Estimating Lexical Diversity Using the Moving Average Type-Token Ratio (MATTR): Pros and Cons

Yves Bestgen

Psychological Science Institute, Université catholique de Louvain

This study is an extended and in-depth version of a point made in Appendix S1 in the supporting information of Bestgen (2024) published in Language Learning. This issue has been the subject of intense discussion with a reviewer and this brief report aims to clarify it.

This is the preprint version. The final version includes an in-depth discussion of Bestgen (2024). With regard to the empirical analyses, that version shows that the findings of the hapax analysis apply to four corpora of learners of four different languages (L2): Czech, English, German, and Italian. If you wish to access this final version, please contact me.

Acknowledgements: Yves Bestgen is Research Associate of the National Fund for Scientific Research (F.R.S.- FNRS).

Declaration of interests: I have nothing to declare

Yves Bestgen, Université catholique de Louvain, Place Cardinal Mercier, 10 B-1348 Louvain-la-Neuve Belgium. yves.bestgen@uclouvain.be

Estimating Lexical Diversity Using the Moving Average Type-Token Ratio (MATTR): Pros and Cons

Abstract

Several recent studies have strongly recommended the use of the moving average type-token ratio (MATTR) to estimate the lexical diversity (LD) of a text because it is the only length-insensitive index that can compare texts of different sizes. After pointing out that a length-insensitive index was proposed in the 1960s and is still being used, I analyse the properties of the MATTR computational procedure that enable it to control for the effects of length. This index is an excellent choice for evaluating the fluctuation of the LD throughout a relatively long text. However, its use for evaluating the overall LD of a text is questionable because the impact of tokens on the score varies according to their position in the text. I illustrate this problem using pseudo-texts and show that this impact is likely to affect a significant proportion of texts by analysing the distribution of hapaxes in texts by learners of Italian (IT) as a second language (L2).

Keywords: Lexical diversity, language assessment, exact probabilistic inference

Introduction

Estimating the lexical diversity (LD) of a text has been one of the main areas of research in quantitative linguistics since the beginning of the 20th century (Thomson & Thompson, 1915). The main reason for this interest is that estimating LD is a much more complex problem than one might first think due to the non-linear relationship between the number of different words and the total number of words contained in a text. Many studies have been conducted to identify LD indices that enable the comparison of texts of different lengths (see Cosette, 1994, and Tweedie & Baayen, 1998, for syntheses). This work is particularly important because LD indices are often used in applied linguistics, especially for the assessment of foreign language learners (e.g., Kyle *et al.*, 2024; Lei & Yang, 2020; Ma *et al.*, 2023; Vidal & Jarvis, 2020). In this field, the texts analysed are almost always relatively short and vary in length. This length is an important predictor of text quality (Bulté & Housen, 2014). It is therefore essential to be able to assess lexical diversity independently of differences in length. A recent series of studies in applied linguistics has strongly recommended the use of the moving average type-token ratio (MATTR) proposed by Covington and McFall (2010). These authors have argued that, unlike other indices such as the hypergeometric distribution D (HD-D), MATTR is the only index that eliminates the impact of text length (Akbary

& Jarvis, 2023; Fergadiotis *et al.*, 2015; Kyle et al., 2024; Kubát, 2014; Lei & Yang, 2020; Shi & Lei, 2022; Vidal & Jarvis, 2020; Wang & Liu, 2018; Zenker & Kyle, 2021). This conclusion is problematic for two reasons. Firstly, Muller proposed a length-insensitive index as early as 1964. This index used the binomial law to determine the number of different words that each of the compared texts would contain if its size were reduced to a given length; that is, at most the shortest of the texts. As confirmed by Cossette (1994) and Baayen (2001), this index is an excellent approximation of the index that can be derived from the hypergeometric distribution, which is better suited to the problem than is the binomial distribution. Baayen (2001: 63-69) provides a mathematical demonstration of the insensitivity of this index to length. This index can be seen as a generalisation of two classic indices, Yule's K and Simpson's D. In applied linguistics, the hypergeometric version of this index, HD-D[1], was rediscovered by McCarthy and Jarvis (2007) during their in-depth study of D, the well-known index proposed by Malvern *et al.* (2004).

