Dispersions and adjusted frequencies in corpora

Stefan Th. Gries
University of California, Santa Barbara

The most frequent statistics in corpus linguistics are frequencies of occurrence and frequencies of co-occurrence of two or more linguistic variables. However, such frequencies in isolation may sometimes be misleading since they do not take into consideration the degree of dispersion of the relevant linguistic variable. Many dispersion measures and adjusted frequency measures have been suggested but are neither widely known nor applied. Another unfortunate aspect of such measures is that many also come with a variety of problems. I pursue three objectives with this article. First, I want to raise awareness of this issue and make the available measures more widely known, so I present an overview of many measures of dispersion and adjusted frequencies. Second, I propose a conceptually simple alternative measure, *DP*, explain and exemplify it, and compare it to previously discussed measures. Third and most importantly, I urge corpus linguists to explore the notion of dispersion in more detail and outline a few proposals which steps to take next.

Keywords: frequency of occurrence, frequency of co-occurrence, dispersion, constructions/patterns, collocations, collostructions

1. Introduction

The most frequently used statistic in corpus linguistics is the frequency of occurrence of some linguistic variable or the frequency of co-occurrence of two or more linguistic variables. The former is usually either invoked for individual words or grammatical patterns or more globally in the form of partial or complete frequency lists. In both of these forms, frequencies are reported, among other things, to indicate the importance of particular words / grammatical patterns for language teaching or to reflect the degree of cognitive entrenchment of particular words / grammatical patterns.

International Journal of Corpus Linguistics 13:4 (2008), **403–437**. DOI 10.1075/ijcl.13.4.02gri
ISSN 1384–6655 / E-ISSN 1569–9811

However, even though this is apparently not recognized much in the field, frequencies of (co-)occurrence may sometimes be incredibly misleading.[1] An instructive example for how raw frequencies can be misleading indicators of the overall importance of words is discussed by Leech et al. (2001). They show that the words *HIV, keeper,* and *lively* are about equally frequent in the British National Corpus (16 occurrences p.m.), which would usually be interpreted as an indication of their overall similar importance. A look at how these words are distributed in the corpus, however, suggests a very different result. While *lively* and *keeper* both occur in 97 of 100 equally-sized corpus parts, *HIV* occurs in only 62, which already indicates that *HIV* is much more specialized. This assessment is supported when Leech et al. compute a more refined measure of dispersion, Juilland et al.'s *D*. Juilland et al.'s (1970) *D* for *lively*, *keeper*, and *HIV* is 0.92, 0.87, and 0.56 respectively, indicating that these three words are far from being equally distributed across the corpus and that, more generally, frequency data should be augmented with information on the dispersion of the items in question. To give just one other example, in the domain of second language acquisition, Ellis and Simson-Vlach (2005) as well as Ellis et al. (2007) demonstrate that the number of academic genres in which a particular *n*-gram appears 4+ times — what they refer to as 'range' — is significantly correlated with processing speed and among the most important determinants of subjects' reading times and, thus, their first sorting criterion in determining *n*-gram lists for learners of (academic) English.

Although dispersion is virtually always only mentioned in the domain of frequencies of occurrence (if at all, that is), it can in fact be equally troublesome when one turns to an area where many people have been concerned with the choice of the right kind of statistics: co-occurrence frequencies or even more complex statistical measures based on co-occurrence frequencies. For example, Stefanowitsch and Gries (2003) introduced an extension of statistical approaches towards collocations or colligations called collexeme analysis. This method quantifies the degree to which particular words are attracted to, or repelled by, syntactically defined slots in grammatical patterns or constructions. Examples include the verb in the verb slot in the passive construction (cf. (1)) or in the ditransitive construction (cf. (2)); cf. Stefanowitsch and Gries (2003) for further examples.

(1) a. *John was shot by Mary*.
    b. verbs attracted to the passive: *base*, *concern*, *use*, *involve*, *publish*, *associate*, …
    c. verbs repelled by the passive: *have*, *think*, *get*, *say*, *want*, *do*, *know*, …

(2) a. *John gave Mary the book*.
    b. verbs attracted to the ditransitive: *give, tell, show, offer, cost, teach, wish, …*
    c. verbs repelled by the ditransitive: *bring*, *play*, *take*, *pass*, *make*, *sell*, …

Crucially, the degree of attraction in collexeme analysis is computed on the basis of co-occurrence frequencies as represented in Table 1.

Table 1. Schematic representation of the table underlying most co-occurrence statistics

	Construction *c*	**Other constructions**	**Row totals**
Verb *v*	*A*	*b*	*a+b*
Other verbs	*C*	*d*	*c+d*
Column totals	*a+c*	*b+d*	*a+b+c+d*

While a variety of different measures are available for the analysis of such tables, Stefanowitsch and Gries (2003) used the $p_{Fisher\text{-}Yates\ exact\ test}$-value; later publications converted this *p*-value into a more easily interpretable negative logarithm to the base of 10 of that *p*-value such that small values and large values indicate degrees of attraction and repulsion respectively. This method and its extensions have provided interesting results in a variety of applications — the syntax-lexis interface (Gries & Stefanowitsch 2004a, b), syntactic priming (Gries 2005; Szmrecsanyi 2005, 2006), second language acquisition (Gries & Wulff 2005) — and it has experimentally been shown to be a better predictor of native speaker performance or judgments than just raw frequencies (cf. Gries et al. 2005, to appear; Ellis & Simpson-Vlach, to appear).

However, in spite of these advantages, like all methods based on raw frequencies of (co-)occurrence it can run into problems when the dispersion of elements is not taken into consideration. Stefanowitsch and Gries's (2003) analysis of the imperative construction in the British component of the International Corpus of English (ICE-GB) yielded the list of verbs in (3) that are most strongly attracted to the imperative (in descending order, with observed frequencies in the imperative in parentheses):[2]

(3) *let* (86), *see* (171), *look* (74), *listen* (26), *worry* (21), *fold* (16), *remember* (35), *check* (21), *process* (15), try (47), *hang on* (17), *tell* (46), *note* (16), *add* (21), *keep* (28)

The problem is that the highly-ranked verbs *fold* and *process* occur in the imperative each in just a single file; the former in a file containing a text on origami, the latter in a file from a cook book. Thus, although there is a relatively high frequency of occurrence of both verbs in the imperative — especially given their overall not so high frequency — and, therefore, a statistically high degree of attraction, this statistical result must be taken with a grain of salt since there is a high likelihood that these high frequencies are not particularly representative of what is going in the corpus as a whole (cf. Stefanowitsch & Gries 2003:237f.).

Gries (2006) returned to this problem and showed that a variety of other verbs from a similar frequency range as *fold* and *process* — the frequency range from 13 to 17 occurrences in the imperative — may yield lower degrees of attraction than *fold* and *process*, but are actually more widely distributed in the corpus as measured by both the number of files in which the verbs occur in the ditransitive construction and Carroll's D_2. I will return to this example below.

Both these examples show that the dispersion of elements or co-occurring elements in a corpus is highly relevant information both in and of itself as well as a factor that can strongly influence many other corpus-linguistic statistics. However, neither is this issue acknowledged in the vast majority of the corpus-linguistic literature nor is it yet clear how to best handle dispersion and its impact on the interpretation of observed frequencies. The following section surveys a variety of measures that have been suggested in the literature as ways to quantify dispersion and adjust observed frequencies on the basis of an elements dispersion throughout a corpus.

2. An overview of dispersion measures and adjusted frequencies

To the best of my knowledge, this section surveys all dispersion measures and adjusted frequencies that have been proposed so far. While this paper will be mostly concerned with the former kind of measure, many measures of the latter kind are derived from the former and the whole topic of adjusted frequencies is inextricably related to matters of dispersion, which is why I will survey both kinds of measures here. To clarify how all of these measures are computed and provide a more unified overview, a few terms need to be introduced.

Let us assume this is our corpus of length $l = 50$, where letters represent words and the pipes the division of the corpus into different, here, $n = 5$ equally-sized parts.[3]

> *b a m n i b e u p k | 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*

The percentages that each of the parts makes up of the whole corpus — in this case all 0.2 — are denoted as s_1 to s_5. Let's assume we are interested in the word *a* in the corpus. The frequencies with which *a* occurs in each part are denoted as v_1, v_2, etc.; as you can see, *a* occurs once in the first part, twice in the second part, and so on, such that $v_1 = 1$, $v_2 = 2$, $v_3 = 3$, $v_4 = 4$, $v_5 = 5$. The vector with all observed frequencies (1, 2, 3, 4, 5) is referred to as v and the sum of all observed frequencies, i.e., the number of occurrences of *a* is referred to as f ($f = \Sigma v = 15$). Note also some other words' distributions: *x* occurs only once in the whole corpus; *e* occurs once in each

part and always in the same position, *b* occurs twice in each file and always in the same positions. In what follows, I will first provide an overview of the dispersion measures (in Section 2.1) and then turn to adjusted frequencies based on corpus parts (in Section 2.2) as well as adjusted frequencies based on distances between successive occurrences (in Section 2.3).[4] Section 2.4 will then discuss a variety of problems of these measures.

2.1 Dispersion measures

The first dispersion measures to be mentioned here are general statistics and not specifically geared to the dispersion of linguistic items in texts but more often used as general indices of variation/dispersion:

(4) range: number of parts containing *a* (*x* times) = 5 (*x* is usually, but need not be, 1)

(5) max-min difference: $max(v) - min(v) = 4$

(6) standard deviation *sd*: $\sqrt{\frac{\sum_{i=1}^{n}(v_i - \bar{v})^2}{n-1}} \approx 1.581$ where $\bar{v} = \frac{f}{n} = 3$

(7) variation coefficient *vc*: $sd/\bar{v} \approx 0.527$

(8) chi-squared χ^2: $\sum_{i=1}^{n} \frac{(observed\ v_i - expected\ v_i)^2}{expected\ v_i} \approx 3.33$ where $expected\ v_i = s_i \cdot f$

Next, a few well-known "classics" from the early 1970s:

(9) Juilland et al.'s (1971) *D*: $1 - vc/\sqrt{n-1} \approx 0.736$

(10) Rosengren's (1971) *S*: $\left(\frac{1}{n}\left(\sum_{i=1}^{n}\sqrt{v_i}\right)^2\right) \cdot \frac{1}{f} \approx 0.937$ where min S = $1/n = 0.2$

(11) Carroll's (1970) D_2: $\left(\log_2 f - \left(\left(\sum_{i=1}^{n} v_i \log_2 v_i\right) \cdot \frac{1}{f}\right)\right) \cdot \frac{1}{\log_2 n} \approx 0.926$

A measure from the domain of information retrieval that was proposed at around the same time and is conceptually similar to an adjusted frequency measure that was proposed by Engvall is inverse document frequency *idf* (cf. Spärck Jones (1972) and Robertson (2004) for discussion). It is computed as is shown in (12).

