

CERVINO: A Large-Scale Eye-Tracking Resource of Italian Legal Texts

Calogero Jerik Scozzaro¹, Matteo Delsanto¹, Sara Pasquet^{2,3*},
Luisa Revelli², Antonio Mastropaolo⁴, Daniele P. Radicioni¹

¹Department of Computer Science, University of Turin

²Department of Human and Social Sciences, University of Aosta Valley

³Department of Management Engineering, University of Bergamo

⁴Department of Economics and Political Sciences, University of Aosta Valley

Correspondence: calogerojerik.scozzaro@unito.it

Abstract

Eye-Tracking (ET) data provide valuable insights into human language processing and have recently been employed across a wide range of fields, from psycholinguistics to NLP. In this work, we introduce CERVINO, the largest Italian eye-tracking-while-reading dataset to date, comprising data from more than 200 participants reading legal texts, recorded at 120Hz. Our analyses show that the dataset captures established psycholinguistic effects: reading times are influenced by word length, lexical frequency, and surprisal. We further evaluate the usefulness of the dataset in downstream tasks by fine-tuning BERT on ET measures. Results show improved alignment between model attention and human reading behavior while preserving downstream task performance. Moreover, combining our data with higher-frequency ET resources yields the strongest alignment results, suggesting that lower-frequency ET data provide useful cognitive signals. Overall, these findings indicate that lower-frequency ET data retain sufficient cognitive information to provide useful supervision for attention alignment and complement existing high-frequency resources.¹

1 Introduction

Eye-Tracking (ET) is a technology used to investigate human behavior. It works by employing cameras and infrared sensors to track the position of the eyes while a person observes a visual stimulus, such as written text. When applied to reading, ET makes it possible to analyze how readers process text. During reading, eye movements are typically characterized by fixations and saccades. Fixations are brief stops that typically occur on words, while saccades are fast movements between two fixations, and are used to reposition the point of focus. Eye

movements during reading have long been studied in psycholinguistics, and many studies have shown that they are closely related to cognitive processing (Rayner, 1998, 2009). In particular, reading times are influenced by several lexical and contextual properties of words. Longer and less frequent words typically require longer fixation times, while more predictable words tend to be processed faster. This latter effect is often explained through surprisal theory (Hale, 2001; Levy, 2008), according to which words that are less expected given the preceding context require greater processing effort. In addition, Pimentel et al. (2023) suggested that reading times may also reflect anticipatory uncertainty about upcoming words, beyond surprisal alone.

In recent years ET data have been used in many Natural Language Processing (NLP) tasks, including linguistic acceptability (Bondar et al., 2025b), named entity recognition (Hollenstein and Zhang, 2019), readability (Scozzaro et al., 2024), language proficiency (Berzak et al., 2018), reading comprehension (Reich et al., 2022; Shubi et al., 2024), text simplification (Gruteke Klein et al., 2025), sentiment analysis (Mishra et al., 2016; Yang and Hollenstein, 2023), and general natural language understanding benchmarks like GLUE (Deng et al., 2023a, 2024). ET data have also proven useful in diagnostic applications, such as dyslexia (Prabha and Bhargavi, 2020; Kasprowski, 2024) and Alzheimer’s disease detection (Barral et al., 2020; Jang et al., 2021). Recently, ET data have been employed to steer language model behavior (Wang et al., 2025; Säuberli et al., 2026) and to align such systems with human preferences (Bondar et al., 2025a; López-Cardona et al., 2025). Finally, they offer a valuable resource for interpretability by analyzing model internals in relation to human attention patterns (Sood et al., 2020; Eberle et al., 2022; Dini et al., 2025a,b). A major limitation of ET-based research is the scarcity of available data, since data collection is

*Work done while at the University of Aosta Valley.

¹The dataset is publicly available on [zenodo](https://zenodo.org) and on [py-movements](https://py-movements.github.io) (Krakowczyk et al., 2023).

costly and difficult to scale. One way to mitigate this limitation is to develop models that generate synthetic scanpaths. Such approaches range from classical cognitive models, such as EZ-Reader (Reichle et al., 1998, 2009) and SWIFT (Engbert et al., 2005), to more recent neural approaches (Deng et al., 2023b; Bolliger et al., 2023, 2025b; Sood et al., 2025). Data scarcity is further amplified by the fact that most prior work relies on high-frequency ET data (typically around 1000 Hz), which require expensive laboratory equipment and highly controlled experimental setups. In contrast, lower-frequency recordings remain underexplored and are often excluded from benchmark settings (Shubi et al., 2025), even though they can be collected with cheaper and more scalable ET devices. This raises an important question: can lower-frequency ET data still provide reliable and useful signals for NLP applications?

Contributions We highlight three main contributions of this work: *i*) we introduce the **Corpus of Eye-movements while Reading Valle d’Aosta Italian NOrms** (CERVINO), the largest publicly available eye-tracking dataset of Italian texts to date, with more than 200 participants; *ii*) we provide eye-tracking data on legal texts, a currently under-represented domain in the literature; and *iii*) we show that lower-frequency eye-tracking data (120 Hz), which are substantially easier and cheaper to collect than high-frequency data, still capture established psycholinguistic effects and retain sufficient information to support existing ET-based NLP pipelines, improving attention alignment while preserving downstream performance.²

2 Related Work

Motivated by the growing use of ET data across different research areas, recent years have seen the development of several ET-while-reading datasets. The availability of such resources varies across languages, with English being the most represented. Widely used resources include Provo (Luke and Christianson, 2018), ZuCo (Hollenstein et al., 2018, 2020), CELER (Berzak et al., 2022), EMTeC (Bolliger et al., 2025a), and OneStop (Berzak et al., 2025). These datasets cover a variety of domains and reading settings, including newswire texts, fiction, popular science, machine-generated texts, and

simplified reading materials. Moving beyond English, two widely used monolingual resources are the Danish CopCo (Hollenstein et al., 2022; Björnsdóttir et al., 2023; Reich et al., 2024) and the German PoTeC (Jakobi et al., 2025a). To the best of our knowledge, the only monolingual Italian ET-while-reading dataset is ReadLet (Ferro et al., 2024), which contains data from 94 children and 55 adults reading four children’s fantasy stories and two short texts for adults. Multilingual resources have also been proposed, such as MECO (Siegelman et al., 2022, 2025), which comprises more than 1,100 participants across 13 languages including 54 Italian participants; and MultiPLEYE (Jakobi et al., 2025b), a large-scale data collection initiative covering more than 30 languages. For a comprehensive overview of available ET resources, see Jakobi et al. (2026).

Despite the growing number of datasets, resources for Italian remain limited, motivating the need for further data collection to support research on reading behavior in this language. Moreover, important domains such as legal texts are largely missing from current ET datasets.

3 Experimental Design

In this Section we describe the experimental design of CERVINO, including text materials, participants, procedure, and preprocessing steps employed to derive the ET measures.

3.1 Text Materials

Legal language is an important application domain for readability assessment, text simplification, and accessibility, yet it is largely absent from existing ET resources. To help fill this gap, the **Corpus of Eye-movements while Reading Valle d’Aosta Italian NOrms** (CERVINO) includes four Italian-language texts derived from two regional laws of the Autonomous Region of Aosta Valley. The first text (LUPO) is based on Regional Law no. 11 of 18 May 2021 concerning the management of issues related to the wolf species. A second variant (LUPO ERR) was created by introducing a set of intentional linguistic anomalies. These included orthographic errors and morpho-syntactic or lexical inconsistencies designed to induce local disruptions during reading while preserving the overall structure of the text. The second text (CICL) is based on Regional Law no. 14 of 29 July 2024 concerning regional support measures for the UCI Moun-

²The code to replicate the experiments is available on [GitHub](#).

