Advancing our understanding of dispersion measures in corpus research

Lukas Sönning (University of Bamberg)

lukas.soenning@uni-bamberg.de

Abstract. This paper offers a survey of recent corpus-based work, which shows that dispersion is typically measured across the text files in a corpus. Systematic insights into the behavior of measures in such distributional settings are currently lacking, however. After a thorough discussion of six prominent indices, we investigate their behavior on relevant frequency distributions, which are designed to mimic actual corpus data. Our evaluation considers different distributional settings, i.e. various combinations of frequency and dispersion values. The primary focus is on the response of measures to relatively high and low sub-frequencies, i.e. texts in which the item or structure of interest is over- or underrepresented (if not absent). We develop a simple method for constructing sensitivity profiles, which allow us to draw instructive comparisons among measures. We observe that these profiles vary considerably across distributional settings. While D, D_A and D_P appear to show the most balanced response contours, our findings suggest that much work remains to be done to understand the performance of measures on items with normalized frequencies below 100 per million words.

1. Introduction

Corpus linguists have at their disposal an ever-growing variety of dispersion measures. With the quantification of dispersion being an active area of methodological research, however, there is as yet relatively little guidance for the choice among them. The aim of the present paper is to contribute to our ongoing efforts to understand dispersion and its measurement, and to support an informed and constructive discourse on this area of corpus-linguistic methodology.

It deserves to be stressed at the outset that the term dispersion has acquired two (partly overlapping) corpus-linguistic senses.[1] First, it may refer to the *pervasiveness* of an item (or structure) in a corpus, i.e. how *widely* it is used. Upon dividing a corpus into units, then, the focus is on whether or not an item occurs in a certain (stretch of) text, without taking into account the number of instances. In its second sense, the word dispersion denotes *evenness* of distribution, i.e. whether the item's occurrence rate is balanced across contexts of language use. This means that in contrast to pervasiveness, which is only concerned with the presence/absence of an item, evenness of distribution compares sub-frequencies across units. It follows that dispersion measures fall into two groups, depending on whether they quantify pervasiveness or evenness of distribution.

The present paper will be (primarily) concerned with evenness measures, where indices usually range from 0 to 1, with low values reflecting a clumpy distribution (occurrence in only few texts or genres) and high values denoting an even distribution. The purpose of the paper is twofold. Our first goal is to survey the corpus-linguistic landscape with regard to the measurement of dispersion. We start with a historical perspective on evolving practices and trace the origin of some prominent measures across time and disciplines. Our focus will be on the recent wave of corpus-linguistic interest in dispersion, which has gained momentum after an influential paper by Gries (2008). To chart current research practices,

[1] I would like to thank Jesse Egbert for drawing my attention to the importance of this distinction.

we examine the facilities provided by popular software packages for corpus analysis. We then conduct a literature survey, to document how corpus linguists typically go about measuring dispersion – which kind of corpus parts they compare, what measures they use, and what kind of structures they study.

This stocktaking exercise sets the scene for the second goal if this paper, which is to advance our understanding of dispersion measures and promote an informed and wider use by researchers. In a first step, we aspire to make transparent the logic underlying a number of relatively complex evenness measures, and to draw attention to differences among them, most notably their sensitivity to certain distributional features. Following this, we field-test indices in data situations that have so far been neglected in the methodological literature. To sketch analysis settings that are likely to be relevant (and therefore informative) for applied work, we draw on the results of our literature survey, which finds that corpus linguists typically measure dispersion across the texts constituting a corpus. We therefore complement existing practices in the methodological literature by monitoring the performance of indices in these distributional situations. To obtain realistic and representative scenarios, our test settings are modeled on actual corpus data. Our analyses reveal a number of insights that have so far not been noted and which may serve to direct future work on the comparative evaluation of dispersion measures.

The two goals of our paper translate into the following outline. Section 2 looks at the history of dispersion in the analysis of textual data and takes stock of software facilities and current research practices. Section 3 then breaks down dispersion measures into their elementary components to make clear their logic and (biased) behavior. Section 4 shows how we can use corpus data to construct data settings for the comparative evaluation of indices. In Section 5, we analyze the behavior of dispersion measures in these data settings and Section 6 closes with a summary and discussion.

2. The measurement of dispersion

Lexical dispersion has been of interest to scholars for more than a century. This section starts with a brief historical outline of its measurement and use in linguistics (Section 2.1), which allows us to contextualize some methodological issues that have recently been raised in corpus linguistics. We pay particular attention to the question of which corpus parts should form the basis of analysis (Section 2.2), and close in Section 2.3 with a survey of current practices in corpus-based work.

2.1. A brief history

Early applications of dispersion measures may be found in the context of language pedagogy, where they were used in the compilation of general vocabulary lists. To identify common words in Spanish, for instance, Keniston (1920: 86) used as a database 'a few score' of texts representing various genres (e.g. plays, novels, newspapers) and then looked at the percentage of texts containing at least one occurrence of a given item. He then formed 8 classes of words; items in class 1, for example, occurred in at least 80% of the texts. We will refer to this measure as *text dispersion* (*TD*) and express it as a proportion (instead of a percentage). The threshold for inclusion in Keniston's (1920) study was .33 (or 33%), which yielded a list of 1,322 words. At much the same time, Thorndike (1921a: iii–iv) compiled a

list of 10,000 English words, ranked by 'importance' or 'commonness'. The ranking was based on 'credit-numbers', which combine information on frequency and dispersion across 41 texts (collections or excerpts) covering different genres such as children's books, literary classics, and newspaper texts (see Thorndike 1921b: 335–340). Vander Beke (1932) carried out similar work on French and went as far as ranking words primarily according to dispersion. His word book lists 6,067 items based on *range* (*R*), i.e. the number of texts (88 in total, drawn from five different genres) in which they appeared. At the outset, then, practitioners were exclusively concerned with the pervasiveness of items across texts.

The 1960s and 1970s saw a period of increased methodological interest in dispersion. This lexicographic wave of research led to the introduction of several new measures (*D*, Juilland & Chang-Rodríguez 1964; D_2, Carroll 1970; *S*, Rosengren 1971). Thus, Juilland & Chang-Rodríguez (1964) were unsatisfied with the lack of sensitivity of *TD* and *R* and essentially argued for pervasiveness to be replaced by evenness as the feature of main interest. They proposed a more nuanced measure, *D*, which captures the evenness of sub-frequencies and is standardized to the unit interval [0;1], with 1 indicating a perfectly even distribution. Their study compared 5 macro-genres of 100,000 words each (drama, fiction, academic writing, news, and essayistic literature).[2]

The early 21st century has seen a renaissance of work on the measurement of dispersion, with an influential paper by Gries (2008) putting it on the corpus-linguistic research agenda. The ongoing discourse between methodology and application has brought about yet further indices (*DP* and DP_{norm}, Gries 2008; D_A, Burch et al. 2017; D_{KL} Gries 2020, 2021), as well as an application of such measures to domains beyond lexis (see Section 2.3). A theme that has received some attention recently is the unit of analysis, i.e. across which units, or stretches of text, occurrence rates should be compared. We will deal with this question in more detail in the next section.

Before we go further, however, let us jot down the most important measures that have emerged over the past 100 years. Figure 1 shows these in chronological order, along with the references indicating where (to our knowledge) they first appeared in the linguistic literature and the distributional feature they measure (pervasiveness vs. evenness). We will deal with the individual measures in more detail in Section 3.

[2] It is of interest to note that the material representing these 'lexical worlds' was based on an elaborate sampling design, which involved the random selection of texts and the systematic sampling of sentences at roughly equal-spaced intervals from these texts (see Juilland & Chang-Rodríguez 1964: XIV–XXVII).

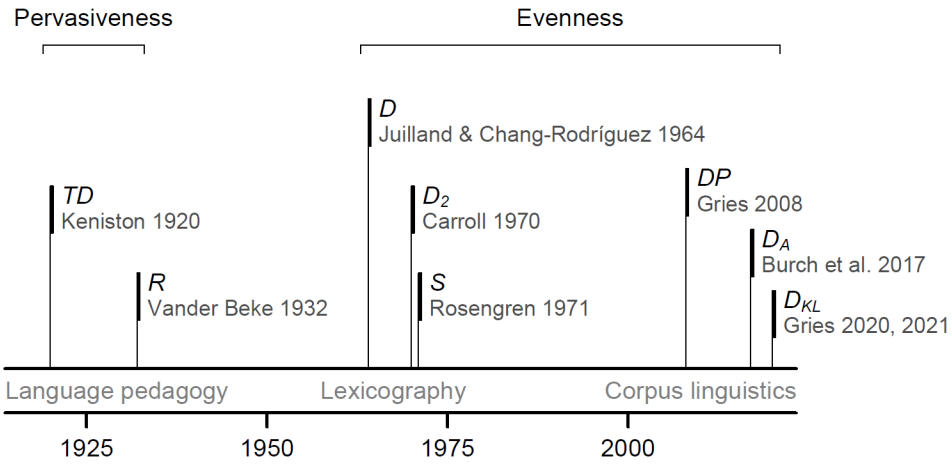

Figure 1. The emergence of dispersion measures across time and disciplines. ⓒⓘ[3]

2.2. The unit of analysis

It has been argued in recent methodological work that linguistically meaningful units should form the basis of dispersion analysis (e.g. Burch et al. 2017; Egbert et al. 2020). Thus, Egbert et al. (2020: 91–92) state that the use of equal-sized corpus parts constitutes 'the traditional approach to measuring lexical dispersion'; 'in practice, dispersion analyses are […] based on arbitrary corpus parts'. Support for this assessment is given by listing 8 studies (published between 2001 and 2016), which pursue pedagogical or lexicographic goals, similar to the 20th-century work referenced in the previous section.