The second problem with recommending the use of MATTR as the LD index is that it is based solely on the traditional question of the sensitivity of indices to the length of the texts that are being compared, and the evaluation of MATTR's mathematical properties has been neglected. It is therefore necessary to ask whether this index has any properties that should be taken into account when using it, which is the purpose of this study. It is worth emphasising that the research question is not the impact of the text length on the index, but the impact on the LD scores on the way in which length is controlled by the algorithm.

How MATTR is Calculated and Used

For a text of *N* tokens, MATTR is the average of the type-token ratios (TTRs) obtained by moving a window of size *n* from token to token, from the first token to the *N*-*n*+1 token. The formula is:

$$\sum_{i=1}^{N-n+1} \frac{V_i}{n \times (N-n+1)}$$

where *n* is the window size and V_i is the number of types in the ith window.

MATTR can be used for two different purposes. Firstly, it allows for the tracking of the fluctuation of the LD in a text using the TTR of each window as an index. The profile of the text can then be represented graphically using a typical moving average chart. The only limitation of MATTR in this regard is that it does not provide information about *n*-1 points in the text. In line

[1] Studies that suggested HD-D was sensitive to text length reached this erroneous conclusion mainly because they employed an inadequate evaluation method (Bestgen, 2024).

with Covington and McFall (2010), we can assume that these are the first *n*-1 tokens. The smaller the *n* that is compared to *N*, the less information is lost. This information lost is therefore negligible when texts are long, as is usually the case for this type of analysis.

Of note, MATTR is most often used to estimate the LD of a text when it is considered as a whole, especially in applied linguistics. It is in this context that the question of the impact of text length arises. Like HD-D, MATTR reduces all texts to the same length and can therefore be used to compare texts of different lengths (Bestgen, 2024). However, MATTR does not perform this reduction using an exact probabilistic inference. It is therefore necessary to analyse the effects of its computational procedure on the estimation of the LD, as will be described in the following sections.

MATTR's Local Approach

The MATTR computational procedure generates two important differences from HD-D. The first is that MATTR determines the total LD of a text based on the local LD, since it does not take repetitions at a distance greater than *n* into account. HD-D is a global measure because it takes all repetitions into account, regardless of how far apart they may be. This is an important difference that has nevertheless received little attention. As pointed out above, MATTR's limited-memory approach is particularly justified for long texts, such as novels, even though the repetition of a single word several hundred words apart can be particularly significant (Eco, 1984).

When the text is a few hundred words long at most and an overall score is required, as was the case in almost all the studies mentioned in the references, the preference for a local or a global index must depend on the LD construct (Jarvis, 2013), which alone can decide whether repetitions at a distance greater than *n* should be taken into account when estimating the LD of a text. This is a complex and intrinsically multidisciplinary issue. While linguists obviously have opinions concerning the determination of the extent to which previously read tokens affect subsequent cognitive processing, so do cognitive psychologists.

Differences in Token Handling

Another consequence of the MATTR computational procedure is that the frequency with which a token is used to compute LD varies according to its position in the sequence. This is obviously not the case with HD-D, which is based on the full type distribution. This property is rarely discussed in the literature (but see Appendix S1 in the supporting information of Bestgen, 2024). The remainder

of this section explains in detail the origin of this difference in treatment and demonstrates its impact on the estimation of the LD using pseudo-texts.

Formulation

Due to the use of a moving-window average, not all the tokens in a text occur in the same number of windows. Tokens from position n to N-n+1 are all involved in n windows. Conversely, the other tokens appear in a smaller number of windows. This number is equal to the difference in absolute value plus one between their position and the closest extreme position (1 or N). The first and last tokens are therefore only involved in one window, while the tokens in positions n-1 and N-n+2 are involved in n-1 windows.