	LUPO	LUPO ERR	CICL	CICL SIMPL
Sentences	19	19	33	45
Sent. Length	26.00 _{32.09}	26.11 _{32.26}	23.79 _{32.82}	18.51 _{19.21}
Words	494	496	785	833
TTR	0.56	0.57	0.44	0.44
Word Length	5.71 _{3.57}	5.72 _{3.57}	5.73 _{3.76}	5.62 _{3.56}
Word Freq.	5.53 _{1.43}	5.53 _{1.43}	5.34 _{1.52}	5.37 _{1.59}
Flesch	23.16	22.90	27.12	38.61
Gulpease	44.06	43.92	46.46	51.23
READ-IT	95.50	97.70	100.00	99.60

Table 1: Descriptive statistics for the four text variants in the CERVINO dataset. Values are reported as mean_{sd}.

tain Bike World Series competitions. A simplified version (CICL SIMPL) was produced by a linguist following plain-language principles. Simplification involved reorganizing information to improve salience, splitting long and syntactically complex sentences, and replacing technical expressions with more frequent lexical alternatives.

Table 1 reports descriptive statistics for each text, including the number of sentences, mean sentence length, total number of words, mean word length, type-token ratio (TTR), word frequencies, and three readability scores. Word frequencies were computed using the wordfreq library³ and expressed on the Zipf scale. Readability was assessed using three measures: *i*) the Italian adaptation of the Flesch Reading Ease score (Franchina and Vacca, 1986); *ii*) Gulpease (Lucisano et al., 1988); and *iii*) READ-IT (Dell’Orletta et al., 2011). For the first two higher values indicate easier texts, whereas higher READ-IT scores correspond to greater difficulty.

Comparing LUPO and LUPO ERR, the introduction of errors leads to a slight decrease in readability. Comparing CICL and CICL SIMPL, the simplified version contains more words due to a higher number of shorter sentences. Importantly, all three readability measures indicate that CICL SIMPL is more readable than the original version.

3.2 Participants

A total of 244 participants (61 per text) were recruited. Participants were adult native speakers of Italian, predominantly university students, and all had normal or corrected-to-normal vision. Each participant was assigned to a single text condition. After data collection, 37 participants were excluded due to technical issues or self-reported reading impairments (e.g., dyslexia), resulting in a final sample of 207 participants. Descriptive statistics for the

	N	Age	Gender	Education
LUPO	51	23.24 _{4.57}	F:35, M:16	14.96 _{2.47}
LUPO ERR	58	23.00 _{4.07}	F:30, M:28	14.10 _{2.01}
CICL	41	21.39 _{1.92}	F:31, M:10	13.37 _{0.99}
CICL SIMPL	57	21.75 _{2.22}	F:13, M:44	13.60 _{1.70}
Total	207	22.31 _{3.45}	F:109, M:98	14.03 _{1.98}

Table 2: Participant demographics by text. Age and education are reported as mean_{sd} (years). Gender is self-reported and given as counts (F = female, M = male).

final sample are reported in Table 2, including age, gender distribution, and years of education. Overall, the mean age is 22.31 ± 3.45 years, ranging from 18 to 33. The dataset comprises 109 female and 98 male participants and the average years of education is 14.03 ± 1.98 .

3.3 Procedure

ET data were recorded using a Tobii Pro Fusion device (120 Hz) with binocular tracking, positioned below the monitor of a laptop with an 15.6-inch display and a resolution of 1920×1080 . The experiment was designed with the Tobii Pro Lab software. Texts were presented in black Easy Reading font on a white background. Each text was displayed across multiple screens (i.e., segments), and reading was self-paced: participants pressed the space key to proceed to the next screen after finishing each segment. Returning to previous screens was not possible. Participants were seated approximately 60 cm from the screen. Head movements were not constrained (no chinrest was used), and participants were instructed to minimize head and body movements during recording. Prior to the experiment, a 9-point calibration procedure was performed. For LUPO and LUPO ERR, participants were informed before reading that the reading material might contain errors. Further details about the experimental setup are provided in Appendix A.1. After reading the text, participants completed a post-reading questionnaire aimed at assessing comprehension. The same protocol was adopted for each pair of text variants, ensuring comparability between original and manipulated versions. For LUPO and LUPO ERR, comprehension was mainly evaluated through open-ended questions. Participants were asked to summarize the content of the law, explain selected legal expressions, and discuss specific terms appearing in the text. For CICL and CICL SIMPL, the protocol was revised to include a more structured comprehension assessment. In addition to summary questions, participants com-

³<https://pypi.org/project/wordfreq/>

pleted 8 true/false statements targeting specific information contained in the text. The entire experiment lasted approximately 20 minutes.

3.4 Preprocessing and Reading Measures

Eye movement events were classified in Tobii Pro Lab using the I-VT (Velocity-Threshold Identification) algorithm with default settings (Salvucci and Goldberg, 2000). Word-level Areas Of Interest (AOIs) (Hooge et al., 2026) were generated within Tobii Pro Lab, and fixations were assigned to AOIs using the software. Initial preprocessing was performed in Tobii Pro Lab to extract scanpaths aligned with AOIs. Subsequent preprocessing was done using custom Python code, where fixations shorter than 100 ms were removed, following Holtenstein et al. (2022). The resulting scanpaths were then used to compute aggregated ET measures. We extracted a set of standard word-level ET measures, grouped as follows (see Table 5 in Appendix A.2 for formal definitions):

Continuous measures (ms)

- first-fixation duration (FFD)
- first duration (FD)
- first-pass reading time (FPRT)
- single-fixation duration (SFD)
- first-reading time (FRT)
- total-fixation time (TFT)
- re-reading time (RRT)
- inclusive regression-path duration (RPD_inc)
- exclusive regression-path duration (RPD_exc)
- right-bounded reading time (RBRT)

Binary measures

- skipped
- first-pass fixation (FPF)
- re-reading (RR)

Count measures

- total fixation count (TFC)
- total count of outgoing regressions (TRC_out)
- total count of incoming regressions (TRC_in)

4 Data Analysis

In this section we focus on a subset of the collected ET measures, namely FFD, FPRT, TFT, RRT, skipped, FPF, and RR. These measures capture different stages of the reading process. Early measures, such as FFD and FPRT, are typically associated with initial lexical access and early processing. In contrast, later measures, such as TFT and RRT,

are associated with global processing and semantic integration. Lastly, binary measures provide complementary information about word skipping, first-pass fixation, and re-reading behavior.

As a point of comparison, we consider the Italian subset of the MultipleEYE data, which includes 51 participants reading 10 texts from different domains, recorded at 1000 Hz. Specifically, in this Section we focus on the two institutional texts, *HumanRights* and *LearningMobility* (Kasperè et al., 2026). These texts are formal, impersonal, and highly structured, and are the closest in style to the legal texts in our dataset.

4.1 Mean Reading Measures

Figure 1 shows the distributions of the chosen measures across our four texts and the two MultipleEYE texts. All continuous measures are reported in ms. Comparing LUPO and LUPO ERR, early measures associated with lexical access (FFD and FPRT) are higher in LUPO ERR. In particular, mean FFD increases from 216.98 to 232.74 (+7%), while FPRT remains comparable (327.77 vs. 325.00). In contrast, later measures decrease in LUPO ERR, with mean TFT dropping from 425.91 to 400.49 (−6%) and RRT from 392.53 to 371.00 (−5%). Regarding binary measures, RR is considerably lower in LUPO ERR than in LUPO, while FPF is higher, indicating more careful first-pass processing. This pattern is compatible with the intended design of the introduced errors, which were meant to primarily affect early, surface-level processing without substantially impacting deeper semantic integration. A different pattern emerges when comparing CICL and CICL SIMPL, where all measures are higher in the simplified version. Specifically, mean FFD increases from 206.28 to 230.45 (+12%), FPRT from 316.36 to 344.30 (+9%), TFT from 435.12 to 505.22 (+16%), and RRT from 406.62 to 467.43 (+15%).⁴

Comparing our data with MultipleEYE, mean FFD in our data ranges from 206 to 232 ms, compared to approximately 215–216 ms in MultipleEYE, while FPRT is considerable higher in our data (316–344 ms vs. 263 ms). Larger differences emerge for measures associated with later processing. Mean TFT is lower in MultipleEYE (334–354 ms vs. 400–505 ms in our data), and the same pattern holds for RRT (281–310 ms vs. 371–467 ms).

