Comparing Reading Behavior across Reader Expertise and Text Complexity: Insights from the French Eye-Tracking Corpus (FETA)

Oksana Ivchenko Natalia Grabar
CNRS, Univ. Lille, UMR 8163 - STL - Savoirs Textes Langage, F-59000 Lille, France
oksana.ivchenko.etu@univ-lille.fr, natalia.grabar@univ-lille.fr

Abstract

This study examines how readers process general and medical texts with varying levels of complexity and how text simplification affects reading behavior. Using eye-tracking data, we compared two participant groups – *common population* and *speech therapy students* – as they read French medical, clinical, and general-domain texts in both original and simplified versions. We applied unsupervised clustering to identify patterns in reading behavior and investigate whether these patterns differ across participant groups, text types and complexity. The analysis identified between two and four clusters per group and condition, revealing distinct reading strategies ranging from effortful re-reading behavior to fluent, streamlined processing. The results reveal that medical and clinical texts elicit longer fixations and more regressions, indicating greater processing effort, while simplification produces shorter fixations and more fluid reading. Speech therapy students generally exhibit more efficient and stable gaze patterns, reflecting greater metalinguistic awareness and familiarity with the field. The dataset is a novel resource for modelling cognitive aspects of text complexity in French.

Keywords: eye-tracking, reading behavior, clustering, French eye-tracking corpus

1. Introduction

Reading is a complex cognitive process that proceeds in real time and involves a dynamic interplay between lexical access, syntactic parsing, discourse integration, and attention control (Rayner and Reichle, 2010). Eye movement tracking has become a robust method for studying text complexity and comprehension because it provides accurate, incremental measures (gaze fixation durations, regressions, saccades, etc.) that reflect the complexity of information processing at each moment in time (Strandberg et al., 2022; Torres et al., 2021). Torres et al. (2021) showed that eye-tracking measures can serve as proxies for text coherence and complexity (e.g. longer fixations in more complex regions), while Strandberg et al. (2022) demonstrated that such measures correlate with comprehension measures in reading tasks.

From the perspective of language resource creation and evaluation, corpora annotated with eye-tracking data are valuable for linking linguistic complexity (e.g., lexical, syntactic features) with empirical reading complexity, thus contributing to readability modeling, text simplification, and the creation of adaptive reading systems (González-Garduño and Søgaard, 2018; Demberg and Keller, 2008).

In this paper, we present an expanded version of an eye-tracking dataset[1] collected from two participant groups (common population vs. speech therapist students) reading three types of texts (on general, medical, and clinical cases), each in original and simplified versions. A first version of this resource, including only the common-population group, was presented in previous work (Ivchenko and Grabar, 2025). This extended design allows us not only to examine group differences in reading behavior, but also to explore how text type and simplification modulate gaze patterns. Our contributions are twofold:

(a) We provide a detailed description of the eye-tracking dataset in French and its associated eye-tracking measures.

(b) We explore whether common population and speech therapist students exhibit distinct reading patterns, and how these patterns vary across text types and levels of complexity.

Specifically, we address the following research questions:

1. Do common population and speech therapist students fall into different groups of reading behavior?
2. Do the discovered clusters correspond to text types (general / medical / clinical) or to text complexity level?
3. How do cluster characteristics (e.g. average fixation duration, regression count) differ across groups and text versions?

In the next sections, we review related work in Section 2, describe the dataset in Section 3, detail the clustering methodology in Section 4, present the results in Section 5, and conclude with future directions in Section 6. In Section 7, we indicate some limitations of the current work.

[1]The dataset and its documentation are available online at https://oksanaivchenko.github.io/feta-corpus/

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 6144–6154
11–16 May 2026.

2. Related Work

We review the related work across three lines: the existing corpora with eye-tracking measures, the link between eye-tracking and text complexity, and clustering of eye-tracking measures.

2.1. Eye-Tracking Corpora

Eye-tracking has long been used to study reading behavior and language processing. Several corpora have made gaze data widely available, providing valuable insights for psycholinguistic and computational analysis.

The Dundee Corpus (Kennedy et al., 2013) contains eye movements from English newspaper text reading, while the GECO Corpus (Cop et al., 2017) provides bilingual English–Dutch data for sentence-level reading. More recent resources, such as ZuCo (Zurich Cognitive Language Processing Corpus) (Hollenstein et al., 2018), combine eye-tracking and EEG recordings for naturalistic text reading, while the Provo Corpus (Luke and Christianson, 2018) includes comprehension measures alongside eye-movement data. Among multilingual eye-tracking resources, MECO is a major reference corpus (Siegelman et al., 2022) and CELER, a large English corpus covering both L1 and L2 reading (Berzak et al., 2022).

These corpora have been instrumental in developing cognitive models of reading and in evaluating NLP systems using human processing data. However, they focus primarily on English and general-domain texts, with limited attention to domain-specific complexity or text simplification.