Impact

The TTRs calculated for the N-n+1 windows are averaged when MATTR is used to calculate the overall LD score of a text. The difference in token handling described previously means that not all tokens have the same impact on this average, as this depends on their position in the text. A token affects all the windows in which it occurs, by adding a type if it is the only occurrence of that type in a window, and adding none if it is not. The greater the number of windows in which a token occurs, the greater its impact. It would be a mistake to think that MATTR's weighting problem can be solved simply by filling in the windows at the end of the text with the tokens from the beginning of the text. Such an approach is problematic because it negates the sequentiality that is the central feature of MATTR. In addition, it will have a greater effect on short texts, creating a dependency on text length. Similarly, weighting the start and end windows differently is not a solution because the tokens in these windows are not used the same number of times.

This difference in token handling affects all types, regardless of their frequency in the text, but it can be better explained by hapaxes. Hapaxes always increase the LD of the windows in which they occur. It follows that, all things being equal, for a text containing a greater proportion of hapaxes amongst the first n-1 tokens and/or amongst the last n-1 tokens, the LD will be underestimated compared to a text containing the same proportion of hapaxes, but with the hapaxes occurring in the middle of the text.

To illustrate this impact, it is necessary to be able to move the tokens in a text without such movement affecting the calculations other than due to the fact that these tokens are at the end or in the middle. This is impossible with natural texts because MATTR is a local index that takes the context in which a token is used into account. Here, ‘context’ refers to all the tokens in the same window. Moving tokens affects this context. A token that was the only occurrence of a type in a

window, which therefore increased its TTR, can be moved to a window containing one or more other occurrences of the same type. In this case, it no longer increases the TTR of the window. It follows that none of the sampling procedures that have been developed to assess the impact of length, such as parallel sampling (Hess *et al.*, 1989), random sampling (Cossette, 1994) and alternate token random sampling (Bestgen, 2024), can be used to study the impact of the difference in token handling.

However, it is possible to study this impact in isolation by using the types of pseudo-texts that Covington and McFall (2010) used to discuss the properties of MATTR. This can be achieved using a 'text' of *N* tokens containing a maximum of *n*-2 hapaxes, with the remainder of the tokens being of the same type (see Table 1 for an example). Regardless of the window in which a hapax occurs, it is obviously a hapax. There will always be at least two repeated tokens in a window; therefore, adding to or removing a repeated token from a window will never alter the number of types it contains. If a repeated token is swapped with a hapax, this increases the windows that contained the repeated token by one type and decreases the windows that contained the hapax by one type, regardless of the other tokens that are present in the windows in question. If there are the same number of windows in both cases, the total MATTR score will not change. However, if the number of windows differs, the MATTR score will be different. As explained above, there will not be the same number of windows when the two tokens that have been exchanged are not at the same distance from the closest extreme text position and at least one is within *n* tokens of the closest extreme position.

The shortest possible pseudo-text to illustrate this impact has been constructed on this basis. Its parameters are *N*=5, *n*=3, #hapax=1 and #repeated=4.

Table 1: MATTR computation for a pseudo-text

```
Text:     a r r r r
Window 1: -----     = 2 types
Window 2:   -----   = 1 type
Window 3:     ----- = 1 type
MATTR = 4 / 3 / 3 = 0.44
```

Sequence	#types in each window	MATTR
arrrr	2 1 1	0.44
rarrr	2 2 1	0.56
rrarr	2 2 2	0.67
rrrar	1 2 2	0.56
rrrra	1 1 2	0.44

As shown in Table 1, the MATTR score for this example ranges from 0.44 to 0.67 depending on the position of the hapax. Figure 1 illustrates the same effect using a pseudo-text of a more usual length in the field, with *N*=200, *n*=50, #hapax=48 and #repeated=152, by moving the continuous set of hapaxes from the beginning of the text to the end thereof. The figure also shows the TTR and the HD-D, which are obviously constant due to their global nature. As can be seen, MATTR ranges between 0.176 (when all the hapaxes are at one of the ends) and 0.338 (when they are all in the median zone). An even lower MATTR score can be achieved by placing half of the hapaxes at the very beginning and the other half at the very end. In this case, MATTR is equal to 0.10.

