# That Looks Hard: Characterizing Linguistic Complexity in Humans and Language Models

**Gabriele Sarti** [1,2] **Dominique Brunato** [2] **Felice Dell'Orletta** [2]

[1] University of Trieste, International School for Advanced Studies (SISSA), Trieste
[2] Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC-CNR), Pisa
ItaliaNLP Lab – *italianlp.it*
gabriele.sarti996@gmail.com
{dominique.brunato, felice.dellorletta}@ilc.cnr.it

## Abstract

This paper investigates the relationship between two complementary perspectives in the human assessment of sentence complexity and how they are modeled in a neural language model (NLM). The first perspective takes into account multiple online behavioral metrics obtained from eye-tracking recordings. The second one concerns the offline perception of complexity measured by explicit human judgments. Using a broad spectrum of linguistic features modeling lexical, morpho-syntactic, and syntactic properties of sentences, we perform a comprehensive analysis of linguistic phenomena associated with the two complexity viewpoints and report similarities and differences. We then show the effectiveness of linguistic features when explicitly leveraged by a regression model for predicting sentence complexity and compare its results with the ones obtained by a fine-tuned neural language model. We finally probe the NLM's linguistic competence before and after fine-tuning, highlighting how linguistic information encoded in representations changes when the model learns to predict complexity.

## 1 Introduction

From a human perspective, linguistic complexity concerns difficulties encountered by a language user during sentence comprehension. The source of such difficulties is commonly investigated using either *offline measures* or *online behavioral metrics*. In the offline framework, complexity ratings can be elicited either by assessing errors in comprehension tests or collecting explicit complexity judgments from readers. Instead, in the online paradigm, cognitive signals are collected mainly through specialized machinery (e.g., MRI scanners, eye-tracking systems) during natural or task-oriented reading. Among the wide range of online complexity metrics, gaze data are widely regarded as reliable proxies of processing difficulties, reflecting both low and high-level complexity features of the input (Rayner, 1998; Hahn and Keller, 2016). Eye-tracking measures have recently contributed to significant improvements across many popular NLP applications (Hollenstein et al., 2019a, 2020) and in particular on tasks related to linguistic complexity such as *automatic readability assessment* (ARA) (Ambati et al., 2016; Singh et al., 2016; González-Garduño and Søgaard, 2018), obtaining meaningful results for sentence-level classification in easy and hard-to-read categories (Vajjala and Lučić, 2018; Evaldo Leal et al., 2020; Martinc et al., 2021). However, readability levels are conceptually very different from cognitive processing metrics since ARA corpora are usually built in an automated fashion from parallel documents at different readability levels, without explicit evaluations of complexity by target readers (Vajjala and Lučić, 2019). A different approach to complexity assessment that directly accounts for the perspective of readers is presented in the corpus by Brunato et al. (2018), where sentences are individually labeled with the *perception of complexity* of annotators, which may better reflect the underlying cognitive processing required by readers to parse the sentence. This consideration is supported by recent results highlighting the unpredictability of outliers in perceived complexity annotations, especially for sentences having complex syntactic structures (Sarti, 2020).

Given the relation between complexity judgments elicited from annotators and online cognitive processing metrics, we investigate whether the connection between the two perspectives can be highlighted empirically in human annotations and language model representations. We begin by leveraging linguistic features associated with a variety of sentence-level structural phenomena and analyzing their correlation with offline and online complexity metrics. We then evaluate the performance of models using either complexity-related explicit



*Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 48–60

features or contextualized word embeddings, focusing mainly on the neural language model ALBERT (Lan et al., 2020). In this context, we show how both explicit features and learned representations obtain comparable results when predicting complexity scores. Finally, we focus on studying how complexity-related properties are encoded in the representations of ALBERT. This perspective goes in the direction of exploiting human processing data to address the interpretability issues of unsupervised language representations (Hollenstein et al., 2019b; Gauthier and Levy, 2019; Abnar et al., 2019). To this end, we rely on the *probing task* approach, a recently introduced technique within the area of NLMs interpretability consisting of training diagnostic classifiers to probe the presence of encoded linguistic properties inside contextual representations (Conneau et al., 2018; Zhang and Bowman, 2018). We observe that fine-tuning on online and offline complexity produces a consequent increase in probing performances for complexity-related features during our probing experiments. This investigation has the specific purpose of studying whether and how learning a new task affects the linguistic properties encoded in pretrained representations. In fact, while pre-trained models have been widely studied using probing methods, the effect of fine-tuning on encoded information was seldom investigated. For example, Merchant et al. (2020) found that fine-tuning does not impact heavily the linguistic information implicitly learned by the model, especially when considering a supervised probe closely related to a downstream task. Miaschi et al. (2020) further demonstrated a positive correlation between the model's ability to solve a downstream task on a specific input sentence and the related linguistic knowledge encoded in a language model. Nonetheless, to our knowledge, no previous work has taken into account sentence complexity assessment as a fine-tuning task for NLMs. Our results suggest that the model's competencies during training are interpretable from a linguistic perspective and are possibly related to its predictive capabilities for complexity assessment.

**Contributions** To our best knowledge, this is the first work displaying the connection between online and offline complexity metrics and studying how they are represented by a neural language model. We a) provide a comprehensive analysis of linguistic phenomena correlated with eye-tracking data and human perception of complexity, addressing similarities and differences from a linguistically-motivated perspective across metrics and at different levels of granularity; b) compare the performance of models using both explicit features and unsupervised contextual representations when predicting online and offline sentence complexity; and c) show the natural emergence of complexity-related linguistic phenomena in the representations of language models trained on complexity metrics.[1]

| Metric Level | Description | Label |
|---|---|---|
| Offline (Perceptual) | Perceived complexity annotation on a 1-to-7 Likert scale. | PC |
| Online (Early) | Duration of the first reading pass in milliseconds. | FPD |
| Online (Late) | Total fixation count | FXC |
| | Total duration of all fixations in milliseconds | TFD |
| Online (Contextual) | Duration of outbound regressive saccades in milliseconds | TRD |

Table 1: Sentence-level complexity metrics. We refer to the entire set of gaze metrics as ET (eye-tracking).

| | Perc. Complexity | Eye-tracking |
|---|---|---|
| domain | news articles | literature |
| aggregation | avg. annotators | words sum + avg. participants |
| filtering | IAA + duplicates | min length |
| # sentences | 1115 | 4041 |
| # words | 21723 | 52131 |
| avg. sent. length | 19.48 | 12.90 |
| avg. word length | 4.95 | 4.60 |

Table 2: Descriptive statistics of the two sentence-level corpora after the preprocessing procedure.