2.2. Link of Eye-Tracking with Reading and Text Complexity

Multiple studies have shown that eye-tracking measures reflect the cognitive effort associated with reading and text complexity. Classic studies by Rayner (1998); Kuperman and Van Dyke (2011) demonstrated that gaze fixation duration and regressions reliably indicate information processing complexity and working memory load. Mézière et al. (2025) investigated whether eye movements recorded during natural, "read-only" reading could predict text comprehension performance. They found that fixation- and saccade-based measures such as reading speed, fixation duration, and regression rate reliably explained individual differences in recall scores. This demonstrates that eye-tracking measures obtained without comprehension tasks can serve as indicators of reading comprehension ability. More recent work by Torres et al. (2021) confirmed that such measures can serve as proxies for text coherence and linguistic complexity, providing a bridge between psycholinguistic theory and computational readability modeling.

2.3. Clustering of Eye-Tracking Metrics

Unsupervised and cluster-based analyses have also been applied to identify the underlying reading patterns. Koornneef and Mulders (2017) showed that dynamics of eye movements (saccade length and regression) can meaningfully distinguish readers who anticipate upcoming information from those who process text retroactively, demonstrating that eye-movement patterns serve as reliable indicators of cognitive reading strategies.

Also, Kucharský et al. (2020) used clustering of gaze measures to distinguish between cognitive strategies during reading. Kucharský's method is designed to discover latent gaze strategies based on where gaze moves between AOIs, e.g. systematic vs top-down scanning patterns. Similarly, Chen et al. (2025) and Koo et al. (2025) applied clustering methods to eye-tracking data to detect subgroups of participants characterized by different patterns of attention or comprehension.

Göbel and Martin (2018) introduced a structured framework for unsupervised clustering of eye-tracking data. Their approach combined feature extraction, clustering with algorithms such as K-Means, Spectral Clustering, and DBSCAN, and validation through both internal metrics (Silhouette score) and visual inspection using t-SNE. Also in the learning-analytics domain, Sáiz-Manzanares et al. (2021) applied unsupervised clustering to eye-tracking data, confirming the applicability of clustering techniques for reading or learning tasks.

Despite the growing body of work in this area, few datasets combine **domain-specific texts** (e.g., medical, clinical, and general texts) with **groups of participants with different levels of expertise**. Furthermore, the impact of text simplification on reading behavior in such domains remains understudied. Importantly, to our knowledge, no such studies have been conducted in the **French language**, limiting our understanding of how these effects manifest in French readers. The present study addresses this gap by providing an eye-tracking dataset that compares the reading strategies of common population and speech-therapist students when reading original and simplified versions of medical and non-medical texts.

3. Dataset Description

In this section, we describe the participants in the study, the texts exploited, the experimental protocol, and the eye-tracking data collected.

Original	**Simplified**
Les *hémocultures* ont permis d'isoler un *Staphylococcus aureus*. "Blood cultures made it possible to isolate a *Staphylococcus aureus*."	Les *hémocultures* (analyses des bactéries éventuelles dans le sang) ont montré la présence de la bactérie *Staphylococcus aureus*. "Blood cultures (tests for possible bacteria in the blood) showed the presence of the *Staphylococcus aureus* bacterium."

Table 1: Example of manual simplification presented in the inline original format.

Text type	#Texts	Version	Screens	Sentences	Tokens
Clinical	4	original	13	60	1187
		simplified	14	97	1435
General	4	original	33	161	3299
		simplified	31	212	3228
Medical	6	original	68	319	6485
		simplified	69	440	6941

Table 2: Summary of Original and Simplified Texts by Type

3.1. Participants

The dataset includes eye-tracking data from two groups of native French speakers: a common population group (CP, $n = 46$) and a speech-therapy student group (ST, $n = 20$). Participants reported normal or corrected-to-normal vision and no neurological or reading disorders. Their age ranged from 18 to 43 years ($M = 23.3, SD = 6.7$) in the CP group and from 20 to 45 years ($M = 23.95, SD = 4.85$) in the ST group.

The educational backgrounds of the CP participants were diverse: bachelor and master students, PhD candidates, post-doctoral researchers, and non-academic employees drawn from multiple disciplines. None of the CP participants had medical training. The ST group included students enrolled in speech therapy programs. Speech therapy students receive specialized paramedical training lasting five years in France, equivalent to a Master's degree. Their curriculum includes courses in linguistics, anatomy and physiology of speech and hearing, neuropsychology, reading and writing acquisition, rehabilitation methods, language disorders, and medical terminology. Overall, this training greatly distinguishes their educational background from that of the common population. Due to this background, ST students are expected to show higher metalinguistic awareness and greater familiarity with medical vocabulary compared to the CP.

A total of 66 participants were included in the dataset.

3.2. Texts

Fourteen texts belonging to three types are exploited: general-language, medical-language and clinical cases. General-language articles from Wikipedia present everyday topics such as *Weekend* or *Camelot*. Medical-language articles from Wikipedia cover specialized topics such as *Vascular Cerebral Accidents* or *Obstetrics*. Finally, clinical cases originate from toxicology and gastroenterology, and describe symptoms, diagnoses, treatments, and follow-ups for individual patients or small cohorts (Grabar et al., 2020). Their narrative structure resembles hospital discharge summaries and is densely packed with specialized terminology and reasoning about therapeutic choices.

