ORAL TEXT READING AS A MULTI-SENSORY TASK

CLAUDIA MARZI ANDREA NADALINI
ALESSANDRO LENTO MANU SRIVASTAVA
ALICE TODESCO VITO PIRRELLI MARCELLO FERRO

ABSTRACT: Reading aloud involves the complex interplay of visual, motor and lexical processes. While eye movements have been extensively investigated in the reading literature, less is known about the coordination of voice, eye and finger movements in oral and finger-point reading. Here we propose a multimodal perspective on these dynamics, emphasising the contribution of integrating eye-tracking, finger-tracking, and voice recording to a more comprehensive understanding of reading proficiency. Our results show that finger and eye movements are strongly coupled in early readers. Conversely, skilled readers show a more flexible coordination of sensorimotor signals and a more adaptive sensitivity to prosodic structures, with voice articulation slowing at key structural points, such as chunk heads and sentence-final boundaries. These findings provide novel insights into how multimodal coordination evolves with reading expertise, contributing to a more fine-grained understanding of reading fluency.

KEYWORDS: reading development, multimodal integration, eye-voice span, finger-voice span, adaptive reading.

1. INTRODUCTION[1]

At the most fundamental, behavioural level of inquiry, oral reading of a connected text requires the fine coordination of eye movements across a line of letter strings, and articulatory movements. The eye starts off the stage of letter decoding that is required for voice articulation to be planned and executed at a relatively constant rate. In turn, articulation provides feedback to oculomotor control for eye movements to be directed when and where processing issues arise in a written text. One specific factor that makes eye-voice coordination hard to manage is the *asynchronicity* of the two time series of movements (Inhoff *et al.* 2011). First, fixation of a written word trivially precedes its articulation.

[1] Authors' roles: conceptualisation & methodology: CM, AN, VP, MF; software: MF; data collection: AN; data annotation & alignment: AN, AL, AT, MS; formal analysis & visualization: CM, AN, MF; draft writing: CM, VP; review and editing: CM.

In addition, the dynamics of eye and voice movements differ substantially. Eye movements are faster than articulators' movements, and are much freer to scan a text forwards and backwards, availing themselves of a wide range of alternative "moves", including long forward saccades, regressions, refixations and word skippings. In contrast, in proficient reading, voice articulation is steady and seamlessly continuous, with very few repetitions or disfluencies, and only occasional voice breaks (pauses) across the boundaries of complex linguistic structures. For all these processes to be optimally coordinated, a reader must rely on a fine motor-control strategy, which requires the buffering of already decoded units into a reader's phonological buffer, integration of the buffered units into larger prosodic chunks, and articulatory planning and execution of the buffered score (De Luca *et al.* 2013; Inhoff *et al.* 2011; Laubrock & Kliegl 2015; Silva *et al.* 2016). In text reading, this is also accompanied by lexical access of word meanings and their online integration into the syntactic scaffolding of a text (Rayner *et al.* 2000; Hirotani *et al.* 2006; Warren *et al.* 2009; Tiffin-Richards & Schroeder 2018).

More recently, Lio *et al.* (2019) studied the connection between eye movements and finger movements in the visual exploration of a picture displayed on a touchscreen. Spatial patterns of finger movements were found to be congruent with patterns of eye fixations on the same image, confirming that tactile exploration of an image can be used as an ecological proxy of visual exploration. Using a simple tablet as a reading book, Nadalini *et al.* (2022) investigated the congruence of eye/finger movements and voice articulation in adults' *finger-point reading*, i.e. when reading is accompanied by the sliding movement of the index finger pointing to the written words being read.

Current developments in information and language technologies have made collections of multi-modal behavioural data in ecological conditions increasingly available. In addition, algorithms for data post-processing turn out to be more accurate if asynchronous time-series of signals are analysed concurrently. At the same time, non-linear statistical techniques can provide dynamic models of how several factors may interact in the execution of a complex multi-sensory task such as reading. We argue that a truly functional approach to reading must investigate this online multi-sensory dynamic, which ultimately rests on the human ability to decode, buffer, integrate and articulate written words forming complex sentences. Here, the potential of such a functional approach will be shown through a comparative analysis of the multi-sensory processes that unfold in the course of children and adults' reading. This way, one can not only investigate the complex range of skills required for reading and understanding a connected text, but can also elucidate how skills interact, and how their interaction may affect reading development and reading strategies.

2. BACKGROUND

Speakers must prepare the phonological units they are about to utter (be them syllables, words or multi-word units) ahead of their articulation (Romani *et al.* 2022). The incremental nature of this preparation phase raises the question of how speakers manage the optimal coordination of phonological planning and articulatory processes when two or more words have to be articulated one after the other without pausing. Scholars have tried to shed light on the *mechanisms* that underlie such a complex coordination (Gambi & Crocker 2017) by presenting subjects with two pictures side by side, and asking to name the pictures as quickly as possible, while controlling for the length of picture names (Griffin 2003). In executing the task, readers usually shift their gaze from the left picture to the right picture as soon as they have retrieved the whole phonological representation of the first word. In addition, they begin retrieving the articulatory code of the first syllable of a word as soon as they complete the phonological processing for this syllable (Figure 1). However, they typically do not start articulating the first word until their gaze shifts to the right picture. In fact, for an optimal phonological plan of two consecutive words to be made, the phonological representation of the second word must also be retrieved. According to Meyer *et al.* (2007), this dynamic explains the so-called *reversed word length effect* (Griffin 2003): the time lag between the gaze shift to the right picture and the articulation onset for the first word (or Eye-Voice Span, henceforth EVS) is observed to be *longer* when the first word is monosyllabic, and *shorter* when the first word is multisyllabic. In the former case, the span includes the entire time needed for retrieving the phonological code of the first word. In contrast, when the first word is multisyllabic, the gaze shift takes place when the phonological code of the first syllable of the first word is already available and its articulation can start (Figure 1).

