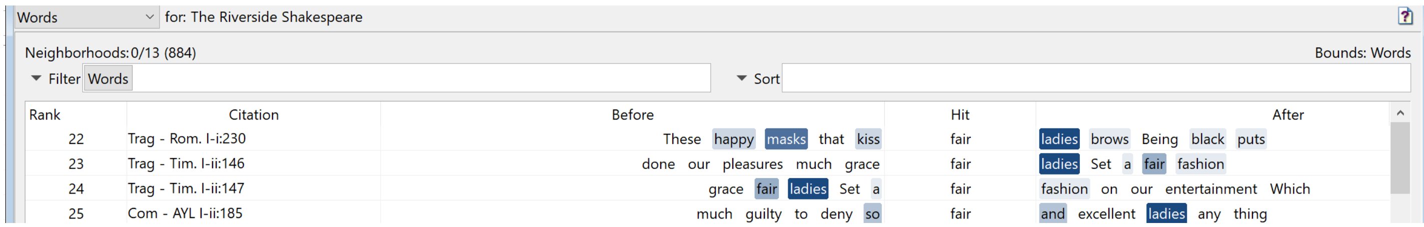
Neighbor or Collocate Statistics

Words tend to appear in combinations (e.g., *day/night*, *brush/teeth, ring/bell*). The meaning of a word (e.g., *ass*) depends on its context (e.g., "don't be an *ass*" and "he was riding an *ass*"). "You shall know a word by the company it keep!"[1]

The **Neighborhood tab** below is a Key Word In Context (KWIC) window that shows the keyword with its neighbors or collocates on each side. You can change the definition of a neighborhood from 5 words before and after (L5–R5).

In Shakespeare's collected works, *fair* occurs 884 times and *ladies* occurs 116 times. *Ladies* is a neighbor of *fair* 12 times. Below are four of the 12 neighborhoods where *ladies* appears. The highlighting identifies the neighbors identified as *friends* because they are seen together more often than expected. The darker the blue highlighting, the more frequently the word (e.g., *ladies*) is a neighbor.


Words for: The Riverside Shakespeare

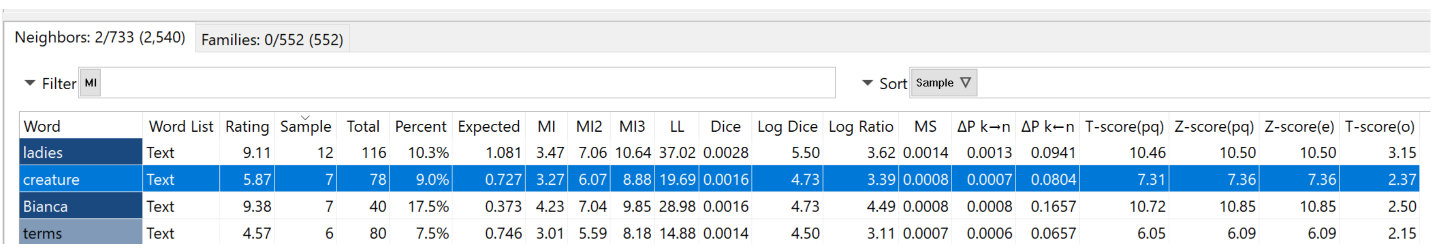
Neighborhoods: 0/13 (884) Bounds: Words

Filter Words Sort

Rank	Citation	Before	Hit	After
22	Trag - Rom. I-i:230	These happy masks that kiss	fair	ladies brows Being black puts
23	Trag - Tim. I-ii:146	done our pleasures much grace	fair	ladies Set a fair fashion
24	Trag - Tim. I-ii:147	grace fair ladies Set a	fair	fashion on our entertainment Which
25	Com - AYL I-ii:185	much guilty to deny so	fair	and excellent ladies any thing

The **Neighbors tab** below shows the neighbors or collocates that are considered *friends*. The *friends* are sorted by the number of times the word appears in a neighborhood. For example, *ladies* is in more neighborhoods than *creature* in the table below. You can click on a neighbor to see each of the neighborhoods it is in.

