Improving reading rate prediction with word length information:

Evidence from Dutch

Marc Brysbaert   Longjiao Sui   Wouter Duyck   Nicolas Dirix

Department of Experimental Psychology

Ghent University, Belgium

Address: Marc Brysbaert

                Department of Experimental Psychology

                Ghent University

                B-9000 Gent, Belgium

                Email: marc.brysbaert@ugent.be

Abstract

Previous research in English has suggested that reading rate predictions can be improved considerably by taking average word length into account. In the present study, we investigated whether the same regularity holds for Dutch. The Dutch language is very similar to English, but words are on average half a letter longer: 5.1 letters per word (in non-fiction) instead of 4.6. We collected reading rates of 62 participants reading 12 texts with varying word lengths, and examined which change in the English equation accounts for the Dutch findings. We observed that predictions were close to the best fitting curve as soon as the average English word length was replaced by the average Dutch word length. The equation predicts that Dutch texts with an average word length of 5.1 letters will be read at a rate of 238 word per minute (wpm). Texts with an average word length of 4.5 letter will be read at 270 wpm, and texts with an average word length of 6.0 letters will be read at a rate of 202 wpm. The findings are in line with the assumption that the longer words in Dutch do not slow down silent reading relative to English and that the word length effect observed in each language is due to word processing effort and not to low-level, visual factors.

Reading rate is an interesting psychological variable. It is used to decide about deficient reading in various forms (dyslexia, vision deficiency, slow reading without a clear cause), to gauge the time required for various tasks (most prominently reading assignments), and to test the quality of new presentation devices and letter fonts. Reading rate is also important for psychological theories about reading processes, individual differences in information processing, language differences, task effects, and metacognition. To make optimal use of the variable, it is important to have information about typical reading rates and the variables influencing them.

A meta-analysis of reading rates (Brysbaert 2019) revealed that English readers have an average reading speed of 238 words per minute (wpm) for non-fiction and 260 wpm for fiction. Brysbaert (2019) argued that the difference in reading rate could be predicted by taking mean word length into account. Average word length in English non-fiction texts is 4.6 letters, against 4.2 letters in fiction. By using reading rate for non-fiction as the baseline, reading rate for fiction could be predicted accurately with the equation *Reading Rate$_{fiction}$ = 238 * 4.6/4.2*, which gives a value of 261 wpm.

Brysbaert's (2019) equation can be used to improve predictions for individual texts. For instance, in the Nelson-Denny test, participants are asked to read texts silently and answer questions about them. The number of words read in the first minute is taken as an estimate of reading rate. Throughout the years, several editions of the test have been published, known as Forms A, B, … J. Each form has a slightly different reading rate in the standardization tables. For instance, average reading rate of university students for Form G is 243 wpm, whereas it is 261 wpm for Form H (Brown et al., 1993). On the basis of Brysbaert's (2019) equation, we expect that average word length will be longer in Form G than in Form H. This is indeed the case: average word length of the first five paragraphs of Form G is 4.4 letters (based on 428 words), whereas it is 4.2 letters in Form H (based on 437 words). Using the equation, the predicted reading rates are 238 * 4.6/4.4 = 249 wpm for Form G, and 238 * 4.6/4.2 = 261 wpm for Form H, reasonably close to the observed values.

Brysbaert's (2019) equation was based on data published by Carver (1971, 1976), Miller and Coleman (1971), and Coke (1974), who all reported that reading rate expressed as words per minute decreased as a function of text difficulty but not when it was expressed as letters per second or syllables per second. These authors therefore suggested that wpm should be replaced by letters or syllables per second as a more appropriate measure. Although these variables have been adopted in specialized research areas, they failed to appeal to the wider public (and researchers in general), arguably because they lack the intuitive understanding of wpm.

Improving reading rate prediction by taking average word length into account has societal value, because society looks at experimental psychology for this type of information. Websites want to know the expected reading time for a specific article, so that they can inform their readers adequately. Solicitors asked to read a particular dossier, want to know how much time will be required. As Yarkoni and Westfall (2017) argued, psychology journals have a tendency to undervalue this role of psychology, because they focus exclusively on the causal mechanisms underlying behaviour. As a result, psychology journals are overpopulated by studies testing intricate theories of psychological mechanisms, and have a shortage of studies providing answers to real-world questions.

