistics, we have a history of doing that, we use that strategy all the time (because all our DMs are so extremely strongly correlated with frequency and because many of our AMs are, too). What we need is a way to measure dispersion that is *not* automatically dominated by, and thus correlated with, frequency, and we need that for two related reasons:

First, such a measure should look at *staining* and at *croaked* and conclude that, *given their frequencies*, they are as clumpily dispersed as they can possibly be, and such a measure would look at *enormous* and conclude that, yes, it is not that frequent, but since I am claiming I want to measure dispersion, not frequency, *given its frequency*, it is about as evenly dispersed as it can be. The "given your frequency" is what holds frequency constant and, thus, assesses dispersion separately. (And ideally, such a measure would not be as primitive as *range* and quantify clumpy/even dispersion only on the basis of whether a word shows up in some corpus part at all, but would also use how often it does and how big the corpus parts are.)

Second, and this is an even more fundamental point: if one's dispersion measure is so highly correlated with frequency (with R^2s in excess of 0.8), then one will never be able to find words for different combinations of frequency and dispersion values. If frequency and dispersion are as highly correlated as they are, we will by definition not be able to find high-frequency-and-high-dispersion words, high-frequency-and-low-dispersion words, low-frequency-and-high-dispersion words, and low-frequency-and-low-dispersion words, because if a word is of low frequency, nearly all existing measures 'condemn' it to also be low dispersion, hence the mismatch between *enormous*'s dispersion values and our recognition that, given its frequency, it could hardly be more evenly distributed.

In what follows, I outline such a measure, which will be based on *DP* and hence be called DP_{nofreq} (but it would not have to be: other bases are possible and I am using *DP* because it includes more information than *range* yet is extremely fast to compute). For a word *w*, it is computed in four steps, which are conceptually very similar to how the gold-standard AM was computed in the previous paper.

First, we compute the regular **observed *DP*-value** for *w* as before; Table 2 above lists them as 0.92799 and 0.998 for *enormous* and *staining* respectively.

Second, we compute the **highest *DP*-value** that any word *with that frequency* can have, i.e. the *DP*-value that represents the clumpiest distribution possible. That value is also extremely easy to compute because it is the *DP*-value that results from the hypothetical scenario that all occurrences of *w* are in the smallest corpus part or file. That hypothetical *DP*-value I will refer to as *upp*, because it is the upper *DP* limit for a word with this frequency; that value here is 0.9981.

Third, we compute the **lowest *DP*-value** (called *low*) any word *with that frequency* can have, i.e. the value that would result from the most even distribution for a word with that frequency, which is more complex. The most even distribution is one where the occurrences of *w* are distributed according to the file sizes as much as possible. While this sounds simple enough, it is actually a complex issue with three possible strategies to consider.

As for strategy 1: With truly incredible computational resources, it would theoretically be possible to generate all possible ways in which the *n* occurrences of *w* are distributed over all corpus parts so that one could compute the *DP*-values for each of them and identify the *low* as the smallest theoretically possible value called *low*. However, with even the smallest corpora, this is already impossible to compute: Even if one just wants to compute this for a word like *a* with 15 occurrences in our small toy corpus with 5 parts, the resulting simplex lattice (generated with `combinat::xsimplex`) has 3876 rows (because there are that many ways to distributed a mere 15 tokens over a mere 5 corpus parts) and the minimal *DP*-value that is returned by this combinatorics approach is 0.02. However, the computation of all ways in which a mere 10 occurrences of a word can be distributed over just 100 files requires >31,500 GB (that is not a typo); thus this theoretical ideal of an approach awaits quantum computing.

As for strategy 2: There is a seemingly convenient heuristic to solve this problem, which is to randomly distribute the occurrences of *w* over the corpus parts such that the sampling is biased by the sizes of the corpus parts, and we do so a certain number of times (e.g. 250 (or 500 or 1000 …) times). That way, we have 250 versions of what the data could look like if the occurrences of *w* were pretty much distributed as the sample sizes would makes us expect. Here is what this would look like if we distribute the 15 occurrences of *a* over the 5 corpus parts 9 times randomly but weighted by corpus part size:

```
##                  CORPUSPARTS
## SAMPLINGITERATION p1 p2 p3 p4 p5
##                 1  2  2  6  4  1
##                 2  3  4  2  4  2
##                 3  4  5  1  3  2
##                 4  1  3  5  1  5
##                 5  4  2  1  2  6
##             6  2  4  1  5  3
##             7  5  1  1  3  5
##             8  2  5  4  2  2
##             9  3  2  2  5  3
```

Then we can compute *DP* for each of these independently sampled rows and find the lowest *DP*-value:

```
## [1] 0.1533333
```

For the 37 occurrences of *enormous* or *staining* in the Brown corpus of 500 very similarly-sized files, the minimal *DP*-value obtained like this is much higher, namely 0.9253413. Crucially, this needs to be done once for every single observed word frequency, because we need to know *low* for each of them; this is not computationally complex, does not really require a lot of RAM, and can be parallelized, but it does increase computing time once corpora have more different word frequencies and more different parts:

- the Brown corpus contains ≈1m word tokens words that occur with 545 different frequencies across 500 parts and this simulation step took 50 seconds on a single thread;
- the spoken component of the BNC contains ≈10m word tokens words that occur with 1555 different frequencies across ≈900 parts and this simulation step took 900 seconds on a single thread;
- the complete BNC contains ≈100m word tokens words that occur with ≈5400 different frequencies across ≈4000 parts and this simulation step took 1515 seconds on 20 threads.

One potential problem with this approach is, therefore, the amount of time it would take for even something like the 10m words of the spoken part of the BNC, but the bigger problem is that, with rare words, but also in general, it is quite possible that none of the random samples hits on the truly most even distribution possible or one that is so close to the most even distribution that the computations resulting from it make no practical difference. In the above example, for instance, given the corpus part sizes of {0.18, 0.2, 0.2, 0.2, 0.22}, the most even distribution of *a*'s 15 instances would be {3, 3, 3, 3, 3}, which would lead to a minimal *DP*-value of 0.02 (as we saw in the combinatorics approach), but the above simulation example with only 9 iterations returns a minimal *DP*-value (i.e. *low*) of 0.1533, meaning it does not even come close to finding the right result. Of course, this is because 9 is a ridiculously small sampling number, but the more worrisome finding is that in this example, the really lowest possible *DP*-value is found only after at least 148 iterations (with the random number seed I used). Thus, the simulation approach helps if the number of iterations is high, but the ideal number of iterations is unknowable and higher numbers aggravate the first problem, namely how resource-intensive this approach is.