A *friend* is a neighbor with an MI score >= 3 and one that is in more than one neighborhood. You can change the definition by selecting another statistic (e.g., LL), setting a new minimum value (e.g., MI >= 5), or changing the minimum frequency (Freq >= 5).


Neighbors: 2/733 (2,540) Families: 0/552 (552)

Filter MI Sort Sample ▽

Word	Word List	Rating	Sample	Total	Percent	Expected	MI	MI2	MI3	LL	Dice	Log Dice	Log Ratio	MS	ΔP k→n	ΔP k←n	T-score(pq)	Z-score(pq)	Z-score(e)	T-score(o)
ladies	Text	9.11	12	116	10.3%	1.081	3.47	7.06	10.64	37.02	0.0028	5.50	3.62	0.0014	0.0013	0.0941	10.46	10.50	10.50	3.15
creature	Text	5.87	7	78	9.0%	0.727	3.27	6.07	8.88	19.69	0.0016	4.73	3.39	0.0008	0.0007	0.0804	7.31	7.36	7.36	2.37
Bianca	Text	9.38	7	40	17.5%	0.373	4.23	7.04	9.85	28.98	0.0016	4.73	4.49	0.0008	0.0008	0.1657	10.72	10.85	10.85	2.50
terms	Text	4.57	6	80	7.5%	0.746	3.01	5.59	8.18	14.88	0.0014	4.50	3.11	0.0007	0.0006	0.0657	6.05	6.09	6.09	2.15

Observed Results

When you click on the detail link by a neighbor in the above table, you will see tables like the following.

OBSERVED	*ladies*	Other Neighborhoods	
fair	o11 = 12	o12 = 872	R1 = 884
Other words	o21 = 104	o22 = 920,071	R2 = 920,175
	C1 = 116	C2 = 920,943	Total = 921,059 [1]

[1] This number uses a word-type filter set to Normal text to omit punctuation.

When the span is greater than one, there is a greater chance of words randomly being in a neighborhood. Some researchers use *corrected* values based on possible neighbors not just possible neighborhoods.

OBSERVED	*ladies*	Other Neighbors	Corrected
fair	o11 = 12	o12c = 8576	R1c = 8588 [2]
Other words	o21 = 104	o22c = 912,367	R2c = 912,471
	C1 = 116	C2 = 920,943	Total = 921,059

[3] If the window size or span is 10 (L5 + R5), the maximum number of neighbors (R1c) is 8840 (R1 * span). In this example the number is smaller because some neighborhoods overlap. Also the first and last words in the document can only have 5 neighbors. Below is an example of 6 overlapping neighborhoods with about 30 possible neighbors instead of 60.

> lord, and to all this fair company! fair desires, in all fair measure, fairly guide them! Especially to you, fair queen, fair thoughts be your fair pillow! *Helen.* Dear lord, you — Comedies, Troilus and Cressida III-i:43–47

Expected Results

When you click on the detail link by a neighbor in the above table, you will also see the four expected values. This is 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 a neighborhood is R1/Total. The number of tries is the number of times a word occurs in the entire book or corpus. Therefore, the expected value for the first cell (e11) is 0.111 (116 * 0.00096).

EXPECTED	***ladies***	**Other Neighborhoods**		**Probability**
fair	e11 = 0.1	e12 = 883.9	R1 = 884	0.00096
Other words	e21 = 115.9	e22 = 920,059.1	R2 = 920,175	0.99904
	C1 = 116	C2 = 920,943	Total = 921,059	

When the span is greater than one, there is a greater chance of words randomly being in a neighborhood. The *corrected* expected values are based on possible neighbors not just possible neighborhoods. Therefore, the expected value for the first cell (e11) is 1.081 (116 * 0.00932) which is about 10 times larger.