## 2 Data and Preprocessing

Our study leverages two corpora, each capturing different aspects of linguistic complexity:

**Eye-tracking** For online complexity metrics, we used the monolingual English portion of GECO (Cop et al., 2017), an eye-tracking corpus based on the novel "The Mysterious Case at Styles" by Agatha Christie. The corpus consists of 5,386 sentences annotated at word-level with eye-movement records of 14 English native speakers. We select four online metrics spanning multiple

[1]Code and data available at https://github.com/gsarti/interpreting-complexity

| Annotation Level | Linguistic Feature Description | Label |
|---|---|---|
| Raw Text | Sentence length (tokens), word length (characters) | n_tokens, char_per_tok |
| | Words and lemmas type/token ratio | ttr_form, ttr_lemma |
| POS Tagging | Distribution of UD and language-specific POS tags | upos_dist_*, xpos_dist_* |
| | Lexical density | lexical_density |
| | Inflectional morphology of auxiliaries (mood, tense) | aux_mood_*, aux_tense_* |
| Dependency Parsing | Syntactic tree depth | parse_depth |
| | Average and maximum length of dependency links | avg_links_len, max_links_len |
| | Number and average length of prepositional chains | n_prep_chains, prep_chain_len |
| | Relative ordering of main elements | subj_pre, subj_post, obj_pre, obj_post |
| | Distribution of dependency relations | dep_dist_* |
| | Distribution of verbal heads | vb_head_per_sent |
| | Distribution of principal and subordinate clauses | princ_prop_dist, sub_prop_dist |
| | Average length of subordination chains | sub_chain_len |
| | Relative ordering of subordinate clauses | sub_post, sub_pre |

Table 3: Description of sentence-level linguistic features employed in our study.

phases of cognitive processing, which are widely considered relevant proxies for linguistic processing in the brain (Demberg and Keller, 2008; Vasishth et al., 2013). We sum-aggregate those at sentence-level and average their values across participants to obtain the four online metrics presented in Table 1. As a final step to make the corpus more suitable for linguistic complexity analysis, we remove all utterances with fewer than 5 words. This design choice is adopted to ensure consistency with the perceived complexity corpus by Brunato et al. (2018).

**Perceived Complexity** For the offline evaluation of sentence complexity, we used the English portion of the corpus by Brunato et al. (2018). The corpus contains 1,200 sentences taken from the Wall Street Journal section of the Penn Treebank (McDonald et al., 2013) with uniformly-distributed lengths ranging between 10 and 35 tokens. Each sentence is associated with 20 ratings of perceived-complexity on a 1-to-7 point scale. Ratings were assigned by English native speakers on the CrowdFlower platform. To reduce the noise produced by the annotation procedure, we removed duplicates and sentences for which less than half of the annotators agreed on a score in the range $\mu_n \pm \sigma_n$, where $\mu_n$ and $\sigma_n$ are respectively the average and standard deviation of all annotators' judgments for sentence $n$. Again, we average scores across annotators to obtain a single metric for each sentence.

Table 2 presents an overview of the two corpora after preprocessing. The resulting eye-tracking (ET) corpus contains roughly four times more sentences than the perceived complexity (PC) one, with shorter words and sentences on average.

## 3 Analysis of Linguistic Phenomena

As a first step to investigate the connection between the two complexity paradigms, we evaluate the correlation of online and offline complexity labels with linguistic phenomena modeling a number of properties of sentence structure. To this end, we rely on the Profiling-UD tool (Brunato et al., 2020) to annotate each sentence in our corpora and extract from it ∼100 features representing their linguistic structure according to the Universal Dependencies formalism (Nivre et al., 2016). These features capture a comprehensive set of phenomena, from basic information (e.g. sentence and word length) to more complex aspects of sentence structure (e.g. parse tree depth, verb arity), including properties related to sentence complexity at different levels of description. A summary of most relevant features in our analysis is presented in Table 3.

Figure 1 reports correlation scores for features showing a strong connection ($|\rho| > 0.3$) with at least one of the evaluated metrics. Features are ranked using their Spearman's correlation with complexity metrics, and scores are leveraged to highlight the relation between linguistic phenomena and complexity paradigms. We observe that features showing a significant correlation with eye-tracking metrics are twice as many as those correlating with PC scores and generally tend to have higher coefficients, except for total regression duration (TRD). Nevertheless, the most correlated features are the same across all metrics. As expected, sentence length (*n_tokens*) and other related fea-

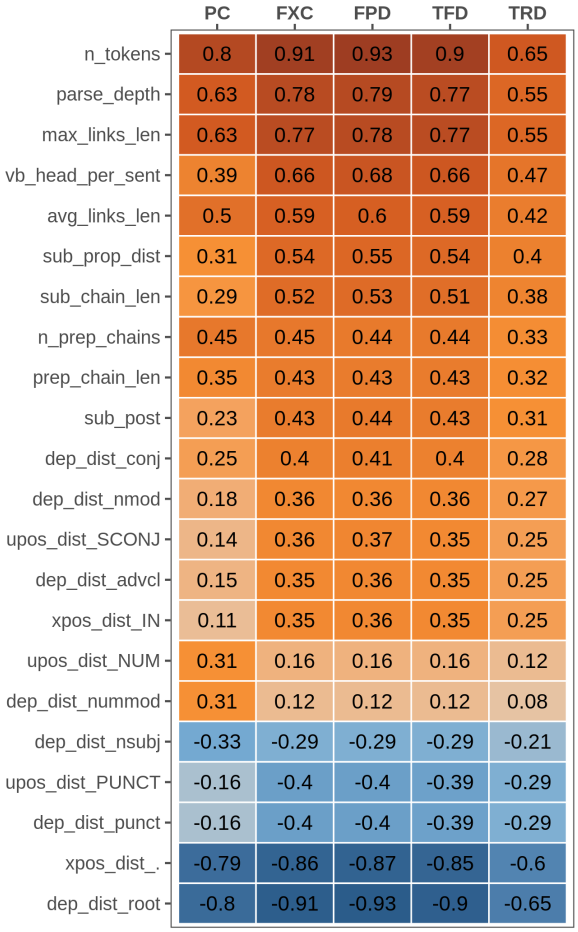


Figure 1: Ranking of the most correlated linguistic features for selected metrics. All Spearman's correlation coefficients have $p < 0.001$.

tures capturing aspects of structural complexity occupy the top positions in the ranking. Among those, we also find the length of dependency links (*max_links_len, avg_links_len*) and the depth of the whole parse tree or selected sub-trees, i.e. nominal chains headed by a preposition (*parse_depth, n_prep_chains*). Similarly, the distribution of subordinate clauses (*sub_prop_dist*, *sub_post*) is positively correlated with all metrics but with stronger effect for eye-tracking ones, especially in presence of longer embedded chains (*sub_chain_len*). Interestingly, the presence of numbers (*upos_NUM*, *dep_nummod*) affects only the explicit perception of complexity while it is never strongly correlated with all eye-tracking metrics. This finding is expected since numbers are very short tokens and, like other functional POS, were never found to be strongly correlated with online reading in our results. Conversely, numerical information has been identified as a factor hampering sentence readability and understanding (Rello et al., 2013).

