JOURNAL OF THE HISTORY OF ECONOMIC THOUGHT
PUBLISHED FOR THE HISTORY OF ECONOMICS SOCIETY

Online Appendix to the JHET Article: "Towards a 'Text as Data' Approach in the History of Economics: An Application to Adam Smith's Classics," by Matthieu Ballandonne and Igor Cersosimo

March 2023 issue (DOI: https://doi.org/10.1017/S1053837222000104)

We provide here additional details on the methods used and further quantitative discussion on the results.

I. TOPICS DETECTION – KEYNESS

Keyness is a measure of how characteristic a word is to a document (indicated as *target*), when compared to all other documents in a corpus (indicated as *reference*). It is obtained from the calculation of the chi-square statistics on the table of the number of occurrences of a word (compared to all other words) in the *target* document and in the rest of the corpus. Words with a very large and positive keyness value are characteristic to the *target* document; words with a large and negative value appear frequently in the corpus (in our case, in the *reference* document) but not in the *target* document; words with keyness values close to zero appear with a similar relative frequency in the *target* document and in the rest of the corpus.

We used keyness to find characteristic words in the *TMS* and *WN*. These words are presented in Table 2 in the article while the corresponding chi-square values are shown in Figure 1 below.

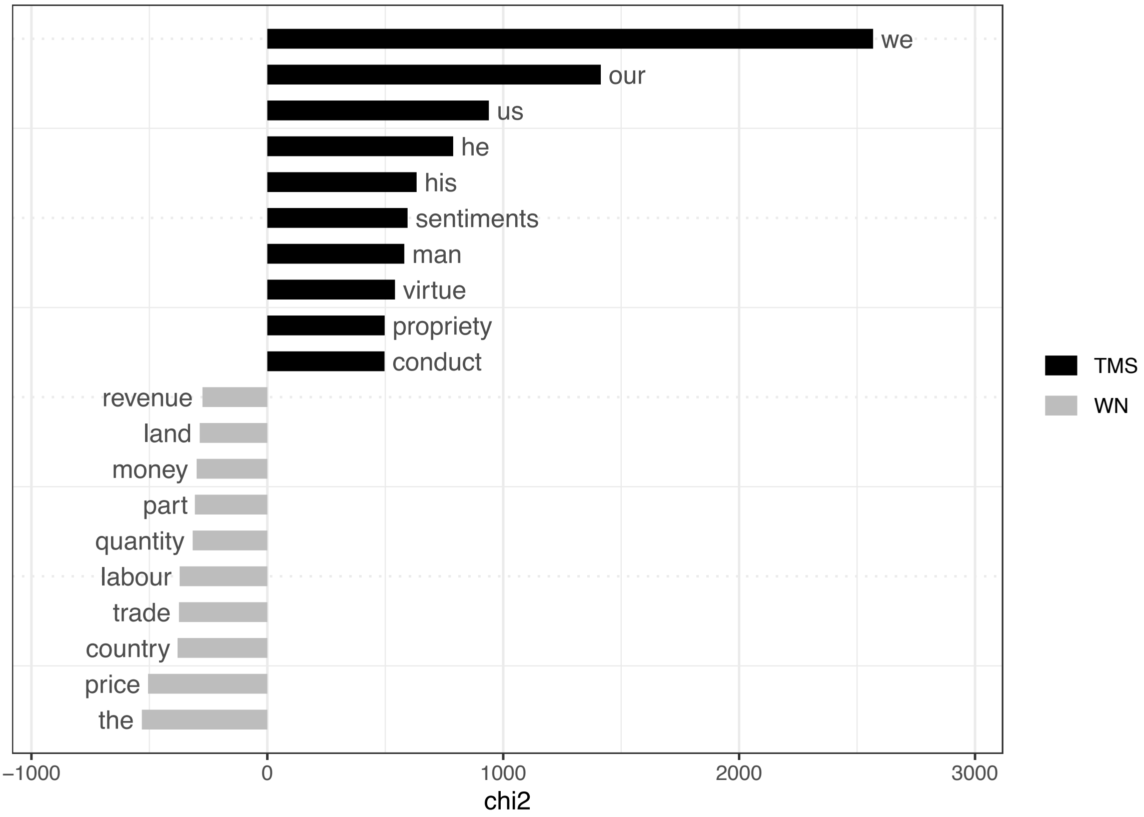

Figure 1: Graphical representation of the keyness values for words in *TMS* and *WN* (with stop words).

We also applied the same method to determine the words that are characteristic when Smith is examining a specific topic, in our case that of "war." The following two graphs represent keyness values for words that are characteristic, respectively, to sentences containing the terms "war" or "wars," versus sentences not containing these terms (Figure 2 and Figure 3).

