

# Reading development at the text level: an investigation of surprisal and embedding-based text similarity effects on eye-movements in Chinese early readers

Xi Fan

Guangzhou Medical University, China  
Maynooth University, Ireland

Ronan Reilly

Maynooth University, Ireland  
Maynooth International Engineering  
College, Fuzhou University

This paper describes the use of semantic similarity measures based on distributed representations of words, sentences, and paragraphs (so-called “embeddings”) to assess the impact of supra-lexical factors on eye-movement data from early readers of Chinese. In addition, we used a corpus-based measure of surprisal to assess the impact of local word predictability. Eye movement data from 56 Chinese students were collected (a) in the students’ 4th grade and (b) one year later while they were in 5th grade. Results indicated that surprisal and some text similarity measures have a significant impact on the moment-to-moment processing of words in reading. The paper presents an easy-to-use set of tools for linking the low-level aspects of fixation durations to a hierarchy of sentence-level and paragraph-level features that can be computed automatically. The study is the first attempt, as far as we are aware, to track the developmental trajectory of these influences in developing readers across a range of reading abilities. The similarity-based measures described here can be used (a) to provide a measure of reader sensitivity to sentence and paragraph cohesion and (b) to assess specific texts for their suitability for readers of different reading ability levels.

---

Keywords: Eye movements, saccades, reading development, text effect


## Introduction

As a reader progresses through a text, depending on their reading goal, they encounter words from which they construct phrases, integrate them into larger sentential and discourse units, and use them to create an isomorphic representation of the writer’s conceptual structure. Information is acquired from words or word clusters and then

integrated into conceptual units of coarser granularity, such as ideas, events, episodes, narratives. There is considerable evidence that the ongoing cognitive processes involved in reading have a direct impact on the lower-level information processing stages involved in eye movement control (see Radach and Kennedy, 2004, 2013; Rayner, 1998, for overviews).

### Measures of eye movements

Using eye-tracking technology, researchers can measure the eye movement characteristics of readers during reading and use the data to probe the underlying perceptual and cognitive process involved in decoding text into a meaning representation. The two relevant eye movement events in reading are fixations, where the eye is relatively still, and saccades, where the eye makes ballis-

Received May 9, 2020; Published September 9, 2020.  
Citation: Fan, X. & Ronan, R. (2020). Reading development at the text level: an investigation of surprisal and embedding-based text similarity effects on eye-movements in Chinese early readers. *Journal of Eye Movement Research*, 13(6):2.  
Digital Object Identifier: 10.16910/jemr.13.6.2  
ISSN: 1995-8692  
This article is licensed under a [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/). 

tic movements across the text. There is a variety of metrics derived from the duration of fixations that are used by reading researchers to provide insight into the time course of reading-related perceptual and cognitive processes. For example, first fixation duration (FFD) is the duration of the first fixation on a word from an incoming rightward saccade and tends to reflect early processing of the word. Refixation duration (RFD) is the sum of subsequent fixations on the word before the eye leaves it and will reflect processing demands related to the frequency of occurrence of a word. Re-reading duration (RRD) measures any subsequent viewing time on the word after the eyes have left it and tends to measure the difficulty the reader has in integrating the word into a larger meaning representation. Another commonly used measure of word processing is gaze duration, which is the sum of FFD and RFD (Rayner, 1998). Figure 1 is a schematic representation of the relationship between these word-based viewing time measures. There are other duration-based measures that aim to capture the processing of larger text regions. However, since our focus is primarily on the earlier stages of word processing, our primary focus will be on the FFD, RFD, and RRD measures (see Radach and Kennedy, 2004, for a further discussion of these measures).

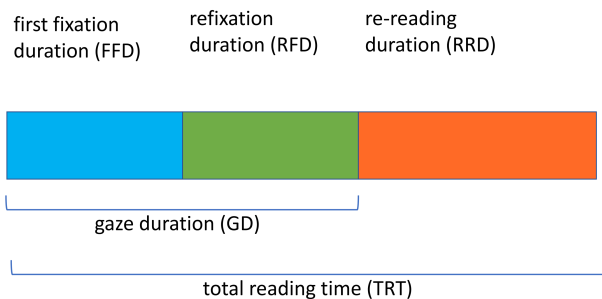


Figure 1. The word-viewing time decomposition used in this paper allows the separate components of first-fixation duration (FFD), refixation duration (RFD), and re-reading duration (RFD) to be combined additively to give the total reading time on the word. The commonly used gaze duration (GD) measure is the sum of FFD and RFD.

### Measures of text relatedness

Because of the structural limitations of the eye, the pickup of word information in reading is a processing bottleneck, but its inherent limitation also presents us with the opportunity to observe the impact of higher level

factors relating to sentence structure and meaning on the deployment of this constrained resource. It is possible, therefore, to observe both the impact of processing the current word, but also of prior context through variations in a variety of eye movement parameters. A major challenge in this area of research is to develop quantitative measures of text complexity at a variety of levels that can be used to predict eye movement behaviour. Ideally, one would like to model faithfully the grammatical and conceptual knowledge that the reader is bringing to bear on the task and use it to predict reading behaviour. Levy (2008) proposed the concept of surprisal as a quantitative measure of the cognitive cost required to process a word in a sentence. Surprisal is a text metric that captures how predictable a word is, given the context of preceding words. Substantial progress has been made in developing the measure with the use of conditional probability distributions over interpretations (Boston et al., 2008). A number of easy-to-calculate proxies for surprisal can be derived from various features of large text corpora. If the corpus is large enough, such as is the case with the Google and Microsoft n-gram corpora (Fang et al., 2010; Michel et al., 2011), then we can obtain usable n-gram frequency counts from unigrams (single words) through to 5-grams (e.g., an n-gram is a sequence of N words, a 5-gram is a five-word sequence). The n-gram frequency counts can be used to calculate surprisal values that can act as approximations to both syntactic and to a lesser extent semantic expectation. These in turn can be assessed as predictors of various eye movement parameters.

While surprisal may capture some of the dynamic local constraints on the reader, clearly there are other things going on during reading. For example, if the reader is dealing with an extended text, he or she has the challenge of integrating meaning across sentences. Depending on the coherence of the text, this can prove difficult. Moreover, the current sentence stands in some relation to the text as a whole to the degree that it's more or less central to the theme. A pioneering attempt to quantify global context effects on text reading was reported by Pynte et al. (2008, 2009) using Latent Semantic Analysis (LSA). LSA is a theory and method for analysing documents to find the underlying meaning or concepts inherent in the documents (Landauer et al., 1998). The central idea is that the aggregate of all the word contexts in which a given word does and does not appear provides a set of mutual constraints that largely determines the simi-

larity of meaning of words and sets of words to each other. LSA is one of the most commonly used methods for word meaning representation. However, in recent years neural-network language "embeddings" have received increasing attention (Arora et al., 2016; Kiros et al., 2015; Mikolov, Chen, et al., 2013; Mikolov, Sutskever, et al., 2013; Pennington et al., 2014). Furthermore, neural-network derived language embeddings tend to have better performance than LSA on very large training corpora (Altszyler et al., 2016)

Neural-network language embeddings exploit statistical properties of text structure to embed text (words, sentences or paragraphs) into numerical vectors of a fixed number of dimensions (usually between 100 and 500), with more dimensions supporting more nuanced discrimination between meanings. The intuition behind the embedding is that it represents text based on its lexical context accumulated over many millions of instances. Consequently, text appearing in similar contexts will have similar embeddings. Embeddings allow us to easily compute semantic similarity between two texts. A typical way of calculating semantic similarity of language items is to measure the cosine of the angle between the high-dimensional vectors representing the language items; the larger the cosine value, the greater the similarity. A significant recent trend has been the development of so-called universal embeddings (Subramanian et al., 2018; Wieting et al., 2015). Universal embeddings are trained on a variety of data sources and use text classification, semantic similarity, clustering, and other natural language tasks to improve their performance by forcing them to incorporate more general word and sentence features. Google's Universal Sentence Encoder is one example of this approach (Cer et al., 2018; Abadi et al., 2015). Figure 2 is intended to show that sentence embeddings can be trivially used to compute sentence level semantic similarity scores that achieve excellent performance on the semantic textual similarity (STS) benchmark (Cer et al., 2018; Abadi et al., 2015). Google's Universal Sentence Encoder is available on TensorFlow Hub (<https://tfhub.dev/google/universal-sentence-encoder>). The website provides easy-to-use code templates that can be used to encode words, sentences or paragraphs into high dimensional vector embeddings. This then allows the straightforward calculation of similarity among words, sentences and paragraphs.