Unsurprisingly, sentence length is the most correlated predictor for all complexity metrics. Since many linguistic features highlighted in our analysis are strongly related to sentence length, we tested whether they maintain a relevant influence when this parameter is controlled. To this end, Spearman's correlation was computed between features and complexity tasks, but this time considering bins of sentences having approximately the same length. Specifically, we split each corpus into 6 bins of sentences with 10, 15, 20, 25, 30 and 35 tokens respectively, with a range of $\pm 1$ tokens per bin to select a reasonable number of sentences for our analysis.

Figure 2 reports the new rankings of the most correlated linguistic features within each bin across complexity metrics ($|\rho| > 0.2$). Again, we observe that features showing a significant correlation with complexity scores are fewer for PC bins than for eye-tracking ones. This fact depends on controlling for sentence length but also on the small size of bins for the whole dataset. As in the coarse-grained analysis, TRD is the eye-tracking metric less correlated to linguistic features, while the other three (FXC, FPD, TFD) show a homogeneous behavior across bins. For the latters, vocabulary-related features (token-type ratio, average word length, lexical density) are always ranked on top (and with a positive correlation) in all bins, especially when considering shorter sentences (i.e. from 10 to 20 tokens). For PC, this is true only for some of them (i.e. word length and lexical density). At the same time, features encoding numerical information are still highly correlated with the explicit perception of complexity in almost all bins. Interestingly, features modeling subordination phenomena extracted from fixed-length sentences exhibit a reverse trend than when extracted from the whole corpus, i.e. they are negatively correlated with judgments. If, on the one hand, we expect an increase in the presence of subordination for longer sentences (possibly making sentences more convoluted), on the other hand, when length is controlled, our findings suggest that subordinate structures are not necessarily perceived as a symptom of sentence complexity. Our analysis also highlights that PC's relevant features are significantly different from those correlated to online eye-tracking metrics when controlling for sentence length. This aspect wasn't evident from the previous coarse-grained analysis. We note that, despite controlling sentence length,

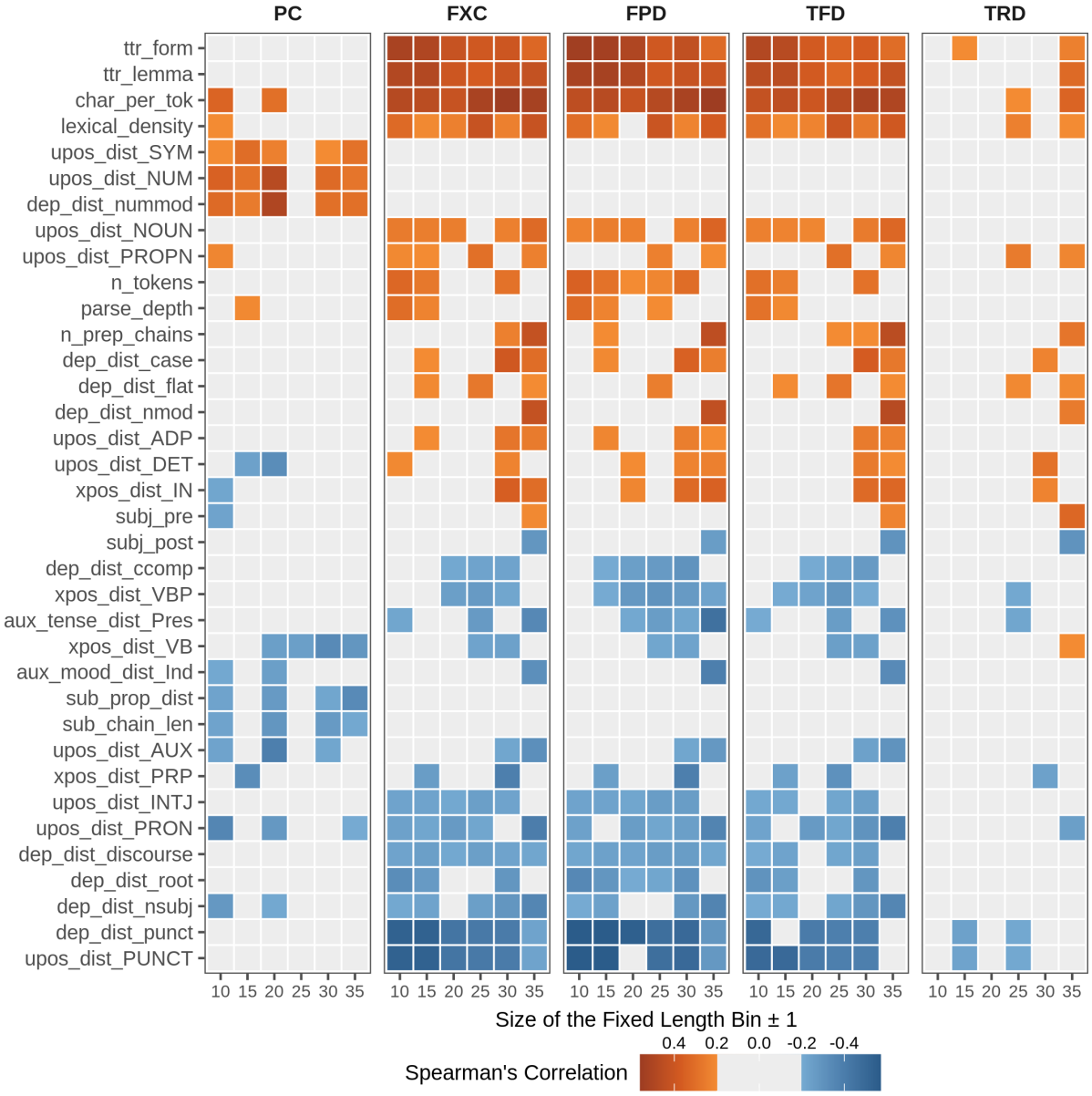


Figure 2: Rankings of the most correlated linguistic features for metrics within length-binned subsets of the two corpora. Coefficients $\geq$ 0.2 or $\leq$ -0.2 are highlighted, and have $p < 0.001$. (Bins from 10 to 35 have sizes of 173, 163, 164, 151, 165, and 147 sentences for PC and 899, 568, 341, 215, 131, and 63 sentences for gaze metrics.)

gaze measures are still significantly connected to length-related phenomena. This can be possibly due to the $\pm 1$ margin applied for sentence selection and the high sensitivity of behavioral metrics to small changes in the input.