As far as I can see, the following, third strategy is best. It consists of deriving the most even distribution in a 'bottom-up' stepwise fashion. Imagine you have a word *w* that occurs three times in a corpus with only four parts, which have the following sizes (in a numeric vector called sizes):

```
## part1 part2 part3 part4
##   0.1   0.2   0.3   0.4
```

The algorithm generates a vector called `hypothetical` of zeros, one for each corpus part, and puts the first occurrence of *w* into the largest corpus part, leading to this distribution:

```
## part1 part2 part3 part4
##     0     0     0     1
```

Then, the algorithm performs the part of the *DP* computation that registers the differences between the observed distribution of *w* relative to *w*'s frequency and the corpus part sizes:

```
(hypothetical/word.freq) - sizes
##       part1       part2       part3       part4
## -0.10000000 -0.20000000 -0.30000000 -0.06666667
```

Clearly, the corpus part with the smallest value – part 3 – is the one in which *w* is currently most underrepresented so that part receives the next instance of *w*; thus, `hypothetical` changes to this:

```
hypothetical[3] <- 1; hypothetical
## part1 part2 part3 part4
##     0     0     1     1
```

Then, the algorithm iterates again:

```
(hypothetical/word.freq) - sizes
##       part1       part2       part3       part4
## -0.10000000 -0.20000000  0.03333333 -0.06666667
```

Now, the corpus part in which *w* is currently most underrepresented is part 2, which correspondingly receives the next instance of *w*, which is also the last instance of *w* we have to allocate (because our small didactic example stipulated a corpus frequency of *w* of 3):

```
hypothetical[2] <- 1; hypothetical
## part1 part2 part3 part4
##     0     1     1     1
```

This leads to a *DP*-value of $^{1}/_{6}$:

```
sum(abs(hypothetical/sum(hypothetical)-sizes))/2
## [1] 0.1666667
```

Once easily check combinatorially that this is indeed the smallest *DP*-value possible when one has to distribute three instances of *w* over four corpus part with the sizes from above:

```
prop.table(t(combinat::xsimplex(p=4, n=3)), 1)        |>
apply(1, \(af) sum(abs(af - c(0.1, 0.2, 0.3, 0.4)))/2) |>
min()
## [1] 0.1666667
```

The even better news is, if we apply the same logic to the word *a* on our above toy corpus of 50 words, this method retrieves the value of 0.02 we know from our combinatorics approach to be the right one, but is also extremely fast. Thus, this is the method we will use here.

Once the observed value, *upp*, and *low* have been computed, the fourth and final step is then, for each word, to

- create a 3-element vector consisting of its theoretical *low* value, its theoretical *upp* value, and the observed *DP*-value;
- transform that vector into the [0,1] interval such that
  - the smallest value of the 3-element vector (likely *low*) becomes 0;
  - the largest value of the 3-element vector (likely *upp*) becomes 1;
  - the last value of the 3-element vector (the observed *DP*-value) becomes whatever corresponds to its proportional position in the [0, 1] interval.

Here are some examples of what this 0–1 transformation does in some fictitious scenarios.

```
zero.to.one <- function (x) { (y <- x - min(x))/max(y) }
zero.to.one(c(low=0.2, upp=0.9, obs=0.55))
## low upp obs
## 0.0 1.0 0.5
zero.to.one(c(low=0.4, upp=0.9, obs=0.4))
## low upp obs
##   0   1   0
zero.to.one(c(low=0.4, upp=0.9, obs=0.8))
## low upp obs
## 0.0 1.0 0.8
zero.to.one(c(low=0.8, upp=0.9, obs=0.88))
## low upp obs
## 0.0 1.0 0.8
```

The last value of each transformed 3-element vector is the new dispersion-without-frequency metric for a word. For *enormous* and *staining*, this measure DP_{nofreq} returns the desired very nice results:

- for the very evenly distributed *enormous*, a value close to the maximally-even threshold of 0, namely 0.03638, and this value would be even smaller if, for instance, 2 of the 37 instances were not in the same file;
- for the very clumpily distributed *staining*, this measure returns a value close to the maximally-clumpy threshold of 1, namely 0.99886, and this value would be even higher if the 37 instances were clumped together not in the file in which they are but in a smaller file.

This measure, I propose, is what the (logged) odds ratio is for association, namely a gold-standard value that reflects dispersion and only dispersion, because its computation involves for every word *w* a normalization based on the possible range of *DP*-values for exactly *w*'s frequency of occurrence.[3] That also means that, since high frequencies are not required for DP_{nofreq} to recognize even dispersion, DP_{nofreq} can now do something that nearly no other DM can do, namely recognize low-frequency-but-high-dispersion words (like *enormous*).

Let us now evaluate this measure from the two perspectives outlined at the beginning.

3.3 Perspective 1: DP_{nofreq} measures dispersion, not frequency

First, we compare correlations of DP_{nofreq}, logged word frequency, and other DMs (in several corpora) with a series of plots, which require some explanation. Figure 1 compares how much *range* is correlated with, and hence already pre-

[3] Of course, if frequencies and DP_{nofreq} are computed for all words in a corpus, they could theoretically still be correlated, but, counter to what one reviewer asked, this is not a contradiction of the above. This is because there are at least two potential reasons for why two variables *x* and *y* (like, here, frequency and dispersion) might be correlated: (i) There could be a correlation because one measures *x* in such a way that the measurement of *x* already includes, or is contaminated with, *y*, meaning the correlation between *x* and *y* is not an empirical finding that might even be surprising, but a design feature of how they are measured: No one would be surprised that people's heights measured in inches and people's heights measured in centimeters are correlated and that would not be an empirical finding. (ii) There could be a correlation because even though *x* and *y* are measured in completely independent ways, the constructs they represent are correlated: At the risk of great simplification, the positive correlation between IQ and income (as, e.g. reported in Zagorsky 2007) is not due to IQ being measured in a way that already statistically includes income (or vice versa), it reflects something else (whatever that is, obviously correlation does not equal causation and obviously income is a function of multiple things). Thus, frequency and DP_{nofreq} could still be correlated, but then that would be an interesting empirical finding rather than a trivial mathematical consequence of how dispersion was measured.