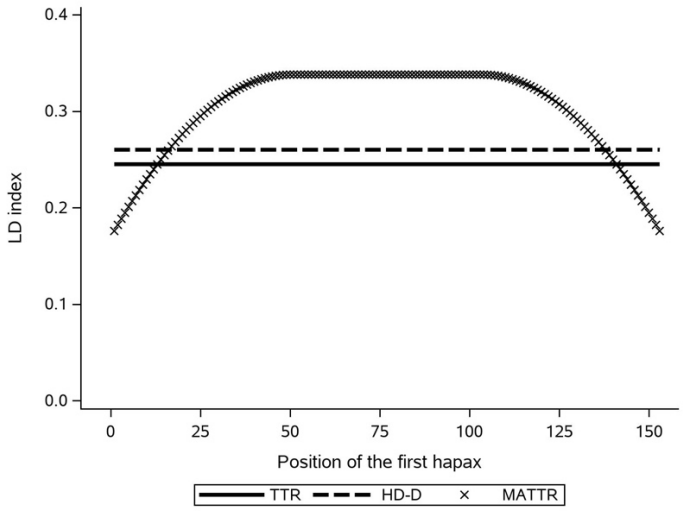

Figure 1: The MATTR scores for a pseudo-text in which 48 hapaxes were moved from the beginning to the end

Since hapaxes represent a significant proportion of the words used in a text, this effect may be quite large. However, this reasoning is based on the hypothesis that hapaxes are not distributed uniformly in texts. To confirm this, I analysed a corpus of texts written by learners of English to compare the proportion of tokens that were hapaxes in the first 49 tokens[2] and in the last 49 tokens to the same proportion that was found in the rest of the text; that is, in the median part of the text.

The 180 texts used in this analysis were taken from the MERLIN corpus (Boyd *et al.*, 2014), and are freely available for research. All these texts contained at least 147 tokens[3] and were written by learners of Italian (IT) as a second language (L2). For each text, the hapaxes were identified and counted in the delimited three sections explained above.

[2] Forty-nine is the number of tokens at the ends of a text that occur in a reduced number of windows when *n* is fixed at 50, which is the usual value.
[3] The minimum length of 147 was chosen to ensure that the middle part contained at least 49 tokens.

Figure 2 shows a comparison of the average proportion of hapaxes in the two sections at the ends of the text to the proportion in the middle. The left-hand side of the figure shows these values for all 180 texts, while the right-hand side only shows the results for the 10% of the texts that had the most positive differences and the 10% with the most negative differences. This figure clearly shows that the hapaxes were not distributed uniformly across all the texts; they were more frequent at the extremities in some texts, while the opposite was true for other texts. The figure on the right-hand side shows that a large effect could be observed in a significant proportion of texts.