⁴Given differences in both the underlying texts and reader populations, statistical comparisons are not directly applicable, and results should be interpreted descriptively.

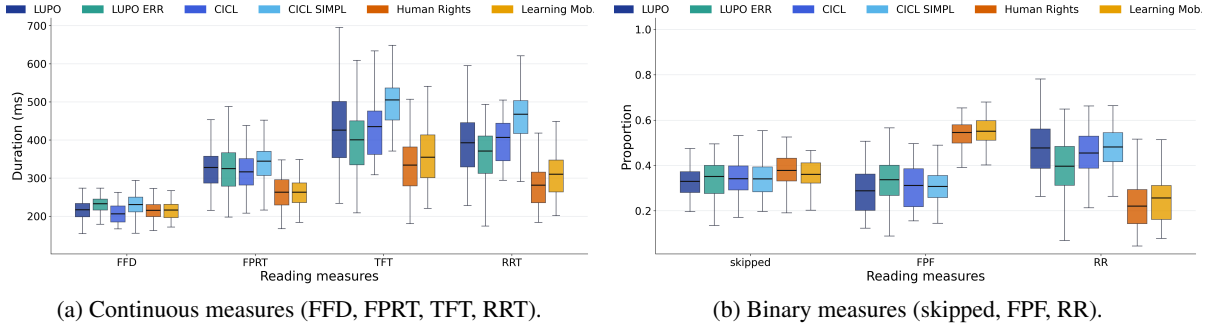


Figure 1: Distribution of reading measures across texts. Values are computed as reader-level means (i.e., the mean of each measure per reader). The black line indicates the median of these reader-level means.

Despite differences in central tendency, interquartile ranges are broadly comparable across datasets, suggesting similar levels of reader-level variability.

Two factors may contribute to the observed differences. First, the text types differ: while both involve formal registers, the institutional texts in MultiPEYE differ from legal texts in structure and complexity, and text genre is known to influence reading measures (Bolliger and Jäger, 2025). Second, differences in recording equipment may contribute, as MultiPEYE data were collected at a higher sampling rate (1000 Hz vs. 120 Hz), potentially yielding more precise fixation estimates.

Finally, we examined whether the increased reading times observed for the simplified text were associated with differences in comprehension. Readers of CICL SIMPL achieved higher comprehension scores than readers of the CICL, with a mean of 5.18 (SD = 1.34) correct answers out of 8, compared to 4.46 (SD = 1.29) for the original text. No significant correlation emerged between comprehension scores and mean reading measures, suggesting that the relationship between ET measures and comprehension may not be straightforward, in line with previous findings (Siegelman et al., 2025).

4.2 Psycholinguistic effects

We analyzed the impact of three well-known psycholinguistic predictors: surprisal, lexical frequency, and word length. Surprisal quantifies how unexpected a word is in context, and is defined as the negative log-probability assigned to a word given its preceding context (Hale, 2001):

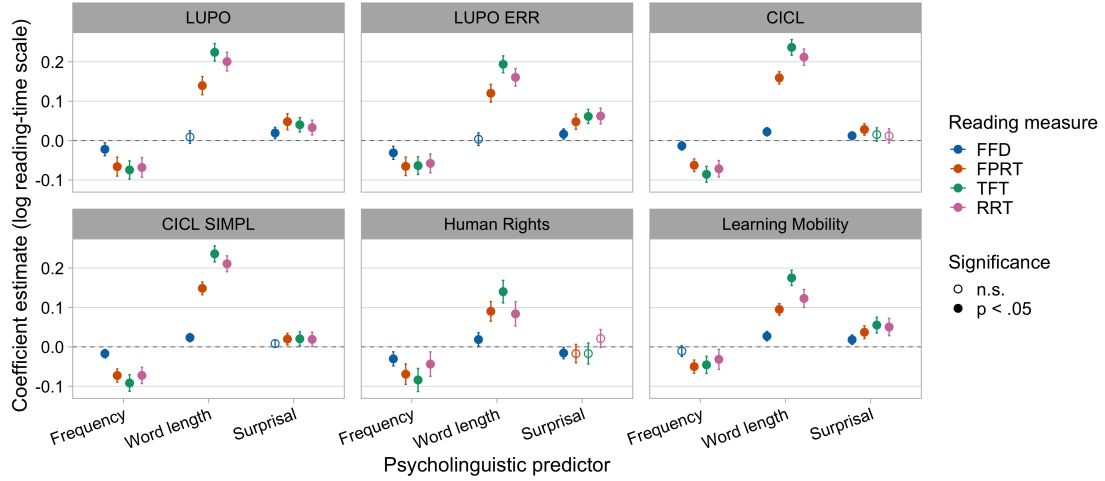
$$\text{surprisal}(w_n) = -\log_2 P(w_n \mid w_1, \dots, w_{n-1}).$$

Higher surprisal values therefore indicate less predictable words. We estimated surprisal values using an Italian version of GPT-2: GePpeTto (De Mattei

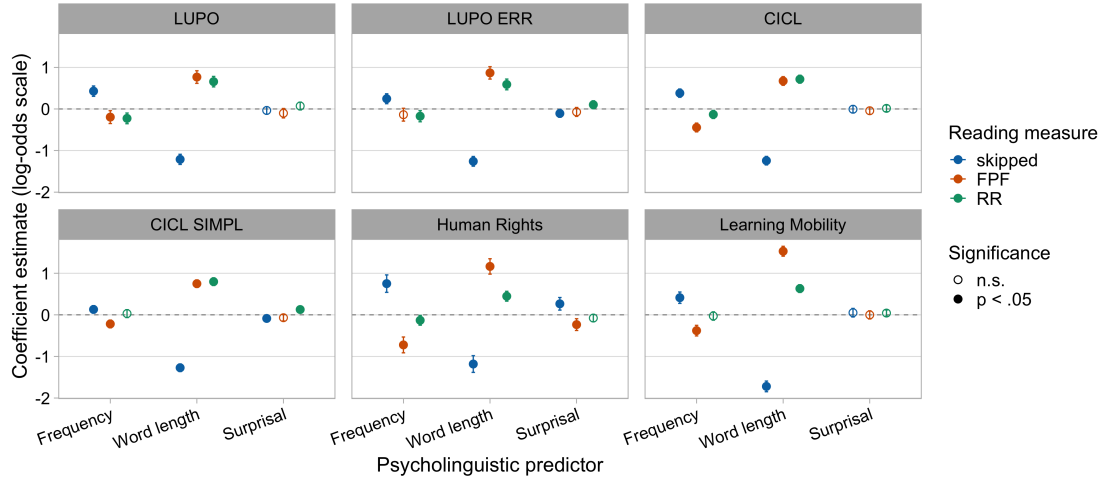
et al., 2020).⁵ Following Bolliger and Jäger (2025), continuous reading-time measures were modeled after excluding zeros and log-transforming the measures. The predictors were word surprisal, lexical Zipf frequency, and word length, all z -scored before model fitting. We fitted linear mixed-effects models using `lmer` and generalized linear mixed-effects models using `glmer` with a Bernoulli-logit link for binary measures. Models included by-subject random intercepts and random slopes for surprisal, and by-item random intercepts (Appendix A.3 for details). Fixed-effect estimates are reported with 95% Wald confidence intervals. Statistical significance was assessed using Satterthwaite-approximated t -tests for continuous models (Satterthwaite, 1946), and Wald z -tests for binary models (Wald, 1943).

Figure 2 shows a similar pattern across the six texts. Overall, word length exhibits the largest effect on reading times among the three predictors considered: longer words are associated with higher reading measures in all texts, with the largest effects observed for global measures (TFT and RRT). Interestingly, the effect of word length is smaller or not statistically significant for FFD. Frequency shows the opposite pattern, with higher-frequency words generally associated with shorter reading times. Surprisal is positively correlated with reading times, although its effect is smaller and not always statistically significant. We also conducted the surprisal analysis using the more recent Minerva-350M (Orlando et al., 2024) and obtained highly consistent results. Across the four CERVINO texts and the four continuous reading-time measures, surprisal coefficients had the same

⁵GePpeTto estimates probabilities at the subword level; these were summed to obtain AOI-level surprisal values, following standard practice (Wilcox et al., 2023).