In light of our historical outline, however, it would seem that the label 'traditional' – if taken literally – may require qualification. After all, early work on lexical dispersion did rely on meaningful units of analysis. Keniston (1920), Thorndike (1921), and Vander Beke (1932) measured dispersion across texts, and the lexicographic wave in the second half of the 20th century also drew on meaningful units: Juilland & Chang-Rodríguez (1964: XVI) compared macro-genres; Carroll (1970: 63, 65) applied measures to the 15 subgenres in Brown and pondered over the possibility of using smaller units (e.g. authors or author-year combinations); and Rosengren (1971: 120) obtained tallies for five newspaper sections (editorials, politics, arts, business, and miscellaneous). A notable exception is the study by Lyne (1985), where a corpus of business correspondence letters, i.e. a relatively uniform collection of texts, was divided into parts of equal size.

It would therefore appear that the use of equal-sized parts may have largely originated in the recent wave of corpus-linguistic work on dispersion – despite the fact that key methodological contributions continue(d) to rely on meaningful units (e.g. Gries 2008, 2010). In search of an explanation, Egbert et al. (2020) note three possible reasons: (i) the computational simplicity associated with using equal-sized parts; (ii) the false belief that partitions must have the same size; and (iii) the assumption that the choice of unit may be inconsequential.

[3] Images with the symbols ⓒⓘ in the figure caption have been published under the Creative Commons Attribution 4.0 license (CC BY 4.0, http://creativecommons.org/licenses/by/4.0) in the accompanying OSF project (https://osf.io/nhw2y). All figures were drawn using the R package 'lattice' (Sarkar 2008).

A further factor related to ease of implementation may be found in the facilities provided by existing corpus analysis tools. Let us therefore examine a number of widely used software packages with regard to the dispersion measure(s) they offer, and the units of analysis that can be formed. Our findings are summarized in Table 1, where resources are listed in alphabetical order:

- *AntConc* (Anthony 2022) calculates dispersion across the corpus files imported into the software, as well as *within* these files based on a user-specified number of equal-sized parts (default: 10)
- *CQPweb* (Hardie 2012) currently only offers experimental support, and it is unclear what units are being formed
- *LancsBox X* (Brezina and Platt 2023) tabulates dispersion scores across the corpus files that are loaded into the software
- *Sketch Engine* (Kilgarriff et al. 2014) only offers a dispersion-adjusted frequency measure (the average reduced frequency; Savický and Hlaváčová 2002), which is based on corpus parts of equal size
- *WMatrix* (Rayson 2009) calculates dispersion across the file folders created using the web interface[4]
- *WordSmith Tools* (Scott 2015) divides the corpus into a user-specified number of equal-sized parts (default: 8)

The summary in Table 1 shows considerable overlap with regard to the measures that are currently implemented (*R*, *TD*, D_P, *D*). In terms of unit of analysis, however, the picture is mixed. Two packages that stand out are *AntConc* and *LancsBox X*, which provide ample support for custom units of analysis, provided these are represented by the imported corpus files.

Table 1. Dispersion analysis options implemented in corpus analysis software

						Unit of analysis	
Software	Version	Measures				Parts	Files
AntConc	4.2	*R*	*TD*	*D*	*DP*	●[a]	●
CQPweb	3.3.17	*R*		*D*	*DP*	?	?
LancsBox X	3	*R*	*TD*	*D*	*DP*		●
SketchEngine	n/a	Average reduced frequency				●	
WMatrix	5	*R*			*DP*		●[b]
WordSmith Tools	6	*R*	*TD*	*D*		●	

Notes. [a] Parts-based measurements are (quite sensibly) only made within corpus files; [b] Feasible only for a relatively small number of units, since files must be uploaded individually

Seeing that some software packages, at least in their latest versions, allow for measurements to be made at the text or genre level, we decided to take another look at existing practices by corpus linguists.

2.3. A survey of corpus-based work

[4] The current version seems to require corpus files to be uploaded individually, which would make it cumbersome to obtain dispersion across individual text files.

To gain an overview of the measurement of dispersion in recent corpus-based work, we examined 730 articles published between 2008 and 2022 across four corpus-linguistic journals (see Sönning 2024 and Web Appendix 1 for details[5]). The search term 'dispersion' narrowed this list down to 121 documents, and 38 of these assessed dispersion in one of the corpus-linguistic senses described in Section 1. We concentrate on studies that feature numerical summaries, which left us with 35 research articles.[6] These were annotated for the following variables:

- The dispersion measure(s) used
- The unit of analysis (e.g. texts, genres, corpus parts)
- The linguistic structure studied (e.g. lexical, phraseological)

Table 2 sets out the findings of our review. Apart from the overall distribution of the three study features, it provides a diachronic perspective on these by subdividing counts into three (5-year) periods. Before we take a closer look at these results, let us briefly consider the overall rate of dispersion reports in the literature, which amounts to 5% of the studies (i.e. 38 out of 730). This relatively low percentage seems to support Gries' (2008; 2020) continuing concerns about the underuse of dispersion in applied work.

Looking at the findings reported in Table 2, the following points are noteworthy:

- While most of the indices listed in Figure 1 appear at least once in our survey, the most popular measures are *R*, *TD*, and D_P. We therefore observe considerable overlap with the options implemented in corpus-analysis software (cf. Table 1). Note that the variety found in period I is primarily due to Gries (2008), who studied the behavior of D_P, *D*, D_2, and *S*.
- The vast majority of studies rely on texts as the unit of analysis; only 11% use arbitrary corpus parts. This suggests that Egbert et al.'s (2020) diagnosis does not extend to the work reported in corpus-linguistic journals. It is also of interest to note that the preference for meaningful units of analysis is stable over time.
- We observe that just over half of the studies are concerned with lexical items. Apparently, the domains of application have diversified over time.

If we accept Table 2 as giving a rough reflection of existing practices in corpus-based research, we can summarize by saying that corpus linguists primarily rely on *R*, D_P, and *TD* and typically apply these measures to meaningful units of analysis.

[5] https://osf.io/nhw2y

[6] We excluded two studies that only relied on a visual inspection of dispersion plots, and one study that applied a measure of dispersion that looks at the median number of words between consecutive occurrences of an item in the corpus. This index is similar to Savický and Hlaváčová's (2002: 220) Average Waiting Time.

Table 2. Dispersion analysis in corpus-based work: Survey of articles in corpus-linguistic journals

			Five-year periods[a]		
Study feature	N	(%)	I	II	III
Dispersion measure[b]					
R	14	(34%)	3	4	7
D_P	11	(27%)	1	4	6
TD	9	(22%)	1	2	6
D	3	(7%)	1	1	1
D_2	2	(5%)	2		
S	1	(2%)	1		
D_A	1	(2%)			1
Unit of analysis					
Text	27	(77%)	5	7	15
Equal-sized corpus part	4	(11%)		2	2
Genre	3	(9%)	1	1	1
Chapter	1	(3%)		1	
Structure[c]					
Lexical	21	(58%)	5	6	10
Phraseological	12	(33%)	1	4	7
Syntactic	3	(8%)	1	1	1
Number of studies	**35**		**6**	**11**	**18**

Notes.

[a] Periods: I (2008–2012), II (2013–2017), III (2018–2022)

[b] Counts do not sum to 35 since three studies report several measures

[c] Counts do not sum to 35 since one study analyzed different structures

3. The nuts and bolts of dispersion measures

In Section 2.1, we touched upon a variety of dispersion indices; let us now consider them in more detail. Our form of exposition deviates from that used in other texts, which often rely quite heavily on (sometimes disconnected) mathematical notation. This style of presentation, despite its explicitness and economy, may not be conducive to an informed use by practitioners. We therefore build on and extend earlier work in two ways. For one, we use unified notation, which streamlines the discussion and reveals parallels between computational procedures. Further, we quite literally look at the inner workings of formulas using, as far as practicable, graphical illustrations; this may allow a wider readership to form a more intuitive understanding of these indices. A constant focus will be on the contribution of individual units (e.g. text files) to the obtained scores, and how measures respond to extreme sub-frequencies, i.e. texts with relatively many or few occurrences of the item in question. Once we have introduced our illustrative data set (Section 3.1), we will discuss the individual measures in chronological order (cf. Figure 1). R code for the following computations and visualizations can be found in the OSF project associated with this article (https://osf.io/nhw2y/).

3.1. Illustrative data

Our illustrative data, which are modeled on the distribution of *which* in the Brown Corpus (Francis and Kučera 1964), contain k = 10 texts of 1,000 words each. In this mini-corpus, we obtain the following text-specific counts for *which*: 0, 1, 2, 4, 4, 5, 8, 8, 11, 17. Since our texts are 1,000 words long, these are at the same time normalized frequencies, i.e. occurrences per thousand words (ptw). We will refer to these as *occurrence rates*. They appear in Figure 2, where each text is represented by a dot.

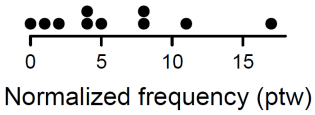

Figure 2. Illustrative set of text-specific occurrence rates.

Next, we define a set of variables and use the subscript i to refer to a specific text. Since each text has 1,000 words, the *word count* (W_i) and the *word share* (w_i) are constant across texts (1,000 words; .10 or 10% of the corpus). For each text, we also have a *token count* (T_i), the number of occurrences of the item. Based on these token counts, which range from 0 to 17, we can determine the *token share* for each text (t_i); it ranges from 0 (0 out of 60) to .28 (17 out of 60). Finally, R_i will denote the occurrence rate of an item in text i. We kindly ask the reader to notice (and accept) upper-case R's double duty in the present paper, as it is also used as a shorthand for *range*. This gives us the following quantities:

- W_i –word count for text i
- T_i – token count for text i
- w_i – text i's share of the total number of words (word share)
- t_i – text i's share of the total number of tokens (token share)
- R_i – occurrence rate for text i

Note how w_i and t_i are the proportional variants of their uppercase counterparts (W_i and T_i). We can define a similar quantity based on the occurrence rates R_i, by first summing over all rates and then dividing each rate by this sum. We refer to this proportional quantity (r_i) as a *rate-based proportion*:

- r_i – text i's share of the sum of all rates (a rate-based proportion).

Table 3 lists these quantities for our hypothetical corpus. Note that, since texts have the same length, t_i is proportional to r_i.

