ReadLet: a Dataset for Oral, Visual and Tactile Text Reading Data of Early and Mature Readers

Marcello Ferro[*], **Claudia Marzi**[*], **Andrea Nadalini**[*]
Loukia Taxitari[†], **Alessandro Lento**[**], **Vito Pirrelli**[*]

[*]Italian National Research Council, Institute for Computational Linguistics, Pisa Italy
{marcello.ferro, claudia.marzi, andrea.nadalini, vito.pirrelli}@cnr.it

[†]Department of Psychology, Neapolis University, Pafos Cyprus
l.taxitari@nup.ac.cy

[**]Biomedical Campus University, Rome Italy
alessandro.lento@unicampus.it

Abstract

The paper presents a time-stamped multimodal dataset for reading research, including multiple time-aligned temporal signals elicited with four experimental trials of connected text reading by both child and adult readers. We illustrate design issues and experimental protocols, as well as the data acquisition process and the post-processing phase of data annotation/augmentation. To evaluate the potential and usefulness of a time-aligned multimodal dataset for reading research, we present a few statistical analyses showing the correlation and complementarity of multimodal time-series of reading data, as well as some results of modelling adults' reading data by integrating different modalities. The total dataset size amounts to about 2.5 GByte in compressed format and is available through the CLARIN infrastructure.

Keywords: text reading, eye movements, finger movements, eye-finger span, synchronisation, parallel processing, multimodality.

1. Introduction

Reading a text for comprehension is a multi-level cognitive task, involving i) decoding and accessing words and their meanings; ii) parsing an entire clause to form a complex meaning unit or *proposition*; iii) connecting the new meaning unit to a growing *network* of propositions forming a *conceptual model* of the text being read; iv) monitoring comprehension and making appropriate inferences (Grabe and Stoller, 2019). Although the most influential eye-tracking research on reading has focused on reading single words or sentences (Rayner, 1998, 2009), increasing concerns with the ecological validity of behavioural language data (Brennan, 2016; Demberg and Keller, 2019; Hasson et al., 2018; Willems, 2015), as well as advances in data recording and processing technologies (Frey et al., 2021; Sato and Mizuhara, 2018; Torres et al., 2021) have gradually shifted the research focus away from specific, highly controlled phenomena, towards real-time processing issues (Jarodzka and Brand-Gruwel, 2017; Kaakinen and Hyönä, 2008; Verhoeven and Perfetti, 2008). In some cases, such a shift turned out to challenge traditional acquisitions from artificially restrained experimental protocols (Coskun et al., in press; Kamienkowski et al., 2016; Kuperman et al., 2013; Wallot et al., 2013), proving that collection of natural reading data at scale can considerably advance our understanding of difficulties in real-life reading.

Another dimension of reading complexity is defined by the inherently multi-sensory nature of reading. Oral reading requires the fine coordination of eye movements and articulatory movements. The eye provides access to the visual stimuli needed for voice articulation to unfold at a relatively constant rate. In turn, articulation can feedback oculomotor control for eye movements to be directed when and where processing difficulties arise. One specific element that makes eye-voice coordination fairly hard to manage is the asynchronicity of the two time series (Inhoff et al., 2011). Eye movements are faster than voice articulation, and are much freer to scout a written text forwards and backwards, availing themselves of a wide range of alternative "moves", including long forward saccades, regressions, refixations and word skippings. A reader must rely on a tight control strategy to ensure that the two processes are optimally coordinated (De Luca et al., 2013; Inhoff et al., 2011; Laubrock and Kliegl, 2015; Silva et al., 2016).

Eye-voice coordination is a crucial cognitive skill in developing reading fluency. A small but robust line of work has examined the knowledge, skills, and behaviours that support the development of *word reading in context* in young readers. This work has shown that the concept of word in print is grounded in letter knowledge and beginning sound awareness, but also in the learners' ability to ac-

LREC-COLING 2024, pages 13595–13609

curately pronounce a printed word in a text line while the index finger is pointing to it (Mesmer and Lake, 2010). The key cognitive insight in developing this ability occurs as learners are able to integrate three emerging sources of information about print and speech: the auditorily anchored understanding of syllables, the linguistic-conceptual knowledge of words, and the unfolding visuospatial understanding of printed words built upon the visual and tactile exploration of the words' spatial dimension (Mesmer and Lake, 2010; Mesmer and Williams, 2015). In attaining an efficient synchronisation between word pointing and the onset of word articulation, the learner must resolve the competing information between the multiple syllables that she hears and feels and individual words that she sees on a printed page (Mesmer and Lake, 2010; Uhry, 1999, 2002).

While most of these reading aspects have been explored and investigated independently, much less work has been conducted so far to study their interaction, also because of the technical difficulty with concurrently recording asynchronous time-series of multimodal signals. In this paper, we present the *ReadLet* dataset, a finely annotated collection of time-stamped, naturalistic text reading data including silent and oral reading sessions by both child and adult readers. Each reading session was either finger-tracked or eye-tracked, and all oral reading sessions were audio-recorded. The resulting dataset is the output of a battery of seamlessly integrated software and hardware language technologies, ranging from automated speech recognition and text readability scoring, to time alignment and convolutional alignment of independent time-series of multimodal signals (Crepaldi et al., 2022; Ferro et al., 2018; Taxitari et al., 2021).

2. Related Work

The analysis of eye movements with eye-tracking data has proven to offer an instrumental window onto the fine spatial and temporal allocation of processing resources over a visual scene, highlighting universal aspects of perception (e.g., Awh et al., 2012; Buschman and Miller, 2007), as well as the influence of goals and processes that are specific of reading (Rayner, 2009). Recently, Lio et al. (2019) studied the connection between eye movements and finger movements in the visual exploration of a picture displayed on a computer touchscreen. Presented with the blurred display of a picture, subjects were instructed to deblur the image by touching the screen area they wanted to inspect in full resolution. A strong correlation was observed between areas deblurred by touching the screen, and subjects' fixation patterns when a full resolution version of the same image was explored only visually. Spatial patterns of finger movements were found to be congruent with patterns of eye fixations on the same image, confirming that (i) eye-hand coordination is a form of developmentally early, natural and accurate motor synergy (Esteve-Gibert and Prieto, 2014), and (ii) tactile exploration of an image can be used as an ecological proxy of visual exploration.

Another familiar situation that exploits the synergy between eye movements and finger movements is when a child is learning to read using the index finger of her dominant hand to point to the letters of written words while reading them out. This is known to help children learn to look at print, and supports basic early reading behaviours such as directional movement, attention focus, and voice-print match (Mesmer and Lake, 2010; Uhry, 2002). Beyond this observational literature on sight reading, most quantitative analyses of finger movements have focused on Braille reading (Hughes et al., 2014; Nonaka et al., 2021), showing that finger sliding across embossed texts is characterised by constant fluctuations in finger velocity through consecutive speed-up and slow-down cycles, mostly due to the bottom-up mechanisms controlling for the programming and execution of slow finger movements.

In reading aloud multiple items, the eyes are observed to lead the voice. The eye is ahead of the spoken words most of the times, as one would expect, since articulation is typically the output of a conscious oculomotor activity. The systematic study of the temporal span between a word's fixation onset and the time the word is articulated (commonly referred to as Eye-Voice Span or *EVS*) can be traced back to Buswell's pioneering work (Buswell, 1920, 1921). His evidence consisted in switching off the light during the reading of a sentence and counting how many words could be articulated after the light was off. The same approach was elaborated a few decades later (Lawson, 1961; Levin and Turner, 1968; Levin and Cohn, 1968; Morton, 1964a,b), when some experimental results appeared to support the view that "subjects tend to read in phrase units" (Levin and Turner, 1968, p. 208), and reading rate and *EVS* were shown to increase with more structured text materials (Morton, 1964a).