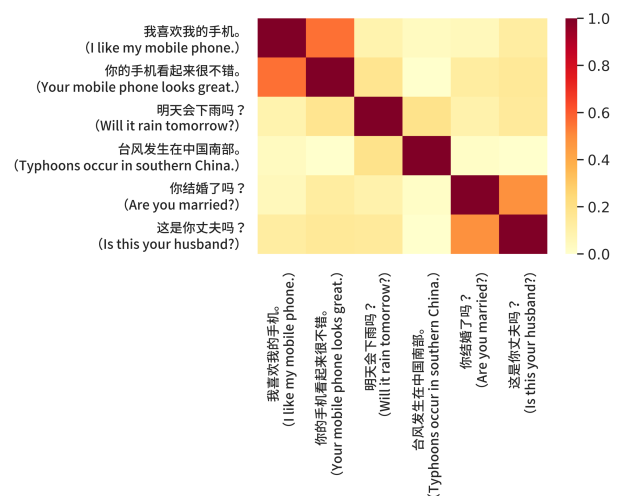


Figure 2. Sentence similarity scores using embeddings from the Chinese version of the universal sentence encoder (Cer et al., 2018). Note that the darker the cell, the greater the sentence similarity. Hence, the diagonal cells are the darkest since they represent self-similarity. The sentence pairs relating to phones and marriage are considered most similar when compared to the pair relating to weather. This may be due to the absence of overlapping words and word stems in the latter two sentences.

### Previous research on reading development

In contrast to the substantial literature on adult reading, the number of research publications on eye movements in developing readers is still quite limited (though see Blythe and Joseph, 2011; Radach et al., 2009, for reviews). Much of the ground work was laid by a set of pioneering studies of text reading (e.g., Buswell, 1922; McConkie et al., 1991; Taylor et al., 1960). These early studies examined elementary school students across various grades, with quantitative analyses mostly restricted to global parameters such as average fixation duration, number of fixations per 100 words, and overall regression frequency. Not surprisingly, this work documented a steady reduction of fixation durations and the number of fixations from grade to grade, whereas the decrease was less pronounced for the proportion of regressive saccades. As Rayner (1985) has pointed out, these developmental changes in eye movements and the analogous differences in the eye movement patterns of readers of differing ability are not the cause of differences in reading fluency. Rather, longer fixation durations, shorter saccades, and more regressions all reflect the fact that reading involves the coordination of many different perceptual and cognitive processes and that this coordination takes time to de-

velop and automatise and some beginning readers are more adept at it than others.

McConkie et al. (1991) pioneered the analysis of children's eye movements at the word level, including fine grained analyses of saccade landing sites within words, and the first quantitative analyses of relations between eye movement parameters and psychometric reading assessments. This was also one of the first studies in which the now common decomposition of viewing times into initial fixation duration, gaze duration and re-reading time was applied to research on developing readers.

Most current studies on eye movements in developing readers have used experimental designs to address specific research questions, usually comparing children with adults or readers at different stages of their development. This work has usually focused on sub-lexical and lexical components of the reading process, such as letter recognition within the perceptual span (Häikiö et al., 2009; Rayner, 1998) or effects of word length and frequency (Blythe et al., 2011; Blythe et al., 2009; Huestegge et al., 2009; Hyönä & Olson, 1995; Joseph et al., 2009; Joseph et al., 2013). In contrast, higher level, post-lexical processing beyond the word level has received very limited attention. The few exceptions include work on semantic plausibility (Connor et al., 2015; Joseph et al., 2008), syntactic ambiguity (Joseph et al., 2013) and local comprehension monitoring (Vorstius et al., 2013). Vander Schoot et al. (2012) examined global text comprehension using a narrative inconsistency task that allowed individual differences to be assessed in terms of the development and updating of a coherent mental representation of a text passage.

### The present study

Complementing this previous experimental work, the present paper will examine supra-lexical effects using surprisal and embedding similarities in a large corpus of eye movement data collected in a longitudinal study of Chinese elementary school students. In the study, the total accumulated time spent on a word during reading was partitioned into three non-overlapping components: first fixation duration (FFD), re-fixation duration (RFD), and the remainder of any viewing time on the word, which will be referred to as re-reading duration (RRD). The central hypothesis of the study is that eye movement metrics are sensitive to surprisal and supra-lexical semantic similarity measures. We should also see develop-

mental changes in sensitivity to these text measures. With respect to local context effects as expressed in surprisal measures, the hypothesis is that surprisal measure will have a stronger effect in the early viewing time phase as it quantifies how predictable a word is, while supra-lexical semantic similarity measures will have a stronger influence in the later viewing time phase as they relate to higher-order semantic features of the sentence and paragraph. Furthermore, over the course of development there should be a gradual increase in the sensitivity of readers to such influences. Therefore, it is predicted that there is a greater likelihood of surprisal effects in 5th graders rather than 4th graders, with a gradual increase as they become more skilled readers. Similarly, with respect to global contextual influences it would be expected to see a growing sensitivity to sentence coherence as a child increased their reading proficiency.

## Methods

### Participants

The study involved two large-scale longitudinal data collections at Huilai Yingnei primary school in Guangdong Province, China. Data was collected from the same cohort of students in 2017 and 2018. Raven's Standard Reasoning Test, normed for China (Zhang & Wang, 1989), and the Literacy Test for the Primary School Students (Wang & Tao, 1993) were administered to all 768 grade 4 students attending the school as intelligence and literacy tests, respectively. The Literacy Test for the Primary School Students is a standardised test to measure pupil's vocabulary size which has high reliability and validity, both of which are 0.8. It has been used in research to assess pupil's reading ability, since how well a child can read is correlated with the size of their vocabulary (Xiong, 2014). The results of the Chinese language end-of-term exam, which was administered by the school just prior to the experiment, was also used as a participant selection criterion. The exam is used as an indicator of participants' reading comprehension as it includes paragraph reading and writing tasks. Fifty-nine participants of different reading ability were chosen from the experiment described for the analysis in this paper. Each experiment took between 30 and 40 minutes per participant.

The "poor" reader group came from 20 students with the lowest literacy test scores and below average Chinese term exam scores. The "good" reader group came from 19 students of similar age and intelligence level but with the highest literacy test scores and above average Chinese term exam scores. The "average" reader group comprised 20 students of similar age and intelligence level but with literacy test scores within  $\pm 0.3$  standard deviation (SD) of the mean and with Chinese term exam scores within  $\pm 0.5$  SD of the mean. Finally, all the students in the study had normal or corrected vision and had not participated in any prior eye movement studies or similar reading tests.

## Materials

Reading material were a translation of age appropriate short stories from the Florida Assessment for Instruction in Reading (FAIR) toolkit (Florida Department of Education, 2009). Five and six stories were used in the first and second data collections, respectively. Each story was presented in multiple paragraphs (3-4 paragraphs), with each paragraph consisting of 5-7 lines (Stenner et al., 2007). Word length ranged from 1 to 5 characters with a total of 4610 characters, 592 unique. Stories were presented on the screen one by one. Each story was followed by three comprehension questions. Paragraphs were displayed in black on a light grey background using a 19.5-inch flat-panel monitor. Display resolution was set to 1024x768 pixels with a refresh rate of 60 Hz. Texts were presented in Xinhei font at 30 px, left aligned, and double-spaced. Viewing distance was adjusted to 68 cm. At this distance, each character subtended approximately  $1^\circ$  of visual angle laterally. Viewing was binocular and eye movements of the participants' right eye were recorded using the EyeLink 1000 (SR Research Ltd., 2010), with a sampling rate of 500Hz.

