Chapter 5
Analyzing Dispersion

Stefan Th. Gries

Abstract This chapter provides an overview of one of the most crucial but at the same time most underused basic statistical measures in corpus linguistics, dispersion, i.e. the degree to which occurrences of a word are distributed throughout a corpus evenly or unevenly/clumpily. I first survey a range of dispersion measures, their characteristics, and how they are computed manually; also, I discuss how different kinds of measures are related to each other in terms of their statistical behavior. Then, I address and exemplify the kinds of purposes to which dispersion measures are put in (i) lexicographic work and in (ii) some psycholinguistic explorations. The chapter then discusses a variety of reasons why, and ways in which, dispersion measures should be used more in corpus-linguistic work, in particular to augment simple frequency information that might be misleading; I conclude by discussing future directions in which dispersion research can go both in terms of how the logic of dispersion measures extends from frequencies of occurrence to co-occurrence and, potentially, even key words and in terms of how dispersion measures can be validated in future research on cognitive and psycholinguistic as well as applied-linguistics applications.

5.1 Introduction

Imagine a corpus linguist looking at a frequency list of the Brown corpus, a corpus aiming to be representative of written American English of the 1960s that consists of 500 samples, or parts, of approximately 2000 words each. Imagine further that corpus linguist is looking at that list to identify verbs and adjectives within a certain frequency range – maybe because he needs to (i) create stimuli for a psycholinguistic experiment that control for word frequency, (ii) identify words from a certain

S. Th. Gries (✉)
University of California Santa Barbara, Santa Barbara, CA, USA

Justus Liebig University Giessen, Giessen, Germany
e-mail: stgries@linguistics.ucsb.edu

M. Paquot, S. Th. Gries (eds.), *A Practical Handbook of Corpus Linguistics*,
https://doi.org/10.1007/978-3-030-46216-1_5

frequency range to test learners' vocabulary, or (iii) compile a vocabulary list for learners, or some other application. Imagine, finally, the frequency range he is currently interested in is between 35 and 40 words per million words and, as he browses the frequency list for good words to use, he comes across an adjective and a verb – *enormous* and *staining* – that he thinks he can use because they both occur 37 times in the Brown corpus (and are even equally long) so he notes them down for later use and goes on.

This is not an uncommon scenario and yet it is extremely problematic because, while that corpus linguist has indeed found words with the same frequency, he has probably not even come close to do what he actually wanted to do. The frequency range of the words he was interested in – 35-40 – or the actual frequency of the two words discussed – 37 – may have been an operationalization for things that might have to do with how fast people can identify the word in a psycholinguistic experiment (as in a lexical decision task) or with how likely a learner would be to have encountered, and thus hopefully know, a word of that kind of rarity. However, chances are that this choice of words is highly problematic: While both words are equally long and equally frequent in one and the same corpus, they could hardly be more different with regard to the topic of this chapter, their *dispersion*, which probably makes them useless for the above-mentioned hypothetical purposes, controlled experimentation, vocabulary testing, or vocabulary lists. This is because

- the word *enormous* occurs 37 times in the corpus, namely once in 35 corpus parts and twice in 1 corpus part;
- the word *staining* occurs 37 times in the corpus, namely 37 times in 1 corpus part.

In other words, given its (relatively low) frequency, *enormous* is pretty much as evenly dispersed as a word with that frequency can possibly be while, given its identical frequency, *staining* is as unevenly dispersed as a word with that frequency can possibly be: *enormous* is characterized by even dispersion, *staining* is characterized by a most uneven dispersion, clumpiness, or, to use Church and Gale's (1995) terms, high burstiness or bunchiness. In the following section, I will discuss fundamental aspects of the notion of dispersion, including some of the very few previous applications as well as a variety of dispersion measures that have been proposed in the past.

5.2 Fundamentals

5.2.1 An Overview of Measures of Dispersion

Corpus linguistics is an inherently distributional discipline: Virtually all corpus-linguistic studies with at least the slightest bit of a quantitative angle involve the frequency or frequencies with which

- an element *x* occurs in a corpus or in a part of a corpus representing a register or variety or something else, ... or
- an element *x* occurs in close proximity (however defined) to an element *y* in a corpus (or in a part of a corpus).

Also, any kind of more advanced corpus statistic – for instance, association measures (see Chap. 7) or key words statistics (see Chap. 6) is ultimately based on the observation of, and computations based upon, such frequencies. However, just like trying to summarize the distribution of any numeric variable using only a mean can be treacherous (especially when the numeric variable is not normally distributed), so is trying to summarize the overall 'behavior' (or the co-occurrence preferences or the keyness) of a word *x* on the basis of just its frequency/frequencies because, as exemplified above, words with identical frequencies can exhibit very different distributional behaviors.

On some level, this fact has been known for a long time. Baron et al. (2009) mention Fries & Traver's assessment that Thorndike was the first scholar to augment frequency statistics with range values, i.e. the numbers of corpus parts or documents in which words were attested at least one. However, this measure of range is rather crude: it does not take into consideration how large the corpus parts are in which occurrences of a word are attested, nor does its computation include how many occurrences of a word are in one corpus part – to have an effect on the range statistic, all that counts is a single instance. Therefore, during the 1970s, a variety of measures were developed to provide a better way to quantify the distribution of words across corpus parts; the best-known measures include Juilland's *D* (Juilland and Chang-Rodriguez 1964, Juilland et al. 1970), Carroll's D_2 (Carroll 1970), and Rosengren's *S* (Rosengren 1971).