(12) *idf*: $\log_2 \frac{n}{\text{number of parts containing } a} = 0$

Lyne's (1985:129, n. 2) D_3 is a preliminarily suggested measure that is based on the chi-square measure for quantifying dispersion:

(13) D_3: $\frac{1-\chi^2}{4f} \approx 0.944$

Much more recently, Zhang et al. (2004) proposed a measure called Distributional Consistency *DC,* which is defined as in (14) and which, in this artificial example, yields the same result as Rosengren's *S*.

(14) Distributional consistency *DC*: $\left(\frac{1}{n}\sum_{i=1}^{n}\sqrt{v_i}\right)^2 \approx 0.937$

The final two measures discussed in this section are rather different from the others in how corpus parts are involved and bridge the gap to measures dealt with further below. First, Quasthoff (2007) proposed a measure on what he calls fractal dimensions of words (and which I will call *FD* here).[5] His approach is very interesting and different from all other ones mentioned so far since while it is based on corpus parts, he does not assume an existing division of the corpus into parts (on the basis of registers, files, etc.). Rather, he appears to propose to divide the corpus into successively smaller parts and count in how many of these parts *a* is attested. For example, he discusses a case where a word occurs in 2 out of $n = 3$ parts, in 4 out of $n = 9$ parts, in 8 out of $n = 27$ parts, etc. From this unpublished ms., however, it is unclear (to me) how exactly the number of divisions *n* is determined and which of the resulting quotients in (15) is then chosen or whether some kind of average is computed, which is why I cannot apply this formula to the present example corpus.

(15) *FD*: $\frac{\text{log number of parts containing } a}{\text{log number of parts}}$ for increasingly larger numbers of parts

Second, Washtell (2007) is an interesting approach based on a measure of spatial dispersion from geography. On the one hand, his approach involves the division of the corpus into parts and counting how often *a* occurs in each of the *n* files, just like all others discussed in this section. On the other hand, his approach does not settle for just using the frequencies with which *a* is observed in each corpus part, but also their distances to each other within these parts. More specifically, in those corpus parts in which *a* is observed more than once, he also utilizes the distances between the different occurrences of *a* such that each occurrence's minimal distance to another occurrence is determined and used within the numerator of the formula in (16).

(16) Washtell's (2007:38) Self Dispersion: $\frac{\frac{1}{g}\cdot\sum_{i=1}^{g}\left(\frac{1}{dist_{a\Leftrightarrow a,i}}\right)}{2\cdot f/l}\approx 0.996$

where g is "the number of instances of a in the corpus (minus those instances which occur only once in a corpus part)" (Washtell, p.c., July 29, 2008), i.e., $g=\sum_{v>1} v$

2.2 Adjusted frequencies based on corpus parts

In the section, I will discuss a few adjusted frequencies that are based on the frequencies of elements in different parts of the corpus. Some of the probably best-known adjusted frequencies are derived from the classic dispersion measures by Juilland et al., Rosengren, and Carroll:

(17) Juilland et al.'s (1971) usage coefficient U: $D\cdot f\approx 11.047$

(18) Rosengren's (1971) Adjusted Frequency AF:[6] $S\cdot\frac{f}{n}\approx 14.053$

(19) Carroll's (1970) U_m: $\left(\sum_{i=1}^{n}\sqrt{v_i}\right)\cdot D_2+(1-D_2)\cdot\frac{f}{n}\approx 14.108$

According to Lyne (1985:101), an adjusted frequency measure suggested by Engvall (1974: 46–55) simply multiplies a's observed frequency with the percentage of corpus parts in which a is observed:

(20) Engvall's measure: $f\cdot$ percentage of parts containing $a=15$

Kromer's (2003) proposes what he considers a psycholinguistically more adequate measure, the psychophysically relations-based usage measure U_R:

(21) Kromer's (2003) U_R: $\sum_{i=1}^{n}(\psi(v_i+1)+C)=8.7$ where $C=0.577215665$.[7]

It must be pointed out, though, that Kromer unfortunately only claims that his measure is psycholinguistically more appropriate — "[i]t is accepted as a working hypothesis that while reading the text, the subjective feeling, caused by a specific word with frequency F, is determined by Equation (6)" (2003:180) — but he does not provide any corroborating psycholinguistic evidence for this claim.

2.3 Adjusted frequencies based on distances

An alternative approach to adjusted frequencies is not based on the frequencies of occurrence of a in different parts of the corpus but solely on the distances between

successive occurrences of *a* in the corpus. For example, Savický and Hlaváčová's (2002) approach to adjusted frequencies differs from nearly all other approaches so far — the exception is Washtell (2007). They propose the three following measures, for which we need another variable: d_1 to d_f refer to the distances between occurrences of *a* in the corpus after the corpus has been shifted such that an occurrence of *a* is the last element of the corpus. In this case, the distances d_1 to d_{15} become (2, 10, 2, 9, 2, 5, 2, 3, 3, 1, 3, 2, 1, 3, 2), which is used for the following measures:

(22) ARF and f_{ARF}: $\frac{f}{l}\sum_{i=1}^{f}\min\left\{d_i, \frac{l}{f}\right\} = 10.8$

(23) AWT: $\frac{1}{2l}\left(l + \sum_{i=1}^{f} d_i^2\right) = \frac{1}{2}\left(1 + \frac{1}{l}\sum_{i=1}^{f} d_i^2\right) = 3.18$ and f_{AWT}: $\frac{1}{l^2}\cdot\sum_{i=1}^{f} d_i^2 \approx 9.328$

(24) ALD: $\frac{1}{l}\sum_{i=1}^{f} d_i \log_{10} d_i \approx 0.628$ and f_{ALD}: $l\cdot 10^{-ALD} = \exp\left(-\sum_{i=1}^{f}\frac{d_i}{l}\ln\frac{d_i}{l}\right) \approx 11.764$

Savický and Hlaváčová (2002:228f.) conclude that, on the whole, *ARF* is the most stable of the three measures, but do not compare their measures to other adjusted frequencies.

2.4 Problems of existing measures

Unfortunately, several of these coefficients — and from now on I will concentrate more on the dispersion measures — suffer from some problems, which will be discussed in the following sections. The main objective of this discussion is not to single out particular measures for criticism — the main objective is to showcase some problems in such a way as to highlight important dimensions of the measures which are in need of further exploration and which may be taken into account for the construction of new (better) measures.

2.4.1 *Problems of parts-based measures*

2.4.1.1 *The sizes of corpus parts.* One problem is that some parts-based measures require the corpus parts for which a dispersion measure is computed to be identically large.[8] However, this is usually not true because corpus files, genre-based corpus parts, or parts from any other easy and/or meaningful corpus divisions are usually not equally-sized, and creating equally-sized corpus parts can be practically difficult and will likely conflate corpus parts that should not be conflated. For example, since the 500 files of the ICE-GB are not exactly equally large, the dispersion values in Gries (2006) can only be interpreted heuristically. A related

problem is that while some measures do not strictly speaking require equally-sized corpus parts — *idf* is one such example — that also means that potentially relevant information about corpus sizes, which will be discussed below, cannot be figured into the computation meaningfully.

For some measures, workarounds have been proposed (which then usually also apply to the adjusted frequencies derived from the respective dispersion measures). Juilland et al.'s *D*, for example, can be adjusted for unequal corpus sizes by computing *vc* not on the basis of the standard deviation of the observed frequencies, but of the standard deviation of the observed relative frequencies. The computation of the mean and the *sd* for Juilland et al.'s *D* changes to (25), the results of which are then inserted into (7), and then into (9). In the case of equally-sized parts s_1 to s_5, the adjusted *D* is of course identical to the unadjusted *D*. A similar adjustment is available for Rosengren's *S*, which changes its formula into (26). Interestingly, these proposed adjustments have hardly been explored, it seems, as they are not even usually referred to in current textbooks or research articles (cf. again note 8).

(25) *sd* for Juilland et al.'s (1971) D_{adj}: $\sqrt{\frac{\sum_{i=1}^{n}\left(\frac{v_i}{s_i}-\bar{v}\right)}{n-1}} \approx 7.906$ where $\bar{v}=\frac{\sum_{i=1}^{n}\frac{v_i}{s_i}}{n}=15$

(26) Rosengren's (1971) S_{adj}: $\frac{(\sum_{i=1}^{n}\sqrt{s_i \cdot v_i})^2}{f} \approx 0.937$ where min $S = 1/n$

2.4.1.2 *Dispersion measures and their defined limits.* Some proposed dispersion measures vary widely in the range of values they may take on. For instance, the range, the standard deviation, the variation coefficient, and chi-square are — unlike some other values — not normalized to fall into a particular convenient range such as 0 to 1, which makes it difficult to compare values across different studies.

A related but smallish issue one may find undesirable is that with five equally-sized corpus parts, all measures cannot get the theoretical maximum value for (2, 2, 2, 2, 1). For example, Carroll's D_2 for the above distribution is 0.967 rather than the theoretical maximum of 1 although there is of course no way that 9 occurrences of a word could be more evenly distributed across 5 equally-sized different corpus parts.

2.4.1.3 *Dispersion measures outside of the defined limits.* Another set of problems, which I have not seen discussed so far, is concerned with the fact that some dispersion measures can take on values outside of their intended range. Both Juilland et al.'s *D* and Lyne's D_3 do not always result in a value within the expected range. They

are supposed to fall between 0 and 1, but as the reader can easily verify, D and D_3 for a word x that occurs only once in one out of six corpus files results in dispersion values of $D = -0.095$ and $D_3 = -0.25$ respectively.

