Actes de CORIA-TALN-RJCRI-RECITAL 2025
Marseille, France, 30 juin au 4 juillet 2025

L'Impact de la complexité textuelle sur le comportement de lecture : une analyse oculométrique et de la surprise des textes français

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

RÉSUMÉ

Cette étude examine comment la complexité du texte affecte les processus de lecture à travers différents types de textes en combinant la méthodologie d'oculométrie avec l'analyse de la surprise. Nous avons créé un corpus en français avec des textes généraux, cliniques et médicaux, dans leurs versions originales et simplifiées, annotés avec des mesures oculométriques complètes provenant de 23 participants. La modélisation linéaire à effets mixtes révèle que la surprise prédit significativement les temps de lecture pour tous les types de textes, les textes médicaux montrant une sensibilité accrue aux mots inattendus. De façon importante, la simplification a des effets différentiels selon le type de texte : bien qu'elle ne réduit pas significativement les temps de lecture pour les textes cliniques, elle diminue considérablement les temps de lecture pour les textes médicaux. De plus, la simplification atténue l'effet de la surprise spécifiquement dans les textes médicaux, réduisant le coût cognitif associé au traitement des mots inattendus.

ABSTRACT

The Impact of Text Complexity on Reading Behaviour : An Eye-Tracking and Surprisal Analysis of French Texts

This study investigates how text complexity affects reading processes across different text types by combining eye-tracking methodology with surprisal analysis. We created a corpus of French general, clinical, and medical texts in both original and simplified versions, annotated with comprehensive eye-tracking measurements from 23 participants. Linear mixed effects modelling reveals that surprisal significantly predicts reading times across all text types, with medical texts showing heightened sensitivity to unexpected words. Importantly, simplification has differential effects depending on text type : while it does not significantly reduce reading times for clinical texts, it substantially decreases reading times for medical texts. Moreover, simplification mitigates the effect of surprisal specifically in medical texts, reducing the cognitive cost associated with processing unexpected words.

MOTS-CLÉS : oculométrie, textes médicaux et généraux, français, traitement cognitif.

KEYWORDS: eye-tracking, medical and general texts, French, cognitive processing.

Actes des 32ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : articles scientifiques originaux, pages 451–467.

1 Introduction

Text readability provides indications of how easily a given text can be read, understood, and used (Pitler & Nenkova, 2008). This becomes particularly important in specialized domains such as medicine, law, or physics, where texts often contain domain-specific terminology, complex sentence structures, and dense informational content. Compared to general language, such texts pose greater challenges for non-expert readers, leading to increased comprehension difficulty (Eklics & Fekete, 2024; Brown, 2008). The measure of text complexity has been addressed by researchers for a long time. Early readability assessment methods, such as the Flesch Reading Ease (Flesch, 1948) or the Gunning Fog Index (Gunning, 1973), attempted to estimate text complexity through simplistic formulas, though their limitations in specific domain have since been widely recognized ((Zheng & Yu, 2017; Kim *et al.*, 2007). With the rise of machine learning and deep learning, data-driven readability models have been developed to improve upon traditional formulas. These approaches use a wide range of features collected at lexical, syntactic and semantic levels, and extracted from large corpora to train classifiers or regressors that predict readability scores (François & Fairon, 2012; Gooding *et al.*, 2021; González-Garduño & Søgaard, 2017). Lately, neural models are being adapted for readability prediction and are used alone or in combination with linguistic features, thus improving overall prediction quality (Nadeem & Ostendorf, 2018; Deutsch *et al.*, 2020; Martinc *et al.*, 2021).

Yet, despite advances in computational metrics for quantifying linguistic complexity, it remains unclear whether the existing measures can reliably predict the reading behaviour of readers across different text types. One recent approach is based on the notion of *surprisal*, an information-theoretic measure that quantifies how unexpected a word is given its preceding context (Hale, 2001; Levy, 2008). Surprisal, being based on contextual predictability rather than surface features, may provide more domain-robust predictions of reading difficulty. Higher surprisal values indicate that a word is less predictable, reflecting a greater cognitive load during sentence processing. Although previous studies have established that higher surprisal values are associated with increased processing difficulty ((Lowder *et al.*, 2018; Goodkind & Bicknell, 2018; Levy, 2013), most research has focused on general reading patterns rather than examining variations across original and simplified texts, or texts from different domains. In our work, we propose to study how eye-tracking indicators correlate with the notion of surprisal.

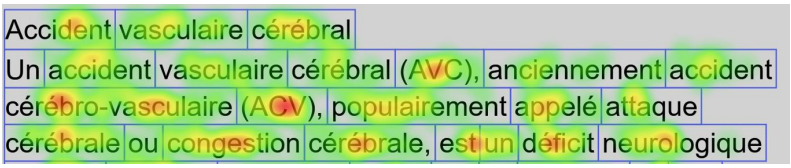

FIGURE 1 – Fixation Heatmap from eye-tracking data. This visualization shows where readers focused on the text : green indicates short fixations, while red highlights longer fixations.

Eye-tracking technology offers objective insights into real-time reading behaviour (Duchowski, 2007), making it a promising method for assessing text complexity. Eye-tracking indeed provides reliable, objective data on the reading process (Ahmed *et al.*, 2016; Cromley *et al.*, 2010). Hence, researches have demonstrated that several key eye-tracking metrics correlate with reading ease and text complexity (Salvucci & Goldberg, 2000; Duchowski, 2007; Radach & Kennedy, 2013) :

— *Fixations* are brief pauses during reading. Longer fixations (see Figure 1) indicate higher cognitive load and processing difficulty.

— The *number of fixations* is important because a greater number of fixations on a word suggests

an increased lexical difficulty or ambiguity.
— *Saccades* are rapid eye movements between fixations, occurring when the eyes move quickly without processing visual information.
— *Regressions* are movements back to earlier parts of the text, often due to comprehension challenges (Rayner, 1998) or the need to reexamine recent content.

Analyzing these indicators enables researchers to identify specific words or passages that contribute to reading difficulty, thereby informing more effective evaluation of text complexity.