Table 3. Quantities for the calculation of dispersion measures

Variable	Label	1	2	3	4	5	6	7	8	9	10
Word count	W_i	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
Word share	w_i	.10	.10	.10	.10	.10	.10	.10	.10	.10	.10
Token count	T_i	0	1	2	4	4	5	8	8	11	17
Token share	t_i	.00	.02	.03	.07	.07	.08	.13	.13	.18	.28
Rate (ptw)	R_i	0	1	2	4	4	5	8	8	11	17
Rate-based proportion	r_i	.00	.02	.03	.07	.07	.08	.13	.13	.18	.28

3.2. TD and R

The calculation of *TD* and *range* (*R*) does not require much explanation: Since the item appears in 9 of the 10 texts, *R* is 9 and *TD* is .90. Even though *R* and *TD* give essentially the same information, we would prefer to list them as separate measures for two reasons. First, it is difficult to interpret and compare *R* scores without knowledge of (or reference to) k, the total number of units. The proportional nature of *TD*, in contrast, yields a standardized score that can in principle be compared across corpora (and studies), and which is on the same scale as most other dispersion indices (i.e. the unit interval [0;1]). Further, *TD* clarifies the unit of analysis, i.e. that dispersion is tabulated across texts. Some scholars prefer to treat *R* and *TD* as one and the same measure, which is also sensible. In any case, the distinction is immaterial for the purposes of the present study, since our focus will be on evenness measures, to which we now turn.

3.3. D

As mentioned above, *D* was proposed by Juilland & Chang-Rodríguez (1964: LIII) as a refinement over *R* and *TD*. It also laid the foundation for what has become the established scaling of dispersion measures, where 0 indicates a maximally uneven and 1 a perfectly balanced distribution. In the following formula, s refers to the (population) standard deviation[7] of the occurrence rates R_i, and m denotes their mean; k is the number of texts (or units):

$$(1)\quad 1 - \frac{s}{m\sqrt{k-1}}$$

Let us take a closer look at (1). Similar to the formulas for most other measures, it contains a fraction. We then pay particular attention to the numerator, since it tells us what determines the magnitude of the measure. The denominator usually only serves to map the score to the unit interval [0;1]. We therefore note that the magnitude of *D* primarily depends on s. Since the fraction is subtracted from 1 (which serves to reverse the scaling, so

[7] The formulas for the population and sample standard deviation (SD) differ. In line with Juilland & Chang-Rodríguez' (1964: LIII), most methodological work on dispersion calculates *D* using the population SD (e.g. Carroll 1970; Lyne 1985, Biber et al. 2016). Gries (2008: 407) uses the sample SD, which explains why he obtains negative values for some distributions (2008: 412). In his later work, Gries (2020: 103) uses the population SD.

that 1 indicates balance), greater values of *s* produce smaller values of *D*: More variable occurrence rates reflect a less even distribution and therefore yield smaller values of *D*.

The standard deviation is calculated based on the squared distance of rates from their mean. This is illustrated in Figure 3, where the area of the individual squares is proportional to the relative contribution of each text to the overall *D* score. The denominator in (1) includes two terms, which have different functions: Dividing by m makes the measure independent of the average occurrence rate, and dividing by $\sqrt{k-1}$ ensures that the measure ranges between 0 and 1.

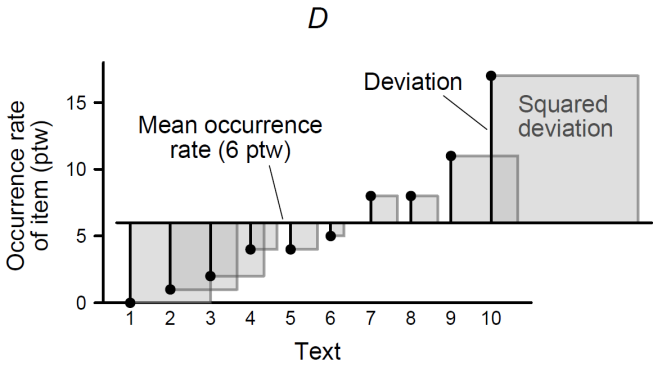

Figure 3. Formula for *D*: The squared deviation of occurrence rates from their mean.

It has been noted that the standard deviation is not an ideal measure of variability for word frequencies, since these tend to have a skewed distribution (e.g. Rosengren 1971: 118). This is because, upon squaring, unusually high occurrence rates will have a considerable effect on the magnitude of *s*. However, since m is likewise sensitive to outliers, their effects tend to cancel out, which means that *D* is not affected by outliers as much. A more serious issue, which has been pointed out by Biber et al. (2016: 443), is the fact that the value of *D* systematically varies with the number of texts *k*: As *k* increases, *D* becomes larger. This behavior, which is due to the term $\sqrt{k-1}$ in the denominator, makes *D* unattractive for dispersion analyses at the text level.[8]

3.4. D_2

Carroll (1970) proposed a measure of dispersion based on the information-theoretic concept of entropy, which describes a distribution of probabilities. High entropy means unpredictability (see Oakes 1998: 58-59): In a situation where there are five possible events, entropy is highest if they are equally likely, since we cannot make a decent prediction as to which event will occur. Applied to dispersion, the predictive task is to forecast where (i.e. in which unit) an item will occur. The number of possible events then corresponds to the number of units (or texts). If an item is evenly dispersed, its location is unpredictable, since

[8] For instance, the COCA-based word frequency data provided on the website www.wordfrequency.info calculates *D* at the text-level. With more than 400,000 texts in the corpus, this produces high dispersion scores even for rare lemmas like *neuroanatomy* (.61) and *coalmine* (.72) (see https://www.wordfrequency.info/samples/lemmas_60k.txt).

(if texts have the same length) it is equally likely to occur in each. High entropy therefore reflects high dispersion.

The formula we will give for D_2 is a slightly rearranged version of the original.[9] Recall that r_i refers to the rate-based proportions (cf. Table 3).

$$(2)\quad \frac{\sum_i^k r_i \log_2 \frac{1}{r_i}}{\log_2 k}$$

Again, the denominator merely ensures that D_2 ranges between 0 and 1. The main action takes place in the numerator, which is the sum of k product terms – one for each text. In a perfectly balanced scenario, each text has the same rate-based proportion: Each r_i is then equal to $\frac{1}{k}$. In Figure 4, all points would then be sitting on the line, which marks $\frac{1}{k}$. In this balanced case, the product term for each text ($r_i \log_2 \frac{1}{r_i}$) reduces to $\log_2 \frac{1}{k}$. The sum of these terms is then equal to the denominator ($\log_2 k$). To appreciate how much a text contributes to D_2, we can compare its product term ($r_i \log_2 \frac{1}{r_i}$) to the balanced analogue ($\log_2 \frac{1}{k}$). The difference reflects the amount by which it deviates from the balanced scenario, and hence its relative contribution to the value of D_2.

In Figure 4, the absolute values of these differences are shown using grey spikes, which are aligned at 0. We observe that D_2 responds differently to text-specific shares above and below the horizontal line. Consider, for instance, texts 5 and 7. Even though they deviate from .10 by the same amount, their relative contribution to D_2 differs: Text 5 affects the value of D_2 more than text 7. For comparison, consider, in Figure 2 above, the size of the grey squares for texts 5 and 7, which are identical – in other words, D treats them even-handedly.

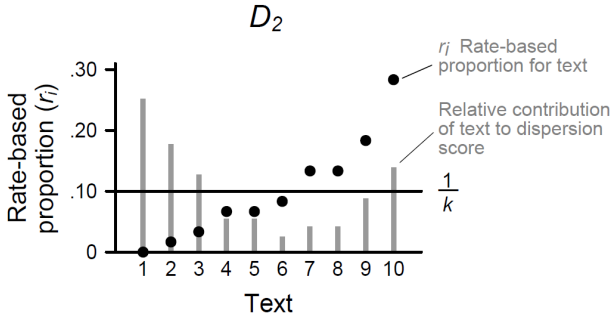

Figure 4. The contribution of each text to the value of D_2. 🅭🅯

[9] The version of D_2 quoted by Gries (2020: 103), using our notation, is $\frac{-\sum_i^k r_i \log_2 r_i}{\log_2 k}$. It is easier to make sense of the formula without the minus sign before the summation in the numerator, and we therefore get rid of it: $\frac{-\sum_i^k r_i \log_2 r_i}{\log_2 k} = \frac{\sum_i^k r_i \log_2 \frac{1}{r_i}}{\log_2 k}$.

This behavior of D_2, which is even more apparent when comparing texts 1 and 10, was pointed out by Lyne (1985: 107, 114), who demonstrated that the measure not only 'penalizes zeros' but also shows a disproportionate sensitivity to 'low sub-frequencies', i.e. text-specific occurrence rates below average. While this behavior of entropy may be useful for information-theoretic purposes, it is not at all clear whether it is desirable for the measurement of dispersion in textual data (see Lyne 1985: 115).

3.5. S

Rosengren (1971) proposed the index *S*; using the shorthand symbols introduced in Table 3, it is calculated using the following formula:

(3) $$\frac{\left(\sum_i^k \sqrt{w_i T_i}\right)^2}{N}$$

Here, *N* is the total number of occurrences of the item (across all texts). Again, our attention centers on the numerator, and we note that each text again contributes one term ($\sqrt{w_i T_i}$) to *S*. In a perfectly balanced setting, in order for the equation to yield a value of 1, the numerator must equal *N*. Working backwards, the sum of the product terms ($\sum_i^k \sqrt{w_i T_i}$) must therefore be equal to the root of *N* (i.e. $\sqrt{N}$), and each term ($\sqrt{w_i T_i}$) must accordingly equal $\sqrt{N}/k$. To see how much each text contributes to *S*, we can compare $\sqrt{w_i T_i}$ to the balanced analogue ($\sqrt{N}/k$) the difference again denoting its relative contribution to *S*.