An altogether different interpretation of the same effect is offered by Griffin (2003), according to whom a reader begins articulating the first word *later* when the word is shorter, in order to have time to complete the second word's articulatory coding before articulation of the first word ends. Delayed articulation is not needed for long words, simply because their articulation time is long enough for the articulatory coding of the second word to be completed in parallel. The same reversed word length effect has also been observed in reading consecutive words (Inhoff *et al.* 2011). Laubrock & Kliegl (2015) provide yet another explanation of the effect. They claim that a reader's ability to delay articulation of more words is constrained by the capacity of the reader's phonological buffer. Since long words take more buffer capacity than short words, the articulation of short words can be delayed for a longer time than the articulation of long words.

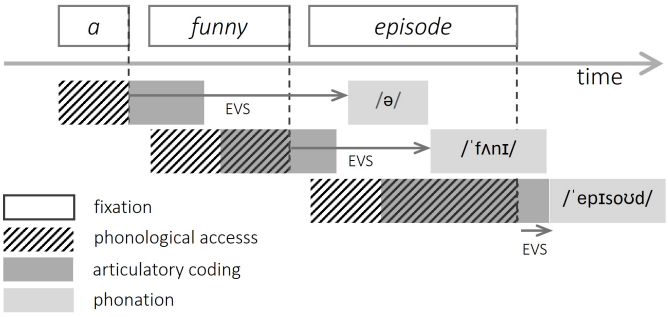

FIGURE 1: TEMPORAL EYE-VOICE SPAN (EVS) IN MULTI-WORD ORAL READING: 1) MONOSYLLABIC WORDS SHOW LONGER EVS THAN MULTISYLLABIC WORDS; 2) A FLUENT READER DELAYS ARTICULATION OF THE FIRST WORD IN A CHUNK (*a*) UNTIL (S)HE STARTS PROCESSING THE CHUNK FINAL WORD (*episode*).

All these accounts share the idea that the explanation of a task effect must be grounded in the need of executing the task as quickly as possible. When it comes to reading, however, functional efficiency does not necessarily amount to maximisation of speed. Optimal reading must fulfil the fundamental need of understanding a text. In oral reading, this is combined with the goal of articulating words with the right intonation and intervening pauses. Ideally, for a reader to be able to read a chunk like *a funny episode* with the right intonation and emphasis, all word tokens making up the chunk must be processed before articulation starts (Figure 1). Nadalini *et al.* (2024) show that this is indeed the case, based on evidence of adults' oral reading of connected texts. Readers tend to slow down their reading pace at the end of chunked multiword units, no matter how long these units are. According to this *adaptive reading hypothesis*, a reader's EVS is neither set by the maximum capacity of the phonological buffer (Laubrock & Kliegl 2015), nor maximised for the sake of reading speed (Silva *et al.* 2016). Rather, it appears to stretch or shrink far enough for the reader to be able to process a larger unit, articulate its words with an appropriate intonation contour, and grasp its meaning. In addition, finger-tracking records of adults' oral finger-point reading show that the finger is ahead of the voice, and that the finger-voice span (FVS) exhibits a similar dynamic as EVS (Nadalini *et al.* 2022).

3. THE READLET PROTOCOL

The *ReadLet* protocol was developed to record eye and finger movements while reading a connected text. It requires a subject to read a text displayed on a PC screen equipped with an eye-tracker, or, alternatively, on a tablet touchscreen.

Reading sessions can be silent or aloud.[2] Overall, each reader is engaged in four experimental tasks: eye-tracked silent reading, eye-tracked oral reading, finger-tracked silent reading and finger-tracked oral reading. Adult readers are asked to complete the entire protocol in one go. Children conduct the eye-tracked and finger-tracked tasks in two separate sessions, at least one day apart. For each experimental task, participants are asked to read a multi-page, multi-episode text, and answer a few questions of reading comprehension upon finishing each episode. Each question consists of a question stem (i.e. the question proper), one correct answer, and three distractors (or incorrect options). Four fantasy stories were specifically designed for child data collection, with each story comprising five self-contained episodes of increasing linguistic complexity. Children read 2 to 5 episodes depending on their grade level (from 2nd to 5th). Adult texts featured excerpts by Saviano and Maffei (2018). Child and adult reading texts were syntactically annotated and chunked. Table 1 summarises key linguistic features of both child and adult texts.