## Procedure

Participants were seated in front of the presentation monitor and received directions for the upcoming task on the display screen in front of them. Participants received identical directions for the reading task, instructing them to read every text so that they understood its meaning and were able to answer comprehension questions. They were advised to read silently. It was also explained that this was not a reading test or contest in order to make them feel more comfortable about the situation. Participants read four three-line practice trials

from one story before the reading experiment, to familiarise them with the calibration routine and eye tracking procedures. A 9-point calibration was performed at the beginning of each story. For some participants extra calibrations were needed during the experiments due to head movements. Mean average position error in an accuracy validation routine did not exceed  $0.33^\circ$  of visual angle. A drift-check before every paragraph ensured accuracy between calibrations. If the drift check showed a deviation of more than  $0.33^\circ$  of visual angle, an additional calibration was performed. These settings (Figure 3) have proven to produce accurate and reliable data in multiple reading studies across different laboratories (see Inhoff & Radach, 1998, McConkie et al., 1991 for detailed discussions of methodological issues). Children could take breaks between tasks or before calibrations, if necessary. Reading was self-paced and children pressed a mouse button to signal that they were done with a trial. The next paragraph or the comprehension question appeared immediately following the mouse press.

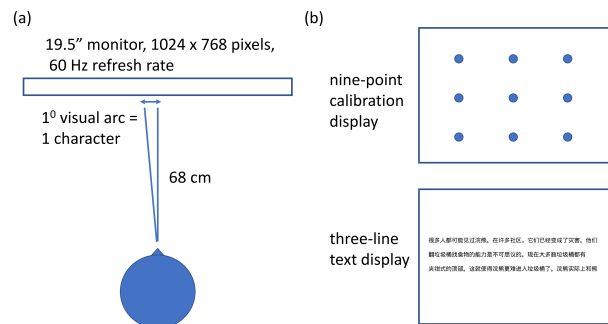


Figure 3. (a) A schematic representation of the experimental setup. Participants sat approximately 68 cm from the display 19.5" display. Characters subtended one degree from visual arc; (b) At the start of each reading session participants' eye movements were calibrated using nine-point grid. Following satisfactory calibration, participants read text three lines at a time.

## Data pre-processing

Fixation in which participants blinked, fixations on the first and last words of each line, the first and last fixations of each trial and switch line fixations were excluded from analyses. Extreme fixation duration values less than 80 ms and of greater than 800 ms were discarded (Inhoff & Radach, 1998). Saccades located exactly at the word boundaries were also excluded. A total of



114,755 fixations contributed to the analysis. Note that the Jieba Chinese text segmentation algorithm was used for word segmentation (Sun, 2013). Given there is sometimes disagreement on word boundaries in Chinese, the Modern Chinese Word Dictionary (2007) was used as an arbitrator in determining the precise length of words in the experimental materials.

## Results and discussion

Linear mixed models (Bates et al., 2014) were used in the analysis of the eye movement data from the study. Statistical analysis was conducted using R 3.5.3 (R Development Core Team, 2018) and the lmerTest R package (Kuznetsova et al., 2017). A set of dependent measures were derived from readers' word viewing times. As mentioned previously, the total accumulated time spent on a word during reading was partitioned into three independent components: first fixation duration (FFD), re-fixation duration (RFD), and the remainder of any viewing time on the word, re-reading duration (RRD). These components of a word's viewing time reflect increasingly more advanced stages in the processing of the text. FFD usually reflects the initial processing of a word and tends to reflect local processing constraints. RFD is a measure of the overall difficulty of a word since it reflects the amount of additional fixations the word received before the reader moves on to the next word. RRD measures the total additional time the reader spends on the word following their initial visit and tends to reflect the difficulty a reader has in integrating a word into an overall understanding of the text. In the study, FFD is expected to reflect the immediate aspects of word processing, RFD somewhat later lexical processing, and RRD higher-level integration processes or more specifically difficulties in performing this integration (see Vorstius et al., 2014 for a recent discussion of these measures).

These three components of viewing time were modelled using the same set of seven fixed independent variables and two random variables as follows.

$$DV \sim \text{lg10WF} + \text{surp} + \text{perp} * (\text{word\_sent} + \text{sent\_sent} + \text{sent\_para} + \text{para\_para}) + (1 | ID) + (1 | \text{word}) \quad (1)$$

In the analysis described in detail below, the dependent variable DV could be one of the three viewing time de-

compositions discussed above. Participant ID and word were treated as random effects. Seven additional fixed independent variables are from four categories: word frequency, conditional probability-based surprisal, text similarity, and reading ability (a detailed description of the fixed-effect independent variables is given in Table 1).

Table 1. Description of the fixed-effect independent variable in the mixed-effects models

Variable category	Variable name	Description
word frequency measure	lg10WF	The word frequency measure was taken from a Chinese language subtitle database (Brysbaert and New, 2009). lg10WF is a log frequency measure reflecting the number of times the word appears in the corpus.
surprisal measure	surp	Surprisal values were calculated by the relevant n-grams from Google Chinese Web 5-gram corpora. High values indicate low predictability.
text similarity measures	word_sent, sent_sent, sent_para, para_para	Four sets of values were calculated from embeddings encoded by Google's Universal Sentence Encoder: (1) word_sent measures semantic similarity between fixated word and sentence in which it is contained; (2) sent_sent measures the semantic similarity between the current sentence to the previous sentence; (3) sent_para measures semantic similarity between current sentence and paragraph; (4) para_para measures the semantic similarity between the current paragraph and the previous one.
reading ability	perf	the scores of the literacy test for good, average and poor readers

Surprisal based on n-gram derived conditional probability was calculated for each word in the texts as follows (Levy, 2008):

$$\text{surprisal}(w_i) = -\log_2 P(w_i | w_1 \dots w_{i-1}, \text{CONTEXT}) \quad (2)$$

Where CONTEXT is the extra-sentential context, which will be ignored in the case of this study.  $P(w_i | w_1 \dots w_{i-1})$  is the conditional probability of the occurrence of  $word_i$  based on its previous words. For example, the conditional probability of  $word_i$  based on the previous word can be calculated by the frequency count of  $word_{i-1}word_i$  divided by the frequency count of  $word_{i-1}$ . The conditional probability of  $word_i$  based on the last two words can be calculated by the frequency count of  $word_{i-2}word_{i-1}word_i$  divided by the frequency count of  $word_{i-2}word_{i-1}$ . Conditional probability based on three or four words can be similarly calculated. As mentioned in the introduction, the conditional probabilities used here were calculated

using the Google Chinese Web 5-gram corpus, which consists of Chinese word n-grams and their observed frequency counts generated from over 800 million text tokens. The length of the n-grams in the corpus ranges from unigrams to 5-grams (Fang et al., 2010). In natural language processing practice, it's more common to use trigrams where the probability of  $word_i$  is conditional on the probability of its co-occurrence with the previous two words (Jurafsky and Martin, 2014). Also, as n-gram length increases, the problem of data sparsity in the corpus increases. Therefore, surprisal measures are limited to those based on the previous two words in a sentence.