dictable from, logged word frequency in the spoken part of the BNC and Figure 2 does the same for DP_{nofreq}. In these two plots,

- each word type is a point at the coordinates of its frequency and dispersion value;
- the red lines are regression lines from GAMs with the R^2 of the GAM indicated in the plot; remember that, for *range*, high values indicate even dispersion whereas for DP_{nofreq}, low values indicate even dispersion;
- the green 'error bars' indicate the ranges of observed dispersion values in each of 10 equally-spaced frequency bins (with the actual numerical range of dispersion values per bin indicated at the top of the plot). That means, for instance, that the *range* values for words with a frequency of around 2^6 (the 4th green bar from the left) only differ by about maximally 0.218 whereas the DP_{nofreq} values for words with that same frequency differ by about 0.619, which in turn means that frequency narrows down *range* immensely, but not DP_{nofreq}: as desired, DP_{nofreq} is able to quantify dispersion much more independently of word frequency;
- the blue line is the cumulative frequency of the word frequencies (same in both panels). That means, for instance, that 85% of all words have a frequency of $\leq 2^5=32$.

In every way, these two plots already show that *range* is fairly clearly determined already by frequency – another way of saying this would be that *range* does not add much information to whatever frequency already provides – whereas DP_{nofreq} is much more independent.

However, the most compelling (though admittedly not intuitive) indication of that is Figure 3, which shows on the *y*-axis how many of all the word types are in frequency bins with a dispersion range for each DM (*R* for *range, D* for DP_{nofreq}). For instance, the upper arrow indicates that 80% of all word types in the corpus are in a frequency bin that has a range of *range* values of a mere 0.022, meaning 80% of all word types in the corpus are in a frequency bin where the word type's frequency already nearly perfectly determines its *range* value. The lower arrow, by contrast, indicates that nearly 60% of all word types in the corpus are in a frequency bin that has a range of DP_{nofreq} values of ≥0.68, meaning nearly 60% of all word types in the corpus are in a frequency bin where the word type's frequency does not predict the word types' DP_{nofreq} values well at all – because we are now truly measuring dispersion independently of frequency.

Space does not permit a similarly detailed representation of the results for other corpora tested here; the results for Brown (comparing DP_{nofreq} to Rosengren's *S*), the ICE-GB (comparing DP_{nofreq} to *DP*), BNC Baby (comparing DP_{nofreq}

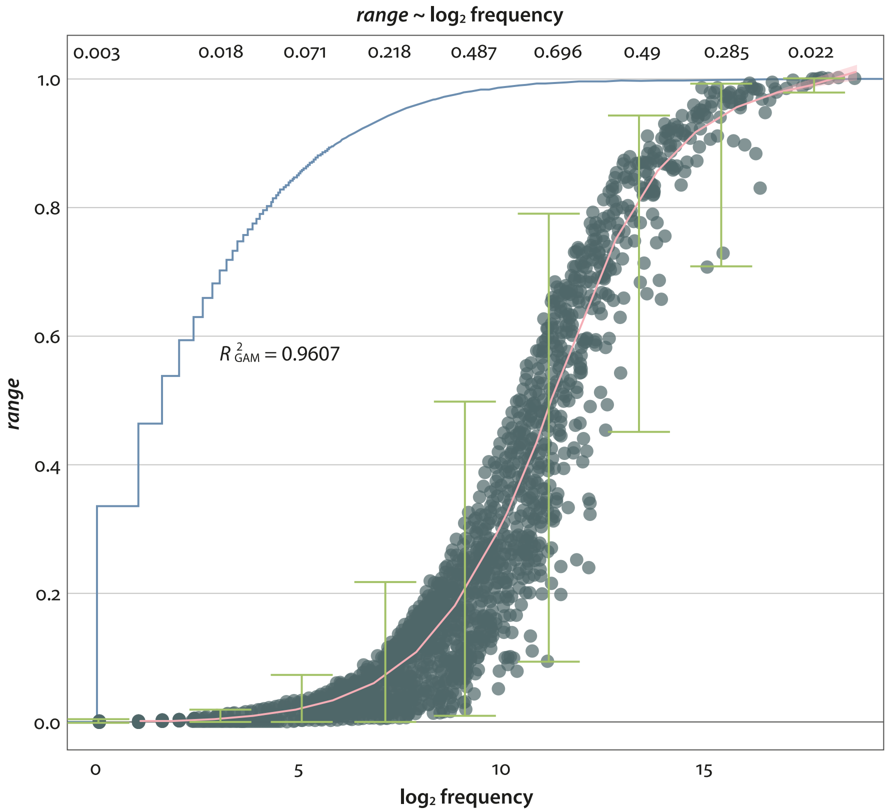

Figure 1. Evaluating *range* against logged frequency in the BNC spoken

to *IDF*), BNC sampler (comparing DP_{nofreq} to the variation coefficient), and the complete BNC (comparing DP_{nofreq} to Juilland's *D*) are, however, conceptually very similar and available as an online supplement *at <http://www.stgries.info/research/2022_STG_DispNoFreq_JSLS.zip>*.