Each text exists in an original and a manually simplified version, following plain-language principles (OCDE, 2015; Saggion, 2017). Simplifications were performed at three levels:

- Syntactic: Long or embedded clauses were segmented into shorter, clause-minimal sentences; passives were converted to actives.
- Lexical: Low-frequency or technical terms were replaced with higher-frequency synonyms or explained in-text.
- Semantic: Contextual or causal information was added to make implicit relations explicit.

We can see an example of simplified sentence in Table 1. This example illustrates a case of lexical simplification combining substitution and explanation. The less frequent and technical expression *ont permis d'isoler (made it possible to isolate)* was replaced with the more common and transparent form *ont montré la présence (showed the presence)*, reducing ambiguity. In addition, the specialized noun *hémocultures (blood cultures)* was supplemented with an explanatory gloss *(analyses des bactéries éventuelles dans le sang (tests for possible bacteria in the blood))* to support readers unfamiliar with medical terminology. These changes preserve the

informational content and accuracy of the original texts, while making the sentence more accessible to lay readers.

Original clinical texts contain approximately 1,180 words, general texts around 3,300 words, and medical texts around 6,480 words. We observe that simplified versions increase the number of sentences by roughly 25–45% due to segmentation and added explanations while preserving meaning. Each text was segmented into several screens, as explained in the next section. A detailed breakdown by screen, sentence, and token is shown in Table 2.

3.3. Eye-Tracking Setup and Experimental Protocol

To ensure balanced exposure to different text types and versions, the 14 texts were partitioned into two equally sized sets: *Set 1* and *Set 2*. Each set contained an identical proportion of medical, clinical, and general texts. Two presentation versions were created for counterbalancing:

- **Version A**: half of the texts appeared in their original version, and the remaining half in a simplified version;
- **Version B**: the assignment was reversed, so that each participant saw each text type once, but never both versions of the same text.

The experiment was conducted using a *Tobii Pro Spectrum* eye-tracker sampling at 600 Hz. Texts were presented on a 24-inch monitor (2880 $\times$ 1620 px), and participants were seated approximately 60 cm from the screen. A five-point calibration was accepted when the average accuracy error was below 0.5° (precision $\leq$ 0.2°). The experimental duration ranged from 50 to 70 minutes, including calibration and a short mid-session break.

Each slide, or screen, displayed one paragraph of the text. For longer texts, the content was split across multiple consecutive slides to maintain optimal legibility. Participants read at their own pace and advanced to the next slide using a mouse click, ensuring natural, self-paced reading.

Comprehension questions with three possible responses (*True / False / I don't know*) appeared after random segments to maintain engagement and check comprehension. These behavioral responses were not included in the present analysis but are available in the dataset.

In summary, the experimental pipeline proceeded as follows:

1. Pre-screening: participants completed an online form collecting demographic and ocular health information, reading habits, and informed consent.
2. Self-evaluation: upon arrival, participants filled in a short questionnaire assessing their perceived ease of understanding medical information.
3. Setup and calibration: eye movements were recorded. Calibration was performed.
4. Familiarisation block: participants read a short sample text and answered two comprehension questions aloud.
5. Main reading block: participants read original or simplified texts according to a counterbalanced Set/Version design. Comprehension questions were interspersed between slides.
6. Mid-session break: a short pause was followed by a second five-point recalibration.
7. Second reading block and debriefing: the remaining texts were presented, followed by brief oral questions about perceived difficulty and comprehension.

Eye-movement data were processed in *Tobii Pro Lab* using the I-VT (Velocity-Threshold Identification) algorithm. Each word was defined as an *Area of Interest (AOI)* in the exported data. Although the Tobii software extracted fixation measures at the word level, the current analysis aggregates these features at the sentence level to capture global reading behavior per sentence.

3.4. Recorded Eye-Tracking Metrics

For each word and participant, the dataset contains a comprehensive set of eye-movement measures automatically extracted in *Tobii Pro Lab*. Fixation events shorter than 60 ms were discarded, and merging parameters were set to a temporal gap ≤ 75 ms and an angular deviation $\leq 0.5°$.

The recorded metrics cover several dimensions:

- Fixation-based measures: total, average, minimum, and maximum fixation durations, number of fixations, and time to first fixation;
- Saccade-based measures: number of saccades within each AOI, time and peak velocity of entry and exit saccades;
- Regression and re-reading measures: regression-path duration, first-pass regression, and re-reading duration;
- Aggregated measures: metrics computed over whole fixations (e.g., Average_duration_of_whole_fixations, Total_duration_of_whole_fixations);

- Temporal and positional indices: time to entry/exit saccade and identification of the last AOI viewed.

Altogether, the dataset includes 27 continuous features per AOI (words) that jointly characterize early (first-pass), integrative, and late (re-reading) processing stages during reading.

4. Methods

In this section, we describe the preprocessing of the eye-tracking data, the feature selection and the clustering approach.

