

Vocabulary Dispersion Statistics

Some people are interested in finding and comparing patterns of word usage in one or more texts.

Frequency: Some reading and language instructors want to know the *high frequency* words in a text or part of a text. They use these lists to introduce vocabulary words to students at appropriate times. For example, *atonement* occurs 81 times in the Old Testament, and only 1 time in the New Testament. Some people look at *low frequency* words. The very low frequency words may be typos. These words can be copied and pasted into a word processor with a spell checker to help identify possible typos.

Linguists are interested in the absolute frequency (Freq.), the relative frequency (Rel.), the \log_{10} frequency (LogF), and the average reduced frequency (ARF) explained below.

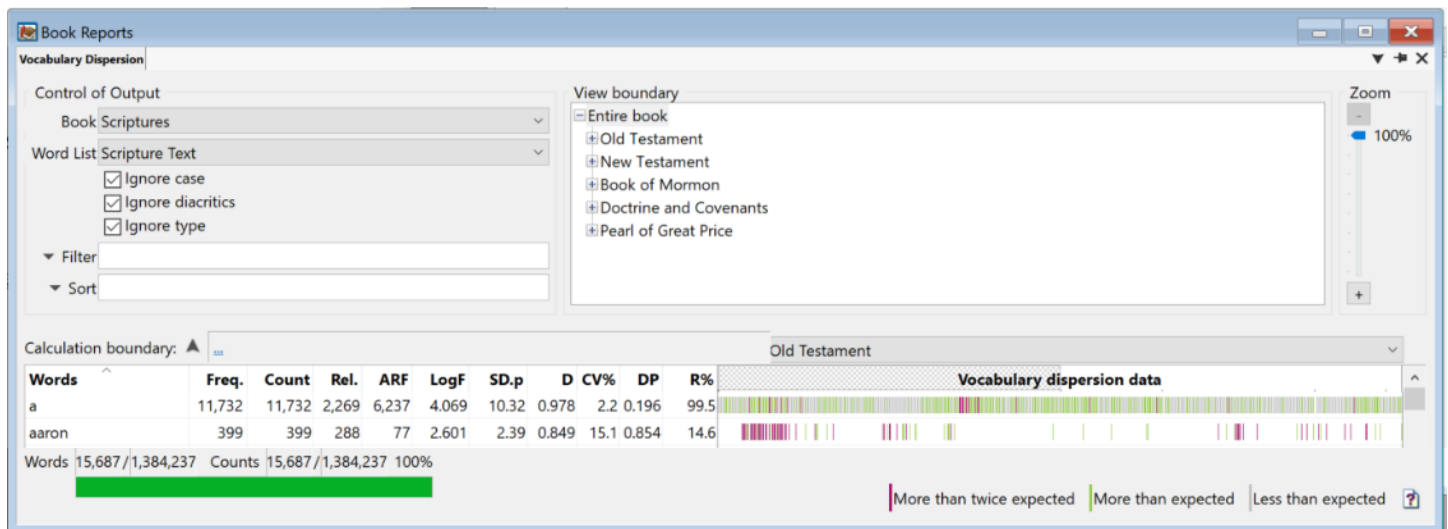
Dispersion: Some people are interested in how words are distributed or dispersed in a text or a collection of texts (corpora). Are they distributed evenly or unevenly across a text or corpus? If a word is only used in a part of a book or a corpus, it would be helpful to introduce its meaning before assigning that part.

If a corpus is organized by decade or genre (e.g., fiction, academic, newspaper, comedy), dispersion statistics help identify words used in many or few sections of the corpus. For example, statistical terms would appear more often in academic research writings than in fiction. Some words like *muggle* are used more frequently and with a different meaning in the decades after the *Harry Potter* books became popular.

Linguists use several different dispersion statistics described below.

Vocabulary Dispersion Report

Select the Book View window you want to analyze. Select Vocabulary Dispersion from the Analyze menu or from the reports menu in the top right corner of each Book View window. This will automatically generate the report and show a green progress bar at the bottom left of the window. In the bottom right corner, you will see an explanation of the colored lines in the “Vocabulary dispersion data” column. In this example, *aaron* occurs primarily in the first five books of the Old Testament, and *a* occurs pretty evenly throughout the Bible.



