Sentence Complexity in Context

Benedetta Iavarone[*◇], **Dominique Brunato**[◇], **Felice Dell'Orletta**[◇]
[*]Scuola Normale Superiore, [◇]Istituto di Linguistica Computazionale "Antonio Zampolli", Pisa
ItaliaNLP Lab – *www.italianlp.it*
benedetta.iavarone@sns.it
{dominique.brunato, felice.dellorletta}@ilc.cnr.it

Abstract

We study the influence of context on how humans evaluate the complexity of a sentence in English. We collect a new dataset of sentences, where each sentence is rated for perceived complexity within different contextual windows. We carry out an in-depth analysis to detect which linguistic features correlate more with complexity judgments and with the degree of agreement among annotators. We train several regression models, using either explicit linguistic features or contextualized word embeddings, to predict the mean complexity values assigned to sentences in the different contextual windows, as well as their standard deviation. Results show that models leveraging explicit features capturing morphosyntactic and syntactic phenomena perform always better, especially when they have access to features extracted from all contextual sentences.

1 Introduction

From a human-based perspective, sentence complexity is assessed by measures of processing effort or performance in behavioral tasks. In this respect, a large part of studies has focused on reading single sentences and correlating syntactic and lexical properties with observed difficulty, being it captured by cognitive signals, such as eye-tracking metrics (Rayner, 1998; King and Just, 1991), or by explicit judgments of complexity given by readers (Brunato et al., 2018). However, models of language comprehension underline the importance of contextual cues, such as the presence of explicit cohesive devices, in building a coherent representation of a text (Kintsch et al., 1975; McNamara, 2001). This implies that a sentence can be perceived as more or less difficult according to the context in which it is presented.

The effect of context on how humans evaluate a sentence has been investigated concerning its acceptability and grammaticality, two properties different from complexity, yet somehow related. In Bernardy et al. (2018) speakers were asked to evaluate the degree of acceptability of sentences from Wikipedia, both in their original form and with some grammatical alterations artificially introduced by a process of round-trip machine translation. Results showed that ill-formed sentences are evaluated as more acceptable when presented within context (i.e. along with their preceding or following sentence) rather than in isolation. More closely related to our study is the one by Schumacher et al. (2016) on readability assessment. In that work, authors gathered pairwise evaluations of reading difficulty on sentences presented with and without a larger context, training a logistic regression model to predict binary complexity labels assigned by humans. They observed that the context slightly modifies the perception of the readability of a sentence, although their predictive models perform better on sentences rated in isolation.

Our study aims to understand how the context surrounding a sentence influences its 'perceived' complexity by humans. As we consider linguistic complexity from the individual's perspective, following Brunato et al. 2018, we use the term complexity as a synonym of difficulty. Also, we assume that sentence complexity is a gradient rather than a binary concept and we operationalize perceived complexity as a score on an ordinal scale. These scores were collected for a new dataset of sentences, where each sentence has been evaluated in three contextual windows, which change according to the position the sentence occupies within them. This enables us to deeply inspect the role of context, allowing us to determine if the perceived complexity of a sentence changes when the context is introduced, and also which contextual window may impact more. To do so, we consider the average complexity score assigned to each sentence as well as the degree of agreement among annotators, calculated in terms of standard deviation. We think

Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, pages 186–199
Online Event, June 10, 2021. ©2021 Association for Computational Linguistics

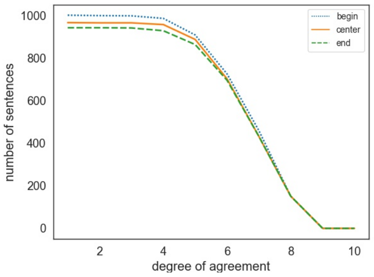


Figure 1: Number of sentences for each degree of agreement.

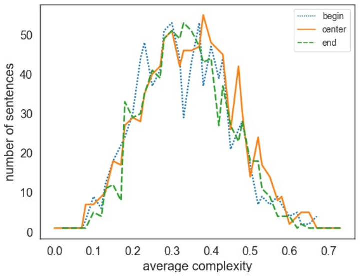


Figure 2: Number of sentences at different average complexity ratings.

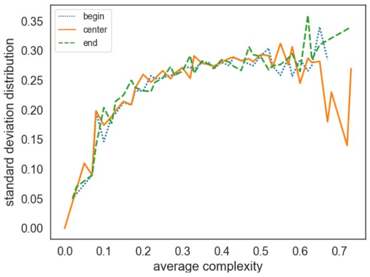


Figure 3: Mean standard deviation at different average complexity ratings.

that this measure is also relevant to comprehend perceived sentence complexity since this is a highly subjective task that cannot be tackled by following specific annotation guidelines. Moreover, knowing that sentence length is a prominent predictor of sentence complexity, we study how complexity scores and annotators' agreement vary for sentences of the same length. Finally, we run experiments to evaluate the accuracy of different regression models in predicting the mean complexity label and standard deviation assigned to a sentence. In particular, we compare models leveraging explicit linguistic features related to a wide set of morphosyntactic and syntactic properties of a sentence, to models exploiting the predictions of a state-of-the-art bidirectional transformer encoder, i.e. BERT (Devlin et al., 2019). To the best of our knowledge, the assessment of sentence complexity in context has never been tackled as a downstream task for evaluating the effectiveness of neural representations in modeling aspects of sentence complexity. Despite the remarkable performance that neural language models have achieved so far in a variety of NLP tasks (Rogers et al., 2020) – also close to ours such as the prediction of perceived sentence acceptability in context (Lau et al., 2020) – our results show that this is not the case as regards to the prediction of sentence complexity: models using explicit linguistic features perform better in all contextual windows, suggesting that information embedded in neural representations could be less effective in modeling the examined task, particularly when few labeled data will be made available.

Contributions. *i)* We release a new dataset of ∼2,900 English sentences rated with human judgments of complexity elicited by presenting sentences in their contexts; *ii)* we model a wide set of morphosyntactic and syntactic phenomena, extracted from the single sentence and contextual ones, and we study which of them are more correlated with sentence complexity judgments in different contextual windows; *iii)* we show that models relying explicitly on these features achieve higher performance in predicting complexity judgments than state–of–the art neural language models, proving the effectiveness of these features to address this task in particularly in a low resource scenario.

All the data discussed here will be made available at: `www.italianlp.it/resources/`.

2 Approach

We first collected an appropriate corpus to evaluate the effect of context on the perception of sentence complexity. We started from the crowdsourced dataset by Brunato et al. (2018), which contains $1,200$ sentences annotated for perceived complexity on a 7-point Likert scale. We similarly built a crowdsourcing task, asking native English speakers to read each sentence and rate its complexity on the same scale. However, while in Brunato et al. all sentences were rated in isolation, in our task sentences were presented in three contextual windows, as illustrated in Section 2.1.