The advent of eye-tracking technology at the services of eye movement research started a prolonged period of little interest in the vocal component of reading, interrupted by Inhoff et al. (2011) and De Luca et al. (2013), and more recently by Laubrock and Kliegl (2015) and Silva et al. (2016). While De Luca et al.'s (2013) data include oral text reading data by 16 dyslexic and 16 non dyslexic children, both Inhoff et al.'s (2011) and Laubrock and Kliegl's (2015) materials consisted in single

sentence reading, and Silva et al.'s (2016) in naming words and non words from a list.

In recent years, eye-tracking technology has become increasingly more affordable and easy to use in different environments, fostering the development of more ecological, multi-line reading data repositories. Several eye-tracking corpora of text reading have been created, such as GECO (Cop et al., 2017), Provo (Luke and Christianson, 2018), Copco (Hollenstein et al., 2022) and MECO (Kuperman et al., 2023). This novel generation of eye-tracking corpora has enabled researchers to focus on reading as we experience it in real life. In addition, it provides data resources that can be used for testing a wide range of alternative hypotheses, without the need to design a new experiment and gather new data, which is considerably time-consuming and may require expensive equipment. Large collections of ecological reading data are also instrumental to train and test increasingly more sophisticated computational reading models (e.g. Coltheart et al., 2001; Dilkina et al., 2010; Engbert et al., 2005; Grainger and Jacobs, 1996; Reichle, 2021), and use machine learning technology for diagnostic purposes (e.g. Gran Ekstrand et al., 2021; Nilsson Benfatto et al., 2016; Prabha and Bhargavi, 2020; Rello and Ballesteros, 2015).

Most recently developed eye-tracking corpora of text reading, however, include data from mature and skilled readers only. The *EyeReadIt* corpus (`https://osf.io/hx2sj/`) represents, to our knowledge, the only recent collection of eye-tracking data including both learner and adult readers of Italian multi-line texts. The ReadLet dataset significantly complements EyeReadIt data by providing the first oral, visual and tactile repository of children and adults' data of *natural* text reading, with texts annotated at multiple levels of linguistic information. As recorded readers were requested to answer a few comprehension questions after reading, ReadLet provides evidence of both online and offline multimodal processing during task execution (Libben et al., 2021).

3. Data Acquisition

3.1. Participants and Protocol

The ReadLet dataset includes silent and oral reading data from learner and mature readers. Participants are i) primary school pupils (from 2 to 5 graders, 50 female and 44 male, mean age = 9, age range = 7-12), and ii) young adults (28 female, 27 male, mean age = 27, age range = 18-39). Children's reading data were collected in two primary schools in Pisa city and province, with a remarkably different socio-economic status. In each school, whole classes, from level 2 to 5, were sampled opportunistically, including pupils with special educational needs, mainly because our data collection campaign (from Winter 2020 to Spring 2021) extensively overlapped with the Covid 19 pandemic, and schools were granting only very limited access to their buildings. Likewise, young adults were recruited internally with leaflets and word of mouth in the Pisa CNR campus and the Sissa campus in Trieste. They were mostly post-graduate and post-doc students and grantees, and received no compensation for their participation in the reading sessions.

Each participant was involved in four experimental tasks: eye-tracked silent reading, eye-tracked oral reading, finger-tracked silent reading and finger-tracked oral reading. For technical reasons, it was not possible to concurrently eye-track and finger-track a reading session.[1] Adult readers were asked to complete the entire protocol in one go. Children conducted the eye-tracked and finger-tracked tasks in two separate sessions, at least one day apart. For each experimental task, participants were asked to read a multi-page, multi-episode text. Upon reading each text episode, each participant was asked to answer two questions of reading comprehension. Each question consisted of a question stem (i.e. the question proper), one correct answer, and three distractors (or incorrect options). Given the different number of episodes read by pupils of different grade levels (see section 3.2), the number of questions ranged accordingly from a minimum of 4 (2^{nd} graders) to a maximum of 10 questions (5^{th} graders)

Order of delivery of the tracking method and reading condition were counterbalanced across participants. Also the presentation of the different reading materials alternated among participants, for them to be equally distributed across experimental conditions. In oral reading sessions, participants were wearing a pair of wireless noise-cancelling headphones with a retractable microphone.

Finger movements were recorded using a tablet in portrait orientation as a reading book. The tablet screen was 14.9cm x 24.5cm, with a resolution 1920 x 1200 pixels. Participants were seated at a distance of approximately 50 cm from the tablet screen, with the tablet being tilted on a lectern at a 45° angle. Finger movements were sampled at a 120Hz rate, approximately corresponding to 24 touch events per syllable when a written word is read at a speed of 5 syllables per second.

Eye movements were recorded with an Eyelink

[1] In a typical experimental setting with an eye-tracker, movements of a finger sliding across the touchscreen of a tablet can block the infrared beams of an eye-tracker's camera positioned below the screen. This prevents concurrent recording of finger and eye movements. We are currently experimenting different technological solutions to address these technical issues.

Portable Duo eyetracker (SR Research, Canada), allowing for head-free eye-tracking with a reported accuracy of 0.25° to 0.50° degrees. Only the right eye of each participant was tracked at a 500 Hz sampling rate. A 9-point calibration was performed at the beginning of each reading session until the average error was below a 0.5° visual angle. Drift correction was performed after each text episode. No chin-rest was used during the experiment in either reading mode. Each participant was seated at about 80 cm from a 24" PC screen, displaying the same text layout used for the tablet at a resolution of 1920x1080. The letter font size was adapted to make the angle required to frame a single letter on a computer screen as close as possible to the angle required for the tablet. Stimulus presentation and eye movements recording were handled with Matlab Psychtoolbox (Brainard, 1997).

3.2. Text Materials

Four child fantasy stories were specifically written for the purposes of ReadLet data collection. A single child story consists of five self-contained episodes, with each episode being linguistically more complex than the previous one.[2] Second graders were asked to read the first two episodes only; third graders read the first three episodes, forth graders the first four episodes and fifth graders the whole story.

Each adult reading text included an excerpt from a Roberto Saviano's tabloid news article, and an excerpt extracted from Lamberto Maffei's popular neuroscience book *Elogio della parola* ('In praise of words', Maffei, 2018). All adult subjects were asked to read both excerpts. Child and adult reading texts were morpho-syntactically annotated (Dell'Orletta et al., 2007), syntactically chunked (Federici et al., 1996; Lenci et al., 2003), and annotated for functional dependency links (Attardi, 2006). Table 1 summarises annotation statistics for all ReadLet texts.

4. The Data

4.1. Raw data

To ensure readers' anonymity in compliance with data protection requirements, the original audio-recordings of oral reading sessions are not made openly accessible. They can be requested for research purposes from the authors' lab through the local data protection officer. Nonetheless, for each word read aloud, we provide open-access information about the onset and offset time of the word's articulation, as computed by a speech-to-text conversion tool (4.2.3). Raw eye-tracking data include the onset and offset time of each fixation event extracted from gaze records using the DataViewer software by SR Research, together with its position coordinates in the screen coordinate space. Likewise, finger tracking records are discretized into *touchmove* events on the surface of the tablet touchscreen, where each event associated with its time onset and the event's position coordinates on the screen. Details of the structure of the ReadLet database can be found in Appendix A.

4.2. Preprocessing and Cleaning

4.2.1. Eye-tracking

Areas of interest were automatically defined as the words' bounding boxes, i.e. the rectangular shapes surrounding each individual word making up the text displayed on the screen. Fixation-to-text alignment was performed using "Warp" (Carr et al., 2021), at present the most reliable algorithm for multi-line reading alignment, with a reported accuracy of 97.9% for mature readers and 97.1% for early readers. Extremely short (50ms for adults, 80ms for children) and long (800ms and 1200ms) fixations were taken out, resulting in the exclusion of 2% and 3.5% of adult and children data respectively. Out of all the recorded sessions, we selected only those with a page coverage of at least 70% of words being associated with at least one fixation. This led to the removal of 16% of data from adult readers, and 12% of data from child readers. The resulting database includes 369 pages, 54635 fixations on 37180 word tokens for adult readers, and 585 pages, 115196 fixations on 61761 word tokens for child readers. Eye movements were finally distilled into standard eye tracking metrics measured either in space (saccadic pattern) or time (fixation duration).