CHILD TEXTS	GRADE 2	GRADE 3	GRADE 4	GRADE 5
word length (letters)	4.03 (2.46)	4.12 (2.54)	4.21 (2.63)	4.29 (2.75)
text length (tokens)	293.0 (1.41)	459.0 (4.69)	628.8 (6.99)	806.8 (11.9)
sentence length (tokens)	13.06 (4.97)	14.68 (5.8)	16.33 (6.94)	17.89 (8.24)
chunk per sentence	6.85 (0.12)	7.66 (0.27)	8.53 (0.2)	9.23 (0.16)
word (log) frequency	4.92 (1.5)	4.87 (1.53)	4.83 (1.55)	4.81 (1.56)
ADULT TEXTS	ALL	SAVIANO	MAFFEI	
word length (letters)	5.17 (3.11)	4.89 (2.95)	5.52 (3.26)	
text length (tokens)	278.75 (37.99)	308.5 (12.79)	249 (26.49)	
sentence length (tokens)	26.99 (18.63)	20.22 (10.98)	47.0 (22.2)	
chunk per sentence	6.28 (3.86)	6.28 (3.86)	9.44 (5.71)	
word (log) frequency	4.32 (1.66)	4.4 (1.77)	4.22 (1.64)	

TABLE 1: MEAN LINGUISTIC FEATURES ACROSS CHILD AND ADULT TEXTS (WITH STANDARD DEVIATIONS).

Finger-tracked reading sessions are recorded using a common tablet in portrait orientation, on a 14.9×24.5 cm (5.87×9.65 in) screen with a resolution of 1920×1200 pixels. Finger movements are 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. Reading sessions are eye-tracked

[2] "Reading to understand: an ICT-driven, large-scale investigation of early grade children's reading strategies" – is a PRIN project (2017W8HFRX) coordinated by the ComphysLab at the Institute for Computational Linguistics, National Research Council (http://www.comphyslab.it). The experimental protocol was formally approved by the CNR Committee for Research Ethics, with the Ethical Clearance Statement 0037523/2021.

with an Eyelink Portable Duo (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. Drift correction was performed after each text episode. Order of delivery of the tracking method and reading condition are counterbalanced across participants. Presentation of the different reading texts is also alternated among participants, for them to be equally distributed across experimental conditions. In oral reading sessions, participants are wearing a pair of wireless noise-cancelling headphones with a retractable microphone. In the following section we consider in some detail how eye-tracking and finger-tracking data are automatically aligned with the reading text.

4. ALIGNING MULTIPLE TIME-SERIES

A common challenge in processing eye-tracking data is represented by the so-called *vertical drift*, i.e. the gradual misalignment of fixation points along the vertical axis of a visual space, due to the progressive loss of eye-tracking calibration. This issue is especially problematic in studies where participants are engaged in reading a connected text, since spatial accuracy is crucial to prevent attribution of a fixation to a wrong line. Vertical drifts are often corrected manually at a post-processing stage, but the approach is labour-intensive, time-consuming, and prone to errors and inconsistencies.

In the literature, several automated post-hoc methods have been proposed to correct alignments in eye-tracking data (Špakov *et al.* 2019). In a recent, comprehensive overview, Carr *et al.* (2021) showed that a technique based on Dynamic Time Warping (*Warp*) appears to outperform other methods. *Warp* is a sequential algorithm that relies on the identification of *return sweeps*, i.e. eye movements from the end of a text line to the start of the next, to split the original scan-path of a reader's eyes into fixation subsequences that are sequentially mapped onto text lines. In spite of their differences, the main source of information exploited by all existing techniques of automated drift correction is the spatial coordinates of both fixations and printed words on a page. This is explained by the largely dominant focus on silent reading of most current reading literature, a bias arguably due the extensive and exclusive use of eye-tracking records as reading data. Here, we show that it is possible to improve the accuracy of state-of-the-art drift correction algorithms by taking into account the dynamic interplay of eye movements and voice articulation in oral reading, and, in particular, variations in EVS. As shown in section 2, in oral reading, EVS may vary as a function of linguistic factors such as word length, word frequency and a word's syntactic role in a sentence, and performance

factors such as a reader's proficiency. If one controls for such factors, EVS variance can be reliably estimated and one can make the reasonable assumption that the most likely fixation point (given a range of candidate points) is the one that minimises the distance from the voice.

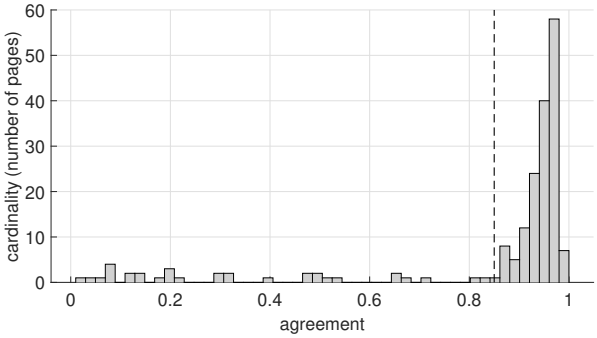

FIGURE 2: DISTRIBUTION OF EYE-TRACKED PAGES FOR THE LINE-BASED AGREEMENT BETWEEN *Warp* AND *Voice-Warp*. THE DASHED VERTICAL LINE INDICATES THE THRESHOLD BELOW WHICH *Warp* FAILS TO ASSIGN THE CORRECT LINE BY MISTAKING IN-LINE REGRESSIONS AS RETURN SWEEPS.