To study which linguistic phenomena may affect annotators' ratings, we represented sentences (the rated one and the contextual ones) with ∼100 linguistic features, based on those described in Brunato et al. (2020). These features model a wide range of sentence properties, which can be viewed as proxies of sentence complexity at different levels of linguistic annotation. The features were first used to study the influence of context-level and sentence-level phenomena on perceived complexity. We did this by analyzing the correlations between the features and the complexity ratings, and the correlations between the features and the standard deviation of complexity. Then, we assessed the automatic prediction of sentence complexity and of the standard deviation of complexity judgments, evaluating if adding information from the context helps in the prediction. We tested two predicting approaches: one based on a linear SVM regression model which leverages the linguistic features discussed so far, and one that employs BERT, one of the most prominent pre-trained neural language model. We compared the accuracy of the models across various scenarios, considering their predictions both for sentences rated in different contextual windows and for sentences distinguished into same-length bins.

2.1 Data Collection

As already mentioned, our dataset was built starting from the sentences collected by Brunato et al. (2018). These sentences were extracted from the Wall Street Journal section of the Penn Treebank and grouped in 6 bins, according to their length in terms of tokens (i.e. 10, 15, 20, 25, 30, 35), as it is well-known that sentence length correlates with complexity. By analyzing sentences with the same length we would understand whether other linguistic features still play an influence on complexity or if their effect is nullified by controlling length. We then proceeded to add context to all sentences, defining context as the sentences that precede and/or follow a given one. For each sentence we created 3 different contextual windows, according to the position occupied by the sentence in relation to the one occupied by the context. In the *begin window*, the sentence appears first and is followed by two contextual sentences; in the *center window*, the sentence is in the middle and is preceded by a contextual sentence and followed by another contextual sentence; in the *end window*, the sentence appears as the last one and is preceded by two contextual sentences. The resulting dataset is composed of 2,913 windows of context: 1,002 for the begin window, 968 for the center window and 943 for the end window.

We carried out a crowdsourcing task to collect complexity ratings through the platform Prolific[1]. For each contextual window, the sentence to be evaluated was highlighted in bold, while the contextual sentences were left in plain style. The windows were randomly ordered and presented on different pages, containing ten windows each. Due to the high number of windows to be evaluated, we split the dataset into smaller sections, containing at most 200 windows each, ending up creating 15 evaluation tasks. For each task, we recruited 10 native English speakers. We then asked participants to read the full paragraph (the whole window of context) and to rate the complexity of the sentence in bold on a 7-point Likert scale, where 1 stands for “very easy" and 7 stands for “very difficult". As complexity perception is very subjective, we then aggregated the ratings to account for the individual bias of annotators, as there could be the case in which a participant always gave low scores, while another one always gave very high scores. Thus, ratings were re-scaled between 0 and 1 and normalized by the range of ratings given by each annotator.

	begin		**center**		**end**	
	judg	*std*	*judg*	*std*	*judg*	*std*
Length 10	.28	.23	.28	.28	.28	.28
Length 15	.27	.23	.32	.28	.30	.28
Length 20	.27	.22	.35	.27	.33	.26
Length 25	.26	.21	.36	.26	.35	.26
Length 30	.26	.22	.38	.26	.36	.25
Length 35	.25	.21	.39	.26	.38	.26
All sents	.26	.22	.35	.27	.33	.27

Table 1: Mean complexity judgment and mean standard deviation on complexity, for all sentences and at different lengths.

2.2 Data Analysis

Firstly, we looked at the *degree of agreement*[2] (DAE) between annotators. Figure 1 reports the number of sentences for every DAE, considering the different sentence positions within the context windows. We found a strong DAE, as most sentences have up to 5 annotators that assigned a complexity judgment within the same range. As the DAE increases, the number of sentences decreases consistently. The highest DAE is found at 8 annotators, but on a small amount of sentences (< 200), while there are no sentences on which 9 or 10 annotators agree. Also, this first examination showed that the sentence position has little to no influence on the DAE, as the numbers for the context windows mostly follow the same trend. To confirm this view, we looked at the distribution of complexity values among the three windows. For each window, we computed the number of sentences that were assigned the same average complexity value. Figure 2 shows that average complexity follows a Gaussian distribution for all the windows of context, as most sentences received an average complexity between 0.2 and 0.4.

[1] www.prolific.co

[2] The degree of agreement is intended as the number of annotators who gave a complexity score within the same range. The range is defined as the standard deviation from the mean of the judgments given to each sentence.

	Zero Variance	**BCE**	**Highest Variance**	**B**	**C**	**E**
Length 10	Tokyo's Nikkei index fell 84.15 points to 35442.40.	.38	Nashua announced the Reiss request after the market closed.	.22	.42	.63
Length 15	Elsewhere in Europe, share prices closed higher in Stockholm, Brussels and Milan.	.23	Last year, the prisons' sales to the Pentagon totaled $336 million.	.62	.32	.20
Length 20	Dow Jones industrials 2645.08, up 41.60; transportation 1205.01, up 13.15; utilities 219.19, up 2.45.	.50	The cash dividend paid on the common stock also will apply to the new shares, the company said.	.12	.12	.55
Length 25	In the nine months, Milton Roy earned $6.6 million, or $1.18 a share, on sales of $94.3 million.	.38	Yesterday, Compaq plunged further, closing at $100 a share, off $8.625 a share, on volume of 2,633,700 shares.	.25	.67	.42
Length 30	SsangYong, which has only about 3% of the domestic market, will sell about 18,000 of its models this year, twice as many as last year.	.32	Though not reflected in the table, an investor should know that the cost of the option insurance can be partially offset by any dividends that the stock pays.	.23	.50	.57
Length 35	In the nine months, net rose 35% to $120.1 million, or $1.64 a share, from $89.20 million, or $1.22 a share, a year earlier.	.48	William Kaiser, president of the Kaiser Financial Group in Chicago, said the decline was almost certainly influenced by the early sell-off in the stock market, which partly reflected a weakening economy.	.45	.23	.58
All sents	Dow Jones industrials 2645.08, up 41.60; transportation 1205.01, up 13.15; utilities 219.19, up 2.45.	.50	The cash dividend paid on the common stock also will apply to the new shares, the company said.	.12	.12	.55

Table 2: Sentences that vary the least or the most within context windows. B, C, and E respectively indicate the begin, center and end windows.

Furthermore, we computed the standard deviation of the complexity judgments that were assigned to each sentence. In Figure 3, we plot the standard deviation of each sentence[3] against the average complexity assigned to that same sentence, for the three windows of context. The standard deviation tends to increase with the average complexity score assigned to sentences. This means that annotators agree more on rating a sentence as simple, suggesting that the perception of a sentence as more complex may be less homogeneous. This trend is quite similar for all contextual windows, though we observe a more uniform behaviour in rating a sentence as more complex when it is surrounded by both contextual sentences (i.e. the center window).

Besides sentence positioning, also sentence length may affect the perception of complexity. Thus, we calculated the average of complexity judgments assigned to sentences of the same length, for all the three context windows, along with the mean standard deviation. As shown in Table 1, for the center and the end window average complexity values tend to increase with the length of the sentences, as expected. On the contrary, standard deviation follows the opposite trend, showing that subjects agree more on the complexity of long sentences (e.g. length 30 and 35), while their perception about shorter sentences is more diversified. It also emerges that when the sentence is at the beginning of the paragraph, it is overall perceived as simpler. This may indicate that the following contextual sentences help annotators in the processing and understanding of the first sentence.