General descriptive statistics for eye and finger movements of adults and children are reported in Table 2. In the table, "first fixation duration" refers to the duration of the first fixation landing on a word, "first-pass duration" measures the amount of time from the onset of a word's first fixation to the onset of the first saccade leaving the word, "total fixation duration" is the sum of the duration of all fixations on a word, including refixations. The table also reports the mean length of a (forward or backward) saccade, the probability for a single word to be skipped, fixated once and fixated more than once. Most of these measures are inherent to the specific nature of the eye-tracking signal and have no equivalent in finger-tracking data (5.2).

[2] Readability levels were automatically computed and controlled using *ReadIt* (Dell'Orletta et al., 2011), a battery of annotation and classification tools scoring Italian texts for levels of reading difficulty.

[3] Word frequencies are extracted from SUBTLEX-IT (Crepaldi et al., 2013)

[4] FT coverage is the ratio between the number of tracked letters and the overall number of letters in a

Table 1: Statistics for adults and children's texts by text type (IPU = Implicit Prosodic Unit).

Adults	all		Saviano		Maffei	
	Mean	SD	Mean	SD	Mean	SD
word length [letters]	5.17	3.11	4.89	2.95	5.52	3.26
text length [words]	278.75	37.99	308.5	12.79	249	26.49
chunk per sentence	6.28	3.86	6.28	3.86	9.44	5.71
PoS type	11.00	0.76	11.50	0.58	10.50	0.58
IPU length [words]	7.11	0.67	6.71	1.43	8.12	0.92
sentence length [words]	26.99	18.63	20.22	10.98	47.00	22.20
dependency length [words]	2.44	3.83	2.18	2.54	2.76	4.93
word log frequency[3]	4.32	1.66	4.40	1.77	4.22	1.64

Children	grade 2		grade 3		grade 4		grade 5	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
word length [letters]	4.03	2.46	4.12	2.54	4.21	2.63	4.29	2.75
text length [words]	293.00	1.41	459.00	4.69	628.75	6.99	806.75	11.90
chunk per sentence	6.85	0.12	7.66	0.27	8.53	0.20	9.23	0.16
PoS type	11.25	1.50	12.00	0.00	12.25	0.50	12.25	0.50
IPU length [words]	7.35	0.53	7.45	0.91	7.73	0.65	7.64	0.35
sentence length [words]	13.06	4.97	14.68	5.80	16.33	6.94	17.89	8.24
dependency length [words]	0.54	2.38	0.63	2.54	0.69	2.72	0.70	3.00
word log frequency	4.92	1.50	4.87	1.53	4.83	1.55	4.81	1.56

4.2.2. Finger-tracking

Text-to-finger alignment was computed using a custom convolutional algorithm finding the closest pattern match between text lines and touch event sequences (see Appendix A). For each continuous time series of touch events falling within a letter bounding box, tracking time was computed as the difference between the last time tick and the first time tick in the series of touch events. Finally, the finger-tracking time for all other units in the text was defined as a summation of the tracking times of the letters each unit spans over.

4.2.3. Speech processing

Speech-to-text conversion of adults' recordings was carried out using Vosk (Shmyrev and Vosk Core Team, 2020), a free open-source toolkit built on Kaldi (Povey et al., 2011). For each word token, Vosk outputs its alphabetic transcription and the associated confidence level, together with onset and offset time-points of the word's articulation. We collected voice data for a total of 54,896 word tokens, out of which 50520 were correctly identified (94% of the data). All cases of word repetition (326 instances overall, $<$ 1% of the data) were dropped, leaving us with 50,194 correctly transcribed tokens. Of these, 23,177 came from ET sessions, and 27,017 from FT sessions.

text. FT tracking is the ratio between the number of letter trackings and the overall number of letters in a text. For each word in a text, when its tracking is larger than its coverage, we take the word to be re-tracked.

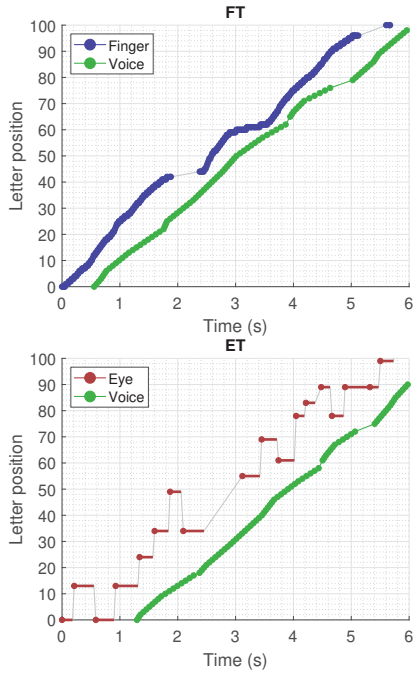

Figure 1: Finger-tracked (FT: top panel) and eye-tracked reading data (ET: bottom panel) recorded by two subjects reading the same sentence.

5. Data Validation

In this section, we present a few quantitative analyses of children and adults' reading data with a view to assessing their independent quality and reliability in both tracking modalities. We will then be concerned with cross-modal aspects of data analysis. At the time of writing this paper, chil-

Table 2: Eye-tracking and finger-tracking statistics for adult and children, in aloud and silent reading.

	Aloud				
	2nd Graders	3rd Graders	4th Graders	5th Graders	Adults
ET first fixation duration [ms]	425 (251)	365 (224)	322 (184)	307 (172)	280 (133)
ET first pass duration [ms]	596 (443)	506 (403)	425 (306)	413 (325)	349 (198)
ET total reading time [ms]	808 (579)	707 (523)	559 (402)	548 (426)	411 (227)
ET forward saccade length [letters]	5.62 (5.12)	5.93 (5.12)	6.35 (4.88)	6.31 (4.80)	7.82 (4.22)
ET backward saccade length [letters]	7.56 (8.4)	6.81 (8.07)	7.02 (8.26)	7.19 (8.31)	5.99 (5.55)
ET word skipping probability	0.16 (0.37)	0.14 (0.35)	0.15 (0.36)	0.15 (0.36)	0.25 (0.43)
ET single fixation probability	0.84 (0.37)	0.86 (0.35)	0.85 (0.36)	0.85 (0.36)	0.75 (0.43)
ET refixation probability	0.60 (0.49)	0.60 (0.49)	0.58 (0.49)	0.59 (0.49)	0.22 (0.42)
ET regression probability	0.24 (.43)	0.26 (0.44)	0.24 (0.43)	0.23 (0.42)	0.21 (0.41)
ET fix per 100 words	163 (32)	171 (33)	150 (21)	152 (28)	117 (15)
ET word fix per minute	61 (21)	77 (19)	96 (24)	98 (21)	158 (18)
FT tracking time [ms]	655 (716)	515 (569)	371 (335)	347 (341)	279 (230)
FT tracking speed [syllables/sec]	2.31 (0.91)	2.81 (0.88)	3.95 (0.99)	4.14 (0.92)	5.71 (0.78)
FT coverage[4]	0.97 (0.10)	0.97 (0.12)	0.97 (0.10)	0.97 (0.12)	0.97 (0.13)
FT tracking [4]	1.06 (0.31)	1.04 (0.29)	1.00 (0.20)	0.99 (0.19)	0.99 (0.13)
FT regression prob. (tracking > coverage) [4]	0.14 (0.35)	0.13 (0.34)	0.05 (0.23)	0.05 (0.22)	0.01 (0.12)
FT tracked words every 100 words	97 (4)	98 (3)	98 (3)	98 (2)	99 (1)
FT tracked word per minute	73 (29)	87 (28)	119 (30)	124 (29)	154 (20)
	Silent				
	2nd Graders	3rd Graders	4th Graders	5th Graders	Adults
ET first fixation duration [ms]	429 (252)	338 (204)	312 (178)	288 (160)	242 (105)
ET first pass duration [ms]	603 (477)	459 (352)	406 (290)	378 (282)	278 (146)
ET total reading time [ms]	824 (668)	685 (536)	526 (399)	520 (415)	325 (199)
ET forward saccade length [letters]	5.30 (4.41)	6.08 (5.18)	6.35 (4.24)	6.52 (4.70)	9.01 (4.51)
ET backward saccade length [letters]	7.16 (7.64)	7.45 (8.64)	6.55 (7.04)	7.33 (8.38)	7.33 (7.32)
ET word skipping probability	0.17 (0.38)	0.14 (0.35)	0.15 (0.36)	0.16 (0.37)	0.30 (0.46)
ET single fixation probability	0.83 (0.38)	0.86 (0.35)	0.85 (0.36)	0.84 (0.37)	0.70 (0.46)
ET refixation probability	0.60 (0.49)	0.59 (0.49)	0.58 (0.49)	0.58 (0.49)	0.15 (0.35)
ET regression probability	0.24 (0.43)	0.27 (0.44)	0.24 (0.43)	0.24 (0.43)	0.19 (0.39)
ET fix per 100 words	160 (40)	174 (31)	143 (22)	149 (32)	96 (17)
ET word fix per minute	63 (26)	80 (21)	105 (30)	100 (25)	223 (54)
FT tracking time [ms]	621 (715)	484 (501)	325 (302)	322 (307)	226 (202)
FT tracking speed [syllables/sec]	2.47 (1.00)	3.14 (1.13)	4.62 (1.49)	4.63 (1.42)	7.14 (1.75)
FT coverage[4]	0.97 (0.12)	0.97 (0.11)	0.97 (0.11)	0.97 (0.11)	0.96 (0.13)
FT tracking[4]	1.01 (0.22)	1.03 (0.41)	0.98 (0.15)	0.98 (0.16)	0.99 (0.16)
FT regression prob. (track > coverage)[4]	0.08 (0.27)	0.08 (0.27)	0.03 (0.16)	0.03 (0.16)	0.02 (0.14)
FT tracked words every 100 words	97 (4)	98 (3)	98 (2)	98 (2)	99 (1)
FT tracked word per minute	79 (32)	97 (36)	140 (46)	139 (44)	193 (48)