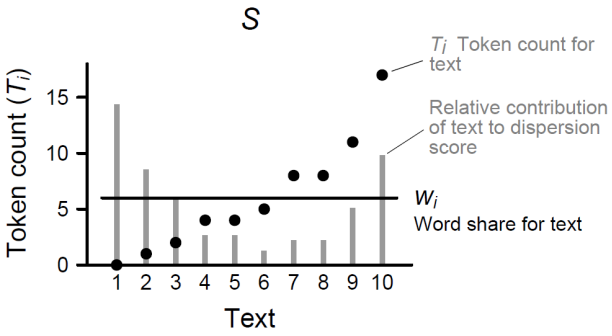

Figure 5. The contribution of each text to the value of *S*. 🅭🅯

In Figure 5, these 'text effects' are again shown using grey spikes. We observe that *S* shows similar behavior to D_2: Texts with a relatively low sub-frequency contribute more to its value. This feature was hinted at by Rosengren (1972: 118) and demonstrated by Lyne (1985: 109) who observed that *S* penalizes zeroes to a greater extent than D_2.

3.6. DP

Gries (2008: 415-419) introduced the 'deviation of proportions' (*DP*) as a dispersion index. While this measure had appeared in other disciplines under different names (e.g. Wilcox 1973: 328), Gries (2008) marks its first application to corpus statistics as well as its generalization to units of different size. By an unfortunate accident of history, Gries (2008)

introduced *DP* with reversed scaling (with 0 indicating evenness), thus deviating from the directionality ingrained in earlier indices (*TD, D, D2, S*). Following Burch et al. (2017: 193), we reverse its scaling and use the subscripted shorthand D_P to set it apart notationally from the inverted original (i.e. *DP*). Over the years, a number of minor modifications have been proposed (e.g. Lijffijt and Gries 2012; Egbert et al. 2020: 99), but these primarily affect *DP*'s behavior in boundary conditions; at its core, the computational procedure has not been altered. We will quote Gries' (2008: 415) original formula (though reversed in scaling, i.e. subtracted from 1), which suffices for our present purposes:[10]

(4) $1 - \frac{\sum_i^k |t_i - w_i|}{2}$

The rationale underlying D_P is best considered visually. Figure 6 shows the text-specific word shares w_i (horizontal line) and token shares t_i (points) for our corpus. If the item were dispersed evenly across the 10 texts, the points would be on the line: Each text's token share would then be equal to its word share. The black vertical lines reflect deviations from a perfectly even distribution.

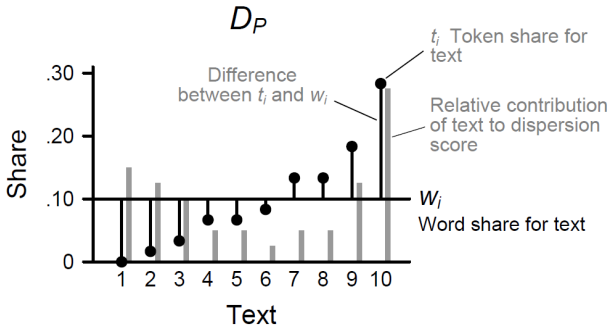

Figure 6. Formula for D_P: The deviation of token shares from the word shares (which are constant across texts here). CC BY

Looking at the numerator of (4), we note that the magnitude of D_P depends on the sum of the absolute values of these deviations. The further the points lie from the horizontal line, the greater this sum, and – since we subtract from 1, similar to *D* – the smaller the value of D_P. This means that the contribution of each text, which is again indicated in Figure 6 with grey spikes, is proportional to these distances. It is evident that D_P does not penalize low or high occurrence rates.

3.7. D_A

The measures considered so far are all based on one score per text. In contrast, D_A (Burch et al. 2017) considers all pairwise distances among texts.[11] This is illustrated for our exemplary data in Figure 7, where texts form a circle. Each text is represented by its occurrence rate (R_i) (in boldface), with its ID appearing in parentheses. The arrows originating from each

[10] For the analyses in Section 5, we use the modified version by Egbert et al. (2020: 99).
[11] While Burch et al. (2017) credit D_A to Wilcox's (1973: 329) *mean difference analog*, the measure in fact represents a special case of Gini's (1921) mean difference.

text are pointed at another text, and the length of these arrows is proportional to the distance between the two texts (on the R_i scale). Thus, a text surrounded by a large ball of arrows is an outlier since its R_i value lies far away from the other occurrence rates.

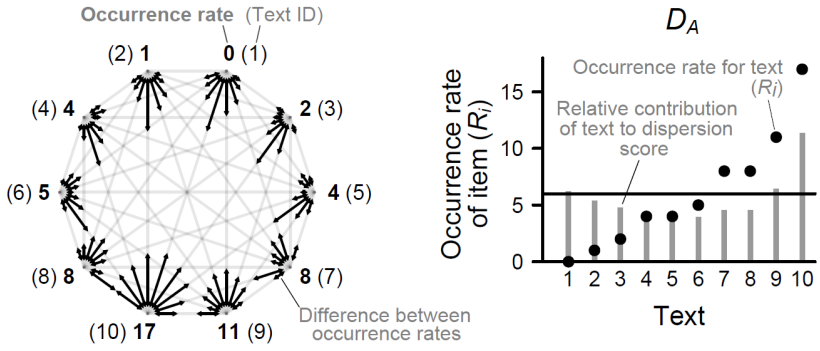

Figure 7. Calculation of D_A: All pairwise distances are considered. CC BY

For our 10 texts, there are 45 distances to consider. Even though it is not immediately apparent from the formula for D_A, which is given in (5) below, the index essentially reflects the mean over these distances. The further these occurrence rates are apart, the greater this average, and – since we subtract from 1, similar to D and D_P – the smaller the value of D_A.

$$(5)\quad 1 - \frac{\sum_{i}^{k-1}\sum_{j=i+1}^{k}|R_i - R_j|}{\frac{k(k-1)}{2}} \times \frac{1}{2\frac{\sum_{i}^{k} R_i}{k}}$$

To recognize the contribution each text makes to D_A, we can mentally average over the bundle of arrows departing from each text in Figure 7: Note how text 10 clearly sticks out. The grey spikes in panel (b) show the relative size of these averages. We find that D_A is similar to D and D_P in that it does not handicap occurrence rates below average (contrary to S and D_2).

Wilcox (1973: 341) and Burch et al. (2017: 210) argue that D_A has some advantages over D_P, and it therefore receives a cautious recommendation from both studies. Gries (2020: 116), on the other hand, points out that D_A is computationally much more expensive, especially when text files are the unit of analysis.[12]

3.8. D_{KL}

The most recently proposed measure, D_{KL} (Gries 2020: 103-104; 2021: 15, 20), is based on the Kullback-Leibler divergence, another information-theoretic measure. Using the notation listed in Table 3, and bringing the scaling in line with that used for the other measures, the formula for D_{KL} is:

$$(6)\quad D_{KL} = e^{-\sum_{i}^{k} t_i \log_2 \frac{t_i}{w_i}}$$

[12] While this certainly applies to the basic formula, Wilcox (1973: 343) gives a computational shortcut, which is much more efficient (see https://lsoenning.github.io/posts/2023-12-11_computation_DA/).

While it is difficult to isolate and visualize the contribution of individual texts via algebraic manipulation, a careful inspection of the formula allows us recognize the underlying logic. To make sense of (6), we get rid of the minus sign before the summation:

(7) $D_{KL} = e^{\sum_i^k t_i \log_2 \frac{w_i}{t_i}}$

Since the exponentiation of a sum (e^{a+b+c}) is equivalent to the product of the exponentiated parts of the sum ($e^a \times e^b \times e^c$), the formula for D_{KL} can be rewritten as

(8) $D_{KL} = e^{t_1 \log_2 \frac{w_1}{t_1}} \times e^{t_2 \log_2 \frac{w_2}{t_2}} \times \ldots \times e^{t_k \log_2 \frac{w_k}{t_k}}$

This means that D_{KL} is the product of k terms, one for each text. If the number of tokens in a text is proportional to its length, i.e. if t_i equals w_i, then $\log_2 \frac{w_1}{t_1}$ reduces to $\log_2(1)$, which is 0. The term $e^{t_1 \log_2 \frac{w_1}{t_1}}$ then becomes e^0, which is 1. This means that for a perfectly balanced distribution, we obtain a D_{KL} score of 1 (i.e. $1 \times 1 \times \ldots \times 1$).

It is difficult, however, to use the formula to isolate the contribution of an individual text like we have done for the other measures. In Section 5, we will use a different method that allows us to quantify the contribution of individual texts to the dispersion score. It will be observed that, for our illustrative set of data, D_{KL} groups with *S* and D_2 in that it imposes a penalty on lower sub-frequencies.

3.9. Summary

We have examined dispersion measures and tried to understand their structure and behavior in visual terms. This has allowed us to see how they process corpus-based counts, and their degree of (im)partiality towards low and high sub-frequencies. We were thereby able to extend the insights provided by Lyne (1985) to indices that have emerged more recently: While *S*, D_2, and D_{KL} inflict a handicap on low sub-frequencies, *D*, D_P, and D_A treat low and high sub-frequencies equal-handedly.

The insights we have gained so far are based on our exemplary data, which resemble the distribution of *which* in the Brown Corpus. While this adds realism to our example, we would like to monitor the behavior of dispersion measures in a wider variety of settings. The next section describes how we approach this goal.

4. Toward realistic settings for evaluation studies

The illustrative data sets used in the methodological literature on dispersion are usually didacticized mini-examples, with a small number of (hypothetical) text categories or genres (e.g. Gries 2008: 406; Burch et al. 2017: 198; Egbert et al. 2020: 101; Gries 2020: 102; Gries 2021: 20; Gries 2022: 178; Nelson 2023: 156). Since frequency distributions across smaller units of analysis differ from such macro-level tallies, it is not clear whether the insights provided by such examples extend to text-level analyses. To develop an intuition for the behavior of measures in such settings, then, we must broaden the set of scenarios. This section describes an approach to constructing controlled yet authentic conditions for the study of dispersion indices when texts form the unit of analysis.