The purpose of our study is to address two key questions :

(1) How do eye-tracking metrics (e.g., fixation durations, regression patterns, and re-reading times) differ when non-experts read various types of texts (original vs. simplified, specialized vs. general) ?

(2) To what extent can surprisal, as a measure of linguistic complexity, account for or predict these differences in reading behaviour ?

By comparing eye-tracking indicators across text types and correlating them with computed surprisal values, we seek to uncover whether surprisal reflects the cognitive effort during reading or if other factors inherent to text complexity play a more decisive role. This investigation, which is positioned at the cross-road of NLP and psycholinguistics, will contribute to a deeper understanding of the interplay between linguistic complexity and reading behaviour, particularly among non-expert audiences.

Besides, a specific accent is put on the textual data processed. Our corpus contains three categories of texts : (1) general-language texts from Wikipedia covering common topics, (2) medical texts from Wikipedia covering medical topics, and (3) clinical texts, that are clinical cases, typically created within clinical healthcare process of patients. The original versions of texts are simplified manually, and both versions (original and simplified) are studied through eye-tracking and surprisal measures.

The remainder of this paper is organized as follows. Section 2 reviews related work in eye-tracking and surprisal research. Section 3 describes our dataset and experimental methodology, including details on the eye-tracking setup. Section 4 outlines our approach to computing surprisal and the subsequent analysis. In Section 5, we present our findings and discuss their implications. Finally, Section 6 concludes the paper and suggests directions for future work.

2 Related Work

2.1 Eye Tracking in (Psycho)linguistics in Relation with Text Complexity

Eye-tracking technology is being widely used for various tasks in the linguistics and NLP domains. We present here some eye-tracking works related with the study of text complexity. These studies are done with different populations and types of texts.

The research by (Singh *et al.*, 2016) uses eye-tracking data to predict the complexity of sentences. First, the authors build a model trained on real human reading data (from the Dundee eye-tracking corpus (Kennedy *et al.*, 2013)) and linguistic features known to affect reading difficulty (word frequency, sentence structure, surprisal, etc.). Then, they apply the model to unseen sentences to automatically detect the sentences that are difficult to read. The study shows that automatically predicted eye-tracking measures can indeed serve as a strong indicator of text complexity and detection of difficult sentences. In another study, the researchers address the reading and processing of health-related texts

by comparing how third-year medical students and residents read and solve two patient cases on cardiac failure and pulmonary embolus, revealing that residents arrive at correct diagnoses more rapidly and with fewer fixations (Vilppu *et al.*, 2016). A differential analysis of reading patterns across original and simplified medical texts has also been proposed (Grabar *et al.*, 2018). The researchers found that simplification leads to fewer and shorter fixations while lengthening saccade amplitude. This indicates that technical medical terminology in original texts typically necessitates multiple fixations, whereas simplified versions enable more efficient and continuous reading processes. However, this experiment was carried out on very short passages and only compared original versus simplified medical texts. Yet another study (Mézière *et al.*, 2023) compares three commonly used reading comprehension tests (YARC, GORT-5, WRAT-4) that vary in silent reading, oral reading, and cloze tasks. Eye-tracking measures explain substantially more variance in comprehension than reading speed alone, although no single measure is predictive for all tests—highlighting how different tests involve distinct cognitive processes and suggesting that eye-tracking can be a valuable tool for assessing reading comprehension.

Eye-tracking technology can also be used with patients. Hence, eye-tracking can assess reading behaviour in real time, revealing subtle eye-movement biomarkers of dyslexia without requiring verbal responses and offering a non-invasive way to identify at-risk children (Rubino & Minden, 1973; Nilsson Benfatto *et al.*, 2016). Also, eye-tracking can be used with individuals on the autism spectrum to detect distinct attention patterns and provide insights into the comparative effectiveness of photographs, symbols, and human-produced easy-read documents (Yaneva *et al.*, 2015).

2.2 Surprisal in Language Processing

Previous studies on surprisal (the unexpectedness of a word given its context) observed that higher surprisal values are associated with increased processing difficulty of texts. Hence, in one work the researchers demonstrated that lexical surprisal significantly predicts reading times (Fernandez Monsalve *et al.*, 2012). Using both word-based (lexicalized) and POS-based (unlexicalized) models (PSGs and RNNs) trained on a large corpus, they found that higher surprisal leads to longer reading times, with notable spill-over effects from preceding words. Another study is related to reading times across languages (de Varda & Marelli, 2022). The researchers use mBERT to obtain bidirectional surprisal estimates from texts and analyze eye-tracking data from the MECO corpus covering 12 languages (Kuperman *et al.*, 2022). Their findings show that higher surprisal robustly predicts longer reading durations across several fixation measures (such as first fixation, gaze duration, and total reading time). Notably, the study reveals that these surprisal effects are stronger in native L1 reading than in non-native L2 reading, highlighting how language proficiency modulates reliance on contextual prediction. Yet another study examines how word predictability and contextual semantic coherence contribute to reading behaviour (Salicchi *et al.*, 2023). The authors compare several regression models : baseline lexical features alone, lexical features augmented with surprisal scores computed with GPT2-xl, semantic cosine similarity from SGNS or BERT, and a hybrid model including all predictors. The authors test the models on the GECO (Colman *et al.*, 2022) and Provo (Luke, 2022) eye-tracking corpora, and show that the hybrid model consistently outperforms those relying on a single factor. Notably, semantic relatedness based on BERT embeddings yields better predictions than static embeddings, highlighting that while surprisal reflects the predictability (syntagmatic relationships) of a word, semantic relatedness captures its contextual coherence (paradigmatic relationships).

In our research we want to investigate whether the surprisal measure can predict variations in reading

behaviour among individuals exposed to different types of texts. By correlating surprisal values with eye-tracking metrics we aim to determine if higher surprisal consistently corresponds to increased cognitive load and distinct reading patterns.

