Legal Text Reader Profiling: Evidences from Eye Tracking and Surprisal Based Analysis

Calogero J. Scozzaro°, Davide Colla°, Matteo Delsanto°, Antonio Mastropaolo•, Enrico Mensa°, Luisa Revelli•, Daniele P. Radicioni°
°Università degli Studi di Torino, Italy; •Università della Valle d'Aosta, Italy
°{calogero.scozzaro47@edu.unito.it; first-name.surname@unito.it}
•{initial.surname@univda.it}

Abstract

Reading movements and times are a precious cue to follow reader's strategy, and to track the underlying effort in text processing. To date, many approaches are being devised to simplify texts to overcome difficulties stemming from sentences obscure, ambiguous or deserving clarification. In the legal domain, ensuring the clarity of norms and regulations is of the utmost importance, as the full understanding of such documents lies at the foundation of core social obligations and rights. This task requires determining which utterances and text excerpts are difficult for which (sort of) reader. This investigation is the aim of the present work. We propose a preliminary study based on eye-tracking data of 61 readers, with focus on individuating different reader profiles, and on predicting reading times of our readers.

Keywords: reader profiling, eye-tracking, surprisal, legal documents, surface errors, semantic errors

1. Introduction

The certainty of law and equality in accessing legal sources are basic pillars of democratic systems: since legal and normative production is predominantly written, the analysis of these sources is crucial, and Natural Language Processing (NLP) may be also central in analyzing legal documents. Various NLP applications have been carried out in the legal domain, including summarizing legal documents, question answering systems, named entity extraction, and various types of judicial support systems. A comprehensive and detailed review and discussion of the relationship between AI (at large, but also including NLP applications) and law has been recently proposed by Villata et al. (2022).

Legislative and regulatory production may contain complex, highly specialized language, lengthy and convoluted sentences that are challenging to grasp. It is featured by specific semiotic and linguistic conventions, vocabulary, semantics, syntax and morphology that may result as difficult to understand by laypeople with no domain expertise. It is thus inherently harder to process than ordinary language: for example, legal documents such as SEC contract clauses (Tuggener et al., 2020) were compared to Simple English Wikipedia (Coster and Kauchak, 2011), and it was observed that legal clauses contain seven times as many tokens than those from Wikipedia, are featured by sentences over three times longer, and by more complex parse trees, as reported by Garimella et al. (2022). Text simplification may then provide valuable insights to legal professionals, and to laypeople lacking of domain expertise, as well. A preliminary issue, connected to textual simplification, is that of characterizing what is either obscure, ambiguous or deserving clarification, thereby needing to be reformulated. Some general readability indexes exist, building on basic parameters such as the number of sentences, the number of words, and the number of syllables, such as, e.g., the Flesch–Kincaid Grade Level (Kincaid et al., 1975; Leroy and Endicott, 2012) —which was also adapted to the Italian language (Piemontese et al., 1996)—, the Dale-Chall scores (Williams, 1972), and more global scoring approaches jointly considering lexical, morpho–syntactic and syntactic features (Dell'Orletta et al., 2011). However, no decisive evidences have been reported, nor models have been proposed able to explain the mechanisms underlying reading comprehension, to predict which elements are most disturbing and undermining for human comprehension, and whether these allow to characterize different classes of readers, e.g., differentiating between expert and non-expert reading performance.

Being able to profile readers, acquiring information on which phrases and sentences mostly impact on texts readability, and whether all readers are equally affected by such sources of difficulty would be therefore highly beneficial for text simplification, and would also allow delivering *ad hoc* paraphrases and rewriting tailored to specific reader groups or user needs.

Rich instruments are to date available to investigate language processing and comprehension in the reading task, by analyzing both readers response and internal properties of texts employed in the reading tasks: in the former case (investigating readers response) we may employ eye-tracking data, and in the latter one (focused on inherent

Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024, pages 114–124

textual properties) we can analyze texts through language models. Eye tracking allows collecting precise data in form of timestamped fixations that describe and to a good extent allow to reconstruct readers' behavior. On the other side, the refinement and spread of language models allows to automatically perform subtle forms of linguistic analysis, such as determining the semantic coherence between a term and its surrounding context, thereby determining the predictability of words given their preceding context.

Several metrics have been proposed to analyze text reading and processing times. While the total reading time (TRT) —the overall duration of eye fixations for each word, including the backward regression movements— is supposed to grasp the time taken by the overall semantic integration (Radach and Kennedy, 2013), two partial and finer-grained measures have been also proposed: the duration of the first fixation (FFD) that allows estimating the cost underlying lexical access (Hofmann et al., 2022), and the number of fixations (NF), which is deemed to report about words integration in the context of what has been read so far (Frazier and Rayner, 1982).

This paper introduces the preliminary results of an experiment targeted at profiling reader's response while dealing with legal texts in Italian. To these ends we collected a corpus containing the normative production from the Aosta Valley Italian Region, composed by the Regional laws dating to the years 1960-2022 and the Regional regulations from the years between 1979 and 2022. In order to be able to gain insights on reader effort in both lexical access and semantic integration, the original utterances were manipulated and two different sorts of errors introduced: surface errors (consisting of morphological variations of terms) and semantic errors (through the introduction of unrelated terms). We present the results of a twofold experimentation: *i*) we report evidences from an eye tracking study involving 61 subjects who read a Law enacted by the Aosta Valley Region. In this setting, based on the analysis of FFDs and NFs we were able to discriminate two reader profiles exhibiting different reading strategies; and *ii*) we report a study targeted at predicting the associated reading times.