Table 3 summarizes the correlations/comparisons across measures and corpora. While there are obvious quantitative differences, the overall picture is clear: Across a range of corpora and in comparisons with different DMs, we always find that the traditional DMs are all very much correlated with word frequency whereas DP_{nofreq} is much less so. More importantly, recall from note 3 that even though the correlations between logged frequency and DP_{nofreq} are still notable, the critical thing is that these correlations are not a design feature of the measure as in most DMs but an actual empirical finding: more frequent words are more likely to be more evenly dispersed *even if* one's definition of dispersion explicitly discards frequency. We also see that for most traditional DMs, when frequency is low, the range of DM-values is very small precisely because these DMs are largely

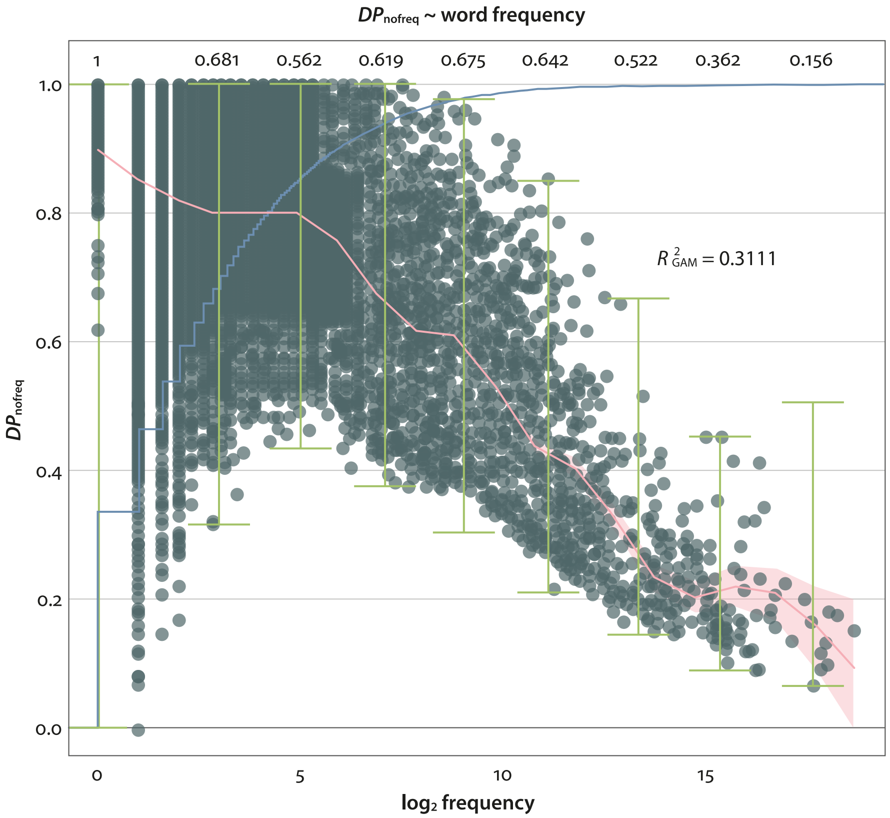

Figure 2. Evaluating DP_{nofreq} against *range* in the BNC spoken

'determined' or contaminated by frequency already, but in panel after panel we also see that even with very low frequencies, DP_{nofreq} can be large or small.

Table 3. R^2_{GAM} for various DMs and DP_{nofreq} across several corpora

Corpus	**Traditional DM**	**Traditional DM ~ freq**	**DP_{nofreq} ~ freq**
BNC spoken	*range*	R^2_{GAM}=0.9607	R^2_{GAM}=0.3111
Brown	Rosengren's *S*	R^2_{GAM}=0.944	R^2_{GAM}=0.5758
ICE-GB	*DP*	R^2_{GAM}=0.9482	R^2_{GAM}=0.5971
BNC	Juilland's *D*	R^2_{GAM}=0.7948	R^2_{GAM}=0.27
BNC Baby	*IDF*	R^2_{GAM}=0.8679	R^2_{GAM}=0.3966
BNC Sampler	*VarCoeff*	R^2_{GAM}=0.7428	R^2_{GAM}=0.2755

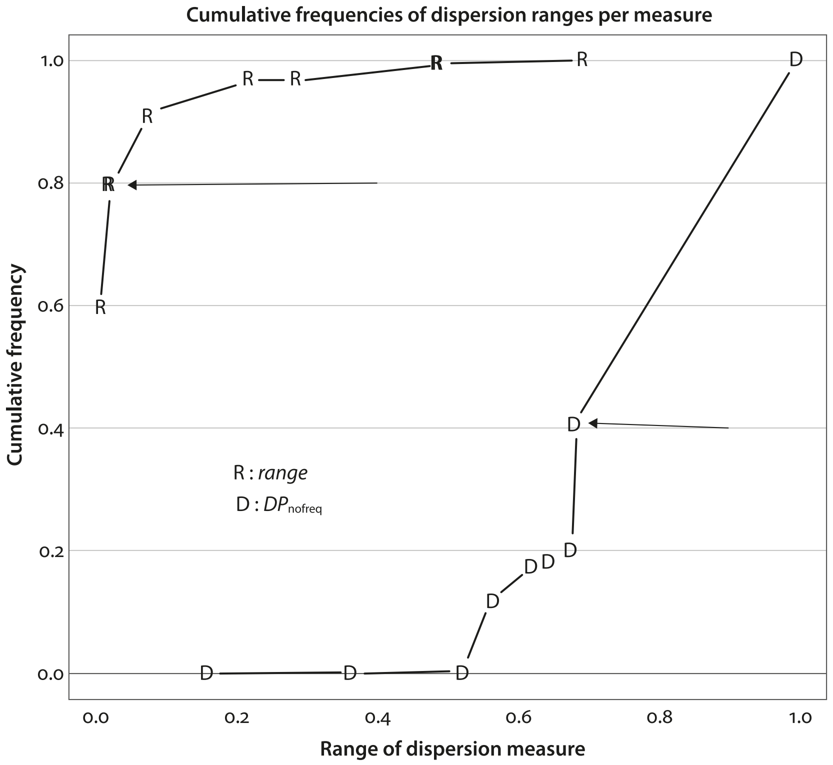

Figure 3. Evaluating DP_{nofreq} (against *range* in the BNC spoken)

Scatterplots and GAM fits (also in the *online supplement*) of each dispersion measure against DP_{nofreq} also show that, while each of the tested traditional DMs is very much predictable from frequency, none of them is predictable well and/or linearly from DP_{nofreq} although that is what those are thought to measure.

Now that we have established that DP_{nofreq} as defined here is different from all other DMs in how it does not most reflect frequency already by design, let us address perspective 2 – the correlation with external evidence – and determine how well it complements frequency as a predictor of lexical decision task reaction times compared to the traditional measures that conflate frequency and dispersion in a single value.