You can click on a column heading to sort the report. You can also **right-click** on a column heading (e.g., ARF) to see a popup menu of possible columns. This allows you to show or hide columns. Each of the columns is explained below.

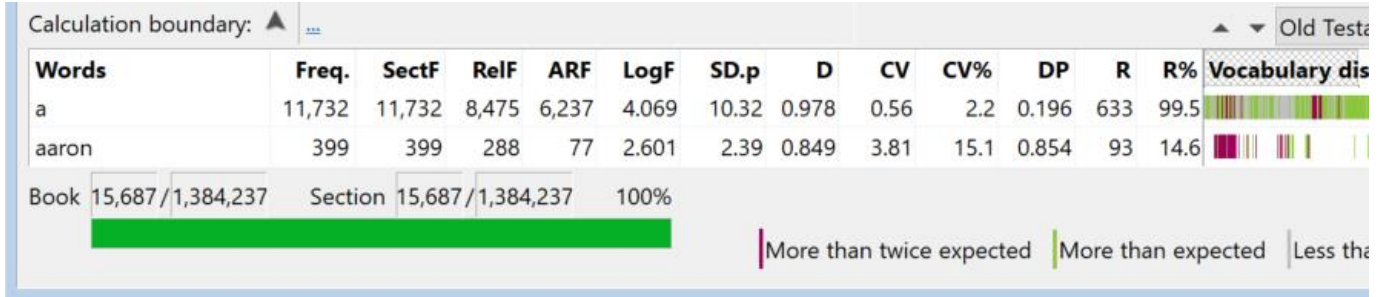
Statistics: For each word, you will see frequency (e.g., Freq, RelF) and dispersion (e.g., D, DP) statistics in the adjacent columns, and a graphical view showing where words occur in the book. Each of these columns is briefly described below. For more information, see [Statistics in Corpus Linguistics: A Practical Guide](#) by Vaclav Brezina, and “[Analyzing dispersion](#),” by Stefan Th. Gries.

Save Results. You can *copy* selected lines from the report. You can also select Export from the File menu to export selected lines or the entire report to a CSV (utf16) file or to a TXT (utf8) file. You can double-click on a CSV file to open it in Excel. To open the TXT file in Excel, click on the Data tab, click on “Get External Data” button, select “From Text,” and follow prompts.

Report Options. When the report is completed, you can change options (i.e., book, word list, and ignore options, filters, sorts, view boundaries, zoom, and calculation boundaries).

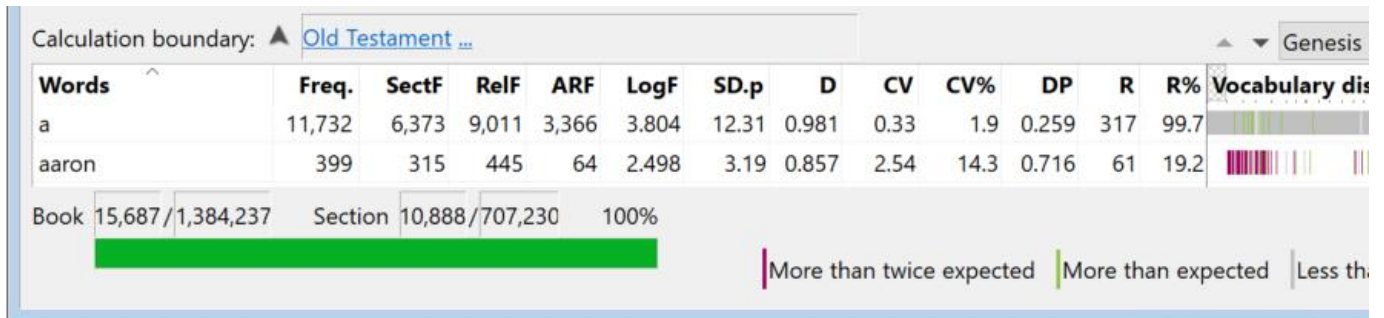
Absolute Frequencies (Freq., SectF)

The *Freq.* column is the number of times each word occurs in the entire book or corpus (e.g., Bible). The Section Frequency (*SectF*) column is the number of times each word occurs in the section specified by the “calculation boundary” (e.g., New Testament). These columns will be the same if there is no calculation boundary. At the bottom of the report you can see the total number of words (types) in the list and the total frequencies (tokens) for the words in the list. In this example, there are 15,687 words that occur 1,384,237 times in the scriptures.