To set up realistic data situations, we rely on information extracted from the Brown Corpus; the data underlying our analyses are available from TROLLing (Sönning 2024). Specifically, we consider the 2,000 most frequent word forms in the corpus and summarize their distributional characteristics using two parameters: a measure of central tendency (similar to the mean), which indicates how often a word form is used; and a measure of spread (similar to the standard deviation), which reflects the extent to which usage rates vary across the 500 texts in the corpus. To capture these features, we use the negative binomial distribution, which has been applied successfully to word frequency data in previous research. For instance, Mosteller and Wallace's (1964) well-known authorship attribution study relied on this distribution, and Church and Gale (1995) demonstrated its capacity for modeling text-level word frequencies.

The negative binomial distribution is an extension of the Poisson distribution and specifically designed to model count variables – the number of events that are observed over a certain period. For word frequency data, the event of interest is the occurrence of the item in question, and the period of observation is the length of the text, measured in the number of running words. In contrast to the Poisson, the negative binomial distribution includes an additional parameter that describes the text-to-text variability of the item's occurrence rate (see, e.g. Ehrenberg 1982: 59-63 for a concise introduction; Winter and Bürkner 2021 for a tutorial aimed at linguists; Long 1997: 217-238 for a thorough treatment; Sönning 2023a for a visual explanation). This additional parameter essentially functions like a standard deviation, and it therefore provides information about the dispersion of an item: If dispersion (in the corpus-linguistic sense) is low, occurrence rates vary widely from text to text; if dispersion is high, on the other hand, text-specific occurrence rates will be very similar.

To make matters concrete, consider Figure 8, which shows the frequency of *which* in the 500 texts in Brown. Each spike represents a text file, and text files are grouped by (sub)genre. The 'hairy' appearance of the graph indicates that *which* is a common word – it appears in almost every document. The token distribution across texts is summarized at the right, using a histogram: We see that most texts feature fewer than 10 instances.

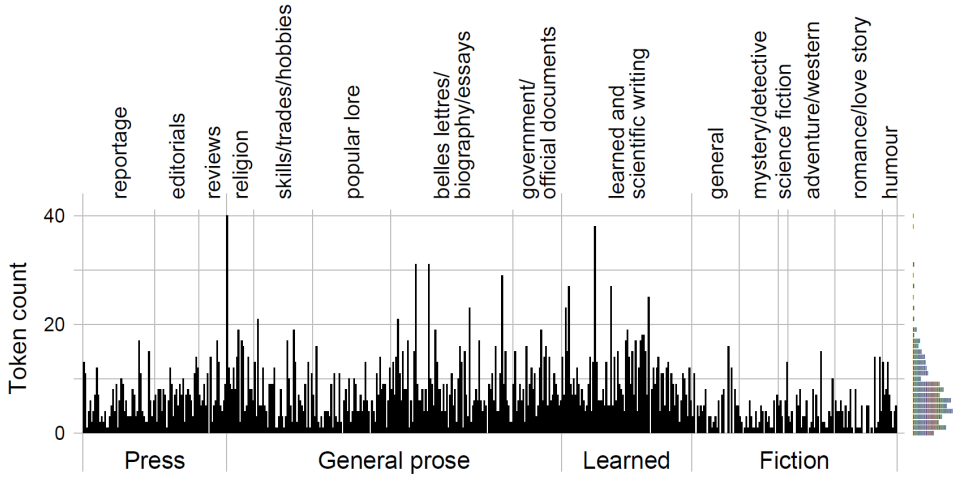

Figure 8. Distribution of *which* in the Brown Corpus. 🅭🅯

Figure 9 compares this histogram to the abstraction yielded by the negative binomial model, which appears as a black trace. Note that frequencies are now shown as occurrence rates (normalized frequency, per thousand words), hence the difference in units on the axes. Based on a visual inspection, the fit is quite good, as the model manages to capture the main features of the distribution.

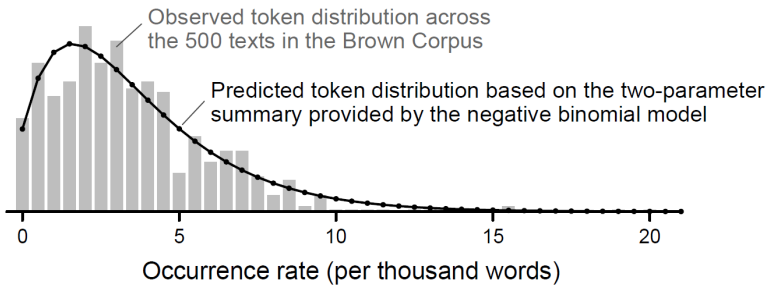

Figure 9. Model check: Observed distribution of occurrence rates across the 500 texts in Brown vs. predicted distribution based on the two-parameter negative binomial model.

We use this procedure to obtain a measure of location and spread for each of the 2,000 most frequent word forms in Brown. The 2,000 pairs of parameters are graphed in Figure 10, where each point represents a word form. The horizontal axis shows the (log-scaled) normalized frequency and the vertical axis shows the (log-scaled) additional parameter of the negative binomial distribution, which describes the variation of word rates across texts. For visual correspondence to the corpus-linguistic dispersion scale, values decrease from bottom to top: Items higher up in the graph are distributed more evenly across the 500 text files (i.e. they have higher dispersion).[13]

Figure 10 shows the distributional space spanned by the 2,000 most frequent word forms in Brown. We use it as a map, to locate a handful of sites that cover a fairly broad (but realistic) range of frequency and dispersion levels. For concreteness, we select representative items at these sites, which are marked in panel (b). These eight word forms (*the*, *for*, *he*, *any*, *my*, *exactly*, *built*, *poet*) will provide the distributional settings in which we will field-test, as it were, our dispersion measures. Before we turn to this task, however, we describe our approach to quantifying the sensitivity of measures to individual texts.

[13] In order to simplify the reading flow, we are skipping a number of details. First, since there are in fact two ways in which the negative binomial distribution can be defined (a direct and an indirect parameterization), there are two candidates for the additional parameter: the gamma scale and the gamma shape parameter (see Sönning 2023b). The R package we have used (gamlss; Rigby and Stasinopoulos 2005) relies on the direct parameterization. Figure 10 therefore shows the gamma scale parameter, which describes the variability of text-specific occurrence rates. The greater its value, the greater the variation of the occurrence rates. The y-axis in Figure 10 has therefore been inverted, since *high* variability in occurrence rates corresponds to low dispersion (in the corpus-linguistic sense); see https://lsoenning.github.io/posts/2023-01-18_dispersion_terminology/.

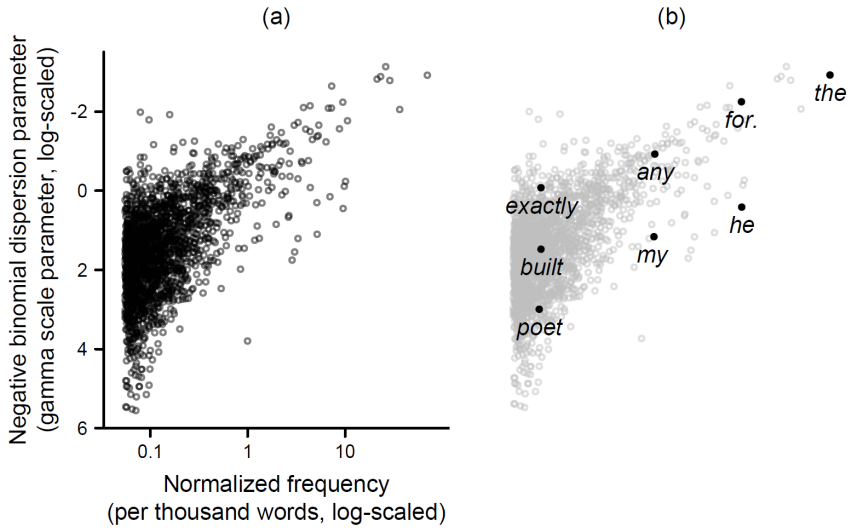

Figure 10. Negative binomial measures of location (x-axis) and spread (y-axis) for the 2,000 most frequent word forms in Brown.

5. Sensitivity of measures to distributional patterns

To sharpen our sense for the behavior of dispersion measures, we analyze their response to individual data points, i.e. the texts in our data. For most indices, our discussion in Section 3 has been able to reveal some interesting features, which had also been noted in earlier work (e.g. Lyne 1985). We now rely on a more general method for isolating the effect of individual texts, which can be applied to any of the measures listed in Figure 1. Section 5.1 describes the method and applies it to our illustrative data set. Section 5.2 then explains how we use it to investigate different sites on our distributional map. In Section 5.3 we present the results.

5.1. Method

Our method is straightforward: To quantify the effect a specific text on a dispersion score, we observe how the score changes if we remove the text from the data. We will interpret this difference as the effect of the text on the dispersion score. In general, we would expect the following consequences:

(a) If we exclude texts with extreme (i.e. very high or very low) frequencies, this will produce a more balanced distribution: Dispersion increases.
(b) If we exclude texts near the center of the distribution, i.e. with a proportionate sub-frequency, the resulting distribution is less balanced: Dispersion should decrease.

Let us now apply the method to our mini-corpus. We start by calculating dispersion scores for the full data set (i.e. all 10 texts); these appear in Figure 11 as dotted horizontal lines. Next, we turn to our data manipulations. Since there are eight unique token counts in our data, we can form eight data subsets (of 9 texts each). The dispersion scores we obtain for these subsets form a U-shaped profile, which makes sense: When excluding texts with extreme occurrence rates (i.e. text 1 or 10), we note an increase in the dispersion statistic.

Some measures show steeper gradients, however, especially at the left margin of the plot. This is how our method reveals 'penalization' of zero and low sub-frequencies. We will return to these patterns shortly.