2. Background and Related Work

Two main eye movements are commonly individuated throughout the reading task, *fixations* and *saccades*. Fixations are brief stops (whose duration ranges from 50 to 1500 ms) that typically occur at each word; sometimes even more stops are needed, depending on words length and difficulty. A saccade is a fast (ranging from 10 to 100 ms) movement between each two fixations, that is used in repositioning the point of focus. In general, it is known from pioneering research in eye-tracking that individual words are fixated differently: e.g., Carpenter and Just (1983) reported that 85% content words and 35% function word get fixations. Among the main variables that impact on eye movements, one must additionally consider *i*) words length: shorter (2-3 letter) words are skipped 75% of the time, while longer (8 letter) words are fixated almost always (Rayner, 1978); and *ii*) syntactic and conceptual difficulty of the text at hand (Jacobson and Dodwell, 1979).

Eye tracking has been exploited to investigate reading at different levels, such as individual words or sentences and whole texts (Jarodzka and Brand-Gruwel, 2017). At the base level, the reading of words/sentences, regressions (backward eye movements) occurring within a single word indicate a processing problem with that word, while regressions between-words indicate comprehension problems at larger scale. A popular experimental technique employs a sliding window where parts of the text are masked (McConkie and Rayner, 1975): on such bases, different processing steps ('first pass' and 'second pass', and 'total reading times') have been hypothesized to underlie fixations and semantic processing (Rayner, 2009). Further cognitive phenomena have been also observed, such as the so-called spill-over effect (the word following an infrequent word is fixated for a longer time, while the previous word is still being processed), and the peripheral vision, that allows to perceive words that are not actually fixated. As regards as the second level, considering whole texts, the analysis typically considers sub-words or words (also AOIs, 'areas of interest') that convey specifically relevant information. An interesting measure in this setting is the 'reading depth', that measures quantities such as how much text is skipped by readers, the width of saccadic movements, and investigates strategies aimed at differentiating reading and scanning texts (such as to search for specific information). Situational models have been proposed to account for the inferential steps performed by readers and for the enrichment of read statements with prior knowledge to enforce semantic coherence (Zwaan et al., 1995). Consistent individual differences between readers also exist, associated to both lexical access and semantic integration. For instance, factors such as previous knowledge and reading expertise/ability are known to affect reading times. At the word/sentence level, good readers are more precise in targeting their regressions to the specific points that caused difficulties in comprehension; while employing prior knowledge proved beneficial for semantic integration purposes.

Most work focused on the processes underlying

lexical access and semantic integration falls into two broad approaches to model context. In the first case we have models concerned with the semantic relatedness between words and their context: in this setting, reading times are predicted based on the similarity between embeddings describing words and their context. Works adopting the second approach mostly rely on a probabilistic framework whereby words may be predicted based on their (left) context. In this view, words predictability should be intended as a function of the probability of a word given the context, and the probability of that word may work, in turn, as a main predictor of reading times: in essence, the less likely the emission of a word, the higher the *surprisal* associated to that word, and the longer the time it requires for readers to process it. Both the approaches based on relatedness and those relying on surprisal are surveyed in detail in (Salicchi et al., 2023).

In the last few years neural language models gained a central role in analyzing reading as well, since they are able to acquire conditional probability distributions over the lexicon that are also predictive of human processing times. While word length and frequency are widely acknowledged as predictors for determining lexical access, different sorts of language models have been recently compared to analyze and explain syntactic and semantic factors (Hofmann et al., 2022): N-gram models have been found to succeed in capturing short-range lexical access, while models based on recurrent neural networks show better fit in predicting the next-word. The role of model features (with focus on parameter size, spanning from 564M to 4.5B parameters) has been investigated in its impact on psychometric quality by de Varda and Marelli (2023), that challenge a widely accepted assumption postulating that the quality of predictions increases as the number of parameters grows. More specifically, also building on previous findings, such as by Shain et al. (2022), de Varda and Marelli (2023) observe that large multilingual Transformer-based models are outperformed by their smaller variants in predicting fixations, and thus are more suited to analyze lexical access and early semantic integration. Importantly enough, the authors make use of a masked language model rather than autoregressive models such as GPT (Devlin et al., 2018), thus accessing to both left and right context. Other studies found that the surprisal scores are strong predictors of reading times and eye fixations obtained through eye-tracking (Smith and Levy, 2008; Goodkind and Bicknell, 2018), along with a substantial linear relationship between models' next-word prediction accuracy and their ability in predicting reading times (Goodkind and Bicknell, 2018; Wilcox et al., 2020, 2023).

The issue of learners' reading ability has been addressed by Paracha et al. (2018), that investigated whether eye-tracking allows discriminating fluent and non-fluent students: skilled readers scan the text quickly, continuously and consistently from comprehension questions to the text, while weak readers read linearly, renouncing to select the most meaningful text elements.

3. Experiment

We start by introducing the data collected for our experiments, and then report about the experimentation: in the first experiment, we present a study on eye-tracking data of 61 persons reading a law from the Italian Region Aosta Valley and investigate their reading style when dealing with regular text, and in response to specific errors. In the second experiment we investigated whether and to what extent the fixations recorded in the former step can be predicted.

