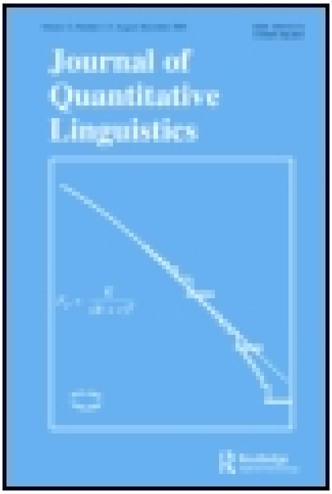

Journal of Quantitative Linguistics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/njql20

Cutting the Gordian Knot: The Moving-Average Type-Token Ratio (MATTR)

Michael A. Covington [a] & Joe D. McFall [a]

[a] Institute for Artificial Intelligence, The University of Georgia , USA

Published online: 14 May 2010.

To cite this article: Michael A. Covington & Joe D. McFall (2010) Cutting the Gordian Knot: The Moving-Average Type-Token Ratio (MATTR), Journal of Quantitative Linguistics, 17:2, 94-100, DOI: 10.1080/09296171003643098

To link to this article: http://dx.doi.org/10.1080/09296171003643098

Journal of Quantitative Linguistics
2010, Volume 17, Number 2, pp. 94–100
DOI: 10.1080/09296171003643098

Routledge
Taylor & Francis Group

Cutting the Gordian Knot: The Moving-Average Type–Token Ratio (MATTR)*

Michael A. Covington and Joe D. McFall
Institute for Artificial Intelligence, The University of Georgia, USA

ABSTRACT

Type–token ratio (TTR), or vocabulary size divided by text length (V/N), is a time-honoured but unsatisfactory measure of lexical diversity. The problem is that the TTR of a text sample is affected by its length. We present an algorithm for rapidly computing TTR through a moving window that is independent of text length, and we demonstrate that this measurement can detect changes within a text as well as differences between texts.

INTRODUCTION

Type–token ratio (TTR), or vocabulary size divided by text length (V/N), is a time-honoured but unsatisfactory measure of lexical diversity, used in literary studies (Holmes, 1985), studies of child language (Richards, 1987), and psychiatry (where perseveration or overassociation is an important symptom [Manschreck et al., 1981]).

The problem is that the TTR of a text sample is affected by its length; obviously, the longer the text goes on, the more likely it is that the next word will be one that has already occurred.

*Address correspondence to: Michael A. Covington, Institute for Artificial Intelligence, The University of Georgia, Athens, GA 30602-7415 USA. E-mail: mc@uga.edu
An earlier version of this paper was presented as a poster at the 2008 Annual Meeting of the Linguistic Society of America. Prototype software to compute MATTR is available at: http://www.ai.uga.edu/caspr. We thank Reinhard Köhler for helpful suggestions.

No solution to this problem has gained universal acceptance. Proposed solutions fall into several main categories:

- Standardizing the length of text samples, unsatisfactory because (for example) the first 1000 words of a 10,000-word text are not semantically or pragmatically comparable to a 1000-word text that stands on its own.
- Transforming the TTR in some way that should make it immune to sample length. The "logarithmic TTR", $\log V/\log N$, of Herdan (1960, 1966) is one popular approach; the function $V/\sqrt{2N}$ of Carroll (1964) is another, and Guiraud (1959) advocates $V/\sqrt{N}$. (For a review, see Wachal and Spreen, 1973.) Hess et al. (1986, 1989) found that none of these adjustments actually makes TTR independent of text length.
- Adjusting the computed value at each point so that (for example) it is based on the number of types and tokens found thus far, plus those expected in the rest of the text if the text is uniform (Köhler & Galle, 1993). This works well for quantities such as verb–adjective ratio, to keep the graph from jumping up and down wildly at the beginning of the text before it stabilizes. However, as Köhler and Galle acknowledge, for TTR this type of correction does not solve the problem because, unlike the properties of being a verb or adjective, the property of being a new token is actually a property of the preceding text ("x is a new token" = "there is nothing like x before this point").
- Measuring the TTR for a variety of text lengths and fitting a parameter describing the relation between vocabulary size and text size. Such parameters include K (Yule, 1944), D or *vocd* (Malvern & Richards, 2002), the Tornquist function of Tuldava (1995) and Panas (2001), or the interpolation function of Müller (2002); for others see Tweedie and Baayen (1998). Such a parameter can be useful for distinguishing short- from long-term repetition, a point to which we shall return, but it incorporates statistical assumptions and is not directly equivalent to type–token ratio.
- Plotting some other cumulative function of vocabulary size vs. text length, such as the vocabulary management profile (VMP) of Youmans (1991). Such a plot is useful for tracking changes within a text but has the obvious disadvantage that the scale or significance of the graph changes as one moves across it.

We cut the Gordian knot by computing and averaging the moving-average type–token ratio (MATTR). This approach was also advocated by Köhler and Galle (1993). We choose a window length (say 500 words) and then compute the TTR for words 1–500, then for words 2–501, then 3–502, and so on to the end of the text. The mean of all these TTRs is a measure of the lexical diversity of the entire text and is not affected by text length nor by any statistical assumptions. Further, the individual TTRs can be compared in order to detect changes within the text.