One issue that arises when comparing the curves shown in Figure 11a is the bounded nature of the scale. Due to floor and ceiling effects (at 0 and 1), variation is constrained near the endpoints. As a result, measures that produce relatively high scores for a specific set of frequencies may appear to be less affected by our manipulation. Measures that yield values near the midpoint of the scale (.50), on the other hand, have more room to vary. Consider D_2, for instance, which produces the highest dispersion scores: Its profile is relatively flat. D_A, on the other hand, shows a more pronounced bend. To remove this scaling artefact and effect better comparability between indices, we will transform scores to the logit scale[14] and redraw the graph. Figure 11b shows the result: The profiles for D_A and D_{KL} have been somewhat flattened, while the vertical range for D_2 and S has been stretched. Since we are primarily interested in comparisons among measures, we will give preference to the logit scale in the following analyses.

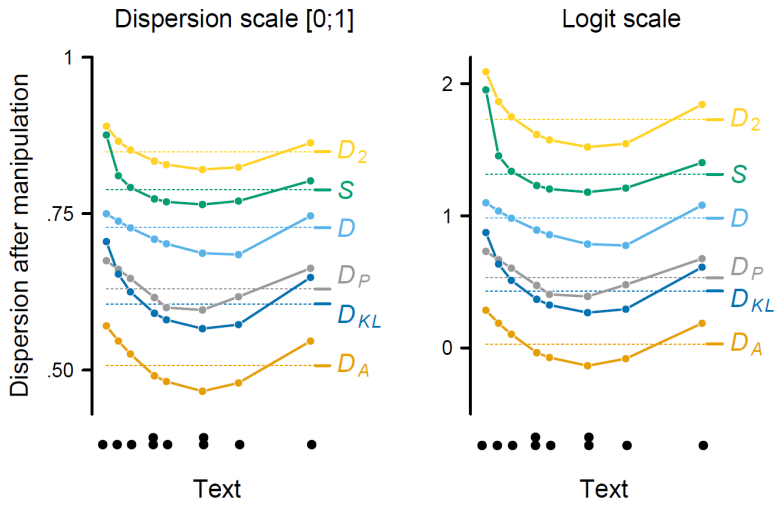

Figure 11. Dispersion scores for the modified data sets on the (a) dispersion scale and (b) logit scale. ⓒⓘ

To be able to make direct comparisons among measures, we must now look for a way to quantify text effects. It turns out that the effect of a specific text on the dispersion score can be assessed by comparing, in Figure 11, the points along the U-shaped profiles to the dotted reference lines. Our next step is therefore to calculate these differences and then compare them across measures. That is, we are asking by how much, and into what direction, scores change if we remove a text. An answer can be given for each measure and token count; we will refer to this difference as the 'text effect'.

Figure 12 shows these differences, or text effects, on the logit scale. We would read this graph as follows: First, we locate the black horizontal line, which marks a difference of zero.

[14] Since the logit-transformation does not accept proportions of 0 and 1, prior to transformation all dispersion scores below .001 and above .999 were replaced with .001 and .999, respectively.

Values near this line indicate that the removal of the text had little effect on the dispersion statistic. Points below this reference line, on the other hand, indicate that a text negatively affects the dispersion score. Thus, if its exclusion yields a more even distribution (producing the U-shapes in Figure 12), this means that its presence in the original data set has a negative influence on the dispersion score. We will refer to the traces shown in Figure 12 as *sensitivity profiles*. We observe the most dramatic differences among measures at the left end of the graph (for text 1), which shows the effect of a text with a sub-frequency of 0.

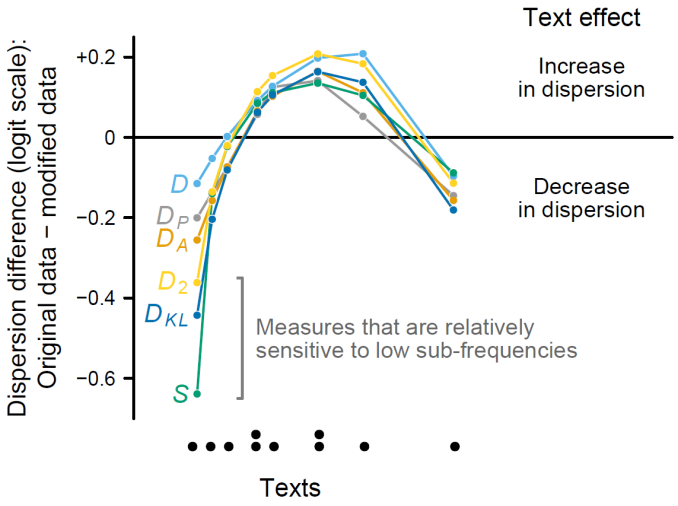

Figure 12. Sensitivity profiles for our illustrative data set.

5.2. Construction of sensitivity profiles for distributional patterns

We would now like to run this kind of sensitivity analysis at a larger scale, i.e. by considering the representative sites identified in Figure 10. The scenario we will study mirrors Brown: a corpus of 500 texts, each 2,000 words long. Our goal is to construct sensitivity profiles similar to those in Figure 12. The remainder of this section describes our method and may be skipped without loss in continuity. Details about the computational implementation may be found in the associated OSF project (https://osf.io/nhw2y/).

In order to obtain smooth sensitivity profiles for each site on our distributional map, we must find a way of averaging out the sampling variation that is inherent in corpus data. After all, like any other corpus, Brown is a sample of language use, in this case consisting of 500 texts. If a different set of texts had been selected, the distributional pattern for a specific word form would be different.

In order to iron out this sampling variation, we need an idealized representation of the distributional features of these items. The idea is to find, for each item, a statistical model that could plausibly have given rise (in a statistical sense) to the token distributions observed in Brown. To this end, we stick to the negative binomial model we have constructed for the eight items (cf. Section 4). To see whether the match between data and model is satisfactory, we ran graphical model checks similar to the one shown in Figure 9,

which may be found in Web Appendix 2. These indicate that the negative binomial model offers a reasonable abstraction.

For our present purposes, then, we will assume that, for each item, the difference between the token distribution in Brown and the underlying statistical model is due to sampling variation. We can then use the statistical model to generate 'alternative Brown corpora'. The token distributions we generate will differ from one another due to sampling variation only. This allows us to get a handle on the likely extent of sampling variability, which we smoothen by averaging over them.

We use the model parameters to generate 50,000 token distributions. A token distribution, in our case, consists of a set of 500 token counts – one count per text in our imaginary corpus. Each of these token distributions is then submitted to the modification procedure described above and we record the effect of individual texts – or more specifically: token counts – on the observed dispersion score. Since we can only do so for the token counts that are attested in the specific token distribution at hand, the results for many simulations will have gaps. To fill these, we run a large number of iterations (i.e. 50,000).

A decision is also needed about the range of token counts to consider. While we are interested in monitoring the behavior of measures when confronted with relatively extreme sub-frequencies, we want to exclude from consideration frequencies that are unlikely to be observed. We put the probabilistic threshold at 1/1000, which means that we exclude those token counts that, based on our statistical model, are expected to occur, on average, at most once per 1,000 texts (of 2,000 words). This means that unusually low counts and unusually high counts are disregarded, which protects us from giving undue attention to unrealistic (or rare) situations.

Using this procedure, we obtain, for each site on our map, a distribution of up to 50,000 'text effects' per dispersion measure and token count. Typically, this number will be (much) smaller, however, due to inevitable gaps and data sparseness in the tails of the token distributions. We then tabulate, for each dispersion measure and distributional setting, the median text effect for every relevant token count. The patterns formed by these median text effects are the sensitivity profiles we will discuss in the following section.

5.3. Results

We come now to the results of our sensitivity analysis, which are shown graphically in Figure 13. What drives our interpretation of these profiles is the question of how dispersion measures respond to low and high sub-frequencies in different distributional milieus. This means that we will pay particular attention to the tails of the profiles. Recall that the token counts we consider for exclusion have a probability of occurrence of at least 0.1% in a 2,000-word text. Since some readers may feel that this likelihood is a rather low, we also add grey shading to Figure 13 to denote token counts that have a probability of 1% of being observed.

Let us first consider the sensitivity of measures to low sub-frequencies (including 0). Essentially, we approach Figure 13 by asking whether the patterns we saw in Figure 12 (and in Section 3), i.e. that D_2, D_{KL}, and S are overly sensitive to lower sub-frequencies, hold in other areas of the distributional space. For convenience, the insights we extract from Figure 13 are summarized in Table 4, where a filled circle (●) indicates relatively strong

penalization and an empty circle (○) indicates relatively weak penalization. There are several things to note:

- In the distributional space we have studied, *D* never penalizes lower sub-frequencies
- D_2, *S*, and D_{KL} inflict a handicap on low sub-frequencies, but this behavior is attenuated for word forms in lower frequency and dispersion bands
- In the lower frequency ranges (< 100 pmw), none of the measures impose a penalty on zero counts

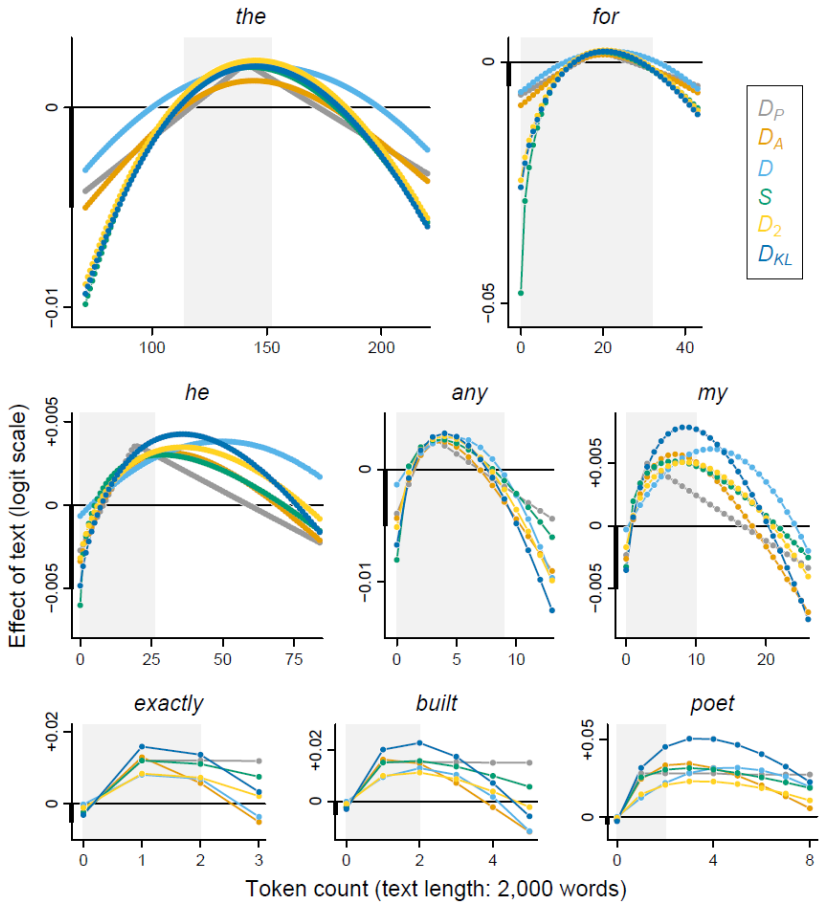

Figure 13. Sensitivity profiles for dispersion measures in different distributional settings. 🅭🅯