EXPECTED	***ladies***	**Other Neighbors**	**Corrected**	**Probability**
fair	e11 = 1.1	e12c = 8586.9	R1c = 8588	0.00932
Other words	e21 = 114.9	e22c = 912,356.1	R2c = 912,471	0.99068
	C1 = 116	C2 = 920,943	Total = 921,059	

Evert (2008) suggests using the corrected values for surface cooccurrence (e.g., in same neighborhood, L5–R5 span), and the uncorrected values for textual (in same sentence, paragraph, document, web page, …) and syntactic (e.g., verb-object [make + decision], adj. + noun [blue + coat]) cooccurrence.[2] The neighborhood report is an example of *surface* cooccurrence or collocation. Proximity searches can help identify *syntactic* cooccurrence, and level searches (e.g., within same paragraph), can help identify *textual* cooccurrence. Most programs including LancsBox use the uncorrected values.[3] WordCruncher uses the corrected values unless you choose the uncorrected values.

Formulas for Observed and Expected Values

The following tables shows how the observed and expected values are calculated. If you know four values (o11, R1, C1, Total), you can calculate the others. When doing When you analyze *textual* or *syntactic* cooccurrence, o11 is the number of search hits. The other numbers (R1, C1, Total) come from the WordWheel frequency column and "Freq. sum" at the bottom.

The detail report for a neighbor or collocate shows these four values and the other calculated values. These four values are usually reported in papers that discuss neighbors or collocates.

	Neighbor	**Other Neighborhoods**		**Probability**
Keywords	o11 $E_{11} = C_1 \times \frac{R_1}{Total}$	o12 = R1 - o11 e12 = R1 - e11	R1	$P_1 = \frac{R_1}{Total}$
Other words	o21 = C1 - o11 e21 = C1 - e11	o22 = o21 – R2 e22 = C2 – e12	R2 = total – R1	$P_2 = 1 - P_1$
	C1	C2 = C1 - total	Total	

When the span is greater than one, the four *corrected* values (o11, R1c, C1, Total) are used to calculate the others. The detail report for a neighbor or collocate shows these four values and the other calculated values.

	Neighbor	**Other Neighbors**	**Corrected**	**Probability**
Keywords	o11 $E_{11c} = C_1 \times \frac{R_{1c}}{Total}$	o12c = R1c - o11 e12c = R1c - e11	R1c [a]	$P_{1c} = \frac{R_1}{Total}$
Other words	o21 = C1 - o11 e21c = C1 - e11c	o22 = o21 – R2c e22c = C2 - e12c	R2c = total – R1c	$P_{2c} = 1 - P_{1c}$
	C1	C2 = C1 - total	Total	

[a] Span= (Ln + Rn). Maximum= (R1 * span) if no overlaps.

Column Headings and Statistics

You will see only the first four columns until you right-click on a column heading and show other columns.

Item	Neighbor	V sort *o11 Freq	Plot	*C1 Total	%	Word List	Filtr >=3 MI	MI2	MI3	LL	Dice	Log Dice	Log Ratio	Min Sens	ΔP k→n	ΔP k←n	Rating	T-scr (pq)	Z-scr (pq)	Z-scr (e)	T-scr (o)	Frq ←	Frq →
0	pride, ...	121																					
1	up	33		3,399	1.0	Scr Text	3.28	8.33	13.37	91.98	0.0144	7.88	3.30	0.0097	0.0247	0.0069	4.27	16.10					
2	lifted	23		237	9.7	Scr Text	6.60	11.13	15.65	167.76	0.0321	9.04	6.75	0.0192	0.0190	0.0969	10.00	46.73					
3	hearts	18		511	3.5	Scr Text	5.14	9.31	13.48	94.21	0.0211	8.43	5.19	0.0150	0.0146	0.0348	7.15	24.48					