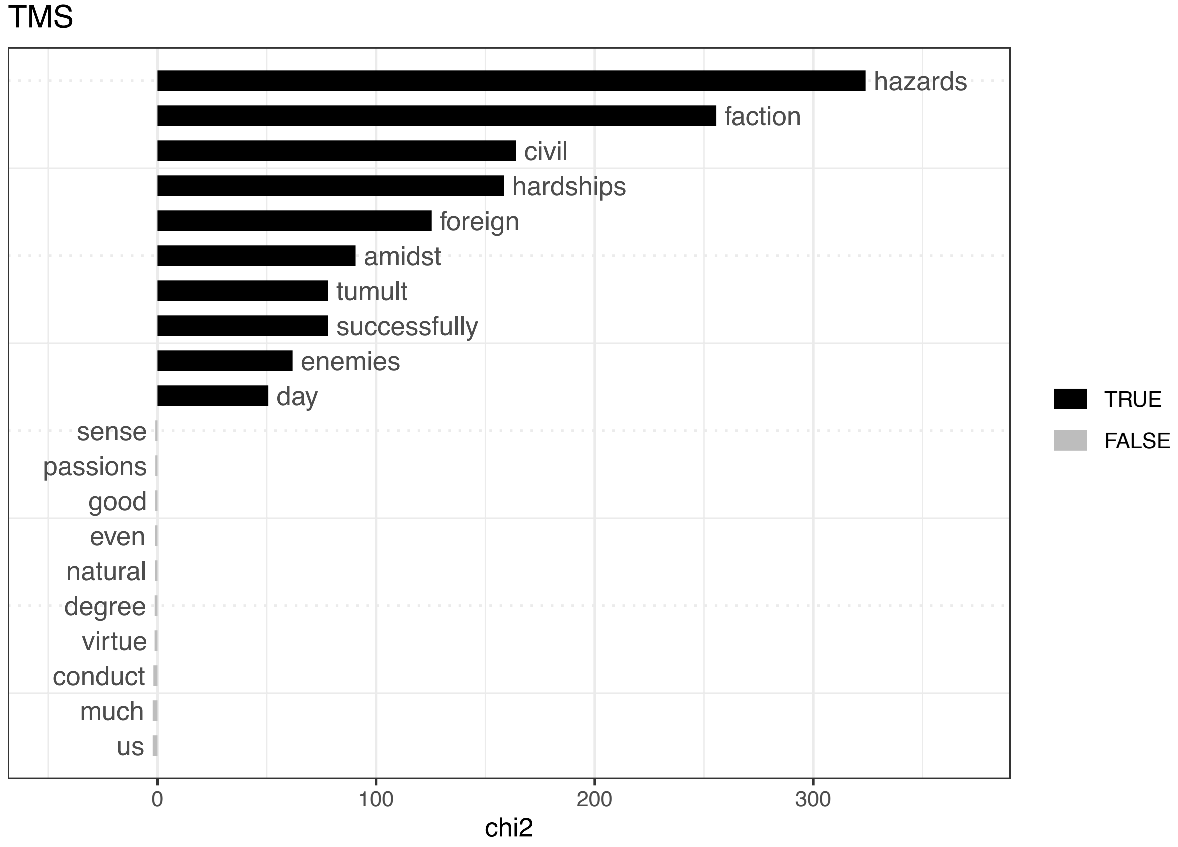

Figure 2: Graphical representation of the keyness values for words associated (TRUE) or not associated (FALSE) with war in the *TMS*.