Words	Freq.	SectF	RelF	ARF	LogF	SD.p	D	CV	CV%	DP	R	R%	Vocabulary dis
a	11,732	11,732	8,475	6,237	4.069	10.32	0.978	0.56	2.2	0.196	633	99.5	
aaron	399	399	288	77	2.601	2.39	0.849	3.81	15.1	0.854	93	14.6	
Book	15,687 / 1,384,237		Section 15,687 / 1,384,237		100%								

If you select a section of the book or corpus, *Freq.* and *SectF* numbers will be different. In this example, there are 10,888 words that occur 707,230 times in the Old Testament.



Words	Freq.	SectF	RelF	ARF	LogF	SD.p	D	CV	CV%	DP	R	R%	Vocabulary dis
a	11,732	6,373	9,011	3,366	3.804	12.31	0.981	0.33	1.9	0.259	317	99.7	
aaron	399	315	445	64	2.498	3.19	0.857	2.54	14.3	0.716	61	19.2	
Book	15,687 / 1,384,237		Section 10,888 / 707,230		100%								

Relative Frequency (RelF)

Some people are interested in comparing the *relative frequency* (RelF) of words used in different texts of unequal sizes. Relative frequency estimates how many times each word would occur if the book or corpus contained exactly one million words. For example, if a word occurred 10 times in a 100 words (10%), and 10 times in 1000 words (1%), the absolute frequencies are the same, but the relative frequencies are 100,000 (10%) and 10,000 (1%) respectively.

The Old Testament (OT) has over 3.4 times as many words as the New Testament (NT) so we would expect the word “*the*” to occur 3.4 times more often in the OT than in the NT. *Lord* occurs 7125 times (1.00%) in the OT and 711 times (0.34%) in the NT. The *absolute* OT frequency is 10 times more than in the NT. However, the *relative* frequency (words per million) in the OT is 10074 compared to 3379 in the NT, or only 3 times the NT.

The table below shows how to calculate the relative frequencies for *Lord* in the Old and New Testaments.

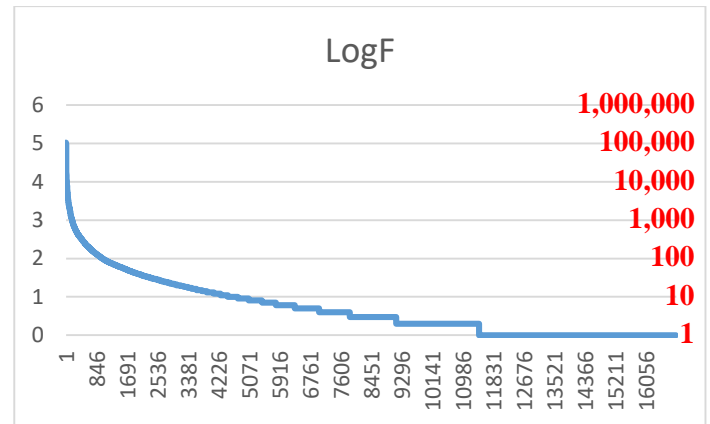
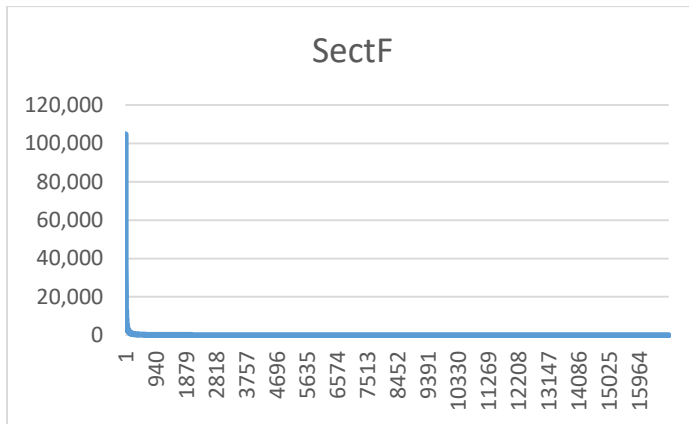
RELATIVE frequencies	Old Testament	New Testament
Lord	RelF = (SectF/SectF _{total}) * 1,000,000 RelF = (7125/707230) * 1,000,000 RelF = 10,074	RelF = (SectF/SectF _{total}) * 1,000,000 RelF = (711/210414) * 1,000,000 RelF = 3,379
All other words	RelF = (700105/707230) * 1,000,000 RelF = 989,926	RelF = (209703/210414) * 1,000,000 RelF = 996,621
Total	RelF = 1,000,000 SectF _{total} = 707,230	RelF = 1,000,000 SectF _{total} = 210,414