4.1. Preprocessing

The eye-tracking data were divided according to the two participant groups: *common population* and *speech therapists*. Analyses were conducted at the sentence level, where each Area of Interest (AOI) aggregated at the sentence level for analysis corresponded to a single sentence. To characterize overall reading patterns, sentence-level measures were averaged per participant, separately for the original and simplified versions of each text. This aggregation produced one set of summary measures per participant and text version, representing their overall reading effort and strategy.

Before clustering, all eye-tracking measures were preprocessed to ensure comparability across variables. Missing values, which may occur when a participant skips a sentence or when no fixation is detected, were replaced with measure-wise means. All numerical variables were then standardized using zero-mean, unit-variance scaling. This choice was motivated by the use of K-Means with Euclidean distance, which is sensitive to differences in variable scale. Since the selected measures include durations, counts, and proportions, standardization ensured that no single feature type disproportionately influenced the clustering process. For example, fixation durations are expressed in milliseconds and typically have much larger numeric ranges than proportions such as regression frequency. The resulting negative values should not be interpreted as negative reading times, but simply as values below the feature mean in standardized space.

4.2. Feature Selection

A data-driven feature selection procedure was employed to identify a concise, non-redundant set of eye-tracking measures for clustering. An initial pool of 27 numeric eye-tracking measures was considered. A Pearson correlation matrix was computed for all measures, and the resulting correlation heatmap revealed clear blocks of highly correlated measures, particularly among duration-based and fixation count variables.

Based on this analysis, and on theoretical considerations regarding different aspects of reading processes (Cook and Wei, 2019; Vasishth et al., 2013), a final set of five non-redundant features was selected for clustering:

- **First-pass first fixation duration** (ms): reflects early lexical processing effort upon first encountering a word;
- **Average duration of fixations** (ms): reflects overall processing effort per word;
- **First-pass regression** (proportion): indicates whether the reader made a backward eye movement from a word during its first encounter, reflecting early integration difficulty;
- **Number of fixations**: represents how many times the reader's gaze landed on the word or sentence (AOI);
- **Regression-path duration** (ms): measures the total time spent re-reading a word and preceding words before moving forward, indicating reanalysis or comprehension monitoring processes.

We can see that the chosen eye-tracking features, in addition to being non-redundant, correspond to various stages of information processing: access, integration, and verification.

4.3. Clustering Procedure

To uncover distinct patterns of reading behavior, an unsupervised clustering analysis was performed separately for the *common population* and the *speech therapists*. Participant-level eye-tracking features were grouped using the K-Means algorithm with Euclidean distance as the similarity measure (Hartigan, 1975).

The optimal number of clusters (k) for each group and text version was determined empirically using a combination of the Elbow Method (Thorndike, 1953) and the Silhouette Score (Rousseeuw, 1987), which jointly assess cluster compactness and separation. This procedure suggested: three clusters for the common population reading original texts, four clusters for the common population reading simplified texts, two clusters for speech therapists reading original texts, and three clusters for speech therapists reading simplified texts.

Separate K-Means models were thus fitted for each participant group and text version, allowing us to compare the resulting reading profiles across populations and text versions.

5. Results

We compared eye-tracking features across text types (medical and general texts, clinical cases and simplified clinical cases) and participant groups (common population vs. speech therapists) to identify reading patterns. We also compared our results with the existing work.

5.1. Comparison of Reading Behavior Across Text Types and Groups

Figure 1 presents the standardized values of five eye-tracking features across four text conditions, shown separately for the *Common population* (left) and *Speech therapists* (right).

Across both groups, reading patterns vary systematically with text type and complexity. In the **Common population**, original medical and clinical texts show higher median values for fixation durations, regression-path durations, and number of fixations, indicating increased processing effort and re-reading. By contrast, the simplified clinical text elicits lower fixation durations and regression measures, reflecting smoother and more efficient reading. General original texts occupy an intermediate position, suggesting that their familiarity and everyday vocabulary reduce processing difficulty and facilitate overall comprehension. For **Speech therapists**, the distributions of fixation-based features are overall lower and more consistent across conditions. This group shows limited variation between original and simplified clinical texts, implying more stable and automatic reading even in complex medical contexts.

5.2. Cluster Profiles

The clustering analysis revealed distinct reading patterns across reader groups and text versions. Figure 2 visualizes the standardized cluster centroids across the five eye-tracking features. Red indicates higher-than-average fixation durations or regressions (reflecting deeper or more effortful reading), whereas blue represents lower-than-average values (fluent or faster reading). This heatmap shows that, for the **common population**, original texts contain clusters with overall longer fixation durations and more regressions, indicating higher processing effort. Simplified texts produce additional clusters characterized by shorter fixations and fewer regressions, suggesting increased fluency and greater strategy diversification. For the **speech therapists**, clusters display generally shorter fixation durations and fewer regressions across both versions, reflecting more efficient and stable reading. Simplified texts further reduce regressions, confirming the facilitating effect of simplification on reading ease.