dren's reading data are being post-processed for speech-to-text conversion, and only adults' reading data have been fully analysed multi-modally. The present cross-modal analyses will thus exclusively focus on adults' data.

5.1. Classical reading effects

We managed to replicate robust effects of word length and frequency on eye fixation and finger tracking duration in both adults and children's reading data. Most notably, modelling finger-tracking times across age ranges of early readers confirmed subtle developmental trends (Marinelli et al., 2013; Marzi et al., 2020). To illustrate, longer words elicit longer eye-fixations and finger-tracking duration, with the effect being larger for younger, typically developing readers than older readers (Fig. 2). Here, for increasing grade levels, slopes are significantly less steep (p-values < 0.001, in both experimental modalities: see detailed model coefficients in Appendix B, Tables 10 and 11). Likewise, the use of sublexical information and serial *n*-gram decoding appears to play a more prominent role in younger readers, suggesting a shortage of fully specified orthographic representations for longer and rarer words in the mental lexicon of less skilled or less mature readers (Zoccolotti et al., 2009).

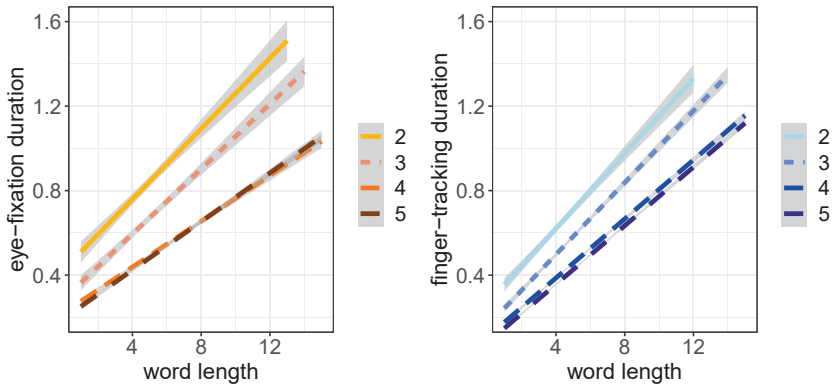

Figure 2: Linear fitting of eye-fixation and finger-tracking times by word-length and grade levels, in aloud reading.

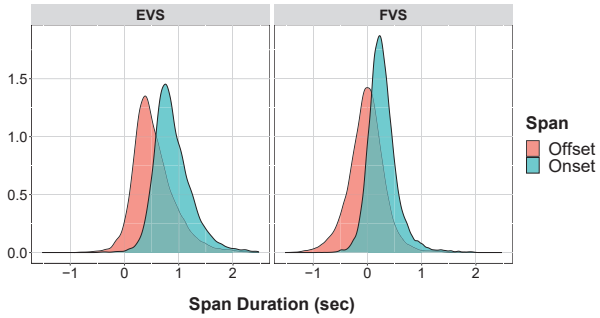

Figure 3: Distribution of offset/onset EVS/FVS in adults' reading (seconds).

5.2. Cross-modal evidence

Fig. 1 pictures a few seconds of the time-bound dynamics of eye and finger movements vs. articulation in oral reading. The acoustic and tactile recordings of a reading session are continuous in time. Finger movements are also continuous in text space, as they tend to fully cover text letters, punctuation marks and blank spaces, with a limited number of orthographic text units being skipped. In the plots, both the eye and the finger start "scanning" the text ahead before voice articulation sets in. The voice delay – known in the literature as Eye-Voice Span (EVS) or Finger-Voice Span (FVS) – is varyingly modulated across the text, and occasionally reversed into negative values, e.g. when the voice is reading out a word at position *n* in the text, while the finger is pointing to (or the eye is fixating) a word at position *n-k*. Fig. 3 depicts the distribution of EVS and FVS, calculated from the onset (cyan bell) and the offset (red bell) of word articulation. Notably the peak of the offset-FVS bell is centred on 0, meaning that, most of the times, the finger leaves a word w_h at the exact moment in time the voice completes the articulation of w_h.

A finger's sliding movement is typically broken at the end of each text line, where the reader lifts her index finger from the screen to shift it backwards across the current line and land it on the beginning of the ensuing line.

This dynamic is in sharp contrast with the series of fixations that are typically made on single words by the eye. Eye fixations are interspersed with relatively instantaneous "jumps" (saccades) whereby the eye leaves the currently fixated word to land on a different word. Notably, the landing site of an eye's saccade may not be the immediately ensuing word in the text line, but a word further away, either in the reading direction or backwards.

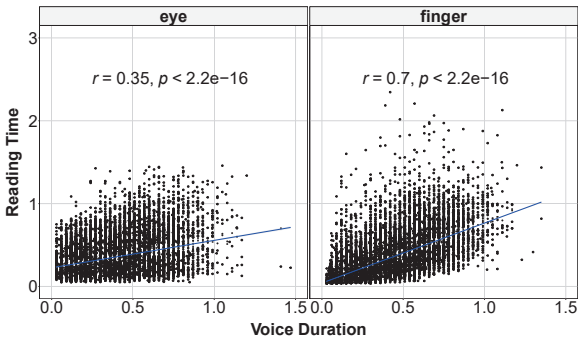

Figure 4: Correlation between speech duration and first-pass duration (left) and finger tracking time (right) across adult readers.

