Phrase Compare Statistics
(N-grams, Keyness)

Some people are interested in finding the most common phrases in a text. For example, the five word phrase or N-gram, "and it came to pass," occurs 87 times in the first two books of the Old Testament—Genesis and Exodus. Apps like Amazon or Google use phrase frequencies to suggest the next words as you type. Writers use phrase frequencies to find better phrases that start with the same words. Students may want to know which phrases they use too often so they can add variety to their writing.

Some people are interested in finding common phrases or phrases that occur *more or less frequently* in one text than in another text. For example, the five word phrase or N-gram "and it came to pass" occurs significantly more often in Genesis and Exodus than it appears in Leviticus, Numbers, and Deuteronomy. This may suggest different types of writing (e.g., historical narrative) are found in the texts.

Other people want to identify phrases used as often in one text as in another text. This might suggest that one text influenced another.

Keyness refers to the comparison of frequencies for individual words in two different books, sections, or corpora. When the maximum length of a phrase is 5, the Phrase Compare report gives information for phrases or N-grams that are 1, 2, 3, 4, and 5 words long. The methods used to compare frequencies of individual words are also helpful in identifying phrases that occur more, less, or about as often in two different books, sections, or corpora.

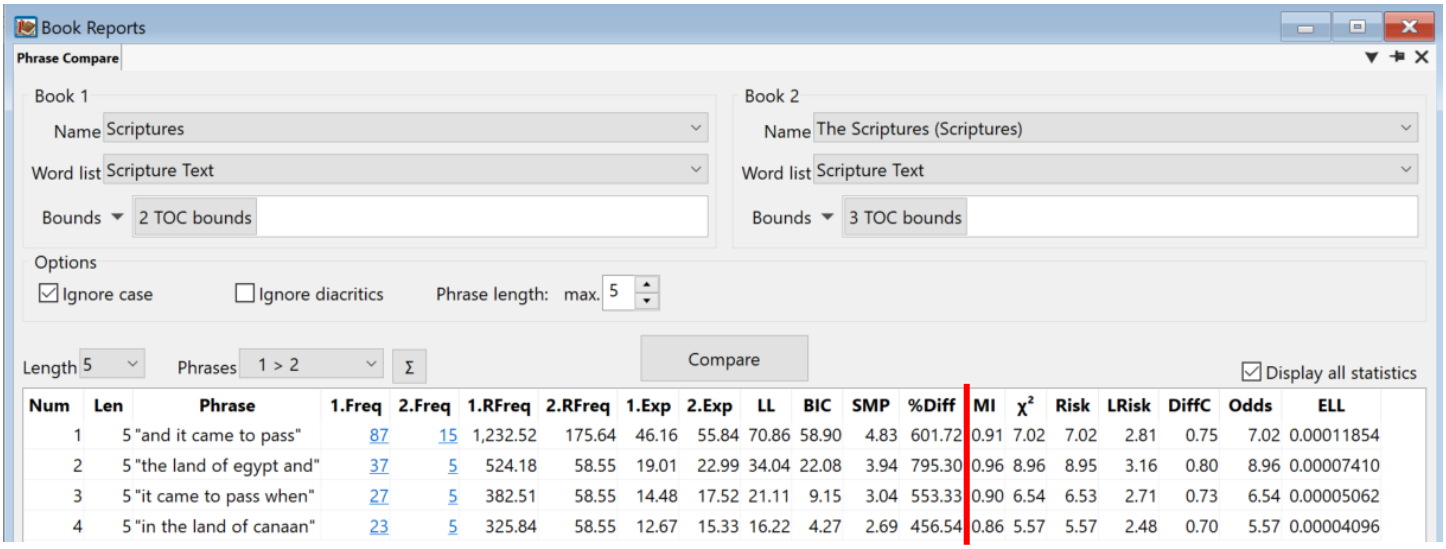
Phrase Compare Report

Book 1. Select the first book, word list, and bounds (optional). In the example below, Genesis and Exodus were selected.