(a) Continuous models.



(b) Binary models.

Figure 2: Fixed-effect estimates for the psycholinguistic predictors across the six texts. Error bars indicate 95% Wald confidence intervals. Filled points denote significant effects ($p < 0.05$), while open points denote non-significant effects.

positive direction in all cases (16/16), and the standardized fixed-effect estimates were strongly correlated with those obtained using GePpeTto (Pearson’s $r = 0.91$, $p < 0.001$).

Among the reading measures, FFD shows the smallest coefficients and the least consistent significance pattern. In contrast, TFT is the most affected measure, followed by RRT, suggesting that psycholinguistic properties have a larger impact on later stages of word processing. Despite differences in absolute values, our data exhibit patterns broadly consistent with those observed in MultiP-EYE, indicating that lower-frequency recordings can still capture meaningful aspects of human reading behavior.

5 Downstream Tasks

A central question of this work is whether ET data collected with lower-frequency (and lower-cost) devices can still provide useful signals for NLP models. This is particularly relevant for scalable resource creation, since portable and affordable eye trackers may enable the collection of larger and more diverse datasets than traditional laboratory setups. Recent work on mouse tracking (Wilcox et al., 2024) suggests that lower-quality behavioral signals can serve as useful proxies before fine-tuning on higher-quality data (Reich et al., 2025). In this section, we aim to achieve two main goals: *i*) replicate findings reported for English by Dini et al. (2025a), showing that fine-tuning on ET data im-

Test set	Model	FFD	FPRT	TFT	RRT	skipped	FPF	RR
HumanRights	BERT	0.20	0.20	0.21	0.19	0.25	0.25	0.23
	BERT _{LearningMobility}	0.20*	0.22	0.24	0.20	0.29	0.29	0.26
	BERT _{CER}	0.20*	0.24	0.26	0.21	0.31	0.31	0.28
	BERT _{ME_NO_HumanRights}	0.20*	0.24	0.27	0.22	0.33	0.33	0.29
	BERT _{CER-LearningMobility}	0.20*	0.25	0.27	0.22	0.33	0.33	0.29
	BERT _{CER-ME_NO_HumanRights}	0.20*	0.26	0.29	0.24	0.36	0.36	0.31
LearningMobility	BERT	0.15	0.19	0.23	0.20	0.24	0.23	0.22
	BERT _{HumanRights}	0.15*	0.19*	0.23*	0.20	0.26	0.26	0.23*
	BERT _{CER}	0.16	0.20	0.25	0.22	0.29	0.29	0.26
	BERT _{ME_NO_LearningMobility}	0.17	0.23	0.27	0.23	0.32	0.32	0.27
	BERT _{CER-HumanRights}	0.16	0.22	0.26	0.23	0.31	0.31	0.27
	BERT _{CER-ME_NO_LearningMobility}	0.17	0.24	0.29	0.25	0.33	0.33	0.29

Table 3: Mean Spearman correlation between model attention and human ET measures across model variants and test sets. Values are averaged over five runs. Subscripts indicate the data used for ET fine-tuning: CER refers to CERVINO; HumanRights and LearningMobility denote single-text fine-tuning; ME denotes MultipleYE; “NO_X” indicates that text *X* is excluded from training; hyphenated variants denote sequential fine-tuning. Bold indicates the highest score per measure and test set. * marks differences from BERT that are not statistically significant (paired Wilcoxon signed-rank test, $p < 0.05$).

proves alignment with human attention without substantially degrading downstream task performance; and *ii*) assess the usefulness of our data in combination with a higher-quality dataset, namely the Italian subset of MultipleYE.

5.1 Attention Patterns

To study the alignment between human and model attention, we used an Italian BERT.⁶ We treated fine-tuning on ET data as a multi-label token-level regression task, where the model predicts the seven ET measures used in Section 4 simultaneously for each token in a text.⁷ To evaluate whether fine-tuning on ET data modifies the model’s attention patterns, we computed the Spearman correlation coefficient between human attention and model attention across layers, following Dini et al. (2025a).⁸ Model attention scores were extracted from the attention values assigned by the [CLS] token to each input token, using only the first sub-token in case of word splitting. Correlation scores were computed per layer and then averaged to obtain a single value for each model-measure pair. Higher values indicate stronger alignment between model attention and human reading behavior. Reader-level mea-

asures were aggregated as described in Section 4.1 and min-max scaled. We consider both single-stage and sequential fine-tuning settings. In the latter, the model is first fine-tuned on the ET measures derived from our four texts and then further fine-tuned on the ET measures from selected MultipleYE texts. We also include settings in which the second stage uses only a single MultipleYE text, simulating a realistic scenario in which only limited high-quality ET data are available. These settings allow us to assess whether lower-frequency ET data can complement higher-quality data. Evaluation was conducted on the *HumanRights* and *LearningMobility* reading measures. To ensure a fair comparison, when using MultipleYE data we excluded the test text from the training set. Experiments were repeated over five runs and results are reported as averages across runs. We compared each fine-tuned model against the pre-trained BERT using paired Wilcoxon signed-rank tests on matched attention-correlation scores, run separately for each of the five seeds. A difference was considered robust only if the test was significant in all five runs. Appendix A.4 reports training details and additional results on ET measure prediction.

Table 3 reports the alignment between model attention and human ET measures across model variants and test sets. Overall, fine-tuning on ET data consistently improves attention alignment over the BERT baseline. Comparing data sources, models trained only on our data (BERT_{CER})⁹ achieve

⁶<https://huggingface.co/dbmdz/bert-base-italian-uncased>

⁷For multi-token words, features were associated only with the first sub-token, following Hollenstein et al. (2021).

⁸Raw attention scores are only one possible proxy for model interpretability and have known limitations. Alternative attribution methods, such as ALTI (Ferrando et al., 2022) and ValueZeroing (Mohebbi et al., 2023), may provide a more faithful characterization of token importance and represent promising directions for future work.

⁹The CER setup includes all four texts collected in our dataset.

slightly lower performance than those trained only on MultipleYE (BERT_{ME_NO_HumanRights} and BERT_{ME_NO_LearningMobility}), likely due to the larger number of texts and higher quality of the latter. Combining the two sources yields the strongest results: sequential fine-tuning on CER followed by a single MultipleYE text matches the performance of using MultipleYE alone on *HumanRights*, and approaches it on *LearningMobility*, while leveraging all available data achieves the best overall performance. These results suggest that lower-frequency ET data, while individually less informative, provide useful signals that can be effectively combined with higher-quality data, even when only limited amounts of the latter are available.

5.2 NLP Benchmarks

We evaluate our models on several established Italian NLP benchmarks, relying on datasets released in EVALITA campaigns,¹⁰ to assess whether ET fine-tuning leads to marked degradation in downstream task performance, i.e., catastrophic forgetting. The tasks include Hate Speech (HS) (Sanguinetti et al., 2020), Irony Detection (IRON) (Cignarella et al., 2018), Named Entity Recognition (NER) (Palmero Aprosio et al., 2023), Part-of-Speech Tagging (POS) (Bosco et al., 2020), and Sentiment Analysis (SA) (Barbieri et al., 2016). All benchmarks provide official training and testing splits. When no validation split was available, we randomly sampled one from the training data. We selected a subset of the models presented in Section 5.1, excluding those fine-tuned on a single MultipleYE text due to their limited gains in attention alignment. To adapt the models to these benchmarks, we removed the additional linear layer used for predicting reading measures. For each task, we fine-tuned the models on the training set for three epochs. We performed a hyperparameter search over the learning rate, selecting the configuration that achieved the best performance on the validation set. Final scores were obtained by averaging results across five runs.