Four measures of text similarity (word-sentence, sentence-sentence, sentence-paragraph, paragraph-paragraph) were calculated using the embeddings derived from Google's Universal Sentence Encoder (Bengio et al., 2003). The Universal Sentence Encoder generates vectors of 128 dimensions for Chinese words, sentences, and paragraphs. The semantic similarity of pairs of language items using their embedding vectors  $x$  and  $y$  were calculated as follows (also see Figure 4):

$$x_{\sim}y = 1 - \arccos(\vec{x} \times \vec{y})/\pi \quad (3)$$

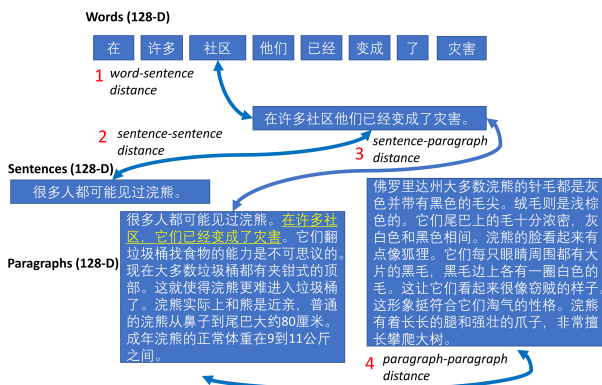


Figure 4. Measures of similarity at multiple scales

The variable  $word\_sent$  measures the semantic similarity between the embeddings representing the fixated word and its containing sentence. This measure would be expected to vary significantly for each word in the sentence. For example, the similarities between function word embeddings and those of the sentence's main content words. The variable  $sent\_sent$  measures the semantic similarity between the sentence in which the fixated word

is contained and the preceding sentence. Sentences that are similar by this measure will tend to be dealing with the same topic, whereas a topic-shift between sentences would lead to a decrease in similarity. The variable  $sent\_para$  measures the semantic similarity between current sentence and its containing paragraph. The measure is intended to quantify the degree to which the current sentence is coherent with its paragraph. It can be viewed as a proxy for the potential impact on the reader of topic shifting or the introduction of new information. The variable  $para\_para$  measures the semantic similarity between the current paragraph and the previous one. All four measures tap into slightly different aspects of coherence, arguably at different levels of processing. Note that all sentence and paragraph similarity measures are identical for each word in a given sentence.

The means and standard deviations of the three dependent measures are summarised in Table 2. As can be seen from the table, there is a decline in the means and standard deviations of first fixation duration (FFD), re-fixation duration (RFD) and rereading duration (RRD) as readers progress from grade 4 to grade 5. Figure 5 below, shows the decomposition of viewing time as a function of reading ability and grade. We can see that the readers classified as poor and average show a slight decline in overall viewing times across grade, while those classified as good show a slight increase in overall viewing times.

Table 2. Numbers of observations, Means and SDs for the three dependent variables (DV's). Note that FFD = first fixation duration; RFD = re-fixation duration; RRD = rereading duration.

DV's	Grade 4			Grade 5		
	N	Mean	St. Dev.	N	Mean	St. Dev.
FFD	32,390	248	133	39,129	236	110
RFD	4,417	334	237	4,536	319	232
RRD	10,443	441	351	13,346	433	347

Table 3 shows the result of the mixed linear models for the three dependent measures. There were overall significant effects of word frequency for all viewing time measures: the higher the frequency, the shorter the viewing time. Reading ability had a statistically significant effect on FFD, with the more able students having shorter viewing times. Surprisal only had significant effect on FFD: the higher the surprisal, the longer the viewing

time. Contrary to what was expected, the text similarity measures only had a statistically significant impact on the early viewing time (FFD): the more similar the current and preceding sentence, the shorter the first fixation duration.

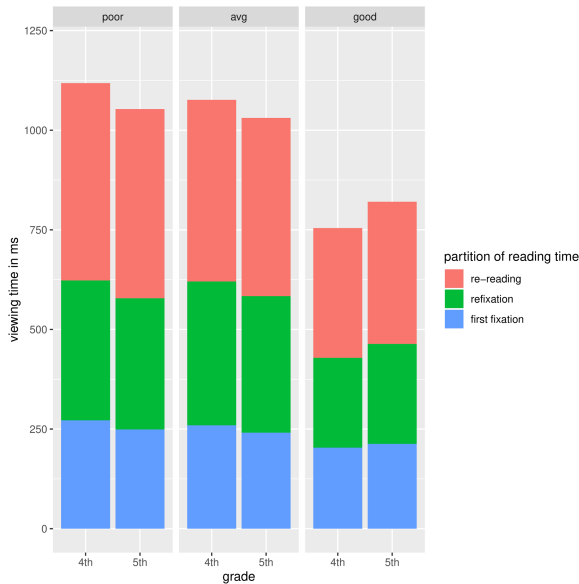


Figure 5. Decomposition of viewing times as a function of grade and reading ability

Table 3. Linear mixed model estimates for the three dependent measures

	Dependent variable		
	FFD	RFD	RRD
lg10WF	<b>-21.4(1.2)***</b>	<b>-19.1 (1.1)***</b>	<b>-43.8(2.5)***</b>
surp	<b>0.4(0.2)**</b>	<b>0.3(0.1)*</b>	-0.3(0.2)
perf	<b>-115.5(61.8)***</b>	-47.5(43.7)	-115.7(94.8)
word_sent	-182.3 (182.9)	159.8 (134.1)	532.6(291.5)
sent_sent	<b>-346.7(159)*</b>	-175.1(116.5)	-210.2(253.2)
para_para	101.4(198.7)	-88.7(145.6)	-571.9(316.5)
sent_para	293.7 (160.3)	66(117.5)	277.3(255.4)
perf: word_sent	62.2 (59.7)	-45 (43.8)	-159.5(95.1)
perf:sent_sent	<b>108.5(52)*</b>	52.2(38.1)	74.5(82.9)
perf:para_para	-23.8 (64.8)	30.9(47.5)	176.9(103.2)
perf:sent_para	-101.7 (52.4)	-20.7(38.5)	-104.5(83.5)
intercept	<b>592.1(189.4)**</b>	236.5(133.9)	<b>636.6(290.5)*</b>

Note: \*p≤0.05; \*\*p≤0.01; \*\*\*p≤0.001

However, the result of the mixed linear models for the three dependent measures done separately for each grade show a slightly different picture of FFD and RRD for fifth graders (Table 4).

### First fixation duration

In the overall analysis, reading ability and word frequency were the dominant factors affecting first fixation duration (FFD) and doing so in the obvious direction - higher word frequency and greater reading ability significantly reduced FFD. Increased surprisal also significantly raised FFD - the more unexpected the word in the sentence context, the longer the FFD. One of the similarity measures, sentence-to-sentence, has a marginally significant effect on FFD, suggesting that if the current and preceding sentence are similar enough, it will shorten FFD for the words in the current sentence. However, this effect is more pronounced for poorer readers, as evidenced by the significant ability-by-similarity interaction. Figure 6 shows FFD as a function of current and preceding sentence similarity and reading ability after removal of between-subject and between-word variance of the dependent variables (Hohenstein & Kliegl, 2014). In this figure, we can see that the source of the significant interaction is a slight floor effect, whereby there is scope for shortening the FFD as result of sentence similarity in the case of less proficient readers, but that there's less room for improvement for the shorter FFDs of more able readers.

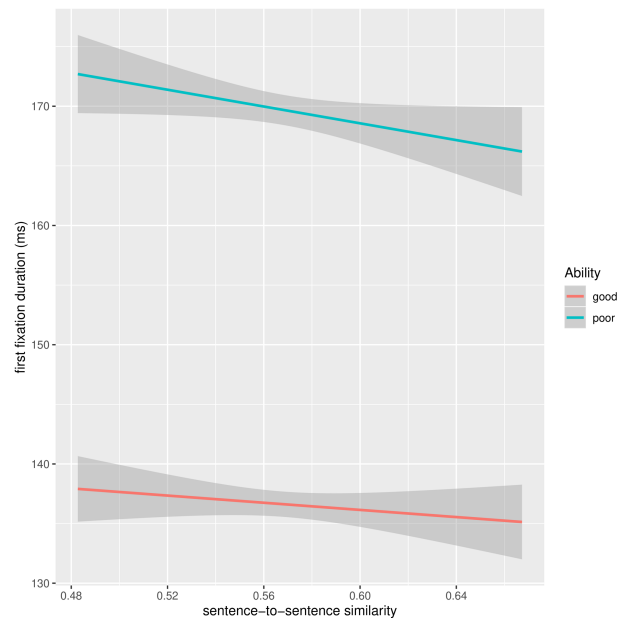


Figure 6. FFD as a function of current and preceding sentence similarity and reading ability: random effects removed



