Phrase Compare TTR Statistics
(Type-Token Ratios, Lexical diversity)

In a book or corpus, some words are used frequently (e.g., and, of, the) while others are used only a few times. Lexical diversity statistics measure the vocabulary of a book, section, or corpus.

Dr. Seuss used 236 unique words (types) to write the 1,626 word (tokens) children's book titled *The Cat in the Hat.* Of the 236 unique words, 54 occur once and 33 occur twice. The most common words (e.g., the, and, I, not) occur more than 40 times.

In her first Harry Potter book, J. K. Rowling used 5,711 different words (types) to write the 80,592 word (tokens) book titled *Harry Potter and the Sorcerer's Stone*. In book 5, she used 11,949 different words to write the 265,944 word book titled *Harry Potter and the Order of the Phoenix*.[1]

When you do a Phrase Compare report, several statistics are computed to help you notice the lexical diversity of a book, section, or corpus. If you click on the ∑ button, you will see the following statistics for book 1 (Genesis and Exodus) and for book 2 (Leviticus, Numbers, and Deuteronomy).

TTR	(1) 0.04795, (2) 0.03977
Total Types	(1) 3402, (2) 3412
Total Tokens	(1) 70947, (2) 85789
CTTR	(1) 9.03134, (2) 8.23717
RTTR	(1) 12.77225, (2) 11.64912
LogTTR	(1) 0.72805, (2) 0.71614
Uber	(1) 17.83775, (2) 17.37957
Maas_a²	(1) 0.05606, (2) 0.05754
Maas_a	(1) 0.23677, (2) 0.23987
MSTTR	(1) 0.27014, (2) 0.25395
MATTR	(1) 0.26941, (2) 0.25224
MATTR StdDev	(1) 0.03289, (2) 0.04659
Segment size	(1) 1000, (2) 1000
Min TTR/segment	(1) 0.16400, (2) 0.09700
Max TTR/segment	(1) 0.36800, (2) 0.39700

Export TTR files. If you click on the "Save results" button near the bottom left corner of the window, you can export the phrase statistics to a CSV (utf16) or to a TXT (utf8). 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.

After the phrase statistics are exported, you will be asked if you want to export the type/token ratio statistics. If you do, (a) a TTRLevels file with the statistics for each level (e.g., book, chapter, paragraph), and (b) a TTRLog file with MATTR segment data will be exported for each book.

Each of the lexical diversity statistics will be introduced below. The following terms will be used in the formulas.

- ***Types*** refer to the unique word forms like *and, the,* or *is*. There are 4 types in the title "The Cat in the Hat."
- ***Tokens*** refer to the total words in a book, section, or corpus. There are 5 tokens in the title "The Cat in the Hat."

Type/Token Ratios (TTR, RTTR, CTTR)

The following measures are based on types and tokens. Each of these statistics are sensitive to the length of the text. They decrease as the text becomes longer and more words get used more often.[2] Therefore, they should only be used to compare books of the same length.

TTR 1957	Type/Token Ratio (TTR) is the percent of total words that are unique word forms. The average word frequency (AWF) is tokens divided by types or 1/TTR. For *The Cat in the Hat*, TTR = 236/1626 = 0.145. Therefore, AWF = 6.89.	$TTR = \frac{types}{tokens}$
RTTR 1960	Root Type/Token Ratio (TTR) For *The Cat in the Hat*, RTTR = 236/√1626 = 5.85.	$RTTR = \frac{types}{\sqrt{tokens}}$
CTTR 1964	Corrected Type/Token Ratio (TTR) For *The Cat in the Hat*, CTTR = 236/√(2*1626) = 4.14.	$CTTR = \frac{types}{\sqrt{2 \times tokens}}$

	Types	Tokens	TTR	AWF	RTTR	CTTR
The Cat in the Hat	236	1,626	0.145	6.89	5.85	4.14
Harry Potter and the Sorcerer's Stone	5,711	80,592	0.071	14.112	20.12	14.22
Genesis, Exodus	3,402	70,947	0.04795	20.854	12.77	9.03
Leviticus–Deut.	3,412	85,789	0.03977	25.143	11.65	8.24

Logarithmic Measures (LogTTR, Uber, Maas)

The following measures are also based on types and tokens. They each use base 10 logarithms. Logarithms reduce the size of a big number more than a small number. This reduces the effect of the number of tokens in very large books or corpora.

LogTTR
1960

Log Type/Token Ratio (LogTTR)

	Types	Tokens	$\log_{10}$ types	$\log_{10}$ tokens	LogTTR
The Cat in the Hat	236	1,626	2.373	3.211	0.739
Harry Potter	5,711	80,592	3.757	4.906	0.766
Gen., Ex.	3,402	70,947	3.532	4.851	0.728
Lev.–Deut.	3,412	85,789	3.533	4.933	0.716

$$\frac{\log_{10} types}{\log_{10} tokens}$$

Maas_a²
1972

Maas a² is a statistic proposed by Heinz-Dieter Maas.[3] He used $\log_{10}$.