3.1. Data Collection: the Aosta Corpus

For our experiments the Aosta corpus was compiled; the corpus is composed of norms and regulations enacted by the Aosta Valley Italian Region. It contains $2,950$ Regional laws dating back to the years between 1960 and 2022, and 131 Regional regulations produced in the year between 1979 and 2022. Laws herein contain on the whole $172,669$ sentences (on average 58.53 sentences per law), $3,462,931$ tokens (on average, $1,173.87$ tokens per law), the Type-Token Ratio (TTR) is 0.546. Regulations contain on the whole $16,009$ sentences (on average 122.21 sentences per regulation), $328,931$ tokens (on average $2,510.92$ tokens per regulation), and the Type-Token Ratio (TTR) is 0.358.

From this corpus we chose the Regional Law 11/2021, 'Measures for prevention and intervention concerning the wolf species'. The choice of the Law was based on the following criteria: *i)* textual structure representative of Regional laws; *ii)* a good deal of linguistic variety ensuring the alternation of long and complex sentences and short and linear sentences; *iii)* reduced length, in order to allow for shorter reading times. By selecting a text of standard length, we would have had to present an extract, and this would have undermined the investigation of the overall understanding with post-reading questions; *iv)* the topic had to be related to a widely and socially relevant subject, rather than targeted to specific social groups. This document contains 3 articles that are further divided into 6 paragraphs, overall 32 sentences, 488 tokens, amounting to $2,783$ characters ($3,240$ including space chars), and its TTR is 0.591. Notably, the tokens were split in the same manner as they were presented to the participants during the reading

experiments, namely based on the AOIs (areas of interest: the areas actually targeted by readers fixations; more on this in Section 3.1). For example, a token such as 'finanziaria)', *financial*, was not split into 'finanziaria' and ')', but was kept as a single token.

The original text was altered to study the response of readers when dealing with errors. Overall 8 words were modified: namely, 4 errors were introduced at the surface level (e.g., a term such as 'urgenza', *urgency*, was changed to 'urrgenza'); and 4 words were replaced with existing words, such that the underlying semantics was affected by the replacement (e.g., in the phrase 'fauna selvatica', *wildlife*, 'selvatica' was changed into 'marina', with the whole meaning turning to *marine fauna*). The resulting expression is loosely related to the context of this regulation, referring to the woodland context, and more generally to the Aosta Valley Region, which is a mountainous region, far from the sea. The former modifications were expected to impact on lexical access, and the latter ones on the semantic integration.

Eye movements were recorded via an SR Research EyeLink 1000 Plus eye-tracker (spatial resolution of $0.01°$), with sampling at 1000 Hz. Participants were seated 60 cm away from a monitor with a display resolution of $1,600 \times 900$, so that approximately three characters subtended $1°$ of visual angle (the monitor was 40×24 deg of visual angle). Although viewing was binocular, eye movements were recorded from the right eye. The experiment was controlled with the SR Research Experiment Builder software.

To collect eye-tracking data 61 participants were recruited on voluntary bases, all native Italian speakers. For each participant we recorded age, level of education, occupation, region of birth/origin, mother tongue, and gender. Neither names nor other private information was asked, so that the authors had no access to information that would allow identifying the individual participants during or after data collection. The gender distribution among participants shows 23 male readers and 38 female readers; their mean age is 40.20 ± 14.70. On average, our participants received 16.41 ± 3.23 years of education. They were all informed about the aim of the eye-tracking experiment, as targeted to investigate readability issues possibly afflicting legal texts, and to individuate specific elements contributing to the difficulty of such text documents. Participants were warned to pay attention to the text meaning, and to try to understand its content, since after the reading phase they would have been interviewed about that text. Before starting they also were informed that the law text had been previously modified, with no further detail. In the first stage, after a brief training step required to calibrate the eye-tracking machinery, they started reading the aforementioned Regional Law 11/2021 from 6 slides employed to display the text through a laptop computer with 16-inch monitor, and their eye movements were recorded. After the recording of participant's eye movements, geometric areas of interest (AOI) were defined using the eye-tracking software. Each AOI is a polygon encompassing an attribute of interest within the image. In the second stage readers were asked whether they had detected any error throughout the reading, and to list the errors they could remember. The interviews were audio-recorded, and meanwhile their answers were collected in structured fashion.

3.2. Reader Profiling

3.2.1. Results

The total number of recorded fixations amounts to $38,022$. Fixations lasting less than 100 milliseconds were removed, as is customarily done in literature (Reisen et al., 2008; Salicchi et al., 2023). Specifically, $2,226$ fixations with a duration of less than 100 milliseconds were filtered out. The final number of fixations considered after the filtering process is $35,796$. Outlier readers were removed from the dataset based on the distribution of gaze plots: three readers were excluded due to an unusually low number of fixations, likely attributed to device errors, while one reader was dropped due to an exceptionally high number of fixations.