Table 4 reports the results on the downstream NLP benchmarks. Overall, fine-tuning on ET data leads to only negligible performance decreases, in line with previous results reported for English (Dini et al., 2025a). Across tasks, ET-enhanced models generally achieve performance comparable to the BERT baseline, with differences often within a few

percentage points. In some cases, they match the baseline performance, such as on IRON and SA, while larger decreases are mainly observed for HS and NER. Notably, models fine-tuned on our data also show only limited performance degradation.

These findings should be interpreted together with the attention-alignment results presented in Section 5.1. Fine-tuning on ET data substantially improves alignment with human reading behavior while largely preserving downstream NLP performance, also when using our dataset. Overall, these results suggest that lower-frequency ET data can provide cognitively meaningful signals without causing catastrophic forgetting.

6 Conclusions

In this work we presented CERVINO, an Italian ET dataset collected during natural reading on legal texts. To the best of our knowledge, this is the largest publicly available ET resource for Italian to date, comprising data from more than 200 participants. It addresses two important gaps in the existing literature: the limited availability of ET resources for Italian and the lack of data covering the legal domain. This resource can enable future work on legal language readability, text simplification, and accessibility.

Our analyses show that the dataset captures known psycholinguistic effects during reading. In particular, word length, lexical frequency, and surprisal affect reading measures in expected directions, with patterns broadly comparable to those observed in higher-frequency resources. We further evaluated the usefulness of the dataset in downstream experiments. Fine-tuning BERT on our ET measures improves the alignment between model attention and human reading behavior. Moreover, when combined with higher-frequency data, it leads to the strongest alignment results. Finally, experiments on Italian NLP benchmarks show that ET fine-tuning improves cognitive alignment with negligible decrease in downstream task performance, replicating findings previously reported for English. The experiments suggest that low-frequency recordings preserve much of the information contained in continuous eye-tracking measures. In contrast, binary measures appear to be more sensitive to the reduced temporal resolution, as reflected by their lower MAE performance when models are fine-tuned exclusively on CERVINO (Table 7). This observation is consistent with the intuition that bi-

¹⁰<https://www.evalita.it/>

Model	HS	IRON	NER	POS	SA	AVG
BERT	0.74 _{0.04}	0.70 _{0.01}	0.89 _{0.01}	0.93 _{<0.01}	0.70 _{0.02}	0.79
BERT _{CER}	0.71 _{0.02}	0.69 _{0.02}	0.87 _{0.01}	0.92 _{<0.01}	0.70 _{0.02}	0.78
BERT _{ME_NO_LearningMobility}	0.72 _{0.03}	0.69 _{0.01}	0.86 _{0.01}	0.92 _{<0.01}	0.70 _{0.02}	0.78
BERT _{ME_NO_HumanRights}	0.72 _{0.01}	0.70 _{0.01}	0.87 _{0.02}	0.92 _{<0.01}	0.69 _{0.02}	0.78
BERT _{CER-HumanRights}	0.71 _{0.03}	0.70 _{0.01}	0.87 _{<0.01}	0.92 _{<0.01}	0.71 _{0.02}	0.78
BERT _{CER-LearningMobility}	0.72 _{0.03}	0.69 _{0.02}	0.87 _{<0.01}	0.92 _{<0.01}	0.70 _{0.02}	0.78
BERT _{CER-ME_NO_LearningMobility}	0.71 _{0.02}	0.70 _{0.01}	0.86 _{0.01}	0.92 _{<0.01}	0.69 _{0.02}	0.78
BERT _{CER-ME_NO_HumanRights}	0.70 _{0.02}	0.68 _{0.03}	0.87 _{<0.01}	0.92 _{<0.01}	0.67 _{0.02}	0.77

Table 4: Results on Italian downstream tasks. Values are reported as mean_{sd} over five runs. Bold indicates the highest score for each task and for the overall average.

nary fixation decisions depend on accurately detecting short fixations, which become harder to measure at lower sampling rates. Although a controlled comparison would require recording the same participants reading the same texts under different acquisition settings, these results suggest that lower-frequency eye tracking may be better suited to modeling continuous reading measures than binary fixation events.

Overall, these findings indicate that data collected with a 120 Hz eye tracker can still capture meaningful psycholinguistic effects and provide informative cognitive signals, particularly for continuous eye-tracking measures. Our results do not suggest that lower-frequency data replace high-frequency ET data; rather, they show that lower-cost recordings can effectively complement existing high-frequency resources for cognitive modeling and NLP, while enabling more scalable data collection.

Limitations

Our study has limitations related to data collection, preprocessing, and model training. A considerable number of recordings (37) were excluded from the final dataset. Of these, approximately 15 correspond to participants with learning disorders; we plan to release these data in future work with appropriate methodological adaptations. Around 10 exclusions were due to errors in the text materials, while the remaining cases were caused by issues related to the ET device. Another limitation concerns the assessment of reading comprehension. During data collection, the comprehension protocol was revised from open-ended questions to a structured questionnaire based on the experience gained in the initial phase of the study, resulting in different assessment protocols across the two pairs of texts. Moreover, the simplified and error-containing con-

ditions are each represented by a single text, limiting the generalizability of our analyses.

While the dataset is overall balanced, participant distributions differ across subsets due to variations in recruitment settings. In particular, CICL is skewed towards female participants, whereas CICL SIMPL is skewed towards male participants.

Event detection and AOI assignment rely on the default I-VT settings provided by the Tobii Pro Lab software, without additional manual correction procedures. This may affect the precision of the fixation estimates.

Lastly, we note a limitation of the ET fine-tuning setup. All models were trained for a fixed number of 100 epochs, without hyperparameter optimization on a validation set. We selected 100 epochs based on preliminary experiments showing that training had largely converged by that point. While this choice is suboptimal and may affect downstream performance, it was necessary due to the limited amount of available ET data. In contrast, downstream NLP benchmarks were tuned using validation sets.

Ethical Considerations

Informed consent was obtained from all participants prior to the experiment, including consent for the reuse of anonymized data for research purposes. The study was reviewed and approved by the Data Protection Officer (DPO) of the University of Aosta Valley. Participation was voluntary and unpaid. All eye-tracking recordings and demographic metadata were anonymized prior to processing and analysis, and no personally identifiable information is included in the released dataset. The resource is intended exclusively for research on reading behavior, cognitive modeling, readability, and NLP applications.

Acknowledgments

This work was supported by the Fondazione CRT under the 2021 Ordinary Grants Programme (Bando Erogazioni Ordinarie 2021), through the project “*L’accessibilità dei testi normativi come dispositivo di inclusione: studio di caso e strumenti applicativi in Valle d’Aosta*”, coordinated by Antonio Mastropaolo and Luisa Revelli at the University of Aosta Valley. The authors also acknowledge the support of the Interdepartmental Research Centre *Human Science and Technologies (HST)* at the University of Turin.