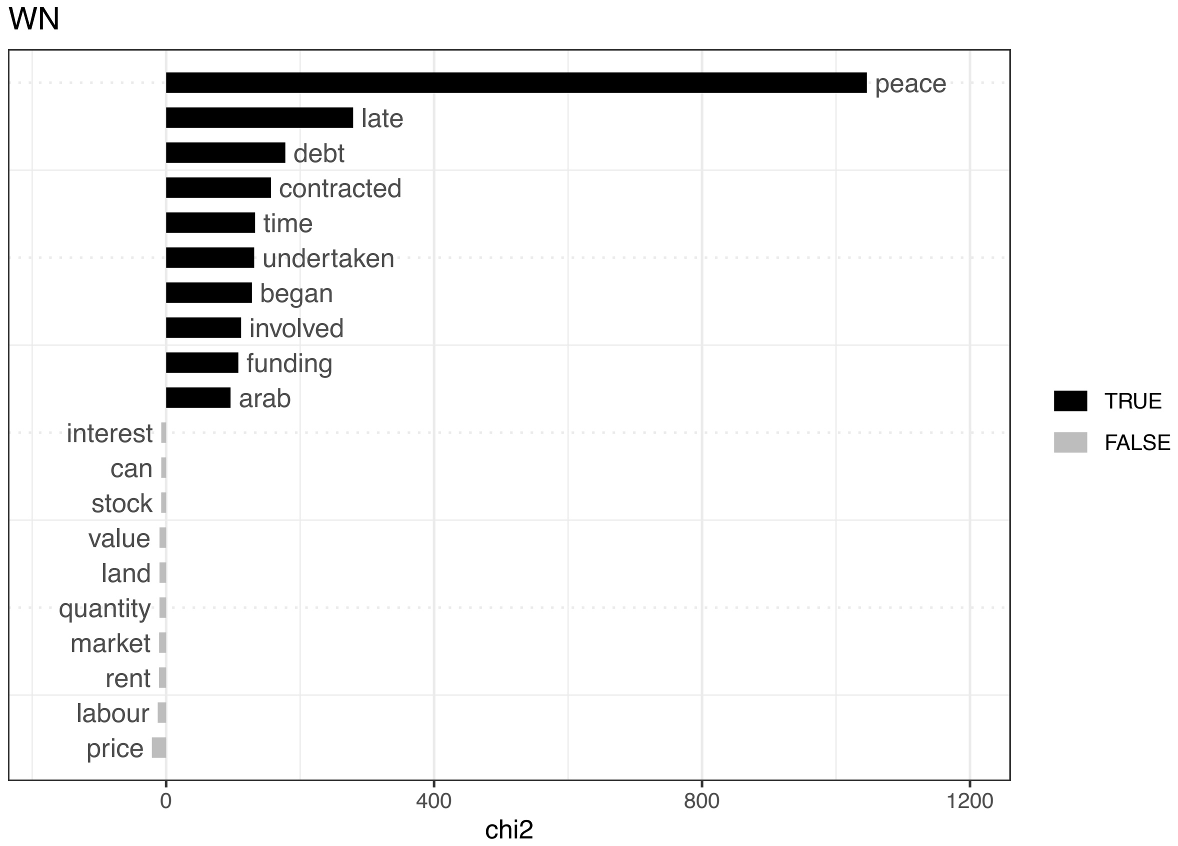

Figure 3: Graphical representation of the keyness values for words associated (TRUE) or not associated (FALSE) with war in the *WN*.

Our discussion of war in the article, based on the keyness results above, is complemented by the analysis of cooccurrences to the term "peace" in the *TMS*. The results of the cooccurrences analysis are presented in Table 1.

Table 1: 10 strongest co-occurrences for "peace" in the *TMS*, considering only words appearing more than 15 times in the book.

word1	word2	correlation
peace	difference	0.115
peace	disturb	0.115
peace	citizens	0.111
peace	distinction	0.108
peace	security	0.104
peace	mutual	0.099
peace	miserable	0.090
peace	society	0.088
peace	family	0.086
peace	order	0.082

II. LANGUAGE DIVERSITY

As we stated in the article, many indicators exist for the study of lexical diversity. We examined three indicators: the type-token ratio (TTR, see Jockers and Thalken 2020), the Hapax richness (Jockers and Thalken 2020) and the moving-average type token ratio (MATTR) (Covington and McFall 2010).

TTR is known to be negatively correlated with the document's length. As Baker (2006, p. 52) puts it, TTR values: "tend to be useful when looking at relatively small text files (say under 5,000 words). However, as the size of the corpus grows the type-token ratio will almost always shrink, because high-frequency grammatical words like 'the' and 'to' tend to be repeated no matter what the size of the corpus is." We confirm this result for both *TMS* and *WN*, with a negative correlation between TTR and the length (number of words) of the sections (see Figure 4).

