A French Eye-Tracking Corpus of Original and Simplified Medical, Clinical, and General Texts - 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

Eye tracking offers an objective window on real-time cognitive processing of information being read: longer fixations, more regressions, and wider pupil dilation reliably index linguistic difficulty. Yet, there is a paucity of the available corpora annotated with eye-tracking features. We introduce in this paper the FETA corpus – a French Eye-TrAcking corpus [1]. It combines three types of texts (general, medical and clinical) in two versions (original and manually simplified). These texts are read by 46 participants, from which we collect eye-tracking data through dozens of eye-tracking features.

1 Introduction

Literacy, when reading general purpose and health-related information, depends critically on a reader's ability to understand such information (Eklics and Fekete, 2024; Brown, 2008). For instance, patients and the general public consult health-related sources – diagnosis leaflets, drug leaflets, web portals – on a daily basis (Fox, 2014), yet these materials are often written at a level well above the average reading proficiency (McCray, 2005). Text simplification (lexical, syntactic, or semantic) has therefore become a central strategy for improving accessibility (Saggion, 2017), but robust *evaluation* of simplification quality remains challenging (Grabar and Saggion, 2022). Eye tracking offers an objective window on real-time cognitive processing: longer fixations, more regressions, and wider pupil dilation reliably index linguistic difficulty (Singh et al., 2016). Employing gaze data to detect complex fragments can guide automatic or human adaptation of text, ultimately facilitating patient-oriented communication. Despite the maturity of eye-movement research in English (Hollenstein et al., 2022; Kuperman et al., 2020; Cop et al., 2017), French still lacks an openly available, large-scale corpus that combines (i) general-language and technical texts, (ii) technical medical and clinical texts, (iii) parallel simplified versions of texts, and (iv) fine-grained eye-tracking annotations.

To fill in this gap we introduce the FETA (French Eye-TrAcking) corpus designed through an eye-tracking experiment that captures reading behaviour across three text types (medical, clinical, general), each paired with manually produced lexical, syntactic, and semantic simplifications. Thus, our work makes several key contributions: it combines three types of texts (general, medical and clinical) in two versions (original and simplified), and gathers eye-tracking data from 46 participants.

In what follows, we describe the corpus texts (original documents and creation of their simplified versions) in Section 2. In Section 3, we describe the experimental protocol and participants. Section 4 is dedicated to the pre-processing of the eye-tracking data and extraction of eye-tracling features. Section 5 introduces the description of the eye-tracking-annotated corpus: metrics for the texts and eye-tracking features. Finally, we conclude in Section 6 and drow up some limitations in Section 7.

2 Corpus Construction

Our study employs a balanced, French-language corpus consisting of 16 texts sourced from two publicly available resources: the CLEAR corpus (Grabar and Cardon, 2018), corpus of clinical cases (Grabar et al., 2020), and general texts from Wikipedia. The set of 14 texts processed spans three text types: general-language articles from Wikipedia present common topics like *Weekend* or *Camelot*, medical-language articles from Wikipedia describe some specialized topics like *Vascular Cerebral Accidents* or *Obstetrics*, and clin-

[1] https://hdl.handle.net/11403/feta

Proceedings of First International Workshop on Gaze Data and NLP associated with RANLP 2025,
pages 37–43, Varna, Bulgaria, Sep 12, 2025.
https://doi.org/10.26615/978-954-452-104-2-005

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.)
Un cathéter a été posé. (A catheter was inserted.)	Un cathéter a été posé *pour évacuer l'urine.* (A catheter was inserted *to drain the urine.*)

Table 1: Examples of manual simplification presented in the inline original format, with translations.

ical cases from toxicology and gastrology. Clinical cases describe symptoms, diagnoses, treatments, and follow-ups for individual patients or small cohorts. Their narrative structure resembles hospital discharge summaries and is densely packed with specialised terminology and reasoning about therapeutic choices. Such texts impose a high cognitive load on lay readers who must comprehend health information relevant to themselves or their relatives. Original clinical texts contain 653 words, general texts contain 1,684 words, and medical texts 2,906 words. A detailed breakdown by screen and sentence is provided in Table 4 (Appendix).

To facilitate controlled eye-tracking experiments, we partitioned the 14 texts into two equally balanced *sets*, *Set 1* and *Set 2*, each containing a uniform mix of medical articles, clinical cases, and general texts, thereby equalising topic distribution and baseline difficulty across sets. Each text has been manually simplified as explained in Section 2.1 and examplified in Table 1. Then, we compose two presentation *versions*:

- *Version A*: half the texts appear in their original form, the remainder in simplified form.
- *Version B*: the original/simplified assignment is reversed, creating a mirror of Version A.