3.4 Perspective 2: DP_{nofreq} helps predicting external data

The question that remains is whether the new measure's main feature – its much greater independence from frequency compared to existing measures – does now also lead to at least some higher degree of predictive power when applied to external data. In the past (e.g. Gries 2010, 2019a), I showed that the measure *DP*, which I proposed in Gries (2008), has a higher degree of predictive power in monofactorial correlations with two small reaction time (RT) data sets (Spieler & Balota 1997, Balota & Spieler 1998, and Baayen 2008) than most traditional measures. Similar results were obtained by both Adelman et al. (2006) and Brysbaert & New (2009), who showed that *range* predicted word processing times better than frequency. Baayen (2010), too, shows that *range* is a better predictor of word processing times and that frequency as a mere repetition counter – i.e. exactly the way that most linguists and psycholinguists in general and cognitive linguists in particular have been endorsing frequency as a causal mechanism – is in fact epiphenomal![4]

However, by now, I find my approach there to be less than ideal because the *DP* measure(s) I proposed back then suffer(s) from the same problem as just about all others, namely that its very high correlation with frequency makes one wonder how much it really contributes above and beyond frequency; the results reported above in Section 3.1 suggest that, if anything, it is the KLD measure that is least correlated with frequency. In addition, the data set studied then was small and did not control for any other predictors (in particular word length). This section aims at putting the new measure DP_{nofreq}, but of course also the others, to a better test, which will be described now.

The RTs I will explore here are from the Massive Auditory Lexical Decision (MALD) database (Tucker et al. 2019). For the purposes of the study here, I retained only the RTs and lengths of all 227,179 word tokens and then merged these data with the words from each of the corpora such that a function would look at each word type in a corpus, check whether there are RT and length values for it in the MALD database and, if so, would add those to the dataframe with the

[4] Unfortunately and unlike Baayen, Adelman et al. (2006) and Brysbaert & New (2009) do not establish a connection to corpus-linguistic measures of dispersion in their work and use the slightly confusing name *contextual diversity* for *range*, when in fact the use of a word in different corpus parts by no means implies that the actual contexts of the word are different: No matter in how many different corpus parts *hermetically* is used, it will probably nearly always be followed by *sealed*. Yes, one could argue that they are using *context* as meaning 'document' or 'text', but (i) that is still not particularly intuitive (in linguistics (rather than information retrieval) we don't usually consider a word's context to be the whole text in which it appears) and (ii) that term makes the connection to range and dispersion even more opaque. Even in Brysbaert et al. (2019), the potential role of dispersion is not recognized, neither just as the proper term for what they are using nor as a predictor in the regression models (although my own analysis shows that *DP* and *KLD* are not insubstantially correlated with their prevalence scores). It seems, sadly, that the recognition of dispersion in psycholinguistics outside of Baayen's work will require a few more decades …

corpus frequencies and dispersions. Table 4 summarizes the number of RT tokens available for correlating it with (logged) frequency and/or dispersion.

Table 4. Numbers of word types used for testing DMs' predictive power

BNC Baby	BNC sampler	BNC spoken	BNC (total)	Brown	ICE-GB
93,708	80,146	89,394	111,975	91,110	68,886

The analytical method used here is based on the logic of proportional reduction of error (PRE) measures and involved the following steps (here described for the BNC Baby). First, I used a random forest to model the words' RTs as a function of their lengths; I used a random forest rather than some sort of regression model because random forests often have higher prediction accuracies than regression models, their prediction are already OOB predictions and, thus, 'cross-validated', and random forests are better at detecting non-linearities than most regression models. I then computed the residuals (observed RTs minus the random forest's predictions), and the median absolute deviation (MAD) of these residuals was considered the baseline, i.e. an overall amount of variability in RTs out of which word length has already been 'partialed out'; that MAD corresponds roughly to what in a regression modeling context might be the null deviance.

Second, I fit 27 different forests on the RTs with different predictors from the BNC Baby data: Each forest predicted the RT data based on length (just like the one used for the baseline) plus one or two additional predictors that are listed below:

- FREQ and FREQLOG;
- RANGE and RANGE + FREQLOG;
- RANGEWITHSIZE and RANGEWITHSIZE + FREQLOG;
- MAXMIN and SD and CHISQ;
- VC and VC + FREQLOG;
- IDF and IDF + FREQLOG;
- JUILLD and JUILLD + FREQLOG;
- ROSGRENS AND ROSGRENS + FREQLOG;
- CARRD2 and CARRD2 + FREQLOG;
- KLD and KLDNORM and KLDNORM + FREQLOG;
- DP and DPNORM;
- DPNOFREQ and FREQ + DPNOFREQ and FREQLOG + DPNOFREQ.[5]

[5] As a reminder: FREQ: raw observed frequency; FREQLOG: frequency logged to the base of 2; RANGE: *range*; RANGEWITHSIZE: range$_{withsize}$; MAXMIN: the difference between the largest and the smallest frequency of a word in a corpus part; SD: standard deviation; CHISQ: chi-squared

The main point to recognize here is that each DM – the traditional ones and DP_{nofreq} – is used as a predictor together with length and with two versions of frequency (the raw one and the logged one) so we can see which DM adds most predictive power to the baseline forest based on length alone.

Third, from each of these 27 random forests I computed (i) the predicted RTs for the data, (ii) the residuals of these predictions (i.e. how much they differed from the actually observed RTs), and then (iii) the MAD of these residuals in exactly the same way as before.

The fourth and final step then consisted of (i) computing the difference between the RT~LENGTH MAD baseline on the one hand and the MAD baseline of each of the 27 random forests on the others and (ii) expressing the improvement as a proportion of the RT~LENGTH MAD baseline (hence, "proportional reduction of error"). One can then sort the forests with their predictors to see which forest(s), and thus, which DM(s), result(s) in the highest improvement/reduction of the deviance.

From a very global perspective, the results are encouraging for the new measure. Figure 4 below shows the median PRE relative to a length-only baseline of all forests across all corpora, crucially, we see at the top that the forests that use length (as the obvious baseline) and then the combination of frequency (or logged frequency) and the new dispersion measure do best; it is also interesting to note that the next two measures are two that are not widely used: the third place is held by the *KLD* measure (see Gries 2020), and the fourth place is held by the improved version of range, $range_{\text{withsize}}$, that I proposed above.