Book 2. Select the second book, word list, and bounds (optional). If you select None as the book, you will find repeated phrases in Book 1. You can select a different book or the same book. In the example below, different parts (Leviticus, Numbers, and Deuteronomy) of the same book (Scriptures) were selected.

Options. Select options like Ignore case and a maximum phrase length like 5 words.

Compare. Click on the Compare button to look for all phrases that occur in both books or parts of books.



Num	Len	Phrase	1.Freq	2.Freq	1.RFreq	2.RFreq	1.Exp	2.Exp	LL	BIC	SMP	%Diff	MI	χ^2	Risk	LRisk	DiffC	Odds	ELL
1	5	"and it came to pass"	87	15	1,232.52	175.64	46.16	55.84	70.86	58.90	4.83	601.72	0.91	7.02	7.02	2.81	0.75	7.02	0.00011854
2	5	"the land of egypt and"	37	5	524.18	58.55	19.01	22.99	34.04	22.08	3.94	795.30	0.96	8.96	8.95	3.16	0.80	8.96	0.00007410
3	5	"it came to pass when"	27	5	382.51	58.55	14.48	17.52	21.11	9.15	3.04	553.33	0.90	6.54	6.53	2.71	0.73	6.54	0.00005062
4	5	"in the land of canaan"	23	5	325.84	58.55	12.67	15.33	16.22	4.27	2.69	456.54	0.86	5.57	5.57	2.48	0.70	5.57	0.00004096

Results. You will see the longest phrases (e.g., 5) that occur significantly more in book 1 than in book 2 (e.g., 1 > 2). You can select different lengths of phrases (1 to 5). You can also select different comparisons. Select "2 > 1" to see phrases that occur significantly more often in book 2 than in book 1. Select "1 ≈ 2" to see phrases that occur statistically about the same number of times in both books.

We use the BIC column to identify phrases that do or do not occur statistically more often in book 1 or book 2. If the BIC number is 2 or greater, the phrase appears in the "1 > 2" list. If BIC is -2 or less, the phrase appears in the "2 > 1" list. If BIC is between -2 and 2, the phrase appears in the "1 ≈ 2" list. Each list is sorted by the SMP_{100} column.

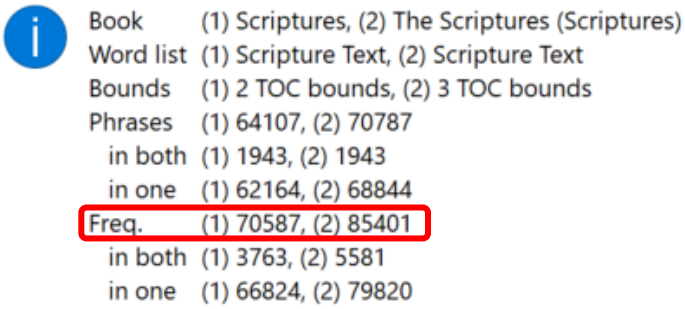
The columns before the red line always appear. Those after the red line only appear if you check the box by "Display all statistics." Each of the columns will be explained briefly below.

Save Results. Click on the Save Results button at the bottom of the results. Select Copy, Print, or Export all. The copy and print options include only the selected phrases.

The *export* option saves all phrases appearing in the results window 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.

Observed Frequencies (1.Freq, 2.Freq)

Phrase Compare statistics are all based on counting how many times each word or phrase is found in the text. If you click on the ∑ button, you will see how many different phrases of the selected length (e.g., 5) are each book.



In the 1.Freq and 2.Freq columns above, "and it came to pass" occurs 87 times in book 1 and 15 times in book 2. Book 1 has 64,107 different five word phrases that occur 70,587 times.

These observed results can be put in a table. The other numbers are calculated by adding or subtracting the observed results shown in bold blue numbers. In the table, o12 refers to the observed times the phrase occurs in book 2. R1 is the number of times the phrase occurs in both books. C1 is the total number of five word phrases in book 1. Total is how many five word phrase occur in both books.