Table 2 shows examples of sentences whose complexity scores vary the least or the most within the different windows of context. In the case of *Zero Variance*, the sentence received the same average complexity, regardless of the relative position in the contextual window (begin, center, end). Instead, sentences with the highest variance received very different average values, according to the position the sentence occupies in the contextual windows. This table also reports the actual average complexity values that the sentences got for each position.

Linguistic Features
Raw Text Properties
Sentence Length
Word Length
Vocabulary Richness
Type/Token Ratio for words and lemmas
Morphosyntactic information
Distribution of UD and language–specific POS
Lexical dens
Inflectional morphology
Inflectional morphology of lexical verbs and auxiliaries
Verbal Predicate Structure
Distribution of verbal heads and verbal roots
Verb arity and distribution of verbs by arity
Global and Local Parsed Tree Structures
Depth of the whole syntactic tree
Average length of dependency links and of the longest link
Average length of prepositional chain and distribution by depth
Clause length
Relative order of elements
Order of subject and object
Syntactic Relations
Distribution of dependency relations
Use of Subordination
Distribution of subordinate and principal clauses
Average length of subordination chain and distribution by depth
Relative order of subordinate clauses

Table 3: Linguistic features.

[3] If more than one sentence was assigned the same average complexity value, we plot the average standard deviation of all the sentences.

3 Correlation between Linguistic Features and Complexity

To detect which linguistic phenomena are more involved in the assessment of sentence complexity, and to verify whether these phenomena capture information about the sentence itself or about the context, we performed a correlation analysis between the complexity score assigned to each sentence and a wide set of linguistic features extracted from the sentence. For each sentence, we computed the Spearman's rank correlation coefficient between the average complexity score and the value of each linguistic feature extracted from *i*) the rated sentence, *ii*) its preceding one and *iii*) its following one, according to the contextual window. We performed the correlation analysis on the sentences altogether and then dividing them into bins according to their length. The same process was repeated correlating the standard deviation of complexity scores with the linguistic features of each sentence. As stated in Section 2, we focused on features that model a wide range of sentence properties extracted from different levels of linguistic annotation, from raw text features (i.e. sentence and word length) to morphosyntactic information (e.g. distribution of verbs according to morphological features such as tense, mood, person), to more complex aspects of the syntactic structure capturing global and local information (e.g. parse tree depth, length of dependency link, use of subordination). Table 3 reports the list of features used for our analysis.

In what follows, we discuss the correlation results for the subset of sentences presented in the *center window*, since this is the only one in which the rated sentence was always surrounded by both a left and a right sentence, allowing us to compare the effect of the two context positions[4].

Considering first the average complexity score, we found statistically significant correlations (p-value< 0.05) with $\rho \geq \pm 0.20$ for 103 features out of the whole set. Among them, 44% belongs to the rated sentence (i.e. 45 features) and 56% to the contextual ones (i.e. 23 and 35 features to the left and the right sentence, respectively). Although we could expect that many features extracted from the rated sentence were correlated to complexity judgments, these results also suggest that humans have paid attention to the whole context when rating the middle sentence, and especially to the following sentence. The influence of context is suggested as well by the fact that we observe much lower coefficients for all correlating features belonging to the rated sentence, unlike those reported by Brunato et al. (2018) for the same sentences evaluated in isolation. Tables 4 shows the top ten features ranked by the correlation score with average complexity, for all sentences and for groups of sentences of the same length. A positive number indicates that the feature is linked to a higher perceived complexity, meaning that linguistic phenomenon makes the sentence more complex in the eyes of annotators. Conversely, a negative number is linked to lower complexity, meaning the linguistic phenomenon helps annotators in the evaluation of the sentence complexity.

Features	**L10**	**L15**	**L20**	**L25**	**L30**	**L35**	**All**
B_dep_aux:pass	-	-	-	−1	-	-	-
B_dep_compound	-	-	5	-	-	-	-
B_dep_compound:prt	−4	-	-	-	-	-	-
B_dep_flat	-	-	-	−5	-	-	-
B_dep_nmod	-	-	-	5	-	-	-
B_dep_nsubj	-	−5	-	-	-	-	-
B_dep_nsubj:pass	-	-	-	−2	-	-	-
B_dep_nummod	-	-	3	-	-	-	-
B _princ_prop	-	-	−4	-	-	-	-
B_verb_root_perc	-	-	−3	-	-	-	-
C_aux_Fin	-	-	−1	-	−4	-	-
C_aux_num_pers_+	−5	-	−5	-	-	-	-
C_aux_Pres	-	-	-	-	−5	-	-
C_avg_max_depth	-	5	-	-	-	-	4
C_avg_max_link	-	-	-	-	-	-	8
C_avg_sub_chain	-	-	-	-	-	−1	-
C_avg_tok_clause	-	-	4	-	-	-	-
C_char_tok	-	-	-	-	-	−5	-
C_dep_aux	-	-	-	-	−2	-	-
C_dep_det	-	−3	-	-	-	-	-
C_dep_nmod	5	-	-	-	-	-	-
C_dep_nummod	-	4	2	-	2	2	5
C_dep_root	-	−1	-	-	-	-	−1
C_dep_xcomp	-	-	-	−3	-	-	-
C_max_link	-	-	-	-	-	-	7
C_n_prep_chain	-	-	-	-	-	-	6
C_n_tok	3	2	-	-	-	-	2
C_tok_sent	4	3	-	-	-	-	3
C_upos_ADJ	-	−4	-	-	-	-	-
C_upos_AUX	-	-	−2	-	−3	−2	−2
C_upos_DET	-	−2	-	-	-	-	-
C_upos_NUM	1	1	1	-	1	1	1
C_upos_PRON	-	-	-	-	-	−3	-
C_upos_SYM	2	-	-	-	-	3	-
C_verb_edge_1	-	-	-	-	−1	-	-
C_verb_Fin	-	-	-	-	3	-	-
C_verb_Ind	-	-	-	-	5	-	-
E_aux_Pres	−3	-	-	-	-	-	-
E_avg_link	−2	-	-	-	-	-	-
E_avg_max_depth	-	-	-	2	-	-	-
E_dep_ccomp	-	-	-	-	-	−4	-
E_dep_nummod	-	-	-	4	-	5	-
E_lexical_dens	-	-	-	−4	-	-	-
E_upos_NUM	-	-	-	1	-	4	-
E_upos_SYM	-	-	-	3	-	-	-
E_verb_edge_4	−1	-	-	-	-	-	-
E_verb_Fin	-	-	-	-	4	-	-

Table 4: Ranking of correlations between the top 10 linguistic features and the average complexity score for all sentences and for all length bins. The number indicates the position the feature occupies in the ranking: the higher the number(positive or negative), the higher the correlation. B_*, C_*,E_* mean that the features characterize the beginning, the central and the ending sentence, respectively.