3 Dataset and Experimental Setup

For this study, we constructed a novel dataset of French texts that spans both general and medical domains. The dataset was built by leveraging excerpts from two corpora, CLEAR (Grabar & Cardon, 2018) and CAS (Grabar *et al.*, 2020), which include a variety of source materials such as Wikipedia articles [1], literature reviews, drug leeflets, and clinical cases. Our dataset is composed of three main categories of texts :

— Medical texts : three Wikipedia articles on medical topics (ulcer, obstetrics, and stroke).
— General-language texts : two Wikipedia articles on common topics (weekend and Camelot).
— Clinical cases : two clinical case reports that describe patients' symptoms, diagnoses, treatments, and follow-ups. Clinical cases are published by medical doctors in medical journals on various clinical issues (procedures, diagnosis, treatments...) of real or fake patients. Clinical cases are anonymized. They resemble hospital discharge summaries, describe typical clinical situations and are rich in medical terminology. The two clinical cases exploited are related to gastroenterology and toxicology.

In addition, all the texts have undergone a process of manual simplification following the guidelines outlined in (OCDE, 2015). This simplification was performed at three levels :

— syntactic : breaking down complex sentences into simpler, more digestible segments,
— lexical : replacing technical terms with synonyms, hyperonyms, or explicit definitions,
— semantic : enhancing the text with additional contextual information through examples, definitions, and clarifications.

Overall, simplified texts typically have a higher number of sentences due to syntactic simplification, achieved primarily through segmentation of complex sentences. Also, simplified versions often contain more words, reflecting lexical substitutions (synonyms, hyperonyms) and semantic clarifications (added definitions or examples). At the lexical and semantic levels, our simplified texts use various simplification strategies, such as providing explanations immediately after the technical term, before the term, or via integrated paraphrase. Table 1 shows an original–simplified pair illustrating one of these strategies : explanation of the term immediately after the term using parenthesis. Generally, we can observe several modifications in simplified versions : including parenthetical explanations *(analyses des bactéries éventuelles dans le sang)/(tests for possible bacteria in the blood)* after *hémocultures/blood cultures* ; term rephrasing (*ont montré la présence/showed the presence of*) instead of a more straightforward *ont permis d'isoler/made it possible to isolate* ; explicit addition of hypernym *bactérie/bacterium* to *Staphylococcus aureus*. Overall, we assume that, through simplification, medical information remains accurate : the technical term *hémocultures/blood cultures* is preserved (with explanation), and the bacteria name is maintained.

The texts are divided into two complementary sets A and B. If a given text appears in its original form in one set, its simplified version is included in the other set. Thus, each participant reads only one version of any text, preventing potential bias from familiarity with both versions of one text. Texts are presented to participants paragraph by paragraph to accommodate the presentation screen. To maintain participant attention, some screens include comprehension questions.

1. https://fr.wikipedia.org

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

TABLE 1 – Examples of simplification

The original clinical corpus comprises 653 words in total, the original general corpus 1,684 words, and the original medical corpus 2,906 words. Table 2 in the Appendix contains the full breakdown by screen and sentence.

A total of 23 native French-speaking participants took part in the eye-tracking experiment (ages ranging from 18 to 39, $Mean = 22.95$, $SD = 5.33$). Participants come from various social backgrounds - including students, doctoral students, and working professionals - but none have medical training. The experiment is conducted using a Tobii Pro Spectrum eye-tracking camera. According to the experimental protocol, each participant silently and naturally reads one of the two text sets. Overall, each participant reads 7 texts (two clinical cases, two general and three medical texts), while each text is read (or annotated) by 11 to 12 different participants.

To investigate the cognitive processes underlying reading, we employ several well-established eye-tracking indicators. These measures capture different stages of reading - from initial lexical processing (word recognition) to later integrative efforts (assimilation of the information read) - and are sensitive to variations in text complexity and processing difficulty. In our analysis, we focus on the following indicators (Hyönä & Kaakinen, 2019; Rayner, 1998) :

— *First-pass first fixation duration* (FPFFD) : the duration of the very first fixation on a word or region during its initial encounter. It reflects the immediate, early processing of the word ;
— *First-pass duration* (FPD) : the total time spent fixating on a word or region during the first pass (i.e., before any regression occurs). This measure captures the initial processing time required for lexical access and early comprehension ;
— *Regression-path duration* (RPD) : the sum of fixation durations from the moment a reader first enters a region until he moves past it to the right, including time spent on regressions to earlier parts of the text. It indicates processing difficulty that causes re-reading and re-analysis ;
— *Re-reading duration* (RRD) : the total time spent fixating on a word or region during re-readings (after the initial pass). This measure reflects additional processing or integration efforts when comprehension is challenged ;
— *Total duration of fixations* (TDF) : the cumulative duration of all fixations on a word or region, including both the first pass and any subsequent re-readings ;
— *Number of fixations* (NoF) : the total count of fixations on a word or region during reading.

4 Methodology

Our methodology relies on the notion of *surprisal*, an information-theoretic measure that quantifies how unexpected a word is given its preceding context (Hale, 2001; Levy, 2008). Mathematically, the surprisal of a word w given its context c is defined as : Surprisal$(w \mid c) = -\log_2 P(w \mid c)$, where $P(w \mid c)$ is the conditional probability of w based on the context c. This can also be understood as the logarithm of the ratio between the prefix probabilities computed before and after a word is processed. Higher surprisal values indicate that a word is less predictable, which is hypothesized to correlate with increased processing difficulty during reading.

To estimate these probabilities for French texts, we employ a pre-trained French GPT-2 model `gpt-fr-cased-base`[2]. This model was specifically trained on a large and diverse corpus of French text (1,017B parameters), making it well-suited for capturing French linguistic patterns and probabilities. As an autoregressive language model, it naturally produces the conditional probabilities needed for surprisal calculations. Although newer language models are available, GPT-2 remains a standard choice for surprisal estimation in psycholinguistic research due to its computational tractability and established validation against human reading data in prior studies. This model uses subword tokenization (e.g., Byte-Pair Encoding) to decompose words into smaller units. Even if this tokenization is not linguistically grounded, it enables to handle rare or morphologically complex words more effectively.