To discuss how these statistics and some other competing ones are computed, I am following the expository strategy of Gries (2008), who surveyed all known dispersion measures on the basis of a small fictitious corpus; ours here consists of the following five parts:

b	**a**	m	n	i	b	e	u	p		
b	**a**	s	**a**	t	b	e	w	q	n	
b	c	**a**	g	**a**	b	e	s	t	**a**	
b	**a**	g	h	**a**	b	e	**a**	**a**	t	
b	**a**	h	**a**	**a**	b	e	**a**	x	**a**	t

This 'corpus' has several characteristics that make it useful for the discussion of dispersion: (i) it is small so all computations can easily be checked manually, (ii) the sizes of the corpus parts are not identical, which is more realistic than if they were, and (iii) multiple corpus-linguistically relevant situations are built into the data:

- the words *b* and *e* are equally frequent in each corpus part (two times and one time per corpus part respectively), which means that their dispersion measures should reflect those even distributions;

- the words *i*, *q*, and *x* are attested in one corpus part each: *i* in the first corpus part (which has 9 elements), *q* in the second corpus part (which has 10 elements), and *x* in the third corpus part (which has 11 elements), which means these words are extremely clumpily distributed, but slightly differently so (because the corpus parts they are in differ in size);
- the word *a*, whose dispersion we will explore below and which is highlighted in bold, is attested in each corpus part, but with different frequencies.

To compute the measures of dispersion to be discussed here, a few definitions are in order; we will focus on the word *a*:

(1) $l = 50$ (the length of the corpus in words)
(2) $n = 5$ (the length of the corpus in parts)
(3) $s = (0.18, 0.2, 0.2, 0.2, 0.22)$ (the percentages of the n corpus part sizes)
(4) $f = 15$ (the overall frequency of a in the corpus)
(5) $v = (1, 2, 3, 4, 5)$ (the frequencies of a in each corpus part 1-n)
(6) $p = ({}^{1}/_{9}, {}^{2}/_{10}, {}^{3}/_{10}, {}^{4}/_{10}, {}^{5}/_{11})$ (the percentages a makes up of each corpus part 1-n)

The most important dispersion measures – because of their historical value and evaluation studies discussed below – are computed as discussed in what follows; see Gries (2008) for a more comprehensive overview. The simplest measure is the *range*, i.e. the number of corpus parts in which the element in question, here *a*, is attested, which is computed as in (7):

(7) *range*: *number of parts containing a* $= 5$

Then, there are two traditional descriptive statistics, the standard deviation of the frequencies of the element in question in all corpus parts (*sd*, see (8)). This measure requires to take every value in v, subtract from it the mean of v (${}^{f}/_{n}$, i.e. 3), square those differences, and sum them up; then one divides that sum by the number of corpus parts n and takes the square root of that quotient:

(8) $sd_{population}$: $\sqrt{\frac{\sum_{i=1}^{n}\left(v_i - \frac{f}{n}\right)^2}{n}} \approx 1.414$ (sd_{sample} has n-1 in the denominator)

A maybe more useful variant of this measure is its 'normalized version, the variation coefficient (*vc*, see (9)); the normalization consists of dividing $sd_{population}$ by the mean frequency of the element in the corpus parts ${}^{f}/_{n}$:

(9) $vc_{population}$: $\frac{sd_{population}(v)}{mean(v)} \approx 0.471$ (vc_{sample} would use sd_{sample})

The version of Juilland's *D* that can handle differently large corpus parts is then computed as shown in (10). In order to accommodate the different sizes of the corpus parts, however, the variation coefficient is not computed using the observed frequencies $v_{1\text{-}n}$ (i.e. 1, 2, 3, 4, 5 in files 1 to 5 respectively, see (5) above) but using

the percentages in $p_{1\text{-}n}$ (i.e. how much of each corpus part is made up by the element in question, i.e. $^{1}/_{9}$, $^{2}/_{10}$, $^{3}/_{10}$, $^{4}/_{10}$, $^{5}/_{11}$, see (6) above), which is what corrects for differently large corpus parts:

(10) Juilland's D: $1 - \frac{sd_{population}(p)}{mean(p)} \times \frac{1}{\sqrt{(n-1)}} \approx 0.785$

Carroll's D_2 is essentially a normalized version of entropy of the proportions of the element in each corpus part, as shown in (11) (see also Gries 2013: Sect. 3.1.3.1 for general applications of this measure). The numerator computes the entropy of the percentages in $p_{1\text{-}n}$ while dividing it by $\log_2 n$ amounts to normalizing it against the maximally possible entropy given the number of corpus parts n.

(11) Carroll's D_2: $\frac{-\sum_{i=1}^{n}\left(\frac{p_i}{\sum p} \times \log_2 \frac{p_i}{\sum p}\right)}{\log_2 n} \approx 0.938$

The version of Rosengren's S that can handle differently large corpus parts is shown in (12). Each corpus part size's in percent (in s) is multiplied with the frequencies of the element in question in each corpus part (in $v_{1\text{-}n}$); of each product, one takes the square root, and those are summed up, that sum is squared, and divided by the overall frequency of the element in question in the corpus (f):

(12) Rosengren's (1971) S_{adj}: $\left(\sum_{i=1}^{n}\sqrt{s_i \cdot v_i}\right)^2 \times \frac{1}{f} \approx 0.95 \ (with \ min \ S{=}1n)$

Finally, Gries (2008, 2010) and the follow-up by Lijffijt and Gries (2012) proposed a measure called *DP* (for deviation of proportions), which falls between 1-*min s* (for an extremely even distribution) and 1 (for an extremely clumpy distribution) as well as a normalized version of *DP*, DP_{norm}, which falls between 0 and 1, which are computed as shown in (13). For *DP*, one computes the differences between how much of the element in question is in each corpus file in percent on the one hand and the sizes of the corpus parts in percent on the other – i.e. the differences between observed and expected percentages. Then, one adds up the absolute values of those and multiplies by 0.5; the normalization then consists of dividing this values by the theoretically maximum value of *DP* given the number of corpus parts (in a way reminiscent of (11)[1]:

(13) DP: $0.5 \times \sum_{i=1}^{n}\left|\frac{v_i}{f} - s_i\right| = 0.18$ and DP_{norm}: $\frac{DP}{1-mins} \approx 0.22$

The final measure to be discussed here is one that, as far as I can tell, has never been proposed as a measure of dispersion, but seems to me to be ideally suited to be one, namely the Kullback-Leibler (or KL-) divergence, a non-symmetric measure that quantifies how different one probability distribution (e.g., the distribution of all the occurrences of a across all corpus parts, i.e. $^{v}/_{f}$) is from another (e.g., the

[1] As pointed out by Burch et al. (2017), DP_{norm} is equivalent to a measure called *ADA* (for average deviation analog) proposed by Wilcox (1973).