In Figure 2, we plot the performance of two alignment algorithms, namely *Warp* and *Voice-Warp*, where the latter is obtained by correcting *Warp* with time-aligned voice data. The overall dataset consists of a total of 188 eye-tracked text pages and 32816 fixations. For each eye-tracked page, we computed how many fixations were assigned to the same line by both *Warp* and *Voice-Warp* (line-based agreement), and then plotted the distribution of the pages in terms of their normalised line-based agreement (in the $[0, 1]$ interval). While the two methods converge in most of the reading sessions, a sizeable amount of pages shows a low line-based agreement.

To comparatively assess the accuracy of the two versions of *Warp*, we visually inspected all pages with a line-based agreement lower than 0.84 (32 individual pages and 5829 fixations). The threshold was set at 0.84 because, on all pages with a lower line-based agreement, misalignments were solely caused by in-line regressions being misidentified as return sweeps (Figure 3). Within this subset, *Warp* failed to assign the correct line in 63% of such cases (3693 fixations), while *Voice-Warp* was always correct. Note, in passing, that misalignments of this type are more frequent when a chin rest is not used, and loss of eye calibration is more likely. Finally, finger-tracking data of finger-point reading sessions are more reliable than eye-tracking data, as they present fewer cases of regressive movements that can be mistaken as return sweeps.

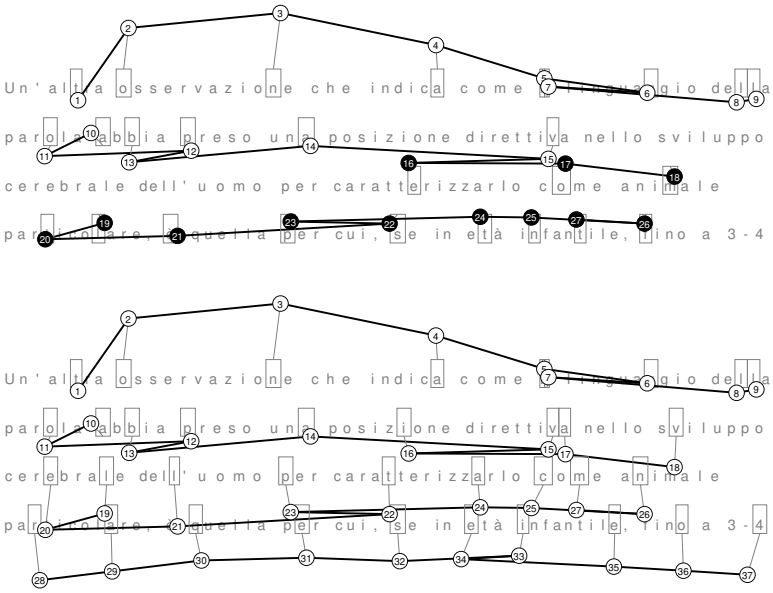

FIGURE 3: AN EXAMPLE OF IN-LINE REGRESSIONS BEING MISTAKEN AS RETURN SWEEPS BY *Warp* (TOP) AND CORRECTLY ASSIGNED BY *Voice-Warp* (BOTTOM). CIRCLES REPRESENT THE RAW COORDINATES OF EYE-TRACKED FIXATIONS, AND RECTANGLES FIXATION POSITIONS AFTER DRIFT CORRECTION. WRONGLY ASSIGNED FIXATIONS ARE MARKED AS BLACK CIRCLES. NUMBERS WITHIN CIRCLES INDICATE THE TIME ORDER OF FIXATIONS.

5. DATA ANALYSIS

In what follows, we propose a quantitative analysis of reading dynamics by focussing on the interaction of the three time-series of data, namely eye fixation speed, finger-tracking speed and token articulation speed. Statistical analyses of multisensory reading data – collected during the aloud experimental tasks – have been modelled with R (R Core-Team 2024) as generalised additive models (*gam* function) and graphed as non-linear regression plots and distribution plots (*ggplot* package, *geom-smooth* function). Data, code and modelling results are available at `https://doi.org/10.5281/zenodo.15276406`.

5.1 Developmental reading

In a developmental perspective, with reading data from 121 participants attending from 2nd to 5th school grade of two elementary schools in the area of Pisa (Italy), we observe the increasing effect of asynchronicity for eye-voice and finger-voice patterns (Figure 4). It is worth noting that, in the developmental perspective, speeds become increasingly greater (see values on the *y* axes – as

confirmed by a permutation-based Jonckheere-Terpstra test, with a significance p-value $< 2e-04$ for all modalities), as well as an increasing divergence of the different temporal dynamics (with p-values $< 2.2e-16$ for all grade levels), suggesting a progressively flexible and effective coordination mechanism that allows the temporal signals to proceed independently without enforcing a strict synchronisation (as confirmed by Kruskal-Wallis *chi-squared* values progressively increasing for increasing grade levels, χ^2 = 564.44, 671.25, 1117.1, 2025.4 for grades 2, 3, 4, 5, respectively).