At first sight, the theoretical implications of the correlation between average word length and reading rate are limited. Although word length is known to play an independent role in visual word recognition (Mandera et al., 2020) and text reading (Dirix et al., 2019), in typical texts it is highly correlated with other variables, such as word frequency, orthographic similarity, word age-of-acquisition, syntactic complexity, test difficulty, and topic familiarity to the reader. So, the correlation is likely to be affected by a multitude of variables affecting word and text difficulty, and average word length is only a practical proxy for word difficulty. Knowing the exact mechanisms behind the correlation would require extensive and detailed testing, involving a comparison of a large variety of texts, differing systematically on the variables listed above (and probably other as well).

In the present article, we follow a different tack. We examine to what extent the relationship observed between average word length and reading rate in English is language-specific or an equation that can be generalized to other languages. For instance, the Dutch language very much resembles English but has words that are on average half a letter longer. Average word length in the English subtitle corpus SUBTLEX-US (Brysbaert & New, 2009) is 3.76 letters, compared to 4.26 letters in the equivalent Dutch subtitle corpus SUBTLEX-NL (Keuleers et al., 2010). A similar analysis of one million sentences from parallel English and Dutch corpora based on newspaper texts and texts collected from the web (Goldhahn et al., 2012) yielded word lengths of 4.81 in English in 5.23 in Dutch.

The longer length of Dutch words allows us to investigate to what extent the word length effect on reading rate is due to low-level visual factors or to language processing factors. If the length effect is caused by low-level visual factors, we must predict that reading rate will be lower in Dutch than in English, because written Dutch words are longer. In contrast, if the length effect is due to language factors (long words are more difficult to process), we would expect to observe very much the same relationship in Dutch as in English. The only parameter that is likely to differ in the equation, is average

word length, which is half a letter longer in Dutch than in English. So, the Dutch equation would be: *Reading rate = 238 * 5.1/WL$_{text}$*, in which 238 is the average reading rate for non-fiction texts, 5.1 is the average word length in Dutch non-fiction texts, and WL$_{text}$ is the word length observed in the text for which one wants to estimate reading rate.

On the basis of current evidence, there are good reasons to expect that word length will be a language factor rather than a visual factor. Language processing speed has been reported to depend more on information density than on the surface properties of a language, both in reading (Liversedge et al., 2016) and in spoken language processing (Coupé et al., 2019). Brysbaert (2019) further reported an average silent reading rate of 228 wpm for Dutch non-fiction, which is close to the English reading rate of 238 wpm, given the low number of Dutch studies (7 against 144 in English).

The availability of a mathematical equation provides us with the possibility of a strong empirical test. Whereas a simple prediction that reading rate will depend on average word length in Dutch does not add much information to what is known, examining whether the very same equation applies constrains the possible explanations in important ways. It will be interesting to see how much the English equation must be changed to account for Dutch findings.

To increase the comparability of our data with those published in English, we used translations of the materials used by Carver (1971) and Miller and Coleman (1971). These comprised expository paragraphs of 150 words compiled by Miller and Coleman (1967) and published by Aquino (1969). The texts varied in average word length, which correlated with text difficulty (readability).

Method

Participants

Participants were 62 students from Ghent University (mainly undergraduate students; 48 females), reporting no reading problems or other issues that might affect their performance. The present experiment was the last part of a larger reading study that lasted two hours and for which the participants were paid €20.

Stimulus materials

The 36 texts compiled by Miller and Coleman (1967) and published by Aquino (1969) were translated by the first author and checked by comparing the original to the Google back translation. The materials for the present study consisted of 12 texts, selected on the basis of their average Dutch word length, ranging from 4.5 to 6.4 letters (average = 5.3). Average word length of the corresponding English texts was 4.7 (range 4.2 – 5.5). The number of words in the Dutch texts was on average the same as that of the English texts (149 words). Each text was followed by four yes/no questions about the main ideas of the texts (in order not to force the participants to pay undue attention to minor details). The Dutch texts and their English originals are available at https://osf.io/ep5fs/. The main characteristics are summarised in Table 1.