Table 5.1 Dispersion measures for several 'words' in the above 'corpus'

	b	*i*	*q*	*x*
Range	5	1	1	1
Sd/vc	0/0	0.4/2	0.4/2	0.4/2
Juilland's *D*	0.968	0	0	0
Carroll's D_2	0.999	0	0	0
Rosengren's *S*	0.999	0.18	0.2	0.22
DP/DP_{norm}	0.02/0.024	0.82/1	0.8/0.976	0.78/0.951
KL-divergence	0.003	2.474	2.322	2.184

corpus part sizes *s*); the KL-divergence is computed as shown in (14) (with $\log_2$s of 0 defined as 0):

(14) KL-divergence: $\sum_{i=1}^{n} \frac{v_i}{f} \times \log_2 \left(\frac{v_i}{f} \times \frac{1}{s_i} \right) \approx 0.137 \; with \; \log_2 0 := 0$

Table 5.1 shows the corresponding results for several elements in the above 'corpus'. The results show that, for instance, *b* is really distributed extremely evenly (since it occurs twice in each file and all files are nearly equally large). Note in particular how the values of Rosengren's *S*, *DP*, and the KL-divergence for *i*, *q*, and *x* differ: all three occur only once in the corpus, only in one corpus part, but what differs is the size of the corpus part, and the larger the corpus part in which the single instance of *i*, *q*, or *x* is attested, the more even/expected that distribution is.

In sum, corpus linguists have proposed quite a few different measures of dispersion, most of which are generally correlated with each other, but that also react differently to the kinds of distributions one finds in corpus data, specifically,

- the (potentially large) number of corpus parts in which an element is not attested;
- the (potentially large) number of corpus parts in which an element is attested much less often than the mean;
- the range of distributions a corpus linguist would consider to be different but that would yield the same dispersion measure(s);
- the number of different corpus parts a corpus linguist would assume and their (even or uneven sizes).[2]

[2]Some dispersion measures do not require a division of the corpus into parts and/or also involve the differences between successive mentions in a corpus parts. These are theoretically interesting alternatives, but there seems to be virtually no research on them; see Gries (2008, 2010) for some review and discussion as well as the Further reading section for a brief presentation of one such study, Savický & Hlaváčová (2002).

5.2.2 Areas of Application and Validation

There are at least a few areas where dispersion information is now considered at least occasionally, though much too infrequently. The area of research/application where dispersion has gained most ground is that of corpus-based dictionaries and vocabulary lists. Leech et al. (2001) discuss dispersion information of words in the British National Corpus (BNC) and remark that, as in the *enormous*/*staining* example above, for instance, the words *HIV*, *lively*, and *keeper* are approximately equally frequent in the corpus, but are very differently dispersed in the corpus and proceed to use Juilland's D as their measure of choice. Similarly, Davies and Gardner (2010) and Gardner and Davies (2014) also use Juilland's D in their frequency dictionary and academic vocabulary list, as does Paquot (2010) for her academic keyword list.

It is worth pointing out in this connection that, especially in this domain of dictionaries/vocabulary lists, researchers have often also computed what is called an *adjusted frequency*, i.e. a frequency that is adjusted downwards depending on the clumpiness/unevenness of the distribution. In the mathematically simplest case, the adjusted frequency is the observed frequency of the word in the corpus times the dispersion value; for instance, Juilland's usage coefficient U is just that: the frequency of the word in the corpus f times Juilland's D, a measure that, for instance, Davies and Gardner (2010) use. In the above case for the word *a*, $U = 15 \times 0.785 = 11.777$ whereas for *q*, $U = 1 \times 0 = 0$; similar adjusted frequencies exist for Carroll's D_2 (the so-called Carroll's U_m) and Rosengren's S (the so-called Rosengren's AF).

Another area where dispersion information has at least occasionally been recognized as important is psycholinguistics, in particular the domain of lexical decision tasks. Consider, for instance, Schmid's (2010:115) concise summary: "frequency is one major determinant of the ease and speed of lexical access and retrieval, alongside recency of mention in discourse." And yes, for many decades now, (logged) frequency of occurrence has been known to correlate with reaction times to word/non-word stimuli. However, compared to frequency, the other major determinant, recency, has been considered much less in cognitive and psycholinguistic work. This is somewhat unexpected because there are general arguments that support the importance of dispersion as a cognitively relevant notion, as the following quote demonstrates:

> 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)

Ambridge et al. do not mention dispersion directly, but what would be its direct corpus-linguistic operationalization. Similarly, Adelman et al. (2006:814) make the valid point that "the extent to which the number of repeated exposures to a particular item affects that item's later retrieval depends on the separation of the exposures in time and context," and of course the corpus-linguistic equivalent to this "separation of the exposures in time and context" is dispersion.

More empirically, there are some studies providing supporting evidence for the role of dispersion when it comes to lexical decision tasks. One such study is in fact Adelman et al. (2006), who study dispersion. Their study has a variety of general problems:

- they only use the crudest measure of dispersion possible (*range*) and do not relate to previous more psychological/psycholinguistic work that also studied the role of range (such as Ellis 2002a, b);
- they do not establish any relation to the notion of dispersion in corpus linguistics and, somewhat worse even, refer to *range* with the misleading label *contextual diversity*, 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*.

Nonetheless, they do show that dispersion is a better and more unique predictor of word naming and lexical decision times than token frequency and they, like Ellis (2011), draw an explicit connection to Anderson's rational analysis of memory. More evidence for the importance of dispersion is offered by Baayen (2010), who includes range in the BNC as a predictor in a multifactorial model that ultimately suggests that the effect of frequency when considered a mere repetition-counter as opposed to some other cognitive mechanism is in fact epiphenomenal and can partly be explained by dispersion, and Gries (2010), who shows that lexical decision times from Baayen (2008) are most highly correlated with *vc* and DP/DP_{norm} (see Box 2 for details).