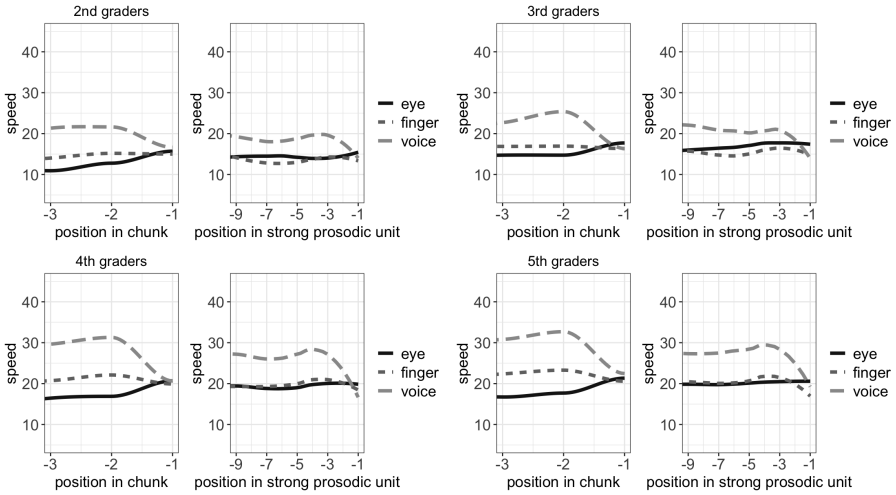

FIGURE 4: NON-LINEAR REGRESSION PLOTS FOR MODALITY SPEED (EYE-FIXATION, FINGER-TRACKING, VOICE ARTICULATION) AS CHARACTERS PER SECONDS, AS A FUNCTION OF TOKEN POSITION FROM THE HEAD OF CHUNKS (LEFT SUBPLOTS) AND FROM THE END OF STRONG PROSODIC UNITS (RIGHT SUBPLOTS), FROM 2ND (TOP LEFT) TO 5TH SCHOOL GRADE (BOTTOM RIGHT).

Interestingly, results show the developmental trend of a reading dynamic which is increasingly attuned to the text's prosodic structure with greater adherence to natural pauses and punctuation. This suggests a more refined and specifically focussed alignment between the temporal dynamics of voice and eye/finger movements driven by the prosody of the text itself. In other words, the observed patterns suggest that reading strategies evolve to prioritise prosodic features of the text to be read, as witnessed by an increasing more efficient chunking and a deeper integration of lexical and prosodic processing mechanisms (see Figure 5). Here, it should be observed that early readers appear not to differentiate between different levels of prosodic boundaries, with weak and strong prosodic units showing a similar pattern (see regression curves of central and bottom plots of Figure 5).

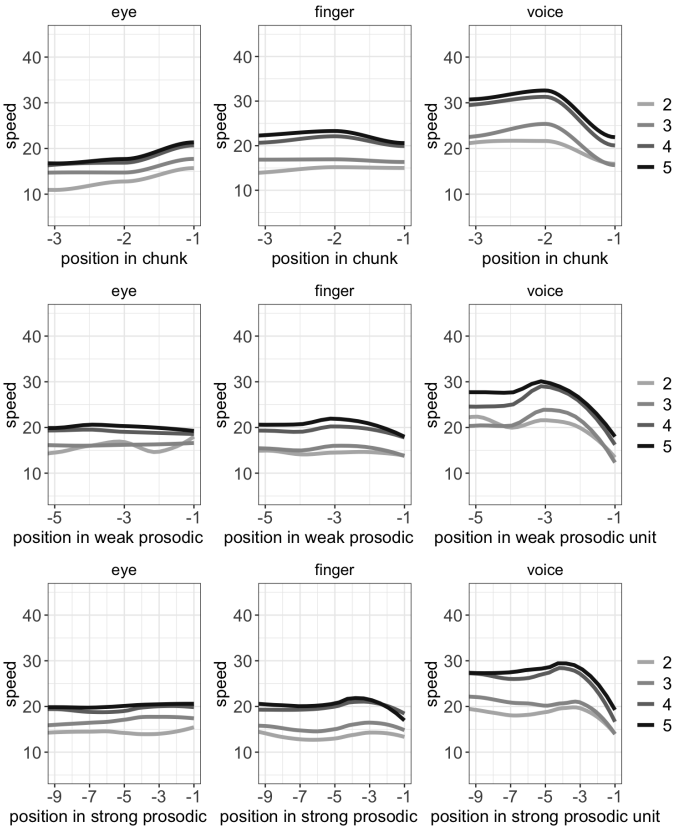

FIGURE 5: NON-LINEAR REGRESSION PLOTS FOR MODALITY SPEED (EYE-FIXATION, FINGER-TRACKING, VOICE ARTICULATION) AS CHARACTERS PER SECONDS, AS A FUNCTION OF GRADE LEVEL (2^{ND}, 3^{RD}, 4^{TH}, 5^{TH}) AND TOKEN POSITION FROM THE HEAD OF CHUNKS (TOP PANELS), FROM THE END OF WEAK (CENTRAL PANELS) AND STRONG (BOTTOM PANELS) PROSODIC UNITS.