Random assignment to Set 1 or 2 gave each participant one version per text, preserving counterbalancing and single exposure

The primary aim of the experiment is to record eye-tracking indicators during natural text reading. To complement these gaze data with a behavioural measure of comprehension, we administer short multiple-choice questions after selected text segments. Each question pertains to the segment that has just been read, and participants respond by choosing *True*, *False*, or *I don't know*. To keep the reading experience as natural as possible and to minimise task interruption, comprehension questions are presented for only a random subset of segments.

2.1 Simplification Pipeline

All texts were manually simplified in respect with the plain-language recommendations (OCDE, 2015) at syntactic, lexical and semantic levels, as examplified in Table 1.

Syntactic level. Syntactic simplification aimed to reduce structural complexity by transforming embedded and multi-clause constructions into shorter, clause-minimal units. Where possible, passive constructions were rewritten in the active voice to make sentence roles (agent, action, patient) more explicit.

In addition, we prioritized the use of direct (subject–verb–object) word order to avoid ambiguity, clarified negations, and systematically avoided gerunds and past participial forms. As a result, simplified versions contain more sentences than their originals.

Lexical level. Lexical simplification involved replacing domain-specific or low-frequency terms with more accessible alternatives to improve comprehensibility. Different strategies were applied: **i)** high-frequency synonyms were used when semantic precision could be maintained, **ii)** hypernyms, or more general terms, were substituted for complex medical terminology, and **iii)** in-text definitions were inserted in parentheses directly after specialised terms to support interpretation. These strategies aimed to retain the intended meaning while lowering lexical complexity for readers without specialised knowledge. In many cases, the original term was kept alongside its explanation to aid familiarity and consistency.

Semantic level. Semantic simplification focused on enriching the text with contextual information to make implicit knowledge more explicit. This was especially important in clinical texts, where technical discourse often assumes prior medical knowledge that general readers may not possess. The goal was to reduce inferential effort by clarifying relationships, causes, effects, and by defining specialized concepts in context. Several semantic strategies were applied: **i)** causal or descriptive links were added to explain the function or consequence of a condition, like in the third example in Table 1, in which the role of catheter is explained; **ii)** integrated paraphrases combined description and terminology to bridge gaps in understanding. These modifications clarify the meaning of complex medical expressions and also anchor them in relatable concepts.

Overall, the simplified corpus exhibits (i) more sentences through syntactic segmentation, (ii) sometimes more lexemes through lexical substitutions, and (iii) richer contextual clues through semantic elaboration. Hence, the material remains fair to the original meaning but is cognitively easier for non-specialist readers.

3 Experimental Protocol and Participants

3.1 Experimental Protocol

The protocol is composed of several steps:

Pre-screening (online). Prior to scheduling, each participant completed a form collecting demographic data (age, gender, highest education), ocular health information (e.g. myopia, astigmatism, corrective lenses), reading habits, and informed-consent details about the study.

Day-of self-evaluation. On arrival, participants filled out a two-pages, self-assessment questionnaire on the perceived difficulty of understanding medical information in daily life, using a four-point Likert scale: *very easy/easy/difficult/very difficult*.

Set-up and calibration. Gaze was recorded with a Tobii Pro Spectrum eye tracker sampling at 600 Hz. Text stimuli were presented on a 24-inch monitor at a native resolution of 2880 × 1620 px; Participants were seated 60 cm from the display (adjusted by $\pm$ 5 cm to accommodate height and optimise calibration), as on Figure 1. A random five-point calibration was accepted when accuracy and precision thresholds of $0.5°$ and $0.2°$, respectively, were met. Calibration quality was manually inspected, and participants who marginally exceeded these thresholds were retained if visual inspection confirmed stable gaze traces.

Figure 1: Experiment set-up and calibration.

Familiarisation block contained several slides: *Slide 1*: introductory instructions; *Slide 2*: a short, easy text common to all participants; *Slides 3–4*: two comprehension questions on this text, answered aloud (*True/False/I don't know*).

Main reading block. Each participant was randomly assigned to *Set 1* or *Set 2* and to *Version A* or *Version B*. The block comprised in average 59 slides: original or simplified texts according to the set–version counter-balancing scheme. Slides advanced via a mouse click. All comprehension questions were to be answered loudly.

Mid-session break. After the first timeline (half of the slides), participants took a short pause. A second five-point recalibration followed the break.

Second timeline and debrief. The remaining slides were presented, after which participants answered some oral questions on perceived text difficulty and comprehension ease. The entire experiment lasted about 50 to 70 minutes, depending on how the participant read.

3.2 Participant Demographics

Forty-six native French speakers (32 women, 14 men; age range 18–43 years, $M = 23.3$, $SD = 6.7$) took part in the study. Their educational backgrounds were diverse but none held a medical or healthcare qualification. All reported normal or corrected-to-normal vision. Participants received an honorarium of €12. To balance exposure to text conditions, they were randomly assigned to four counter-balanced groups: Set 1-A (n=12), Set 1-B (n=11), Set 2-A (n=11), and Set 2-B (n=12). More detailed information is provided in the Table 3.