In spite of all the effort that has apparently gone into developing measures of dispersion and in spite of uneven dispersion posing a serious threat to the validity of virtually all corpus-based statistics, it is probably fair to say that dispersion is still far from being routinely included in both (more) theoretical research and (more) practical applications. One early attempt to study the behavior of these different measures is Lyne (1985), who compared *D*, D_2, and *S* to each other using 30 words from the French Business Correspondence Corpus, which for that application was divided into 5 equally large parts; on the basis of testing all possible ways in which 10 words can be distributed over 5 corpus parts, Lyne concludes that Juilland's *D* performs best; see also Lyne (1986), but there is little research that includes dispersion on a par with frequency or other corpus statistics and even less work that attempts to elucidate which measures are best (for what purpose); two studies that begin to work on this important issue are summarily discussed below.

Representative Study 1

Biber D., Reppen, R., Schnur, E., and Ghanem, R. 2016. On the (non)utility of Juilland's D to measure lexical dispersion in large corpora. *International Journal of Corpus Linguistics* 21(4): 439–464.

Starting out from observations in Gries (2008), Biber et al. (2016) is one of the most comprehensive tests, if not the most comprehensive one, of how the perceived default of Juilland's *D* behaves in particular with contemporary corpora that are large and have many different corpus parts, i.e. high values of *n*.

They begin by discussing the mathematical characteristics of Juilland's *D*, in particular the fact that the formula shown above in (10) increases "degrees of uniformity" (i.e. evenness of distribution/dispersion across corpus parts) "as the number of corpus parts is increased" (Biber et al. 2016:443); thus, the larger the corpora one considers, the more likely one uses a relatively large number of corpus parts (for reasons of statistical sampling), and the more Juilland's *D* is reduced, which "inflat[es] the estimate of uniformity, and overall, greatly reduc[es] the effective range of values for *D*" (p. 444).

Biber et al. then proceed with two case studies. The first one explores *D*-values of a set of words in the British National Corpus, which, for the purpose of testing what effect the numbers of corpus parts *n* one assumes, was divided into $n = 10$, 1000, and 1000 equal-sized parts; crucially, the words explored were words for which theoretical considerations would lead an analysis to expect fairly different *D*-values, contrasting words such as *at*, *all*, or *time* (which should be distributed fairly evenly) with words such as *erm*, *ah*, and *urgh* (which, given their preponderance in spoken data, should be distributed fairly unevenly). Specifically, they analyzed 153 words in 10 categories emerging from crossing (i) several different word frequency bands and (ii) expected distribution (uniform, writing-skewed, and speech-skewed).

In this first case study, they find the expected high *D*-values for higher-frequency words that would be uniformly-distributed or skewed towards writing (i.e. the 90% majority of the BNC) regardless of *n*. However, they also discover that the *D*-values for lower-frequency writing-skewed words are quite sensitive to variations of *n*. Their concern that these results are *not* due to the larger sampling sizes reflecting the dispersions more accurately is supported by what they find for the speech-skewed words, namely "extremely large discrepancies even for the most frequent speech-skewed words" (p. 450). More precisely, *D*-values for high-frequency speech-skewed words can vary between very high (e.g. 0.885 for *yeah* with $n = 1000$) and very low (e.g.

(continued)

0.286 for *yeah* with $n = 10$). Even more worryingly, "[t]hese discrepancies become even more dramatic as [they] consider moderate and lower-frequency words" (p. 452), with differences in *D*-values frequently exceeding 0.5 just because of varying *n*, which on a scale from 0 to 1 of course corresponds to what seems to be an unduly large effect. Their main conclusion of the first case study is that "*D* values based on 1,000 corpus parts completely fail to discriminate among words with uniform versus skewed distributions in naturalistic data" (p. 454).

In their second case study, Biber et al. created different data sets with, therefore, known distributions of target words across different numbers of corpus parts, but the bottom line of this more controlled case study is in fact the same as that of the first. Their maybe most extreme, and thus worrying, result is that

> the exact same distribution of a target word – a uniform distribution across 10% of a corpus – can result in a D value of 0.0 when the computation is based on a corpus split into 10 parts, versus a D value of 0.905 when the computation is based on a corpus split into 1000 parts. (p. 457)

As a more useful alternative, they propose to use Gries's (2008) *DP*. They recommend *DP* because it is conceptually simple, can easily handle unequally large corpus parts, and "it seems to be a much more reliable estimate of dispersion (and uniformity) in large corpora divided into many corpus parts" (p. 459). In a direct comparison with Juilland's *D*, they show that *DP* not only returns values from a more useful wider range of values when given a diverse set of differently dispersed words, but it also reacts differently to larger numbers of *n*: (1-*DP*) values are consistently lower for corpus divisions into many parts, which Biber et al. interpret as being desirably compatible with the expected benefits of the finer-grained sampling that comes with increasing *n*:

> Theoretically, we would expect more conservative estimates of dispersion based on a large number of corpus parts. For example, it is more likely that a word will occur in 6 out of 10 corpus parts than for that same word to occur in 600 out of 1000 corpus parts. The values for 1-*DP* seem to reflect this fact, resulting in consistently lower values when computations are based on a large number of corpus parts. In summary, *DP* is clearly more effective than *D* at discriminating between uniform versus skewed distributions in a corpus, especially when it is computed based on a large number of corpus-parts. (Biber et al. 2016:460)

Biber et al. conclude with a plea for more validation and triangulation when it comes to developing corpus-linguistic statistics and/or more general methods.

Representative Study 2

Gries, S.T. 2010. Dispersions and adjusted frequencies in corpora: further explorations. In *Corpus linguistic applications: current studies, new directions*, eds. Gries S.T., Wulff S., and Davies, M., 197–212. Rodopi, Amsterdam.

The second representative study to be discussed here is concerned with dispersion and its role in psycholinguistic contexts. Gries (2010) is an attempt to provide at least a first glimpse at how different dispersion measures are behaving statistically and predictively when studied in conjunction with psycholinguistic (reaction time) data. To that end, he conducted two kinds of case studies: First, he explored the degree to which the many existing measures capture similar kinds of dispersion information by exploring their intercorrelations; second, he computed the correlations between raw frequency, all dispersion measures, and all adjusted frequencies on the one hand and experimentally-obtained reaction time data from lexical decision tasks in psycholinguistics; in what follows, I briefly discuss these two case studies.