When all sentences are considered, we observe

[4] We report in the Appendix the whole tables of correlation results for all contextual windows.

Features	L10	L15	L20	L25	L30	L35	All
B_aux_Inf	2	-	-	-	-	-	-
B_dep_compound:prt	−5	-	-	-	-	-	-
B_subj_pre	-	-	-	−5	-	-	-
B_upos_SYM	-	-	-	−3	-	-	-
B_verb_edge_1	-	−2	-	-	-	-	-
B_verb_Past	-	−1	-	-	-	-	-
C_avg_sub_chain	-	-	-	-	-	-	−1
C_char_tok	-	-	-	-	-	-	−5
C_dep_aux	-	-	-	-	−1	-	-
C_dep_nummod	-	-	-	-	-	-	2
C_dep_punct	-	-	-	−4	-	-	-
C_princ_prop	-	-	-	-	-	−2	-
C_sub_prop	-	-	-	-	-	3	-
C_upos_AUX	-	-	-	-	-	-	−2
C_upos_NUM	-	-	-	-	-	-	1
C_upos_PRON	-	-	-	-	-	-	−3
C_upos_PUNCT	-	-	-	−2	-	-	-
C_upos_SYM	-	-	-	-	-	-	3
C_verb_edge_1	-	-	-	-	-	2	-
C_verb_root_perc	-	-	-	-	-	−1	-
E_avg_link	−4	-	-	-	-	-	-
E_avg_max_link	−2	-	-	-	-	-	-
E_dep_aux	−6	-	-	-	-	-	-
E_dep_ccomp	-	-	-	-	-	-	−4
E_dep_nummod	-	-	-	-	-	-	5
E_dep_parataxis	−7	-	-	-	-	-	-
E_dep_root	1	-	-	-	-	-	-
E_max_link	−3	-	-	-	-	-	-
E_upos_ADV	-	-	1	-	-	-	-
E_upos_NUM	-	-	-	-	-	-	4
E_verb_edge_3	-	-	-	−1	-	1	-
E_verb_Past	3	-	-	-	-	-	-
E_verb_Pres	−1	-	-	-	-	-	-

Table 5: Ranking of correlations between the top 10 linguistic features and complexity standard deviation for all sentences and for all length bins. Feature labels and ranking numbers are used as in 4.

that the first ten ones all belong to the middle sentence and refer to features modeling linguistic phenomena of different nature, although we can distinguish two main groups, positively correlated with the perception of sentence complexity. The first group is related to the presence of numerical information (i.e. literal numbers in the sentence), as conveyed by both POS and syntactic features (*C_upos_NUM, C_dep_nummod*). The second one, as more expected, concerns sentence length (*C_tok_sent, C_dep_root*) and features still related to length but capturing aspects of structural complexity, e.g. the depth of the whole parse tree and specific sub-trees, i.e. nominal chain headed by a preposition (*C_avg_max_depth, C_n_prep_chain*). Notably, the effect of sentence length is observed only for the middle sentence, while the length of contextual sentences is never correlated with judgments. Again, the correlation is much lower with respect to the one obtained by sentences judged in isolation (i.e. 0.31 vs 0.84 reported in the previous study). Within bins of same-length sentences, we notice a more prominent role of features from the context, as suggested by the presence of features characterizing both the sentence preceding and following the rated one in the first ten position of the ranking. Interestingly, for all bins numerical information turned out to be the feature most correlated with complexity score, being it extracted from the rated or from contextual sentences (specifically, the right sentence, for the bin composed by sentences with 25 tokens).

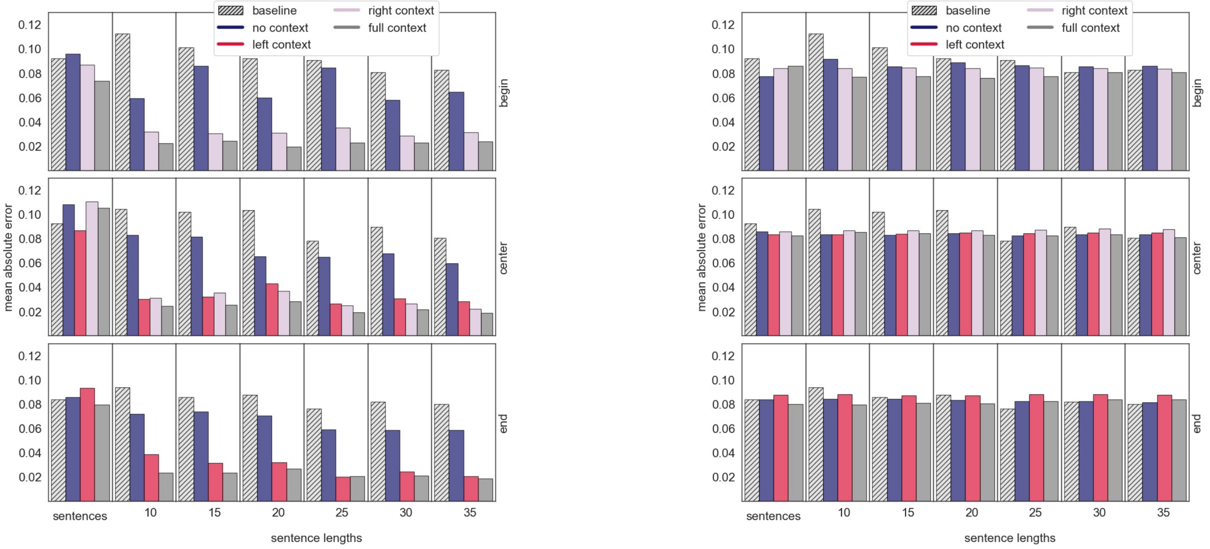
For standard deviation, we found 29 statistically significant ($p < 0.05$) features with correlation $\rho \geq 0.20$. These include 24% of features belonging to the rated sentence (i.e. 7 features), while the remaining features belong to the contextual sentences (i.e. 6 features for the left sentence, 15 for the right one). In this case we found far less correlations, with most features being significant for the length bins but not when considering the sentences altogether. These results confirm that humans have paid attention to the whole context when evaluating the sentence, but also that standard deviation, and thus annotators' agreement, is a phenomenon harder to describe and subjective to factors that linguistic features cannot fully detect. Similarly to what done for average complexity, in Table 5 we report the first ten features mostly correlated with standard deviation, for all rated sentences and for sentences of the same length. As we can see, the ranking is mostly different from the one resulting from correlating feature values and average complexity scores.

4 Predicting Sentence Complexity

The results of the correlation analysis have shown that linguistic information of the context affects the perception of sentence complexity and the extent to which this perception is shared by annotators. We thus proceed to assess the contribution of the context from a modeling standpoint. We built two regression tasks, one to predict the average complexity value assigned to each sentence, and one to predict the standard deviation of complexity for each sentence. In both scenarios, we employed two different models: the first is a linear SVM regression model with standard parameters that leverages the explicit linguistic features presented in Table 3, the second is obtained by fine-tuning the BERT base model (i.e. bert-base-uncased) on our dataset using the FARM[5] regression implementation. Both models were evaluated with a 5-fold cross validation for each of the three windows of context.