Next, we consider the upper tail of the distributions. For the moment, our focus remains on penalization (i.e. the symbols ● ○). The summary in Table 4 shows that the picture is more mixed:

- Overall, *D* and D_P show the most balanced performance, producing at most weak penalization against high sub-frequencies
- In lower frequency bands (< 100 pmw), none of the measures impose a penalty on high counts

In general, then, issues of penalization appear to be more of a concern for items with normalized frequencies exceeding roughly 100 pmw. Typically, the number of word forms clearing this threshold is somewhere around 1,000: In Brown, roughly 1,000 forms have a

frequency greater than 100 pmw; in Leech et al.'s (2001) report for the BNC, about 800 word forms exceed this threshold in the spoken part, and 1,100 forms in the written part.

In the distributional space below this approximate threshold, all dispersion measures show markedly different behavior. Most notably, low sub-frequencies exert no appreciable leverage on the observed scores. This may be because sub-frequencies of 0 account for the greatest share in these frequency bands: *exactly* is absent from 83% of the texts, *built* from 86%, and *poet* from 92% of the texts. Since low sub-frequencies are so typical, a removal of one text (i.e. 1% of the data) appears to carry no weight.

The sensitivity profiles for *exactly*, *built*, and *poet* show two further features of interest. First, in the sensitivity ranges considered here, texts with non-zero token counts typically lead to an increase in dispersion. While this response may be quite sensible for sub-frequencies of 1, texts with a token count of 2 (or higher) should actually yield a decrease in dispersion. This is because counts of 2 are already quite high for the three items considered: In Brown, only 2.4% of texts feature 2 or more occurrences of *exactly*, and this figure is 3.8% for *built* and 2.6% for *poet*. In other words, the behavior of dispersion measures appears counterintuitive when confronted with high sub-frequencies of items with an overall occurrence rate of roughly 100 pmw. This feature is indicated in Table 4 using open triangles (△) – apparently, none of the dispersion measures is immune to this issue.

Table 4. Summary of patterns in Figure 13

			Low sub-frequencies						High sub-frequencies					
Item	Frequency	Dispersion	D_A	D_P	D	D_2	S	D_{KL}	D_A	D_P	D	D_2	S	D_{KL}
the	++	++				●	●	●				●	●	●
for	+	+				●	●	●				○	○	○
he	+		○	○		○	●	●						
any		+	○	○		○	●	●	○		○	○		●
my		−	○	○		○	○	○	●	○		○		●
exactly	−									▲		△	△	△
built	−	−								▲			△	
poet	−	−−							△	▲	△		△	△

Note. Key to symbols: ● relatively strong penalization; ○ relatively weak penalization; ▲ unresponsive to high sub-frequencies; △ counterintuitive response to relatively high sub-frequencies

The second peculiarity we note concerns the behavior of D_P for the low-frequency word forms (i.e. *exactly*, *built*, and *poet*). Reference to the numbers underlying Figure 13 confirms that the D_P score virtually flatlines[15] across non-zero sub-frequencies. This behavior suggests that D_P may lack sensitivity to the kinds of data manipulations we have implemented: It appears to show no (or little) response to the removal of a text with relatively high sub-frequencies.[16] In Table 4, this feature of D_P is denoted using filled triangles (▲).

[15] A similar pattern emerges when calculating the mean (instead of the median) across the simulated data sets.
[16] We would like to thank an anonymous reviewer for strongly suggesting that this observation be explained on a theoretical level; the following elaborations represent our attempts to do so.

In fact, it has been noted in previous work that there are conditions under which D_P does not take into account all sub-frequencies in the data. While this was stated quite clearly in Wilcox (1973: 339), the nature of these "conditions" is concealed in the computational formulas listed in the Appendix of that paper (ibid: 343). Burch et al. (2017: 213) elaborate on this feature of D_P and discuss a formula given in Wilcox (1973: 343). The basic insight that emerges is that if texts have the same length, D_P can in fact be computed from a subset of the texts: If we determine the average occurrence rate across all texts and then arrange them by occurrence rate in decreasing order, D_P takes as input only those texts whose occurrence rate falls below the overall average. This means that the value of D_P does not depend on (or reflect) texts with a sub-frequency above average. It follows that it does not matter how far above the average a specific occurrence rate is, as D_P is unresponsive to its magnitude. This is why the sensitivity profile for this index flatlines across sub-frequencies greater than 0.

To make this feature of D_P more concrete for the data at hand, we can identify areas on our distributional map where the index can lack sensitivity to the magnitude of high sub-frequencies. We identify these regions using simulation. The idea is to generate prototypical token distributions across our map (cf. Figure 10), and then monitor whether the removal of non-zero sub-frequencies leads to a change in the D_P score. If the removal of a sub-frequency does not alter the dispersion score, we have identified a spot on the map where D_P can show the kind of unresponsiveness to high sub-frequencies that we observed for *exactly*, *built*, and *poet*.

We use the negative binomial distribution to generate idealized token distributions across a grid of location and spread values. These prototypical distributions[17] aim to approximate the negative binomial model as closely as possible. In Figure 9, for example, this would be observed in a close correspondence between the grey bars and the black profile. Importantly, we have to settle on the number and length of texts to produce these token distributions. We choose these to reflect the make-up of Brown (500 texts à 2,000 words). This means that our map will only tell us about the behavior of D_P on this specific corpus design.

The results of our simulation are summarized in Figure 14, which covers the same area as Figure 10. For orientation, the top 2,000 word forms in Brown also appear in the graph. To get a better sense of the nature of token distributions across the map, thin lines show the expected percentage of texts that feature 0 occurrences of the item. This share increases (i) toward the left, i.e. as the occurrence rate of the item drops, and (ii) toward the bottom, as the dispersion of the item increases. Our attention, however, is centered the grey shading, which denotes areas where D_P may lack sensitivity to the removal of high sub-frequencies. Different shades of grey reflect how likely it is that D_P will show this type of unresponsiveness, i.e. how likely it is that the index fails to discriminate between token distributions that differ in the upper tail. These probability bands are arranged like onion shells and the likelihood extends up to .30 (or 30%).

[17] As documented in the R scripts that are available in the OSF project (https://osf.io/nhw2y/), we use a quantile-based approach to construct these prototypical distributions. It is quite difficult, however, to faithfully represent the upper tail of many distributions using 500 data points (the number of texts in our simulation), especially if frequency and/or the text-to-text variability in occurrence rates is high, as the negative binomial distribution can have a very long and very thin tail in these settings. The grey shading in Figure 14 should therefore be considered an approximation.

The most striking pattern in Figure 14 is the vertical incision at 0.5 per thousand words, or, expressed in term of occurrences in a 2,000-word text (as in Brown), at an average number of 1 token per text. This threshold follows from the points we have noted above, i.e. the fact that D_P is blind to sub-frequencies above average. If this average is above 1 token per text (the vertical incision in Figure 14), D_P will take into account those texts that have a token count of 0 or 1. If we drop a text with a high sub-frequency from the data, the average token count can shift below 1, in which case D_P will change accordingly: Texts with a count of 1 are then disregarded. If, on the other hand, the average token count is below 1 already, dropping high sub-frequencies will have no effect on the subset of texts that D_P embraces: Still only those texts with counts of 0 will feed into the dispersion score.

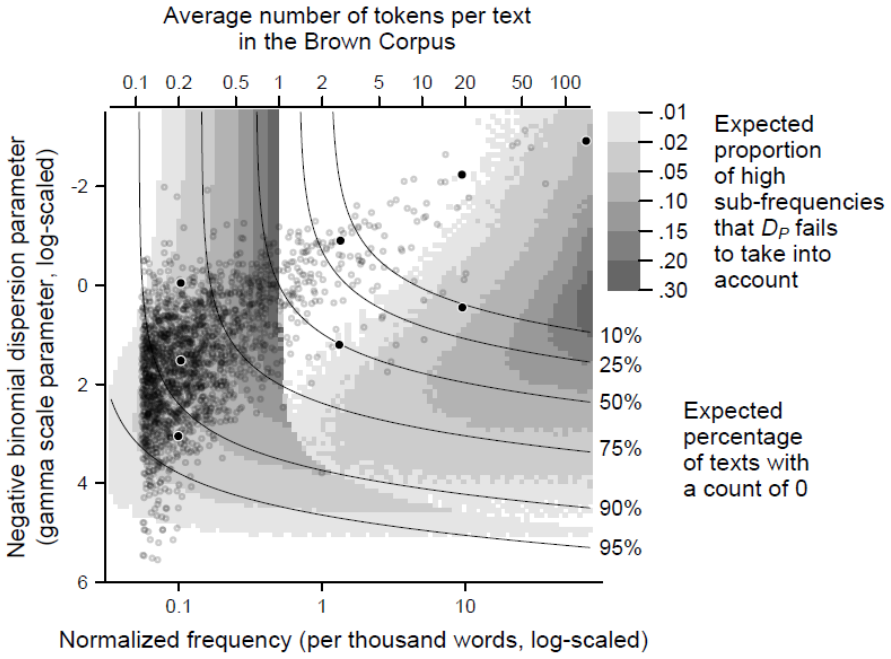

Figure 14. Approximate regions where D_P is unresponsive to high sub-frequencies in the Brown Corpus. Grey shades denote the probability of unresponsiveness. ⓒⓘ

6. Summary and conclusion

The overarching goal of the present paper has been to advance our understanding of dispersion and its measurement using corpus data. We started out with an overview of the gradual evolution of dispersion measures, their implementation in corpus analysis software, and their use in current corpus-based work. Our literature survey showed that corpus linguists rely on a handful of indices, which they typically apply to meaningful corpus units such as text files and (to a lesser extent) genres. This points to a discrepancy between applied work and the kinds of hypothetical data settings found in the methodological literature. Thus, previous evaluations of indices have dealt with data distributions that mirror frequencies observed across groups of texts such as genres. Token distributions across text files, however, differ markedly from such aggregated tallies: The number of units is (usually much) larger, and units are shorter in length.