As for the first case study, he extracted all word types from the spoken component of the BNC that occur 10 or more times – there are approx. 17,500 such types – and computed all 29 dispersion measures and adjusted frequencies cataloged in the most recent overview article of Gries (2008). All measures were z-standardized (to make their different scales more comparable) and then used as input to both hierarchical agglomerative cluster analyses (see Chap. 18) and principal component analyses (see Chap. 19) separately for dispersion measures and adjusted frequencies. For the former, he used 1-Pearson's r (see Chap. 17) as a similarity measures and Ward's method as an amalgamation rule.

The results from both analyses revealed several relatively clear groupings of measures. For instance, the following clusters/components were well established in both the cluster and the principal components analysis:

- Rosengren's *S*, *range*, and a measure called *Distributional Consistency* (Zhang et al. 2004);
- Juilland's *D*, Carroll's D_2, and a measure called *D3* based on chi-squared (Lyne 1985); and, more heterogeneously,
- *DP*, DP_{norm}, *vc*, and *idf* (inverse document frequency, see Spärck Jones 1972 and Robertson 2004);
- frequency, the maxmin measure (the difference between $\max(v_{1\text{-}n})$ and $\min(v_{1\text{-}n})$), and *sd*.

In fact, the principal components analysis revealed that just two principal components capture more than 75% of the variance in the 16 dispersion

(continued)

measures explored: many measures behave quite similarly and fall into several smaller groups. Nevertheless, the results also show that the groups of measures also sometimes behave quite dissimilarly: "different measures of dispersion will yield *very* different (ranges of) values when applied to actual data" (Gries 2010:204, his emphasis).

With regard to the adjusted frequencies, the results are less diverse and, thus, more reassuring. All measures but one behave relatively similarly, which is mostly interesting because it suggests that (i) the differences between the adjusted frequencies are less likely to yield very different results, but also that (ii) the computationally very intensive distance-based measures that have been proposed (see in particular Savický and Hlaváčová 2002 as well as Washtell 2007) do not appear to lead to fundamentally different results; given that these measures computing time can be 10 times as long or much much longer for large corpora, this suggests that the simpler-to-compute 'classics' might do the job well enough.

The second case study in this paper involves correlating dispersion measures and adjusted frequencies with response time latencies from several psycholinguistic studies, specifically with (i) data from young and old speakers from Spieler and Balota (1997) and Balota and Spieler (1998), and (ii) data from Baayen (2008). All dispersion measures and adjusted frequencies were centered and then correlated with these reactions times (using Kendall's τ, see Chap. 17). For the Balota/Spieler data, the results indicate that some measures score best (including, for instance, *AF*, *U*, and *DP*), but that most measures' correlations with the reaction times are very similar. However, for the reaction times of Baayen (2008), a very different picture emerges: While *DP* scores very well, only surpassed by *vc*, there is a distinct cline such that some measures really exhibit only very low and/or insignificant correlations with the psycholinguistic comparison data.

Gries concludes with some recommendations: Many dispersion measures are relatively similar, but if one is uncertain what measure to trust, it would be useful to compute measures that his cluster/principal component analyses considered relatively different to get a better picture of the diversity in the data; at present and until more data have been studied, it seems as if the computationally more demanding measures may not be worth the effort. Trivially, more analyses (than Lyne's really small study) are needed, in particular of larger data sets and, along the lines of what Biber et al. (2016) did, of data sets with known distributional characteristics.

5.3 Critical Assessment and Future Directions

The previous sections already touched upon some recommendations for future work. It has hopefully become clear that dispersion is as important an issue as it is still neglected or even completely ignored. While every corpus linguist with only the slightest bit of statistical knowledge knows to never present a mean or median without a measure of dispersion, the exact same advice is hardly ever heeded when it comes to frequencies and dispersions in corpus data: There are really only very few studies that report frequency data and dispersion or, just as importantly, report frequencies and association measures and dispersion, although Gries (2008) has shown that the computation of association measures is just as much at risk as frequencies when dispersion information is not also considered. Thus, the first desideratum is that more research takes the threat of underdispersion/clumpiness much more seriously; strictly speaking, reviewers should *always* request dispersion information so that readers can more reliably infer what reported frequencies or association measures really represent or whether they represent what they purport to represent.

Second, we need more studies of the type discussed in the representative studies boxes so that we better understand the different measures' behavior in actual but also controlled/designed data. One issue, for instance, has to do with how corpora are divided into how many parts and how this affects dispersion measures (see for example Biber et al.'s 2016 discussion of the role of the denominator in Juilland's *D*, which features the number of corpus parts). Another is how dispersion measures relate to issues outside of corpus linguistics such as, again, psycholinguistically- or cognitively-informed approaches. This is particularly relevant for measures that are advertised as having certain characteristics. To discuss just one example, Kromer (2003:179) promotes his adjusted frequency measure by pointing to its interdisciplinary/psycholinguistic utility/validity:

> From our point of view, all usage measures considered above have one common disadvantage: their introduction and application are not based psycholinguistically. A usage measure, free from the disadvantage mentioned, is offered below.

However, the advantage is just asserted, not demonstrated, and in Gries (2010) at least, the only study I am aware of testing Kromer's measure, his measure scored worse than most others when explicitly compared to psycholinguistic reference data. While that does of course not mean Kromer's measures has been debunked, it shows what is needed: more and explicit validation.