Ultimately, the less experienced a reader is, the less finely modulated to align with the prosodic structure of a text is its reading speed. In addition, it is worth noting that younger readers show a stronger synchronisation between eye-tracking and finger-tracking speeds, suggesting that at earlier stages of reading development, the finger serves as a visual anchor to support the coordination of eye movements along the text. As readers grow older and become more proficient, the finger's movement becomes more closely aligned with the voice speed (post-hoc *Dunn* tests for multiple comparisons revealed significant increasing reading speed across modalities, with voice $>$ finger $>$ eye, for grades 3, 4, 5, with all p-values < 0.001, weaker for grade level 2). This shift likely reflects a more advanced integration of motor and cognitive processes,

where finger-tracking supports the pacing of articulation rather than the visual scanning of the text. Interestingly, the eye-finger contrast becomes increasingly pronounced with age (with post-hoc *Dunn* test $Z = 2.78$ for grade 2 and 10.92 for grade 5, where Z represents the standardised distance in reading speed between eye and finger), suggesting a progressive reorganisation of reading coordination towards a more proficient pattern.

5.2 Adult reading

With a goal of modelling highly proficient reading dynamics, we collected reading data from 59 young adults (in the 18–39 age range) in the premises of the CNR research area of Pisa and the SISSA of Trieste (Italy). Texts submitted to adults are linguistically more complex than texts administered to children (see the comparative overview of linguistic features across the different texts as reported in Table 1). Nevertheless, one can appreciate a greater ability in adults than children to quicker integrate the sources of information about print and speech (Figure 6). Adults, in fact, read significantly faster than children, in both eye/finger-based and voice-based speed (p-value $< 2.2e-16$). In addition, adult readers exhibit a more nuanced sensitivity to prosodic structure, showing a graded adjustment of speed at weak and strong prosodic units (see central and right plots in Figure 6). This pattern indicates that proficient readers integrate prosodic information in a hierarchically structured manner, allowing for fluid yet expressive aloud reading (in line with evidence on the contribution of prosodic sensitivity to reading comprehension, among others, Tong *et al.* 2023; Schwanenflugel *et al.* 2004; Koriat *et al.* 2002).

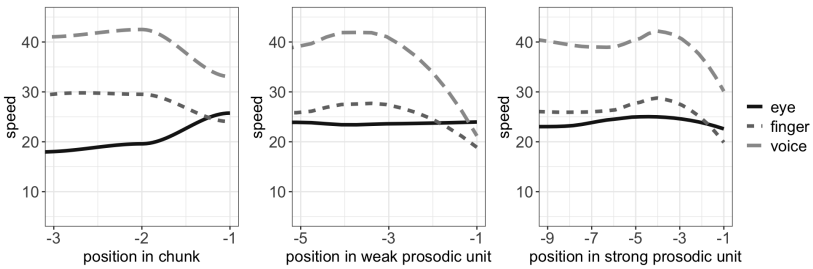

FIGURE 6: NON-LINEAR REGRESSION PLOTS FOR MODALITY SPEED (EYE-FIXATION, FINGER-TRACKING, VOICE ARTICULATION) AS CHARACTERS PER SECONDS IN ADULTS, AS A FUNCTION OF TOKEN POSITION FROM THE HEAD OF CHUNKS (LEFT PLOT) AND FROM THE END OF WEAK (CENTRAL PLOT) AND STRONG PROSODIC UNITS (RIGHT PLOT).

Here, the speed values of all time-series – eye, finger and voice – are notably higher compared to developing readers (see Figure 4), as a measure of their

high proficiency. In addition, they are also more distinct and differentiated, suggesting a less tight coupling and a functional specialisation. While eye-fixation moves ahead to facilitate anticipatory processing, consistent with parafoveal preview and predictive text processing, the finger maintains a steady pacing dynamic, often providing a motor anchor for voice articulation. This pattern is confirmed by a post-hoc *Dunn* test, showing markedly larger contrasts between modalities ($Z = 20.53$ for eye–finger, $Z = 62.79$ for eye–voice, $Z = 51.87$ for finger–voice), supporting the idea of signal-specific functional roles in proficient readers. The voice, in turn, exhibits the strongest modulation, slowing down on prosodically significant elements such as chunk heads and sentence-final boundaries marked by punctuation, as confirmed by *Levene*'s test on variance in speed across prosodic units ($F(2,n) = 9.38, p = .002$). No such effect was found for eye- or finger-tracking. This suggests that voice articulation is more dynamically adjusted to the prosodic structure of the text, likely reflecting an advanced stage of integration between linguistic parsing and motor planning.

Such findings align with evidence of prosodic and syntactic integration in skilled reading, where proficient readers optimise the interplay between linguistic, motor and cognitive processes, supporting fluent and context-sensitive reading (Inhoff & Radach 1998; Cutler & Foss 1977). In the context of eye-voice span (EVS) – the temporal lag between eye fixation on a word token and its subsequent vocal articulation during aloud reading (Inhoff *et al.* 2011; Crepaldi *et al.* 2022; Nadalini *et al.* 2024) – increasing proficiency levels are characterized by a larger EVS, as more skilled readers rely on parafoveal and preview-based processing to anticipate upcoming words, thereby decoupling visual and articulatory processes (Rayner 1998).