Log₁₀ Frequencies (LogF)

A book or corpus has a few very high frequency words (e.g., the, and, of) and many low frequency words. In the scriptures, 200 unique words occur between 750 and 93,000 times, while about 5,000 words only occur one time. When we sort and graph the SectF frequency, it looks like a straight line down and a straight line across the bottom.

When we calculate the log₁₀ frequency of each word, 100,000 becomes 5.0, and 10 becomes 1.0 as shown in the table below. It is easier to notice low frequency differences in the LogF graph.

Frequency	100,000	10,000	1,000	100	10	3	2	1
Log₁₀ frequency	5.0	4.0	3.0	2.0	1.0	0.48	0.30	0.0



Average Reduced Frequency (ARF)

ARF combines absolute frequency and dispersion. Suppose we have two words that occur 100 times in a book or corpus. One of these words only occurs in part of the book, and the other occurs in all parts of the book. The first word will have a lower ARF than the second word. For example, if a word w occurs 5 times in a text of 100 words, they might be close to each other or exactly 20 words apart. The subscripts indicate the position of the word in the 100 word book.

...W₂₀.W₂₄.W₂₈.W₃₂.W₃₆...

W₁...W₂₁...W₄₁...W₆₁...W₈₁...

To calculate the ARF, we calculate the average distance or number of words between each occurrence of w . We then calculate the actual distance (d) between consecutive occurrences, compare it to the average, and add up the *smaller* or *minimum* of the two.

$v = (\text{SectF}_{\text{total}} / \text{SectF}) = 100 / 5 = 20$ $d_1 = w_2 - w_1 = 24 - 20 = 4$ $d_2 = w_3 - w_2 = 28 - 24 = 4$ $d_5 = w_1 - w_4 = (100 - 36) + (0 + 20) = 84$	$ARF = \frac{1}{v} \times \sum_{i=1}^f \min(d_i, v)$
--	--

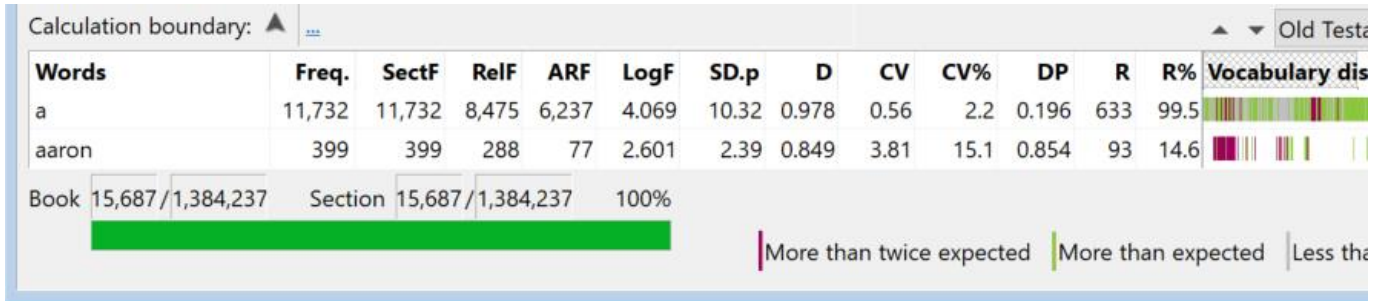
In the first example above, the distance between each of the first four words and the next one is less than 20, and the distance from the last one to the first one (w_{36} to w_{20}) is 84 (64 to the end plus 20 from the beginning to the first word).