Then, we merge surprisal values computed for subtokens to get a single surprisal value per word. Specifically, if a word is split into subtokens $t_i, t_{i+1}, \ldots, t_j$, we sum the surprisal values for those subtokens to produce a single word-level surprisal. This approach sums surprisal values (measured in bits as $-\log_2(P)$), not probabilities themselves, which ensures mathematical validity since surprisal values are additive by definition. These word-level surprisal measures are aligned with eye-tracking data to explore how surprisal predictability affects reading behaviour observed with eye-tracking. For instance, we examine whether words with higher surprisal (i.e., lower predicted probability) are associated with eye-tracking indicators suggesting greater processing difficulty, like longer fixation durations or increased regression counts. In subsequent statistical analyses, correlations using Spearman methods (Spearman, 1904) are computed between surprisal values and eye-tracking measures, while controlling for factors such as text version (original vs. simplified) and text genre differences. We also employ LMEMs (Linear mixed effects models) to analyze how surprisal, text type, and simplification influence reading times while controlling for participant-specific variation. LMEMs are statistical models that account for both fixed effects (e.g., experimental manipulations) and random effects (e.g., individual variability) simultaneously. Specifically, we model reading measures (e.g., first-pass duration, total fixation times) as a function of surprisal (centered), text type (medical vs. clinical), and version (original vs. simplified), including all possible interactions as fixed effects.

5 Results and their Discussion

We present the results across three lines : global descriptive statistics, statistical correlation analysis between surprisal and eye-tracking measures, and LMEM analysis.

2. https://huggingface.co/asi/gpt-fr-cased-base

Descriptive Statistics. Tables 3 and 4 in the Appendix present the descriptive statistics for six eye-tracking measures across different text types (clinical, general, and medical) in both original and simplified versions. Overall, the data reveals several notable patterns. (1) There are evident processing differences between original and simplified texts : in most cases, simplified versions show reduced reading times across multiple measures compared to their original counterparts, suggesting improved readability, as has been observed in a previous study with experimental data comparable to ours (Grabar *et al.*, 2018). (2) Text type variations are also observed, with clinical texts generally exhibiting longer fixation durations and more fixations than general texts, indicating higher processing demands. (3) Regarding measure variability, regression-path duration (RPD) and re-reading duration (RRD) show considerably higher standard deviations compared to other measures, reflecting the variable nature of regression and re-reading behaviours among readers. (4) For fixation patterns, the mean number of fixations (NoF) ranges from 1.10 to 1.81 across different text types, with clinical texts generally requiring more fixations than general texts. Thus, the clinical text related to gastroenterology shows the most pronounced difference between original and simplified versions, with substantial reductions in all eye-tracking measures after simplification. For instance, the mean first-pass duration decreases from 197.88 ms to 116.79 ms, and the total duration of fixations from 404.38 ms to 247.74 ms. Medical texts show consistent but more moderate improvements with simplification, particularly for the *ulcer* text, where the total fixation duration is reduced from 282.05 ms to 245.39 ms after simplification. Finally, general texts display the smallest differences between original and simplified versions, suggesting that these texts may be more accessible in their original form compared to specialized medical and clinical texts. By comparing these descriptive statistics across text types and versions, we can gain a preliminary understanding of how linguistic complexity and text simplification may affect reading behaviour.

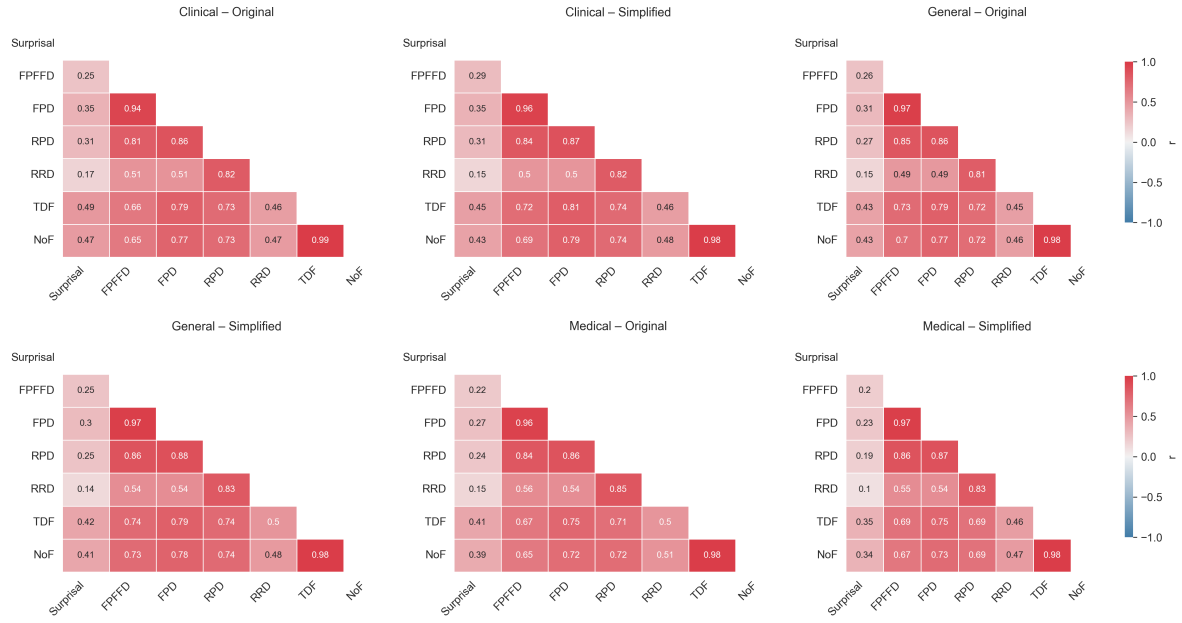

FIGURE 2 – Correlation of eye-tracking measures with surprisal for the three text types and their two versions. 1st line : clinical original, clinical simplified, general original. 2nd line : general simplified, medical original, medical simplified.