OBSERVED frequencies	*Book 1*	Book 2	
and it came to pass	o11 = **87**	o12 = **15**	R1 = 102
Other phrases	o21 = 70,500	o22 = 85,386	R2 = 115,886
	C1 = **70,587**	C2 = **85,401**	Total = 115,988

Relative Frequencies (1.RFreq, 2.RFreq)

Relative frequencies estimate how many times each phrase would occur if both books contained exactly one million phrases. For example, if a phrase occurred 10 times in a 100 words (10%), and 10 times in 1000 words (1%), the observed frequencies are the same, but the relative frequencies are 100,000 and 10,000 respectively.

In the above example, book 2 is about 20% longer than book 1. The table below calculates the relative frequencies by dividing 87 by the total phrases (e.g., o11/C1 and o12/C2) to find the percent. This percent is then multiplied by 1,000,000 to find the relative frequency or frequency per million.

RELATIVE frequencies	*Book 1*	Book 2	
and it came to pass	r11 = (o11/C1) * 1,000,000 r11 = (87/70587) * 1,000,000 r11 = **1232.52**	r12 = (o12/C2) * 1,000,000 r12 = (15/85401) * 1,000,000 r12 = **175.64**	R1 = 1408.16
Other phrases	r21 = 998,767.47	r22 = 999824.36	R2 = 1,998,591.84
	C1 = 1,000,000	C2 = 1,000,000	Total = 2,000,000

Expected Frequencies (1.Exp, 2.Exp)

Expected frequencies are based on the probability of a result times the number of tries. For example, if you flip a coin 10 times, the *expected* number of *heads* would be 5 since the probability of *heads* is 50%.

The probability of being in either book is R1/Total. The number of tries is the number of phrases in each book (C1 or C2). Therefore, the expected value for the first cell (e11) is 46.16 (102 * 0.000654).

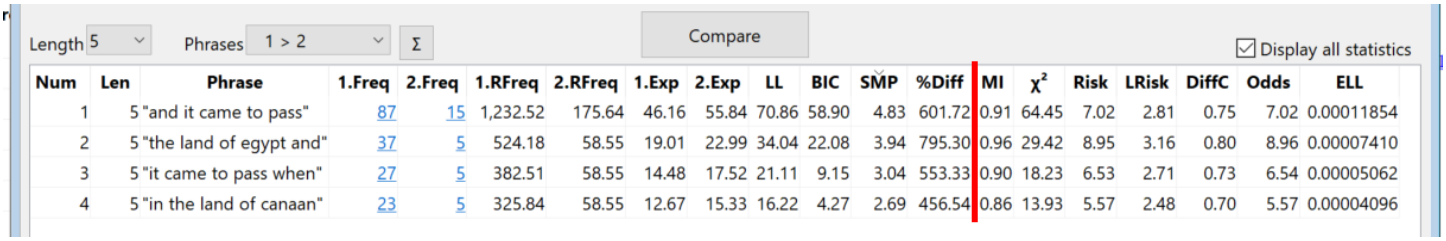
EXPECTED frequencies	Book 1	Book 2		Probability
and it came to pass	e11 = 46.16	e12 = 55.84	R1 = 102	0.000654
Other phrases	e21 = 70540.84	e22 = 85345.16	R2 = 115,886	0.999346
	C1 = 70,587	C2 = 85,401	Total = 115,988	

The following table shows how the expected values are calculated.

	Book 1	Book 2		Probability
Phrase	$e_{11} = C_1 \times \frac{R_1}{Total}$	e12 = R1 - e11	R1	$P_1 = \frac{R_1}{Total}$
Other phrases	e21 = C1 - e11	e22 = C2 – e12	R2 = total – R1	$P_2 = 1 - P_1$
	C1	C2	Total = C1 + C2	

Column Headings and Statistics

The last six columns after the red line are shown if "Display all statistics" is checked. These statistics are discussed below.