Table 4. Linear mixed model estimates for the three dependent measures across grade

	Dependent variable					
	FFD <sub>G4</sub>	FFD <sub>G5</sub>	RFD <sub>G4</sub>	RFD <sub>G5</sub>	RRD <sub>G4</sub>	RRD <sub>G5</sub>
lg10WF	<b>-21.3 (1.6)***</b>	<b>-22.6 (1.3)***</b>	<b>-19.6(1.3)***</b>	<b>-17.4(1.2)***</b>	<b>-39.7(3.2)***</b>	<b>-44(2.8)***</b>
surp	<b>0.6(0.2)*</b>	0.3(0.2)	0.3(0.2)	0.2(0.1)	-0.4(0.4)	-0.1(0.3)
perf	<b>-254.7(101.6)*</b>	<b>-179.8(74.8)*</b>	-113.1(75.5)	-49.6(52.5)	<b>-315.6(154)*</b>	-53.4(120.6)
word_sent	-49.5 (285.9)	-393.8 (232.6)	201.1(217.2)	93.2(168.2)	<b>1081.7(443.1)*</b>	38.8(381.6)
sent_sent	-21.3(251.6)	-249.7(204)	-272.4 (191.1)	23.7 (147.4)	-181.4(389.8)	152.9(334.5)
para_para	<b>-1095.9 (389)**</b>	-385.7 (229.9)	-477.3 (295.4)	-201 (166.5)	<b>-1630.8(602.7)**</b>	-706(337.1)
sent_para	293.7(246.8)	384.5(207.6)	164(187.4)	-2.7(150.1)	-104.4(382.4)	635.6 (340.5)
perf:word_sent	8.8 (93.3)	141.8 (75.9)	-62.4(70.9)	-19.9(54.9)	<b>-342.1(144.5)*</b>	2(124.5)
perf:sent_sent	-21.2(82.4)	94.6(66.7)	86.5 (62.6)	-9 (48.2)	51.6(127.6)	-27.1(109.4)
perf:para_para	<b>332.1 (126.6)**</b>	112.7 (75)	141.3 (96.2)	65.6 (54.2)	<b>518.2(196.2)**</b>	202.7(123)
perf:sent_para	-96(80.6)	<b>-135.7(67.9)*</b>	-50.3(61.2)	-0.5(49.1)	28.6(124.8)	<b>-232.9(113.3)*</b>
intercept	<b>1111.9(311.1)***</b>	<b>784.8(229.1)***</b>	<b>473.9 (231.2)*</b>	233.3(161)	<b>1261.9(471.9)**</b>	463.8(369.9)

Note:

\*p<0.05; \*\*p<0.01; \*\*\*p<0.001

In Table 4, grade 4 values are compared to grade 5 and apart from consistent frequency and ability effects across grades, it appears that greater paragraph similarity benefits grade 4 students more than grade 5. Moreover, reading ability also plays a role in this effect, with the less able readers benefiting more from paragraph similarity as indicated by the significant interaction with ability. This interaction is visualised in Figure 7, which shows FFD for 4th and 5th grade readers as a function of current and preceding paragraph similarity and reading ability after removal of between-subject and between-word variance of the dependent variables. The greater benefit for the less able readers may be another instance of the floor effect mentioned above, where there is an uncompressible lower bound on FFD in grade 4, which limits the amount of improvement that can be gained from exploiting paragraph similarity. The pattern of data for grade 5 in the same figure shows an overall decrease in FFD, but this time with no statistically significant interaction with reading ability.

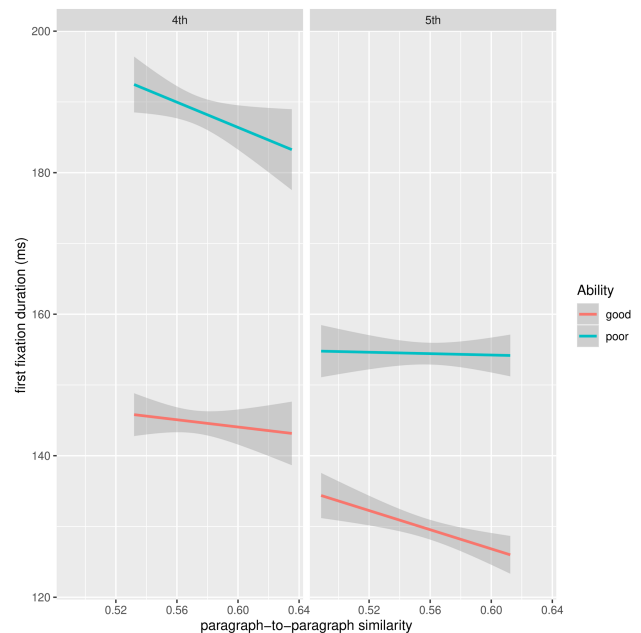


Figure 7. FFD for 4th and 5th grade as a function of current and preceding paragraph similarity and reading ability: random effects removed

### Refixation duration

In the global analysis of refixation duration (RFD), while there is a significant effect of word frequency and a marginal effect of surprisal, no other independent measures reach statistical significance. When we look at the analysis by grade, there is only a significant effect of word frequency. The absence of significant effects may be a consequence of the relatively small number of words refixations in the data. This probably arises from the unique characteristics of Chinese text. In the texts used for the study, 54% of the words consisted of just one character, while 42% comprised two characters. More generally, 70% of Chinese words comprise two characters, 20% one character, and 10% three or more characters (Modern Chinese Frequency Dictionary, 1986). The shorter average word length in the texts reflects the fact that they were designed for reading by children. Furthermore, given the spatial compactness of the Chinese writing system, one fixation will tend to suffice in most cases for the satisfactory identification of words. Overall, therefore, the RFD measure may not be as reliable a metric of word processing in the case of Chinese reading as it is for alphabetic writing systems with more heterogeneous word lengths. In effect, the RRD measure for Chinese reading might tend to incorporate the RFD metric that normally would be distinct in alphabetic reading.

### Re-reading duration

In the global analysis, only word frequency had a significant effect on re-reading duration (RRD) and in the expected direction. No other independent measures reach overall statistical significance.

However, when RRD is analysed separately by grade, there is a significant sensitivity to various similarity measures among grade 4 readers. The strongest effects are for paragraph similarity, with high similarity reducing RRD. Figure 8 shows RRD for 4th and 5th grade as a function of current and preceding paragraph similarity and reading ability after removal of between-subject and between-word variance in the dependent variables. The source of the interaction between ability and paragraph similarity in 4th grade readers appears this time to be the greater benefit good readers derive from paragraph similarity compared to less able readers. A similar pattern is seen for 5th grade readers, though not a statistically significant one.

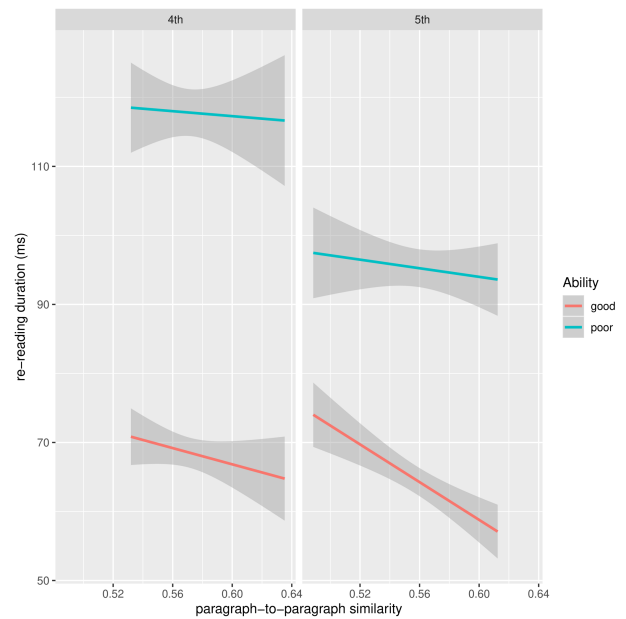


Figure 8. RRD for 4th and 5th grade as a function of current and preceding paragraph similarity and reading ability: random effects removed