Correlation. The Spearman correlation matrices (Figure 2) reveal systematic relationships between word surprisal and eye-tracking measures across text types, confirming that less predictable words

also require more effort during reading. The strength of these relationships varies by text type : clinical texts show the strongest correlations (first-pass duration : $r = 0.36$; total fixation duration : $r = 0.47$), followed by general texts (first-pass duration : $r = 0.30$; total fixation duration : $r = 0.42$), with medical texts showing the weakest correlations (first-pass duration : $r = 0.25$; total fixation duration : $r = 0.38$). Re-reading duration consistently shows the weakest correlation with surprisal across all text types ($r = 0.13$-0.16), suggesting that later reading processes are driven by factors beyond local word predictability. Strong intercorrelations exist among reading measures themselves, with first-pass measures highly correlated ($r > 0.95$), as are total fixation duration and number of fixations ($r > 0.97$). Regression-path duration correlates with both early and late measures ($r > 0.82$), suggesting it captures both initial processing of words and difficulties with the integration of words. These findings support information-theoretic accounts of language processing, with the stronger effect in clinical texts potentially reflecting challenges in processing specialized terminology. The weaker correlation with re-reading duration suggests that multiple reading stages may be differently affected by word predictability.

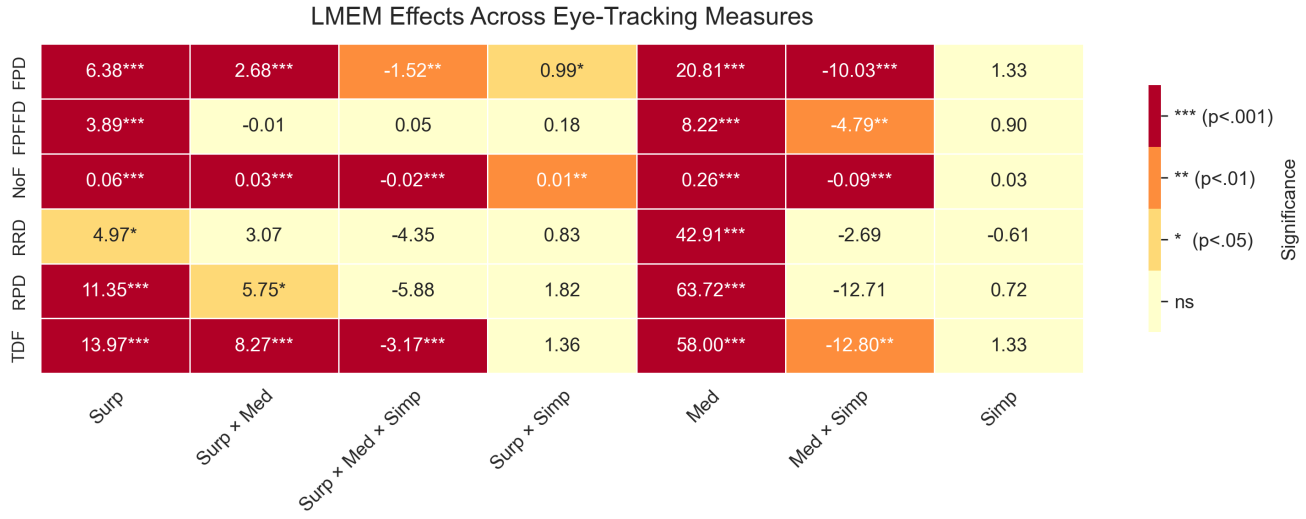

FIGURE 3 – Coefficient values and significance levels for Linear Mixed Effects Models across six eye-tracking measures.

Surprisal Effects Across Eye-Tracking Measures. To obtain a comprehensive understanding of how surprisal affects reading processes, we conduct Linear Mixed Effects Models (LMEMs) on the six eye-tracking measures. These measures capture different aspects of reading, from early word recognition to later integration processes. All models include surprisal, text type, and simplification as fixed effects, with all possible interactions, while controlling for participant variability as a random effect. The analyses (Figure 3) reveal several consistent patterns across measures. First, surprisal significantly predicts reading times across all measures (all $p < 0.001$), confirming that less predictable words consistently require more processing resources. This effect is particularly strong for total duration of fixations ($\beta = 13.970$) and regression-path duration ($\beta = 11.349$), compared to first-pass measures (first-pass first fixation duration (FPFFD) and first-pass duration (FPD)). Second, medical texts consistently show longer reading times than clinical texts across all measures, with the largest effects observed for regression-path duration ($\beta = 63.718$) and total duration of fixations ($\beta = 57.998$). Interestingly, simplification alone does not show a significant main effect for most measures, suggesting that simplification benefits are context-dependent. It is important to note that, in

our case, simplification does not mean eliminating difficult medical terms. The simplified versions still contain technical terminology, but these terms are accompanied by explanations or synonyms. This means that readers still encounter specialized vocabulary, but with additional support. Participants often skim or skip over dense jargon in short clinical cases, which can lead to lower measured fixation times per word than in longer, more uniformly processed medical texts. Indeed, the purpose of simplification is to find the most effective strategies for making text more readable while preserving essential information.

The interaction effects reveal more nuanced patterns across different reading measures. Hence, the interaction between text type and surprisal is significant for most measures, indicating that medical texts show stronger sensitivity to surprisal than clinical texts. For example, in first-pass duration, surprisal increases reading time by 6.38ms per unit in clinical texts, and by up to 9.06ms in medical texts ($\beta = 2.681$, $p < 0.001$). This pattern is even more pronounced for total duration of fixations ($\beta = 8.270$, $p < 0.001$). We also find a significant link between text versions for several measures, most notably first-pass duration ($\beta = -10.028$, $p < 0.001$), total duration of fixations ($\beta = -12.799$, $p = 0.001$), and number of fixations ($\beta = -0.086$, $p < 0.001$). This indicates that simplification substantially reduces reading times specifically for medical texts compared to clinical texts.