## 4 Predicting Online and Offline Linguistic Complexity

Given the high correlations reported above, we proceed to quantify the importance of explicit linguistic features from a modeling standpoint. Table 4 presents the RMSE and $R^2$ scores of predictions made by baselines and models for the selected complexity metrics. Performances are tested with a 5-fold cross-validation regression with fixed random seed on each metric. Our baselines use average metric scores of all training sentences (Average) and average scores of sentences binned by their length in # of tokens (Length-binned average) as predictions. The two linear SVM models leverage explicit linguistic features, using respectively only n_tokens (SVM length) and the whole set of $\sim$100 features (SVM feats). Besides those, we also test the performances of a state-of-the-art Transformer neural language model relying entirely on contextual word embeddings. We selected ALBERT as a

| | PC | | FXC | | FPD | | TFD | | TRD | |
|---|---|---|---|---|---|---|---|---|---|---|
| | RMSE | $R^2$ | RMSE | $R^2$ | RMSE | $R^2$ | RMSE | $R^2$ | RMSE | $R^2$ |
| Average | .87 | .00 | 6.17 | .06 | 1078 | .06 | 1297 | .06 | 540 | .03 |
| Length-binned average | .53 | .62 | 2.36 | .86 | 374 | .89 | 532 | .85 | 403 | .45 |
| SVM length | .54 | .62 | 2.19 | .88 | 343 | .90 | 494 | .86 | 405 | .45 |
| SVM feats | **.44** | .74 | **1.77** | **.92** | **287** | **.93** | **435** | .89 | 400 | .46 |
| ALBERT | **.44** | **.75** | 1.98 | .91 | 302 | **.93** | **435** | **.90** | **382** | **.49** |

Table 4: Average Root-Mean-Square Error and $R^2$ for complexity predictions of two average baselines, two SVMs relying on explicit features and a pretrained language model with contextualized word embeddings using 5-fold cross-validation. ALBERT learns eye-tracking metrics in a multitask setting over parallel annotations.

lightweight yet effective alternative to BERT (Devlin et al., 2019) for obtaining contextual word representations, using its last-layer [CLS] sentence embedding as input for a linear regressor during fine-tuning and testing. We selected the last layer representations, despite having strong evidence on the importance of intermediate representation in encoding language properties, because we aim to investigate how final layers encode complexity-related competences. Given the availability of parallel eye-tracking annotations, we train ALBERT using multitask learning with hard parameter sharing (Caruana, 1997) on gaze metrics.[2]

From results in Table 4 we note that: i) the length-binned average baseline is very effective in predicting complexity scores and gaze metrics, which is unsurprising given the extreme correlation between length and complexity metrics presented in Figure 1; ii) the SVM feats model shows considerable improvements if compared to the length-only SVM model for all complexity metrics, highlighting how length alone accounts for much but not for the entirety of variance in complexity scores; and iii) ALBERT performs on-par with the SVM feats model on all complexity metrics despite the small dimension of the fine-tuning corpora and the absence of explicit linguistic information. A possible interpretation of ALBERT's strong performances is that the model implicitly develops competencies related to phenomena encoded by linguistic features while training on online and offline complexity prediction. We explore this perspective in Section 5.

As a final step in the study of feature-based models, we inspect the importance accorded by the SVM feats model to features highlighted in previous sections. Table 5 presents coefficient ranks produced by SVM feats for all sentences and for the 10±1 length bin, which was selected as the broadest subset. Despite evident similarities with the previous correlation analysis, we encounter some differences that are possibly attributable to the model's inability in modeling non-linear relations. In particular, the SVM model still finds sentence length and related structural features highly relevant for all complexity metrics. However, especially for PC, lexical features also appear in the top positions (e.g. *lexical density*, *ttr_lemma*, *char_per_tok*), as well as specific features related to verbal predicate information (e.g. *xpos_dist_VBZ,_VBN*). This holds both for all sentences, and when considering single length-binned subsets. While in the correlation analysis eye-tracking metrics were almost indistinguishable, those behave quite differently when considering how linguistic features are used for inference by the linear SVM model. In particular, the fixation count metric (FXC) consistently behaves in a different way if compared to other gaze measures, even when controlling for length.

## 5 Probing Linguistic Phenomena in ALBERT Representations

As shown in Table 4, ALBERT performances on the PC and eye-tracking corpora are comparable to those obtained using a linear SVM with explicit linguistic features. To investigate if ALBERT encodes the linguistic knowledge that we identified as strongly correlated with online and perceived sentence complexity during training and prediction, we adopt the *probing task* testing paradigm. The aim of this analysis is two-fold: i) probing the presence of complexity-related information encoded by ALBERT representations during the pre-training

[2] Additional information on parameters and chosen training approach is presented in Appendix A.