Text presented	Text M&C	English Nwords	Letts/word	Dutch Nwords	Letts/word
1	16	150	4.2	144	4.5
2	12	150	4.2	148	4.6
3	23	150	4.3	167	4.7
4	31	150	4.7	169	4.8
5	24	150	4.6	161	4.9
6	15	150	4.2	149	5.1
7	27	150	4.9	154	5.3
8	28	150	4.7	137	5.5
9	34	150	5.1	155	5.7
10	26	150	5.0	126	6.0
11	35	150	5.5	143	6.2
12	33	150	5.3	141	6.4
	Average	150.0	4.7	149.5	5.3

Table 1: Characteristics of the texts used: Text number in Miller & Coleman (1967), number of words and average word lengths in the original English texts and in the Dutch translations.

Procedure

Texts were presented on a computer screen. Participants were asked to read the text, shown on a single screen, as they would normally do. They were told that each text would be followed by four yes/no questions, in order to ensure reading for comprehension. No feedback was given about the correctness of the responses. Each participant got a different, random order of the texts. The present experiment

happened after a similar experiment also involving text reading and answering questions. This meant that participants had already been reading for more than an hour (with some breaks in-between). Brysbaert (2019) showed that longer testing times lead to more stable results, because participants resort to their habitual reading rate, rather than trying to be overly fast or overly accurate (slow). Time was registered when the text appeared and when participants pressed the bar to answer questions, giving us an indication of text reading time.

Results

Participants answered 87% of the questions correctly (range: 75% - 98%; on a total of 48 questions per person; guessing rate = 50%). Reading times were converted to reading rates expressed as words per minute. Average reading rate was 228 wpm (SD across participants: 55). Differences in reading rates between participants were very stable, with both Cronbach's alpha and McDonald's omega .96.

Figure 1 shows the reading rates for the different texts, ordered by average word length (going from short to long). As expected, reading rate decreased as average word length increased. Pearson correlation between average reading rate and average word length was -.80 (N = 12, p < .01), indicating that 64% of the variance in reading rates between texts could be accounted for by average word length.
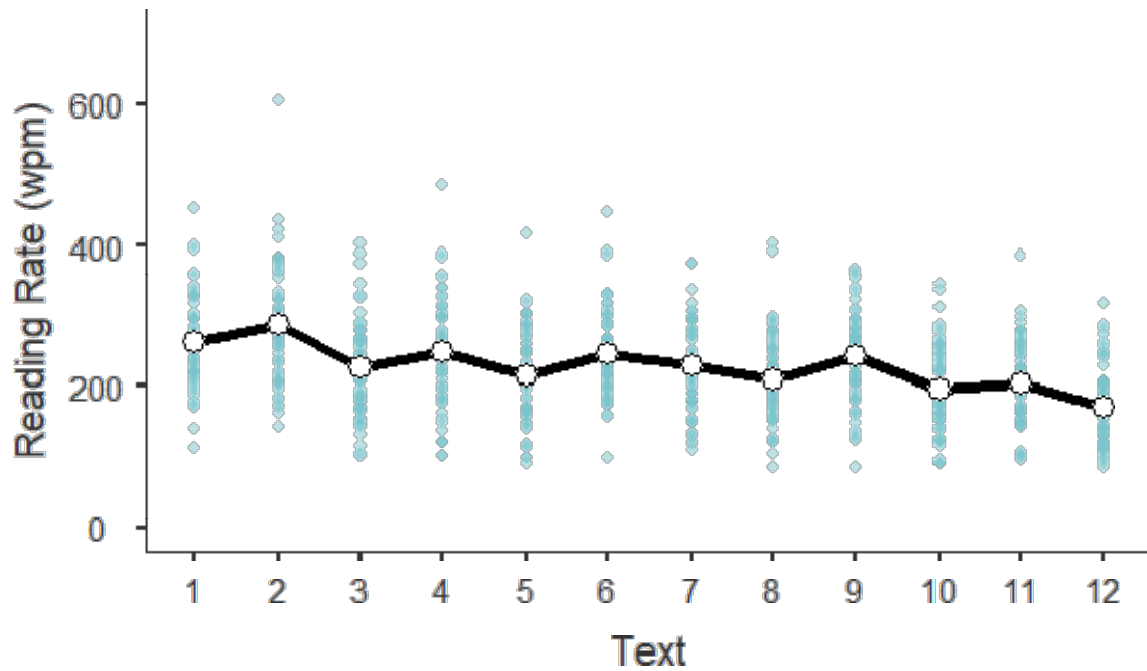
Figure 1: Reading rate as a function of average word length. Texts ordered from short words (average length = 4.5) to long words (average word length = 6.4). Blue dots indicate the raw data of the participants.