The most theoretically significant finding is the three-way interaction between surprisal, text type, and simplification, which is significant for first-pass duration ($\beta = -1.521$, $p = 0.008$), total duration of fixations ($\beta = -3.175$, $p < 0.001$), and number of fixations ($\beta = -0.016$, $p < 0.001$). This interaction reveals that simplification reduces the effect of surprisal specifically in medical texts. For instance, in original medical texts, surprisal increases first-pass duration by 9.06ms per unit, but in simplified medical texts, this effect decreases to 7.54ms. The effect is not significant for early measures like first-pass first fixation duration ($p = 0.898$) or later measures like re-reading duration ($p = 0.217$), suggesting that simplification impact on surprisal processing occurs primarily during initial word processing but not during the very earliest stages of word recognition or later re-analysis.

Comparison of early and late measures reveals that the effects of surprisal and the benefits of simplification are most pronounced in measures that capture overall reading behaviour (total duration and number of fixations) rather than just initial processing. This suggests that simplification not only facilitates initial word recognition, but also reduces the need for multiple fixations and extended processing time. Our findings, showing that surprisal significantly predicts reading times across all measures, align with established works (Fernandez Monsalve *et al.*, 2012). However, our study extends previous research through the novel discovery that simplification benefits vary significantly by text type, with the three-way interaction revealing that simplification specifically reduces surprisal effects in medical texts but not clinical texts - a finding that refines our understanding of how text adaptation strategies should be tailored to specific domains.

6 Conclusion and Future Work

This study examined how text simplification affects reading processes across different text types through a combination of eye-tracking methodology and surprisal analysis, which quantifies how unexpected a word is given its preceding context. We created a corpus of French general, clinical, and medical texts in both original and simplified versions, annotated with comprehensive eye-tracking measurements from 23 participants. This corpus will be made publicly available, providing a resource for researchers investigating reading processes, and domain-specific language processing.

Our analysis across multiple eye-tracking measures revealed that surprisal significantly predicts reading behaviour in all text types, with medical texts showing heightened sensitivity to unexpected words. We found that simplification benefits are not uniform across text types. While simplification had minimal effects on clinical texts, it substantially reduced reading times for medical texts and mitigated the impact of surprisal specifically in these texts. This differential effect was consistent across multiple eye-tracking measures, particularly in those capturing overall reading behaviour rather than just initial word recognition.

We plan to expand the corpus by increasing both the number of annotated texts and the number of participants. Particularly important will be the inclusion of pathology speech students as an additional participant group, which will allow to examine how domain expertise modulates the effects of surprisal and simplification. This comparison could reveal whether experts process specialized terminology differently and whether simplification benefits vary with expertise level.

We note that simplified texts in our corpus tend to be longer and contain a higher proportion of content words, which could influence surprisal estimates. We plan to apply post-hoc POS-based filtering of function versus content words in future work to isolate this effect.

We also intend to employ multiple language models for surprisal calculation to ensure robustness and to investigate whether different models capture different aspects of predictability in specialized texts. Furthermore, future analyses will explore additional psycholinguistic phenomena (such as spillover effects (Shvartsman *et al.*, 2014), entropy (Arora *et al.*, 2022), perplexity (Jurafsky & Martin, 2025), etc.). These more detailed analyses may reveal subtler aspects of how predictability influences reading across different text types and simplification versions.

Acknowledgement

This work was partially funded by the French National Agency for Research (ANR) as part of the CLEAR project (Communication, Literacy, Education, Accessibility, Readability), ANR-17-CE19-0016-01. The authors would like also to thank the reviewers for their helpful comments and questions that permitted to improve the overall quality of the paper.

Références

AHMED Y., FRANCIS D. J., YORK M., FLETCHER J. M., BARNES M. & KULESZ P. (2016). Validation of the direct and inferential mediation (dime) model of reading comprehension in grades 7 through 12. *Contemporary Educational Psychology*, **44-45**, 68–82. DOI : 10.1016/j.cedpsych.2016.02.002.

ARORA A., MEISTER C. & COTTERELL R. (2022). Estimating the entropy of linguistic distributions. In S. MURESAN, P. NAKOV & A. VILLAVICENCIO, Éds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, p. 175–195, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-short.20.

BROWN J. (2008). How clinical communication has become a core part of medical education in the uk. *Medical Education*, **42**(3), 271–278. DOI : https ://doi.org/10.1111/j.1365-2923.2007.02955.x.

COLMAN T., FONTEYNE M., DAEMS J., DIRIX N. & MACKEN L. (2022). GECO-MT : The ghent eye-tracking corpus of machine translation. In N. CALZOLARI, F. BÉCHET, P. BLACHE, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 29–38, Marseille, France : European Language Resources Association.

CROMLEY J. G., SNYDER-HOGAN L. E. & LUCIW-DUBAS U. A. (2010). Reading comprehension of scientific text : A domain-specific test of the direct and inferential mediation model of reading comprehension. *Journal of Educational Psychology*, **102**(3), 687–700. DOI : 10.1037/a0019452.

DE VARDA A. & MARELLI M. (2022). The effects of surprisal across languages : Results from native and non-native reading. In Y. HE, H. JI, S. LI, Y. LIU & C.-H. CHANG, Éds., *Findings of the Association for Computational Linguistics : AACL-IJCNLP 2022*, p. 138–144, Online only : Association for Computational Linguistics. DOI : 10.18653/v1/2022.findings-aacl.13.

DEUTSCH T., JASBI M. & SHIEBER S. (2020). Linguistic features for readability assessment. In J. BURSTEIN, E. KOCHMAR, C. LEACOCK, N. MADNANI, I. PILÁN, H. YANNAKOUDAKIS & T. ZESCH, Éds., *Proc of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, p. 1–17, Seattle, WA, USA → Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.bea-1.1.

DUCHOWSKI A. (2007). *Eye Tracking Methodology : Theory and Practice*. DOI : 10.1007/978-1-84628-609-4.

EKLICS K. & FEKETE J. (2024). From a simulated patient interview to a case presentation.