MATTR is more informative than the mean segment TTR (MSTTR) introduced by Johnson (1944), advocated by Schach (1987), cited by Köhler and Galle (1993), discussed by Malvern and Richards (2001), and implemented in *WordSmith 4.0* (Oxford University Press). MSTTR is computed on successive non-overlapping segments of the text whereas MATTR uses a smoothly moving window. Thus MATTR yields a value for every point in the text except for those less than one window length from the beginning, while MSTTR is only a stepwise approximation to this. Thus MATTR is better for tracking changes within texts, and MATTR is not affected by accidental interactions between segment boundaries and text unit boundaries.

ALGORITHM FOR RAPID COMPUTATION

Crucially, the computation of the $N - W + 1$ individual TTRs, for a text of length N with window size W, is not $N - W + 1$ times as much work as computing just one of them. It is appreciably less because, each time the window moves one step, only one word enters it and one word leaves it. Having computed a word-frequency table for the first window position, one only needs to adjust two items in it every time the window advances.

Our implementation is a C# program that uses the built-in hashtable data structure of Microsoft .NET Framework (Microsoft, 2007; the newer Dictionary data structure would work equally well). Each element of the hashtable is a key-value pair; in our case, a word and its frequency. Thanks to hash coding, elements can usually be added, retrieved, and removed in constant time.

Let P denote the position of the current window, initially 1 (i.e. starting with the first word). The first step is to compute a full word-frequency table for the first W words (i.e. words P to $P+W-1$) and store it in the hashtable.

Then, increment P by 1 repeatedly until $P=N-W+1$ (the final or rightmost window position). At each step:

- Word $P-1$ has just left the window. Decrement its word count in the hashtable. If the result is zero, remove the hashtable entry.
- Word $P+W-1$ has just entered the window. Increment its word count in the hashtable, or if it does not have one, make a new hashtable entry for it.

The TTR at every position P is the number of distinct hashtable entries divided by W. Finally, take the mean of all the TTRs computed.

Using a 2.66-GHz dual-core Pentium computer and a 100-word window size, our prototype implementation processes text at a rate of more than 100,000 words per second. Clearly, the computation of MATTR is not prohibitively slow.

WINDOW SIZE

Obviously, the moving-average TTR of a text varies with the window size more or less the same way that the conventional TTR varies with the text length. Empirically, for typical English text, MATTR $\approx 2\ W^{-0.2}$, so with window sizes of 100 and 500 words, typical MATTRs are 0.8 and 0.6 respectively.

Thus, for reproducible results, a standard window size must be chosen. How big should it be? Smaller than the smallest text to be processed, but large enough to provide a meaningful measure of style. If $W=1$, the MATTR is always 1.0 and is useless. We suggest a window size of 500 words for stylometric analysis. If the primary goal is to determine the size of the author's vocabulary, a much larger window, as large as 10,000 words, may be more appropriate, provided the length of the texts permits it (Tuldava, 1995, pp. 133–134).

A short window, perhaps as short as 10 words, is appropriate if the goal is to detect repetition of immediately preceding words or phrases due to dysfluent production. In fact, the ratio of MATTRs with two different window sizes is a potentially useful indication of whether repetition occurs over short or long spans. To see why this is so, consider the two sequences:

a b c d e f g a b c d e f g

a a b b c c d d e e f f g g

Clearly, both have the same TTR as a whole, and nearly so when measured with a sufficiently large window, but with a small window, the TTR of the first text rises to 1 and that of the second text does not. In general, a high ratio of small-window to large-window MATTR indicates that the TTR of a text is being lowered by short-span repetition rather than topic perseveration or small vocabulary.

The window size can interact with periodicities present in the text itself. Consider the artificial text:

a a a a b b b b a a a a b b b b a a a a b b b b

With a window size of four or less, the MATTR periodically dips to a very low value as the window coincides with a series of identical tokens;

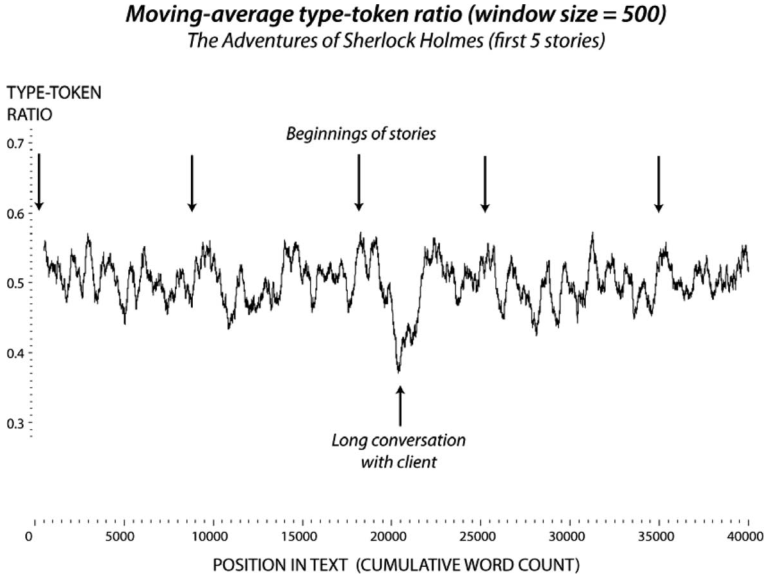

Fig. 1. Plot of the moving-average type–token ratio (MATTR) for a literary text.