At first glance, these results may not seem very impressive: The PREs in general are low and it's not like the forests involving the new measure 'blow all other measures out of the sky'. However,

- the fact remains that they do score highest (and let's not forget that newly-developed corpus-linguistic measures are often not tested against truly *external* data/standards like this in the first place; see, e.g., Kromer 2003 stating his adjusted frequency measure is psycholinguistically more adequate than others without any evidence for that assessment);
- I would actually already have considered the new measure somewhat of a success even if it had only been as good as the average of the existing DMs sim-

statistic; VC: variation coefficient; IDF: inverse document frequency; JUILLD: Juilland's *D* (for unequal corpus part sizes); ROSGRENS: Rosengren's *S* (for unequal corpus part sizes); CARRD2: Carroll's D_2; KLD: Kullback-Leibler divergence; KLDNORM: Kullback-Leibler divergence normalized to the interval [0,1]; DP: Deviation of Proportions; DPNORM: Deviation of Proportions normalized to the interval [0,1]; DPNOFREQ: $DP_{\text{withoutfrequency}}$.

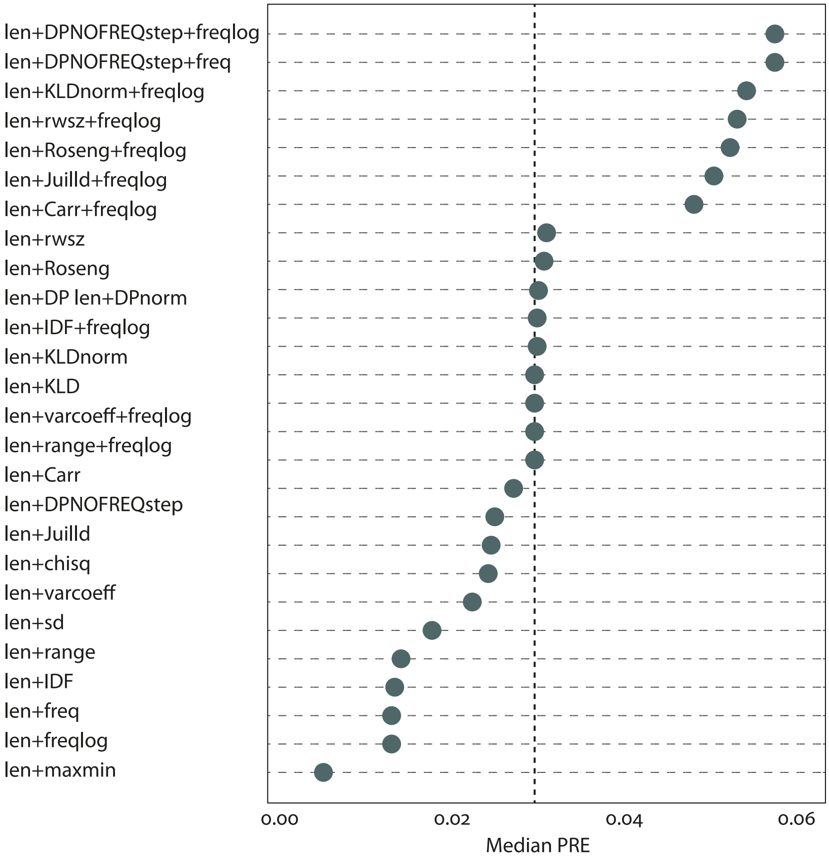

Figure 4. Median PRE-scores for 27 random forests

ply because it is 'clean' and, unlike others, controls for frequency (by holding it constant), i.e. it is not merely an amalgam of mostly frequency and 'some unknown amount of something else'.

On a corpus-by-corpus basis, the results hardly change the overall picture:

- the forests with DPNOFREQ score the highest PRE for all corpora but the complete BNC: the BNC Baby, the BNC Sampler, the spoken part of the BNC, the Brown corpus, and the ICE-GB;
- the forests with DPNOFREQ score a very close second-highest PRE for the BNC, with only a very small difference to Juilland's *D*.

And at least some such differences are to be expected, given how much genre effects can affect lexical predictors on processing (see Baayen et al. 2016 for discussion). Still, the bottom line is that DP_{nofreq} is a conceptually cleaner DM in how it can be measured independently of frequency and the forests with it outperform all other tested measures in 5 out of 6 corpora and in the overall aggregate in terms of the PRE/deviance of the lexical decision times.

4. Two short excurses

In this brief section, I want to briefly mention two consequences of the above discussion, which I cannot discuss in more detail, given space considerations.

4.1 Excursus 1 *range*$_{\text{nofreq}}$

As a first side remark, it is worth pointing out that the general approach used here for *DP* can also be used for other measures, where by “the general approach” I mean the notion of

- computing an observed value;
- computing the largest and smallest theoretically possible values given a word’s overall frequency;
- relativizing the observed value against the theoretically possible range.

For example, this can easily be applied to *range* or even *range*$_{\text{withsize}}$. Imagine we have a corpus with 5 parts with these sizes and we are interested in two words that have the following frequencies in those 5 parts:

```
rm(list=ls(all=TRUE))
corpus.part.sizes.rel <- c(0.1, 0.2, 0.25, 0.35, 0.1)
   names(corpus.part.sizes.rel) <- paste0(“part”, 1:5)
no.of.corpus.parts <- length(corpus.part.sizes.rel)
freq.of.word1.in.parts <- c(0, 1, 0, 1, 0)
freq.of.word2.in.parts <- c(0, 2, 3, 2, 1)
```

The *range*-values for *word*$_1$ and *word*$_2$ would be 0.4 and 0.8 respectively:

```
c(“range for word1”=(obs.word1 <- mean(freq.of.word1.in.parts>0)),
   “range for word2”=(obs.word2 <- mean(freq.of.word2.in.parts>0)))
## range for word1 range for word2
##             0.4             0.8
```

The value for *word*$_1$ is on the lower side of things, but it is again clear that, *for a word with that frequency*, it has the highest possible range because a word with two occurrences cannot be in more than two parts, i.e. `obs` is already `upp`. Hence, since we already have the observed ranges, we can now also compute `low` (which is always $^1/_{\text{no of corpus parts}}$):

```
low <- 1/no.of.corpus.parts)
## [1] 0.2
```