FERNANDEZ MONSALVE I., FRANK S. L. & VIGLIOCCO G. (2012). Lexical surprisal as a general predictor of reading time. In W. DAELEMANS, Éd., *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, p. 398–408, Avignon, France : Association for Computational Linguistics.

FLESCH R. (1948). A new readability yardstick. *Journ Appl Psychol*, **23**, 221–233.

FRANÇOIS T. & FAIRON C. (2012). An "AI readability" formula for French as a foreign language. In J. TSUJII, J. HENDERSON & M. PAŞCA, Éds., *Proc of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, p. 466–477, Jeju Island, Korea : Association for Computational Linguistics.

GONZÁLEZ-GARDUÑO A. V. & SØGAARD A. (2017). Using gaze to predict text readability. In J. TETREAULT, J. BURSTEIN, C. LEACOCK & H. YANNAKOUDAKIS, Éds., *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, p. 438–443, Copenhagen, Denmark : Association for Computational Linguistics. DOI : 10.18653/v1/W17-5050.

GOODING S., BERZAK Y., MAK T. & SHARIFI M. (2021). Predicting text readability from scrolling interactions. In A. BISAZZA & O. ABEND, Éds., *Proceedings of the 25th Conference on Computational Natural Language Learning*, p. 380–390, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.conll-1.30.

GOODKIND A. & BICKNELL K. (2018). Predictive power of word surprisal for reading times is a linear function of language model quality. In A. SAYEED, C. JACOBS, T. LINZEN & M. VAN SCHIJNDEL, Éds., *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, p. 10–18, Salt Lake City, Utah : Association for Computational Linguistics. DOI : 10.18653/v1/W18-0102.

GRABAR N. & CARDON R. (2018). Clear – simple corpus for medical French. In *Workshop on Automatic Text Adaption (ATA)*, p. 1–11.

GRABAR N., DALLOUX C. & CLAVEAU V. (2020). CAS : corpus of clinical cases in French. *Journal of BioMedical Semantics*, **11**(1), 1–7.

GRABAR N., FARCE E. & SPARROW L. (2018). Study of readability of health documents with eye-tracking approaches. In A. JÖNSSON, E. RENNES, H. SAGGION, S. STAJNER & V. YANEVA, Éds., *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, p. 10–20, Tilburg, the Netherlands : Association for Computational Linguistics. DOI : 10.18653/v1/W18-7003.

GUNNING R. (1973). *The art of clear writing*. New York, NY : McGraw Hill.

HALE J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

HYÖNÄ J. & KAAKINEN J. (2019). *Eye Movements During Reading*, p. 239–274. DOI : 10.1007/978-3-030-20085-5_7.

JURAFSKY D. & MARTIN J. H. (2025). *Speech and Language Processing : An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. 3rd édition. Online manuscript released January 12, 2025.

KENNEDY A., PYNTE J., MURRAY W. S. & PAUL S.-A. (2013). Frequency and predictability effects in the dundee corpus : An eye movement analysis. *Quarterly Journal of Experimental Psychology*, **66**(3), 601–618. PMID : 22643118, DOI : 10.1080/17470218.2012.676054.

KIM H., GORYACHEV S., ROSEMBLAT G., BROWNE A., KESELMAN A. & ZENG-TREITLER Q. (2007). Beyond surface characteristics : a new health text-specific readability measurement. In *AMIA Annual Symposium Proceedings*, volume 2007, p. 418–422 : American Medical Informatics Association. PMID : 18693870 ; PMCID : PMC2655856.

KUPERMAN V., SIEGELMAN N., SCHROEDER S., ACARTURK C., ALEXEEVA S., AMENTA S., BERTRAM R., BONANDRINI R., BRYSBAERT M., CHERNOVA D., FONSECA S., DIRIX N., DUYCK W., FELLA A., FROST R., GATTEI C., KALAITZI A., LÕO K., MARELLI M. & USAL K. (2022). Text reading in english as a second language : Evidence from the multilingual eye-movements corpus. *Studies in Second Language Acquisition*, **45**, 1–35. DOI : 10.1017/S0272263121000954.

LEVY R. (2008). Expectation-based syntactic comprehension. *Cognition*, **106**(3), 1126–1177. DOI : 10.1016/j.cognition.2007.05.006.

LEVY R. (2013). Memory and surprisal in human sentence comprehension. In *Sentence Processing*, p. 78–114. Psychology Press, 1st edition édition.

LOWDER M. W., CHOI W., FERREIRA F. & HENDERSON J. M. (2018). Lexical predictability during natural reading : Effects of surprisal and entropy reduction. *Cognitive Science*, **42**(4), 1166–1183. DOI : 10.1111/cogs.12597.

LUKE S. G. (2022). The provo corpus : A large eye-tracking corpus with predictability norms. Retrieved from osf.io/sjefs.

MARTINC M., POLLAK S. & ROBNIK-ŠIKONJA M. (2021). Supervised and unsupervised neural approaches to text readability. *Computational Linguistics*, **47**(1), 141–179. DOI : 10.1162/coli_a_00398.

MÉZIÈRE D. C., YU L., REICHLE E. D., VON DER MALSBURG T. & MCARTHUR G. (2023). Using eye-tracking measures to predict reading comprehension. *Reading Research Quarterly*, **58**(3), 425–449.

NADEEM F. & OSTENDORF M. (2018). Estimating linguistic complexity for science texts. In ACL, Éd., *Workshop on Innovative Use of NLP for Building Educational Applications*, p. 45–55, New Orleans, LA. DOI : https ://doi.org/10.18653/v1/W18-0505.

NILSSON BENFATTO M., ÖQVIST SEIMYR G., YGGE J., PANSELL T., RYDBERG A. & JACOBSON C. (2016). Screening for dyslexia using eye tracking during reading. *PLoS ONE*, **11**(12), e0165508.

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

PITLER E. & NENKOVA A. (2008). Revisiting readability : A unified framework for predicting text quality. In M. LAPATA & H. T. NG, Éds., *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, p. 186–195, Honolulu, Hawaii : Association for Computational Linguistics.