with a window size of five or more, the MATTR is uniform. This is an exaggerated example of something that could happen in a text with a repetitious internal structure, such as a book of essays or news stories of uniform length, or perhaps a book of sonnets alternating with translations of the same sonnets into a different language.

TRACKING CHANGES WITHIN A TEXT

Figure 1 shows a plot of the moving-average TTR (with window size 500) of the first five stories in *The Adventures of Sherlock Holmes* (A. Conan Doyle, 1892, text from www.gutenberg.org). As expected, the MATTR rises at the beginning of every story, since new vocabulary is introduced there, but also rises equally high elsewhere. More strikingly, there is a passage with low MATTR which turns out to be Holmes' long conversation with a client (in "A case of identity"), discussing a single situation at length, using relatively simple language and not introducing new vocabulary. This demonstrates that MATTR is a useful measure of changes of style within a text.

REFERENCES

Carroll, J. B. (1964). *Language and Thought*. Englewood Cliffs, NJ: Prentice-Hall.

Guiraud, P. (1959). *Problèmes et méthodes de la statistique linguistique*. Dordrecht: Reidel.

Herdan, G. (1960). *Type-token Mathematics*. The Hague: Mouton.

Herdan, G. (1966). The advanced theory of language as choice and chance. In *Kommunikation und Kybernetik in Einzeldarstellungen*, Band 4. New York: Springer.

Hess, C. W., Sefton, K. M., & Landry, R. G. (1986). Sample size and type-token ratios for oral language of preschool children. *Journal of Speech and Hearing Research*, *29*, 129–134.

Hess, C. W., Haug, H. T., & Landry, R. G. (1989). The reliability of type-token ratios for the oral language of school age children. *Journal of Speech and Hearing Research*, *32*, 536–540.

Holmes, D. I. (1985). The analysis of literary style – a review. *Journal of the Royal Statistical Society*, *148* (part A), 328–341.

Johnson, W. (1944). Studies in language behavior. I: A program of research. *Psychological Monographs 56.2* (255), 1–15. (Reprinted in his *People in Quandaries: The Semantics of Personal Adjustment*, pp. 499–518. New York: Harper.)

Köhler, R., & Galle, M. (1993). Dynamic aspects of text characteristics. In L. Hřebíček & G. Altmann (Eds), *Quantitative Text Analysis* (pp. 46–53). Quantitative Linguistics, vol. 52. Trier: Wissenschaftlicher Verlag Trier (WVT).

Malvern, D., & Richards, B. (2002). Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing*, *19*, 85–104.

Manschreck, T. C., Maher, B. A., & Ader, D. N. (1981). Formal thought disorder, the type-token ratio and disturbed voluntary motor movement in schizophrenia. *British Journal of Psychiatry*, *139*, 7–15.

Microsoft Corporation (2007). Hashtable Class. Retrieved March 9, 2010, from http://msdn2. microsoft.com/en-us/library/system.collections.hashtable.aspx

Müller, D. (2002). Computing the type token relation from the *a priori* distribution of types. *Journal of Quantitative Linguistics*, *9*, 193–214.

Panas, E. (2001). The generalized Torquist: Specification and estimation of a new vocabulary text-size function. *Journal of Quantitative Linguistics*, *8*, 233–252. ("Torquist" here = "Tornquist" in Tuldava 1995.).

Richards, B. (1987). Type/token ratios: what do they really tell us? *Journal of Child Language*, *14*, 201–209.

Schach, E. (1987). Empirische Eigenschaften der TTR bet ausgewählten Texten. In K. R. Wagner (Ed.), *Wortschatz-Erwerb* (pp. 102–114). Arbeiten zur Sprachanalyse, vol. 6. Bern: Lang.

Tuldava, J. (1995). On the relation between text length and vocabulary size. In J. Tuldava (Ed.), *Methods in Quantitative Linguistics* (pp. 131–149). Quantitative Linguistics, vol. 54. Trier: Wissenschaftlicher Verlag Trier (WVT).

Tweedie, F., & Baayen, R. H. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities*, *32*, 323–352.

Wachal, R. S., & Spreen, O. (1973). Some measures of lexical diversity in aphasic and normal language performance. *Language and Speech*, *16*, 169–181.

Youmans, G. (1991). A new tool for discourse analysis: The vocabulary-management profile. *Language*, *67*, 763–789.

Yule, G. U. (1944). *The Statistical Study of Literary Vocabulary*. Cambridge: Cambridge University Press.