For every window, we carried out different runs of the models, varying the amount of contextual features to be considered. For the *begin window*

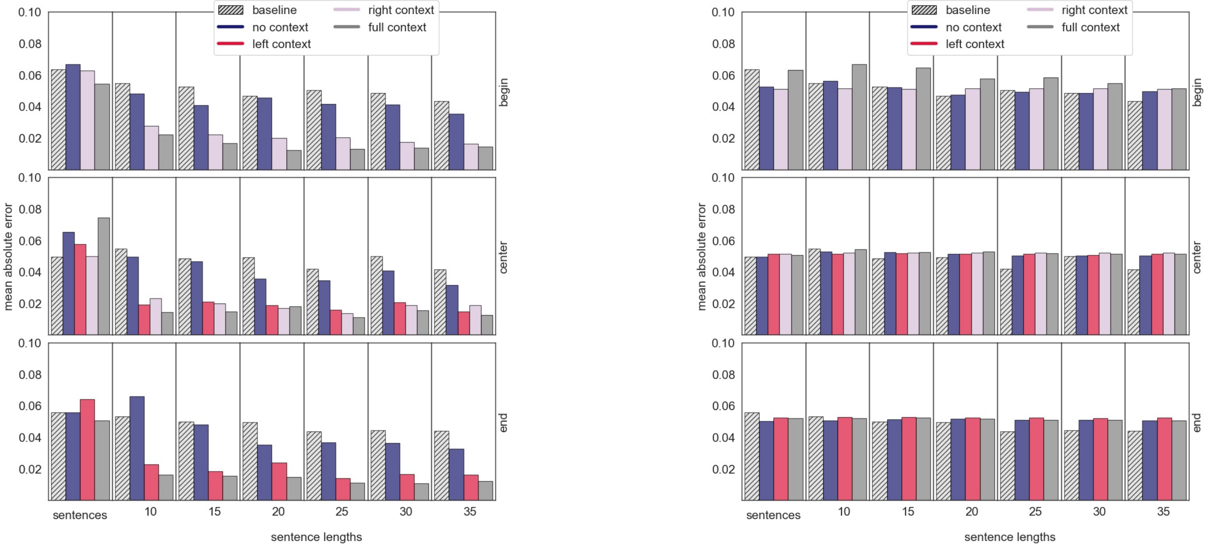
[5] github.com/deepset-ai/FARM



(a) SVM models

(b) BERT models

Figure 4: Performance (MAE) of SVM regression model on avg complexity ratings prediction. In different windows of context and with different context spans, for all sentences and at different sentence lengths.



(a) SVM models

(b) BERT models

Figure 5: Performance (MAE) of SVM regression models and BERT models in the prediction of complexity standard deviation. In different windows of context and with different context spans, for all sentences and at different sentence lengths.

and the *end window* we ran the models with *i*) the features of the single sentence (no context), *ii*) the features of the sentence + the features of the next sentence (right context) for the *begin window*, or + the features of the previous sentence (left context) for the *end window*, *iii*) the features of all the three sentences (full context, i.e. the whole window of context); for the *center window*, we trained the models with *i*) no context features, *ii*) left or right context features, *iii*) full context features. We measured the performance of the models in terms of *mean absolute error* (MAE), evaluating their accuracy in predicting the same average judgment of complexity assigned by humans and the standard deviation of the complexity judgments. We then repeated the same experiments grouping the sentences according to their length. The baseline for the models evaluation was calculated (i) in the case of all sentences, by giving in input to the linear regression model only the length of the sentence as feature for the prediction, (ii) in the case of different lengths (binned sentences), by having the model always assigning the average complexity value (calculated on the whole set of sentences) to each sentence.

Figure 4 reports the results for the prediction of the average complexity, showing the average MAE obtained after the 5-fold validation, both for SVM models and BERT models. The SVM models with linguistic features outperform BERT models overall. BERT models remain close to the baseline in all cases, despite the amount of context considered and the length of the sentences. Instead, the SVM models show significant differences as appropriate. In the case of all sentences, the performances of the model are close to the baseline. Adding contextual features slightly helps in the case of the begin and the end window, while performances worsen in the case of the center window. When considering sentences of the same length, the performance of the model is always helped by the presence of contextual features and best results are achieved when the full context is taken into account, for all the windows of context. This behavior confirms on one side that linguistic characteristics of the context are indeed very influential on complexity, on the other side that the length of the sentence plays an important role on the perception of complexity, as it is only by binning the sentences that we can exploit the effect of context in predicting complexity.

Figure 5 shows the results for the prediction of the standard deviation of complexity, for SVM models and BERT models. As in the previous case, BERT models obtain results that are in line with the baseline and that are not influenced by different amounts of context. When looking at the results obtained with the explicit linguistic features, the outcome is quite different. For the all sentences case, the SVM model cannot predict the standard deviation of complexity, although the error gets lower for the begin window and the end window when the full context is used. Conversely, the model shows large improvement when working on sentences of the same length. In all windows and for all lengths, using the features of the whole context significantly decreases the error in the prediction of standard deviation. When running the model with the features of the single sentence (i.e. no context), the performances of the model are in general close to the ones of the baseline. This suggests that the context is particularly relevant in predicting how people will agree on their perception of complexity.

Overall, our results show that information about the complexity of a sentence is better encoded in its explicit linguistic features, thus its syntactic and morphosyntactic structures. On the other hand, although BERT has been proven to embed a wide range of linguistic properties, including syntactic ones (Tenney et al., 2019; Miaschi et al., 2020), our findings seem to suggest that this model does not exploit these kind of features to solve a downstream task like ours, for which few data are available. Indeed, it has been shown that BERT performs better on datasets larger than ours (Kumar et al., 2020). Thus, it is fair to assume that more data may be needed for BERT to detect phenomenon about perceived complexity.

Moreover, our results show that the presence of context plays an important role on complexity. As the SVM models are always helped by the contextual features, it is fair to assume that annotators have taken into account the whole context when expressing their judgment upon complexity, and that the presence of the context has strongly influenced their perception. Also, contextual linguistic phenomena are the ones that impact more on the variation of complexity perception between annotators as they are the ones that help more in the prediction of this variation.

5 Conclusion

We studied how the context surrounding a sentence influences the perception of its complexity by humans. Starting from a newly collected dataset, we investigated which linguistic phenomena, among a wide set of lexical, morphosyntactic, and syntactic ones, are more correlated with complexity judgments and the degree of agreement between annotators. From a modeling standpoint, we observe that models using explicit linguistic features achieve higher accuracy than state-of-the-art neural language models in predicting the average complexity score assigned to a sentence, as well as the variation among scores. This is especially true when they use explicit linguistic features from all contextual sentences in addition to the linguistic features of the sole rated sentence.