That being said, a certain frequent trend in corpus linguistic research should be resisted and this is best explained with a very short excursus on association measures (see Chap. 7), where the issue at hand has been recognized earlier than it has in the little existing dispersion research. For several decades now, corpus linguists have discussed dozens of association measures that are used to rank-order, for instance, collocations by the attraction of their constituent words. Some of these measures are effect sizes in the sense that they do not change if the co-occurrence tables from which they are computed are increased by some factor (e.g., the odds ratio), others are based on significance tests, which means they conflate both sample size/actual

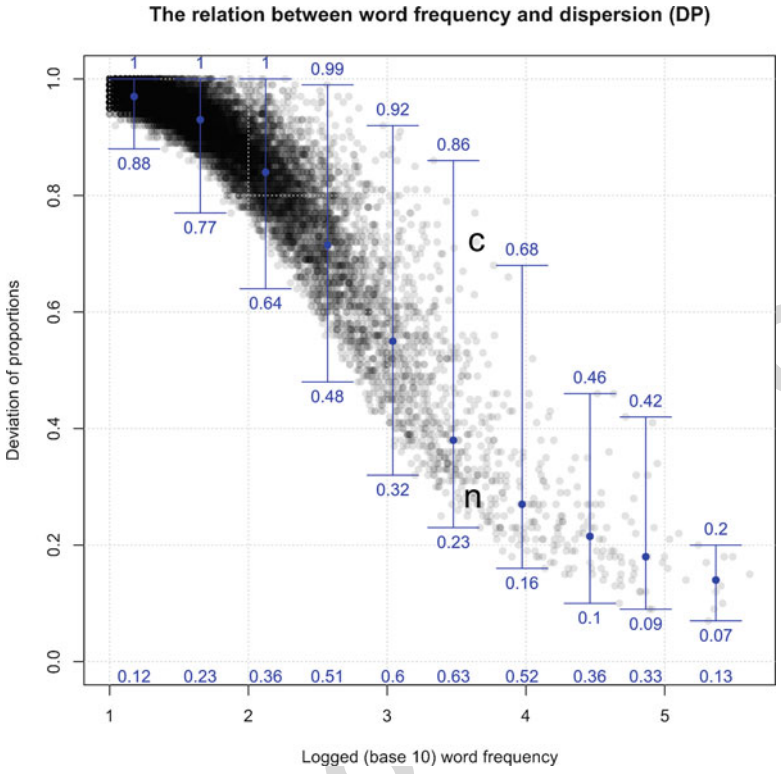

Fig. 5.1 The correlation of frequency and *DP* of words in the spoken BNC

observed frequencies and effect size (e.g., the probably most widely-used measure, the log-likelihood ratio).

This is relevant in the present context of dispersion measures because we are now facing a similar issue in dispersion research, namely when researchers and lexicographers also take two dimensions of information – frequency and the effect size of dispersion – and conflate them into one value such as an adjusted frequency (e.g., by multiplication, see above Juilland's *U*). To say it quite bluntly, this is a mistake because, frequency and dispersion are two different pieces of information, which means conflating them into a single measure loses a lot of information. This is true even though frequency and dispersion are correlated, as is shown in Fig. 5.1 and Fig. 5.2. Both have word frequency on the *x*-axis (logged to the base of 10) and a dispersion measure (*DP* in Fig. 5.1, *range* in Fig. 5.2) on the *y*-axis, and have words represented by grey points. Also, in both plots, the words have been divided into 10 frequency bins, for each of which a blue whisker and the numbers above and below it represent the range of the dispersion values in that frequency bin. For example, in Fig. 5.1, the 6th frequency bin from the left includes words with frequencies

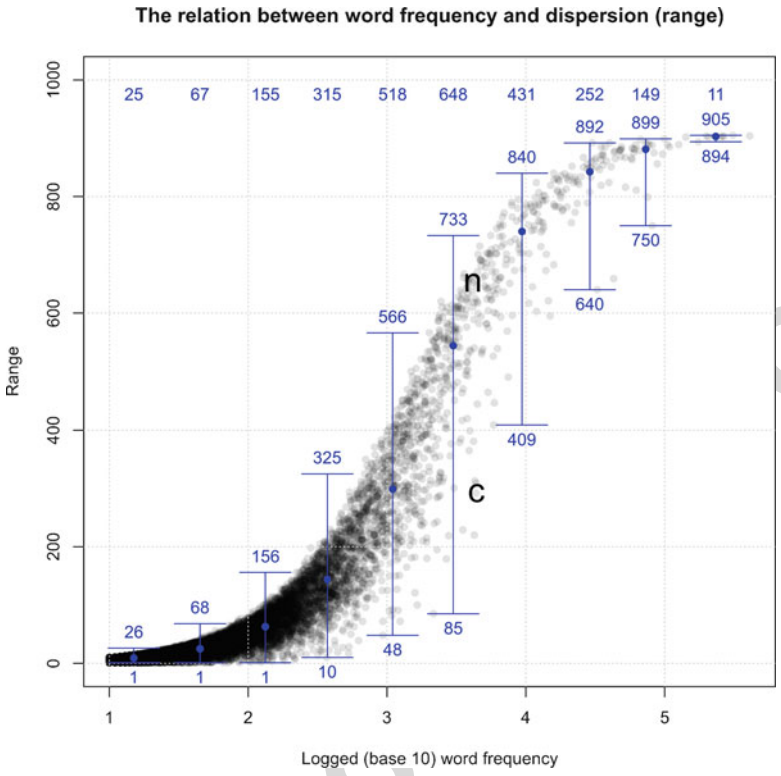

Fig. 5.2 The correlation of frequency and *range* of words in the spoken BNC

between 2036 and 5838 and *DP* values between 0.23 and 0.86, i.e. a *DP*-range of 0.63 also noted in blue at the bottom of the scatterplot.

Obviously, there are the expected correlations between frequency and dispersion (R^2 = 0.832 for logged frequency and *DP*), but just as obviously, especially in the middle range of frequencies – 'normal content words' with frequencies between 1000 and 10,000 – words can have extremely similar frequencies but still extremely different dispersions. This means several things: First, even though there is the above-mentioned overall correlation between frequency and dispersion, this correlation can be very much weakened in certain frequency bins. For example, in the 6th frequency bin, R^2 for the correlation between frequency and dispersion is merely 0.086.

Second, a relatively 'specialized' word like *council* is in the same (6th) frequency bin (*freq* = 4386, *DP* = 0.72, range = 292 out of 905) as intuitively more 'common/widespread' words like *nothing*, *try*, and *whether* (*freqs* = 4159, 4199, 4490; *DP*s = 0.28, 0.28, 0.32; ranges = 652, 664, 671 out of 905); in both plots, the positions of *council* and *nothing* are indicated with the *c* and the *n* respectively plotted into the graph.