4 Pre-processing and Annotation

Text–Gaze Alignment. Text presentation and word-level AOIs (one per word) were handled automatically in Tobii Pro Lab. Fixations were matched to those AOIs within the software, eliminating custom tokenization or manual ID assignment.

Feature Extraction and Normalisation. Eye-movement events were classified in Tobii Pro Lab using the I-VT (Velocity-Threshold Identification) algorithm with the following settings:
Eye selection: average of both eyes.
Noise reduction: moving median (window = 3 samples).
Velocity calculator: window length = 20 ms.
I-VT threshold: 30 deg/s.
Fixation merging: max. gap = 75 ms; max. angle = 0.5°.
Discard short fixations: min. fix. duration = 60 ms.

5 Dataset Statistics

5.1 Text Metrics

Table 4 in the Appendix compares original and simplified versions for each text (number of screens, sentences, and words). We can see that, across the corpus, simplification primarily increased the number of sentences due to syntactic simplification, with more modest changes in word counts.

By domain, clinical texts rose from 32 to 42 sentences (+31.3%) and from 653 to 805 words (+23.3%). General texts showed a strong sentence increase (73 → 107, +46.6%) but virtually no change in word count (1 684 → 1 691, +0.4%), reflecting many short sentence splits without added explanations. Medical texts increased from 144 to 179 sentences (+24.3%) and from 2,906 to 3,081 words (+6.0%).

5.2 Gaze Metrics

For every word–participant pair we release ten eye-movement features. The full list of features is in Appendix 8.1.

Figure 2 shows the difference in reading original and simplified clinical text for the *Total_duration_of_fixations* feature.

Table 2 reports the median and inter-quartile range (IQR) of each feature, aggregated by domain (clinical, medical, general) and by version (original, simplified). The headline metric, *Total Fixation Duration* (TFD), shows a clear reduction in reading effort (Figure 3): in clinical texts, the median TFD drops –17%; in medical texts –6%; in general texts –3%.

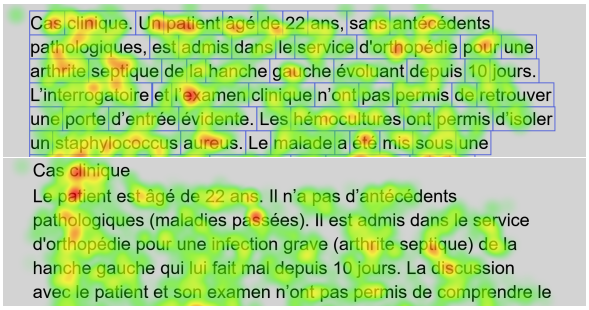

Figure 2: Original clinical case (top) and Simplified clinical case (bottom)

Table 2: Median and inter-quartile range (IQR) of word-level total fixation duration (in milliseconds).

Domain	Version	Median (ms)	IQR (ms)
Clinical	Original	225	435
	Simplified	187	342
Medical	Original	203	375
	Simplified	193	357
General	Original	183	333
	Simplified	182	323

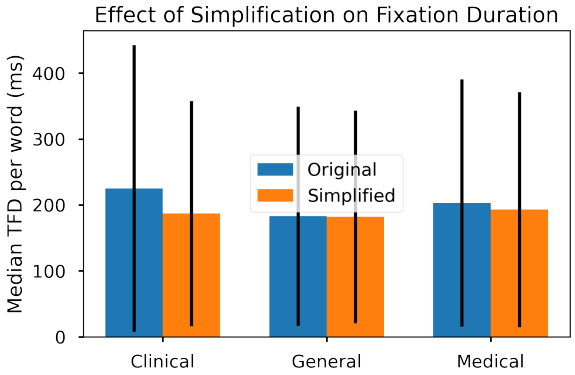

Figure 3: Median word-level fixation duration by domain and version.

6 Conclusion and Future Work

We introduced the FETA (French Eye-TrAcking) corpus, built with general-language and health documents in two versions (technical and manually simplified), thus covering several topics and genres. This corpus was read by 46 participants, through a precise experimental protocol. This permitted to collect several eye-tracking features, of which 10 are provided as part of the FETA corpus.

Besides, eye-tracking data are also being collected from speech-language pathology students,

which will permit to compare the reading from non-specialised and specised participants.

7 Limitations

Although all recordings met our calibration criteria, occasional attentional shifts or transient tracker losses may have gone undetected. Consequently, some fixations – especially at line breaks – could be mis-assigned, lowering word-level accuracy.

To preserve natural reading, only eleven multiple-choice questions were randomly inserted across the 50 slides. This design prevents us from verifying comprehension on every individual slide, which means that local misunderstandings might therefore remain unnoticed.