Problems:

- Maas (1972) used $\log_{10}$ but some authors[4] use $\log_e$ instead.
- Some sources report a^2, but Maas (1972) reports a.
- As TTR increases, Maas a² decreases.

$$a^2 = \frac{\log_{10} tokens - \log_{10} types}{(\log_{10} tokens)^2}$$

Maas_a
1972

Maas (1972) reports a not a^2.

$$a = \sqrt{a^2}$$

Uber
1978

Uber, Dugast (Uber)

Uber = 1/Maas_a² and therefore increases as TTR increases.

$$\frac{(\log_{10} tokens)^2}{\log_{10} tokens - \log_{10} types}$$

	Types	Tokens	LogTTR	Uber	Maas a²	Maas a
The Cat in the Hat	236	1,626	0.739	12.302	0.081	0.285
Harry Potter and the Sorcerer's Stone	5,711	80,592	0.766	20.940	0.048	0.219
Genesis, Exodus	3,402	70,947	0.728	17.838	0.056	0.237
Leviticus–Deut.	3,412	85,789	0.716	17.380	0.058	0.240

Equal Segment Measures (MSTTR, MATTR)

The following statistics help compare TTR values from texts of different sizes because the same segment size (e.g., 1000 tokens) is used for each text.

Mean Segmental Type-Token Ratio (MSTTR)

MSTTR is the average TTR for each *non-overlapping* segment of equal size. If the segment size is 10 and we ignore case, the first 7 segments in KJV Genesis are shown below. Note: (a) Punctuation and verse numbers are automatically ignored, and (b) the last segment is ignored if it has less than 10 tokens.

Seg	Tokens	Tokens	Types	TTR
1	[1] in the beginning god created the heaven and the earth	10	8	.8
2	[2] and the earth was without form and void and darkness	10	8	.8
3	was upon the face of the deep and the spirit	10	8	.8
4	of god moved upon the face of the waters [3] and	10	8	.8
5	god said let there be light and there was light	10	9	.9
6	[4] and god saw the light that it was good and	10	9	.9
7	god divided the light from the darkness	7	6	ignore

MSTTR = sum of the complete segment TTRs divided by the number of complete segments.

MSTTR = (0.8 + 0.8 + 0.8 + 0.8 + 0.9 + 0.9) / 6
= 5.0 / 6 = 0.833

Moving Average Type-Token Ratio (MATTR)

MATTR is the average TTR for all possible *overlapping* segments of equal size. If the segment size is 10 and we ignore case, the first 7 segments in KJV Genesis are shown below. Each segment contains all but one word found in the previous segment. We have almost as many segments as tokens in the document. As before, we omit segments at the end of the document that have less than 10 tokens.

Seg	Tokens	Tokens	Types	TTR
1	[1] in the beginning god created the heaven and the earth	10	8	.8
2	the beginning god created the heaven and the earth [2] and	10	7	.7
3	beginning god created the heaven and the earth [2] and the	10	7	.7
4	god created the heaven and the earth [2] and the earth	10	6	.6
5	created the heaven and the earth [2] and the earth was	10	6	.6
6	the heaven and the earth [2] and the earth was without	10	6	.6
7	heaven and the earth [2] and the earth was without form	10	7	.7

MATTR = sum of the complete segment TTRs divided by the number of the complete segments.

MATTR = (0.8 + 0.7 + 0.7 + 0.6 + 0.6 + 0.6 + 0.7) / 7
= 4.7 / 7 = 0.67

In the examples below, the segment size is 1000. For MATTR you will also see the standard deviation as well as the minimum and maximum TTR values for individual segments.

	Types	Tokens	MSTTR	MATTR	StdDev	Minimum	Maximum
Genesis, Exodus	3,402	70,947	0.270	0.269	0.0329	0.164	0.368
Leviticus–Deut.	3,412	85,789	0.254	0.252	0.0466	0.097	0.397

More TTR Information: If you want to see more TTR detail, you can click on the "Save results" button and select "Export all (Phrase Compare)." After the phrases are exported, click Yes when you see "Export Type-Token-Ratio (TTR) files?" This will export two files for Book 1 (TTRLog-1, TTRLevels-1) and for Book 2 (TTRLog-2, TTRLevels-2).

The TTRLevel files show the TTR information for segments starting in each level (e.g., book, chapter, section, paragraph). They also show the citation of the first segment containing the minimum TTR (Ex. 36:30) and maximum TTR (Gen. 48:18).

The TTRLog files show TTR information for all possible segments. This file shows the z-score for each segment.