To better characterize the patterns, each cluster was assigned a descriptive label based on its eye-tracking profile (Table 3). For the **common population** reading original texts, three distinct profiles emerged: *effortful deep processors* with prolonged fixations and thorough processing (Cluster 1), *struggling readers* showing frequent backtracking (Cluster 2), and *efficient fluent readers* with smooth processing and minimal difficulty (Cluster 3). When reading simplified texts, this group exhibited more diverse strategies, with four clusters ranging from *rapid scanners* to *minimal-effort readers*, confirming that simplification induces more varied reading behaviors. *Selective-attention* readers often focus on certain words and may skip others, regardless of whether those words are simple or complex. **Speech therapists** demonstrated more consistent and efficient patterns with fewer clusters. Their reading of original texts revealed *balanced professionals* and *analytical reviewers*, while simplified texts elicited *quick professional scanning* and *verification* behaviors. Reduced cluster number among speech therapists compared to the common population indicates more homogeneous, expertise-driven reading strategies across text types.

5.3. Distance Between Group Centroids

To quantify differences in reading behavior between the two participant groups, we computed the Euclidean distances between the cluster centroids of the *Common population* and *Speech therapists*, separately for original and simplified texts (Figure 3). Smaller distances (light cells) indicate similar reading strategies with comparable fixation and regression patterns, while larger distances (dark cells) correspond to distinct reading behaviors.

Original texts. The smallest distance (0.58) occurs between CP C2 (*Struggling/Backtracking Readers*) and ST C2 (*Analytical Reviewers*), suggesting similar high-regression reading patterns at the behavioral level, although the underlying cognitive processes may differ. In contrast, the largest distance (3.43) appears between CP C1 (*Effortful/Deep Processors*) and ST C2 (*Analytical Reviewers*), reflecting markedly different reading approaches. Notably, CP C3 (*Efficient/Fluent Readers*) is relatively close to ST C1 (*Balanced Professionals*) with a distance of 0.69, suggesting that fluent readers in the common population exhibit reading patterns comparable to those of speech therapists.

Simplified texts. The smallest distance (0.49) occurs between CP C1 (*Rapid Scanners*) and ST C1 (*Quick Professional Scanners*), suggesting that text simplification may enable both groups to adopt similarly efficient, quick scanning behaviors. The largest distance (3.79) appears between CP C2 (*Engaged/Careful Processors*) and

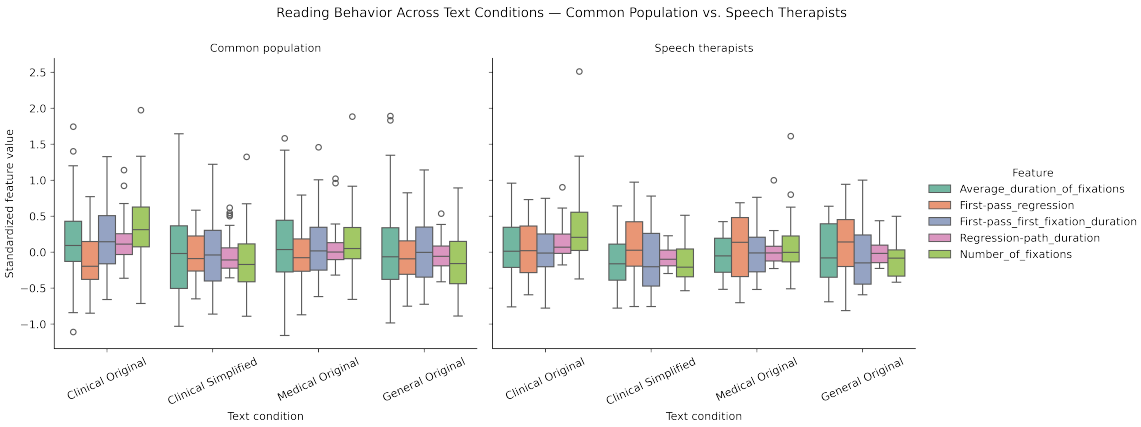

Figure 1: Comparison of standardized eye-tracking features across text types for both groups.