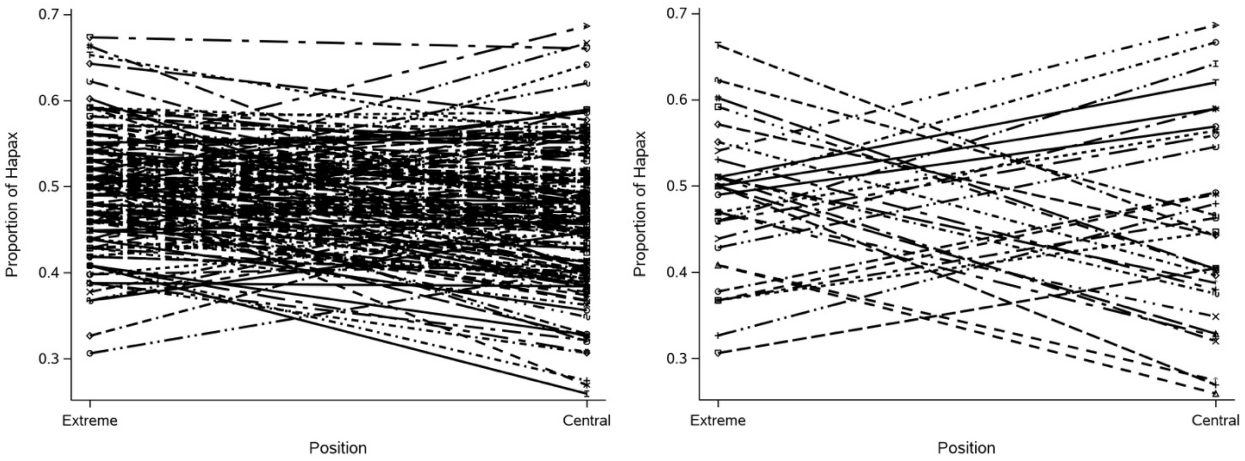

Figure 2: Proportion of hapaxes at the ends of a text and in the middle thereof

Discussion and Conclusion

The aim of this study was to evaluate the advantages and disadvantages of MATTR for analysing the LD of texts. MATTR's computational procedure makes it an excellent index for evaluating the fluctuation of the LD throughout a text, particularly in sufficiently long texts; that is, when the length is much greater than is that that of the window that is being moved. Conversely, using MATTR to estimate the LD of a text as a whole is only justified when researchers intend to investigate a local point of view. If this property is not relevant, HD-D should be preferred, since it allocates the same importance to all the tokens in a text, unlike MATTR. It is worth noting that HD-D is as easy to interpret as MATTR. Whereas MATTR is the average TTR of a continuous text segment of length n, HD-D is the average TTR of all samples of n tokens that can be extracted from that text. This study highlighted the interest in further evaluating the computational properties of LD indices such as MATTR, as well as the mean segmental type-token ratio (MSTTR) and the measure of textual lexical diversity (MTLD). All these indices are local, and evaluate the LD on the basis of text segments. The relevance of indices with a limited span for estimating the LD of a text is, to my knowledge, a field of research that has received virtually no attention. However, it is an

important question for the definition of the LD construct itself. This question could be answered by determining whether this local factor influences the human ratings of the LD of natural or linguistically manipulated texts (Jarvis, 2017). Local indices can also provide other potentially useful information. For example, Kubát (2014) proposed comparing the LD of texts based on the distributions of scores obtained in the different MATTR windows of each text, and argued that this type of information was much more useful than was an overall score for each text. In summary, local indices, such as MATTR, are useful tools for the study of LD in applied linguistics, but they need to be used according to their properties.

References

Akbary, M. & Jarvis, S. (2023). Lexical diversity as a predictor of genre in TV shows. *Digital Scholarship in the Humanities*, *38*, 921–936. https://doi.org/10.1093/llc/fqad004

Baayen, R. H. (2001). *Word frequency distributions*. Springer. https://doi.org/10.1007/978-94-010-0844-0

Bestgen, Y. (2024), Measuring lexical diversity in texts: The twofold length problem. *Language Learning*, *74*(3), 638-671. https://doi.org/10.1111/lang.12630

Boyd, A., Hana, J., Nicolas, L., Meurers, D., Wisniewski, K., Abel, A., Schone, K., Stindlova, B. & Vettori, C. (2014). The MERLIN corpus: Learner language and the CEFR. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, 1281–1288.

Bulté, B. & Housen, A. (2014). Conceptualizing and measuring short-term changes in L2 writing complexity, *Journal of Second Language Writing, 26*, 42–65. doi: 10.1016/j.jslw.2014.09.005

Cossette, A. (1994). *La Richesse lexicale et sa mesure*. Genève-Paris: Slatkine-Champion.