The different dynamic between the "reading" finger and the "reading" eye makes the two time series of movements asynchronous. How far ahead the eye goes is a function of several factors, including the reader's articulatory rate and phonological working memory, the length and frequency of a fixated word, the large meaningful syntactic units

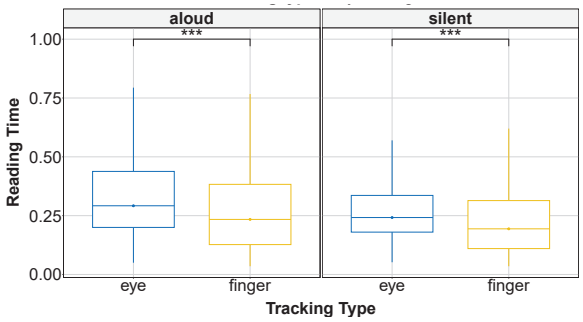

Figure 5: Eye (blue) and finger's (yellow) word tracking times in aloud and silent reading.

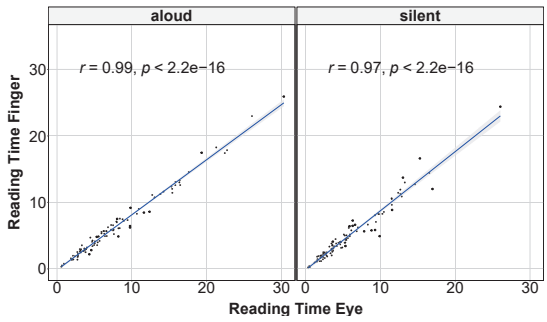

Figure 7: Scatter plot of adult sentence reading times as measured by eye- and finger-tracking.

where words occur (Laubrock and Kliegl, 2015; Silva et al., 2016; Nadalini et al., 2024).

Although the finger somehow "chases" after the eye, adult readers tend to keep their finger as temporally and spatially close as possible to the currently articulated word (Nadalini et al., 2024). Such a self-monitoring process accounts for the robust correlation between finger-tracking times and word articulation times (Fig. 4), causing the finger's pace to slow down when the span between the finger and the voice is longer than one word or two. This is similar to what has been extensively observed for the eye-voice span (Inhoff et al., 2011; Laubrock and Kliegl, 2015), as shown by the bell-shaped distributions of Figure 3. Conversely, the asynchrony between eye and finger movements accounts for the apparently paradoxical observation that word fixations are, on average, longer than word finger-tracking times (Fig. 5). The token-level gap between eye-tracking and finger-tracking times is nonetheless made up for at the sentence level, where time differences virtually disappear (Figure 6) and their correlation (Pearson *r*) approaches 1 in both modalities (Figure 7).

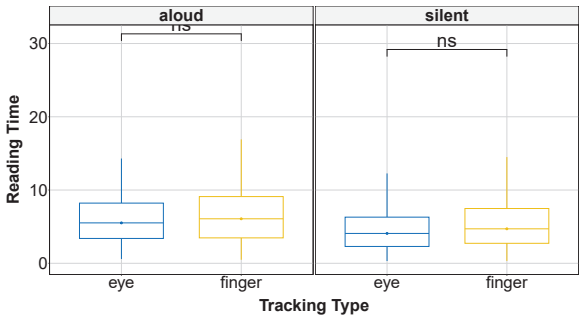

Figure 6: Eye (blue) and finger's (yellow) sentence tracking times in aloud and silent reading.

The evidence suggests that, differences in local dynamic notwithstanding, both tracking modalities reflect the same high level processing constraints, shedding light on a unique underlying dynamic, and possibly complementing each other at low processing levels. In fact, although the two time series happen to be more weakly (but nonetheless significantly) correlated at the token level, considerable information about which words are fixated in eye reading can be gained from the observation of alternating speed-up and slow-down patterns of finger movements (Nadalini et al., 2022). Most notably, an increase in finger-tracking time is shown to correlate with decreasing word skipping probabilities when we control for word length (Figure 8).

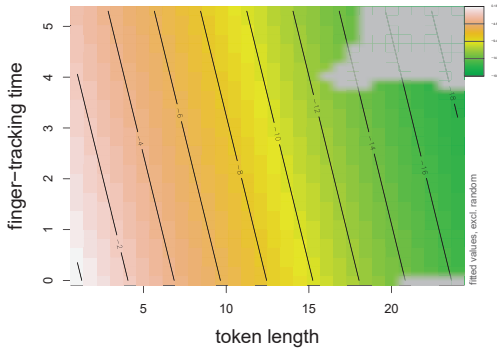

Figure 8: Contour plot of effects of word length and finger- tracking time on word skipping log odds (from white = high to green = low: model details in Appendix B).

6. Discussion and Outlook

Of late, in both linguistic and cognitive domains there has been a growing interest in the potential of research on cross-modal interaction (Lio et al., 2019), with a view to interdisciplinary synergy. Reading research has considerably benefited from this convergence (De Luca et al., 2013; Inhoff et al., 2011; Laubrock and Kliegl, 2015; Silva et al., 2016) and recent, dramatic advances in digital technologies have played a fundamental role in fostering the process. The ReadLet dataset is a further step in this direction, paving the way to

cross-modal data collection at scale for reading research and education.

There are several reasons to recommend cross-modal data collection for reading research. As recently emphasised by Libben et al. (2021) "[...] a key challenge in the design of psycholinguistic research on lexical processing is to create experiments that have ecological validity and at the same time are sufficiently controlled so that specific variables and hypotheses regarding their effects can be examined." A reading task should thus be evaluated according to two dominant parameters: i) whether the task allows investigators to collect evidence of online processing ease/difficulty for the reader, and ii) whether the reading task is modelled in a "natural" way. The ReadLet dataset addresses both concerns. The use of a friendly and widely accepted electronic device such as a tablet as a reading book allows investigation of connected text reading in highly ecological conditions, combining the benefits of large naturalistic data collection, with technological portability, accuracy and lack of intrusiveness.

A further bonus of having multiply time-aligned multi-modal data streams is that, in processing raw data, noise in one channel can be reduced by integrating synchronous information coming from a less noisy channel. For example, the vertical drift of an eye-tracking signal in a particular time window, can be spotted and corrected by using the voice signal sampled and text-aligned in the same time window, as the latter can provide reliable information about which text line the reader is currently processing. This is expected to offer better eye-tracking, finger-tracking and spoken data, which can be aligned more reliably both individually with the text being read, and with each other.

In principle, this also speaks in favour of integrating more data streams through the ReadLet protocol, for example by adding EEG data recording and yet other, potentially very informative biological signals. Although nothing prevents from pursuing this line of development in more controlled experimental settings, it should be appreciated that ReadLet was originally designed and intended as a non-intrusive, ecological, and ubiquitous protocol for the collection of large data repositories. For this reason, we are currently more interested in exploring a trade-off between data quantity and the ecological validity of our experimental tasks. In fact, the great potential of mobile information technology and cloud computing for huge data collection and analysis makes finger-tracking especially suitable for extensive reading assessment activities in primary schools. The ReadLet computing framework supports highly parallel and distributed processes of data acquisition, which can be delivered in real time to research, clinical and education centres as terminals for data modelling and quantitative analysis. Large-scale studies can be conducted, paving the way to more generalisable results than ever in the past. In addition, the possibility to take single-subject measurements on more occasions and in different settings makes finger-tracking evidence suitable not only for group studies, but also for individual diagnostic purposes and large developmental studies.

At present, the ReadLet database includes Italian reading data only. Preliminary experiments have been conducted with reading Modern Standard Arabic, English, French, German, Modern Greek, and Hebrew, with yet other languages (e.g. Bulgarian) and scripts (Cyrillic) being currently experimented with. In fact, the portability and language-independent nature of our technology for data collection, makes the ReadLet database easily scalable multilingually, paving the way to large-scale screening of children's reading skills across languages. In the end, we believe this technology to have the potential to define a converging perspective between cognitive (Pollatsek and Treiman, 2015), computational (Reichle, 2021) and educational (Grabe and Stoller, 2019) approaches to reading research, both within and across modalities (e.g. sight reading vs. Braille reading).