2.4.1.4 *Dispersion measures and the number of corpus parts.* Third and relatedly, some measures' ranges are dependent on the number of corpus parts.[9] For example, in the above corpus, the word *e* occurs equally frequently in each of the 5 equally-sized corpus parts and the parts-based measures that are supposed to fall between 0 and 1 — Juilland et al.'s D, Carroll's D_2, Rosengren's S, DC — return their maximal value, 1, to indicate a maximally even distribution. However, x is maximally underdispersed in that it occurs only once in the corpus and that is problematic for some measures. The parts-based measures with the exception of Carroll's D_2 do not return their theoretical minimal value of 0 but 0.2, i.e., 1 divided by the number of corpus parts n, which means that the value decreases with increasing numbers of corpus parts. In addition, Juilland et al.'s D even breaks down and returns a negative value, which it is not supposed to return (cf. Section 4.1 below for more discussion).

2.4.1.5 *Lack of sensitivity.* Some measures appear to be not as sensitive to distributional differences as needed or desired, or too sensitive. As for a lack of sensitivity, measures may be too insensitive in the sense that they cannot pick up potentially relevant differences.[10] For example, as Francis and Kučera (1982:463), citing Muller (1965), discuss, Juilland et al.'s D does not distinguish between the following two distribution vectors for both of which $D = 0.526$: (4, 2, 1, 1, 0) and (3, 3, 2, 0, 0). In addition, as Rosengren (1971:117) himself has shown, Rosengren's S does not distinguish between the pairs of the vectors in the following two distributions (4, 4, 4, 4, 0) and (9, 4, 1, 1, 1) (both $S = 0.8$) as well as (9, 9, 4, 0, 0) and (16, 4, 1, 1, 0) (both $S = 0.582$) although in both cases intuitively the latter member of each pair is less equally spread (assuming equal corpus sizes, that is); the same is actually true of DC.

Second, measures may be too insensitive such that they uniformly output their extreme value when all occurrences of a word a are in one and the same corpus part irrespective how large that corpus part is compared to the others. In an admittedly hypothetical extreme case, one of three corpus parts may account for 98% of the corpus and if then all instances of a occurred in that part, this would be the expected natural case: the tokens of a are found in a large part of the corpus, which is what large dispersion is all about. (Parts-based measures, but not necessarily distance-based measures, would still face the problem that if the sizes of corpus parts are very heterogeneous as in this example, then measures based on corpus parts cannot take the dispersion of the expression in question within the

large corpus part(s) into account.) By contrast, if all occurrences of *a* occurred in one small part, this would be very underdispersed/clumpy, which is why the sizes of the corpus parts must be figured into the equation (cf. Section 4.1 below for exemplification).

2.4.1.6 *Oversensitivity.* As for being oversensitive, Lyne (1985:107–9), for example, shows how Juilland et al.'s *D* is impartial about zeroes — i.e., corpus parts not containing the word *a* in question — while Carroll's D_2 penalizes them and Rosengren's *S* penalizes them even more strongly, which Lyne considers objectionable on the basis of his data.[11] Also, some measures such as max-min difference, the standard deviation *sd*, or chi-square are very sensitive in the sense that few extreme values or very small expected frequencies can distort results considerably. (This is a well-known disadvantage that rules out the use of chi-square in many corpus-linguistic situations.)

2.4.2 *Problems of distance-based measures*

Given the different nature of the distance-based measures, the problems they come with are somewhat different from those of parts-based ones. For example, under certain (admittedly marginal) circumstances and from a certain perspective, the measures by Savický and Hlaváčová (2002) can lack sensitivity. Let us assume a small corpus containing one file with spoken dialog of two interlocutors in which each interlocutor produces exactly one ten-word sentence on each turn. Let us further assume that one interlocutor consistently produces word *a* at position 5 in each turn and the other one does not. In this case, the measures proposed by Savický and Hlaváčová (2002) would, since they are based on distances between successive occurrences and not on corpus parts, yield the result that *a* is perfectly evenly dispersed in the corpus: all distances are 20. A parts-based measure, on the other hand, is theoretically able to see that the one file consists of two parts — two parts each of which contains all sentences for each interlocutor — and that *a* only shows up in the part for one interlocutor and accordingly output a more adequate value. Similar comments apply to corpora with different genre parts etc.[12]

Another issue is the question of how such measures handle different corpus parts, which results in a real catch-22 for these measures. For example, in the above corpus example, the computation of Savický and Hlaváčová's measures will simply ignore corpus parts and their distances may therefore in fact cross document boundaries. That may be disadvantageous in the sense that it is linguistically counterproductive, but it may also be advantageous in the sense that their measure is more widely applicable as it avoids the methodological problems Washtell's approach runs into. His measure is linguistically more intuitive as it respects document boundaries, but then methodologically more problematic because it cannot

handle many lower-frequency elements: since his measures require within-part neighbors for its computation, by definition it must disregard all instances where *a* occurs just once in a part. Thus, it cannot even be applied at all to distributions where the element in question occurs just once in the corpus (cf. *x* in the above example where it should output most extreme underdispersion) or where the element in question occurs just once in each corpus part (cf. *e* in the above example), and even when the element in question occurs in corpus parts more than once, the measure still becomes more unreliable the more occurrences of the element in question are the only ones in their respective parts.

Then, unlike parts-based measures, distance-based measures can of course be sensitive to order effects. They output different values both when the corpus parts are arranged differently or when the words in the corpus parts are arranged differently. For example, Savický and Hlaváčová's adjusted frequencies, but not Washtell's dispersion measure, of *a* in the above example will change if the five corpus parts are arranged in a different order. This is not necessarily desirable: in the case of the BNC, this means that the adjusted frequencies computed for the BNC would only be comparable if everybody used the same ordering of the files, which in the case of the BNC would probably mean one would have to stick to, for example, the arbitrary order of file names. Similarly, if the order of the five parts of the corpus above remains the same but the three occurrences of *a* in the third part are moved to the end of the third part and the four occurrences of *a* are moved to the beginning of the fourth part, the measure changes drastically, too. Again, this may not be desirable because on the level of granularity of the five corpus parts nothing has changed in the data. On the other hand, one may just as well argue that the finer resolution of their measures is intended to detect and reflect such changes. Ultimately, the decision appears to boil down to whether one would be willing to treat a corpus as one homogeneous string of words devoid of any structure (in the form of turns, file parts, files, genre/register parts, etc.) or not.

2.5 Interim summary

In sum, it seems as if there are few if any dispersion measures that provide unproblematic measures for equally- and unequally-sized parts. In his comparative review of the classics — *D*, D_2, and *S* — Lyne (1985:117) concludes that *D* is the overall most adequate measure. In the previous section, however, I hope to have already shown that there are many issues to be considered when it comes to evaluating dispersion measures and adjusted frequencies and that, sometimes at least, (some of) these issues can counteract each other.

In the following section, I will propose for discussion a conceptually very simple alternative measure *DP* (for deviation of proportions), which (i) allows to

quantify the dispersion of lexical items just like the above, (ii) does not rely on the unwarranted assumption of equally-sized corpus parts, (iii) is, as I see it, neither too nor too little sensitive, (iv) is not a measure of statistical significance and thus avoids theoretical problems of the hypothesis-testing paradigm,[13] and (v) theoretically at least ranges from 0 to 1. In the following section, I will explain how *DP* is computed and how it behaves in certain distributionally basic and/or interesting situations. Then, I will exemplify this measure *DP* on the basis of dispersions of words from different frequency bands and with different degrees of dispersion in the British National Corpus Sampler.

3. An alternative measure of dispersion: *DP* / *DPnorm*

To determine the degree of dispersion *DP* of word *a* in a corpus with *n* parts, one needs to take three simple steps.

i. Determine the sizes s_{1-n} of each of the *n* corpus parts, which are normalized against the overall corpus size and correspond to expected percentages which take differently-sized corpus parts into consideration
ii. Determine the frequencies v_{1-n} with which *a* occurs in the *n* corpus parts, which are normalized against the overall number of occurrences of *a* and correspond to an observed percentage.
iii. Compute all *n* pairwise absolute differences of observed and expected percentages, sum them up, and divide the result by two. The result is *DP*, which can theoretically range from approximately 0 to 1, where values close to 0 indicate that *a* is distributed across the *n* corpus parts as one would expect given the sizes of the *n* corpus parts. By contrast, values close to 1 indicate that *a* is distributed across the *n* corpus parts exactly the opposite way one would expect given the sizes of the *n* corpus parts.

Let me illustrate this on the basis of a set of fictitious distributions. Imagine a corpus consisting of three 200-word parts, i.e. 600 words. Imagine further one is interested in a word *a* that occurs 9 times in the corpus, 3 times in each of the three corpus parts. In this case, the computation of the three steps can be summarized as in Table 2. Step 1 results in the leftmost column: if *a* is distributed as one would expect given the sizes of the *n* corpus parts, *a*'s frequency in each file should be one third of its overall frequency in the corpus: $^{200}/_{600} = 0.33$. Step 2 results in the second column from the left: in each row, i.e. for each corpus part, $^{3}/_{9} = 0.33$. Step 3 requires to compute the *n* row-wise absolute differences (shown in the third column), sum them up (shown in the fourth column), and divide by 2; the result is

DP. The result in the rightmost column shows that *a* is distributed perfectly evenly in the corpus, namely in exact accordance with how the corpus parts look like.

Table 2. Computation of *DP*; example 1

Step 1	**Step 2**	**Step 3**		
Expected %	Observed %	Abs. difference	Sum of abs. diff.	Divide by 2
0.33	0.33	0		
0.33	0.33	0	0	0
0.33	0.33	0		

For comparison, imagine now the same corpus, but the occurrences of *a* are all found in one of the *n* equally-sized corpus parts. The computation changes as represented in Table 3.