RADACH R. & KENNEDY A. (2013). Eye movements in reading : Some theoretical context. *Quarterly journal of experimental psychology (2006)*, **66**. DOI : 10.1080/17470218.2012.750676.

RAYNER K. (1998). Eye movements in reading and information processing : 20 years of research. *Psychological Bulletin*, **124**(3), 372–422. DOI : 10.1037/0033-2909.124.3.372.

RUBINO C. & MINDEN H. (1973). Analysis of eye-movements in children with reading disability. *Cortex*, **9**, 217–220.

SALICCHI L., CHERSONI E. & LENCI A. (2023). A study on surprisal and semantic relatedness for eye-tracking data prediction. *Frontiers in Psychology*, **14**, 1112365. DOI : 10.3389/fpsyg.2023.1112365.

SALVUCCI D. D. & GOLDBERG J. H. (2000). Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications*, ETRA '00, p. 71–78, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/355017.355028.

SHVARTSMAN M., LEWIS R. & SINGH S. (2014). Computationally rational saccadic control : An explanation of spillover effects based on sampling from noisy perception and memory. In *Proceedings of the Fifth Workshop on Cognitive Modeling and Computational Linguistics*, p. 1–9, Baltimore, Maryland, USA : Association for Computational Linguistics. DOI : 10.3115/v1/W14-2001.

SINGH A. D., MEHTA P., HUSAIN S. & RAJAKRISHNAN R. (2016). Quantifying sentence complexity based on eye-tracking measures. In D. BRUNATO, F. DELL'ORLETTA, G. VENTURI, T. FRANÇOIS & P. BLACHE, Éds., *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, p. 202–212, Osaka, Japan : The COLING 2016 Organizing Committee.

SPEARMAN C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, **15**(1), 72–101.

VILPPU H., MIKKILÄ-ERDMANN M., SÖDERVIK I. & ÖSTERHOLM MATIKAINEN E. (2016). Exploring eye movements of experienced and novice readers of medical texts concerning the cardiovascular system in making a diagnosis. *Anatomical Sciences Education*, **10**(1), 23–33.

YANEVA V., TEMNIKOVA I. & MITKOV R. (2015). Accessible texts for autism : An eye-tracking study. In ACM, Éd., *Int ACM SIGACCESS Conference on Computers & Accessibility*, p. 49–57.

ZHENG J. & YU H. (2017). Readability formulas and user perceptions of electronic health records difficulty : A corpus study. *Journal of Medical Internet Research*, **19**(3), e59. DOI : 10.2196/jmir.6962.

7 Appendix

TABLE 2 – Comparison of Original and Simplified Texts

text_type	text_name	version	total_screens	total_sentences	total_words
clinical	toxico	original	4	19	398
		simplified	5	29	469
clinical	gastro	original	3	13	255
		simplified	3	13	336
general	weekend	original	9	31	844
		simplified	9	49	811
general	camelot	original	8	42	840
		simplified	8	58	880
medical	obstetrique	original	12	57	1104
		simplified	12	65	1202
medical	avc	original	3	10	276
		simplified	3	22	328
medical	ulcere	original	15	77	1526
		simplified	15	92	1551

TABLE 3 – Descriptive Statistics of Eye-Tracking Measures (Part I) : FPFFD = First-pass first fixation duration ; FPD = First-pass duration ; RPD = Regression-path duration.

text_type	text_name	version	FPFFD		FPD		RPD	
			mean	std	mean	std	mean	std
clin.	toxico	or.	101.21	128.87	151.34	266.28	331.09	1500.39
		simp.	112.13	130.73	159.34	242.63	287.43	870.36
clin.	gastro	or.	115.26	132.53	197.88	347.11	410.41	1438.92
		simp.	93.27	118.73	116.79	169.15	251.28	1086.57
gen.	camelot	or	101.68	128.99	131.28	194.47	259.19	1040.46
		simp.	93.37	123.81	114.97	173.14	242.14	1060.20
gen.	weekend	or.	93.38	121.40	112.32	159.28	230.88	904.47
		simp.	100.30	120.81	123.28	166.54	234.55	714.41
med.	avc	or	95.30	123.90	119.36	173.99	305.87	1197.42
		simp	94.08	124.07	112.69	170.27	293.71	1351.04
med.	obstetrique	or	97.05	124.70	132.02	206.52	300.07	1260.04
		simp	93.84	124.99	125.84	208.63	277.25	1244.52
med.	ulcere	or	104.85	129.98	140.09	210.28	283.81	1164.95
		simp.	93.79	126.52	114.28	177.77	248.94	1186.59

TABLE 4 – Descriptive Statistics of Eye-Tracking Measures (Part II) : RRD = Re-reading duration ; TDF = Total duration of fixations ; NoF = Number of fixations.

text_type	text_name	version	RRD		TDF		NoF		
			mean	std	mean	std	count	mean	std
clin.	toxico	or.	179.74	1463.75	328.82	470.65	4378	1.50	1.89
		simp.	128.10	829.77	285.45	335.17	5628	1.32	1.37
clin.	gastro	or.	212.53	1396.26	404.38	474.86	3060	1.81	1.88
		simp.	134.50	1060.52	247.74	311.72	3696	1.20	1.38
gen.	camelot	or.	127.92	1011.99	257.49	295.63	9240	1.20	1.21
		simp.	127.16	1030.35	240.90	307.84	8800	1.13	1.29
gen.	weekend	or.	118.55	873.70	227.47	259.78	9281	1.10	1.16
		simp.	111.26	683.13	233.03	245.22	9732	1.13	1.09
med.	avc	or.	186.51	1173.32	304.49	316.07	3312	1.46	1.40
		simp.	181.02	1314.81	291.10	347.67	3608	1.37	1.42
med.	obstetrique	or.	168.05	1220.99	296.61	379.03	12144	1.41	1.65
		simp.	151.41	1217.82	275.50	323.60	14424	1.31	1.36
med.	ulcere	or.	143.72	1137.19	282.05	313.80	18312	1.30	1.29
		simp.	134.66	1158.02	245.39	313.17	17061	1.15	1.30