Figure 2 shows the observed reading rates with the predicted values on the basis of the equation *Reading rate = 238 * 5.1/WL$_{text}$*, in which 238 is the expected reading rate for texts with word length of 5.1 letters and WL$_{text}$ is the word length of the text under consideration. For comparison purpose, the figure also includes the best fitting quadratic function based on the observed values. The a priori predictions were close to the best possible fit. R^2 of the best fitting quadratic function was .66 (compared to .64 for the linear function). R^2 for the a priori predicted values was .62 (based on the squared deviations between the a priori predicted values and the observed values, relative to the total variability in the observed values).
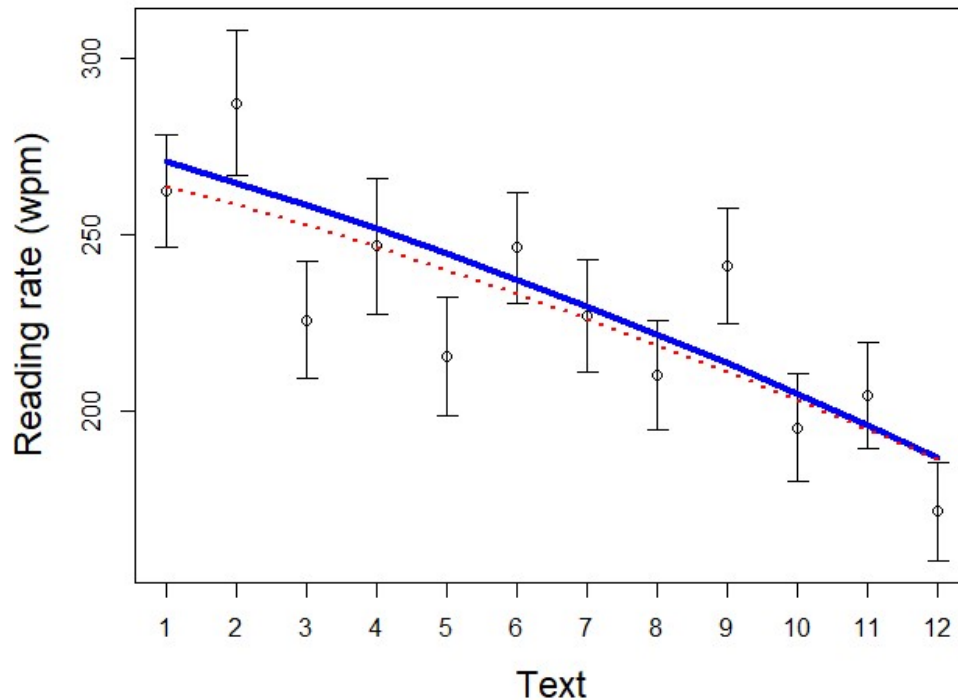
Figure 2: Observed reading rates and 95% confidence intervals for the texts with different word lengths. Also shown are the predictions based on Brysbaert's (2019) equation – thick solid line – and the best fitting quadratic curve – thin dotted line.

Discussion

Reading rate differs between texts (and between participants). Various authors (Carver, 1971, 1976; Coke, 1974; Miller & Coleman, 1971) have argued that the text differences can be captured quite well by taking average word length of a text into account. They proposed letters per second or syllables per second as better reading rate measures than words per minute (wpm) or they suggested to recalculate observed word lengths to a standardised word length of five letters. None of these reading measures have had much impact beyond specialised research areas, because wpm is intuitively easier to understand and also easier to calculate.