Table 3. Computation of *DP*; example 2

Step 1	**Step 2**	**Step 3**		
Expected %	Observed %	Abs. difference	Sum of abs. diff.	Divide by 2
0.33	1	0.67		
0.33	0	0.33	1.33	0.67
0.33	0	0.33		

Note here one important characteristic in which *DP* differs from some measures, a characteristic I mentioned briefly above. In a case like the one shown in Table 3, *DP* and the other standard parts-based measures output their extreme values since all occurrences of *a* are in one corpus part. However, the extreme value is not the theoretical minimum 0 or the theoretical maximum 1, as I mentioned above in Section 2.4.1.2. One might criticize *DP* for this characteristic, but my response to this would be that, first, as shown above all the classic parts-based measures with the exception of Carroll's D_2 behave the same way so if this is a valid point of critique, then it applies to more than just *DP*. Second, I do not know what other scholars' motivation for designing their dispersion measures were, but with regard to *DP* the idea is to have it not output its theoretically maximal value here because the size of *DP* accounts for the fact that while all occurrences of *a* do occur in one and the same part, a particular proportion of *a* was expected to occur in there anyway. Third, this issue of course only arises noticeably with very small numbers of corpus parts *n*.

Given this legitimate concern, however, let me clarify related aspects of *DP*'s behavior. Imagine a corpus whose parts' sizes are extremely heterogeneous. It consists of three parts, two of which each account for 1% of the corpus while the last part accounts for the remaining 98% of the corpus. However, the first corpus

part contains 98% of the occurrences of *a*, while the remaining two corpus parts each contain only 1% of all occurrences of *a*. In this case, *a* is obviously extremely underdispersed since just about all of its occurrences are only in one small part of the corpus and any dispersion measure should reflect this. The resulting computation for *DP* is summarized in Table 4 and yields the high value of *DP*=0.97, which reflects that the distribution of *a* is very uneven in the sense of being completely at odds with the sizes of the corpus parts.

Table 4. Computation of *DP*; example 3

Step 1	**Step 2**	**Step 3**		
Expected %	Observed %	Abs. difference	Sum of abs. diff.	Divide by 2
0.01	0.98	0.97		
0.01	0.01	0	1.94	0.97
0.98	0.01	0.97		

A nearly opposite kind of distribution is shown in Table 5. While the proportions of the corpus parts are the same as in the previous example — 0.98, 1, 1 — now all occurrences of *a* are in the largest part. This means that *a* is well dispersed because it is spread out nicely across most of the corpus (at least in the highly artificial way the corpus parts are defined here), which is indicated by *DP*'s correspondingly low value.

Table 5. Computation of *DP*; example 3

Step 1	**Step 2**	**Step 3**		
Expected %	Observed %	Abs. difference	Sum of abs. diff.	Divide by 2
0.01	0	0.01		
0.01	0	0.01	0.04	0.02
0.98	1	0.02		

Note that others' dispersion measures which take the sizes of corpus parts into account also return values that mark the distribution in Table 5 as well dispersed: the occurrences of the word in question are simply exactly where they would be given an equal distribution.[14]

Let me now briefly return to a shortcoming of some measures mentioned above and discuss two examples where the expected frequencies are more realistic than in the above examples, which serves to highlight the behavior of *DP* in extreme situations. In the following two examples, the corpus again consists of three parts, which make up 45%, 35%, and 20% of the corpus. First, consider the case where all occurrences of *a* are in the first corpus part, which is the largest of the three parts; the computation in Table 6 results.

This is interesting in comparison with Table 7, which represents the case in which all occurrences of *a* are in the second corpus part, which is the second largest. As Table 7 shows, *DP* is now higher than in Table 6. This is relevant because — as before — many measures such as *D*, D_2, *DC*, and *idf* would not reflect that difference but would rather assign the same extreme value to both distributions. By contrast, *DP* reflects this difference because it takes into consideration the fact that, while both distributions are extreme, the second one in Table 7 is more extreme because it has the same observed percentage of 100%, but in an even smaller part of the corpus.

Table 6. Computation of *DP*; example 4

Step 1	**Step 2**	**Step 3**		
Expected %	Observed %	Abs. difference	Sum of abs. diff.	Divide by 2
0.45	1	0.55		
0.35	0	0.35	1.1	0.55
0.2	0	0.2		

Table 7. Computation of *DP*; example 5

Step 1	**Step 2**	**Step 3**		
Expected %	Observed %	Abs. difference	Sum of abs. diff.	Divide by 2
0.45	0	0.45		
0.35	1	0.65	1.3	0.65
0.2	0	0.2		

Before we look at some real data, let me point out a few other positive aspects of *DP* apart from its ability to distinguish between different degrees of "most extreme distributions" and also propose a normalization to *DP*. First, unlike Juilland et al.'s *D* or Lyne's D_3 it cannot even in the most extreme distributions fall outside of the range of 0 to 1. Second, *DP* can also distinguish distributions that some other measures cannot. Assuming equal sizes of corpus parts and looking at the distributions Juilland et al.'s *D* treated equally, *DP* for (4, 2, 1, 1, 0) is 0.35 whereas *DP* for (3, 3, 2, 0, 0) is 0.4. The same is true of the distributions Rosengren's *S* and *DC* do not distinguish: assuming equal sizes of corpus, *DP* for (4, 4, 4, 4, 0) is 0.2 whereas *DP* for (9, 4, 1, 1, 1) is 0.4125; similarly, *DP* for (9, 9, 4, 0, 0) is 0.4182 whereas *DP* for (16, 4, 1, 1, 0) is 0.5273. Thus, not only can *DP* distinguish all these pairs of distributions, but the value for the latter distributions of the two pairs scores a higher value reflecting a less homogeneous distribution. Third, while *DP* has a greater discriminatory power than even adjusted *D*, it shares with *D* the property of not necessarily penalizing zeros overly strongly, an undesirable characteristic of D_2 and *S* according to Lyne. Fourth, an observation for which I am grateful to

Petr Savický: *DP* has the attractive characteristic that, if a part of the corpus is split into, say, two parts of smaller sizes in some proportion and the occurrences of the considered word split in the same proportion, then *DP* does not change. Finally, while *DP* can also not attain its maximum value of 1 for the distribution (2, 2, 2, 2, 1) with equally-sized corpus parts, just like *D* and *S* and unlike D_2, it can when then sizes of the corpus parts are proportional to (2, 2, 2, 2, 1).

Finally, let me return briefly to the issue of the range of dispersion values. While I have argued in favor of *DP* as it is, those who prefer a dispersion measure that returns the theoretically possible minimal and maximal values (which Carroll's D_2 does) and that is at the same time as theoretically simple as *DP* (by being based on simple percentage differences), there is a simple normalization step that changes *DP*'s behavior with respect to this issue:

iv. Divide *DP* by $1 - (1/n)$ to yield DP_{norm}.

If this step is added to the computation exemplified in Table 3, the computation changes to that represented in Table 8. Note, first, how DP_{norm} now takes on the maximal value of 1 but at the same time does not anymore account for the fact that one third of the occurrences were expected in a corpus part that makes up one third of the corpus. Note, second, that as the numbers of corpus parts increase, the impact of this normalization will decrease (since $1/n$ will become smaller).

Table 8. Computation of *DPnorm*; example 1

Step 1	**Step 2**	**Step 3**			**Step 4**
Expected %	Observed %	Abs. difference	Sum of abs. diff.	Divide by 2, =*DP*	Divide by $1-1/n$
0.33	1	0.67			
0.33	0	0.33	1.33	0.67	1
0.33	0	0.33			

Let us now study the behavior of *DP* when applied to real data from the BNC Sampler.

4. Applications

In this section, I will look at three small case studies to explore *DP*'s behavior when applied to real data. Section 4.1 explores results when *DP* is applied to simple frequencies of occurrence while Section 4.2 explores *DP*'s performance with regard to co-occurrence frequencies.

4.1 The dispersion of words: *DP*'s results

In this section, I will discuss results from using *DP* to look at a pseudorandom sample of words from the 2-million word BNC Sampler.[15] I first generated a frequency list of the BNC Sampler and then chose 68 words from five different frequency bands; cf. Appendix 1 for the list of words included in the analysis as well as a precise characterization of how the words were obtained. For each of these words, I computed all the dispersion measures from Section 2 above as well as *DP* and DP_{norm} using an R function, which is available from my website for readers to use; cf. Section 5.3 below for links and explanations. The corpus parts I assumed were the individual files. In what follows, I will discuss the results of this analysis (if, for reasons of space, only summarily).

First, some general descriptive information. Given the above sampling procedure, it is reassuring to see that, like all other measures, *DP* exhausts nearly the complete range of possible values, as can be seen clearly in the small sample of dispersion measures presented in Section 2. (For measures where adjustments for unequally-sized corpus parts are available, only these are provided). (Note in passing how again D_3 yields many negative values outside of the range into which it is supposed to fall.) In addition to the comparable spread of values, Table 10 shows that *DP* also behaves "well" when it comes to the words scoring the highest, most intermediate, and lowest values. The first three columns list a well-known common set of function words and light verbs; the last three columns list words many of which I have never seen before; the three middle columns list words which I think most would intuitively agree are certainly well-known to all native speakers and advanced learners of English, but which also one would not necessarily expect to see everywhere and evenly everywhere.