	Description	Formula
Item	This is a sequential number. If you click on the item number, you will see a detail report showing how the statistics were calculated for that neighbor or collocate.	
Neighbor	This shows the neighbors or collocates (e.g., *ladies*) of your search word (e.g., *fair*) shown at the top of the list (see item 0).	
Freq	Frequency of collocations or the number of times a neighbor is in a neighborhood of the search word.	$O_{11} = O_{11\,before} + O_{11\,after}$
Freq ←	Frequency or number of times a neighbor appears ***before*** the search word in the neighborhoods.	$O_{11\,before}$
Freq →	Frequency or number of times a neighbor appears ***after*** the search word in the neighborhoods.	$O_{11\,after}$
Plot	This is a bar graph of the "Freq" column.	
Word Total	This is the number of times the neighbor (e.g., *ladies*) appears in the book or corpus (e.g., Riverside Shakespeare).	C_1
%	This is the percent of times a neighbor (e.g., *ladies*) is in the neighborhood of the search word (e.g., *fair*).	$\frac{O_{11}}{C_1} \times 100$
Word List	This column shows the word list (e.g., text, preface, footnotes) of the neighbor.	

Types of statistical measures

Statistical *association measures* are used to measure the attraction between words. High scores identify words that are good *friends* that are strongly attracted to each other. Low scores identify words with low attraction that are together only by chance.

Some statistics below measure the *effect size* (e.g., MI, Dice) to answer questions like "how strongly attracted are the words?" or "how much bigger is O_{11} than E_{11} (e.g., O_{11} / E_{11})."

Other statistics measure *significance* (e.g., LL, T-score, Z-score) to answer questions like "how much evidence of positive attraction exists regardless of the effect size?" and "is O_{11} significantly bigger E_{11}?"

Effect Size Measures

	Description of Effect Size Measures	Formula
MI Mutual Information	Mutual Information is the most well-known measure of *effect-size* (how tightly linked the words are) and is easily interpreted. If observed over expected (O / E) is 10, the two words occur together 10 times more often than expected by chance. Since O / E can be very big, log_2 O/E is used. If MI=1, O occurs 2 times more than expected. If MI=2 , O occurs 4 times more. If MI=8, O occurs 256 times more. Negative MI scores suggest the words 'repel' each other. **Problem:** Very low frequency words can have a high MI score. A descending sort of MI scores puts low frequency words near the top of the list. **Solutions:** • Remove low frequency words (O_{11} < 2–5). • Use other statistics (e.g., MI2, MI3). • Use MI to select *friends* and use a descending frequency sort to put low frequency words last.	$\log_2 \frac{O_{11}}{E_{11}}$
MI2	MI2 uses $O_{11}{}^2$ (O_{11} x O_{11}) to increase the score for higher frequency collocations. If O=1, O^2=1. If O=3, O^2=9. Thus, words that occur together more often, will have higher MI2 scores.	$\log_2 \frac{{O_{11}}^2}{E_{11}}$
MI3	MI3 uses $O_{11}{}^3$ (O_{11} x O_{11} x O_{11}) to increase the score for higher frequency collocations. If O=1, O^3=1. If O=3, O^3=27. Thus, words that occur together more often, will have higher MI3 scores.	$\log_2 \frac{{O_{11}}^3}{E_{11}}$
Dice proposed by Lee R. Dice[4]	The Dice score measures effect-size and is always between 0 and 1. It focuses on strong association not independence. It is the *harmonic mean* of the ratios O_{11}/R1 (the percent of all *fair* near *ladies*) and O_{11}/C1 (the percent of all *ladies* near *fair*). The Dice score is close to 1 if there is a strong prediction in both directions, from *fair* to *ladies* and vice versa. The score is much lower if the relation between the words goes in only one direction. Threshold: 0.1 (Smadja)	$\frac{2 \times \left(\frac{O_{11}}{R_1}\right) \times \left(\frac{O_{11}}{C_1}\right)}{\left(\frac{O_{11}}{R_1}\right) + \left(\frac{O_{11}}{C_1}\right)} = \frac{2 \times O_{11}}{R_1 + C_1}$
Log Dice	Dice gives good results, but the scores are usually very small numbers. The maximum *log dice* is 14 but most scores are less than 10. If the log dice value is 0, there is less than 1 cooccurrence per 16,000 R_1 or C_1. If one score is 1 more than another, it is two times bigger than the other.[5]	$14 + \log_2 \frac{2 \times O_{11}}{R_1 + C_1}$
Log Ratio	Log ratio[6] divides the probability of *ladies* occurring with *fair* by the probability of *ladies* **not** occurring with *fair*. If the log ratio is 2, *ladies* is 2 times more likely to occur with *fair* than not. If the log ratio is 4, *ladies* is 16 times more likely to occur with *fair* than not.	$\log_2 \frac{O_{11} \times R_2}{O_{21} \times R_1} = \log_2 \frac{\frac{O_{11}}{R_1}}{\frac{O_{21}}{R_2}}$
Min Sens	Minimum Sensitivity is the minimum of two ratios: (a) the probability of *ladies* occurring with *fair*, and the probability of *fair* occurring with *ladies*. If first ratio were 1, *ladies* would always occur with *fair*. If the second ratio were 1, *fair* would always occur with *ladies*.	$\min(\frac{O_{11}}{C_1}, \frac{O_{11}}{R_1})$