	...W ₂₀ .W ₂₄ .W ₂₈ .W ₃₂ .W ₃₆ ...	W ₁ ...W ₂₁ ...W ₄₁ ...W ₆₁ ...W ₈₁ ...
$v = (\text{SectF}_{\text{total}} / \text{SectF})$	100 / 5 = 20	100 / 5 = 20
Total distance = sum of $\min(w_i, v)$	4 + 4 + 4 + 4 + 20 = 36	20 + 20 + 20 + 20 + 20 = 100
ARF = (total distance) / v	36 / 20 = 1.8	100 / 20 = 5

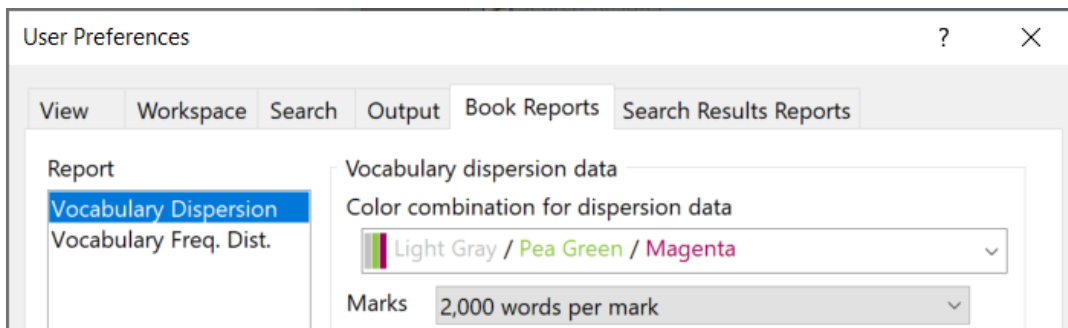
The ARF for a word will be less than SectF unless the word occurs only one time. For words distributed evenly in the section, ARF will be closer to SectF as shown in the last column of the table above. For unevenly distributed words, ARF will be much smaller than SectF as shown in the middle column above.

Dispersion Statistics

Marks and Parts: Most dispersion statistics measure how many parts of a book or corpus contain a word. When you use some programs, the parts may be a different text files of unequal sizes. In WordCruncher, the parts are equal divisions of the entire book or corpus. The graphical chart has one line or mark for each part. Each mark represents one part of the book or section. The color of each mark indicates the presence of more or less than the expected frequency.



The color of each mark and the minimum size or number of words in each part are specified in the User Preferences. You can select minimum size of 1000, 2000, 3000, 4000, 5000, or 10000 words.



The actual size of each part for this book is calculated as shown.

Minimum = 2000

N = 1,273,498 Total number of “normal” words (tokens) in the section or book.

P = int(N / minimum size) = 1,273,498 / 2000 = int(636.7) = 636 Complete parts containing the minimum size.

Part size = N / P = 1,273,498 / 636 = 2002.4

Note: The minimum number of parts is 2 and the maximum is 100,000.

The colors are relative to the expected frequency (SectF/P) for each part. For “a” the expected or average frequency in each part is 11732/636=18.4 and for “aaron” it is 399/636=0.6.

In the first column and tooltips, — means *even* distributions across all parts and | means *uneven* distributions.