I hope I have been able to show in this section that *DP* does what it is supposed to do: when applied to a random frequency-stratified sample of words from the BNC Sampler, it nearly fully exhausts the possible range of values,[16] provides intuitively very reasonable output when high, intermediate, and low *DP* values are inspected, and it is overall consistent with some of the best-known dispersion

Table 9. Descriptive summary statistics for selected dispersion measures

Statistic	*DP*	*D*	D_2	*S*	D_3	*DC*	*idf*
Minimum	0.08	−0.0027	0	0.0014	−44.75	0.0054	0
1st quartile	0.2668	0.6424	0.4415	0.0651	−3.6002	0.0543	0.0829
Median	0.6187	0.8384	0.7526	0.3499	−0.1143	0.3038	1.593
Mean	0.5965	0.6755	0.6126	0.4463	−9.2311	0.4081	2.46
3rd quartile	0.9252	0.9424	0.9256	0.8627	0.7793	0.7771	4.2016
Maximum	0.9986	0.9809	0.9684	0.9886	0.918	0.9196	7.5256

Table 10. The words with the fifteen maximal, most intermediate, and minimal *DP* values

Minimal *DP's*			**Intermediate *DP's***			**Maximal *DP's***		
Word	*DP*	Freq	Word	*DP*	Freq	Word	*DP*	Freq
a	0.08	39,119	*definition*	0.795	102	*macari*	0.999	10
to	0.103	46,187	*includes*	0.716	102	*mamluks*	0.998	10
and	0.106	53,216	*thousands*	0.714	102	*lemar*	0.996	10
with	0.155	11,138	*plain*	0.709	102	*sem*	0.994	10
but	0.158	10,569	*formal*	0.708	102	*hathor*	0.994	10
in	0.159	32,198	*anywhere*	0.645	102	*tatars*	0.989	10
not	0.165	9,211	*properly*	0.625	102	*scallop*	0.989	10
this	0.166	9,651	*excuse*	0.612	102	*malins*	0.988	10
the	0.168	104,248	*hardly*	0.585	102	*ft*	0.986	102
have	0.178	11,928	*er*	0.556	9,721	*defender*	0.98	10
be	0.207	12,735	*each*	0.474	1,007	*scudamore*	0.98	10
are	0.223	9,770	*lot*	0.472	1,032	*pre*	0.945	10
that	0.227	29,280	*house*	0.453	1,024	*diamond*	0.941	102
there	0.243	9,243	*tell*	0.414	1,023	*carl*	0.938	102
of	0.249	44,276	*came*	0.412	1,013	*proclaimed*	0.934	10

measures. In the following section, I will briefly show how *DP* can also be applied to co-occurrence information of the kind exemplified in Section 1.

4.2 The dispersion of words in constructions/patterns: *DP*'s results

In Section 1, I used the example of the verbs attracted to the imperative in the ICE-GB to point out that, contrary to what a look at the literature might make us believe, the issue of dispersion is also highly relevant to co-occurrence data: statistics based on frequencies and co-occurrence frequencies — in the above example, $p_{Fisher\text{-}Yates\ exact}$-values as measures of collocational attraction/repulsion — can suffer from very similar problems as raw frequencies of occurrence alone. Whatever dispersion measure one adopts would therefore ideally be extendable to handle co-occurrences and their dispersion. In this section, I will discuss two examples in which the logic underlying *DP* is applied to co-occurrences. In Section 4.2.1, I will briefly return to the imperative construction mentioned above; in Section 4.2.2, I will then look at the ditransitive pattern.

4.2.1 *The imperative in the ICE-GB*

As a first example, I will very briefly revisit the imperative in the 1-million word ICE-GB as discussed by Stefanowitsch and Gries (2003). This is how I retrieved the data for this case study:

- using the fuzzy-tree fragment search facility of ICE-CUP, I generated and saved a concordance of all instances of auxiliaries used in the imperative (category: aux, feature: imp) as well as a concordance of all instances of imperative clauses (category: CL, feature: imp) containing a verb phrase containing a verb in the imperative (category: V, feature: imp);
- then, I retrieved all the auxiliary and verb forms from above and generated a table the columns of which contain all aux/verb lemmas attested in the imperative, the rows of which contain all corpus files, and each cell of which states how much in percent of all imperative occurrences of this column's verb occurs in this row's corpus file. To clarify this a little, the verb lemma *trust* is used in the imperative three times (once in S1A-028, once in W1B-004, and once in W2D-009). Thus, the column for *trust* in this table contains 497 zeros (one for each file in which *trust* is not attested in the imperative) and three times ⅓, namely in the rows for the above files. Thus, this table contains all observed percentages.
- I then created a vector that lists for each file the number of verb and aux tokens it contains and converted that into a vector of percentages (that sum to 1). Thus, this vector states how many verbs in percent each file contributes to the corpus; i.e., this vector contains the expected percentages.

The computation of *DP* for each verb was then performed as above: for each verb, I subtracted the 500 observed percentages from the 500 expected proportions, summed the absolute differences and divided the sum by 2. The results are rather clear: the 20 verbs with the lowest and highest *DP*-values are given in (27) and (28) as is the *DP*-value for the verb *fold* that was discussed above in Section 1; *DP-v*alues are listed in parentheses.

(27) *let* (0.676), *see* (0.845), *look* (0.885), *take* (0.882), *go* (0.886), *come* (0.912), *try* (0.94), *tell* (0.914), *get* (0.918), *be* (0.929), *make* (0.946), *have* (0.925), *do* (0.949), *remember* (0.942), *put* (0.952), *give* (0.957), *keep* (0.948), *say* (0.948), *use* (0.969), *ask* (0.96)

(28) *recall, resolve, cancel, cf, manipulate, employ, chips, enquire, unmask, love, fly, contrast, scrape, lend, melt, process, break, pat, chill, season (*all *DP*≈0.999; *fold*'s *DP*=0.998, too)

Two things are noteworthy. First, the results are well correlated with Stefanowitsch and Gries's (2003: Section 3.3.2) study of verbs that are typical for the imperative, and they make intuitive sense when, as we usually do in corpus linguistics, inspect the results in terms of the ranking of elements they provide, the prevalent practice in our interpretation of, say, measures of collocational or collostructional strength. For example, verbs that we would expect to occur without much con-

straint in the imperative construction show up first, esp. *let*. The verb *see* is also quite natural there, given how *see* is often used sentence-initially as a discourse marker and many other contexts: *See what's on the other side*, *See what it says here*, *See, and all because …*, *See leaflets NI 230 Unemployment Benefit …*, etc, etc. Most following verbs are ones that would not be surprising to find in an imperative. Another example is *have*: examples include *Have a seat*, various versions of *Have some* [FOOD], *Have a* [MODIFIER] *birthday!* etc. Some other verbs in (27) are similarly widespread in the imperative and are arguably in the process of becoming discourse markers: *tell* and *say* (as in intonation unit-initial *tell me, …!* or *say, …!*), or *remember* (cf. Tao 2001). Similarly, the verbs in (28) are certainly ones whose distribution in the imperative we would not necessarily expect to be particularly regular. Note in particular that the problematic cases of *process* and *fold* score extremely high values, which reflects the fact that they are only attested in imperatives in a single file, just as we would want a dispersion measure to detect.

Second, the *DP*-values do not extend across the whole range of values (from 0 to 1) anymore. Rather, the values are rather high and approach 0.9 and higher fast. However, this is less reason for concern than one might think. First, there is a long tradition in corpus linguistics to evaluate distributional statistics (such as collocational strengths) in terms of the ranking of words in comparison to other words rather than their absolute values in isolation, and the ranking has enough discriminatory power to provide meaningful results out of all 387 verbs whose dispersion was included. Second, the *DP* values become large fast because the observed frequencies in this small corpus become small fast: the verb with the 20th-highest *DP*-value occurs only 27 times in the imperative in the corpus. Third and relatedly, this is exactly the reason why some other dispersion measures produce results very similar to *DP*. (Note in passing that Juilland et al.'s *D* and Lyne's D_3 again do not fall within the range they are supposed to fall.)

- results for *let*: Carroll's $D_2 = 0.766$, Rosengren's $S = 0.288$, Lyne's $D_3 = -0.26$, $DC = 0.27$, Juilland et al.'s $D = 0.904$;
- results for *expect*: Carroll's $D_2 = 0.112$, Rosengren's $S = 0.004$, Lyne's $D_3 = -61.25$, $DC = 0.004$, Juilland et al.'s $D = 0.292$;
- results for *fold*: Carroll's $D_2 = 0$, Rosengren's $S = 0.002$, Lyne's $D_3 = -123.75$, $DC = 0.002$, Juilland et al.'s $D = -0.001$.

In sum, these results suggest that *DP* can be applied to identify underdispersed words in co-occurrence relations, an issue which we will also look at in the following section.

4.2.2 *The ditransitive in the ICE-GB*

As a second example, I will look at verbs in the ditransitive (or double object construction) in the ICE-GB. The data for this case study were retrieved and processed in virtually the same as those for Section 4.1: using the fuzzy-tree fragment search facility of ICE-CUP, I generated and saved a concordance of all instances of verbs used ditransitively (category: 'verb', feature: 'ditr'); retrieved all verb forms from the marked matches together with the file in which they occurred, lemmatized all the verb forms and generated the same kind of table as before, and computed *DP* as before. The results are rather clear: the 20 verbs with the highest and lowest *DP*-values are given in (29) and (30); *DP*-values are listed in parentheses.

(29) *give* (0.43), *tell* (0.51), *ask* (0.83), *show* (0.86), *send* (0.9), *offer* (0.92), *get* (0.95), *cost* (0.96), *allow* (0.96), *teach* (0.97), *convince* (0.97), *remind* (0.97), *inform* (0.97), *buy* (0.97), *do* (0.97), *take* (0.98), *pay* (0.98), *promise* (0.98), *warn* (0.98), *lend* (0.98)

(30) *render, deal, permit, profit, vote, feed, file, sell, instruct, prescribe, keep, command, draw, rent, loan, bet, supply, build, cut, overpay* (all $DP \approx 1$)

On all three points discussed above, we basically get the same kind of results as before: again, the results correspond well to previous corpus-based work on ditransitives and their semantics, and the result makes intuitive sense: verbs that we tend to associate with the ditransitive construction a lot (because their semantics are highly compatible; cf. Stefanowitsch & Gries 2003: Section 3.2.2) show up first and esp. *give* and *tell*, the ditransitive verbs par excellence, occupy the first two places (in that order), but also all following verbs are ones that would not be surprising to find in a ditransitive (*remind* or *inform*, for example, as in *remind me what age he was* or *informing him what I've done* respectively can be metaphorically construed as involving transfer of the requested information from the reminder to the remindee). Similarly, the verbs in (30) are certainly ones whose distribution in the ditransitive we would not expect to be particularly regular: for example, *sell* is more significantly more associated with the prepositional *to*-dative (cf. Gries & Stefanowitsch 2004: 106).