7. Acknowledgements and Credits

We gratefully acknowledge the *ReadLet* Italian National Strategic Research Grant (PRIN 2017W8HFRX: "Reading to understand: an ICT driven, large-scale investigation of early grade children's reading strategies" (2019-23) from the Ministry of University and Research, and the *ReadGround* grant, from the Italian National Research Council program "Supportive technologies assisting frail people" (2021-24). Experimental protocols were designed and conducted in compliance with the ethical principles for research involving human subjects stated in the Helsinki Declaration and were formally approved by the CNR Committee for Research Ethics and Deontology (Ethical Clearance Statement no. 0037523/2021). Authors' roles: Conceptualization & Methodology (AN, CM, LT, MF, VP); Validation, Formal Analysis & Visualization (AN, CM, MF); Data collection (AN, LT); Resources & Data curation: (AL, AN, LT, MF); Software (MF); Original draft preparation (VP); Draft review and editing: (CM, MF, VP); Supervision: (CM, MF, VP); Funding acquisition: (CM, VP). Alessandro Lento is a PhD student enrolled in the *National PhD in Artificial Intelligence*, XXXVII cycle, course on Health and Life sciences, organized by Università Campus Bio-Medico of Rome.

Fully anonymised data are made available through the CLARIN infrastructure.

8. Bibliographical References

Giuseppe Attardi. 2006. Experiments with a multilanguage non-projective dependency parser. In *The 10th International Conference on Computational Natural Language Learning (CoNLL-X2006)*, pages 166–170, New York.

Edward Awh, Artem V Belopolsky, and Jan Theeuwes. 2012. Top-down versus bottom-up attentional control: A failed theoretical dichotomy. *Trends in cognitive sciences*, 16(8):437–443.

David H. Brainard. 1997. The psychophysics toolbox. *Spatial vision*, 10(4):433–436.

Jonathan Brennan. 2016. Naturalistic sentence comprehension in the brain. *Language and Linguistics Compass*, 10(7):299–313.

Timothy J Buschman and Earl K Miller. 2007. Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices. *Science*, 315(5820):1860–1862.

Guy Thomas Buswell. 1920. An experimental study of the eye-voice span in reading. *Journal of Educational Psychology*, 4(12):217–227.

Guy Thomas Buswell. 1921. The relationship between eye-perception and voice-response in reading. *Journal of Educational Psychology*, 12(4):217–227.

Jon W Carr, Valentina N Pescuma, Michele Furlan, Maria Ktori, and Davide Crepaldi. 2021. Algorithms for the automated correction of vertical drift in eye-tracking data. *Behavior Research Methods*, pages 1–24.

Max Coltheart, Kathleen Rastle, Conrad Perry, Robyn Langdon, and Johannes Ziegler. 2001. Drc: a dual route cascaded model of visual word recognition and reading aloud. *Psychological review*, 108(1):204.

Uschi Cop, Nicolas Dirix, Denis Drieghe, and Wouter Duyck. 2017. Presenting geco: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior research methods*, 49:602–615.

Melda Coskun, Victor Kuperman, and Jay Rueckl. in press. Long-lag repetition priming in natural text reading: No evidence for morphological effects.

Davide Crepaldi, Marcello Ferro, Claudia Marzi, Andrea Nadalini, Vito Pirrelli, and Loukia Taxitari. 2022. *Finger Movements and Eye Movements During Adults' Silent and Oral Reading*, pages 443–471. Springer International Publishing.

Davide Crepaldi, Emmanuel Keuleers, Pavel Mandera, and Michael Brysbaert. 2013. SUBTLEX-IT: A frequency list based on movie subtitles. (unpublished manuscript).

Maria De Luca, Maria Pontillo, Silvia Primativo, Donatella Spinelli, and Pierluigi Zoccolotti. 2013. The eye-voice lead during oral reading in developmental dyslexia. *Frontiers in human neuroscience*, 7:696.

Felice Dell'Orletta, Marcello Federico, Alessandro Lenci, Simonetta Montemagni, and Vito Pirrelli. 2007. Maximum Entropy for Italian PoS Tagging. *Intelligenza Artificiale*, 4(2).

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2011. READ–IT: Assessing readability of Italian texts with a view to text simplification. In *Proceedings of the second workshop on Speech and Language Processing for Assistive Technologies*, pages 73–83.

Vera Demberg and Frank Keller. 2019. Cognitive models of syntax and sentence processing. *Human language: From genes and brains to behavior*, pages 293–312.

Katia Dilkina, James L McClelland, and David C Plaut. 2010. Are there mental lexicons? the role of semantics in lexical decision. *Brain research*, 1365:66–81.

Ralf Engbert, Antje Nuthmann, Eike M Richter, and Reinhold Kliegl. 2005. Swift: a dynamical model of saccade generation during reading. *Psychological review*, 112(4):777.

Núria Esteve-Gibert and Pilar Prieto. 2014. Infants temporally coordinate gesture-speech combinations before they produce their first words. *Speech Communication*, 57:301–316.

Stefano Federici, Simonetta Montemagni, and Vito Pirrelli. 1996. Shallow Parsing and Text Chunking: a View on Underspecification in Syntax. *Proceedings of ESSLLI'96 Workshop on Robust Parsing*, pages 35–44.

Marcello Ferro, Claudia Cappa, Sara Giulivi, Claudia Marzi, Ouafae Nahli, Franco Alberto Cardillo, and Vito Pirrelli. 2018. Readlet: Reading for understanding. In *Proceedings of 5th IEEE Congress on Information Science & Technology (IEEE CiST'18)*, Marrakech, Morocco.

Markus Frey, Matthias Nau, and Christian F Doeller. 2021. Magnetic resonance-based eye tracking using deep neural networks. *Nature neuroscience*, 24(12):1772–1779.

William Grabe and Fredricka L Stoller. 2019. *Teaching and researching reading*, 3rd edition. Routledge.

Jonathan Grainger and Arthur M Jacobs. 1996. Orthographic processing in visual word recognition: a multiple read-out model. *Psychological review*, 103(3):518.

Anna Carin Gran Ekstrand, Mattias Nilsson Benfatto, and Gustaf Öqvist Seimyr. 2021. Screening for Reading Difficulties: Comparing Eye Tracking Outcomes to Neuropsychological Assessments. In *Frontiers in Education*, volume 6, page 643232. Frontiers Media SA.

Uri Hasson, Giovanna Egidi, Marco Marelli, and Roel M Willems. 2018. Grounding the neurobiology of language in first principles: The necessity of non-language-centric explanations for language comprehension. *Cognition*, 180:135–157.

Nora Hollenstein, Maria Barrett, and Marina Björnsdóttir. 2022. The copenhagen corpus of eye tracking recordings from natural reading of danish texts. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 1712–1720.

Barry Hughes, Amber McClelland, and Dion Henare. 2014. On the nonsmooth, nonconstant velocity of braille reading and reversals. *Scientific Studies of Reading*, 18(2):94–113.

Albrecht W Inhoff, Matthew Solomon, Ralph Radach, and Bradley A Seymour. 2011. Temporal dynamics of the eye–voice span and eye movement control during oral reading. *Journal of Cognitive Psychology*, 23(5):543–558.

Halszka Jarodzka and Saskia Brand-Gruwel. 2017. Tracking the reading eye: Towards a model of real-world reading.

Johanna K Kaakinen and Jukka Hyönä. 2008. Perspective-driven text comprehension. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 22(3):319–334.

Juan E Kamienkowski, M Julia Carbajal, Bruno Bianchi, Mariano Sigman, and Diego E Shalom. 2016. Cumulative repetition effects across multiple readings of a word: Evidence from eye movements. *Discourse Processes*, 55(3):256–271.

Victor Kuperman, Denis Drieghe, Emmanuel Keuleers, and Marc Brysbaert. 2013. How strongly do word reading times and lexical decision times correlate? combining data from eye movement corpora and megastudies. *Quarterly Journal of Experimental Psychology*, 66(3):563–580.