On average over AOIs, recorded total reading time (TRT) amounts to 276.64 ms, the mean number of fixations (NF) is 1.21, while the mean first fixation duration (FFD) lasted 159.77 ms; the standard deviations complementing these data are 234.18 (TRT), 0.96 (NF) and 118.64 (FFD). Such values are comparable to those in the Provo Corpus (Luke and Christianson, 2018), whose mean values (standard deviations) are 198.14 (173.03) for TRT, 0.95 (0.76) for NF, and 139.80 (107.11) for FFD (Luke and Christianson, 2018). The reliability of recorded data is also supported by the ratio between standard deviation and mean values: for our dataset these are 84.65%, 79.34%, 74.26% (for TRT, NF and FFD, respectively), and 87.33%, 80.00%, 76.62% for the Provo data. The slight increase in the average values of our dataset is likely influenced by the specialized nature of the text and the particularity of the legal domain, while the Provo Dataset contains 55 short English texts covering various topics.

By inspecting NF and FFD data —TRT was considered as a measure dependent on the previous ones—, readers can be categorized into four classes based on their mean NF and FFD values:

- class 1: readers with FFD above average and NF below average (10 subjects);

	TRT (std)	NF (std)	FFD (std)
average	276.64 (234.18)	1.21 (0.96)	159.77 (118.64)
class 1	281.23 (191.59)	1.10 (0.70)	183.32 (119.59)
class 2	371.67 (264.23)	1.57 (1.06)	186.54 (120.00)
class 3	200.74 (168.22)	0.94 (0.76)	133.21 (98.48)
class 4	274.83 (193.54)	1.34 (0.91)	142.00 (84.71)

Table 1: Mean values (and standard deviations) for total reading times (TRT), number of fixations (NF) and first fixation durations (FFD) featuring our corpus.

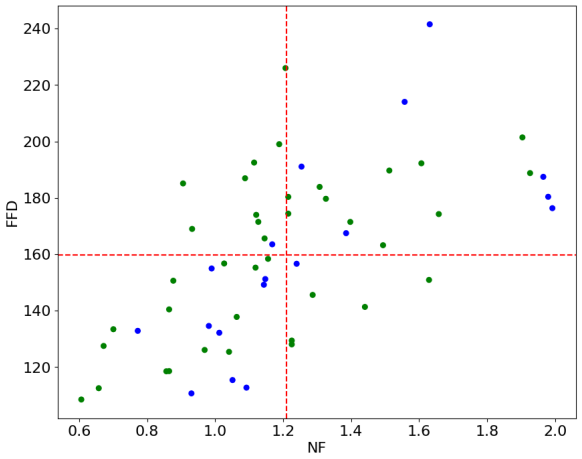

Figure 1: Plot of readers based a two-dimensional space representing NF and FFD values, with red lines indicating mean values. Class 1 is positioned on the top-left, class 2 on the top-right, class 3 on the bottom-left, and class 4 on the bottom-right. Blue points indicate readers that individuated at least 2 errors, green points those that found at most 1 error.

- class 2: with above-average FFD and above-average NF (18 subjects);
- class 3: with below-average FFD and below-average NF (23 subjects);
- class 4: with below-average FFD and above-average NF (6 subjects);

The mean values for the whole dataset and all classes are presented in Table 1; in Figure 1 we provide the plot of our readers arranged into the four classes. Classes 2 and 3 are of particular interest: class 2 identifies readers whose strategy involves higher number of fixations and longer first fixation times, while class 3 identifies readers spending less time for first fixations, and employing less fixations to read the text. After the eye-tracking session, readers were interviewed and requested to report about any errors: in this introspective effort participants were able to remember from 0 to 4 errors. Remarkably, readers that individuated at least 2 errors are mostly located either in class 2 or 3 (39% and 50%, respectively): this datum seems to suggest that the shorter the first fixation and the fewer the number of fixations, the greater the ability to identify errors. Also, 64% readers aged over 40 belong to either class 3 or 4 —thus featured by smaller FFD—, while readers under 40 are mainly (62%) found in classes 1 and 2. A correlation test was run to check whether FFD and age are (inversely) correlated, obtaining a limited Pearson correlation $\rho = -0.25$, $p < 0.058$ and a Spearman correlation $r = -0.29$, $p < 0.029$.

	CONTENT	FUNCTION
average	38.96 (21.79)	110.04 (23.33)
class 1	32.90 (21.78)	110.90 (12.85)
class 2	22.94 (10.90)	87.67 (20.07)
class 3	52.70 (19.05)	129.57 (11.66)
class 4	44.50 (19.70)	100.83 (10.75)

Table 2: Average number (std) of skips recorded in correspondence of AOIs containing content and function words.

Our categorization seems to be corroborated by the analysis of skipped AOIs: while readers from class 2 skip few (less than average) function words and few content words, almost all class 3 readers skip more function words than readers from other classes, and most of them are above average also for skipping AOIs associated to content words. By considering the number of skips, we observe that readers from class 2 consistently skip less function and content words, while those in class 3 are well above the average, as illustrated in Table 2. The regression analysis also supports our categorization: on average, we recorded 110 regressions per reader, lasting around 219 ms. The reading strategy of class 3 readers involves less (below average) and shorter (also below average) regressions, while conversely class 2 readers are featured by more and longer regressions. To complete the picture, readers from class 1 exhibit below average regressions, but lasting above average, while class 4 readers are featured by shorter but numerous regressions. These data, paired with the higher success rate in recognizing errors, seem to qualify readers from class 3 as expert readers.

The differential behavior of readers on content and function words shows that the total reading times for class 1 and 4 readers are close to the average values over all classes (which is 123.9 ms per content word syllable, and 101.86 ms per function word syllable). Readers from class 2 employ some 30% longer time than average readers to read content words and 46% on function words. Readers from class 3 save around 25% reading time on content words and 37% on function words. Detailed figures are reported in Table 3.