	Description of Dispersion Measures	Formula
R R% 0= 100=—	<p>Range (R) and Range percent (R%) statistics are very basic measures of dispersion. Range tells us how many parts of the text contain the word. In the example above, each part contains at least 2000 words. If the distribution is even (—), R% is close to 100%. If it is uneven (), R% is close to 0%.</p> <ul style="list-style-type: none"> “a” occurs in 633 of 636 parts (R% = 99.5%). If we had used the “All text” word list instead of “Scripture Text,” “a” would have been in all 636 parts (R% = 100.0%). “aaron” occurs in 93 of 636 parts (R% = 14.6%). <p>Problem: R and R% only count the number of parts and ignore word frequencies in each part. If we were using parts with unequal parts, range would ignore the size differences.</p> <p>Solution: Use R and R% for a simple exploration of the data. Use more sensitive dispersion measures for further analysis.</p>	$R\% = \frac{R}{\text{parts}}$ <ul style="list-style-type: none"> R = number of parts containing at least one instance of a word Parts = number of parts containing at least the user specified number of words.

<p>SD.p 0 = —</p>	<p>The population Standard Deviation (SD.p or σ) indicates how much a word frequency in each part differs from the average frequency (μ) per part. If SD.p is about the same as the average, there is a lot of variation with many parts having many more or less than the average frequency.</p> <p>Problem: SD.p cannot be used to compare dispersions of words that occur with different frequencies because SD.p needs to be interpreted relative to the average frequency per part of a word.</p> <p>Solution: Use other measures like (a) CV and D that compare SD.p with the average, or (b) DP that doesn't use SD.p.</p>	$\sigma = \sqrt{\frac{\sum(x - \mu)^2}{n}}$ <ul style="list-style-type: none"> • $\mu = \text{SectF} / \text{parts} = \text{average frequency}$ • $x = \text{word frequency in a part}$ • $n = \text{number of parts}$ • $\text{SD.p} = \sigma$
<p>CV 0 = —</p>	<p>The Coefficient of Variation (CV) is the amount of variation relative to the average or mean frequency of a word in a book. This standardized measure enables you to compare different words in a corpus. If CV is close to 0, the word is evenly distributed in the book or corpus.</p>	$\text{CV} = \frac{\sigma}{\mu}$ <ul style="list-style-type: none"> • $\mu = \text{SectF} / \text{parts} = \text{average frequency}$ • $\sigma = \text{SD.p}$
<p>CV% 0 = — 100 = </p>	<p>CV% divides CV by the maximum value of CV.</p> <p>If the word occurs in all parts, CV% will be close to 0% (— = even distribution). If a word occurs in only one part, CV% is 100% (= uneven distribution).</p>	$\text{CV\%} = \left(\frac{\text{CV}}{\text{CVmax}}\right) \times 100$ <ul style="list-style-type: none"> • $\text{CVmax} = \sqrt{\text{parts} - 1}$
<p>D 0 = 1 = —</p>	<p>Juillard's D is based on CV. It is a number between 0 (= uneven distribution) and 1 (— = even distribution).</p> <p>Problem: With a large number of parts (e.g., 1000), the practical range of D becomes restricted.</p>	$D = 1 - \left(\frac{\text{CV}}{\text{CVmax}}\right)$
<p>DP 0 = — 100 = </p>	<p>Deviation of Proportions (DP) was proposed by Gries (2008). It compares the expected distribution of a word with the actual distribution. It is a number between 0 (— = even distribution) and 1 (= uneven distribution).</p> <p>DP is the sum of the <i>absolute</i> differences between the observed and expected frequencies for each part of a book or corpus divided by 2. Words with DP closer to 0 are more evenly distributed (—). Words with DP closer to 1 are unevenly distributed ().</p> <p>In the scriptures, DP=0.191 for “a” which is evenly distributed and DP=0.851 for “aaron” which is primarily found in the first five books of the OT. The maximum DP is $1 - (1/\text{parts}) = 0.998$.</p>	$DP = \frac{\sum o\% - e\% }{2}$ <ul style="list-style-type: none"> • $o\% = f / \text{SectF}$ where f is the word frequency in a part. • $e\% = \text{partSize} / N$ where partSize is the number of words in each part and N is the number of words in the section. Since all parts are of equal size in WordCruncher, $e\% = 1/\text{parts}$. • DP max = $1 - \min(e\%)$