Second, the *DP*-values again do not extend across the whole range of values (from 0 to 1) for the same reason as above (the frequencies in this small corpus become small fast: the verb with the 20th-highest *DP*-value occurs only 12 times in the ditransitive in the corpus), but also again the rank ordering of the verbs is still sensible and, third, other dispersion measures produce very similar results:

– for *give*: Carroll's $D_2 = 0.889$, Rosengren's $S = 0.568$, Lyne's $D_3 = 0.633$, $DC = 0.551$, Juilland et al.'s $D = 0.946$;

– for *lend*: Carroll's $D_2 = 0.307$, Rosengren's $S = 0.017$, Lyne's $D_3 = -21.319$, $DC = 0.015$, Juilland et al.'s $D = 0.598$;
– for *accord*: Carroll's $D_2 = 0.177$, Rosengren's $S = 0.005$, Lyne's $D_3 = -40.417$, $DC = 0.006$, Juilland et al.'s $D = 0.422$.[17]

Again, *DP* allows for identifying the underdispersed words and outputs as most regularly dispersed verbs a range of words that fit both our expectation of the distribution and the semantics of the construction perfectly.

5. Conclusions and a (brief) outlook

In this paper, I have taken a few first steps towards a from my point of view overdue program of research. Dispersion and adjusted frequencies are an essential tool in a discipline that is so dependent on distributional data and, accordingly, in the past 30–40 years, a large number of dispersion measures and adjusted frequencies has been proposed. However, apart from Lyne's early work there is virtually no systematic exploration, comparison, or even comprehensive introduction of these kinds of statistics in both corpus linguistic research and textbooks. It is actually surprising to, on the one hand, see the wealth of immensely interesting literature on collocational statistics (much of which is unfortunately still underutilized), but, on the other hand, see that the distributional statistic of dispersion has remained relatively ignored in spite of the fact that particular kinds of dispersion can completely mess up even the most careful collocational statistics.

While I do not want to lay claim to having provided a great many solutions or earth-shaking observations, I did provide the most comprehensive overview of dispersion measures and adjusted frequencies as well as their characteristics, pros, and cons to date and I proposed a new measure of dispersion for the analysis of corpus data. While the latter may appear as just adding one more measure to an already sizable list of measures, I believe this measure has several appealing characteristics that make it worth considering:

– flexibility: *DP* is able to handle differently sized corpus parts;
– simplicity: *DP* is conceptually extremely simple and straightforward, as it is based on something anyone can understand immediately — differences between observed percentages of words and percentages that corpus parts make up of a corpus;
– extendability: *DP* can immediately be applied to other kinds of data / scenarios such as co-occurrence frequencies;
– high sensitivity: unlike some other measures, *DP* does not blindly output extreme values for extreme distributions but since it includes the expected pro-

portion of occurrences into the equation, it can distinguish cases where all observed occurrences of a word are in the smallest file from cases where all observed occurrences of a word are in the largest file; also, *DP* can distinguish distributions other measures fail to distinguish;
– not too high sensitivity: unlike some other measures, at the same time, it does not over-penalize zeros, does not output extremely high values when low expected frequencies come into play, and is insensitive to the order of corpus parts.

However, given the present state of the art, much remains to be done, and the following sections mention a few such issues.

5.1 The integration of frequencies and dispersion

The first area requiring work is about how to best integrate frequency and dispersion information. The probably most widely-used approach so far has been to compute adjusted frequencies of the above-mentioned sort, where sometimes the adjusted frequency is just the product of the observed frequency and a dispersion value. However, to me at least it is often unclear what these values "mean": frequencies as such are straightforward to understand but the product of a frequency and some dispersion value has usually no such straightforward interpretation. I am not (yet) certain I am in a position to make a full-fledged proposal about how this issue can be addressed, but let me mention briefly one way of handling this: for many linguistic research purposes — as opposed to, say, lexicographical purposes where practical constraints of time and money dictate procedures — a two-dimensional representation of the kind of Figure 1 may be more useful. The product of a frequency and a dispersion value ultimately loses information — does the adjusted frequency result from $x \cdot y$ or from $10x \cdot 0.1y$? The two-dimensional plot, by contrast, preserves each word's frequency and dispersion and is thus more informative; cf. Figure 1 for a partial representation of the words investigated in Section 3.

5.2 Refining and normalizing measures

A second area of concern is that we need to explore in more detail how measures behave under different circumstances and how they should be tweaked, normalized, weighted etc. Let me discuss only one example here: most measures are based on corpus parts, but there is virtually no exploration of how differently-sized corpus parts can distort the picture. Above, we discussed the hypothetical case that, in a corpus with three parts, one part contains 98% of all the words and how

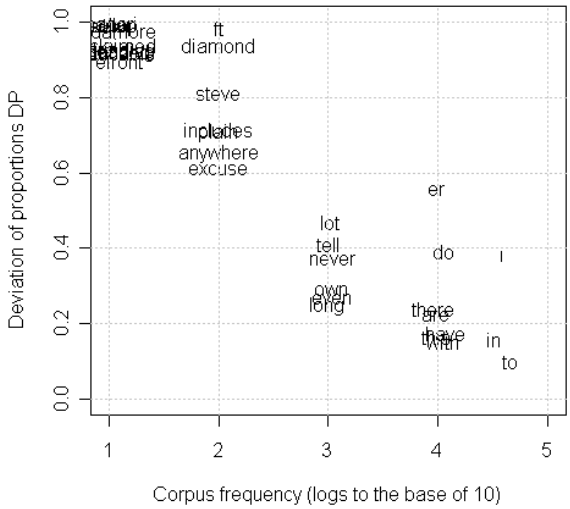

Figure 1. Selected words' *DP* and $\log_{10}$ frequency

parts-based measures (such as *DP*) and distance-based measures handle such distributions. One possibility is, therefore, that one should provide with every parts-based measure an indication of how unequal the corpus parts' sizes are on which the dispersion measures or adjusted frequency is based (e.g., relative entropy). A second possibility would be to compute the measures of dispersion and the adjusted frequencies on the basis of the existing corpus divisions (with all their unequal corpus part sizes) but then compare them to the measures one gets when one assumes that the corpus consists of the same number of parts, but now assumes they are equally-sized. Maybe the fractal dimension approach can help get important insights into how different corpus part sizes can be interpreted …

5.3 Validating dispersion measures and adjusted frequencies

One of the if not the most crucial issues is that both dispersions measures and adjusted frequencies are in need of corpus-external validation. It is one thing to devise statistics that are theoretically motivated and make intuitive sense when applied to corpus data, and above I myself did just that. However, very often measures of dispersion and adjusted frequencies serve practical purposes that make them very relevant even to linguists who may otherwise see no relevance in these matters. For example, the psycholinguist or psychologist who wishes to formulate experimental items in such a way as to avoid frequency or familiarity effects should ultimately not just choose words with particular frequencies — we have seen above that dispersion must be taken into consideration, too. That means, however, that if our dispersion measures or adjusted frequencies are supposed to be operationalizations

of familiarity or likelihood of encounter, then we must validate them against non-corpus-based evidence, i.e., psycholinguistic experimentation.

For instance, it is well-known that (logs of) observed frequencies are good proxies towards the familiarity of words given the strong correlations of frequencies with processing speed — cf. Howes and Solomon (1951) for recognition times, Oldfield and Wingfield (1965) and Forster and Chambers (1973) for naming times, and Ellis (2002a, b) as well as Jurafsky (2003) for overviews. However, there is very little work on the predictive power of dispersion measures and adjusted frequencies although one could, strictly speaking, of course argue that as long as we corpus linguists do not show that our dispersions and adjusted frequencies actually correspond to something outside of our corpora, we have failed to provide even the most elementary aspect of a new measure: its validation.

How could we come up with such evidence? Two research strategies are most obvious. First, we can reanalyze published psycholinguistic data, and Gries (2008, to appear a) are first attempts to correlate different adjusted frequencies with response time latencies. Second, we can of course perform experiments ourselves. For example, one could run experiments (i) on the fictitious distributions discussed in the first part of the article to determine whether our measures should actually be able to distinguish them or not (cf. Lyne 1985:115) and (ii) to determine which measures' results on large balanced corpora are most compatible with subjects' intuitions regarding the words' overall centrality in a language. Recent laudable work including dispersion is recent experimental work by Ellis (Ellis & Simpson-Vlach), who show that the range has significant predictive power above and beyond raw frequency of occurrence, and it is this kind of evidence we must provide in order to show our efforts are more than devising clever equations.

To conclude, given the still overwhelming reliance on unweighted frequency data and given the evidence on how misleading results based on frequencies alone can be discussed above, I hope that this paper (i) minimally motivates more researchers to not just rely on frequencies alone and maybe — once they decide to include/report dispersions — also adopt some kind of dispersion measure that can handle differently large corpus parts and (ii) maximally stimulates more researchers to try to come to grips with the distributional peculiarities of our trade. In order to facilitate this program of research, I am making some resources available myself (for the first two you will need to install the open source software R on your computer; cf. R Development Core Team 2008):

– an R function called `dispersions1`: This function requires three arguments: the first is a vector with all the words in the corpus in their order in the corpus; the second is a vector that states for each word which file it occurs in; the third is the element in square quotes for which dispersion measures are required.

Thus, if the BNC sampler is in a vector called `corpus` (e.g., Lebanon, leader, builds, cabinet, By, …) and the names of the files where each word occurs in that order are in a vector called `corpus.parts` (e.g., A7V, A7V, A7V, A7V, A7V, …), then you can enter the following two lines of code at the R prompt to get all dispersion measures and adjusted frequencies discussed above for the word form *understand* in the files of the BNC sampler (cf. Appendix 2 for an example of what the output looks like):

```
source("http://www.linguistics.ucsb.edu/faculty/stgries/research/
dispersion/_dispersion1.r")¶
dispersions1(corpus, corpus.parts, "understand")¶
```

– an R function called `dispersions2`: This function requires two arguments: the first is the vector of observed frequencies of the element in questions; in the example discussed above in Section 2, this vector would be (1, 2, 3, 4, 5); the second argument is a vector of corpus part sizes (in percent); in the above case of five equally-sized corpus parts, this would be (0.2, 0.2, 0.2, 0.2, 0.2). You can then enter the following two lines at the R prompt to get all dispersion measures and adjusted frequencies discussed above but the distance-based measures for the element whose frequencies you entered and the corpus whose part sizes you specified:

```
source("http://www.linguistics.ucsb.edu/faculty/stgries/research/
dispersion/_dispersion2.r")¶
dispersions2(c(1,2,3,4,5), c(0.2,0.2,0.2,0.2,0.2))¶
```

– (zipped) text files, OpenOffice.org Calc files, and .RData files that contain all of the above dispersion measures and adjusted frequencies for all word forms in the BNC Sampler, all word forms in the BNC Baby, all word forms in the spoken part of the BNC XML, and all word forms in the ICE-GB that occur more than 9 times. These files together with readme files explaining how they were generated can be downloaded from http://www.linguistics.ucsb.edu/faculty/stgries/research/dispersion/links.html.