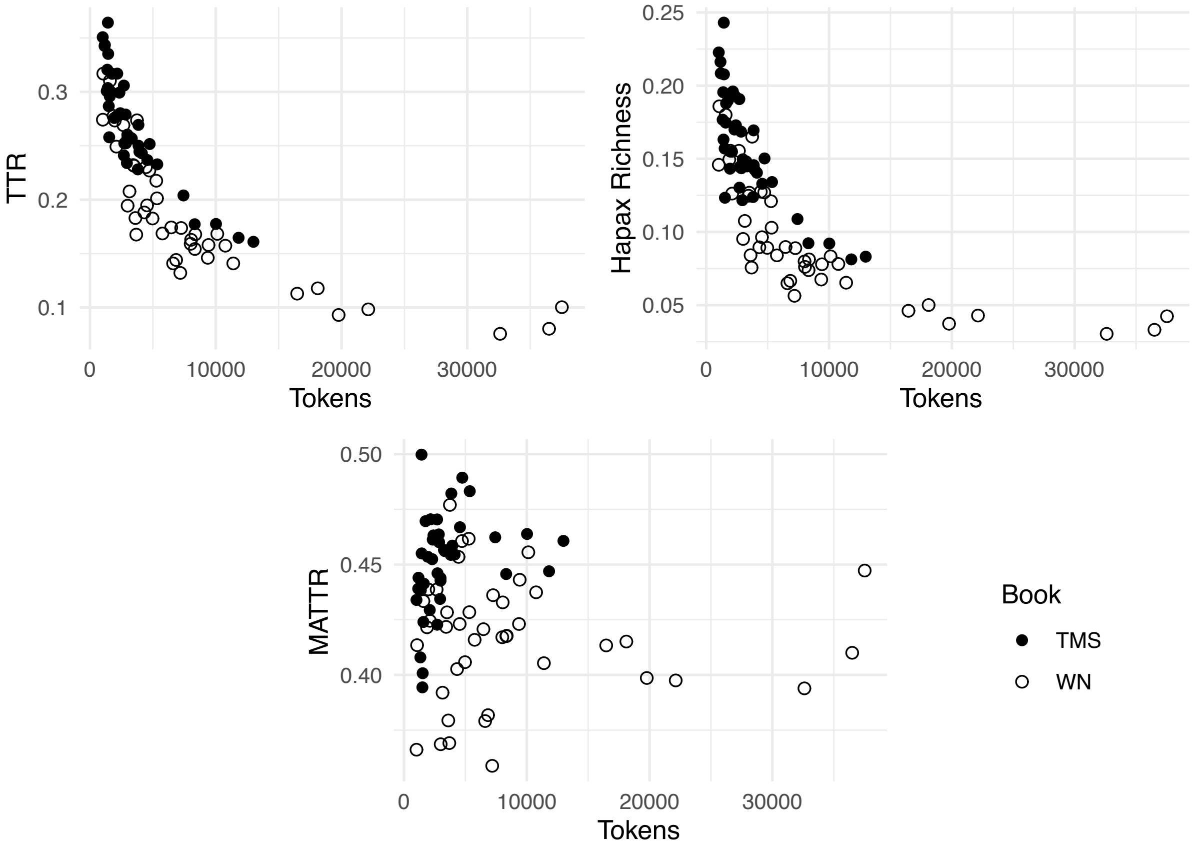

Figure 4: Relationship between the length of text (*x*-axis) and the different lexical richness measures (*y*-axis), for each section of the two books. The graph shows the negative relationship between the length of text and TTR or hapax richness, while there is no relationship in the case of MATTR.

Concerning hapax richness, Fan (2010) has shown that a U-shaped curve should be expected, with a negative relationship between hapax richness and text length until a 3 million words threshold, followed by a positive relationship for longer texts. All sections in Smith's books are shorter than Fan's threshold and, in fact, we do not observe the aforementioned U-shaped relationship. We observe instead a dynamic identical to that of TTR, with a negative relationship between hapax richness and text length.

Given the above issue, both TTR and Hapax richness are not suitable for our analysis. Smith's books are, in fact, of very different lengths. An analysis by book section would also be impacted, considering that the lengths of sections span between 1005 and 42,610 tokens (we removed sections shorter than 1000 tokens for the analysis – see also below).

The third measure of lexical diversity that we consider is the moving-average type token ratio (MATTR) (Covington and McFall 2010). As noted in the article, MATTR calculates TTRs over a fixed length moving window of consecutive words, and finally averages these TTRs to obtain the final measure. The decision of the length of the window is left to the researcher. We followed Covington and McFall (2010, p. 97) who recommend using a window of 500 words for stylometric studies. Also, we decided to consider only sections of at least 1000 words.

Figure 4 and statistical tests confirm that MATTR values are not correlated with the length of text and, therefore, this indicator is indeed better suited for the study of texts of varying length, as those in our analysis.