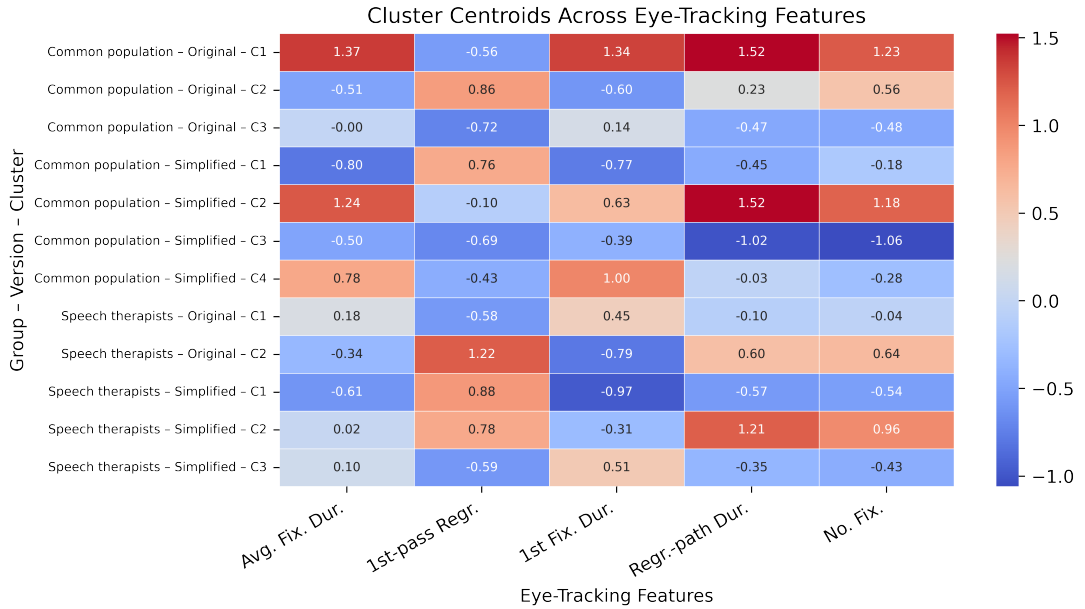

Figure 2: Heatmap of standardized cluster centroids across five eye-tracking features. Red indicates higher-than-average fixation durations or regressions (more effortful reading), blue indicates more fluent reading.

ST C1 (*Quick Professional Scanners*), indicating that some common-population readers maintain thorough, engaged processing even with simplified texts, whereas speech therapists tend to adopt rapid assessment strategies. Interestingly, CP C4 (*Selective Attention*) shows the closest alignment with ST C3 (*Expert Fluent Readers*) at a distance of 0.93, suggesting that strategic, focused readers in the common population may approach the fluency of speech therapists when reading simplified texts.

Overall, while the minimum distance decreased slightly from 0.58 to 0.49, the maximum distance increased from 3.43 to 3.79, suggesting that simplification may produce convergence for some reading strategies while accentuating differences for others. However, this comparison should be interpreted cautiously, since the clustering solutions were estimated independently for each group and do not share the same number or configuration of clusters. As a result, centroid distances are not strictly equivalent across groups and may partly reflect differences in partition structure rather than only differences in reading behavior.

5.4. Comparison with the Existing Work

Only limited comparison with previous work is possible, mainly because the tasks previously explored are different from the present text-reading: they may address tasks like learning process (Sáiz-Manzanares et al., 2021) or map reading (Göbel and Martin, 2018). The closest work used two eye-tracking measures (saccade length and regression)

Ver.	Cluster Label	Key Characteristics of the Reading Profiles
		Common population
O	C1: *Effortful / Deep Processors*	Very high fixation duration, regression-path duration, and number of fixations; low regressions; thorough, careful reading with sustained attention.
O	C2: *Struggling / Backtracking Readers*	High first-pass regressions, moderate fixations; short fixation durations; difficulty with comprehension, frequent re-reading.
O	C3: *Efficient / Fluent Readers*	Low regressions, balanced fixation patterns; fewer fixations overall; smooth processing with minimal difficulty.
S	C1: *Rapid Scanners*	Very short fixations, high regressions, fewer fixations; quick scanning behavior, checking back for confirmation.
S	C2: *Engaged / Careful Processors*	High fixation duration and regression-path duration; more fixations; thorough engagement despite simplified text.
S	C3: *Minimal Effort Readers*	Extremely low across all metrics; very easy processing, possibly disengaged or text too simple.
S	C4: *Selective Attention*	High fixation duration and first-fixation duration; fewer fixations and regressions; strategic, focused processing.
		Speech therapists
O	C1: *Balanced Professionals*	Near-neutral across all metrics; efficient expert reading with professional familiarity.
O	C2: *Analytical Reviewers*	Very high regressions, moderate fixations; critical evaluation mode, checking for therapeutic relevance.
S	C1: *Quick Professional Scanners*	Short fixations, high regressions, fewer overall fixations; rapid professional assessment of modifications.
S	C2: *Verification Readers*	High regressions and regression-path duration; more fixations; detailed checking of simplification quality.
S	C3: *Expert Fluent Readers*	Low regressions, balanced fixation patterns; effortless professional reading of simplified content.

Table 3: Cluster profiles (*C1, C2*...) and characteristics based on eye-tracking metrics for two populations (Common Population and Speech Therapists) and two versions of texts (original O and simplified S).

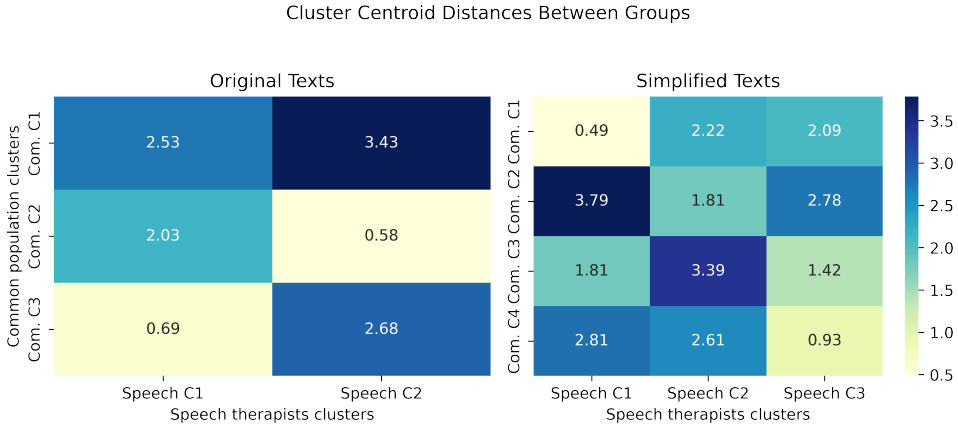

Figure 3: Heatmap of distances between cluster centroids for the two groups. Smaller values (light cells) indicate similar reading profiles; higher values (dark cells) indicate distinct reading behaviors.

to distinguish two reading profiles: proactive (long saccades, many regressions) and conservative (short saccades, few regressions) readers (Koornneef and Mulders, 2017). In our work, we provide a more fine-grained analysis of the reading patterns and a deeper understanding of this cognitive process.