Directional Effect Size Measure

DeltaP is a *directional measure* of attraction that identifies words like *red herring* where the probability of *red* being a neighbor of *herring* is greater than the probability of *herring* being a neighbor of *red*.

	Description of Directional Effect Size Measure	**Formula**
ΔP k→n	DeltaP forward (ΔP k→n) is the probability of the neighbor (e.g., *ladies*) being in a neighborhood of the search word (e.g., *fair*) minus the probability of the neighbor (e.g., *ladies*) **not** being in the neighborhood. **Problem:** If corrected formula based on neighbors (R_{1c}, R_{2c}) not neighborhoods (R_1, R_2) is used, ΔP_c k→n is lower. For example, with a window size or span of 10 (L5–R5), ΔP_c k→n for *fair* → *ladies* is about one tenth of ΔP k→n.	$\frac{O_{11}}{R_1} - \frac{O_{21}}{R_2}$
ΔP k←n	DeltaP backward (ΔP k←n) is the probability of the search word (e.g., *fair*) having a neighbor (e.g., *ladies*) minus the probability of it having different neighbors.	$\frac{O_{11}}{C_1} - \frac{O_{12}}{C_2}$

Significance Measures

Statistical *significance measures* are based on hypothesis tests. For example, is there less than a 5% probability that the words (e.g., *fair* and *ladies*) would be neighbors by chance? Below are two Z-score and two T-score formulas that differ in how they compute the standard deviation or number on the bottom.

	Description of Significance Measures	**Formula**
LL	Log-likelihood is the most popular *significance* measure in computational linguistics. 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 *friends* and then sort by Frequency or another measure. • Multiply LL by -1 if $O_{11} < E_{11}$ for sorting, but use the absolute value |LL| for significance.[7]	$2 \times \left(\begin{array}{c} O_{11} \times \log_e \frac{O_{11}}{E_{11}} + O_{21} \times \log_e \frac{O_{21}}{E_{21}} + \\ O_{12} \times \log_e \frac{O_{12}}{E_{12}} + O_{22} \times \log_e \frac{O_{22}}{E_{22}} \end{array} \right)$
Z-score (e)	A |*z-score*| greater than 1.96 ($p < 0.05$) or 3.29 ($p < 0.001$) are generally considered statistically significant. However, most word pairs are highly significant.[8] **Problem:** Word pair data violates the z-score normality assumption since E is often less than 1. This results in inflated z-scores and a low-frequency bias similar to MI. **Solutions:** • Z-scores are used to rank or select word pairs, not to determine statistical significance. • Use other statistics (e.g., t-score).	$\frac{O_{11} - E_{11}}{\sqrt{E_{11}}}$