References

- Francesco Barbieri, Valerio Basile, Danilo Croce, Malvina Nissim, Nicole Novielli, Viviana Patti, and 1 others. 2016. Overview of the evalita 2016 sentiment polarity classification task. In *CEUR workshop proceedings*, volume 1749. CEUR-WS.
- Oswald Barral, Hyeju Jang, Sally Newton-Mason, Sheetal Shajan, Thomas Soroski, Giuseppe Carenini, Cristina Conati, and Thalia Field. 2020. Non-invasive classification of alzheimer’s disease using eye tracking and language. In *Machine Learning for Healthcare Conference*, pages 813–841. PMLR.
- Yevgeni Berzak, Boris Katz, and Roger Levy. 2018. Assessing language proficiency from eye movements in reading. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1986–1996.
- Yevgeni Berzak, Jonathan Malmaud, Omer Shubi, Yoav Meiri, Ella Lion, and Roger Levy. 2025. Onestop: A 360-participant english eye tracking dataset with different reading regimes. *Scientific Data*.
- 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*, 6:41–50.
- Marina Björnsdóttir, Nora Hollenstein, and Maria Barrett. 2023. Dyslexia prediction from natural reading of danish texts. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 60–70.
- Lena Bolliger, David Reich, Patrick Haller, Deborah Jakobi, Paul Prasse, and Lena Jäger. 2023. Scandl: A diffusion model for generating synthetic scanpaths on texts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15513–15538.
- Lena S Bolliger, Patrick Haller, Isabelle CR Cretton, David R Reich, Tannon Kew, and Lena A Jäger. 2025a. Emtec: A corpus of eye movements on machine-generated texts. *Behavior Research Methods*, 57(7):189.
- Lena S Bolliger, David R Reich, and Lena A Jäger. 2025b. Scandl 2.0: A generative model of eye movements in reading synthesizing scanpaths and fixation durations. *Proceedings of the ACM on Human-Computer Interaction*, 9(3):1–29.
- Lena Sophia Bolliger and Lena Ann Jäger. 2025. Genre matters: How text types interact with decoding strategies and lexical predictors in shaping reading behavior. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 7470–7487.
- Anna Bondar, David Robert Reich, and Lena Ann Jäger. 2025a. Aleyegment: Leveraging eye-tracking-while-reading to align language models with human preferences. In *Proceedings of the First International Workshop on Gaze Data and Natural Language Processing*, pages 58–70.
- Anna Bondar, David Robert Reich, and Lena Ann Jäger. 2025b. Colagaze: A corpus of eye movements for linguistic acceptability. In *Proceedings of the 2025 Symposium on Eye Tracking Research and Applications*, pages 1–9.
- Cristina Bosco, Silvia Ballarè, Massimo Cerruti, Eugenio Gorla, Caterina Mauri, and 1 others. 2020. Kipos@ evalita2020: overview of the task on kiplara part of speech tagging. In *CEUR WORKSHOP PROCEEDINGS*, pages 1–8.
- Marco Canali, Christina Bachmann, and Federico Alfonsetti. 2021. A highly legible font for all. In *Congress of the International Ergonomics Association*, pages 172–179. Springer.
- Alessandra Teresa Cignarella, Simona Frenda, Valerio Basile, Cristina Bosco, Viviana Patti, Paolo Rosso, and 1 others. 2018. Overview of the evalita 2018 task on irony detection in italian tweets (ironita). In *CEUR Workshop Proceedings*, volume 2263, pages 1–6. CEUR-WS.
- Lorenzo De Mattei, Michele Cafagna, Felice Dell’Orletta, Malvina Nissim, and Marco Guerini. 2020. Geppetto carves italian into a language model. In *Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020)*, pages 97–104.
- Felice Dell’Orletta, Simonetta Montemagni, and Giulia Venturi. 2011. READ-IT: Assessing readability of Italian texts with a view to text simplification. In *Proceedings of the second workshop on speech and language processing for assistive technologies*, pages 73–83.
- Shuwen Deng, Paul Prasse, David Reich, Tobias Schefter, and Lena Jäger. 2023a. Pre-trained language

- models augmented with synthetic scanpaths for natural language understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6500–6507.
- Shuwen Deng, Paul Prasse, David Reich, Tobias Scheffer, and Lena Jäger. 2024. Fine-tuning pre-trained language models with gaze supervision. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 217–224.
- Shuwen Deng, David R Reich, Paul Prasse, Patrick Haller, Tobias Scheffer, and Lena A Jäger. 2023b. Eyettention: An attention-based dual-sequence model for predicting human scanpaths during reading. *Proceedings of the ACM on Human-Computer Interaction*, 7(ETRA):1–24.
- Luca Dini, Lucia Domenichelli, Dominique Brunato, and Felice Dell’Orletta. 2025a. [From human reading to NLM understanding: Evaluating the role of eye-tracking data in encoder-based models](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17796–17813, Vienna, Austria. Association for Computational Linguistics.
- Luca Dini, Luca Moroni, Dominique Brunato, and Felice Dell’Orletta. 2025b. [In the eyes of a language model: A comprehensive examination through eye-tracking data](#). *Neurocomputing*, 650:130617.
- Oliver Eberle, Stephanie Brandl, Jonas Pilot, and Anders Søgaard. 2022. Do transformer models show similar attention patterns to task-specific human gaze? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4295–4309.
- Ralf Engbert, Antje Nuthmann, Eike M Richter, and Reinhold Kliegl. 2005. Swift: a dynamical model of saccade generation during reading. *Psychological review*, 112(4):777.
- Javier Ferrando, Gerard I Gállego, and Marta R Costa-Jussà. 2022. Measuring the mixing of contextual information in the transformer. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8698–8714.
- Marcello Ferro, Claudia Marzi, Andrea Nadalini, Loukia Taxitari, Alessandro Lento, and Vito Pirrelli. 2024. Readlet: a dataset for oral, visual and tactile text reading data of early and mature readers. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13595–13609.
- Valerio Franchina and Roberto Vacca. 1986. Adaptation of Flesch readability index on a bilingual text written by the same author both in Italian and English languages. *Linguaggi*, 3:47–49.
- Keren Gruteke Klein, Omer Shubi, Shachar Frenkel, and Yevgeni Berzak. 2025. [The effect of text simplification on reading fluency and reading comprehension in 11 english speakers](#). *Proceedings of the Annual Meeting of the Cognitive Science Society*, 47(0).
- John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *Second meeting of the north american chapter of the association for computational linguistics*.
- 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.
- Nora Hollenstein, Federico Pirovano, Ce Zhang, Lena Jäger, and Lisa Beinborn. 2021. [Multilingual language models predict human reading behavior](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 106–123, Online. Association for Computational Linguistics.
- Nora Hollenstein, Jonathan Rotsztein, Marius Troendle, Andreas Pedroni, Ce Zhang, and Nicolas Langer. 2018. Zuco, a simultaneous eeg and eye-tracking resource for natural sentence reading. *Scientific data*, 5(1):180291.
- Nora Hollenstein, Marius Troendle, Ce Zhang, and Nicolas Langer. 2020. Zuco 2.0: A dataset of physiological recordings during natural reading and annotation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 138–146.
- Nora Hollenstein and Ce Zhang. 2019. Entity recognition at first sight: Improving ner with eye movement information. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1–10.
- Ignace TC Hooge, Roy S Hessels, Diederick C Niehorster, Richard Andersson, Marta K Skrok, Robert Konklewski, Patrycjusz Stemplewski, Maciej Nowakowski, Szymon Tamborski, Anna Szkulmowska, and 1 others. 2024. Eye tracker calibration: How well can humans refixate a target? *Behavior Research Methods*, 57(1):23.
- Ignace TC Hooge, Marcus Nyström, Diederick C Niehorster, Richard Andersson, Tom Foulsham, Antje Nuthmann, and Roy S Hessels. 2026. The fundamentals of eye tracking part 6: Working with areas of interest. *Behavior Research Methods*, 58(3):65.
- Deborah N. Jakobi, Thomas Kern, David R. Reich, Patrick Haller, and Lena A. Jäger. 2025a. [Potec: A german naturalistic eye-tracking-while-reading corpus](#). *Behavior Research Methods*, 57(8):211.