Also, even just in the sixth frequency band, the extreme *range* values that are observed are $^{85}/_{905} = 9.4\%$ vs. $^{733}/_{905} = 81\%$ of the corpus files, i.e. huge differences between words that in a less careful study that ignores dispersion would simply be considered 'similar in frequency'.

Finally, these graphs also show that forcing frequency and dispersion into one value, e.g. an adjusted frequency, would lose a huge amount of information. This is obvious from the visual scatter in both plots, but also just from simple math: If a researcher reports an adjusted frequency of 35 for a word, one does not know whether that word occurs 35 perfectly evenly distributed times in the corpus (i.e., frequency = 35 and, say, Juilland's $D = 1$) or whether it occurs 350 very unevenly distributed times in the corpus (i.e., frequency = 350 and, say, Juilland's $D = 0.1$). And while this example is of course hypothetical, it is not as unrealistic as one might think. For instance, the products of observed frequency and 1-*DP* for the two words *pull* and *chairman* in the spoken BNC are very similar – 375 and 368.41 respectively – but they result from very different frequencies and *DP*-values: 750 and 0.5 for *pull* but 1939 and 0.81 for *chairman*. Not only is it the dispersion value, not the frequency one, that reflects our intuition (that *pull* is more basic/widely-used than *chairman*) much better, but this also shows that we would probably not want to treat those two cases as 'the same' as we would if we simply computed and reported some conflated adjusted frequency. Thus, keeping frequency and dispersion separate allows researchers to preserve important information and it is therefore important that we do not give in to the temptation of 'a single rank-ordering scale' and simplify beyond necessity/merit – what is needed is more awareness and sophistication of how words are distributed in corpora, not blunting our research tools.

In all fairness, even if one decides to keep the two dimensions separate, as one definitely should, there still is an additional unresolved question, namely what kind of threshold value(s) to choose for (frequency and) dispersion. It is unfortunately not clear, for instance, what dispersion threshold to adopt to classify a word as 'evenly dispersed enough for it to be included in a dictionary': $DP = 0.4/D = 0.8$? $DP = 0.45/D = 0.85$? In the absence of more rigorous comparisons of dispersion measures to other kinds of reference data, at this point any cut-off point is arbitrary (see Oakes and Farrow 2007:92 for an explicit admission of this fact). Future research will hopefully both explore which dispersion measures are best suited for which purpose and how their relation to frequency is best captured. In order to facilitate this necessary line of research, an R function computing dispersion measures and adjusted frequencies is provided at the companion website of this chapter, see Sect. 5.4; hopefully, this will inspire more research on this fundamental distributional feature of linguistic elements and its impact on other corpus statistics such as association measures, key (key) words, and others.

5.4 Tools and Resources

Dispersion is a corpus statistic that has not been implemented widely into existing corpus tools and arguably it is in fact a statistic that, unlike others, is less obvious to implement, which is why all implementations of dispersion in such general-purpose tools probably leave something to be desired. This is for two main reasons. First, most tools offer only a very small number of measures, if any, and no ways to implement new ones or tweak existing ones. Second, most existing dispersion measures require a division of the corpus into parts and the decision of how to do this is not trivial. While ready-made corpus tools such as WordSmith Tools or AntConc might assume for the user that the corpus parts to be used are the *n* (a user-defined number) equally-sized parts a corpus can be divided into or the separate files of the corpus, this may actually not be what is required for a certain study if, for instance, sub-divisions in files are to be considered as well (as might be useful for some files in the BNC) or when groupings of files into (sub-)registers are what is of interest.

To mention a few concrete examples, WordSmith Tools offers a dispersion plot as well as range and Juilland's *D*-values (without explicitly stating that that is in fact the statistic that is provided) while AntConc offers a version of a dispersion plot separately for each file of a corpus, which is often not what one needs. The COCA-associated website https://www.wordfrequency.info/ (accessed 22 May 2019) provides data that went into Davies and Gardner (2010), which means they provide Juilland's *D* for the corpus when split up into 100 equally-sized parts. As is obvious, the range of features is extremely limited and virtually non-customizable.

By far the best – in the sense of most versatile and powerful – approach to exploring issues of dispersion is with programming languages such as R or Python (see Chap. 9), because then the user is not dependent on measures and settings enshrined in ready-made software but can customize an analysis in exactly the way that is needed, develop their own methods, and/or run such analysis on data/annotation formats that none of the above tools can handle. This chapter comes with some companion code for readers to explore as well as an R function to compute a large number of dispersion measures for data provided by a user. This function is an update of the function provided in Gries (2008), which adds the KL-divergence as a dispersion measure, updates the computation of some measures, cleans up the code, and drastically speeds up all computations; see the companion website for how to use it.

Further Reading

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

Burch et al. (2017) is a study that introduces another dispersion measure D_A (or *MDA* in Wilcox's 1973 terminology) and compares it to the historically most

widely-used dispersion measure of Juilland's *D* and to the recently-proposed measure of Gries's *DP*. They define D_A and test its performance by, for instance, a simulation study of three different scenarios by creating randomly sampled corpora and comparing the three different dispersion statistics. Also, they correlate the dispersion statistics for 150 words taken from the British National Corpus using scatterplots and pairwise differences of dispersion statistics. It is worth pointing out, as the authors also do, that (i) this study is based on the overall probably less realistic scenario that all corpus parts are equally large, which is not that likely when corpus parts are considered to be files (e.g., in the BNC) or (sub-)registers (e.g. in the ICE-GB) and that (ii) computing D_A can take literally thousands more time than *D* or *DP* even though its non-linear correlation R^2 with *DP* exceeds 0.99. That being said, their study is nonetheless a good example of exactly the kind of study we need more of to further our understanding of (i) how different dispersion measures react to corpus-linguistic data and (ii) how they react to certain kinds of potentially extreme input data.

Savický, P., and Hlaváčová, J. 2002. Measures of word commonness. *Journal of Quantitative Linguistics* 9(3):15–31.