III. READABILITY

Two popular indicators for measuring readability are the Flesch (1948) reading ease score, and the Flesch-Kincaid (Kincaid et al. 1975) grade level score. Both indicators are linear combinations of the length of sentences and the length of words, with different intercepts and coefficients.

$$Flesch\,Reading\,Ease\,Score = 206.835 - 1.015\frac{N.\,of words}{N.\,of sentences} - 84.6\,\frac{N.\,of syllables}{N.\,of words}$$

$$Flesch\text{–}Kincaid\,Grade\,Level = 0.39\frac{N.\,of words}{N.\,of sentences} + 11.8\,\frac{N.\,of syllables}{N.\,of words} - 15.59$$

The scores of the two indicators are not bound to a specific interval. For instance, in the case of the FleschKincaid score, it is possible to obtain negative values (extremely easy to read) or values larger than 30 (extremely difficult), that would complicate the use of a U.S. grade level as frame of reference. In our paper, we only use the Flesch-Kincaid score because it is expressed in a familiar measuring unit and is, therefore, easier to interpret.

The weight associated to the length of sentences is comparatively larger in the Flesh-Kincaid than in the Flesh indicator. Thus, the Flesh-Kincaid indicator considers the length of a sentence to be a more relevant factor in determining its readability. The two indicators can therefore return different readability results, especially in cases in which the length of sentences is very different among documents. Nonetheless, there is a strong relationship between the indicators.

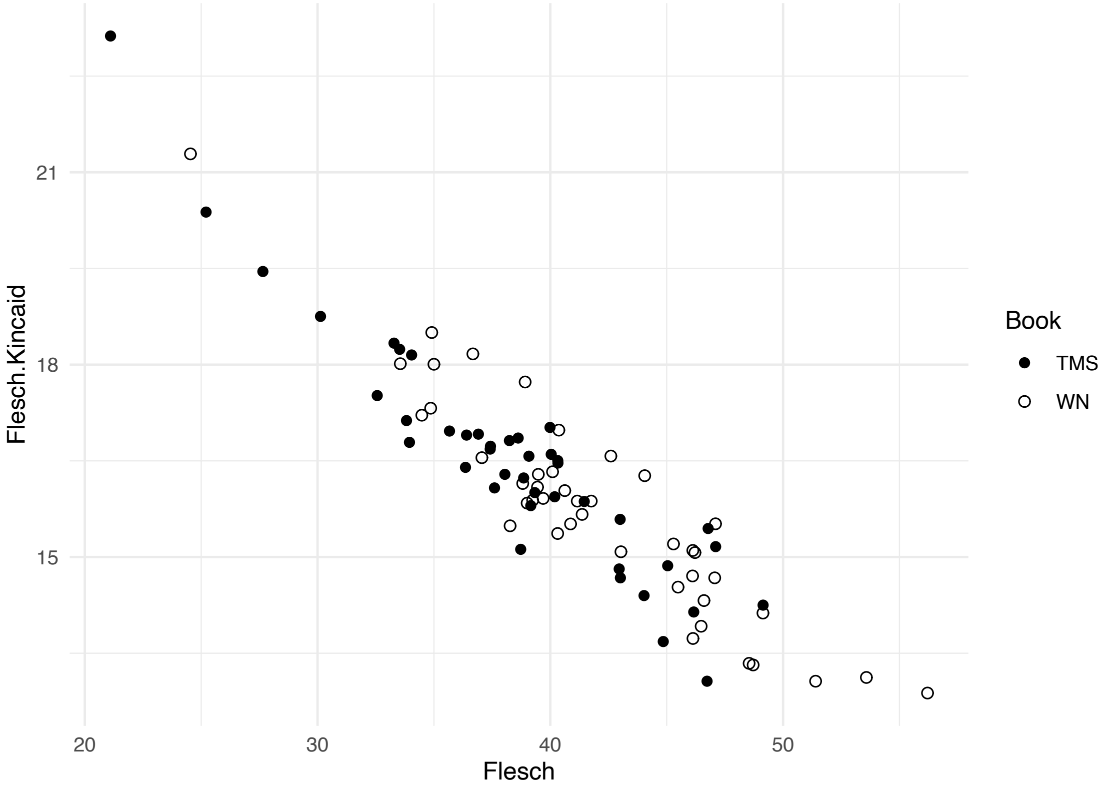

Figure 5: Relationship between Flesch and Flesch-Kincaid scores calculated on the sections of the *TMS* and *WN*.

Figure 5 shows the relationship between Flesch and Flesch-Kincaid scores calculated for Smith's books. In this case, we do not see any section that diverges from the trend and we therefore expect that the qualitative interpretation in terms of readability would remain the same with both metrics. Figure 6 confirms this, the only caveat being that the results will appear upside down, because the sign of the weights for the lengths of words and sentences in the two indicators is inverted.