We investigated the response of readers when dealing with errors: for both surface and semantic errors, we observe total reading times consistently higher than for the rest of the text (please refer to Table 4). Mean total reading times are similar for both

content w	TRT (std)	NF (std)	FFD (std)
average	123.89 (93.66)	0.54 (0.38)	70.29 (43.24)
class 1	128.04 (78.41)	0.49 (0.28)	82.26 (45.56)
class 2	161.29(103.21)	0.67(0.41)	78.51 (42.21)
class 3	93.40 (70.29)	0.43 (0.31)	61.33 (37.86)
class 4	121.68 (80.77)	0.59 (0.37)	59.99 (30.42)
function w	**TRT (std)**	**NF (std)**	**FFD (std)**
average	101.86 (130.35)	0.46 (0.54)	73.24 (91.27)
class 1	102.66 (108.08)	0.42 (0.40)	84.08 (87.14)
class 2	148.94 (150.89)	0.64 (0.61)	100.51 (97.19)
class 3	64.09 (88.19)	0.31 (0.42)	51.78 (69.61)
class 4	104.10 (104.31)	0.52 (0.50)	74.60 (70.09)

Table 3: Mean values (and standard deviations), expressed in ms for TRT and FFD, characterizing fixations for content words (top) and function words (bottom); reported figures are normalized by the number of syllables.

surface	TRT (std)	NF (std)	FFD (std)
average	608.73 (468.31)	2.19 (1.64)	213.88 (149.06)
class 1	598.18 (463.38)	1.88 (1.55)	247.48 (180.93)
class 2	789.64 (519.72)	2.68 (1.70)	245.82 (173.62)
class 3	481.74 (353.50)	1.89 (1.44)	186.41 (98.64)
class 4	570.42 (356.12)	2.42 (1.55)	167.33 (64.93)
semantic	**TRT (std)**	**NF (std)**	**FFD (std)**
average	613.31 (498.64)	2.51 (2.00)	207.80 (110.43)
class 1	670.23 (438.33)	2.48 (1.59)	234.85 (133.65)
class 2	782.83 (579.59)	3.28 (2.52)	221.97 (105.42)
class 3	418.89 (319.64)	1.82 (1.17)	180.28 (94.56)
class 4	755.17 (590.17)	2.92 (2.20)	225.66 (82.11)

Table 4: Reading times relative to words containing surface (on top, tagged as 'morph.') or semantic (bottom, 'sem.') errors. Values averaged over all readers and over the four reader classes are reported.

kinds of error for the average reader: more specifically, dealing with both surface and semantic errors involved higher FFD and more fixations (NF), resulting in twice as longer total reading times (TRT) with respect to the average over the whole text (please refer to Table 1). As expected, the growth of average FFD (which is mostly concerned with lexical access) is in percentage analogous for both kinds of error; conversely, semantic errors were responsible for more consistent growth in the NF value: we recorded on average 1.21 NF per word in the overall data, which raises to 2.19 for words with surface errors, and to 2.51 for words violating the semantic/contextual integrity of the surrounding sentence. As regards as the response of readers in the four classes to the introduced errors, readers from class 3 dealing with surface errors reveal the most consistent increase over the four classes, both in the FFD values and in the average NF. It is noteworthy that half readers that correctly individuated at least 2 errors belong to this class: so readers that in general are featured by smallest FFD and NF (placed in the bottom-left corner in Figure 1) are also those with highest accuracy in identifying errors, and whose reading strategy was influenced most by errors. By recording the average number of regressions to AOIs containing errors, we observe that class 2 readers conduct an equal number of regressions compared to average readers on surface errors, and 17% more regressions on semantic errors; conversely, individuals from class 3 perform 9% more regressions than average on surface errors, and 10% less than average on semantic errors. By computing the ratio between the average number of regressions associated to AOIs containing words with errors and the average number of regressions in all other AOIs we create an index to analyze the growth of regressions corresponding to words with errors. Looking at such index, we realize that readers from class 2 conduct 1.23 (1.80) as many regressions on surface (semantic) errors, while those in class 3 conduct 2.35 (2.43) as many regressions on surface (semantic) errors.

	TRT (std)	NF (std)	FFD (std)
average	1,077.35 (816.44)	3.12 (2.34)	244.44 (226.08)
class 1	1,259.20 (1,032.94)	3.20 (2.82)	338.30 (294.52)
class 2	1,324.72 (855.07)	3.56 (2.29)	282.11 (297.22)
class 3	805.04 (550.71)	2.57 (1.66)	172.78 (88.38)
class 4	1,076.00 (821.70)	3.83 (3.18)	249.67 (85.90)

Table 5: Reading times recorded for the token 'd'urrgenza' for all readers, and the four reader classes.

In Table 5 we present the values relating to the impact of one of the four surface anomalies introduced *ad hoc*: the orthographic rendering of the 'd'urrgenza' syntagm in which the double 'r' was unduly introduced. While on average, Classes $1, 2$ and $3, 4$ exhibit comparable first fixation time duration (by construction: please refer to Table 1), in correspondence of such error, readers from classes 2 and 3 show —over the four classes— the smallest increase in their FFD, which was 1.5 times longer than for the rest of text for Class 2, and 1.3 times longer for Class 3.