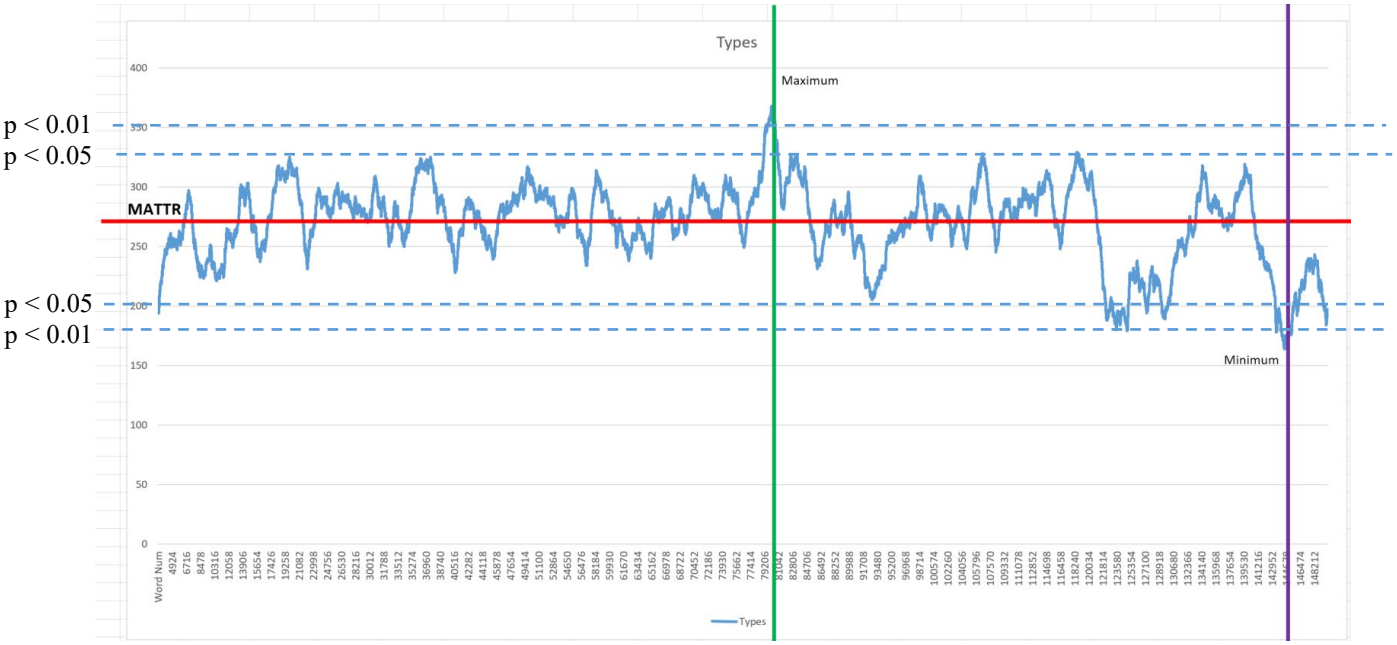
$$\text{z-score} = (\text{TTR} - \text{MATTR}) / \text{StdDev}$$

For the first segment in this example, the z-score is -2.232 which means the first TTR (0.196) is over 2 standard deviations below the MATTR (0.269). The following symbols indicate the statistical significance of the z-scores.

- >>> for z-scores > 3.291 ($p < 0.001$).
- >> for z-scores > 2.576 ($p < 0.01$).
- > for z-scores > 1.960 ($p < 0.05$).
- < for z-scores < -1.960 ($p < 0.05$).
- << for z-scores < -2.576 ($p < 0.01$).
- <<< for z-scores < -3.291 ($p < 0.001$).

The $p < 0.05$ value means only about 5% of the segments will have at least one > or < symbol in this column.
The $p < 0.01$ value means only about 1% of the segments will have at least two > or < symbols in this column.

Below is an Excel chart showing the TTR values for each segment in Genesis and Exodus. The solid lines indicate the minimum and maximum values as well as the average or MATTR (0.269). The dashed lines show the statistical significance or critical p-values. This helps identify parts of the text with significantly more or less unique words per 1000 and consequently less or more repetition of those words.



NOTE: Excel spreadsheets can only have about one million rows. Therefore, if you open a TTRLog file for a book with more than one million tokens, it will only show the first million segments.

---

[1] Shaffer, J.D. "The potential use of the Harry Potter book series for incidental vocabulary acquisition." *The Journal of Shizuoka University Education*. 2018 (14), 67–82.

[2] Torruella and Capsada, "Lexical Statistics and Tipological Structures: A Measure of Lexical Richness," Procedia: Social and Behavioral Sciences, Volume 95, 25 October 2013, pages 449.

[3] See page 78, formula 2.13 and results on page 79 of Maas, H. D. (1972). Zusammenhang zwischen Wortschatzumfang und Länge eines Textes. *Zeitschrift für Literaturwissenschaft und Linguistik*, vol. 2, Iss. 8, (Jan. 1,1972), 73-79.

[4] Malvern, Richards, Chipere, and Durán. *Lexical Diversity and Language Development*, (2004), p. 185 (in back matter). Torruella and Capsada, "Lexical Statistics and Tipological Structures: A Measure of Lexical Richness," Procedia: Social and Behavioral Sciences, Volume 95, 25 October 2013, pages 447–454.