And then we compute upp (which is always ’all instances of the word are maximally spread out’) and transform the three values per word to the [0,1] interval to get range values that do not by design reflect frequency, too:

```
(upp.word1 <- min(2, no.of.corpus.parts)/no.of.corpus.parts)
## [1] 0.4
(range.wout.freq.word.1 <- zero2one(c(low, upp.word1, obs.word1))[3])
## [1] 1
```

However, consider now *word*$_2$, which could be more evenly dispersed because it occurs often enough to potentially occur at least once in every corpus part, and our new approach can see that:

```
(upp.word2 <- min(8, no.of.corpus.parts)/no.of.corpus.parts)
## [1] 1
(range.wout.freq.word.2 <- zero2one(c(low, upp.word2, obs.word2))[3])
## [1] 0.75
```

Again we see that, now, the DM really measures dispersion and recognizes the spread of a word *given its frequency*.

4.2 Excursus 2: *fast bowler* vs. *fast food*

In the previous paper on AMs, I showed how, when collocates of *fast* are studied on the basis of both their co-occurrence frequency and their association to *fast* (using an AM untainted by frequency), we find that *fast bowler* is only a bit less frequent than *fast food* but the association strength of *fast bowler* is notably greater than that of *fast food*, which is probably not what most people consider useful: Certainly, *fast food* is the more interesting/important collocation of the two. In that paper, I pointed at the ‘solution’: This result is due to us using ‘only’ pure frequency and pure association, but not also dispersion: *fast bowler* is concentrated in a much smaller number of files than *fast food*, i.e. its *range* is smaller. But, as mentioned above, *range* is a very crude measure that takes neither corpus part sizes nor the frequencies of words in the parts into consideration so what happens if we look at a more fine-grained measure such as *DP*? *DP* for *fast food* and *fast bowler* is 0.9533425 and 0.962413 respectively. This means that *fast food* is indeed more evenly dispersed, but (i) the difference between the values is very small (0.0090705 on a scale from 0 to 1) and (ii) both values are very close to the theoretical maximum of *DP* of 1 because that measure, like most traditional DMs, is overly influenced by the relatively low frequency of both collocations.

What happens if we apply our new measure to this case and, thus, partial out the effect that the slightly higher frequency of *fast food* may have, i.e. if we *really* only look at dispersion? The results change quite a bit: *DP*$_{\text{nofreq}}$ for *fast food* and *fast bowler* are now 0.7169702 and 0.7544027 respectively. Not only are the values nowhere nearly as maximal even for a collocation like *fast food* whose 154 instances are after all attested in 95 files, but also the difference to *fast bowler*, whose 134 instances are only attested in nearly 47 files, is now more pronounced: the difference between the two collocations' DP_{nofreq} is more than 4 times as large as the difference between their *DP*s. While the exact size of that numerical difference proves very little, I do think that not only does DP_{nofreq} capture the frequency-and-range ratios of the two collocations better than *DP* did, but I also prefer to not have near maximal dispersion values of >0.95 for words with such ranges, plus the bigger difference between the two collocations is certainly also more compatible with the much higher utility of *fast food* as a collocation than *fast bowler*. Be that as it may, the main point of this excursus was to show that the new dispersion-only measure can of course, and probably should, also be utilized in the identification of collocations so as to make sure that we do contextualize the frequencies and AM results properly (see again Gries 2019b for more discussion of this 'tupleization').

5. Concluding remarks

This paper tried to make three main points:

(1) Much of corpus linguistics is still using statistical measures that prioritize convenience (e.g., of sortability along one dimension) over the 'cleanness' of that dimension: unlike what about 50 years of publications on dispersion might make one expect, nearly all of our DMs reflect frequency more than they do dispersion (just as some of our most widely-used AMs reflect frequency more than they do association).

(2) I motivated and defined a DM called DP_{nofreq} that by design controls for frequency and, thus, does not conflate frequency and dispersion in its output; it is, therefore, a real, conceptually clean/pure DM, which can return very high or very low dispersion values for words regardless of the frequencies of the words.

(3) I showed that this measure coupled with the now independent notion of frequency nearly always has a higher degree of predictive power than that of previous less clean measures and even than that of previous less clean measures together with frequency.

Obviously and as always, more validation and testing will be necessary and the result of this might well be that true dispersion on its own has less predictive power than it might seem right now – but at least we would then know that and would be calling a spade a spade rather than, what I have done myself, promote the value of a measure of 'dispersion', which really derives most of its merit from frequency.

References

Adelman, James S., Gordon D.A. Brown, & José F. Quesada. 2006. Contextual Diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science 19*(9). 814–823. https://doi.org/10.1111/j.1467-9280.2006.01787.x

Baayen, R. Harald. 2008. *Analyzing linguistic data: a practical introduction to statistics with R.* Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511801686

Baayen, R. Harald. 2010. Demythologizing the word frequency effect: A discriminative learning perspective. *The Mental Lexicon 5*(3). 436–461. https://doi.org/10.1075/ml.5.3.10baa

Baayen, R. Harald, Petar Milin, & Michael Ramscar. 2016. Frequency in lexical processing. *Aphasiaology 30*(11). 1174–1220. https://doi.org/10.1080/02687038.2016.1147767

Balota, David A. & Daniel H. Spieler. 1998. The utility of item level analyses in model evaluation: a reply to Seidenberg and Plaut. *Psychological Science 9*(3). 238–240. https://doi.org/10.1111/1467-9280.00047

Bestgen, Yves & Sylviane Granger. 2009. Quantifying the development of phraseological competence in L2 English writing: An automated approach. *Journal of Second Language Writing 26*. 28–41. https://doi.org/10.1016/j.jslw.2014.09.004

Brysbaert, Marc & Boris New. 2009. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods 41*(4). 977–990. https://doi.org/10.3758/BRM.41.4.977

Brysbaert, Marc, Pawel Mandera, Samantha F. McCormick, & Emmanuel Keuleers. 2019. Word prevalence norms for 62,000 English lemmas. *Behavior Research Methods 51*. 467–479. https://doi.org/10.3758/s13428-018-1077-9

Carroll, John B. 1970. An alternative to Juilland's usage coefficient for lexical frequencies and a proposal for a standard frequency index. *Computer Studies in the Humanities and Verbal Behaviour 3*(2). 61–65.