| | All Sentences | | | | | Bin 10±1 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | PC | FXC | FPD | TFD | TRD | PC | FXC | FPD | TFD | TRD |
| n_tokens | 1 | 1 | 1 | 1 | 1 | -36 | 5 | 1 | 1 | 2 |
| char_per_tok | 2 | 2 | 12 | 10 | 16 | 3 | 1 | 3 | 3 | 19 |
| xpos_dist_VBN | 5 | -37 | 76 | 77 | 75 | 28 | 9 | 26 | 21 | 42 |
| avg_links_len | 6 | -6 | 7 | 7 | 7 | 11 | -8 | -23 | -30 | -46 |
| n_prep_chains | 7 | 3 | 10 | 9 | 8 | -44 | 16 | 50 | 41 | 48 |
| dep_dist_compound | 9 | 7 | 58 | 61 | 49 | 13 | 12 | 60 | 51 | 47 |
| vb_head_per_sent | 10 | 4 | 4 | 6 | 3 | 2 | -9 | 31 | 36 | -33 |
| max_links_len | 56 | 5 | 2 | 2 | 2 | -32 | -30 | 36 | 30 | -39 |
| parse_depth | 34 | -36 | 3 | 3 | 4 | -17 | -1 | 22 | 24 | 12 |
| sub_post | 28 | -33 | 8 | 8 | 9 | -28 | -40 | 33 | 34 | 48 |
| dep_dist_conj | 17 | 31 | 11 | 13 | 10 | 37 | -37 | 46 | 56 | -48 |
| upos_dist_NUM | 15 | 39 | 70 | 72 | 72 | 4 | / | / | / | / |
| ttr_form | -42 | 28 | 77 | 74 | -26 | 17 | 2 | 3 | 2 | 1 |
| prep_chain_len | 53 | 12 | 16 | 16 | 14 | -48 | -23 | 43 | 39 | 42 |
| sub_chain_len | 24 | -14 | 19 | 19 | 32 | -30 | -43 | 56 | 55 | 35 |
| dep_dist_nsubj | 11 | -16 | -8 | -8 | -9 | -2 | 31 | -18 | -19 | -29 |
| upos_dist_PRON | -16 | -13 | -7 | -6 | -8 | -44 | -21 | -5 | -8 | -38 |
| dep_dist_punct | -21 | -3 | -4 | -4 | -4 | -20 | -3 | -2 | -2 | -2 |
| dep_dist_nmod | -20 | -2 | 55 | 50 | 50 | -9 | 3 | 28 | 17 | 15 |
| xpos_dist_. | -11 | 15 | -1 | -1 | -1 | -6 | 43 | -24 | -30 | 32 |
| xpos_dist_VBZ | -9 | 20 | 82 | -33 | -30 | 24 | 14 | 20 | 40 | -47 |
| dep_dist_aux | -8 | 17 | -30 | -29 | 77 | 32 | 27 | 39 | 31 | 45 |
| dep_dist_case | -7 | -34 | 25 | 22 | 34 | 8 | -6 | 62 | 44 | -21 |
| ttr_lemma | -4 | 21 | -22 | -28 | -11 | -4 | -45 | 4 | 4 | 9 |
| dep_dist_det | -3 | 52 | 42 | 40 | 21 | -27 | -36 | 17 | 14 | 5 |
| sub_prop_dist | -2 | 29 | 6 | 5 | 5 | 26 | 28 | 63 | 59 | 21 |
| lexical_density | -1 | -1 | 26 | 25 | 20 | -37 | -5 | 5 | 6 | 10 |

Table 5: Rankings based on the coefficients assigned by SVM feats for all metrics. Top ten positive and negative features are marked with orange and cyan respectively. "/" marks features present in less than 5% of sentences.

process, especially in relation to analyzed features; and ii) verifying whether, and in which respect, this competence is affected by a fine-tuning on complexity assessment tasks.

To conduct the probing experiments, we aggregate three UD English treebanks representative of different genres, namely: EWT, GUM and ParTUT by Silveira et al. (2014); Zeldes (2017); Sanguinetti and Bosco (2015), respectively. We thus obtain a corpus of 18,079 sentences and use the Profiling-UD tool to extract $n$ sentence-level linguistic features $\mathcal{Z} = z_1, \ldots, z_n$ from gold linguistic annotations. We then generate representations $A(x)$ of all sentences in the corpus using the last-layer [CLS] embedding of a pretrained ALBERT base model without additional fine-tuning, and train $n$ single-layer perceptron regressors $g_i : A(x) \rightarrow z_i$ that learn to map representations $A(x)$ to each linguistic feature $z_i$. We finally evaluate the error and $R^2$ scores of each $g_i$ as a proxy to the quality of representations $A(x)$ for encoding their respective linguistic feature $z_i$. We repeat the same evaluation for ALBERTs fine-tuned respectively on perceived complexity (PC) and on all eye-tracking labels with multitask learning (ET), averaging scores with 5-fold cross-validation. Results are shown on the left side of Table 6.

As we can see, ALBERT's last-layer sentence representations have relatively low knowledge of complexity-related probes, but the performance on them highly increases after fine-tuning. Specifically, a noticeable improvement is obtained on features that were already better encoded in base pretrained representation, i.e. sentence length and related features, suggesting that fine-tuning possibly accentuates only properties already well-known by the model, regardless of the target task. To verify that this isn't the case, we repeat the same experiments on ALBERT models fine-tuned on the smallest length-binned subset (i.e. 10±1 tokens) presented in previous sections. The right side of Table 6 presents these results. We know from our length-binned analysis of Figure 2 that PC scores are mostly uncorrelated with length phenomena,