Since we have seen above that dispersion measures and adjusted frequencies are sensitive to different corpus part sizes, the functions and the data also provide an index of how equal the corpus part sizes are to each other. When this index, relative entropy, is large, the files are similarly large, and the dispersions and adjusted frequencies are likely to be reliable — when it is low, the corpus files are very differently large, and dispersions and adjusted frequencies are less likely to be reliable.

As I see it, these tools and lists will be useful to a variety of different researchers: corpus linguists who wish to further investigate measures of dispersion,

adjusted frequencies, and maybe work on corpus comparison and homogeneity; lexicographers who are interested in checking the frequency information for words that they consider including in dictionaries; psycholinguistics and psychologists who use them to generate experimental stimuli that control for frequency and dispersion in a more suitable way than the Brown corpus can afford; computational or text linguists who use dispersion for information retrieval and text-linguistic purposes etc. I think the range of applications is vast and hope that these resources are one small step to help advance our knowledge and quality of practical work.

Notes

1. This judgment is supported by the fact that the number of papers at four recent corpus-linguistic conferences — ICAME 2007, Corpus Linguistics 2007, AACL 2008, and ICAME 2008 — which supported their frequency results with at least one index of dispersion is vanishingly small.

2. Cf. http://www.ucl.ac.uk/english-usage/projects/ice-gb/ for information on the ICE-GB.

3. Contrary to Savický and Hlaváčová (2002:216), the corpus parts need not correspond to genre groups.

4. Unless indicated otherwise, all retrieval operations, computations, and graphs were performed with R for Windows 2.7.0 (R Development Core Team 2008).

5. My discussion of this approach is rather vague. As far as I know, this proposal has not been published so far and my knowledge of it is only based on a four-page Powerpoint slideshow, which, as far as I understand it, does not explain the measure in great detail. I chose to include the measure anyway because its logic appears to be different from all existing ones and its discussion thus serves to highlight the range of parameters future research on dispersion may want to explore.

6. This measure is also known as Francis and Kučera's (1982:463f.) adjusted frequency *AF*.

7. Note that this formula is not exactly the one provided by Kromer (2003:181, formula 8) as I have inserted a closing bracket after "*C*", which is missing in his paper. Ψ refers to the first derivative of the logarithm of the gamma function; *C* is the Euler-Mascheroni constant computed as follows: $c = \lim_{n\to\infty}\left[\left(\sum_{k=1}^{n}\frac{1}{k}\right) - \ln n\right]$ or, in R: `z<-1:10000000; sum(1/z)-log(length(z))¶`.

8. Interestingly, there seems to be disagreement and/or confusion regarding this issue. Rosengren (1971) discusses the coefficients *U*, *AF*, and Juilland et al.'s *D* and states "[a]ll the measures discussed hitherto presuppose equally large categories" (p. 119). He then goes on to discuss adjustments for these measures that allow using unequal sizes of corpus parts as well as his own measure *S*. Oakes (1998:191), on the other hand, discusses a variety of dispersion measures as well as Lyne's (1985, 1986) comparison of *D*, D_2, and *S*, and in the immediately following paragraph states "[i]t is *not* possible to use these measures if the corpus is subdivided into sections of different sizes [my emphasis, STG]." Similarly, Piao (2002:212f.) states that *D* requires equally-

sized corpus parts whereas D_2 and S do not in spite of the published proposals refinements to be discussed presently. Finally, Forsbom (2006:5) simply states that Savický and Hlaváčová's (2002) measures require equally-sized corpus parts, which is simply wrong.

9. This issue has been brought up by Paul Rayson and both anonymous reviewers.

10. The discussion of this point of critique needs to be qualified since it has to be admitted that it is not always intersubjectively obvious which distributions should be distinguished by dispersion measures. For example, Lyne (1985) discusses how D cannot distinguish between (3, 2, 1, 1, 1) and (2, 2, 2, 2, 0) while D_2 and S yield quite marked differences and comments:

> As we have just seen, the reasons why S and D_2 "demote" certain distributions compared with D is that they contain one or more low sub-frequencies, (not necessarily zero). But surely it is not the business of a dispersion measure to discriminate against particular distributions in this manner. In the above example [A: (3, 2, 1, 1, 1) vs B: (2, 2, 2, 2, 0)], which is [Rosengren's] own, it seems to us quote proper that B should be rated as highly as A, because the presence of a single low sub-frequency, 0, in B is balanced by the perfectly even distribution across the remaining four sections. (Lyne 1985:115)

11. Some may actually find penalizing zeros attractive, but I agree with Lyne in the sense that I would not want to place too much emphasis on zeros. First, while zeros do signal underdispersion, that underdispersion must take into consideration the expected frequency (e.g., on the basis of the size of the corpus part) and especially if that expected frequency is small zeros should not have too much leverage on the dispersion value. Second, zeros may be due to sampling variation, and given that there is this whole field of research on providing better estimates of the frequency of unseen items (with, say, Good-Turing estimates), I think one should be reluctant to overemphasize zeros.

12. This is not to say that this is a very realistic assumption or application of how parts-based measures have so far been applied, but in a comparative review all pros and cons should be pointed out. Also, note that Savický and Hlaváčová's (2002) measures could provide the same information when the corpus is resorted, but then this may be difficult to motivate when one does not already assume different corpus parts whose integrity the sorting would restore: if one's measure assumes that a corpus does not come in parts and only the distances between successive occurrences are important, why would one want to destroy the natural order of turns in a dialog corpus?

13. For a discussion of problems of the null-hypothesis-significance testing paradigm cf., e.g., Loftus (1991, 1996).

14. This can be best understood on the basis of a non-linguistic example. Imagine three buckets so close to each other that any coin thrown in their direction must land in one of the them. Imagine further that of the one bucket takes of 98% of the space and the other two equally share the remaining 2%. If one now tossed 100 coins randomly in the direction of the buckets, then the result in the absence of any patterning would be that approximately 98 of the coins land in the big bucket. The absence of patterning here corresponds to "no particular distributional bias that (all) dispersions measures pick up", which is why the measures that take the corpus sizes into consideration label the result where 98 coins end up in the largest bucket the "normal" distribution.

15. Cf. http://www.natcorp.ox.ac.uk/corpus/index.xml.ID=products#sampler for information on the BNC sampler.

16. One may erroneously suspect that *DC* appears to be a better measure because "its scores are more evenly distributed." At present, however, this argument does not follow. First, the values for DC are not necessarily more evenly distributed: *DC*'s interquartile range is larger, but its range is just about the same as that of *DP* (or, say, *S*). Second, how would one know that an even distribution of dispersion values is one or oven the best criterion? It may be, but it may just as well not be since one could just as well argue that the fact that the median and the mean of *DP* are closer to 0.5 than those of *DC* shows that *DP* is better. I explicitly do not do that because there is no way of knowing which criterion is better, which is why corpus linguists will ultimately need to look at corpus-external converging evidence (cf. Gries 2008). Third, even if *DP* and *DC* were to yield results of exactly the same quality (however measured), *DP* is still more appealing because it is conceptually simpler (Occam's razor) and easier to understand — everybody can understand differences of percentages. Finally, the dispersion data for tens of thousands of words from different corpora that I make available with this paper (cf. Section 4) will allow other researchers to follow up on this and arrive at their own conclusions.

17. I also computed *DP* in another way to test the validity of the results. In this second way, the expected baseline percentages were not the percentages of verbs of each corpus file, but the percentages of all words of each corpus file. The results are for all intents and purposes the same as those obtained on the basis of the verbs.

Acknowledgements

I am very grateful to Petr Savický and especially Justin Washtell for a lot of input and stimulating discussion that influenced and shaped my thinking about these issues considerably. Finally, I thank audiences at Corpus Linguistics 2007 and AACL 2008 for feedback. The usual disclaimers apply.

References

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. (1995a). Inverse document frequency: a measure for deviations from Poisson. In D. Yarowsky & K.W. Church (Eds), *Proceedings of the Third Workshop on Very Large Corpora* (pp. 121–130). Cambridge, MA: MIT.

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

Ellis, N. C. (2002a). Frequency effects in language processing and acquisition. *Studies in Second Language Acquisition*, 24 (4), 143–188.

Ellis, N. C. (2002b). Frequency effects in language processing and acquisition. *Studies in Second Language Acquisition*, 24 (4), 297–339.

Ellis, N. C. & Simpson-Vlach, R. (2005). An academic formulas list (AFL): extraction, validation, prioritization. Paper presented at Phraseology 2005, Université Catholique Louvain-la-Neuve.

Ellis, N. C. & Simpson-Vlach, R. (to appear). Formulaic language in native speakers: triangulating psycholinguistics, corpus linguistics, and education. *Corpus Linguistics and Linguistic Theory*.

Ellis, N. C., Simpson-Vlach, R. & Maynard, C. (2007). The processing of formulas in native and L2 speakers: psycholinguistic and corpus determinants. Paper presented at the Symposium on Formulaic Language, University of Wisconsin-Milwaukee.

Engwall, G. (1974). *Fréquence et distribution du vocabulaire dans un choix de romans français.* Stockholm: Skriptor.

Forsbom, E. (2006). Deriving a base vocabulary pool from the Stockholm-Umeå Corpus. Unpubl. paper, NGSLT course "Soft Computing".

Forster, K. I. & Chambers, S. M. (1973). Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior*, 12 (6), 627–635.

Francis, W. N. & Kučera, H. (1982). *Frequency analysis of English usage: lexicon and grammar*. Boston: Houghton Mifflin.

Gries, St. Th. (2006). Some proposals towards more rigorous corpus linguistics. *Zeitschrift für Anglistik und Amerikanistik*, 54 (2), 191–202.