As many NLP applications are concerned with the analysis of linguistic complexity, particularly for text readability and text simplification purposes, we think that our results emphasize the importance of considering contextual information both in the creation of gold benchmarks, which are typically based only on data paired at sentence level and in the development of cognitively inspired evaluation systems driven by how people perceive complexity.

References

Jean-Philippe Bernardy, Shalom Lappin, and Jey Han Lau. 2018. The influence of context on sentence acceptability judgements. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 456–461, Melbourne, Australia. 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 7147–7153, 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.

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.

Jonathan King and Marcel Adam Just. 1991. Individual differences in syntactic processing: The role of working memory. *Journal of Memory and Language*, 30(5):580 – 602.

W. Kintsch, E. Kozminsky, W.J. Streby, G. McKoon, and J.M. Keenan. 1975. Comprehension and recall of text as a function of content variables. *Journal of Verbal Learning and Verbal Behavior, 14(2).*

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. In *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*, pages 18–26.

Jey Han Lau, Carlos S Armendariz, Shalom Lappin, Matthew Purver, and Chang Shu. 2020. How furiously can colourless green ideas sleep? sentence acceptability in context. *arXiv preprint arXiv:2004.00881.*

Danielle S. McNamara. 2001. Reading both high-coherence and low-coherence texts: Effects of text sequence and prior knowledge. *Canadian Journal of Experimental Psychology*, 55(1):51–62.

Alessio Miaschi, Dominique Brunato, Felice Dell'Orletta, and Giulia Venturi. 2020. Linguistic profiling of a neural language model. *arXiv preprint arXiv:2010.01869.*

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

Anna Rogers, O. Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works. *ArXiv*, abs/2002.12327.

Elliot Schumacher, Maxine Eskenazi, Gwen Frishkoff, and Kevyn Collins-Thompson. 2016. Predicting the relative difficulty of single sentences with and without surrounding context. In *Conference on Empirical Methods in Natural Language Processing, pages 1871–1881.*

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

A Appendix. Results of Correlations between Linguistic Features and Complexity Average Scores (*judg*) and between Linguistic Features and Complexity Standard Deviation (*std*).

Features	Length 10		Length 15		Length 20		Length 25		Length 30		Length 35		All sents	
	judg	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*
B_aux_+	−0.20	-	-	-	-	-	-	-	-	-	-	-	-	-
B_aux_Fin	−0.29	-	-	-	−0.25	-	-	0.22	-	-	-	-	-	-
B_aux_Ind	−0.27	-	-	-	-	-	-	-	-	-	-	-	-	-
B_avg_link	0.31	-	-	-	-	-	-	-	-	-	-	-	0.21	-
B_avg_max_depth	0.25	-	0.23	-	-	-	-	-	-	-	-	-	0.29	-
B_avg_max_link	0.36	-	-	-	-	-	-	-	-	-	-	-	0.26	-
B_avg_prep_chain	-	-	-	-	-	-	-	-	-	0.20	-	-	-	-
B_avg_sub_chain	-	-	-	-	-	-	−0.23	-	-	-	-	-	-	-
B_avg_tok_clause	−0.20	-	0.25	-	-	-	-	-	-	-	-	-	-	-
B_char_tok	-	-	−0.24	-	-	-	-	-	-	-	-	-	-	-
B_dep_advmod	-	-	-	-	0.20	-	-	-	-	-	-	-	-	-
B_dep_amod	−0.25	-	-	-	-	-	-	-	-	-	-	-	-	-
B_dep_appos	0.54	-	-	-	-	-	-	-	-	-	-	-	-	-
B_dep_compound	0.27	-	-	-	-	−0.22	-	-	0.20	-	0.21	-	0.22	-
B_dep_cop	−0.25	-	-	-	-	-	-	-	-	-	-	-	-	-
B_dep_det	−0.33	-	-	-	-	-	-	-	-	-	-	-	−0.21	-
B_dep_nsubj	−0.43	-	-	-	-	-	-	-	-	-	-	-	−0.22	-
B_dep_nummod	0.39	-	0.20	-	0.30	-	0.23	-	0.33	-	0.35	-	0.33	-
B_dep_obl:tmod	-	-	-	-	-	-	-	-	-	−0.26	-	-	-	-
B_dep_punct	-	-	-	-	0.20	-	-	-	-	-	-	-	-	-
B_dep_root	−0.34	-	−0.33	-	-	-	-	-	-	-	-	-	−0.32	-
B_dep_xcomp	-	-	-	-	-	-	−0.25	-	-	-	-	-	-	-
B_lexical_dens	-	-	-	-	−0.22	-	−0.21	-	-	-	-	-	-	-
B_max_link	0.36	-	-	-	-	-	-	-	-	-	-	-	0.26	-
B_n_prep_chain	-	-	-	-	-	-	0.20	-	-	-	-	-	0.24	-
B_n_tok	0.34	-	0.33	-	-	-	-	-	-	-	-	-	0.32	-
B_obj_post	-	-	0.22	-	-	-	-	-	-	-	-	-	-	-
B_princ_prop	−0.27	-	-	-	-	-	-	-	-	-	-	-	-	-
B_sub_1	−0.24	-	-	-	-	-	-	-	-	-	-	-	-	-
B_sub_prop	-	-	-	-	-	-	−0.22	-	-	-	-	-	-	-
B_subj_pre	−0.42	-	-	-	-	-	-	-	-	-	-	-	-	-
B_tok_sent	0.34	-	0.33	-	-	-	-	-	-	-	-	-	0.32	-
B_ttr	−0.20	-	-	-	-	-	-	-	-	-	-	-	-	-
B_ttr_lemma	−0.21	-	-	-	-	-	-	−0.20	-	-	-	-	-	-
B_upos_ADJ	−0.26	-	-	-	−0.24	-	-	-	-	-	-	-	-	-
B_upos_ADP	−0.20	-	-	-	-	-	-	-	-	-	-	-	-	-
B_upos_AUX	−0.29	-	-	-	-	-	-	0.23	-	-	-	-	-	-
B_upos_DET	−0.33	-	-	-	-	-	-	-	-	-	-	-	−0.21	-
B_upos_NUM	0.40	-	0.30	-	0.33	-	0.30	-	0.34	-	0.30	-	0.34	-
B_upos_PART	-	-	-	-	-	-	-	-	-	-	−0.20	-	-	-
B_upos_PRON	−0.25	-	−0.24	-	-	-	-	-	-	-	-	-	-	-
B_upos_PUNCT	-	-	-	-	0.20	-	-	-	-	-	-	-	-	-
B_upos_SYM	0.30	-	0.22	-	-	-	0.27	-	0.29	-	0.31	-	0.28	-
B_upos_VERB	−0.30	-	-	-	-	-	-	-	-	-	-	-	-	-
B_verb_edge_0	-	-	−0.25	-	-	-	-	-	-	-	-	-	-	-
B_verb_head_sent	−0.42	-	-	-	-	-	−0.21	-	-	-	-	-	-	-
B_verb_root_perc	−0.43	−0.22	-	-	-	-	-	-	-	-	-	-	-	-
C_aux_+	-	-	-	-	-	−0.21	-	-	-	-	-	-	-	-
C_aux_Fin	−0.31	-	-	-	−0.21	-	-	-	-	-	-	-	-	-
C_aux_Ind	−0.32	-	-	-	-	-	-	-	-	-	-	-	-	-
C_aux_Pres	−0.23	-	-	-	-	-	-	-	-	-	-	-	-	-
C_aux_Sing+3	−0.29	-	-	-	-	-	-	-	-	-	-	-	-	-
C_avg_link	−0.23	-	-	-	-	-	-	-	-	-	-	-	-	-
C_avg_sub_chain	−0.24	-	-	-	-	-	-	-	-	-	-	-	-	-
C_avg_tok_clause	−0.33	-	-	-	-	-	-	-	-	-	-	-	-	-
C_avg_verb_edge	−0.30	-	-	-	−0.21	-	-	-	-	-	-	-	-	-
C_char_tok	0.28	-	-	-	-	-	-	-	-	-	-	-	-	-
C_dep_aux	-	-	-	-	−0.21	-	-	-	-	-	-	-	-	-
C_dep_aux:pass	-	-	-	-	-	-	-	-	-	0.23	-	-	-	-
C_dep_cc	−0.23	-	-	-	−0.20	-	-	-	-	-	-	-	-	-
C_dep_ccomp	−0.23	-	-	-	-	-	-	-	-	-	-	-	-	-
C_dep_compound	0.21	-	-	-	0.24	-	-	-	-	-	-	-	-	-
C_dep_nmod:poss	−0.24	-	-	-	-	-	-	-	-	-	-	-	-	-
C_dep_nsubj	-	-	-	-	−0.31	-	-	-	-	-	-	-	-	-
C_dep_nsubj:pass	-	-	-	-	-	-	-	-	-	0.23	-	-	-	-
C_dep_nummod	-	-	-	-	0.28	-	0.27	-	0.22	-	0.26	-	0.23	-
C_dep_obj	−0.21	-	-	-	-	-	-	-	-	-	-	-	-	-
C_dep_obl	−0.25	-	-	-	-	-	-	-	-	-	-	-	-	-
C_dep_root	0.30	-	-	-	-	-	-	-	-	-	-	-	-	-
C_n_tok	−0.30	-	-	-	-	-	-	-	-	-	-	-	-	-
C_obj_post	−0.24	-	-	-	−0.21	-	-	-	-	-	-	-	-	-
C_prep_3	0.37	-	-	-	-	-	-	-	-	-	0.22	-	-	-
C _princ_prop	−0.27	-	-	-	-	-	-	-	-	-	-	-	-	-
C_sub_1	−0.30	-	-	-	-	-	-	-	-	-	-	-	-	-
C_sub_post	−0.28	-	-	-	-	-	-	-	-	-	-	-	-	-
C_sub_pre	-	-	-	-	−0.24	-	-	-	-	-	-	-	-	-