Durrant, Phil & Norbert Schmitt. 2009. To what extent do native and non-native writers make use of collocations? *International Review of Applied Linguistics 47*. 157–177. https://doi.org/10.1515/iral.2009.007

Ellis, Nick C. 2007a. Language acquisition as rational contingency learning. *Applied Linguistics 27*(1). 1–24. https://doi.org/10.1093/applin/ami038

Ellis, Nick C. 2007b. The Associative-Cognitive CREED. In Bill VanPatten & Jessica Williams. (eds.), *Theories of second language acquisition: an introduction*, 77–95. Mahwah, NJ: Lawrence Erlbaum.

Ellis, Nick C., Rita Simpson-Vlach, & Carson Maynard. 2008. Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly 42*(3). 375–396. https://doi.org/10.1002/j.1545-7249.2008.tb00137.x

Evert, Stefan. 2009. Corpora and collocations. In Anke Lüdeling & Merja. Kytö. (eds.), *Corpus Linguistics: An International Handbook*, Vol. 2, 1212–1248. Berlin & New York: Mouton de Gruyter.

Fu, M. & Shaofeng, Li. 2019. The associations between individual differences in working memory and the effectiveness of immediate and delayed corrective feedback. *Journal of Second Language Studies 2*(2). 233-257 (25) https://doi.org/10.1075/jsls.19002.fu

Gries, Stefan Th. 2008. Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics 13*(4). 403–437. https://doi.org/10.1075/ijcl.13.4.02gri

Gries, Stefan Th. 2010. Dispersions and adjusted frequencies in corpora: further explorations. In Stefan Th. Gries, Stefanie Wulff, & Mark Davies. (eds.), *Corpus linguistic applications: current studies, new directions*, 197–212. Amsterdam: Rodopi. https://doi.org/10.1163/9789042028012_014

Gries, Stefan Th. 2019a. *Ten lectures on corpus-linguistic approaches: Applications for usage-based and psycholinguistic research*. Leiden & Boston: Brill. https://doi.org/10.1163/9789004410343

Gries, Stefan Th. 2019b. 15 years of collostructions: some long overdue additions/corrections (to/of actually all sorts of corpus-linguistics measures). *International Journal of Corpus Linguistics 24*(3). 385–412. https://doi.org/10.1075/ijcl.00011.gri

Gries, Stefan Th. 2020. Analyzing dispersion. In Magali Paquot & Stefan Th. Gries. (eds.), *A practical handbook of corpus linguistics*, 99–118. Berlin & New York: Springer. https://doi.org/10.1007/978-3-030-46216-1_5

Gries, Stefan, Th. 2021. What do (some of) our association measures measure (most)? Association? *Journal of Second Language Studies*. Available online: 12 November 2021. https://doi.org/10.1075/jsls.21028.gri

Juilland, Alphonse G., Dorothy R. Brodin, & Catherine Davidovitch. 1970. *Frequency dictionary of French words*. The Hague: Mouton de Gruyter.

Kromer, Victor. 2003. An usage measure based on psychophysical relations. *Journal of Quantitative Linguistics 10*(2). 177–186. https://doi.org/10.1076/jqul.10.2.177.16718

Oakes, Michael P. & Malcolm Farrow. 2007. Use of the Chi-Squared Test to examine vocabulary differences in English language corpora representing seven different countries. *Literary and Linguistic Computing 22*(1). 85–99. https://doi.org/10.1093/llc/fql044

Pecina, Pavel. 2009. Lexical association measures and collocation extraction. *Language Resources and Evaluation 44*(1–2). 137–158. https://doi.org/10.1007/s10579-009-9101-4

Robertson, Stephen. 2004. Understanding Inverse Document Frequency: on theoretical arguments of IDF. *Journal of Documentation 60*(5). 503–520. https://doi.org/10.1108/00220410410560582

Rosengren, Inger. 1971. The quantitative concept of language and its relation to the structure of frequency dictionaries. *Études de linguistique appliquée (Nouvelle Série) 1*. 103–127.

Savický, Petr & Jaroslava Hlaváčová. 2002. Measures of word commonness. *Journal of Quantitative Linguistics 9*(3), 215–231. https://doi.org/10.1076/jqul.9.3.215.14124

Schmid, Hans Joerg. 2010. Entrenchment, salience, and basic levels. In Dirk Geeraerts & Hubert Cuyckens. (eds.), *The Oxford Handbook of Cognitive Linguistics*, 117–138. Oxford: Oxford University Press.

Siyanova-Chanturia, Anna. 2015. Collocation in beginner learner writing: A longitudinal study. *System 53*. 148–160. https://doi.org/10.1016/j.system.2015.07.003

Spärck Jones, Karen. 1972. A statistical interpretation of term specificity and its application in information retrieval. *Journal of Documentation 28*(1). 11–21. https://doi.org/10.1108/eb026526

Spieler, Daniel H. & David A. Balota. 1997. Bringing computational models of word naming down to the item level. *Psychological Science 8*(6). 411–416. https://doi.org/10.1111/j.1467-9280.1997.tb00453.x

Tucker, Benjamin V., Daniel Brennerm, D. Kyle Danielson, Matthew C. Kelley, Filip Nenadić, & Michelle Sims. 2019. The Massive Auditory Lexical Decision (MALD) database. *Behavior Research Methods 51*. 1187–1204. https://doi.org/10.3758/s13428-018-1056-1

Zagorsky, Jay L. 2007. Do you have to be smart to be rich? The impact of IQ on wealth, income and financial distress. *Intelligence 35*(5). 489–501. https://doi.org/10.1016/j.intell.2007.02.003

Address for correspondence

Stefan Th. Gries
Department of Linguistics
University of California Santa Barbara
Santa Barbara, CA 93106-3100
United States
stgries@gmail.com

Publication history

Date received: 14 July 2021
Date accepted: 13 October 2021
Published online: 30 November 2021