- Deborah N. Jakobi, David R. Reich, Paul Prasse, Jana M. Hofmann, Lena S. Bolliger, and Lena A. Jäger. 2026. [Eye-tracking-while-reading: A living survey of datasets with open library support.](#)
- Deborah Noemie Jakobi, Maja Stegenwallner-Schütz, Nora Hollenstein, Cui Ding, Ramune Kaspere, Ana Matić Škorić, Eva Pavlinusić Vilus, Stefan Frank, Marie-Luise Müller, Kristine M Jensen de López, and 1 others. 2025b. Multipleye: Creating a multilingual eye-tracking-while-reading corpus. In *Proceedings of the 2025 Symposium on Eye Tracking Research and Applications*, pages 1–11.
- Hyeju Jang, Thomas Soroski, Matteo Rizzo, Oswald Barral, Anuj Harisinghani, Sally Newton-Mason, Safirin Granby, Thiago Monnerat Stutz da Cunha Vasco, Caitlin Lewis, Pavan Tutt, and 1 others. 2021. Classification of alzheimer’s disease leveraging multi-task machine learning analysis of speech and eye-movement data. *Frontiers in Human Neuroscience*, 15:716670.
- Ramunė Kasperė, Anna Bondar, Sergiu Nisioi, Maja Stegenwallner-Schütz, Hanne B. Søndergaard Knudsen, Ana Matić, Eva Pavlinusić Vilus, Dorota Klimek-Jankowska, Chiara Tschirner, Not Battesta Soliva, Deborah N. Jakobi, Cui Ding, Dima Abu Romi, Cengiz Acarturk, Matilda Agdler, Anton Marius Alexandru, Mohd Faizan Ansari, Annalisa Arcidiacono, Elizabete Ausma Velta Barisa, and 87 others. 2026. [The multipleye text corpus: Towards a diverse and ever-expanding multilingual text corpus.](#) In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 6706–6721, Palma, Mallorca, Spain. European Language Resources Association (ELRA).
- Pawel Kasprowski. 2024. Utilizing gaze self similarity plots to recognize dyslexia when reading. In *Proceedings of the 2024 Symposium on Eye Tracking Research and Applications*, pages 1–5.
- Daniel G Krakowczyk, David R Reich, Jakob Chwastek, Deborah N Jakobi, Paul Prasse, Assunta Süß, Oleksii Turuta, Paweł Kasprowski, and Lena A Jäger. 2023. Pymovements: A python package for eye movement data processing. In *Proceedings of the 2023 Symposium on Eye Tracking Research and Applications*, pages 1–2.
- Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.
- Ángela López-Cardona, Carlos Segura, Alexandros Karatzoglou, Sergi Abadal, and Ioannis Arapakis. 2025. [Seeing eye to AI: Human alignment via gaze-based response rewards for large language models.](#) In *The Thirteenth International Conference on Learning Representations*.
- Pietro Lucisano, Maria Emanuela Piemontese, and 1 others. 1988. GulpEase: una formula per la predizione della leggibilità di testi in lingua italiana. *Scuola e città*, pages 110–124.
- Steven G Luke and Kiel Christianson. 2018. The provo corpus: A large eye-tracking corpus with predictability norms. *Behavior research methods*, 50:826–833.
- Abhijit Mishra, Diptesh Kanojia, Seema Nagar, Kuntal Dey, and Pushpak Bhattacharyya. 2016. Leveraging cognitive features for sentiment analysis. In *Proceedings of the 20th SIGNLL conference on computational natural language learning*, pages 156–166.
- Hosein Mohebbi, Willem Zuidema, Grzegorz Chrupała, and Afra Alishahi. 2023. Quantifying context mixing in transformers. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3378–3400.
- Riccardo Orlando, Luca Moroni, Pere-Lluís Huguët Cabot, Simone Conia, Edoardo Barba, Sergio Orlan-dini, Giuseppe Fiameni, and Roberto Navigli. 2024. Minerva llms: The first family of large language models trained from scratch on italian data. In *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 707–719.
- Alessio Palmero Aprosio, Teresa Paccosi, and 1 others. 2023. Nermud at evalita 2023: overview of the named-entities recognition on multi-domain documents task. In *CEUR WORKSHOP PROCEEDINGS*, volume 3473. CEUR-WS.
- Tiago Pimentel, Clara Meister, Ethan G. Wilcox, Roger P. Levy, and Ryan Cotterell. 2023. [On the Effect of Anticipation on Reading Times.](#) *Transactions of the Association for Computational Linguistics*, 11:1624–1642.
- A Jothi Prabha and Renta Bhargavi. 2020. Predictive model for dyslexia from fixations and saccadic eye movement events. *Computer Methods and Programs in Biomedicine*, 195:105538.
- Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3):372.
- Keith Rayner. 2009. The 35th sir frederick bartlett lecture: Eye movements and attention in reading, scene perception, and visual search. *Quarterly journal of experimental psychology*, 62(8):1457–1506.
- David R Reich, Shuwen Deng, Marina Björnsdóttir, Lena Jäger, and Nora Hollenstein. 2024. Reading does not equal reading: Comparing, simulating and exploiting reading behavior across populations. In *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024)*, pages 13586–13594.
- David Robert Reich, Cui Ding, Lena Sophia Bolliger, Patrick Haller, Paul Prasse, and Lena A Jäger. 2025. Proxy-based pre-training for eye-tracking applications. In *Proceedings of the 2025 Symposium on Eye Tracking Research and Applications*, pages 1–7.

- David Robert Reich, Paul Prasse, Chiara Tschirner, Patrick Haller, Frank Goldhammer, and Lena A Jäger. 2022. Inferring native and non-native human reading comprehension and subjective text difficulty from scanpaths in reading. In *2022 Symposium on eye tracking research and applications*, pages 1–8.
- Erik D Reichle, Alexander Pollatsek, Donald L Fisher, and Keith Rayner. 1998. Toward a model of eye movement control in reading. *Psychological review*, 105(1):125.
- Erik D Reichle, Tessa Warren, and Kerry McConnell. 2009. Using ez reader to model the effects of higher level language processing on eye movements during reading. *Psychonomic bulletin & review*, 16(1):1–21.
- Dario D Salvucci and Joseph H Goldberg. 2000. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 symposium on Eye tracking research & applications*, pages 71–78.
- Manuela Sanguinetti, Gloria Comandini, Elisa Di Nuovo, Simona Frenda, Marco Stranisci, Cristina Bosco, Tommaso Caselli, Viviana Patti, and Irene Russo. 2020. Haspeede 2@ evalita2020: Overview of the evalita 2020 hate speech detection task. *Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*.
- Franklin E Satterthwaite. 1946. An approximate distribution of estimates of variance components. *Biometrics bulletin*, 2(6):110–114.
- Andreas Säuberli, Darja Jepifanova, Diego Frassinelli, and Barbara Plank. 2026. [Controlling reading ease with gaze-guided text generation](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2383–2397, Rabat, Morocco. Association for Computational Linguistics.
- Calogero J Scozzaro, Matteo Delsanto, Antonio Mastropalo, Enrico Mensa, Luisa Revelli, and Daniele P Radicioni. 2024. On the reform of the italian constitution: an interdisciplinary text readability analysis. In *CEUR WORKSHOP PROCEEDINGS*, 3877. CEUR-WS.
- Omer Shubi, Yoav Meiri, Cfir Avraham Hadar, and Yevgeni Berzak. 2024. Fine-grained prediction of reading comprehension from eye movements. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3372–3391.
- Omer Shubi, David Robert Reich, Keren Gruteke Klein, Yuval Angel, Paul Prasse, Lena Ann Jäger, and Yevgeni Berzak. 2025. [Eyebench: Predictive modeling from eye movements in reading](#). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Noam Siegelman, Sascha Schroeder, Cengiz Acartürk, Hee-Don Ahn, Svetlana Alexeeva, Simona Amenta, Raymond Bertram, Rolando Bonandrini, Marc Brysbaert, Daria Chernova, and 1 others. 2022. Expanding horizons of cross-linguistic research on reading: The multilingual eye-movement corpus (mecor). *Behavior research methods*, 54(6):2843–2863.
- Noam Siegelman, Sascha Schroeder, Yaqian Borogjoon Bao, Cengiz Acartürk, Niket Agrawal, Lena S Bolliger, Jan Brasser, César Campos-Rojas, Denis Drieghe, Dušica Filipović Đurđević, and 1 others. 2025. Wave 2 of the multilingual eye-movement corpus (mecor): New text reading data across languages. *Scientific data*, 12(1):1183.
- Ekta Sood, Prajit Dhar, Enrica Troiano, Rosy Southwell, and Sidney K D’Mello. 2025. Scanez: Integrating cognitive models with self-supervised learning for spatiotemporal scanpath prediction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1132–1142.
- Ekta Sood, Simon Tannert, Diego Frassinelli, Andreas Bulling, and Ngoc Thang Vu. 2020. Interpreting attention models with human visual attention in machine reading comprehension. In *Proceedings of the 24th conference on computational natural language learning*, pages 12–25.
- Abraham Wald. 1943. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical society*, 54(3):426–482.
- Xintong Wang, Jingheng Pan, Liang Ding, Longyue Wang, Longqin Jiang, Xingshan Li, and Chris Bie-mann. 2025. Cogsteer: Cognition-inspired selective layer intervention for efficiently steering large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25507–25522.
- Ethan G Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P Levy. 2023. Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, 11:1451–1470.
- Ethan Gotlieb Wilcox, Cui Ding, Mrinmaya Sachan, and Lena Ann Jäger. 2024. Mouse tracking for reading (motr): A new naturalistic incremental processing measurement tool. *Journal of Memory and Language*, 138:104534.
- Duo Yang and Nora Hollenstein. 2023. [Plm-as: Pre-trained language models augmented with scanpaths for sentiment classification](#). *Proceedings of the Northern Lights Deep Learning Workshop*, 4.