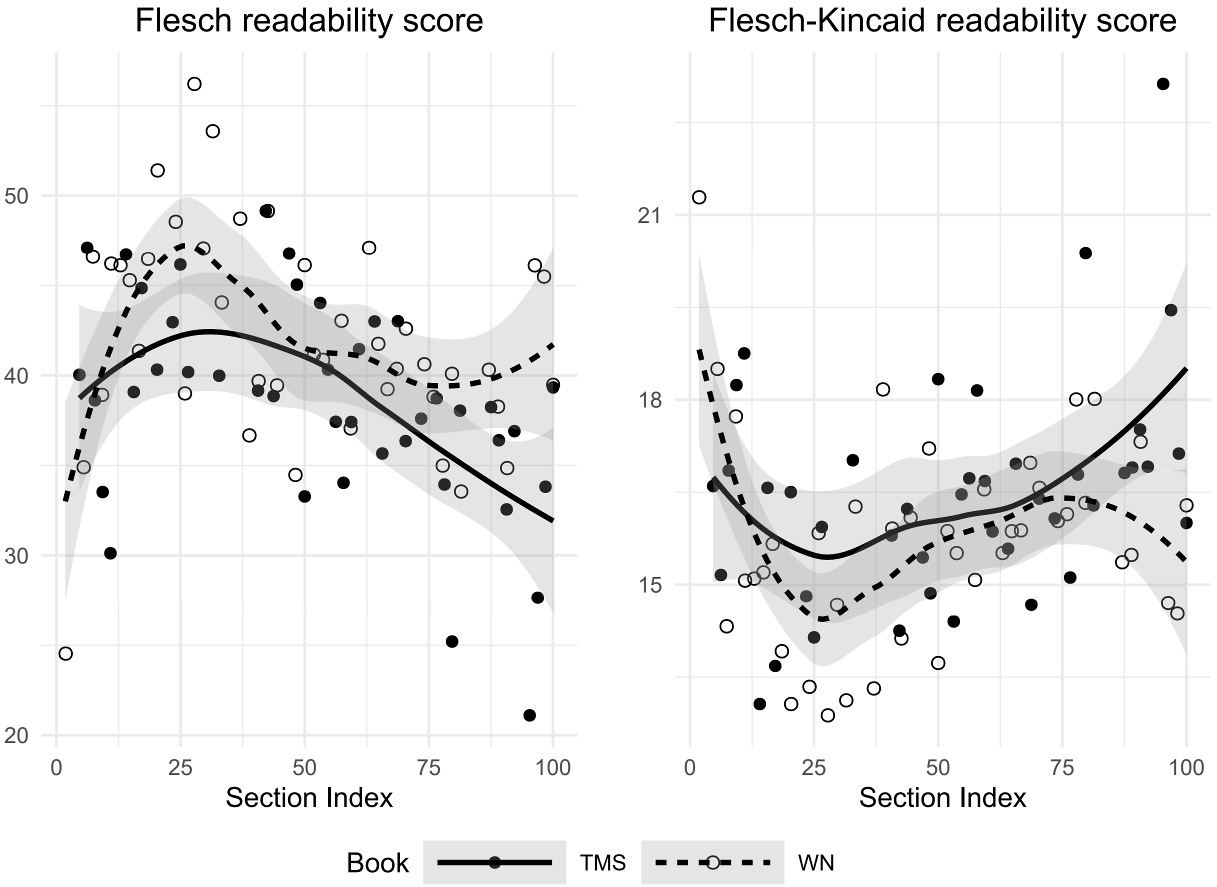

Figure 6: Indicators of readability calculated throughout *TMS* and *WN*.

IV. SENTIMENT ANALYSIS

We mostly used a refined version of the sentiment analysis method, as implemented in the R package *sentimentr*, version 2.9.0 (Rinker 2021). The package provides a set of functions to analyze the sentiment of a text. Nonetheless, its application to Smith's writings required some modifications regarding the lexicon used and the graphical representation of the results.

The sentiment analysis process starts by splitting the text into single sentences and then single words. Then, a function performs a lexicon lookup, where a lexicon is a specific dictionary in which words are associated with a sentiment score between –1 (negative sentiment) and +1 (positive sentiment). In this simple version, the process is evaluating the balance between words carrying a positive sentiment and words carrying a negative one.

The implementation offered in *sentimentr* package has instead the advantage of considering valence shifters, i.e., specific words that influence the original sentiment of an expression. The valence shifters in *sentimentr* are divided into three categories, depending on their effect on the sentiment score:

- Inverting effect: words and expressions such as "not" or "don't";
- Amplifying effect: words and expressions such as "very" or "extremely";
- Weakening effect: words and expressions such as "somewhat", "slightly" or "barely".