Victor Kuperman, Noam Siegelman, Sascha Schroeder, Cengiz Acartürk, Svetlana Alexeeva, Simona Amenta, Raymond Bertram, Rolando Bonandrini, Marc Brysbaert, Daria Chernova, et al. 2023. Text reading in English as a second language: Evidence from the Multilingual Eye-Movements Corpus. *Studies in Second Language Acquisition*, 45(1):3–37.

Jochen Laubrock and Reinhold Kliegl. 2015. The eye-voice span during reading aloud. *Frontiers in psychology*, 6:1432.

Everdina A Lawson. 1961. A note on the influence of different orders of approximation to the english language upon eye-voice span. *Quarterly Journal of Experimental Psychology*, 13(1):53–55.

Alessandro Lenci, Simonetta Montemagni, and Vito Pirrelli. 2003. “Chunk-it”. An Italian Shallow Parser for Robust Syntactic Annotation. *Linguistica Computazionale*, XVIII-XIX:353–386.

Harry Levin and Julie A Cohn. 1968. Effects of instruction on the eye-voice span. In Harry Levin, Eleanor J. Gibson, and Jack J. Gibson, editors, *The Analysis of Reading Skills: A Program of Basic and Applied Research. Final Report*, pages 254–283. Cornell University, Ithaca, New York.

Harry Levin and Elizabeth Ann Turner. 1968. Sentence structure and the eye-voice span. In Harry Levin, Eleanor J. Gibson, and Jack J. Gibson, editors, *The Analysis of Reading Skills: A Program of Basic and Applied Research. Final Report*, pages 196–220. Cornell University, Ithaca, New York.

Gary Libben, Jordan Gallant, and Wolfgang U. Dressler. 2021. Textual effects in compound processing: A window on words in the world. *Frontiers in Communication*, 6:42.

Guillaume Lio, Roberta Fadda, Giuseppe Doneddu, Jean-René Duhamel, and Angela Sirigu. 2019. Digit-tracking as a new tactile interface for visual perception analysis. *Nature Communications*, 10(5392):1–13.

Steven G Luke and Kiel Christianson. 2018. The Provo Corpus: A large eye-tracking corpus with predictability norms. *Behavior research methods*, 50:826–833.

Lamberto Maffei. 2018. *Elogio della parola*. il Mulino, Bologna.

Chiara Valeria Marinelli, Daniela Traficante, Pierluigi Zoccolotti, and Cristina Burani. 2013. Orthographic neighborhood-size effects on the reading aloud of italian children with and without dyslexia. *Scientific Studies of Reading*, 17(5):333–349.

Claudia Marzi, Anna Rodella, Andrea Nadalini, Loukia Taxitari, and Vito Pirrelli. 2020. Does finger-tracking point to child reading strategies? In *Proceedings of 7th Italian Conference on Computational Linguistics*, volume 2769, Bologna.

Heidi Anne E. Mesmer and Karen Lake. 2010. The role of syllable awareness and syllable-controlled text in the development of finger-point reading. *Reading Psychology*, 31(2):176–201.

Heidi Anne E Mesmer and Thomas O Williams. 2015. Examining the role of syllable awareness in a model of concept of word: Findings from preschoolers. *Reading Research Quarterly*, 50(4):483–497.

John Morton. 1964a. The effects of context upon speed of reading, eye movements and eye-voice span. *Quarterly Journal of Experimental Psychology*, 16(4):340–354.

John Morton. 1964b. A model for continuous language behaviour. *Language and Speech*, 7(1):40–70.

Andrea Nadalini, Marcello Ferro, Alessandro Lento, Vito Pirrelli, and Claudia Marzi. 2022. Evidence for saccadic reading dynamic with finger-tracking speed rates. Canada. The Mental Lexicon Conference.

Andrea Nadalini, Claudia Marzi, Marcello Ferro, Loukia Taxitari, Alessandro Lento, Davide Crepaldi, and Vito Pirrelli. 2024. Eye-voice and finger-voice spans in adults' oral reading of connected texts. implications for reading research and assessment. *The Mental Lexicon*.

Mattias Nilsson Benfatto, Gustaf Öqvist Seimyr, Jan Ygge, Tony Pansell, Agneta Rydberg, and Christer Jacobson. 2016. Screening for dyslexia using eye tracking during reading. *PloS one*, 11(12):e0165508.

Tetsushi Nonaka, Kiyohide Ito, and Thomas A Stoffregen. 2021. Structure of variability in scanning movement predicts braille reading performance in children. *Scientific reports*, 11(1):1–12.

Alexander Pollatsek and Rebecca Treiman. 2015. *The Oxford handbook of reading*. Oxford University Press.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. 2011. The Kaldi Speech Recognition Toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society. IEEE Catalog No.: CFP11SRW-USB.

A Jothi Prabha and Renta Bhargavi. 2020. Predictive model for dyslexia from fixations and saccadic eye movement events. *Computer Methods and Programs in Biomedicine*, 195:105538.

Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3):372.

Keith Rayner. 2009. The 35th Sir Frederick Bartlett Lecture: Eye movements and attention in reading, scene perception, and visual search. *Quarterly journal of experimental psychology*, 62(8):1457–1506.

Erik D Reichle. 2021. *Computational models of reading: A handbook*. Oxford University Press.

Luz Rello and Miguel Ballesteros. 2015. Detecting readers with dyslexia using machine learning with eye tracking measures. In *Proceedings of the 12th International Web for All Conference*, pages 1–8.

Naoyuki Sato and Hiroaki Mizuhara. 2018. Successful encoding during natural reading is associated with fixation-related potentials and large-scale network deactivation. *eNeuro*, 5(5).

Nickolay V. Shmyrev and Vosk Core Team. 2020. Vosk Speech Recognition Toolkit: Offline speech recognition API for Android, iOS, Raspberry Pi and servers with Python, Java, C# and Node. `https://github.com/alphacep/vosk-api`.

Susana Silva, Alexandra Reis, Luís Casaca, Karl M. Petersson, and Luís Faísca. 2016. When the Eyes no longer lead: Familiarity and Length Effects on Eye-Voice Span. *Frontiers in Psychology*, 7.

Loukia Taxitari, Claudia Cappa, Marcello Ferro, Claudia Marzi, Andrea Nadalini, and Vito Pirrelli. 2021. Using mobile technology for reading assessment. In *Proceedings of 6th IEEE Congress on Information Science & Technology (IEEE CiST'20)*, Agadir, Morocco.

Débora Torres, Wagner R Sena, Humberto A Carmona, André A Moreira, Hernán A Makse, and José S Andrade Jr. 2021. Eye-tracking as a

proxy for coherence and complexity of texts. *PloS one*, 16(12):e0260236.

Joanna K Uhry. 1999. Invented spelling in kindergarten: The relationship with finger-point reading. *Reading and Writing*, 11:441–464.

Joanna K. Uhry. 2002. Finger-point reading in kindergarten: The role of phonemic awareness, one-to-one correspondence, and rapid serial naming. *Scientific Studies of Reading*, 6(4):319–342.

Ludo Verhoeven and Charles Perfetti. 2008. Advances in text comprehension: Model, process and development. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 22(3):293–301.

Sebastian Wallot, Geoff Hollis, and Marieke van Rooij. 2013. Connected text reading and differences in text reading fluency in adult readers. *PloS one*, 8(8):e71914.

Roel M Willems, editor. 2015. *Cognitive neuroscience of natural language use*. Cambridge University Press.

Pierluigi Zoccolotti, Maria De Luca, Gloria Di Filippo, Anna Judica, and Marialuisa Martelli. 2009. Reading development in an orthographically regular language: Effects of length, frequency, lexicality and global processing ability. *Reading and Writing*, 22(9):965–992.