Num	Len	Phrase	1.Freq	2.Freq	1.RFreq	2.RFreq	1.Exp	2.Exp	LL	BIC	SMP	%Diff	MI	χ^2	Risk	LRisk	DiffC	Odds	ELL
1	5	"and it came to pass"	87	15	1,232.52	175.64	46.16	55.84	70.86	58.90	4.83	601.72	0.91	64.45	7.02	2.81	0.75	7.02	0.00011854
2	5	"the land of egypt and"	37	5	524.18	58.55	19.01	22.99	34.04	22.08	3.94	795.30	0.96	29.42	8.95	3.16	0.80	8.96	0.00007410
3	5	"it came to pass when"	27	5	382.51	58.55	14.48	17.52	21.11	9.15	3.04	553.33	0.90	18.23	6.53	2.71	0.73	6.54	0.00005062
4	5	"in the land of canaan"	23	5	325.84	58.55	12.67	15.33	16.22	4.27	2.69	456.54	0.86	13.93	5.57	2.48	0.70	5.57	0.00004096

Types of statistical measures

Statistical *association measures* are used to compare the frequencies of words or phrases in two different books, sections, or corpora. High scores identify words or phrases that occur more often in one book than another. Low scores identify words or phrase with different frequencies that may be due to chance.

Some statistics below measure the *effect size* (e.g., SMP, %Diff) to answer questions like "how much bigger is o_{11} than e_{11} (e.g., o_{11} / e_{11})" or "how much bigger is r_{11} than r_{12}."

Other statistics measure *statistical significance* (e.g., LL, BIC, χ^2) to answer questions like "is O_{11} significantly bigger E_{11}?"

For more information, see *Statistics in Corpus Linguistics: A Practical Guide* by Vaclav Brezina, "Log-likelihood and effect size calculator," its downloadable spreadsheet (http://ucrel.lancs.ac.uk/people/paul/SigEff.xlsx), and "Keyness Analysis: nature, metrics and techniques" by Costas Gabrielatos.

Significance Measures

Statistical *significance measures* are based on hypothesis tests. For example, is there less than a 5% probability that the difference in frequencies for a phrase in two books would be due to chance?

	Description of Significance Measures	Formula
LL	Log-likelihood (LL) is used to determine if differences in observed frequencies are statistically significant or due to chance. LL is interpreted using the χ^2 distribution with one degree of freedom: 3.84 ($p < 0.05$), 6.63 ($p < 0.01$), 10.83 ($p < 0.001$), and 15.13 ($p < 0.0001$). **Problem:** LL is a two-sided test that assigns high positive scores when observed results (O_{11}) are much greater ***or less*** than expected (E_{11}). **Solutions:** • Use LL to identify *phrases* that occur significantly more in one book or the other. • Multiply LL by -1 if $O_{11} < E_{11}$ for sorting, but use the absolute value |LL| for significance.[1]	$2 \times \left(o_{11} \times \log_e \frac{o_{11}}{e_{11}} + o_{12} \times \log_e \frac{o_{12}}{e_{12}} \right)$ • If $O_{11} = 0$, the two parts of the formula with O_{11} are set to 0. The same applies for O_{12}.
BIC	The Bayesian Information Criterion (BIC) of Bayes Factor is based on the Log Likelihood and can be interpreted as shown. < 0: No difference 0–2: Small difference 2–6: Different 6–10: Very different > 10: Very, very different **Problem:** BIC scores greater than 2 indicate a difference, but they don't indicate which book has a greater frequency. **Solutions:** • The list "1 > 2" includes phrases with BIC >= 2, and the relative frequency in book 1 was greater than in book 2. • The list "1 < 2" means BIC >= 2, and the relative frequency in book 2 was greater than in book 1. • The list "1 ≈ 2" means BIC < 2, and the relative frequency in book 2 was greater than in book 1.	$LL - \log_e Total$
χ^2	**χ^2** or chi-square is a widely used significance test. Chi-square is interpreted using the χ^2 distribution with one degree of freedom: 3.84 ($p < 0.05$), 6.63 ($p < 0.01$), 10.83 ($p < 0.001$), and 15.13 ($p < 0.0001$). If **χ^2** is greater than 15.13, the difference is significant at the 0.0001 level ($p < 0.0001$). **Problem:** A large chi-square value identifies phrases that occur more ***or less*** times in book 1. Chi-square is not reliable if o_{11} or o_{12} are small (e.g., < 5) especially if $C_1 + C_2$ is very large. This happens frequently in WordCruncher books and in corpus linguistics. **Solutions:** • Use chi-square, LL, or BIC to select significant differences. • Sort the list using SMP or another effect statistic that shows *more* or *less*.	$\frac{(o_{11} - e_{11})^2}{e_{11}} + \frac{(o_{12} - e_{12})^2}{e_{12}} + \frac{(o_{21} - e_{21})^2}{e_{21}} + \frac{(o_{22} - e_{22})^2}{e_{22}}$ • This formula does not show the *Yates correction* that is added.