3.3. Prediction of Reading Times

In this Section we describe the different models devised for the regression task aimed at predicting the three metrics TRT, NF, and FFD, and provide the obtained results.

3.3.1. Procedure

We implemented three different regression models.

– The first one is our baseline model (BL) with word-related statistics that are known to influence sentence and word processing (i.e., word frequency, word length, word position within the sentence, previous word frequency, previous word length), similar to the approach adopted by Salicchi et al. (2023).

– The second model (BL-SUR) also includes baseline features and adds surprisal scores, computed by employing a language model which is an adaptation to Italian of an English GPT-2 model (de Vries and Nissim, 2021).[1] Surprisal associated to a word w_n is defined as the negative logarithm of the probability of emitting w_n given its history $h = \{w_0, w_1, \ldots, w_{n-1}\}$: $\text{SUR}(w) = -\log P(w_n|w_0, w_1, \ldots, w_{n-1})$ (Hale, 2016).
– The third model (BL-SUR-FT) incorporates baseline features along with surprisal, computed using a fine-tuned version of the GPT-2 model obtained by exposing the language model to the laws and regulations in the AOSTA corpus, excluding 'Regional Law 11/2021'.

The regressor used is the LightGBM regressor,[2] based on the gradient boosting framework, which proved successful in the CMCL 2021 Shared Task on Eye-Tracking Prediction (Hollenstein et al., 2021; Bestgen, 2021). Gradient boosting is an ensemble learning technique based on weak learners, typically decision trees, with the objective of minimizing a given loss function. Key features of LightGBM include its leaf-wise tree growth strategy, which means that the algorithm grows the tree by expanding the leaf with the maximum delta loss instead of growing it level by level. Such strategy allows the model to find optimal split points more quickly. Moreover, a binning approach was adopted, aimed at computing optimal split points: instead of evaluating every possible split point for each feature, this strategy groups together the feature values into bins, which allows for more efficient computation. To optimize the performance of the LightGBM regressor, a comprehensive search for optimal hyperparameters was performed using a grid search technique.

The hyperparameters considered for optimization include:

– `num_leaves`: The maximum number of leaves in each tree. A range of values, such as $[4, 5, 8, 10, 20, 30]$ was explored to identify the optimal balance between model complexity and generalization.
– `learning_rate`: The step size at each iteration during training. Different learning rates $(0.1, 0.05, 0.005)$ were investigated to speed up convergence.
– `n_estimators`: The number of trees to be built. Various values $(50, 100, 200, 500)$ were tested in this setting to determine the optimal number of trees to achieve a balance between underfitting and overfitting.
– `max_depth`: The maximum depth of each tree. Values such as $[-1, 3, 5]$ were explored to control the complexity of individual trees.

The optimization process specifically targeted the mean absolute error (MAE). The evaluation of different parameter combinations was performed through a 5-fold cross-validation strategy during the grid search. This approach guarantees robustness and reliability in evaluating the model's generalization capabilities, while explicitly focusing on minimizing the MAE for optimal predictive accuracy.

[1] https://huggingface.co/GroNLP/gpt2-small-italian.

[2] https://lightgbm.readthedocs.io

	TRT (std)	NF (std)	FFD (std)
avg	20.80 (17.61)	25.20 (19.99)	12.00 (8.91)
class 1	21.38 (14.57)	23.22 (14.78)	13.94 (9.09)
class 2	28.26 (20.09)	33.22 (22.43)	14.18 (9.12)
class 3	14.69 (12.31)	18.97 (15.34)	9.76 (7.22)
class 4	20.90 (14.72)	28.31 (19.23)	10.80 (6.44)

Table 6: Figures obtained after scaling the data reported in Table 1: TRT and FFD (that are expressed as ms) were scaled based on the maximum value of TRT, while NF values were scaled based on their maximum.

3.3.2. Results

To evaluate our models we computed the Mean Absolute Error (MAE), which is a standard measure in this setting. That is, given n as the number of tokens, y_i as the actual value for i, and $\hat{y}_i$ as the predicted value for i, $\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$. We also report MAE/mean scores. In fact, while MAE grasps the average difference between predicted and actual values, which is an absolute value, the latter metric scales such figures with respect to mean values, thus informing on the proportional magnitude of the error. Before computing the MAE, our features were scaled between 0 and 100, following the methodology described by Hollenstein et al. (2021).[3] The final scaled values are provided in Table 6.

We found that our best-performing model is the BL-surprisal with fine-tuning (BL-SUR-FT), whose error estimates are presented in Table 7.

By looking at the four reader classes, we obtained most favorable prediction of reading times on class 3, where we observe lowest MAE through all three measures, with MAE/mean also confirming that the predictions on readers from this class are more reliable than those on subjects from other classes. Through all classes lexical access seems to be more easily predicted than the semantic integration: consistent with the findings by Hollenstein

[3] TRT and FFD were jointly scaled as they are both measured in milliseconds (but we diverged from the approach used in the aforementioned study, due to the absence of the "go-past-time" (GPT) feature in the present setting, where we used TRT), while NF was independently scaled.