| | Base | | PC | | ET | | PC Bin 10±1 | | ET Bin 10±1 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | RMSE | $R^2$ | RMSE | $R^2$ | RMSE | $R^2$ | RMSE | $R^2$ | RMSE | $R^2$ |
| n_tokens | 8.19 | .26 | 4.66 | **.76** | 2.87 | **.91** | 8.66 | .18 | 6.71 | **.51** |
| parse_depth | 1.47 | .18 | 1.18 | **.48** | 1.04 | **.60** | 1.50 | .16 | 1.22 | **.43** |
| vb_head_per_sent | 1.38 | .15 | 1.26 | **.30** | 1.14 | **.42** | 1.44 | .09 | 1.30 | **.25** |
| xpos_dist_. | .05 | .13 | .04 | **.41** | .04 | **.42** | .04 | .18 | .04 | **.38** |
| avg_links_len | .58 | .12 | .53 | **.29** | .52 | **.31** | .59 | .10 | .56 | **.20** |
| max_links_len | 5.20 | .12 | 4.08 | **.46** | 3.75 | **.54** | 5.24 | .11 | 4.73 | **.28** |
| n_prep_chains | .74 | .11 | .67 | **.26** | .66 | **.29** | .72 | .14 | .69 | **.21** |
| sub_prop_dist | .35 | .09 | .33 | .13 | .31 | **.22** | .34 | .05 | .32 | .15 |
| upos_dist_PRON | .08 | .09 | .08 | .14 | .08 | .07 | .07 | **.23** | .08 | .15 |
| upos_dist_NUM | .05 | .08 | .05 | .06 | .05 | .02 | .05 | **.16** | .05 | .06 |
| dep_dist_nsubj | .06 | .08 | .06 | .10 | .06 | .05 | .05 | **.17** | .06 | .11 |
| char_per_tok | .89 | .07 | .87 | **.12** | .90 | .05 | .82 | **.22** | .86 | .14 |
| prep_chain_len | .60 | .07 | .57 | **.17** | .56 | **.19** | .59 | .12 | .56 | **.18** |
| sub_chain_len | .70 | .07 | .67 | **.15** | .62 | **.26** | .71 | .04 | .66 | **.16** |
| dep_dist_punct | .07 | .06 | .07 | .06 | .07 | **.14** | .07 | .06 | .07 | **.14** |
| dep_dist_nmod | .05 | .06 | .05 | .07 | .05 | .06 | .05 | .09 | .05 | .09 |
| sub_post | .44 | .05 | .46 | **.12** | .44 | **.18** | .47 | .05 | .45 | **.14** |
| dep_dist_case | .07 | .05 | .06 | .06 | .07 | .08 | .07 | .07 | .07 | .10 |
| lexical_density | .14 | .05 | .13 | .03 | .13 | .03 | .13 | **.13** | .13 | **.13** |
| dep_dist_compound | .06 | .04 | .06 | .05 | .06 | .03 | .06 | .10 | .06 | .07 |
| dep_dist_conj | .04 | .03 | .04 | .04 | .04 | .04 | .05 | .02 | .04 | .03 |
| ttr_form | .08 | .03 | .08 | .05 | .08 | .05 | .08 | .05 | .08 | .05 |
| dep_dist_det | .06 | .03 | .06 | .02 | .06 | .04 | .06 | .03 | .06 | .03 |
| dep_dist_aux | .04 | .02 | .04 | .01 | .04 | .01 | .04 | .06 | .04 | .04 |
| xpos_dist_VBN | .03 | .01 | .03 | .00 | .03 | .00 | .03 | .01 | .03 | .00 |
| xpos_dist_VBZ | .04 | .01 | .04 | .01 | .04 | .02 | .04 | .02 | .04 | .02 |
| ttr_lemma | .09 | .01 | .09 | .06 | .09 | .06 | .09 | .04 | .09 | .03 |

Table 6: RMSE and $R^2$ scores for diagnostic regressors trained on ALBERT representations, respectively, without fine-tuning (Base), with PC and eye-tracking (ET) fine-tuning on all data (left) and on the $10 \pm 1$ length-binned subset (right). **Bold** values highlight relevant increases in $R^2$ from Base.

while ET scores remain significantly affected despite our controlling of sequence size. This also holds for length-binned probing task results, where the PC model seems to neglect length-related properties in favor of other ones, which were the same highlighted in our fine-grained correlation analysis (e.g. word length, numbers, explicit subjects). The ET-trained model confirms the same behavior, retaining strong but lower performances for length-related features. We note that, for all metrics, features that were highly relevant only for the SVM predictions, such as those encoding verbal inflectional morphology or vocabulary-related ones (Table 5), are not affected by the fine-tuning process. Despite obtaining the same accuracy of a SVM, the neural language model seem to address the task more similarly to humans when accounting for correlation scores (Figure 2). A more extensive analysis of the relation between human behavior and predictions by different models is deemed interesting for future work.

To conclude, although higher probing tasks performances after fine-tuning on complexity metrics should not be interpreted as direct proof that the neural language model is exploiting newly-acquired morpho-syntactic and syntactic information, they suggest an importance shift in NLM representation, triggered by fine-tuning, that produces an encoding of linguistic properties able to better model the human assessment of complexity.

## 6 Conclusion

This paper investigated the connection between eye-tracking metrics and the explicit perception of sentence complexity from an experimental standpoint. We performed an in-depth correlation analysis between complexity scores and sentence-level properties at different granularity levels, highlighting how all metrics are strongly connected to sentence length and related properties, but also revealing different behaviors when controlling for length. We then evaluated models using explicit

linguistic features and unsupervised word embeddings to predict complexity, showing comparable performances across metrics. We finally tested the encoding of linguistic properties in the contextual representations of a neural language model, noting the natural emergence of task-related linguistic properties within the model's representations after the training process. We thus conjecture that a relation subsists between the linguistic knowledge acquired by the model during the training procedure and its downstream performances on tasks for which the morphosyntactic and syntactic structures play a relevant role. For the future, we would like to test comprehensively the effectiveness of tasks inspired by the human language learning as intermediate steps to train more robust and parsimonious neural language models.

## 7 Broader Impact and Ethical Perspectives

The findings described in this work are mostly intended to evaluate recent efforts in the computational modeling of linguistic complexity. This said, some of the models and procedures described can be clearly beneficial to society. For example, using models trained to predict reading patterns may be used in educational settings to identify difficult passages that can be simplified, improving reading comprehension for students in a fully-personalizable way. However, it is essential to recognize the potentially malicious usage of such systems. The integration of eye-tracking systems in mobile devices, paired with predictive models presented in this work, could be used to build harmful surveillance systems and advertisement platforms using gaze predictions for extreme behavioral manipulation. In terms of research impact, the experiments presented in this work may provide useful insights into the behavior of neural language models for researchers working in the fields of interpretability in NLP and computational psycholinguistics.

## References


Samira Abnar, Lisa Beinborn, Rochelle Choenni, and Willem Zuidema. 2019. Blackbox meets blackbox: Representational similarity & stability analysis of neural language models and brains. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 191–203, Florence, Italy. Association for Computational Linguistics.

Bharat Ram Ambati, Siva Reddy, and Mark Steedman. 2016. Assessing relative sentence complexity using an incremental CCG parser. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1051–1057, San Diego, California. Association for Computational Linguistics.

Dominique Brunato, Andrea Cimino, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2020. Profiling-UD: a tool for linguistic profiling of texts. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 7145–7151, Marseille, France. European Language Resources Association.

Dominique Brunato, Lorenzo De Mattei, Felice Dell'Orletta, Benedetta Iavarone, and Giulia Venturi. 2018. Is this sentence difficult? do you agree? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2690–2699, Brussels, Belgium. Association for Computational Linguistics.

Rich Caruana. 1997. Multitask learning. *Machine Learning*, 28:41–75.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

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

Deepset. 2019. FARM: Framework for adapting representation models. GitHub repository: `https://github.com/deepset-ai/FARM`.

Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193 – 210.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Sidney Evaldo Leal, João Marcos Munguba Vieira, Erica dos Santos Rodrigues, Elisângela Nogueira Teixeira, and Sandra Aluísio. 2020. Using eye-tracking

data to predict the readability of Brazilian Portuguese sentences in single-task, multi-task and sequential transfer learning approaches. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5821–5831, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jon Gauthier and Roger Levy. 2019. Linking artificial and human neural representations of language. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 529–539, Hong Kong, China. Association for Computational Linguistics.

Ana Valeria González-Garduño and Anders Søgaard. 2018. Learning to predict readability using eye-movement data from natives and learners. In *AAAI Conference on Artificial Intelligence 2018*. AAAI Conference on Artificial Intelligence.

Michael Hahn and Frank Keller. 2016. Modeling human reading with neural attention. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 85–95, Austin, Texas. Association for Computational Linguistics.

Nora Hollenstein, Maria Barrett, and Lisa Beinborn. 2020. Towards best practices for leveraging human language processing signals for natural language processing. In *Proceedings of the Second Workshop on Linguistic and Neurocognitive Resources*, pages 15–27, Marseille, France. European Language Resources Association.

Nora Hollenstein, Maria Barrett, Marius Troendle, Francesco Bigiolli, Nicolas Langer, and Ce Zhang. 2019a. Advancing nlp with cognitive language processing signals. *arXiv preprint arXiv:1904.02682*.

Nora Hollenstein, Antonio de la Torre, Nicolas Langer, and Ce Zhang. 2019b. CogniVal: A framework for cognitive word embedding evaluation. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 538–549, Hong Kong, China. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A Lite BERT for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. 2021. Supervised and Unsupervised Neural Approaches to Text Readability. *Computational Linguistics*, 47(1):141–179.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal Dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria. Association for Computational Linguistics.

Amil Merchant, Elahe Rahimtoroghi, Ellie Pavlick, and Ian Tenney. 2020. What happens to BERT embeddings during fine-tuning? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 33–44, Online. Association for Computational Linguistics.

Alessio Miaschi, Dominique Brunato, Felice Dell'Orletta, and Giulia Venturi. 2020. Linguistic profiling of a neural language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 745–756, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

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

Luz Rello, Susana Bautista, Ricardo Baeza-Yates, Pablo Gervás, Raquel Hervás, and Horacio Saggion. 2013. One half or 50%? an eye-tracking study of number representation readability. In *Human-Computer Interaction – INTERACT 2013*, pages 229–245, Berlin, Heidelberg. Springer Berlin Heidelberg.

Manuela Sanguinetti and Cristina Bosco. 2015. *PartTUT: The Turin University Parallel Treebank*, pages 51–69. Springer International Publishing, Cham.

Gabriele Sarti. 2020. UmBERTo-MTSA @ AcCompl-It: Improving complexity and acceptability prediction with multi-task learning on self-supervised annotations. In *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, Online. CEUR.org.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Chris Manning. 2014. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2897–2904, Reykjavik, Iceland. European Language Resources Association (ELRA).

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

Sowmya Vajjala and Ivana Lučić. 2018. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 297–304, New Orleans, Louisiana. Association for Computational Linguistics.

Sowmya Vajjala and Ivana Lučić. 2019. On understanding the relation between expert annotations of text readability and target reader comprehension. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 349–359, Florence, Italy. Association for Computational Linguistics.

Shravan Vasishth, Titus von der Malsburg, and Felix Engelmann. 2013. What eye movements can tell us about sentence comprehension. *Wiley interdisciplinary reviews. Cognitive science*, 4 2:125–134.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Amir Zeldes. 2017. The GUM corpus: creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51:581–612.

Kelly Zhang and Samuel Bowman. 2018. Language modeling teaches you more than translation does: Lessons learned through auxiliary syntactic task analysis. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 359–361, Brussels, Belgium. Association for Computational Linguistics.

| | Model & Tokenizer Parameters | | |
|---|---|---|---|
| heads dimension | 1-Layer Dense | | |
| max seq. length | 128 | | |
| embed. dropout | 0.1 | | |
| seed | 42 | | |
| lowercasing | ✗ | | |
| tokenization | SentencePiece | | |
| vocab. size | 30000 | | |
| | **Training Parameters** | | |
| | **PC** | **ET** | **Probes** |
| fine-tuning | standard | multitask | multitask |
| freeze LM $w$ | ✗ | ✗ | ✓ |
| weighted loss | - | ✓ | ✗ |
| CV folds | 5 | 5 | 5 |
| early stopping | ✓ | ✓ | ✗ |
| training epochs | 15 | 15 | 5 |
| patience | 5 | 5 | - |
| evaluation steps | 20 | 40 | - |
| batch size | 32 | 32 | 32 |
| learning rate | 1e-5 | 1e-5 | 1e-5 |

Table 7: Model, tokenizer and training parameters used for fine-tuning ALBERT on complexity metrics.

## A Parametrization and Fine-tuning Details for ALBERT

We leverage the pretrained albert-base-v2 checkpoint available in the HuggingFace's Transformer framework (Wolf et al., 2020) and use adapted scripts and classes from the FARM framework (Deepset, 2019) to perform multitask learning on eye-tracking metrics. Table 7 presents the parameters used to define models and training procedures for experiments in Sections 4 and 5.

During training we compute MSE loss scores for task-specific heads for the four eye-tracking metrics ($\ell_{FXC}$, $\ell_{FPD}$, $\ell_{TFD}$, $\ell_{TRD}$) and perform a weighted sum to obtain the overall loss score $\ell_{ET}$ to be optimized by the model:

$$\ell_{ET} = \ell_{FXC} + \ell_{FPD} + \ell_{TFD} + (\ell_{TRD} \times 0.2)$$

The use of $\ell_{TRD}$ was shown to have a positive impact on the overall predictive capabilities of the model only when weighted to prevent it from dominating the $\ell_{ET}$ sum.

Probing tasks on linguistic features are performed by freezing the language model weights and training 1-layer heads as probing regressors over the last-layer [CLS] token for each feature. In this setting no loss weighting is applied, and the regressors are trained for 5 epochs without early stopping on the aggregated UD dataset.

## B Examples of Sentences from Complexity Corpora

Table 8 presents examples of sentences randomly selected from the two corpora leveraged in this study. We highlight how eye-tracking scores show a very consistent relation with sentence length, while PC scores are much more variable. This fact suggests that the offline nature of PC judgments makes them less related to surface properties and more connected to syntax and semantics.