continued on next page

continued from previous page

Features	**Length 10**		**Length 15**		**Length 20**		**Length 25**		**Length 30**		**Length 35**		**All sents**	
	judg	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*
C_sub_prop	−0.22	-	-	-	-	-	-	-	-	-	-	-	-	-
C_subj_pre	−0.39	-	-	-	-	-	-	-	-	-	-	-	-	-
C_tok_sent	−0.30	-	-	-	-	-	-	-	-	-	-	-	-	-
C_upos_AUX	−0.22	-	-	-	−0.23	-	-	-	-	-	-	-	-	-
C_upos_CCONJ	−0.21	-	-	-	-	-	-	-	-	-	-	-	-	-
C_upos_DET	−0.23	-	-	-	-	-	-	-	-	-	-	-	-	-
C_upos_NUM	-	-	-	-	0.24	-	0.35	-	0.23	-	0.26	-	0.24	-
C_upos_PART	−0.25	-	-	-	-	-	-	-	-	-	-	-	-	-
C_upos_PRON	−0.27	-	−0.23	-	-	-	-	-	-	-	-	-	-	-
C_upos_VERB	−0.29	-	-	-	−0.27	-	-	-	-	-	-	-	-	-
C_verb_edge_5	−0.31	-	-	-	-	-	-	-	-	-	-	-	-	-
C_verb_head_sent	−0.38	-	-	-	−0.28	-	-	-	-	-	-	-	-	-
C_verb_Ind	−0.23	-	-	-	-	-	-	-	-	-	-	-	-	-
C_verb_Inf	−0.21	-	-	-	-	−0.20	-	-	-	-	-	-	-	-
C_verb_Part	−0.24	-	-	-	-	-	-	-	-	-	-	-	-	-
C_verb_Past	−0.29	-	-	-	-	-	-	-	-	-	-	-	-	-
C_verb_Pres	-	-	-	-	−0.26	-	-	-	-	-	-	-	-	-
C_verb_root_perc	−0.44	-	-	-	−0.29	-	-	-	-	-	-	-	-	-
C_verb_Sing+3	-	-	-	-	−0.25	-	-	-	-	-	-	-	-	-
E_aux_Fin	−0.29	-	-	-	−0.21	-	-	-	-	-	-	-	-	-
E_aux_Ind	−0.22	-	-	-	−0.22	-	-	-	-	0.22	-	-	-	-
E_aux_Pres	-	-	-	-	−0.24	-	−0.21	-	-	-	-	-	-	-
E_avg_link	0.32	-	-	-	0.31	-	-	-	-	-	-	-	-	-
E_avg_max_link	-	-	-	-	0.29	-	-	-	-	-	-	-	-	-
E_avg_prep_chain	0.26	-	-	-	-	-	-	-	-	-	0.23	-	-	-
E_avg_verb_edge	−0.23	-	-	-	-	-	-	-	-	-	-	-	-	-
E_dep_advmod	-	-	−0.25	−0.24	-	-	-	-	-	-	-	-	-	-
E_dep_appos	0.37	-	-	-	0.21	-	0.23	-	-	-	-	-	-	-
E_dep_det	−0.25	-	-	-	-	-	-	-	-	-	-	-	-	-
E_dep_list	-	-	-	-	0.22	-	-	-	-	-	-	-	-	-
E_dep_nmod	0.35	-	-	-	-	-	-	-	-	-	0.25	-	-	-
E_dep_nsubj	−0.29	-	-	-	−0.25	-	-	-	-	-	-	-	-	-
E_dep_nummod	0.40	-	-	-	-	-	-	-	-	-	-	-	0.21	-
E_dep_obj	−0.31	-	-	-	-	-	-	-	-	-	-	-	-	-
E_lexical_dens	−0.30	-	-	-	-	-	-	-	-	-	-	-	-	-
E_max_link	-	-	-	-	0.29	-	-	-	-	-	-	-	-	-
E_n_prep_chain	0.32	-	-	-	0.22	-	-	-	-	-	0.20	-	-	-
E_obj_post	−0.28	-	-	-	-	-	-	-	-	-	-	-	-	-
E_prep_1	0.29	-	-	-	-	-	-	-	-	-	-	-	-	-
E _princ_prop	−0.23	-	-	-	-	-	-	-	-	-	-	-	-	-
E_sub_pre	-	−0.21	-	-	-	-	-	-	-	-	-	-	-	-
E_subj_pre	−0.45	-	-	-	-	-	-	-	-	-	-	-	-	-
E_ttr	−0.31	-	-	-	−0.29	-	-	-	-	-	-	-	-	-
E_ttr_lemma	−0.30	-	-	-	−0.29	-	-	-	-	-	-	-	-	-
E_upos_ADP	0.22	-	-	-	-	-	-	-	-	-	-	-	-	-
E_upos_AUX	−0.24	-	-	-	−0.27	-	-	-	-	-	-	-	-	-
E_upos_DET	−0.27	-	-	-	-	-	-	-	-	-	-	-	-	-
E_upos_NUM	0.39	-	-	-	0.23	-	-	-	-	-	0.21	-	0.23	-
E_upos_PRON	−0.34	-	-	-	-	-	-	-	-	-	-	-	-	-
E_upos_VERB	−0.38	-	-	-	-	-	-	-	-	-	−0.24	-	-	-
E_verb_edge_1	−0.29	-	-	-	-	-	-	-	-	-	-	-	-	-
E_verb_edge_3	−0.23	-	-	-	-	-	-	-	-	-	-	-	-	-
E_verb_Ger	-	-	-	-	0.20	-	-	-	-	-	-	-	-	-
E_verb_head_sent	−0.30	-	-	-	-	-	-	-	-	-	-	-	-	-
E_verb_root_perc	−0.41	-	-	-	−0.21	-	-	-	-	-	-	-	-	-