Covington, M.A. & McFall, J.D. (2010). Cutting the Gordian knot: The moving-average type-token ratio (MATTR). *Journal of Quantitative Linguistics, 17* (2), 94-100. https://doi.org/10.1080/09296171003643098

Eco, U. (1984). Postille a Il nome della rosa. Bompiani.

Fergadiotis, G., Wright, H.H. & Green, S.B. (2015). Psychometric evaluation of lexical diversity indices: Assessing length effects. *Journal of Speech, Language, and Hearing Research, 58* (3), 840-852. https://doi.org/10.1044/2015_JSLHR-L-14-0280

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* (3), 536-540. https://doi.org/10.1044/jshr.3203.536

Jarvis, S. (2013). Capturing the diversity in lexical diversity. *Language Learning, 63*(S1), 87-106. https://doi.org/10.1111/j.1467-9922.2012.00739.x

Jarvis, S. (2017). Grounding lexical diversity in human judgments. *Language Testing, 34*(4), 537-553. https://doi.org/10.1177/0265532217710632

Kubát, M. (2014). Moving window type-token ratio and text length. In Altamnn, G., Čech, R., Mačutek, J., Uhlířová, L. (Eds.), *Empirical Approaches to Text and Language Analysis*. Lüdenscheid: RAM, 105–113.

Kyle, K., Sung, H. Eguchi, M. & Zenker, F. (2024). Evaluating evidence for the reliability and validity of lexical diversity indices in L2 oral task responses. *Studies in Second Language Acquisition*, *46*, 278-299 https://doi.org/10.1017/S0272263123000402

Lei, S. & Yang, R. (2020). Lexical richness in research articles: Corpus-based comparative study among advanced Chinese learners of English, English native beginner students and experts. *Journal of English for Academic Purposes*, *47*, 100894. https://doi.org/10.1016/j.jeap.2020.100894

Ma, H., Wang, J., & He, L. (2023). Linguistic features distinguishing students' writing ability aligned with CEFR levels, *Applied Linguistics,* 1–21. https://doi.org/10.1093/applin/amad054

Malvern, D., Richards, B., Chipere, N. & Durán, P. (2004). *Lexical Diversity and Language Development: Quantification and Assessment*. Palgrave MacMillan. https://doi.org/10.1057/9780230511804

McCarthy, P.M. & Jarvis, S. (2007). vocd: A theoretical and empirical evaluation. *Language Testing, 24* (4), 459-488. https://doi.org/10.1177/0265532207080767

Muller, C. (1964). Calcul des probabilités et calcul d'un vocabulaire. Reproduit dans *Langue française et linguistique quantitative*. Genève-Paris: Slatkine-Champion, 167-176.

Shi, Y. & Lei, L. (2022). Lexical Richness and Text Length: An Entropy-based Perspective, *Journal of Quantitative Linguistics*, *29*(1), 62-79. https://doi.org/10.1080/09296174.2020.1766346

Thomson, G. H., and Thompson, J. R. (1915). Outlines of a method for the quantitative analysis of writing vocabularies. *British Journal of Psychology*, *8*, 52-69.

Tweedie, F.J. & Baayen, R.H. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers & the Humanities, 32*, 323-352. https://doi.org/10.1023/A:1001749303137

Vidal, K., & Jarvis, S. (2020). Effects of English-medium instruction on Spanish students' proficiency and lexical diversity in English. *Language Teaching Research*, *24*(5), 568-587. https://doi.org/10.1177/1362168818817945

Wang, Y. & Liu, H. (2018). Is Trump always rambling like a fourth-grade student? An analysis of stylistic features of Donald Trump's political discourse during the 2016 election. *Discourse & Society*, *29*(3) 299–323.

Zenker, F. & Kyle, K. (2021). Investigating minimum text lengths for lexical diversity indices. *Assessing Writing, 47*, Article 100505. https://doi.org/doi:10.1016/j.asw.2020.100505