Savický and Hlaváčová (2002) is another interesting reading. Their study starts out from the question of how to identify "common" words to be included in a universal dictionary. However, they propose to approach dispersion in ways that do not require a division of a corpus in parts – rather, the corpus is treated as a single sequence or vector of words and then dispersion is used to compute corrected frequencies that are close to the actual observed frequencies when a word is very evenly distributed and (much) small when it is not. They propose three different corrected frequencies – one based on Average Reduced Frequency (f_{ARF}), one based on Average Waiting Time (f_{AWT}), and one based on Average Logarithmic Distance (f_{ALD}) – and proceed to apply them to data from the Czech National Corpus to test the measures' stability (how much do they vary when applied to different parts of the overall corpus?) and to exemplify the kinds of words that the measures return as highly unevenly distributed. While these dispersion measures can take much longer to compute than the parts-based measures reported on above and adjusted frequencies are problematic for the reasons discussed above, this paper is nonetheless noteworthy and interesting for the novel, non-parts-based approach to dispersion.

References

Adelman, J. S., Brown, G. D. A., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science, 19*(9), 814–823.

Ambridge, B., Theakston, A. L., Lieven, E. V. M., & Tomasello, M. (2006). The distributed learning effect for children's acquisition of an abstract syntactic construction. *Cognitive Development, 21*(2), 174–193.

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to R*. Cambridge: Cambridge University Press.

Baayen, R. H. (2010). Demythologizing the word frequency effect: A discriminative learning perspective. *The Mental Lexicon, 5*(3), 436–461.

Balota, D. A., & Spieler, D. H. (1998). The utility of item level analyses in model evaluation: A response to Seidenberg and Plaut. *Psychological Science, 9*(3), 238–240.

Baron, A., Rayson, P., & Archer, D. (2009). Word frequency and keyword statistics in historical corpus linguistics. *Anglistik: International Journal of English Studies, 20*(1), 41–67.

Biber, D., Reppen, R., Schnur, E., & Ghanem, R. (2016). On the (non)utility of Juilland's D to measure lexical dispersion in large corpora. *International Journal of Corpus Linguistics, 21*(4), 439–464.

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

Carroll, J. 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.

Church, K. W., & Gale, W. A. (1995). Poisson mixtures. *Journal of Natural Language Engineering, 1*(2), 163–190.

Davies, M., & Gardner, D. (2010). *A frequency dictionary of contemporary American English: Word sketches, collocates and thematic lists*. London/New York: Routledge, Taylor and Francis.

Ellis, N. C. (2002a). Frequency effects in language processing and acquisition: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition, 24*(2), 143–188.

Ellis, N. C. (2002b). Reflections on frequency effects in language acquisition: A response to commentaries. *Studies in Second Language Acquisition, 24*(2), 297–339.

Ellis, N. C. (2011). Language acquisition as rational contingency learning. *Applied Linguistics, 27*(1), 1–24.

Gardner, D., & Davies, M. (2014). A new academic vocabulary list. *Applied Linguistics, 35*(3), 305–327.

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

Gries, S. T. (2010). Dispersions and adjusted frequencies in corpora: Further explorations. In S. T. Gries, S. Wulff, & M. Davies (Eds.), *Corpus linguistic applications: Current studies, new directions* (pp. 197–212). Amsterdam: Rodopi.

Gries, S. T. (2013). *Statistics for linguistics with R* (2nd rev. and ext. ed, 359). Berlin/Boston: De Gruyter Mouton.

Juilland, A. G., & Chang-Rodriguez, E. (1964). *Frequency dictionary of Spanish words*. The Hague: Mouton de Gruyter.

Juilland, A. G., Brodin, D. R., & Davidovitch, C. (1970). *Frequency dictionary of French words*. The Hague: Mouton de Gruyter.

Kromer, V. (2003). An usage measure based on psychophysical relations. *Journal of Quantitative Linguistics, 10*(2), 177–186.

Leech, G. N., Rayson, P., & Wilson, A. (2001). *Word frequencies in written and spoken English: Based on the British National Corpus*. London: Longman.

Lijffijt, J., & Gries, S. T. (2012). Correction to "Dispersions and adjusted frequencies in corpora". *International Journal of Corpus Linguistics, 17*(1), 147–149.

Lyne, A. A. (1985). Dispersion. In *The vocabulary of French business correspondence* (pp. 101–124). Geneva/Paris: Slatkine-Champion.

Lyne, A. A. (1986). In praise of Juilland's D. In *Méthodes quantitatives et informatiques dans l'Études des textes, vol. 2* (pp. 589–595). Geneva/Paris: Slatkine-Champion.

Oakes, M., & Farrow, M. (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.

Paquot, M. (2010). *Academic vocabulary in learner writing: From extraction to analysis*. London/New York: Continuum.

Robertson, S. (2004). Understanding inverse document frequency: On theoretical arguments of IDF. *Journal of Documentation, 60*(5), 503–520.

Rosengren, I. (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, P., & Hlaváčová, J. (2002). Measures of word commonness. *Journal of Quantitative Linguistics, 9*(3), 15–31.

Schmid, H. J. (2010). Entrenchment, salience, and basic levels. In D. Geeraerts & H. Cuyckens (Eds.), *The Oxford handbook of cognitive linguistics* (pp. 117–138). Oxford: Oxford University Press.

Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in information retrieval. *Journal of Documentation, 28*(1), 11–21.

Spieler, D. H., & Balota, D. A. (1997). Bringing computational models of word naming down to the item level. *Psychological Science, 8*(6), 411–416.

Washtell, J. (2007). *Co-dispersion by nearest-neighbour: Adapting a spatial statistic for the development of domain-independent language tools and metrics*. Unpublished, M.Sc. thesis, School of Computing, Leeds University.

Wilcox, A. R. (1973). Indices of qualitative variation and political measurement. *The Western Political Quarterly, 26*(2), 325–343.

Zhang, H., Huang, C., & Yu, S. (2004). *Distributional consistency: As a general method for defining a core lexicon*. Paper presented at language resources and evaluation 2004, Lisbon, Portugal.