Effect Size Measures

<table>
<tr><th></th><th>Description of Effect Size Measures</th><th>Formula</th></tr>
<tr><td>SMP

SMP₁
SMP₁₀
SMP₁₀₀
SMP₁₀₀₀</td><td>The Simple Maths Parameter (SMP) compares the relative frequency of each word or phrase in two books or corpora. A constant (k) is added to each frequency to avoid division by 0 and to focus on phrases that with a relative frequency greater than k.

If k = 100, the SMP₁₀₀ highlights relative frequencies greater than 100 as shown in the table below. The constant k is typically 1, 10, 100, or 1000. See columns SMP₁, SMP₁₀, SMP₁₀₀, and SMP₁₀₀₀.

<table>
<tr><th colspan="2">Relative Freq.</th><th colspan="4">K</th></tr>
<tr><th>r11</th><th>r12</th><th>1</th><th>10</th><th>100</th><th>1000</th></tr>
<tr><td>2</td><td>1</td><td>1.50</td><td>1.09</td><td>1.01</td><td>1.00</td></tr>
<tr><td>20</td><td>10</td><td>1.91</td><td>1.50</td><td>1.09</td><td>1.01</td></tr>
<tr><td>200</td><td>100</td><td>1.99</td><td>1.91</td><td>1.50</td><td>1.09</td></tr>
<tr><td>2000</td><td>1000</td><td>2.00</td><td>1.99</td><td>1.91</td><td>1.50</td></tr>
</table>

The Phrase Compare filter “1 > 2” includes phrases if BIC >= 2, and then it is sorted by SMP₁₀₀ scores to see phrases that occur more in book 1 than in book 2.</td><td>$\frac{r_{11}+k}{r_{12}+k}$</td></tr>
<tr><td>%Diff</td><td>%Diff is the percent difference between the relative frequency of a phrase in book 1 and in book 2. Each of the SMP examples above would have the same %Diff of 100.0.

<table>
<tr><th colspan="2">Relative Freq.</th><th rowspan="2">%Diff</th></tr>
<tr><th>r11</th><th>r12</th></tr>
<tr><td>100</td><td>100</td><td>0</td></tr>
<tr><td>200</td><td>100</td><td>100.0</td></tr>
<tr><td>100</td><td>400</td><td>-75.0</td></tr>
<tr><td>800</td><td>100</td><td>700.0</td></tr>
</table></td><td>$\frac{(r_{11}-r_{12})*100}{r_{12}}$</td></tr>
<tr><td>Risk</td><td>Relative Risk or Ratio estimates the likelihood of a phrase being in book 1 or book 2. If Risk = 1, the likelihood is the same. If Risk > 1, the phrase is more likely to be in book 1. If Risk = 2, the phrase occurs twice as often in book 1.

<table>
<tr><th colspan="2">Relative Freq.</th><th rowspan="2">Risk</th></tr>
<tr><th>r11</th><th>r12</th></tr>
<tr><td>100</td><td>100</td><td>1</td></tr>
<tr><td>200</td><td>100</td><td>2</td></tr>
<tr><td>100</td><td>400</td><td>0.25</td></tr>
<tr><td>800</td><td>100</td><td>8</td></tr>
</table></td><td>$\frac{r_{11}}{r_{12}}$

• If r_{11} is 0, $0.5/C_1$ is used instead.
• If r_{12} is 0, $0.5/C_2$ is used instead.</td></tr>
<tr><td>LRisk</td><td>Log Relative Risk (LRisk) or Log Ratio focuses on the magnitude of the difference. If a phrase occurs more in book 2, LRisk is negative.