	TRT	NF	FFD
	MAE ($\frac{\text{MAE}}{\text{mean}}$)	MAE ($\frac{\text{MAE}}{\text{mean}}$)	MAE ($\frac{\text{MAE}}{\text{mean}}$)
average	4.14 (0.20)	4.52 (0.18)	1.81 (0.15)
class 1	5.90 (0.28)	5.70 (0.25)	3.25 (0.23)
class 2	6.64 (0.24)	6.77 (0.20)	2.53 (0.18)
class 3	3.43 (0.23)	4.29 (0.23)	1.84 (0.19)
class 4	6.94 (0.33)	8.54 (0.30)	2.81 (0.26)

Table 7: MAE (MAE/mean) values obtained through the BL-SUR-FT model implementing the baseline enriched with surprisal scores computed through a model fine-tuned on the AOSTA corpus.

et al. (2021), FFD confirms to be more accurately predicted than TRT and NF, that are acknowledged to grasp reader's effort throughout the semantic processing stage.

3.4. Discussion

A basic reader profiling was performed by partitioning readers based on their average number of fixations and on the duration of their first fixations. It is known that such measures can be considered as a proxy for different significant stages in linguistic processing.

As regards as the first task, aimed at reader profiling, two main reader classes were identified, that cover around 72% of those who participated in our experiments: if we wanted to resort to simplistic labels, we found fast and slow readers. We closely examined our data, and found that different views on data suggest that two main approaches to reading may be individuated: those employing less and faster fixations, slightly more accurate in individuating errors, skipping more words than average reader (possibly adapting skips to function and content words), employing less and shorter regressions even when dealing with errors in the text. In the other class we have a reading style involving more and longer fixations, less accurate in individuating errors, that are not familiar with skipping words, employing more and longer regressions, with reduced differences between content and function words, less sensitive to errors, and to the different types of error. Furthermore, we found an interesting (though weak) correlation of some variables with socio-demographic descriptors, such as that between FFD and readers age. Such elements might be helpful in refining reader profiles, and in investigating reading effort: such investigation will be addressed in future work.

As regards as the second task, aimed at predicting reading times, a thorough comparison with results available in literature can be hardly obtained, since differences may stem from factors that cannot be accounted for, such as the intrinsic properties of texts at hand. The recorded error on the number of fixations prediction is in line with the results in literature, e.g. by Hollenstein et al. (2021), but the documents in our corpus differ from those employed in the cited work: we dealt with the Italian Language (whose structure differs from English, with longer sentences and even different word lengths (Smith, 2012)), and our corpus includes Italian laws and regulations, against sentences from movie reviews borrowed from the Stanford Sentiment Treebank (Socher et al., 2013) and Wikipedia (Culotta et al., 2006). Additionally, our documents contain both surface and semantic errors that made more complex the task of predicting reading times, and individuals not necessarily expert in legal language were recruited. The greater difficulty of these texts is evidenced by the average NF featuring our data: after scaling this amounts to 25.2 (please refer to Table 6), while in the paper by Hollenstein et al. (2021) this datum is 15.1. Predicting reading times for the four reader classes turned out to be very challenging: MAE (and MAE/mean, too) is always higher than for average readers. Among classes, reading times of subjects in class 3 were those predicted with minimum error. Probabilistic language modeling, as a device able to describe the incremental mechanisms underlying language processing should be helpful to investigate the different reading strategies. Such strategies are basically concerned with planning and handling expectations on what follows, and on evaluating how these match with actual stimuli (Levy, 2008); surprisal was plugged into our models to support the prediction of reading times by also accounting for the difficulty of predicting words. Although it contributed to refining the baseline model, especially after the fine-tuning step, further work is needed to further improve the accuracy in the prediction of reading times.

4. Conclusions

In this work we have introduced a new dataset collecting Regional laws and regulations in Italian. One of these laws was modified by inserting 8 errors, and used for an eye-tracking experiment in which 61 readers were tracked. Collected data were utilized for reader profiling purposes and to predict their reading times. In the former case we individuated two main groups exhibiting rather different reading styles to cope with general text and with errors therein. In the latter experiment we applied an approach based on the gradient boosting framework; our best performing model also makes use of surprisal scores obtained through an Italian porting of a GPT-2 model fine-tuned on the set of Regional legal documents, consistent with the document used for experimentation. While the prediction of reading proved to be in line with results reported in literature, predicting the reading times of

the subjects in the two main classes individuated in the former experiment revealed a very challenging task.

Since the Aosta Valley is a bilingual (Italian and French) Region, and its body of regulations and laws is thus a naturally parallel corpus, in future work we will collect French documents and eye-tracking data on these. We will also investigate whether text difficulty and errors interact with cognitive load and how such temporal factors affect readers' performance, by examining how fixations and regressions vary through time. Finally, by considering the entire Aosta Corpus from 1960 to 2022, it would be interesting to analyze the evolution of the legal lexicon and language from a diachronic perspective, and to investigate whether older and more recent language differently impact on the reading task.

Acknowledgments

This work was carried out in the frame of the project 'The accessibility of regulatory texts as a tool for inclusion: case study and applicative tools in Valle d'Aosta', based at the University of Valle d'Aosta,[4] financed by the CRT Foundation, 2021 and 2022.

The eye-tracking data collection was made possible by the Human Science and Technologies laboratories from the University of Turin;[5] we are specially grateful to Prof. Francesca Garbarini, Prof. Olga Dal Monte and Dr. Monia Cariola for their keen and generous support.

5. Bibliographical References

Yves Bestgen. 2021. LAST at CMCL 2021 shared task: Predicting gaze data during reading with a gradient boosting decision tree approach. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 90–96, Online. Association for Computational Linguistics.