Since little is known about the behavior of dispersion measures in text-level analysis settings, we have directed attention to frequency profiles that mimic text-level token distributions. To this end, we relied on the Brown Corpus to build realistic scenarios that span a relatively broad but representative portion of the distributional space covered by the 2,000 most frequent word forms in the corpus. Our interest then centered on the sensitivity of measures to relatively high and low sub-frequencies. We relied on a simple form of data manipulation to document text effects and introduced the notion of a sensitivity profile, which allowed us to quantify these effects and compare them across measures. This allowed us to extend the insights provided by Lyne (1985) to a wider range of measures and distributional milieus.

Our evaluation study allowed us to identify indices with desirable properties. Since there appears to be no noticeable justification for penalizing lower (or higher) sub-frequencies, practitioners may decide to give preference to measures that treat higher and lower sub-frequencies on a par (i.e. D or D_P). It should be stressed that over-sensitivity to zeroes is a rather problematic feature in text-level analyses. This is because texts may (and often do) differ in length. The shorter the text, the greater the likelihood of observing a count of zero – for *any* item. This may distort dispersion assessments, which appears to be particularly problematic when comparing (sub-)corpora across which text lengths differ systematically.

Perhaps more importantly, the sensitivity profiles we observed for the lower-frequency items *exactly*, *built*, and *poet* have demonstrated that the behavior of dispersion measures in these areas of the distributional space requires further research. In particular, the responses we noted when removing texts with high sub-frequencies appears to be at odds with our expectations – extreme sub-frequencies should depress dispersion scores. The behavior we did observe, i.e. no notable depression of scores, may arise from the correlation that has been observed to hold more generally between frequency and dispersion. Thus, the higher the overall frequency of an item, the higher its dispersion (Gries 2022). This correlation may help explain the behavior of measures when confronted with high sub-frequencies of low-frequency items (e.g. *exactly*, *built*, *poet*): If we remove a text with a high sub-frequency, the overall (average) frequency of the item drops. Due to the association between frequency and dispersion, this will also deflate the dispersion score. And this frequency-induced decrease may counteract the increase in evenness we observe in the data. Whether the behavior of *exactly*, *built*, and *poet* in fact reflects these opposing forces is a question that remains to be explored more fully in future work. As long as we are lacking a thorough understanding of such artifacts and their patterned occurrence, we must remain cautious when interpreting response profiles observed in these distributional domains.

Our sensitivity analysis has also drawn our attention to a peculiarity in the behavior of D_P, which had been noted in earlier work (Wilcox 1973; Burch et al. 2017). The distributional space we have constructed using the negative binomial model allowed us to figure out the nature of D_P's unresponsiveness to high sub-frequencies, and to see which areas are likely to be affected. For the corpus design at hand (500 texts, each 2,000 words long), we were able to make sense of the patterns suggested by our simulations. Whether and how these insights extend to corpus layouts where texts differ in length is a key area for follow-up research.

Finally, it must also be noted that we have restricted our attention to distributional patterns found among the top 2,000 word forms in Brown. This means that the coverage provided by our distributional map in Figure 10 is incomplete. Future work must venture into lower-

frequency domains, which would seem essential for forming an understanding of sensitivity patterns for the bulk of lexical items that may be subject to dispersion analysis. We have hinted at approximate thresholds below which measures may show a change in behavior. However, these observations have been rather coarse and indeed rest on insecure grounds. A more systematic and thorough coverage of the distributional space may allow us to derive more elaborate response profiles and delineate areas where dispersion measures may be expected to exhibit a shift in response style. This kind of cartographic work may produce valuable and practicable insights that are relevant to methodologists and applied researchers alike.

References

Anthony, Laurence. 2022. AntConc (Version 4.2.0) [Software]. Tokyo: Waseda University. Available from https://www.laurenceanthony.net/software

Biber, Douglas, Randi Reppen, Erin Schnur & Romy Ghanem. 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.

Brezina, Vaclav & William Platt. 2023. #LancsBox X (version 3.0.0) [software]. Available from: http://lancsbox.lancs.ac.uk

Burch, Brent, Jesse Egbert & Douglas Biber. 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, 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.

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

Egbert, Jesse, Brent Burch & Douglas Biber. 2020. Lexical dispersion and corpus design. *International Journal of Corpus Linguistics* 25(1). 89–115.

Ehrenberg, Andrew S. C. 1982. *A primer in data reduction.* Chichester, UK: John Wiley & Sons.

Francis, W. Nelson & Henry Kučera. 1964. *Manual of information to accompany a standard corpus of present-day edited American English for use with digital computers.* Providence, Rhode Island: Department of Linguistics, Brown University. http://khnt.hit.uib.no/icame/manuals/brown/index.htm.

Gini, Corrado. 1921. Measurement of inequality of incomes. *Economic Journal* 31(121). 124–126.

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

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.

Gries, Stefan Th. 2020. Analyzing dispersion. In Magali Paquot & Stefan Th. Gries (eds.), *A practical handbook of corpus linguistics*, 99–118. New York: Springer.

Gries, Stefan Th. 2021. A new approach to (key) keywords analysis: Using frequency, and now also dispersion. *Research in Corpus Linguistics* 9(2). 1–33.

Gries, Stefan Th. 2022. What do (most of) our dispersion measures measure (most)? Dispersion? *Journal of Second Language Studies* 5(2). 171–205.

Hardie, Andrew. 2012. CQPweb - combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics* 17 (3). 380–409.

Juilland, Alphonse G. & Eugenio Chang-Rodríguez. 1964. *Frequency dictionary of Spanish words.* The Hague: Mouton de Gruyter.

Keniston, Hayward. 1920. Common words in Spanish. *Hispania* 3(2). 85–96.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý & Vít Suchomel. 2014. The Sketch Engine: Ten years on. *Lexicography* 1. 7–36.

Leech, Geoffrey, Paul Rayson & Andrew Wilson. 2001. *Word frequencies in written and spoken English based on the British National Corpus*. London: Longman.
Lijffijt, Jefrey & Stefan Th. Gries. 2012. Correction to Stefan Th. Gries' 'Dispersions and adjusted frequencies in corpora'. *International Journal of Corpus Linguistics* 17(1). 147–149.
Long, J. Scott. 1997. Regression models for categorical and limited dependent variables. Thousand Oaks, CA: Sage.
Lyne, Anthony A. 1985. *The vocabulary of French business correspondence*. Paris: Slatkine-Champion.
Mosteller, Frederick & David L. Wallace. 1984. *Applied Bayesian inference: The case of The Federalist Papers*. New York: Springer.
Nelson, Robert N. 2023. Too noisy at the bottom: Why Gries' (2008, 2020) dispersion measures cannot identify unbiased distributions of words. *Journal of Quantitative Linguistics* 30(2). 153-166.
Oakes, Michael P. 1998. *Statistics for corpus linguistics*. Edinburgh: Edinburgh University Press.
Rayson, Paul. 2009. Wmatrix: a web-based corpus processing environment, Computing Department, Lancaster University. http://ucrel.lancs.ac.uk/wmatrix/
Rigby, Robert A. & Mikis D. Stasinopoulos. 2005. Generalized additive models for location, scale and shape. *Applied Statistics* 54(3). 507–54.
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.
Sarkar, Deepayan. 2008. *Lattice: Multivariate data visualization with R*. New York: Springer.
Savický, Petr and Hlaváčová, Jaroslava. 2002. Measures of word commonness. *Journal of Quantitative Linguistics* 9. 215–231.
Scott, Mike. 2015. WordSmith Tools manual (version 6.0) [software]. Lexical Analysis Software Ltd. Stroud, Gloucestershire, UK.
Sönning, Lukas. 2023a. The negative binomial distribution: A visual explanation. *Statistics for linguist(ic)s blog*. December 12. https://lsoenning.github.io/posts/2023-11-16_negative_binomial/
Sönning, Lukas. 2023b. Different parameterizations of the negative binomial distribution. *Statistics for linguist(ic)s blog*. December 13. https://lsoenning.github.io/posts/2023-12-13_negative_binomial_parameterization/
Sönning, Lukas 2024. *Background data for: Advancing our understanding of dispersion measures in corpus research*. DataverseNO, V1. https://doi.org/10.18710/FVHTFM
Thorndike, Edward L. 1921a. *The teacher's word book*. Teachers College, Columbia University.
Thorndike, Edward L. 1921b. Word knowledge in the elementary school. *Teachers College Record* 22(4). 1–27.
Vander Beke, George E. 1932. *French word book*. New York: Macmillan.
Wilcox, Allen R. 1973. Indices of qualitative variation and political measurement. *The Western Political Quarterly* 26 (2). 325–343.
Winter, Bodo & Paul-Christian Bürkner. 2021. Poisson regression for linguists: A tutorial introduction to modelling count data with brms. *Language and Linguistics Compass* e12439.