Likewise, finger-voice span (FVS) suggest an analogous coordination pattern, with early readers – as opposed to skilled readers – showing eye and finger movements tightly coupled. This suggests that the finger pointing to the text to be read may reinforce visual attention and serial scanning for letters and syllables (Uhry 2002; Mesmer & Lake 2010). As proficiency increases, FVS evolves from supporting gaze alignment to articulation time, with the finger movement becoming more aligned with the pace of voice articulation.

6. GENERAL DISCUSSION

Evidence of multisensory and cross-modal data collected in oral reading of connected texts contributes to a better understanding of the developmental trajectory of reading and its relation to text structure. In skilled readers, eye, finger and voice signals become increasingly distinct in their dynamics, with the voice slowing down at chunk heads and sentence boundaries. As reading

skills develop, the eye moves increasingly ahead of the voice, reflecting greater efficiency in text processing and lexical anticipation. Reading development appears to be characterised by a growing ability to flexibly adjust reading speed based on prosodic cues, moving beyond a stage of syllabic or lexical processing, towards a more adaptive, anticipatory processing of text structure. Accordingly, our results emphasise the importance of prosodic and syntactic sensitivity for proficient reading, reinforcing the notion that skilled readers adaptively balance the demands of linguistic, motor and cognitive processing.

We suggest that the increasing divergence between EVS and FVS for increasing levels of reading proficiency is the hallmark of a flexible coordination strategy. Skilled readers apportion their cognitive and motor resources in ways that optimally support fluency in voice articulation, prosodic and syntactic chunking and, ultimately, text comprehension. Evidence that the interplay between eye, voice, and finger evolves from a tightly coupled system to a more differentiated but functionally coordinated one, reflects the maturation of a multisensory reading strategy. Proficient reading is not simply characterised by greater speed, but requires the optimal interaction of motor and cognitive processes. While early reading is highly dependent on motor anchoring (with finger pointing guiding eye fixation), skilled reading reflects a more hierarchical, adaptive coordination, where each modality contributes independently to fluency and comprehension.

Our findings highlight the importance of a cross-modal perspective on reading research. By concurrently investigating eye/finger movements and voice articulation, we can gain a more comprehensive understanding of how different sensorimotor and cognitive mechanisms interact in reading development. Traditional studies of reading fluency have primarily focused on eye movements or voice articulation in isolation, but analysis of multimodal time-series can capture the dynamic interplay between visual, motor and lexical processes and their changes with reading proficiency. The observed shifts in EVS and FVS suggest that fluency is tightly related with how efficiently these modalities are coordinated. Crucially, the observed developmental trajectory towards an efficient reading strategy not only illustrates how it naturally evolves, but also how it is expected to evolve. Understanding how these complex multimodal dynamics evolve can be instrumental in identifying potential markers of atypical reading development, providing valuable support for fostering a better integration of visual, motor and articulatory processes.

REFERENCES

Carr, J.W., V.N. Pescuma, M. Furlan, M. Ktori & D. Crepaldi (2021). Algorithms for the automated correction of vertical drift in eye-tracking data. *Behavior Research Methods*. 1–24.

Crepaldi, D., M. Ferro, C. Marzi, A. Nadalini, V. Pirrelli & L. Taxitari (2022). Finger movements and eye movements during adults' silent and oral reading. In R. Levie, A. Bar-On, O. Ashkenazi, E. Dattner & G. Brandes (eds.) *Developing Language and Literacy: Studies in Honor of Dorit Diskin Ravid*, 443–471. Springer International Publishing. https://doi.org/10.1007/978-3-030-99891-2_17.

Cutler, A. & D.J. Foss (1977). On the role of sentence stress in sentence processing. *Language and speech*, 20(1). 1–10.

De Luca, M., M. Pontillo, S. Primativo, D. Spinelli & P. Zoccolotti (2013). The eye-voice lead during oral reading in developmental dyslexia. *Frontiers in human neuroscience*, 7. 696.

Gambi, C. & M. Crocker (2017). How do speakers coordinate planning and articulation? evidence from gaze-speech lags. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*.

Griffin, Z.M. (2003). A reversed word length effect in coordinating the preparation and articulation of words in speaking. *Psychonomic bulletin & review*, 10(3). 603–609.

Hirotani, M., L. Frazier & K. Rayner (2006). Punctuation and intonation effects on clause and sentence wrap-up: Evidence from eye movements. *Journal of Memory and Language*, 54(3). 425–443.

Inhoff, A.W. & R. Radach (1998). Definition and computation of oculomotor measures in the study of cognitive processes. In G. Underwood (ed.) *Eye Guidance in Reading and Scene Perception*, 29–53. Amsterdam: Elsevier Science Ltd. https://doi.org/10.1016/B978-008043361-5/50003-1.

Inhoff, A.W., M. Solomon, R. Radach & B.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.

Koriat, A., H. Kreiner & S.N. Greenberg (2002). The extraction of structure during reading: Evidence from reading prosody. *Memory & cognition*, 30. 270–280.