Due to Tobii Pro Lab's limitations in processing large datasets, raw data export proved challenging. We will include additional eye-tracking features and raw data as data processing continues.

8 Ethical Considerations

Participation in this study is voluntary, with informed consent obtained from all participants, ensuring compliance with the European General Data Protection Regulation (EU) 2016/679 and the modified French Data Protection Act of January 6, 1978. All personal data collected in the course of this research are anonymized to protect participant privacy and are accessible only by the designated project manager. This study has been registered in the University of Lille registry under reference 2022-075, affirming our commitment to upholding the highest standards of data protection and participant rights.

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.

References

Jo Brown. 2008. How clinical communication has become a core part of medical education in the uk. *Medical Education*, 42(3):271–278.

Uschi Cop, Nicolas Dirix, Denis Drieghe, and Wouter Duyck. 2017. Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods*, 49(2):602–615.

Kata Eklics and Judit Fekete. 2024. From a simulated patient interview to a case presentation.

Susannah Fox. 2014. The social life of health information. Technical report, Pew Internet & American Life Project, Washington DC.

Natalia Grabar and Rémi Cardon. 2018. CLEAR – simple corpus for medical French. In *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, pages 3–9, Tilburg, the Netherlands. Association for Computational Linguistics.

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

Natalia Grabar and Horacio Saggion. 2022. Evaluation of automatic text simplification: Where are we now, where should we go from here. In *TALN-RECITAL 2022*, pages 453–463, Avignon, France.

Nora Hollenstein, Maria Barrett, and Marina Björnsdóttir. 2022. The copenhagen corpus of eye tracking recordings from natural reading of Danish texts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1712–1720, Marseille, France. European Language Resources Association.

Victor Kuperman, Noam Siegelman, Sascha Schroeder, Alina Alexeeva, Cengiz Acarturk, Simona Amenta, Raymond Bertram, Rolando Bonandrini, Marc Brysbaert, Daria Chernova, et al. 2020. Text reading in English as a second language: Evidence from the multilingual eye-movements corpus (MECO). *Studies in Second Language Acquisition*.

A McCray. 2005. Promoting health literacy. *J of Am Med Infor Ass*, 12:152–163.

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

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

Abhinav Deep Singh, Poojan Mehta, Samar Husain, and Rajkumar Rajakrishnan. 2016. Quantifying sentence complexity based on eye-tracking measures. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 202–212, Osaka, Japan. The COLING 2016 Organizing Committee.

8.1 Appendices

Table 3: Participant overview by experimental group.

Set	*n*	F/M	Age	BA	MA	PhD	Emp
1A	12	9 / 3	19–42	9	2	2	5
1B	11	8 / 3	18–35	8	2	1	4
2A	11	7 / 4	18–43	9	1	1	4
2B	12	8 / 4	18–36	8	3	1	2
Total	**46**	**32 / 14**	18–43	34	8	5	15

BA = Bachelor's (Licence); MA = Master's; PhD = Doctorate; Emp = "Employed" counts anyone working (including student+worker). Levels are those held at the time of the experiment.

Features provided.
Duration_of_first_fixation: time (ms) of the first fixation on a word.
First-pass_duration: cumulative fixation time from first entering the word until leaving it to the right.
First-pass_first_fixation_duration: first-fixation duration restricted to the first-pass window.
First-pass_regression: binary flag (1 = gaze exits the word to the left during first pass).
Maximum_duration_of_fixations / *Minimum_duration_of_fixations*: longest and shortest single fixations on the word.
Number_of_fixations: count of fixations on the word.
Re-reading_duration: fixation time accumulated after the first pass.
Regression-path_duration: time from first entering the word until leaving it to the right *after* any regressions.
Total_duration_of_fixations: sum of all fixation durations on the word (early + late).

8.2 Appendices

Table 4: Comparison of Original and Simplified Texts

Text type	Text name	Version	Screens	Sentences	Tokens
Clinical	gastro	original	3	17	285
		simplified	3	19	323
	obGyn	original	3	11	249
		simplified	3	24	307
	toxico	original	4	19	398
		simplified	5	29	469
	gastro	original	3	13	255
		simplified	3	25	336
General	camelot	original	8	42	840
		simplified	8	58	880
	quince	original	7	44	751
		simplified	7	54	785
	popcorn	original	9	44	865
		simplified	7	51	752
	weekend	original	9	31	843
		simplified	9	49	811
Medical	autopsy	original	10	39	943
		simplified	9	65	925
	stroke	original	3	10	276
		simplified	3	22	328
	chikungunya	original	21	102	1983
		simplified	21	138	1975
	erytheme	original	7	34	653
		simplified	9	58	960
	obstetrics	original	12	57	1104
		simplified	12	65	1202
	ulcer	original	15	77	1526
		simplified	15	92	1551