6. Conclusion and Future Work

This study introduced a novel French eye-tracking corpus combining domain-specific texts (medical, clinical, and general) in original and simplified versions, with two reader populations with different levels of medical-domain expertise: the common population and speech therapy students. Using

unsupervised clustering of fixation- and regression-based features, we identified distinct reading profiles that reveal how text type, complexity, and reader expertise shape reading behavior.

Our analysis yielded three main findings. (1) Medical and clinical texts elicited greater processing effort, with longer fixations and more regressions, while simplification facilitated smoother, more fluent reading in both groups. (2) Speech therapy students exhibited more efficient and stable gaze patterns, reflecting higher metalinguistic awareness and resulting in fewer, more homogeneous clusters. (3) Simplification diversified reading strategies, particularly among the common population, creating convergence for some profiles (e.g., *Rapid Scanners*) but amplifying differences for others, leading to greater overall heterogeneity rather than uniform simplification effects. The resulting dataset is a valuable new resource for modeling the cognitive aspects of text complexity in French.

Future work will leverage these results to gain insight into the need for targeted, individualized simplification, as the increase in differences after text modification indicates that effective simplification must take into account the reader's prior knowledge and specific cognitive strategy. Furthermore, we aim to link these behavioral patterns with comprehension outcomes and to develop predictive models capable of estimating eye-tracking features directly from text, which will contribute to the study of cognitive indicators of text complexity in French.

7. Limitations

Among the limitations of this work, we can mention:

- The total sample (66 participants), while substantial for an eye-tracking experiment, may still limit the generalizability of the clustering results and between-group comparisons.
- The two participant groups (common population vs. speech therapy students) also differ in size and background diversity, which may influence the observed variability in reading patterns.
- Furthermore, averaging features at the sentence level smooths out within-sentence variability and may obscure fine-grained word-level effects. This aspect will be examined in future work.
- Finally, although comprehension questions were administered during the experiment, the present analysis focuses exclusively on eye-tracking measures and does not include the behavioral comprehension scores (True / False / I don't know). These questions were primarily designed to maintain attention during reading and were not sufficiently numerous or fine-grained to provide a robust measure of sentence-level comprehension. Future work will examine the relationship between comprehension performance and gaze-based reading profiles.

8. Ethical Considerations

In this study, we use eye-tracking data collected from participants while they read texts. The eye movements are recorded using a specialized eye-tracking camera. The study protocol has been approved by the Data Protection Officer (DPO) of the institution. Participants receive an information letter and a consent form before the experiment begins. Participation is entirely voluntary: the participants are free to choose whether to participate and can withdraw from the experiment at any time. Additionally, participants can request the removal of their data without providing any justification. Participants receive either financial compensation of €12 or pedagogical compensation in the form of course credits for their participation in the experiment.

Acknowledgement

This work was partially funded by the French National Research Agency (ANR) through the CLEAR project (Communication, Literacy, Education, Accessibility, Readability), ANR-17-CE19-0016-01 and by the CDP PRIME-NEXTGEN project of the Université de Lille. We gratefully acknowledge the Fédération de Recherche Sciences et Cultures du Visuel (FR CNRS 2052 SCV), as well as the state support managed by the Agence Nationale de la Recherche under the Programme d'Investissements d'Avenir (grant ANR-21-ESRE-0030 – Equipex+ Continuum), for providing access to the eye-tracking equipment and laboratory facilities used in this study. We also thank the reviewers for their helpful comments and questions, which improved the overall quality of the paper.

9. Bibliographical References

S. Chen, J. Feng, and M. Carl. 2025. Exploring preparatory reading in bidirectional sight and written translation through clustering analysis of eye-tracking data. *PLOS ONE*, 20(8):e0329858.

Anne E. Cook and Wei Wei. 2019. What can eye movements tell us about higher level comprehension? *Vision*, 3(3).

Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109:193–210.

Ana González-Garduño and Anders Søgaard. 2018. Learning to predict readability using eye-movement data from natives and learners. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32.

Fabian Göbel and Henry Martin. 2018. Unsupervised clustering of eye tracking data.

John A. Hartigan. 1975. *Clustering Algorithms*, 99th edition. John Wiley & Sons, Inc., USA.

S. J. Koo, E. J. Cha, J. E. Min, et al. 2025. Eye tracking based clustering using the korean version of the reading the mind in the eyes test. *Scientific Reports*, 15:3929.