## C Models' Performances on Length-binned Sentences

Similarly to the approach adopted in Section 3, we test the performances of models on length-binned data to verify if performances on length-controlled sequences are consistent with those achieved on the whole corpora. RMSE scores averaged with 5-fold cross validation over the length-binned sentences subsets are presented in Figure 3. We note that ALBERT outperforms the SVM with linguistic features on nearly all lengths and metrics, showing the largest gains on intermediate bins for PC and gaze durations (FPD, TFD, TRD). Interestingly, overall performances of models follow a length-dependent increasing trend for eye-tracking metrics, but not for PC. We believe this behavior can be explained in terms of the high sensibility to length previously highlighted for online metrics, as well as the variability in bin dimensions (especially for the last bin containing only 63 sentences). We finally observe that the SVM model based on explicit linguistic features (SVM feats) performs poorly on larger bins for all tasks, sometimes being even worse than the bin-average baseline. While we found this behavior surprising given the positive influence of features highlighted in Table 4, we believe this is mostly due to the small dimension of longer bins, which negatively impacts the generalization capabilities of the regressor. The relatively better scores achieved by ALBERT in those, instead, support the effectiveness of information stored in pretrained language representations when a limited number of examples is available.

| Length bin | Sentence | PC Score |
|---|---|---|
| Bin 10±1 | It hasn't made merger overtures to the board. | 2.15 |
| Bin 15±1 | For most of the past 30 years, the marriage was one of convenience. | 1.45 |
| Bin 20±1 | Shanghai Investment & Trust Co., known as Sitco, is the city's main financier for trading business. | 3.35 |
| Bin 25±1 | For fiscal 1988, Ashland had net of $224 million, or $4.01 a share, on revenue of $7.8 billion. | 4.55 |
| Bin 30±1 | C. Olivetti & Co., claiming it has won the race in Europe to introduce computers based on a powerful new microprocessor chip, unveiled its CP486 computer yesterday. | 4.25 |
| Bin 35±1 | The White House said he plans to hold a series of private White House meetings, mostly with Senate Democrats, to try to persuade lawmakers to fall in line behind the tax cut. | 2.9 |

| Length bin | Sentence | FPD | FXC | TFD | TRD |
|---|---|---|---|---|---|
| Bin 10±1 | Evidently there was a likelihood of John Cavendish being acquitted. | 1429 | 7.69 | 1527 | 330 |
| Bin 15±1 | I come now to the events of the 16th and 17th of that month. | 1704 | 9.71 | 1979 | 467 |
| Bin 20±1 | Who on earth but Poirot would have thought of a trial for murder as a restorer of conjugal happiness! | 2745 | 15.38 | 3178 | 1003 |
| Bin 25±1 | He knew only too well how useless her gallant defiance was, since it was not the object of the defence to deny this point. | 3489 | 19.77 | 4181 | 1012 |
| Bin 30±1 | I could have told him from the beginning that this obsession of his over the coffee was bound to end in a blind alley, but I restrained my tongue. | 3638 | 21.36 | 4190 | 1010 |
| Bin 35±1 | There was a breathless hush, and every eye was fixed on the famous London specialist, who was known to be one of the greatest authorities of the day on the subject of toxicology. | 4126 | 23.14 | 4814 | 1631 |

Table 8: Example of sentences selected from all the length-binned subset for the Perceived Complexity Corpus (top) and the GECO corpus (bottom). Scores are aggregated following the procedure described in Section 2. Reading times (FPD, TFD, TRD) are expressed in milliseconds.

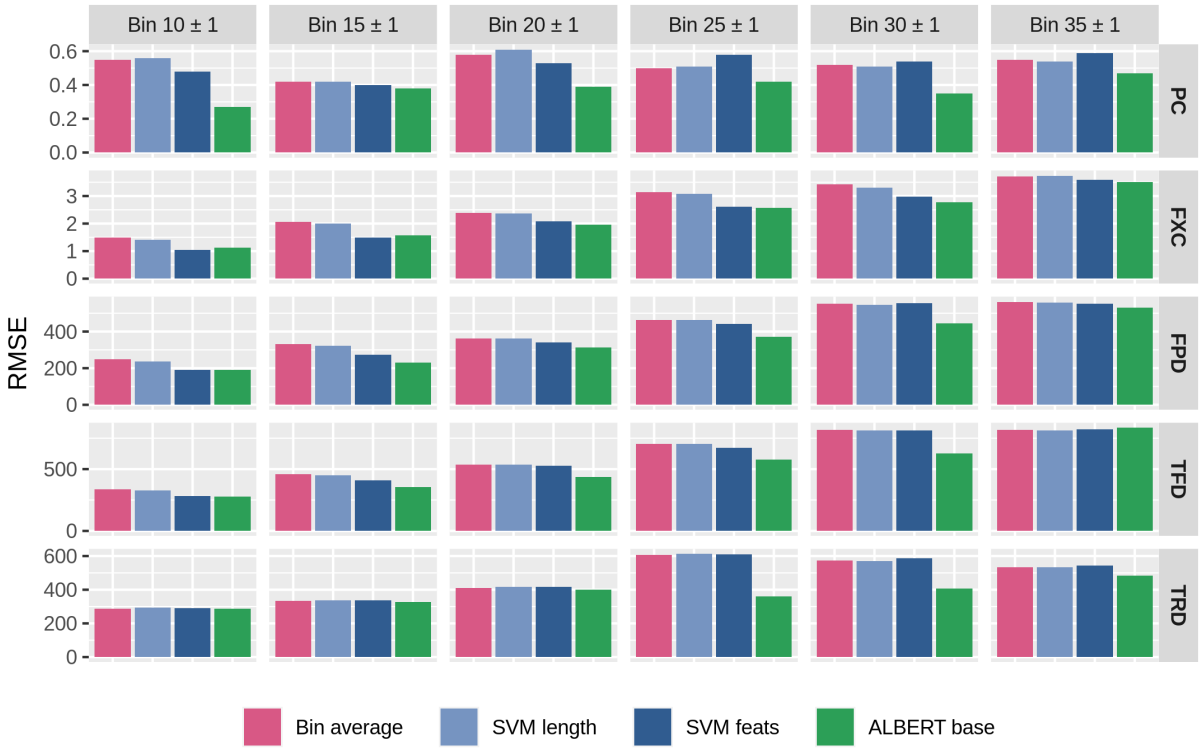


Figure 3: Average Root Mean Square Error (RMSE) scores for models in Table 4, performing 5-fold cross-validation on the same length-binned subsets used for the analysis of Figure 2. Lower scores are better.