Laubrock, J. & R. Kliegl (2015). The eye-voice span during reading aloud. *Frontiers in psychology*, 6. 1432.

Lio, G., R. Fadda, G. Doneddu, J.R. Duhamel & A. Sirigu (2019). Digit-tracking as a new tactile interface for visual perception analysis. *Nature Communications*, 10(5392). 1–13.

Maffei, L. (2018). *Elogio della parola*. Bologna: il Mulino.

Mesmer, H.A.E. & K. Lake (2010). The role of syllable awareness and syllable-controlled text in the development of finger-point reading. *Reading Psychology*, 31(2). 176–201.

Meyer, A.S., E. Belke, C. Häcker & L. Mortensen (2007). Use of word length information in utterance planning. *Journal of Memory and Language*, 57(2). 210–231.

Nadalini, A., M. Ferro, A. Lento, V. Pirrelli & C. Marzi (2022). Evidence for saccadic reading dynamic with finger-tracking speed rates. Canada: The Mental Lexicon Conference.

Nadalini, A., C. Marzi, M. Ferro, L. Taxitari, A. Lento, D. Crepaldi & V. Pirrelli (2024). Eye-voice and finger-voice spans in adults' oral reading of connected texts. Implications for reading research and assessment. *The Mental Lexicon*. `https://benjamins.com/catalog/ml.00025.nad`.

R Core-Team (2024). *R: A language and environment for statistical computing.* Foundation for Statistical Computing. Vienna: Austria.

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

Rayner, K., G. Kambe & S.A. Duffy (2000). The effect of clause wrap-up on eye movements during reading. *The Quarterly Journal of Experimental Psychology: Section A*, 53(4). 1061–1080.

Romani, C., P. Silverstein, D. Ramoo & A. Olson (2022). Effects of delay, length, and frequency on onset rts and word durations. *Cognitive Neuropsychology*, 39(3-4). 170–195.

Schwanenflugel, P.J., A.M. Hamilton, M.R. Kuhn, J.M. Wisenbaker & S.A. Stahl (2004). Becoming a fluent reader: reading skill and prosodic features in the oral reading of young readers. *Journal of educational psychology*, 96(1). 119.

Silva, S., A. Reis, L. Casaca, K.M. Petersson & L. Faísca (2016). When the Eyes no longer lead: Familiarity and Length Effects on Eye-Voice Span. *Frontiers in Psychology*, 7. `https://doi.org/10.3389/fpsyg.2016.01720`.

Špakov, O., H. Istance, A. Hyrskykari, H. Siirtola & K.J. Räihä (2019). Improving the performance of eye trackers with limited spatial accuracy and low sampling rates for reading analysis by heuristic fixation-to-word mapping. *Behavior research methods*, 51. 2661–2687.

Tiffin-Richards, S.P. & S. Schroeder (2018). The development of wrap-up processes in text reading: A study of children's eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 44(7). 1051.

Tong, S.X., K. Lentejas, Q. Deng, N. An & Y. Cui (2023). How prosodic sensitivity contributes to reading comprehension: A meta-analysis. *Educational Psychology Review*, 35(3). 78.

Uhry, J.K. (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.

Warren, T., S.J. White & E.D. Reichle (2009). Investigating the causes of wrap-up effects: Evidence from eye movements and e–z reader. *Cognition*, 111(1). 132–137.

Claudia Marzi
National Research Council, Institute for Computational Linguistics (CNR-ILC)
via G. Moruzzi 1 - 56124 Pisa, Italy
e-mail: claudia.marzi@ilc.cnr.it
https://orcid.org/0000-0002-3427-2827

Andrea Nadalini
National Research Council, Institute for Computational Linguistics (CNR-ILC)
via G. Moruzzi 1 - 56124 Pisa, Italy
e-mail: andrea.nadalini@ilc.cnr.it
https://orcid.org/0000-0001-8859-9449

Alessandro Lento
National Research Council, Institute for Computational Linguistics (CNR-ILC)
via G. Moruzzi 1 - 56124 Pisa, Italy
e-mail: alessandro.lento@ilc.cnr.it
https://orcid.org/0009-0002-0825-1827

Manu Srivastava
National Research Council, Institute for Computational Linguistics (CNR-ILC)
via G. Moruzzi 1 - 56124 Pisa, Italy
e-mail: manu-srivastava@ilc.cnr.it
https://orcid.org/0009-0000-5667-0538

Alice Todesco
National Research Council, Institute for Computational Linguistics (CNR-ILC)
via G. Moruzzi 1 - 56124 Pisa, Italy
e-mail: alice.todesco@ilc.cnr.it
https://orcid.org/0009-0007-2763-5814

Vito Pirrelli
National Research Council, Institute for Computational Linguistics (CNR-ILC)
via G. Moruzzi 1 - 56124 Pisa, Italy
e-mail: vito.pirrelli@ilc.cnr.it
https://orcid.org/0000-0002-5581-7451

Marcello Ferro
National Research Council, Institute for Computational Linguistics (CNR-ILC)
via G. Moruzzi 1 - 56124 Pisa, Italy
e-mail: marcello.ferro@ilc.cnr.it
https://orcid.org/0000-0002-1324-3699