<table>
<tr><th colspan="2">Relative Freq.</th><th rowspan="2">LRisk</th></tr>
<tr><th>r11</th><th>r12</th></tr>
<tr><td>100</td><td>100</td><td>0</td></tr>
<tr><td>200</td><td>100</td><td>1</td></tr>
<tr><td>100</td><td>400</td><td>-2</td></tr>
<tr><td>800</td><td>100</td><td>3</td></tr>
</table></td><td>$\log_2(\frac{r_{11}}{r_{12}})$</td></tr>
</table>

<table>
<tr>
<td>DiffC</td>
<td>Difference Coefficient (DiffC) looks at the difference in relative frequencies tor the same phrase in different books.

DiffC varies between +1 and -1. A positive value indicates $r_{11} > r_{12}$; a negative value indicates $r_{11} < r_{12}$. If a phrase occurs only in book 1, DiffC = 1. If it occurs only in book 2, DiffC = -1.

<table>
<tr><th colspan="2">Relative Freq.</th><th rowspan="2">DiffC</th></tr>
<tr><th>r11</th><th>r12</th></tr>
<tr><td>100</td><td>100</td><td>0</td></tr>
<tr><td>200</td><td>100</td><td>0.33</td></tr>
<tr><td>100</td><td>400</td><td>-0.66</td></tr>
<tr><td>800</td><td>100</td><td>0.78</td></tr>
</table>
</td>
<td>$$\frac{r_{11} - r_{12}}{r_{11} + r_{12}}$$</td>
</tr>
<tr>
<td>Odds</td>
<td>Odds Ratio is the odds of a word being in book 1 divided by the odds of being in book 2.

In statistics the odds of an event is number of times it happens (o_{11}) divided by the number of times it doesn't happen (o_{21}). If the odds of my team winning a game are 2 to 1, the probability of my team winning (o_{11}=66.7%) is twice the probability of losing (o_{21}=33.3%).</td>
<td>$$\frac{\frac{o_{11}}{o_{21}}}{\frac{o_{12}}{o_{22}}} = \frac{\frac{o_{11}}{C_1 - o_{11}}}{\frac{o_{12}}{C_2 - o_{12}}} = \frac{o_{11} \times o_{22}}{o_{21} \times o_{12}}$$
If o_{12} is 0, 0.00001 is used instead.
</td>
</tr>
<tr>
<td>MI</td>
<td>Mutual Information (MI) is a well-known measure of effect-size and is easily interpreted. If observed over expected (o_{11}/e_{11}) is 10, the word or phrase occurs 10 times more often than expected by chance. Since o_{11}/e_{11} can be very big, $log_2\, o_{11}/e_{11}$ is used. If MI=1, o_{11} occurs 2 times more than expected. If MI=2, o_{11} occurs 4 times more. If MI=8, o_{11} occurs 256 times more than expected.

Positive MI scores identify words or phrases that occur more often in book 1. Negative MI scores identify words or phrases that occur more often in book 2.

Problem: Very low frequency words can have high MI scores. A descending sort of MI scores puts low frequency words or phrases near the top of the list. MI is very useful for identifying friends or collocates. However, it is not as useful as other columns for identifying phrases occurring more in book 1 than in book 2.

Solutions:

Remove low frequency words ($O_{11} < 2$–5).
Use MI to select phrases with MI > 3 or some other number. Sort the selected phrases by frequency, SMP, etc.

</td>
<td>$$\log_2 \frac{o_{11}}{e_{11}}$$</td>
</tr>
<tr>
<td>ELL</td>
<td>Effect size for Log Likelihood (ELL) varies between 0 and 1.

ELL is LL divided by the maximum possible LL. This gives a proportion of the maximum difference between observed and expected frequencies.</td>
<td>$$\frac{LL}{Total \times \log_e(\min(e_{11}, e_{12}))}$$
If e_{11} or e_{12} is 0, 0.0000001 is the minimum value used.
</td>
</tr>
</table>

[1] Evert, 2008, p. 21.