A. Appendix: Overview of Dataset Structure

A.1. Raw data

An abridged version of the table structure for raw tracking data is provided in Table 1, where only finger-tracking (FT) events are shown. Here, the type (*eventType*), timing (*timeOffset*) and duration measured in seconds (*dt*) of each touch event is associated with a reading session (*idSession*) and the event's *x* and *y* coordinates in the touchscreen space. The last two rows in the table contain information about the association of a single touch event with a specific letter box (*bid*) and its embedding word token (*tid*) after the alignment of tracking data with the reading text is computed. The same table structure is used to expose eye-tracking data.

Automated text-to-finger alignment is enforced through a convolutional algorithm that finds the nearest spatial match between text lines and touch event sequences. First, finger-tracking data along the time axis are projected onto a static image. After elimination of possible outliers, the image is shifted vertically (i.e. convoluted) onto the bounding-box image of the text, until a point is reached that maximises the overlap of the two images. This convolution operation is repeated across a search space that includes scaling and rotation of the finger-tracking image, to compensate for any spatial drift of the tracking signal relative to the positioning of the text on the page.

Reading sessions are grouped into data acquisition campaigns, where information about the protocol being administered (e.g. text material and layout options) is provided, as shown in Table 2.

Information about individual readers (*idUser*) and reference to reading sessions (*idSession*) are provided in Table 3 and 4 respectively, with the latter containing information about the text being read (*idDoc*), tracking modality (*idDevice*), and reading modality (*idReadingType*). The association between readers (*idUserData*) and campaigns (*idCampaign*) shown in Table 5 makes provision for longitudinal data being encoded.

Table 1: Column headings for FT data by **touch events**.

heading	type	gloss	example
idSession	*num*	reading session id	2263
eventType	*string*	type of touch event	touchstart
timeOffset	*num*	time of touch event	13.2396
dt	*num*	tracking duration	0.017
x	*num*	event *x*-coordinate	0.058263
y	*num*	event *y*-coordinate	0.311007
tid	*num*	word token position	7
bid	*num*	letter token position	39

Table 2: Column headings by **measurement campaigns**.

heading	type	gloss	example
idCampaign	*num*	campaign id	103
title	*string*	label	CNR - adults
options	*json*	default parameters	{...}

Table 3: Column headings by **subjects**.

heading	type	gloss	example
idUser	*num*	subject id	XXYY
idHand	*num*	subject's dominant hand	id_r
idGender	*num*	subject's gender	id_m

Table 4: Column headings by **reading sessions**.

heading	type	gloss	example
idSession	*num*	reading session id	2263
idCampaign	*num*	campaign id	103
idDevice	*num*	tracking modality	1 (finger)
idUser	*num*	subject id	XXYY
idDoc	*num*	text identifier	38
idReadingType	*num*	reading mode	id_silent
dt	*num*	total tracking time	160.778
questTime	*num*	time spent for questions	16.84
coverage	*num*	tracked letter coverage	0.945
tracks	*num*	tracked letter rate	0.948
questAcc	*num*	comprehension acc.	1.0

Table 5: Column headings by **user data**.

heading	type	gloss	example
idUserData	*num*	subject data id	15
idUser	*num*	subject id	XXYY
idCampaign	*num*	measurement campaign id	103
gradeLevel	*num*	subject's grade level	18
age	*string*	subject's age	28;5
notes	*string*	subject's notes	

A.2. Post-processed data

Table 6 epitomises the structure for post-processed finger-tracking data, viewed by word tokens and reading sessions. The tracking onset time (*t*) gives the time point when the tracking of a word starts (measured in seconds elapsed from the beginning of the reading session). The total tracking duration (*dt*) indexes the total amount of time a single word token was tracked, including possible regressive movements. A more time-bound view of word tokens' data can be extracted along the timeline, grouping raw data (Table 1) by consecutive touch events on individual word tokens.

Table 7 contains the output of the Vosk transcriber by word tokens. Here, information about the articulation onset (*timeOnset*) and offset (*timeOffset*) of each automatically recognised token is provided, together with an indication of the reading session (*idSession*), the associated token in the text (*tid*), and the transcriber's level of *confidence*, ranging from 0 (no confidence) to 1 (full confidence).

Table 6: Column headings by **word tokens**.

heading	type	gloss	example
idSession	*num*	reading session id	2263
tid	*num*	position of word token	7
token	*string*	word token string	*di*
lemma	*string*	word token lemma	di
dt	*num*	total tracking duration	0.367
t	*num*	tracking onset time	13.2396
len	*num*	word token length	2
freq	*num*	word token frequency	2202526
FPOS	*string*	word token part of speech	E

Table 7: Headings by **transcribed tokens**.

heading	type	gloss	example
idSession	num	reading session id	2004
tid	num	position of word token	7
token	string	word token string	*di*
timeOnset	num	articulation onset	3.4202
timeOffset	num	articulation offset	3.5402
confidence	num	transcription confidence	1

B. Appendix: Data Modelling

Logistic regression models fitting word skipping probabilities with finger-tracking time and word length (Table 8) and finger-tracking time and word frequency (Table 9) as predictors, and adult readers and word tokens entered as random effects.

Table 8: Logistic regression model coefficients: eye skipping $\sim$ finger-tracking time + word len, (adult users = re), (token = re).

	estimate	st. err	z value	p-value
intercept (aloud)	0.91	0.19	4.85	< 0.001
tracking time (aloud)	-0.54	0.04	-12.87	$< 2e-16$
silent reading	0.39	0.01	32.48	$< 2e-16$
tracking time (silent)	0.48	0.05	8.73	$< 2e-16$
word length	-0.72	0.03	-26.41	$< 2e-16$
random effects	$< 2e-16$		R^2	0.38

Table 9: Logistic regression model coefficients: eye skipping $\sim$ finger-tracking time + word log frequency, (adult users = re), (token = re).

	estimate	st. err	z value	p-value
intercept (aloud)	-9.98	0.35	-28.78	$< 2e-16$
tracking time (aloud)	-0.56	0.04	-13.25	$< 2e-16$
silent reading	0.39	0.01	32.51	$< 2e-16$
tracking time (silent)	0.47	0.05	8.57	$< 2e-16$
word log frequency	0.65	0.04	18.15	$< 2e-16$
random effects	$< 2e-16$		R^2	0.38

Generalised Additive Models (GAMs) fitting eye-fixation and finger-tracking times as a function of word length and grade level, with word tokens and children as random effects, are reported in Table 10 and Table 11.

Table 10: GAM coefficients: eye-fixation time $\sim$ word length $*$ grade levels, (child users = re), (token = re).

	estimate	st. err	z value	p-value
intercept (2nd grade)	0.45	0.06	7.50	< 0.001
word length (2nd grade)	0.09	0.01	26.73	$< 2e-16$
intercept (3rd grade)	-0.14	0.09	-1.65	> 0.05
word length (3rd grade)	-0.02	0.01	-2.63	< 0.05
intercept (4th grade)	-0.20	0.08	-2.42	< 0.05
word length (4th grade)	-0.04	0.01	-10.11	$< 2e-16$
intercept (5th grade)	-0.20	0.08	-2.63	< 0.01
word length (5th grade)	-0.04	0.01	-10.05	$< 2e-16$
random effects	< 0.001		R^2	0.36

Table 11: GAM coefficients: finger-tracking time $\sim$ word length $*$ grade levels, (child users = re), (token = re).

	estimate	st. err	z value	p-value
intercept (2nd grade)	0.30	0.07	4.19	< 0.001
word length (2nd grade)	0.10	0.01	42.06	$< 2e-16$
intercept (3rd grade)	-0.14	0.10	-1.43	> 0.05
word length (3rd grade)	-0.01	0.01	-4.80	< 0.001
intercept (4th grade)	-0.19	0.09	-2.08	< 0.05
word length (4th grade)	-0.03	0.01	-11.22	$< 2e-16$
intercept (5th grade)	-0.21	0.09	-2.29	< 0.05
word length (5th grade)	-0.03	0.01	-12.04	$< 2e-16$
random effects	$< 2e-16$		R^2	0.42