Gries, St. Th. (2008). Measures of dispersion in corpus data: a critical review and a suggestion. Paper presented at the conference "American Association of Corpus Linguistics" 2008, Brigham Young University.

Gries, St. Th. (to appear a). Dispersions and adjusted frequencies in corpora: further explorations.

Gries, St. Th. (to appear b). *Quantitative corpus linguistics with R: a practical introduction*. New York: Routledge.

Gries, St. Th., Hampe, B. & Schönefeld, D. (2005). Converging evidence: bringing together experimental and corpus data on the association of verbs and constructions. *Cognitive Linguistics*, 16 (4), 635–676.

Gries, St. Th., Hampe, B. & Schönefeld, D. (to appear). Converging evidence II: more on the association of verbs and constructions. In J. Newman & S. Rice (Eds.), *Experimental and empirical methods in the study of conceptual structure, discourse, and language.* Stanford, CA: CSLI.

Gries, St. Th. & Stefanowitsch, A. (2004). Extending collostructional analysis: a corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics*, 9 (1), 97–129.

Gries, St. Th. & Wulff, St. (2005). Do foreign language learners also have constructions? Evidence from priming, sorting, and corpora. *Annual Review of Cognitive Linguistics* 3, 182–200.

Hofland, K. & Johansson, S. (1982). *Word frequencies in British and American English*. The Norwegian Computing Centre for the Humanities, Bergen, Norway.

Howes, D. H. & Solomon, R. L. (1951). Visual duration threshold as a function of word-probability. *Journal of Experimental Psychology*, 41 (6), 401–410.

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

Jurafsky, D. (2003). Probabilistic modeling in psycholinguistics. In R. Bod, J. Hay & S. Jannedy (Eds.), *Probabilistic linguistics* (pp. 39–96). Cambridge, MA: The MIT Press.

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.

Loftus, G. R. (1991). On the tyranny of hypothesis testing in the social sciences. *Contemporary Psychology*, 36 (2), 102–5.

Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions on Psychological Science*, 5(6), 161–71.

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.

Muller, C. (1965). Fréquence, dispersion et usage: à propos des dictionnaires de fréquence. *Cahiers de Lexicologie*, 7 (2), 33–42.

Oakes, M. P. (1998). *Statistics for corpus linguistics*. Edinburgh: Edinburgh University Press.

Oakes, M. P. & 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.

Oldfield, R. & Wingfield, A. (1965). Response latencies in naming objects. *Quarterly Journal of Experimental Psychology A*, 17 (4), 273–281.

Piao, S. S. (2002). Word alignment in English-Chinese parallel corpora. *Literary and Linguistic Computing*, 17 (2), 207–230.

Quasthoff, U. (2007). Fraktale Dimension von Wörtern. Unpubl. ms.

R Development Core Team. (2008). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

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, 103–27.

Savický, P. & Hlaváčová, J. (2002). Measures of word commonness. *Journal of Quantitative Linguistics*, 9 (3), 215–31.

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

Stefanowitsch, A. & Gries, St. Th. (2003). Collostructions: investigating the interaction between words and constructions. *International Journal of Corpus Linguistics*, 8 (2), 209–243.

Szmrecsanyi, B. (2005). Language users as creatures of habit: a corpus-linguistic analysis of persistence in spoken English. *Corpus Linguistics and Linguistic Theory*, 1 (1), 113–150.

Szmrecsanyi, B. (2006). *Morphosyntactic persistence in spoken English. A corpus study at the intersection of variationist sociolinguistics, psycholinguistics, and discourse analysis.* Berlin, Heidelberg, New York: Mouton de Gruyter.

Tao, H. (2001). A usage-based approach to argument structure: *remember* and *forget* in spoken English. *International Journal of Corpus Linguistics*, 8 (1), 75–95.

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

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.

Author's address

Stefan Th. Gries
Department of Linguistics
University of California, Santa Barbara
Santa Barbara, CA 93106-3100
United States of America

Appendix 1

The words that were included in the BNC Sampler analysis in Section 3 (corpus frequencies and ranges in parentheses) are the following: *the* (104248, 184), *and* (53216, 184), *to* (46187, 184), *of* (44276, 184), *a* (39119, 184), *i* (38445, 167), *it* (33968, 184), *in* (32198, 183), *you* (30538, 164), *that* (29280, 183), *be* (12735, 183), *he* (12424, 168), *have* (11928, 184), *do* (11374, 179), *with* (11138, 184), *but* (10569, 182), *are* (9770, 182), *er* (9721, 105), *this* (9651, 183), *there* (9242, 182), *not* (9211, 183), *never* (1104, 147), *even* (1100, 170), *own* (1066, 172), *lot* (1032, 135), *end* (1031, 171), *house* (1024, 141), *tell* (1023, 141), *came* (1013, 152), *each* (1007, 144), *both* (1001, 163), *long* (993, 164), *anywhere* (102, 60), *carl* (102, 8), *definition* (102, 33), *diamond* (102, 10), *egypt* (102, 12), *excuse* (102, 62), *formal* (102, 40), *ft* (102, 2), *hardly* (102, 62), *includes* (102, 45), *pink* (102, 29), *plain* (102, 42), *properly* (102, 57), *russians* (102, 11), *steve* (102, 31), *thousands* (102, 45), *defender* (10, 1), *enhanced* (10, 10), *forefront* (10, 10), *frustrated* (10, 10), *grind* (10, 10), *hathor* (10, 1), *lemar* (10, 1), *macari* (10, 1), *malins* (10, 1), *mamluks* (10, 1), *misleading* (10, 10), *practicable* (10, 10), *pre* (10, 10), *prevailing* (10, 10), *proclaimed* (10, 10), *scallop* (10, 1), *scudamore* (10, 1), *sem* (10, 1), *tatars* (10, 1), *verdict* (10, 10).

The following exposition explains in very much detail how these words were chosen in (up to the level of providing Perl-compatible regular expressions). The reason for this level of detail is that the default settings of different concordance programs can yield different outputs even when given the same search strings (cf. Gries, to appear b). In order to guarantee the replicability of the retrieval procedure, the kind of meticulous characterization common in other disciplines has to be adopted here, too. I wrote an R script (cf. R Development Core Team 2008) that

- loaded each corpus file, converted it to lower case, and retained only the lines that contained sentence numbers (regexp: "<s n");
- deleted all sequences of spaces, non-word tags, and the material they refer to (regex: "·*<(?!w·.*?>).*?>[^<]*") as well as a formula and number tags and what they refer to (regex: "·*<w·(fo|m).*?>[^<]*");
- split up the remaining data at sequences of optional spaces and word tags (regex: "·*<w.*?>").

The result of this was a word list containing 50,354 types / 1,944,548 tokens. In order to choose an appropriate set of words with which to test *DP*, I determined the maximal frequency of a word in the corpus, which turned out to be 104,248 (of *the*). I then computed the logarithm to the base of 10 of 104,253 (5.018068), divided that number by five, multiplied it with the numbers from 1 to 5, and antilogged them to arrive at a frequencies at every order of magnitude from 10 to 104,248. (The exact function that was used probably makes the procedure clearer than the prose description: round(10^(1:5*(*log10*(104254)/5))).) As a result, I obtained the following frequencies to be inspected: 10, 102, 1,025, 10,338, and 104,248. However, the numbers of words with these or very similar corpus frequencies were very uneven, which is why I chose to investigate

- the ten most frequent words;
- the word with the frequency closest to 10,338 as well as the next five more frequent and less frequent words;
- the word with the frequency closest to 1,025 as well as the next five more frequent and less frequent words;
- all sixteen words with a corpus frequency of 102;
- ten words with a corpus frequency of 10 that occurred in 10 files each and ten words with a corpus frequency of 10 that occurred in 1 file each.

Appendix 2

```
R Console
File  Edit  Misc  Packages  Help

> dispersions1(corpus, corpus.part.sizes, element)
$element
[1] "a"

$`observed overall frequency`
[1] 15

$`sizes of corpus parts / corpus expected proportion`
[1] 0.2 0.2 0.2 0.2 0.2

$`relative entropy and variation coefficient of all sizes of the corpus parts`
[1] 1 0

$range
[1] 5

$maxmin
[1] 4

$`standard deviation`
[1] 1.581139

$`variation coefficient`
[1] 0.5270463

$`Chi-square`
[1] 3.333333

$`Juilland et al.'s D (based on equally-sized corpus parts)`
[1] 0.7364769

$`Juilland et al.'s D (not requiring equally-sized corpus parts)`
[1] 0.7364769

$`Carroll's D2`
[1] 0.925634

$`Rosengren's S (based on equally-sized corpus parts)`
[1] 0.9368466

$`Rosengren's S (not requiring equally-sized corpus parts)`
[1] 0.9368466

$`Lyne's D3 (not requiring equally-sized corpus parts)`
[1] 0.9444444

$`Distributional consistency DC`
[1] 0.9368466
```

```
$`Inverse document frequency IDF`
[1] 0

$`Engvall's measure`
[1] 15

$`Juilland et al.'s U (based on equally-sized corpus parts)`
[1] 11.04715

$`Juilland et al.'s U (not requiring equally-sized corpus parts)`
[1] 11.04715

$`Carroll's Um (based on equally sized corpus parts)`
[1] 14.10761

$`Rosengren's Adjusted Frequency (based on equally sized corpus parts)`
[1] 14.0527

$`Rosengren's Adjusted Frequency (not requiring equally sized corpus parts)`
[1] 14.0527

$`Kromer's Ur`
[1] 8.7

$`Savický and Hlavácová's distances`
 [1]  2 10  2  9  2  5  2  3  3  1  3  2  1  3  2

$`Savický and Hlavácová's ARF / f_ARF`
[1] 10.8

$`Savický and Hlavácová's AWT`
[1] 3.18

$`Savický and Hlavácová's f_AWT`
[1] 9.328358

$`Savický and Hlavácová's ALD`
[1] 0.628417

$`Savický and Hlavácová's f_ALD`
[1] 11.76395

$`Washtell's Self Dispersion`
[1] 0.9960317

$`Deviation of proportions DP`
[1] 0.2

$`Deviation of proportions DP (normalized)`
[1] 0.25
```