Patricia A Carpenter and Marcel Adam Just. 1983. What your eyes do while your mind is reading11this research was partially supported by grant g-79-0119 from the national institute of education and grant mh-29617 from the national institute of mental health. *Eye movements in reading*, pages 275–307.

William Coster and David Kauchak. 2011. Simple english wikipedia: a new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 665–669.

Aron Culotta, Andrew McCallum, and Jonathan Betz. 2006. Integrating probabilistic extraction models and data mining to discover relations and patterns in text. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 296–303.

Andrea de Varda and Marco Marelli. 2023. Scaling in cognitive modelling: a multilingual approach to human reading times. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 139–149, Toronto, Canada. Association for Computational Linguistics.

Wietse de Vries and Malvina Nissim. 2021. As good as new. how to successfully recycle english gpt-2 to make models for other languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Lyn Frazier and Keith Rayner. 1982. Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive psychology*, 14(2):178–210.

Aparna Garimella, Abhilasha Sancheti, Vinay Aggarwal, Ananya Ganesh, Niyati Chhaya, and Nanda Kambhatla. 2022. Text simplification for legal domain: Insights and Challenges. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 296–304.

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th workshop on cognitive modeling and computational linguistics (CMCL 2018)*, pages 10–18.

John Hale. 2016. Information-theoretical complexity metrics. *Language and Linguistics Compass*, 10(9):397–412.

[4] Original title: 'L'accessibilità dei testi normativi come dispositivo di inclusione: studio di caso e strumenti applicativi in Valle d'Aosta con sede presso l'Università della Valle d'Aosta'.

[5] https://www.hst.unito.it

Markus J Hofmann, Steffen Remus, Chris Biemann, Ralph Radach, and Lars Kuchinke. 2022. Language models explain word reading times better than empirical predictability. *Frontiers in Artificial Intelligence*, 4:730570.

Nora Hollenstein, Emmanuele Chersoni, Cassandra L. Jacobs, Yohei Oseki, Laurent Prévot, and Enrico Santus. 2021. CMCL 2021 shared task on eye-tracking prediction. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 72–78, Online. Association for Computational Linguistics.

J Zachary Jacobson and Peter C Dodwell. 1979. Saccadic eye movements during reading. *Brain and Language*, 8(3):303–314.

Halszka Jarodzka and Saskia Brand-Gruwel. 2017. Tracking the reading eye: Towards a model of real-world reading.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and Flesch reading ease formula) for navy enlisted personnel.

Gondy Leroy and James E Endicott. 2012. Combining nlp with evidence-based methods to find text metrics related to perceived and actual text difficulty. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, pages 749–754.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Steven G Luke and Kiel Christianson. 2018. The provo corpus: A large eye-tracking corpus with predictability norms. *Behavior research methods*, 50:826–833.

George W McConkie and Keith Rayner. 1975. The span of the effective stimulus during a fixation in reading. *Perception & Psychophysics*, 17:578–586.

Samiullah Paracha, Ayaka Inuoue, and Sania Jehanzeb. 2018. Detecting online learners' reading ability via eye-tracking. In *Optimizing Student Engagement in Online Learning Environments*, pages 163–185. IGI Global.

Maria Emanuela Piemontese, M Piemontese, et al. 1996. Capire e farsi capire. teorie e tecniche della scrittura controllata.

Ralph Radach and Alan Kennedy. 2013. Eye movements in reading: Some theoretical context. *The Quarterly journal of experimental psychology*, 66(3):429–452.

Keith Rayner. 1978. Eye movements in reading and information processing. *Psychological bulletin*, 85(3):618.

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.

Nils Reisen, Ulrich Hoffrage, and Fred W Mast. 2008. Identifying decision strategies in a consumer choice situation. *Judgment and decision making*, 3(8):641–658.

Lavinia Salicchi, Emmanuele Chersoni, and Alessandro Lenci. 2023. A study on surprisal and semantic relatedness for eye-tracking data prediction. *Frontiers in Psychology*, 14:1112365.

Cory Shain, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger Philip Levy. 2022. Large-scale evidence for logarithmic effects of word predictability on reading time.

Nathaniel J Smith and Roger Levy. 2008. Optimal processing times in reading: a formal model and empirical investigation. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 30.

Reginald Smith. 2012. Distinct word length frequencies: distributions and symbol entropies. *Glottometrics*, 23:7–22.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Don Tuggener, Pius Von Däniken, Thomas Peetz, and Mark Cieliebak. 2020. Ledgar: A large-scale multi-label corpus for text classification of legal provisions in contracts. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1235–1241.

Serena Villata, Michal Araszkiewicz, Kevin Ashley, Trevor Bench-Capon, L Karl Branting, Jack G Conrad, and Adam Wyner. 2022. Thirty years of artificial intelligence and law: the third decade. *Artificial Intelligence and Law*, 30(4):561–591.

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, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy. 2020. On the predictive power of neural language models for human real-time comprehension behavior. *CoRR*, abs/2006.01912.

Robert T Williams. 1972. A table for rapid determination of revised dale-chall readability scores. *The Reading Teacher*, 26(2):158–165.

Rolf A Zwaan, Mark C Langston, and Arthur C Graesser. 1995. The construction of situation models in narrative comprehension: An event-indexing model. *Psychological science*, 6(5):292–297.