T-score (o)	The *t-score* is smaller than z-scores for low frequency words since O_{11} is always 1 or more. **Problem:** Word pair data violates the t-score normality assumption.[9] **Solutions:** • Interpret a t-score like a z-score without overestimating scores of low-frequency words.	$\frac{O_{11} - E_{11}}{\sqrt{O_{11}}}$
Z-score (pq)	The *z-score$_{pq}$* is based on binomial probability like flipping a coin with *n* trials (flips) and two outcomes (heads or tails). P is probability of success (heads or collocation), and Q (or 1-P) is the probability of failure. The mean (E_{11}) is n * P, and the variance is n * P * (1-P).[10] **Problem:** For low-frequency words (e.g., $E_{11} < 1$), multiplying E_{11} by 1-P makes it even smaller which results in higher z-scores.	$\frac{O_{11} - E_{11}}{\sqrt{E_{11} \times \left(1.0 - \left(\frac{C_1}{Total}\right)\right)}}$
T-score (pq)	The *t-score$_{pq}$* multiplies Bessel's correction,[11] N/(N-1), times the variance in the *z-score$_{pq}$* formula. If a word occurs only two times in a corpus, this multiples the variance by 2.0. If it occurs only once, it is multiplied times a very large number to simulate infinity (∞). This lowers the *t-score* for low-frequency words. A word that occurs once in a corpus will have a t-score of about 0. Earlier versions of WordCruncher used this formula.	$\frac{O_{11} - E_{11}}{\sqrt{\left(\frac{C_1}{C_1 - 1}\right) \times E_{11} \times \left(1.0 - \left(\frac{C_1}{Total}\right)\right)}}$
Rating	A *rating* based on *t-score$_{pq}$* was used to identify *friends* of search words in earlier versions of WordCruncher. A *z-score* of all of the *t-scores$_{pq}$* was calculated for each neighbor. Ratings greater than 10 were set to 10. Ratings less than -10 were set to -10. Each rating indicates how much a t-score$_{pq}$ is above or below the average t-score. A *friend* had to have a rating greater than 0 and appear in more than 1 neighborhood ($O_{11} > 1$). Neighbors were sorted by Rating and 10 shades of blue were used to highlight the friends in the neighborhood display.	$Rating = \frac{Tscore - Tscore_{avg}}{\sigma_{tscore}}$

[1] Firth, J. R. 1957. "A Synopsis of Linguistic Theory, 1930–1955," 11. In *Studies in Linguistic Analysis*. Special Volume of the Philological Society., pp. 1–32.
[2] Evert, Stefan. 2008, "Corpora and collocations," extended manuscript, 13 October 2007, p. 41. http://www.stefan-evert.de/PUB/Evert2007HSK_extended_manuscript.pdf. See also PDF of examples: http://www.stefan-evert.de/SIGIL/sigil_R/materials/collocations.slides.pdf.
[3] Brezina, McEnery, and Wattam. 2015. "Collocations in context: A new perspective on collocation networks," Appendix 1. 169–171. https://benjamins.com/catalog/ijcl.20.2.01bre/fulltext/ijcl.20.2.01bre.pdf
[4] Dice, Lee R. (1945). "Measures of the Amount of Ecologic Association between Species." *Journal of Ecology*, 26: 297–302.
[5] Rychlý, Pavel. "A Lexicographer-Friendly Association Score," p. 9. In P. Sojka & A. Horak (Eds.) *Proceedings of Recent Advances in Slavonic Natural Language Processing*, RASLAN 2008, pp. 6–9, 2008.
[6] http://cass.lancs.ac.uk/log-ratio-an-informal-introduction/
[7] Evert, 2008, p. 21.
[8] Evert, 2008, p. 20.
[9] Evert, Stefan. 2004, *The Statistics of Word Cooccurrences: Word Pairs and Collocations*, Dissertation, p. 83. https://elib.uni-stuttgart.de/bitstream/11682/2573/1/Evert2005phd.pdf
[10] Pascual Cantos Gomez, *Statistical Methods in Language and Linguistic Research*, 2013, p. 203. See also stattrek.com/probability-distributions/binomial.aspx (accessed 10/22/2019) for an explanation of binomial experiments and probability.
[11] https://en.wikipedia.org/wiki/Bessel%27s_correction