A Appendix

A.1 Experimental Setup

This section provides additional details about the experimental setup. Texts were presented in black

Measure	Abbr.	Definition
<i>Continuous measures (ms)</i>		
first-fixation duration	FFD	Duration of the first fixation on a word if it is fixated during first-pass reading, otherwise 0.
first duration	FD	Duration of the first fixation on a word (identical to FFD if not skipped in the first pass).
first-pass reading time	FPRT	Sum of the durations of all first-pass fixations on a word (0 if skipped in first pass).
single-fixation duration	SFD	Duration of the only first-pass fixation on a word; 0 if skipped or multiple fixations occur.
first-reading time	FRT	Sum of fixation durations from the first fixation on a word until leaving it for the first time.
total-fixation time	TFT	Sum of all fixations on a word (FPRT + RRT).
re-reading time	RRT	Sum of fixation durations after first-pass reading (TFT – FPRT).
inclusive regression-path duration	RPD_inc	Sum of fixations from first-pass fixation until moving to a word to the right, including regressions.
exclusive regression-path duration	RPD_exc	Same as RPD_inc excluding fixations on the current word.
right-bounded reading time	RBRT	Sum of fixations on a word until a word to the right is fixated.
<i>Binary measures</i>		
skipped	skipped	1 if the word is skipped, otherwise 0.
first-pass fixation	FPF	1 if the word is fixated during first-pass reading, otherwise 0.
re-reading	RR	1 if the word is fixated after first-pass reading, otherwise 0.
<i>Count / ordinal measures</i>		
total fixation count	TFC	Total number of fixations on a word.
total count of outgoing regressions	TRC_out	Total number of regressions initiated from the word.
total count of incoming regressions	TRC_in	Total number of regressions landing on the word.

Table 5: Overview of ET measures. Definitions adapted from Table 12 in [Jakobi et al. \(2025a\)](#).

Easy Reading font on a white background with wide line spacing. The font used is a highly legible dyslexia-friendly typeface specifically designed for the Italian language, with the aim of facilitating reading and reducing perceptual crowding ([Canali et al., 2021](#)). The physical display size was 34.5×19.5 cm. The Area of Interest around each word was set to 2 cm, and the approximate character height ranged from 0.5 to 1 cm. Before the reading task, the interviewer introduced the study and its objectives. After obtaining informed consent, participants were asked to provide socio-demographic information, including age, educational background, geographical area of residence, and native language. These data were collected to support subsequent analyses relating eye-tracking measures to the main socio-demographic characteristics of the sample.

Before each recording session, a standard calibration procedure was performed, yielding an average calibration error of 0.281° of visual angle, in line with the recommendations by [Hooge et al. \(2024\)](#), who suggest calibration errors below approximately 0.5° .

A.2 Eye-Tracking Measures

Table 5 reports the ET measures extracted in our preprocessing pipeline together with their definitions. The resulting feature representation is compatible with the ET measure format adopted in the MultipIEYE corpus ([Jakobi et al., 2025b](#)).

A.3 Psycholinguistic Analyses

Below we report the lme4 specifications used for the psycholinguistic analyses. Continuous models

Hyperparameter	ET FT	Benchmark FT
Optimizer	AdamW	AdamW
Learning rate	2×10^{-5}	$\{10^{-4}, 10^{-5}, 2 \times 10^{-5}, 5 \times 10^{-5}\}$
Batch size	8	64
Epochs	100	3
Sequence length	256	128
Loss function	MSELoss	CrossEntropyLoss

Table 6: Hyperparameters used for ET and downstream fine-tuning experiments.

were specified as:

$$\begin{aligned} \log(Y) \sim & 1 + \text{word_length}_z + \text{freq}_z \\ & + \text{surprisal}_z \\ & + (1 + \text{surprisal}_z \parallel \text{subject}) \\ & + (1 \mid \text{item}) \end{aligned}$$

The term $(1 + \text{surprisal}_z \parallel \text{subject})$ is shorthand for a by-subject random intercept and a by-subject random slope for surprisal, with the intercept–slope correlation constrained to zero. The binary models used the same fixed- and random-effects structure with a Bernoulli-logit link:

$$\begin{aligned} Y \sim & 1 + \text{word_length}_z + \text{freq}_z \\ & + \text{surprisal}_z \\ & + (1 + \text{surprisal}_z \parallel \text{subject}) \\ & + (1 \mid \text{item}) \end{aligned}$$

A.4 Fine-Tuning Details and ET Prediction Results

Table 6 reports the hyperparameters used for ET and downstream fine-tuning experiments with the

Test	Model	FFD	FPRT	TFT	RRT	skipped	FPF	RR
HumanRights	BERT _{LearningMobility}	0.083	0.087	0.087	0.086	0.125	0.129	0.116
	BERT _{ME_NO_HumanRights}	0.037	0.027	0.033	0.040	0.076	0.081	0.086
	BERT _{CER}	0.040	0.043	0.060	0.073	0.119	0.258	0.236
	BERT _{CER-LearningMobility}	0.071	0.080	0.078	0.075	0.105	0.108	0.098
	BERT _{CER-ME_NO_HumanRights}	0.035	0.025	0.032	0.038	0.075	0.080	0.087
LearningMobility	BERT _{HumanRights}	0.133	0.125	0.148	0.110	0.161	0.165	0.199
	BERT _{ME_NO_LearningMobility}	0.037	0.027	0.043	0.046	0.093	0.098	0.141
	BERT _{CER}	0.042	0.048	0.072	0.086	0.120	0.277	0.232
	BERT _{CER-HumanRights}	0.107	0.102	0.114	0.081	0.120	0.122	0.177
	BERT _{CER-ME_NO_LearningMobility}	0.036	0.025	0.043	0.046	0.088	0.096	0.136

Table 7: MAE across model variants and test sets. ET measures were min-max scaled to the $[0, 1]$ range before training and evaluation; Lower values indicate better performance. CER refers to CERVINO; ME denotes MultipleYE; “NO_X” indicates that text X was excluded from training; hyphenated variants denote sequential fine-tuning. Bold indicates the best score for each measure and test set.

Italian BERT model (110M parameters). All experiments were repeated across five runs on an NVIDIA T4 GPU.

For ET fine-tuning, models were trained for 100 epochs using the AdamW optimizer, a learning rate of 2×10^{-5} , a batch size of 8, and a sequence length of 256.

For downstream fine-tuning, models were trained for 3 epochs using the AdamW optimizer, a sequence length of 128, and a batch size of 64. Learning rates were selected through grid search over $\{10^{-4}, 10^{-5}, 2 \times 10^{-5}, 5 \times 10^{-5}\}$. Validation performance was monitored at each epoch, and the best-performing checkpoint was used for testing.

Table 7 reports the Mean Absolute Error (MAE) obtained by different model variants when predicting ET measures on the test sets. Models fine-tuned on ET data from a single text generally show higher errors than those trained on larger resources. The best overall results are obtained by sequentially fine-tuning on CERVINO and MultipleYE while excluding the target test text from training. This pattern is particularly evident for the continuous ET measures. For the binary measures, models fine-tuned exclusively on CERVINO perform worse than those trained on MultipleYE, suggesting that binary fixation events may be more sensitive to the lower temporal resolution of the recordings.