We calculate the sentiment for each sentence in the book and then look for patterns that would show the evolution of the sentiment along the text. Figure 7 shows the standard plot provided by the *sentimentr* package. The plot is built by smoothing and rescaling the sentiment score of sentences along the books and plotting the smoothed values.

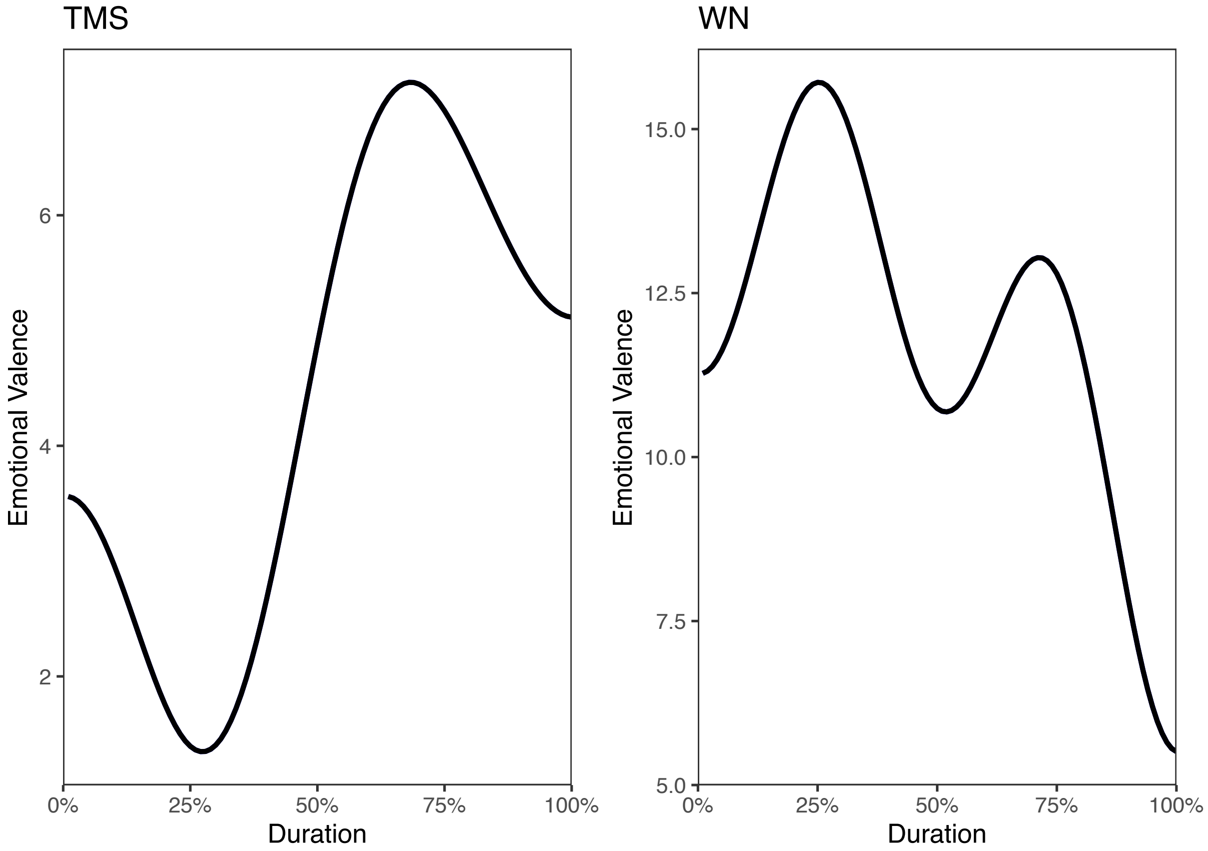

Figure 7: Graphical representation of the sentiment of *TMS* and *WN* generated by the default plot function of *sentimentr* package.

We found this graphical representation to be difficult to read, because of the transformation (rescaling) applied to sentiment scores, and potentially misleading, because of the exceedingly aggressive smoothing that leads to pronounced patterns and hides the variability of the original scores. Therefore, we compared *sentimentr*'s default plot displayed above with our own graph, which includes a smoothing line within a plot (Figure 8) that shows the sentiment scores for all sentences in the two books, giving a clearer representation of the variability of scores. This second plot shows how the smoothing lines are relatively close to zero, indicating mildly positive sentiment, without major shifts.