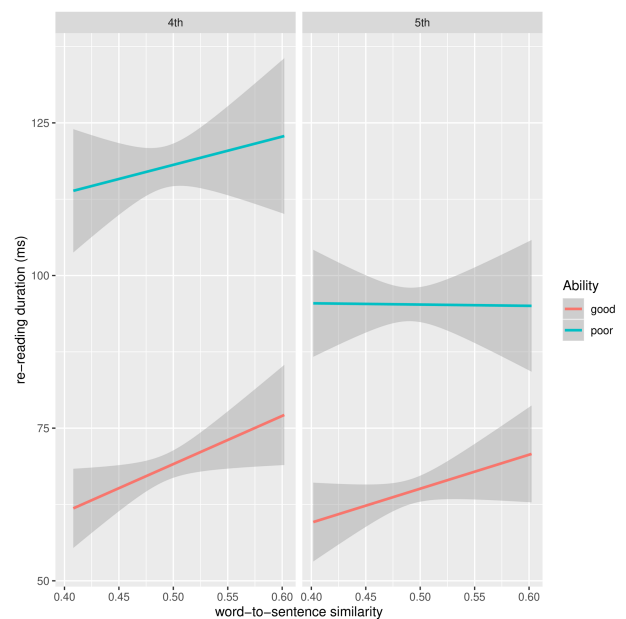


Figure 9. RRD for 4th and 5th grade as a function of word and sentence similarity and reading ability: random effects removed

Figure 9 shows the marginally significant interaction between reading ability and word-sentence similarity for grade 4 readers. The source of the interaction is not entirely clear and, moreover, it is not clear why greater word-sentence similarity should lead to an increase in RRD for both ability levels. One possibility is that the effect is due to the reader’s integration efforts. Recall that RRD measures repeated visits to a word, so perhaps greater similarity gives rise to an increase in confirmatory fixations. The elevated trend in RRD for good readers is also apparent in grade 5, though without reaching significance.

### Path analysis of semantic similarity measures

Given the, albeit relatively small, colinearity of the family of semantic similarity measures used in the preceding analysis and the potential complexity of interpretation of their influence as discussed above, it was decided to unpack the pattern of inter-dependence among the measures in more detail using path analysis (Rosseel, 2012). Specifically, the degree to which these measures directly and indirectly influenced two of the three dependent measures, FFD and RRD, was explored. RFD was omitted from these analyses because of the relatively low numbers of observations involved after partitioning the data and also the problematic nature of RFD given the homogeneous word length of Chinese, particularly in the children’s texts used. Another focus of the analysis was to see if the patterns of influence varied as a function of reading ability. It was hypothesised that FFD, as an early viewing time measures, would show less direct and indirect influence from the sentence and paragraph similarity measures. On the other hand, the later RRD measure was more likely to show influences of broader contextual factors. Furthermore, poorer readers should show less direct and indirect influence of extra-sentential similarities than better readers, based on the assumption that less skilled readers are more narrowly focused on local lexical context and less on the sentential and paragraph level.

Table 5 is the output from an overall path analysis examining the pattern of influences (direct and indirect) on FFD and RRD. Examination of the pattern of significant estimates shows the significant paths for FFD are also significant for RRD, with RRD having two additional significant paths of influence from paragraph related

measures. The pattern of significant paths is visualised in Figure 10, where Figure 10 (a) represents all possible paths of influence and (b) and (c) represent the paths with a significance of  $P < 0.05$  for FFD and RRD, respectively. It is apparent that RRD shares the same pattern of influences as FFD but with additional pathways from sentence-paragraph similarity. This difference can be viewed as a progression from sentence level to paragraph level influences to which FFD and RRD are expected to be differentially sensitive. Note that the dashed paths represent negative estimates, indicating a reduction in viewing time as a function of increased similarity, whereas positive estimates indicate an increase in viewing time. Note also that word frequency and surprisal are included in the path analysis but are not shown in the path diagrams. The inclusion of surprisal explains the absence word\_sentence similarity effects in this and all subsequent path analyses. When surprisal is omitted from the path analysis model, all of the paths of influence on RRD are mediated through the word\_sentence variable. However, it was decided to use the full model to derive the paths in order to align better the analysis with the preceding linear mixed-model analyses.

Table 5. Estimates (and their standard errors) for the path analyses of the impact of the similarity measures on FFD and RRD dependent measures. Note that pp = paragraph\_paragraph similarity; sp = sentence\_paragraph similarity; ss = sentence\_sentence similarity; ws = word\_sentence similarity.

	FFD	RRD
pp→	0.206 (0.115)	-0.176 (0.105)
ss→	<b>-0.736 (0.097)***</b>	<b>-0.191 (0.088)*</b>
sp→	0.045 (0.097)	<b>-0.188(0.089)*</b>
ws→	-0.145 (0.111)	-0.000 (0.101)
pp→sp→ss→ws→	0.001 (0.000)	0.000(0.000)
pp→sp→ws→	-0.001 (0.001)	-0.000(0.000)
pp→sp→ss→	<b>-0.07 (0.009)***</b>	<b>-0.018 (0.008)*</b>
pp→ss→ws→	-0.001 (0.001)	-0.000 (0.001)
pp→ss→	<b>0.139 (0.018)***</b>	<b>0.036(0.017)*</b>
pp→ws→	-0.002 (0.001)	-0.000 (0.001)
pp→sp→	0.015 (0.032)	<b>-0.062 (0.029)*</b>
sp→ss→ws→	0.002 (0.001)	0.000 (0.001)
sp→ws→	-0.002(0.002)	-0.000 (0.001)
sp→ss→	<b>-0.212 (0.028)***</b>	<b>-0.055 (0.025)*</b>
ss→ws→	0.006 (0.005)	0.000 (0.004)

Note: \* $p \leq 0.05$ ; \*\* $p \leq 0.01$ ; \*\*\* $p \leq 0.001$

The strength of various paths' influence on reading time measures can be considered to reflect efforts by the reader to integrate what they are reading into their developing understanding of the text. Therefore, an analysis of the strength of paths for readers of different reading ability should tell us something about readers' growing sensitivity to higher-order properties of the text. The main measures where we would expect to see differences

are in the FFD and RRD viewing times, since the preceding LMM analyses have shown that semantic similarity measures have a significant effect on these two variables.

Table 6 provides a partitioning of the path analysis into good and poor readers for the FFD and RRD measures. While there is little difference in the pattern of significant paths on the basis of reading ability for the FFD measure, there is a clearer divergence in the case of RRD. More able readers show viewing time benefits from greater sentence and paragraph similarity, while poorer readers show no significant benefits. The significant paths for RRD for the more able readers are illustrated in Figure 11. An equivalent figure for poorer students would show no visible paths. The one inconsistent result in these data is the significant positive estimate in Table 6 for the FFD of the less able students. It's unclear why greater similarity of paragraphs should have an impact on FFD, let alone for less able readers and best explored further within an experimental rather than corpus-based paradigm.