Brysbaert (2019) proposed a simple equation to correct reading rate in wpm for differences in word length (WL) and thereby improve predictions for individual texts. The equation started from the observation that, across studies, English reading rate is 238 wpm for texts with an estimated WL of 4.6 letters. By taking the equation $238 * 4.6/WL_{text}$, the estimate is adapted for the text at hand. Texts with average WL below 4.6 (fiction, spoken language) result in higher estimates; texts with longer words (scientific texts, some expository texts) result in lower estimates. This allows users to refine predictions for specific texts, as shown for the Nelson-Denny test in the introduction.

In the present study we investigated how the English equation must be adapted for Dutch. Dutch is an interesting comparison language because it resembles English but the written words are on average half a letter longer. If the word length effect were due to low-level visual factors, we would find that reading in Dutch is slower than in English. In the most extreme case, we could find that the typical reading rate for Dutch non-fiction texts is 238 * 4.6/5.1 or 215 wpm (4.6 = English WL, 5.1 = Dutch WL). In contrast, based on existing evidence, we expected that differences in average word length between languages would not matter. Long words in a language take longer to process because they are more difficult (e.g., contain more information or are less familiar), but differences in word length between languages are unimportant as long as the words convey the same amount of information. In that scenario the time required to read a Dutch translation of an English text is the same as the time needed to read the original text. Therefore, we predicted that Dutch reading rates would be predicted well by the equation *Reading rate = 238 * 5.1/WL$_{text}$*.

As shown in Figure 2, the expected equation indeed closely predicted the results we obtained. As a matter of fact, the predictions are so close that no further adaptations are needed. This is in line with other studies suggesting that language processing speeds do not depend on the surface characteristics of the languages. What counts is the information conveyed by the message, not the specific words used in different languages (Coupé et al., 2019; Liversedge et al., 2016).

Similar to the English equation, the Dutch equation can be used to predict reading rates and reading times for Dutch texts. It says that Dutch texts with WL = 4.0 will be read at a rate of 238 * 5.1/4.0 = 303 wpm, whereas Dutch texts with WL = 6.5 will be read at a rate of 238 * 5.1/6.5 = 187 wpm. To see how well the Dutch equation improves predictions for other studies, we can look at data collected by Mak and Willems (2018). These authors asked 109 participants to read three Dutch stories (between two thousand and three thousand words each), while eye movements were tracked. After each story the participants were asked three general questions to make sure they had read the stories. Reading rate

was 304 wpm for the first story, 253 wpm for the second, and 222 wpm for the third. This strongly suggests that the texts differed in average word length. Upon inspection of the materials, that was indeed the case: The respective word lengths were 4.4 letters, 4.5 letters, and 5.1 letters. Based on the Dutch equation, the predicted reading rates for the three texts are: 238 * 5.1/4.4 = 276 wpm, 238 * 5.1/4.5 = 270 wpm, and 238 * 5.1/5.1 = 238 wpm. Although the predictions fall short of accounting for the complete range of reading rates observed, they reduce the average prediction error from 32 wpm for the uncorrected reading rate to 20 wpm for the corrected one.

As mentioned in the introduction, we do not interpret the correlation between average word length and reading rate as evidence that word length is the primary variable affecting language processing speed. Average word length is an interesting variable because it is straightforward to calculate[1] and can easily be implemented in reading time calculators, but in all likelihood derives a large part of its power from the fact that it is highly correlated with other language-related variables influencing text readability, such as average word frequency, word age-of-acquisition, sentence length, syntactic complexity, topic familiarity, and a rift of other variables (for reviews, see Chen & Meurers, 2016; Crossley et al., 2017; De Clercq & Hoste, 2016). Further research will have to indicate how much further gain can be made by taking these text difficulty factors into account. These studies will have to make use of many more texts than the current 12, and of texts that systematically vary all variables researchers are interested in.

Another interesting question is how well the equation between average word length and reading rate generalizes to other languages. We are optimistic that the equation will apply to all languages using the Roman alphabet. If true, for these languages all we need is an estimate of average written word length. More uncertainty exists as to what extent the equation will generalize to languages using other alphabets (Cyrillic, Greek), orthographies omitting vowels (Arabic, Devanagari, Hebrew, Hangul, Thai), orthographies representing syllables (kana), and orthographies representing words (Chinese). It seems likely that reading rate in these languages will also depend on the number of symbols needed to write the words, but the equations will have to be determined empirically. To some extent, the inclusion of word length already happens in Chinese. Because this language does not have spaces between words, the number of characters read per minute is a more straightforward measure of reading rate than the number of words read per minute (e.g., Bai et al., 2008; Fraser, 2007).