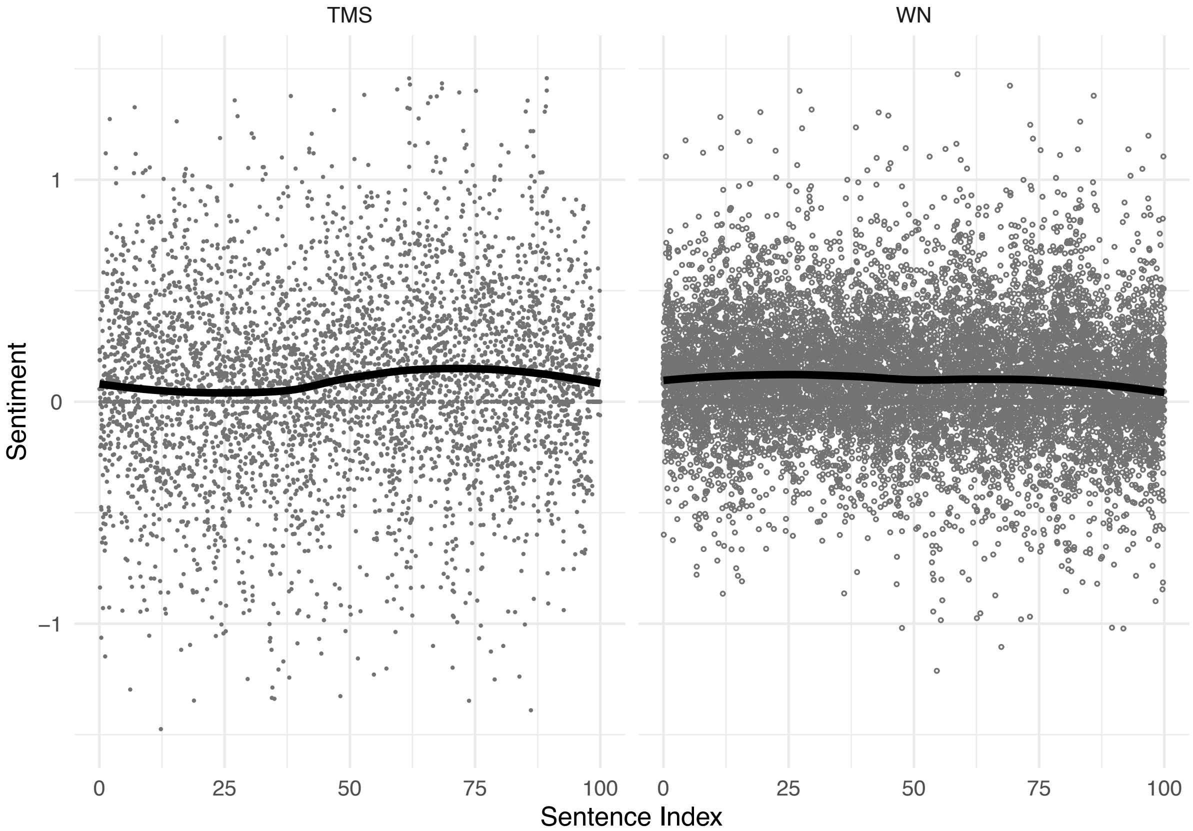

Figure 8: Sentiment scores for all sentences of *TMS* and *WN*. The *y*-axis indicates the direction (positive-negative) and the strength of the sentiment measured in each sentence. Values close to zero are essentially neutral.

Crucially, this second plot, together with the boxplot presented in the article, allows us to appreciate the large variability of the results. We provide some possible explanations for this variability in the article.

REFERENCES

Baker, Paul. 2006. *Using Corpora in Discourse Analysis*. London - New York: Continuum.

Covington, Michael A., and Joe D. McFall. 2010. "Cutting the Gordian Knot: The Moving-Average Type–Token Ratio (MATTR)." *Journal of Quantitative Linguistics* 17 (2): 94–100.

Fan, Fengxiang. 2010. "An Asymptotic Model for the English Hapax/Vocabulary Ratio." *Computational Linguistics* 36 (4): 631–37.

Flesch, Rudolph. 1948. "A New Readability Yardstick." *Journal of Applied Psychology* 32 (3): 221–33.

Jockers, Matthew L., and Rosamond Thalken. 2020. *Text Analysis with R: For Students of Literature*. Cham: Springer Nature.

Kincaid, J. Peter, Robert P. Fishburne Jr., Richard L. Rogers, and Brad S. Chissom. 1975. *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel*. Defense Technical Information Center. Accession Number: ADA006655. Available at: https://apps.dtic.mil/sti/pdfs/ADA006655.pdf.

Rinker, T. W. (2021). *sentimentr: Calculate Text Polarity Sentiment.* Version 2.9.0. https://github.com/trinker/sentimentr