Arnout Koornneef and Iris Mulders. 2017. Can we 'read' the eye-movement patterns of readers? unraveling the relationship between reading profiles and processing strategies. *Journal of Psycholinguistic Research*, 46.

Šimon Kucharský, Ingmar Visser, Gabriela-Olivia Truțescu, Paulo G. Laurence, Martina Zaharieva, and Maartje E. J. Raijmakers. 2020. Cognitive strategies revealed by clustering eye movement transitions. *Journal of Eye Movement Research*, 13(1):1–20.

V. Kuperman and J. A. Van Dyke. 2011. Effects of individual differences in verbal skills on eye-movement patterns during sentence reading. *Journal of Memory and Language*, 65(1):42–73.

Diane C. Mézière, Lili Yu, Titus von der Malsburg, Erik D. Reichle, and Genevieve McArthur. 2025. Using eye movements from a "read-only" task to predict text comprehension. *Reading Research Quarterly*, 60(3):e70023. E70023 RRQ-2024-10-0290.

OCDE. 2015. *Guide de style de l'OCDE Troisième édition: Troisième édition*. OECD Publishing.

Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3):372–422.

Keith Rayner and Erik D. Reichle. 2010. Models of the reading process. *WIREs Cognitive Science*, 1(6):787–799.

Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Horacio Saggion. 2017. *Automatic Text Simplification*, volume 32 of *Synthesis Lectures on Human Language Technologies*. Morgan & Claypool, University of Toronto.

Andrea Strandberg, Mattias Nilsson, Per Östberg, and Gustaf Öqvist Seimyr. 2022. Eye movements during reading and their relationship to reading assessment outcomes in swedish elementary school children. *Journal of Eye Movement Research*, 15(4):1–16.

María Consuelo Sáiz-Manzanares, Ismael Ramos Pérez, Adrián Arnaiz Rodríguez, Sandra Rodríguez Arribas, Leandro Almeida, and Caroline Françoise Martin. 2021. Analysis of the learning process through eye tracking technology and feature selection techniques. *Applied Sciences*, 11(13).

Robert Thorndike. 1953. Who belongs in the family? *Psychometrika*, 18(4):267–276.

Débora Torres, Wagner R. Sena, Humberto A. Carmona, André A. Moreira, Hernán A. Makse, and José S. Andrade. 2021. Eye-tracking as a proxy for coherence and complexity of texts. *PLOS ONE*, 16(12):e0260236.

Shravan Vasishth, Titus von der Malsburg, and Felix Engelmann. 2013. What eye movements can tell us about sentence comprehension. *WIREs Cognitive Science*, 4(2):125–134.

10. Language Resource References

Yevgeni Berzak, Chie Nakamura, Amelia Smith, Emily Weng, Boris Katz, Suzanne Flynn, and Roger Levy. 2022. Celer: A 365-participant corpus of eye movements in l1 and l2 english reading. *Open Mind*, 6:1–10.

U. Cop, N. Dirix, D. Drieghe, and W. Duyck. 2017. Presenting geco: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods*, 49(2):602–615.

Natalia Grabar, Clément Dalloux, and Vincent Claveau. 2020. CAS: corpus of clinical cases in French. *Journal of BioMedical Semantics*, 11(1):1–7.

N. Hollenstein, J. Rotsztejn, M. Troendle, et al. 2018. Zuco, a simultaneous eeg and eye-tracking resource for natural sentence reading. *Scientific Data*, 5:180291.

Oksana Ivchenko and Natalia Grabar. 2025. A French eye-tracking corpus of original and simplified medical, clinical, and general texts - FETA. In *Proceedings of the First International Workshop on Gaze Data and Natural Language Processing*, pages 37–43, Varna, Bulgaria. INCOMA Ltd., Shoumen, BULGARIA.

A. Kennedy, J. Pynte, W. S. Murray, and S. A. Paul. 2013. Frequency and predictability effects in the dundee corpus: an eye movement analysis. *Quarterly Journal of Experimental Psychology*, 66(3):601–618. Epub 2012 May 29.

S. G. Luke and K. Christianson. 2018. The provo corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods*, 50:826–833.

Noam Siegelman, Sascha Schroeder, Cengiz Acartürk, Hyeonjeong D. Ahn, Svetlana Alexeeva, Sara Amenta, Raymond Bertram, Riccardo Bonandrini, Marc Brysbaert, Daria Chernova, Susana M. Da Fonseca, Nicolas Dirix, Wouter Duyck, Agnieszka Fella, Ram Frost, Carolina A. Gattei, Athanasia Kalaitzi, Nahyun Kwon, Kaidi Lõo, Marco Marelli, Timoklis C. Papadopoulos, Athanassios Protopapas, Spyridon Savo, Diego E. Shalom, Natalia Slioussar, Roni Stein, Liyun Sui, Alvina Taboh, Veronica Tønnesen, Kübra A. Usal, and Victor Kuperman. 2022. Expanding horizons of cross-linguistic research on reading: The Multilingual Eye-movement Corpus (MECO). *Behavior Research Methods*, 54(6):2843–2863.