---

[1] A simple way to calculate average word length of a text is to use MS Word. Take away all punctuation marks from the text (except for ' and - in words such as don't and make-up) and use the REVIEW button to calculate the number of letters without spaces and the number of words. Divide the former by the latter.

Availability

Texts and data are freely available at https://osf.io/ep5fs/.

References

Aquino, M. (1969). The validity of the Miller-Coleman readability scale. *Reading Research Quarterly, 4*, 342-357.

Bai, X., Yan, G., Liversedge, S. P., Zang, C., & Rayner, K. (2008). Reading spaced and unspaced Chinese text: Evidence from eye movements. *Journal of Experimental Psychology: Human Perception and Performance*, 34(5), 1277-1287.

Brown, J.A., Fishco, V.V., & Hanna, G. (1993). *Nelson–Denny Reading Test: Manual for Scoring and Interpretation, Forms G & H*. Rolling Meadows, IL: Riverside Publishing.

Brysbaert, M. (2019). How many words do we read per minute? A review and meta-analysis of reading rate. *Journal of Memory and Language, 109*, 104047.

Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*, 977-990.

Carver, R. P. (1971). Readers' Dialogue: Evidence for the Invalidity of the Miller-Coleman Readability Scale. *Journal of Reading Behavior, 4(3)*, 42-47.

Carver, R. P. (1976). Word length, prose difficulty, and reading rate. *Journal of Reading Behavior, 8(2)*, 193-203.

Chen, X., & Meurers, D. (2016). Characterizing text difficulty with word frequencies. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 84-94). Available at https://www.aclweb.org/anthology/W16-0509.pdf.

Coke, E. U. (1974). The effects of readability on oral and silent reading rates. *Journal of Educational Psychology, 66(3)*, 406-409.

Coupé, C., Oh, Y. M., Dediu, D., & Pellegrino, F. (2019). Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche. *Science Advances, 5(9)*, eaaw2594.

Crossley, S. A., Skalicky, S., Dascalu, M., McNamara, D. S., & Kyle, K. (2017). Predicting text comprehension, processing, and familiarity in adult readers: New approaches to readability formulas. *Discourse Processes, 54(5-6)*, 340-359.

De Clercq, O., & Hoste, V. (2016). All mixed up? Finding the optimal feature set for general readability prediction and its application to English and Dutch. *Computational Linguistics, 42(3)*, 457-490.

Dirix, N., Brysbaert, M., & Duyck, W. (2019). How well do word recognition measures correlate? Effects of language context and repeated presentations. *Behavior Research Methods, 51*, 2800-2819.

Fraser, C. A. (2007). Reading rate in L1 Mandarin Chinese and L2 English across five reading tasks. *The Modern Language Journal, 91(3)*, 372-394.

Goldhahn, D., Eckart, T., & Quasthoff, U. (2012). Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. *LREC, 29*, 31-43.

Liversedge, S. P., Drieghe, D., Li, X., Yan, G., Bai, X., & Hyönä, J. (2016). Universality in eye movements and reading: A trilingual investigation. *Cognition, 147*, 1-20.

Mak, M., & Willems, R. M. (2018).* Mental simulation during literary reading: Individual differences revealed with eye-tracking. *Language, Cognition and Neuroscience, 34(4)*, 511-535.

Mandera, P., Keuleers, E., & Brysbaert, M. (2020). Recognition times for 62 thousand English words: Data from the English Crowdsourcing Project. *Behavior Research Methods, 52*, 741-760.

Miller, G. R. and Coleman, E. B. (1967). A set of thirty-six passages calibrated for complexity. *Journal of Verbal Learning and Verbal Behavior, 6*, 851-854.

Miller, G. R., & Coleman, E. B. (1971). The measurement of reading speed and the obligation to generalize to a population of reading materials. *Journal of Reading Behavior, 4(3)*, 48-56.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science, 12(6)*, 1100-1122.