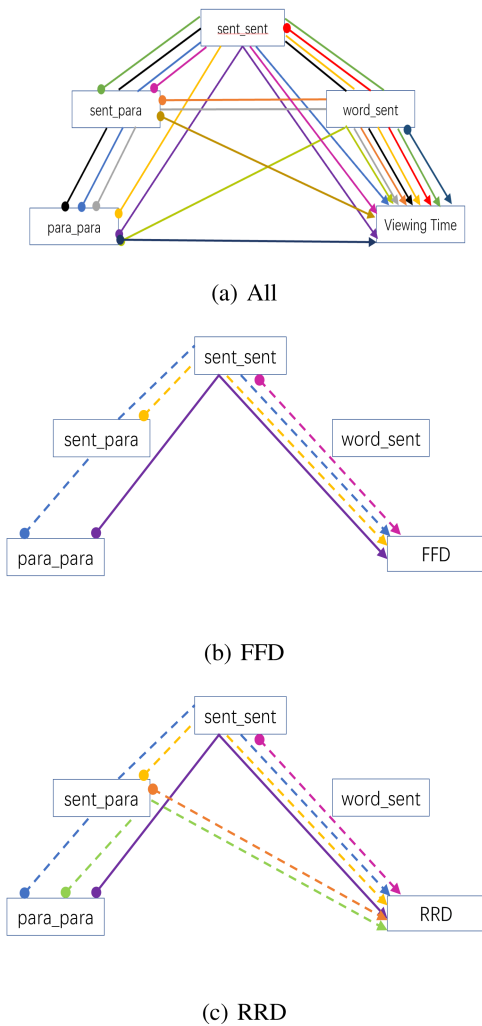


Figure 10. Path analysis showing the patterns of influence of the four text similarity measures used in the analysis of viewing time data: (a) all possible paths; (b) significant paths influencing first fixation duration (FFD); and (c) significant paths influencing re-reading duration (RRD). Distinct paths are illustrated with different colours, where the start of the path is represented by a filled circle and its end by an arrowhead. In (b) and (c) only paths with  $p < 0.05$  are graphed, dashed paths indicate a negative estimate, and solid lines a positive one.

Table 6 Estimates (and their standard errors) for the path analyses of the impact of the similarity measures on the first fixation duration (FFD) and re-reading duration (RRD) as a function of reading ability. Note that pp = paragraph\_paragraph similarity; sp = sentence\_paragraph similarity; ss = sentence\_sentence similarity; ws = word\_sentence similarity.

	FFD <sub>poor</sub>	FFD <sub>good</sub>	RRD <sub>poor</sub>	RRD <sub>good</sub>
pp→	<b>0.484 (0.198)*</b>	-0.24 (0.198)	-0.136 (0.197)	<b>-0.443(0.158)**</b>
ss→	<b>-0.831 (0.165)***</b>	<b>-0.744 (0.170)***</b>	-0.058 (0.164)	-0.143(0.135)
sp→	0.251 (0.166)	0.085 (0.169)	-0.105(0.165)	<b>-0.271 (0.134)*</b>
ws→	-0.272 (0.19)	-0.106 (0.193)	-0.088 (0.189)	0.152(0.154)
pp→sp→ss→ws→	0.001 (0.001)	0.000 (0.001)	0.000 (0.001)	-0.001 (0.001)
pp→sp→ws→	-0.001(0.001)	-0.000(0.001)	-0.000(0.001)	0.001(0.001)
pp→sp→ss→	<b>-0.078 (0.016)***</b>	<b>-0.068 (0.016)***</b>	-0.006 (0.015)	-0.013 (0.012)
pp→ss→ws→	-0.002(0.002)	-0.001(0.002)	-0.001 (0.002)	0.001 (0.001)
pp→ss→	<b>0.158 (0.032)***</b>	<b>0.154 (0.035)***</b>	-0.001 (0.002)	0.001 (0.001)
pp→ws→	-0.003 (0.003)	-0.001 (0.002)	-0.001 (0.002)	0.002(0.002)
pp→sp→	0.083 (0.055)	-0.027 (0.053)	-0.035(0.055)	<b>-0.086 (0.043)*</b>
sp→ss→ws→	0.003 (0.002)	0.001 (0.002)	0.001(0.002)	-0.002(0.002)
sp→ws→	-0.003(0.003)	-0.002(0.003)	-0.001(0.002)	0.002(0.002)
sp→ss→	<b>-0.237 (0.047)***</b>	<b>-0.215 (0.049)***</b>	-0.017(0.047)	-0.041(0.039)
ss→ws→	0.012 (0.008)	0.004 (0.008)	0.004 (0.008)	-0.006(0.005)

Note: \* $p \leq 0.05$ ; \*\* $p \leq 0.01$ ; \*\*\* $p \leq 0.001$

Overall, the path analysis results have highlighted the potential for using the emergence of sensitivity to high-order similarity measures (e.g., sentence and paragraph similarity) as indices of reading development. Alternatively, a lack of such sensitivity may serve as an indicator of readers at risk. It is important to note that the current

study was a corpus-based analysis, using naturally occurring patterns of word, sentence, and paragraph similarity and cohesion. The next step is to design controlled experiments where we probe the finer detail of the sensitivity uncovered in this preliminary exploration. Such experiments would involve, for example, controlling the specific text features that serve to increase or decrease similarity. In addition, there may be identifiable sub-groups of readers who vary in their sensitivity to specific similarity configurations.

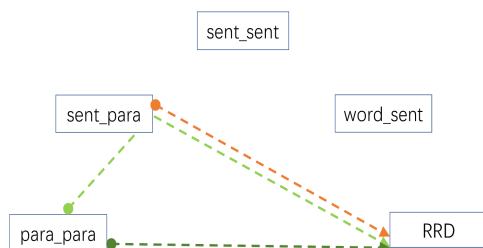


Figure 11. Results of a path analysis showing the significant patterns of influence ( $p < 0.05$ ) of the four similarity measures used in the analysis of RRD for good readers. Distinct paths are illustrated with different colours, where the start of the path is represented by a filled circle and its end by an arrowhead. Dashed paths indicate a negative estimate and a reduction in viewing time, solid lines are positive estimates that contribute to an increase in viewing time.

## Conclusion

The results described in this paper demonstrate that text similarity measures have a significant impact on moment-to-moment processing of words in reading. Previous research has demonstrated that n-gram and LSA contextual measures have an impact on word viewing times in adult readers (Pynte et al., 2008, 2009). However, this is the first attempt to track the developmental trajectory of these influences in Chinese early readers as well as readers with differing reading abilities.

While the underlying factors driving the developmental change in response to supra-lexical properties of the text clearly need to be explored further, on the basis of what has been found one should be cautious about attributing the main changes in the developing reader merely to changes in the efficiency of lexical processing (cf. Engbert et al., 2005; McDonald et al., 2005; Reichle et al., 2003; Reilly and Radach, 2006 for arguments along

these lines). The study finds that there are robust contextual effects impacting on the local processing of words during a fixation and the nature of these effects changes as the reader becomes more proficient.

Another contribution this paper is to present an easy-to-use set of tools for linking the low-level aspects of fixation durations to a hierarchy of sentence-level and paragraph level features that can be computed automatically. The use of the decomposition of word viewing times into immediate and later components combined with measures of sentence and paragraph coherence illuminates the time-course of the reader's processing of a text. Similar to the study by Radach et al. (2008), broader contextual constraints have been shown to impact on low-level aspects of the reading process.

Finally, the similarity-based measures could also be used to assess text for their suitability for readers of different levels of ability. While there are text complexity measures such as "Lexile" (Lexile, 2019) available for English texts, nothing comparable exists for Chinese. The measures described here could be applied to any language for which there is a large text corpus. Moreover, the Lexile measure is primarily calculated as a function of word frequency and sentence length. The text coherence measures described here could usefully augment a lexile-like measure to provide quantitative measures of the semantic characteristics of the sentences and texts in addition to word frequency and sentence length.

## Ethics and Conflict of Interest

The authors declare that the contents of the article are in agreement with the ethics described in <http://biblio.unibe.ch/portale/elibrary/BOP/jemr/ethics.html> and that there is no conflict of interest regarding the publication of this paper.

## References

- Altszyler, E., Sigman, M., Ribeiro, S., & Slezak, D. F. (2016). Comparative study of lsa vs word2vec embeddings in small corpora: A case study in dreams database. arXiv preprint arXiv:1610.01520.
- Arora, S., Liang, Y., & Ma, T. (2016). A simple but thought-beat baseline for sentence embeddings.



- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.
- Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of machine learning research*, 3(Feb), 1137–1155.
- Blythe, H. I., Häikiö, T., Bertam, R., Liversedge, S. P., & Hyönä, J. (2011). Reading disappearing text: Why do children refixate words? *Vision research*, 51(1), 84–92.
- Blythe, H. I., & Joseph, H. S. (2011). Children's eye movements during reading.
- Blythe, H. I., Liversedge, S. P., Joseph, H. S., White, S. J., & Rayner, K. (2009). Visual information capture during fixations in reading for children and adults. *Vision research*, 49(12), 1583–1591.
- Boston, M., Hale, J., Kliegl, R., Patil, U., & Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the potsdam sentence corpus. *The Mind Research Repository (beta)*, (1).
- 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(4), 977–990.
- Buswell, G. T. (1922). *Fundamental reading habits, a study of their development*. Chicago, Ill. : The University of Chicago.
- Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Et al. (2018). Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.
- Connor, C. M., Radach, R., Vorstius, C., Day, S. L., McLean, L., & Morrison, F. J. (2015). Individual differences in fifth graders' literacy and academic language predict comprehension monitoring development: An eye-movement study. *Scientific Studies of Reading*, 19(2), 114–134.
- Engbert, R., Nuthmann, A., Richter, E. M., & Kliegl, R. (2005). Swift: A dynamical model of saccade generation during reading. *Psychological review*, 112(4), 777.
- Fang, L., Yang, M., & Lin, D. (2010). Chinese web 5-gram version 1. Retrieved March 4, 2019, from <https://catalog.ldc.upenn.edu/LDC2010T06>
- Florida Department of Education, a. (2009). *Florida assessments for instruction in reading administration manual: Grades 3-12 (tech. rep.)*. State of Florida, Department of Education.
- Häikiö, T., Bertram, R., Hyönä, J., & Niemi, P. (2009). Development of the letter identity span in reading: Evidence from the eye movement moving window paradigm. *Journal of experimental child psychology*, 102(2), 167–181.
- Hohenstein, S., & Kliegl, R. (2014). Semantic preview benefit during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(1), 166.
- Huestegge, L., Radach, R., Corbic, D., & Huestegge, S. M. (2009). Oculomotor and linguistic determinants of reading development: A longitudinal study. *Vision Research*, 49(24), 2948–2959.
- Hyönä, J., & Olson, R. K. (1995). Eye fixation patterns among dyslexic and normal readers: Effects of word length and word frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21(6), 1430.
- Inhoff, A. W., & Radach, R. (1998). Definition and computation of oculomotor measures in the study of cognitive processes, In *Eye guidance in reading and scene perception*. Elsevier.
- Joseph, H. S., Liversedge, S. P., Blythe, H. I., White, S. J., Gathercole, S. E., & Rayner, K. (2008). Children's and adults' processing of anomaly and implausibility during reading: Evidence from eye movements. *The Quarterly Journal of Experimental Psychology*, 61(5), 708–723.
- Joseph, H. S., Liversedge, S. P., Blythe, H. I., White, S. J., & Rayner, K. (2009). Word length and landing position effects during reading in children and adults. *Vision Research*, 49(16), 2078–2086.
- Joseph, H. S., Nation, K., & Liversedge, S. P. (2013). Using eye movements to investigate word frequency effects in children's sentence reading. *School Psychology Review*, 42(2), 207–223.
- Jurafsky, D., & Martin, J. H. (2014). *Speech and language processing (Vol. 3)*. Pearson London.
- Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Skip-thought vectors, In *Advances in neural information processing systems*.

- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). Lmertest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13).
- Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, 25(2-3), 259–284.
- Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106(3), 1126–1177.
- Lexile. (2019). The lexile framework for reading.
- Martin Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Jia, Y., Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, . . . Xiaoqiang Zheng. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems [Software available from tensorflow.org]. <https://www.tensorflow.org/>
- McConkie, G. W., Zola, D., Grimes, J., Kerr, P. W., Bryant, N. R., & Wolff, P. M. (1991). Children's eye movements during reading. *Vision and visual dyslexia*, 13, 251–262.
- McDonald, S. A., Carpenter, R., & Shillcock, R. C. (2005). An anatomically constrained, stochastic model of eye movement control in reading. *Psychological review*, 112(4), 814.
- Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Et al. (2011). Quantitative analysis of culture using millions of digitized books. *science*, 331(6014), 176–182.
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality, In *Advances in neural information processing systems*.
- Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation, In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)*.
- Pynte, J., New, B., & Kennedy, A. (2008). On-line contextual influences during reading normal text: A multiple regression analysis. *Vision research*, 48(21), 2172–2183.
- Pynte, J., New, B., & Kennedy, A. (2009). On-line contextual influences during reading normal text: The role of nouns, verbs and adjectives. *Vision research*, 49(5), 544–552.
- R Development Core Team. (2018). R: A language and environment for statistical computing [ISBN 3-900051-07-0]. ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria. <http://www.Rproject.org>
- Radach, R., Huestegge, L., & Reilly, R. (2008). The role of global top-down factors in local eye-movement control in reading. *Psychological research*, 72(6), 675–688.
- Radach, R., & Kennedy, A. (2004). Theoretical perspectives on eye movements in reading: Past controversies, current issues, and an agenda for future research. *European Journal of Cognitive Psychology*, 16(1- 2), 3–26.
- Radach, R., & Kennedy, A. (2013). Eye movements in reading: Some theoretical context. *The Quarterly Journal of Experimental Psychology*, 66(3), 429–452.
- Radach, R., Schmitt, C., Glover, L., & Huestegge, L. (2009). How children read for comprehension: Eye movements in developing readers. *Beyond decoding: The biological and behavioral foundations of reading comprehension*, 75–106.
- Rayner, K. (1985). The role of eye movements in learning to read and reading disability. *Remedial and Special Education*, 6(6), 53–60.
- Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3), 372.
- Reichle, E. D., Rayner, K., & Pollatsek, A. (2003). The ez reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and brain sciences*, 26(4), 445–476.
- Reilly, R. G., & Radach, R. (2006). Some empirical tests of an interactive activation model of eye movement control in reading. *Cognitive Systems Research*, 7(1), 34–55.

- Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2), 1–36. <http://www.jstatsoft.org/v48/i02/>
- Stenner, A., Burdick, H., Sanford, E., & Burdick, D. (2007). The lexile framework for reading technical report. Durham, NC: Metametrics, 6.
- Subramanian, S., Trischler, A., Bengio, Y., & Pal, C. J. (2018). Learning general purpose distributed sentence representations via large scale multi-task learning. arXiv preprint arXiv:1804.00079.
- Sun, J. (2013). Jieba chinese text segmentation. GitHub, Inc.
- Taylor, S. E., Frackenpohl, H., & Pettee, J. L. (1960). Grade level norms for the components of the fundamental reading skill. Educational Developmental Laboratories.
- Van der Schoot, M., Reijntjes, A., & van Lieshout, E. C. (2012). How do children deal with inconsistencies in text? an eye fixation and self-paced reading study in good and poor reading comprehenders. *Reading and Writing*, 25(7), 1665–1690.
- Vorstius, C., Radach, R., & Lonigan, C. J. (2014). Eye movements in developing readers: A comparison of silent and oral sentence reading. *Visual Cognition*, 22(3- 4), 458–485.
- Vorstius, C., Radach, R., Mayer, M. B., & Lonigan, C. J. (2013). Monitoring local comprehension monitoring in sentence reading. *School Psychology Review*, 42(2), 191.
- Wang, X., & Tao, B. (1993). Primary school literacy test and evaluation scale. Shanghai education press.
- Wieting, J., Bansal, M., Gimpel, K., & Livescu, K. (2015). Towards universal paraphrastic sentence embeddings. arXiv preprint arXiv:1511.08198.
- Xiong, J. (2014). The oculomotor characters of chinese developmental dyslexia.
- Zhang, H., & Wang, X. (1989). Standardization research on raven's standard progressive matrices in china. *Acta Psychologica Sinica*, 21(02), 3. [http://journal.psych.ac.cn/xlxb/CN/abstract/article\\_758.shtml](http://journal.psych.ac.cn/xlxb/CN/abstract/article_758.shtml)