Table 6: Values of correlation for statistically significant (p-value<0.05) linguistic features with $\rho \geq 0.20$ that correlate with either the average judgment of complexity or the complexity standard deviation. For the *begin context window*, for all sentences and for sentences divided according to their length.

Features	**Length 10**		**Length 15**		**Length 20**		**Length 25**		**Length 30**		**Length 35**		**All sents**	
	judg	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*
B_aux_+	−0.21	-	-	-	-	-	-	-	-	-	-	-	-	-
B_aux_form_Ger	-	-	0.21	-	-	-	-	-	-	-	-	-	-	-
B_aux_form_Inf	-	0.20	-	-	-	-	-	-	-	-	-	-	-	-
B_aux_Pres	-	-	-	-	-	-	-	-	−0.21	-	-	-	-	-
B_avg_prep_chain	-	-	-	-	-	-	0.23	-	-	-	-	-	-	-
B_avg_sub_chain	−0.24	-	-	-	-	-	-	-	-	-	-	-	-	-
B_dep_aux	-	-	-	-	-	-	-	-	-	-	−0.28	-	-	-
B_dep_aux:pass	-	-	-	-	-	-	−0.32	-	-	-	-	-	-	-
B_dep_compound	-	-	0.20	-	0.21	-	-	-	-	-	0.22	-	0.21	-
B_dep_flat	-	-	-	-	-	-	−0.22	-	-	-	-	-	-	-
B_dep_nmod	-	-	-	-	-	-	0.25	-	-	-	-	-	-	-
B_dep_nsubj	-	-	−0.24	-	-	-	-	-	-	-	−0.21	-	-	-
B_dep_nsubj:pass	-	-	-	-	-	-	−0.29	-	-	-	-	-	-	-
B_dep_nummod	-	-	0.27	-	0.23	-	-	-	-	-	0.26	-	-	-
B_n_prep_chain	-	-	-	-	-	-	0.23	-	-	-	-	-	-	-

continued on next page

continued from previous page

Features	**Length 10**		**Length 15**		**Length 20**		**Length 25**		**Length 30**		**Length 35**		**All sents**	
	judg	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*
B_princ_prop	-	-	-	-	−0.24	-	-	-	-	-	-	-	-	-
B_sub_post	−0.22	-	-	-	-	-	-	-	-	-	-	-	-	-
B_subj_pre	-	-	-	-	-	-	-	−0.20	-	-	-	-	-	-
B_upos_NUM	-	-	0.22	-	-	-	-	-	-	-	0.29	-	-	-
B_upos_PRON	-	-	-	-	-	-	-	-	−0.22	-	-	-	-	-
B_upos_PROPN	-	-	0.21	-	-	-	-	-	-	-	-	-	-	-
B_upos_SYM	-	-	-	-	-	-	-	−0.21	-	-	0.21	-	-	-
B_upos_VERB	-	-	−0.21	-	−0.22	-	-	-	-	-	-	-	-	-
B_verb_edge_1	-	-	-	−0.20	-	-	-	-	-	-	-	-	-	-
B_verb_Past	-	-	-	−0.21	-	-	-	-	-	-	-	-	-	-
B_verb_root_perc	-	-	-	-	−0.25	-	-	-	-	-	-	-	-	-
C_aux_+	−0.24	-	-	-	−0.24	-	-	-	-	-	−0.20	-	-	-
C_aux_form_Fin	−0.21	-	-	-	−0.28	-	-	-	−0.22	-	-	-	-	-
C_aux_Ind	-	-	-	-	−0.23	-	-	-	-	-	-	-	-	-
C_aux_Past	-	-	-	-	−0.21	-	-	-	-	-	-	-	-	-
C_aux_Pres	−0.21	-	-	-	-	-	-	-	−0.22	-	−0.24	-	-	-
C_avg_max_depth	0.21	-	0.31	-	-	-	-	-	-	-	-	-	0.29	-
C_avg_max_link	-	-	-	-	-	-	-	-	-	-	-	-	0.25	-
C_avg_sub_chain	-	-	-	-	-	-	-	-	−0.20	-	−0.38	-	-	-
C_avg_tok_clause	-	-	-	-	0.22	-	-	-	-	-	0.26	-	-	-
C_char_tok	-	-	-	-	-	-	-	-	-	-	−0.30	-	-	-
C_dep_amod	-	-	-	-	−0.24	-	-	-	-	-	-	-	-	-
C_dep_aux	-	-	−0.22	-	−0.21	-	-	-	−0.28	−0.25	−0.29	-	-	-
C_dep_case	-	-	-	-	-	-	-	-	-	-	0.22	-	-	-
C_dep_ccomp	-	-	-	-	-	-	-	-	-	-	−0.27	-	-	-
C_dep_det	-	-	−0.28	-	−0.22	-	-	-	-	-	-	-	-	-
C_dep_mark	-	-	-	-	-	-	-	-	-	-	−0.23	-	-	-
C_dep_nmod	0.23	-	-	-	-	-	-	-	-	-	-	-	-	-
C_dep_nsubj	−0.22	-	-	-	-	-	-	-	-	-	-	-	-	-
C_dep_nummod	0.21	-	0.32	-	0.25	-	-	-	0.26	-	0.35	-	0.29	-
C_dep_punct	-	-	-	-	-	-	-	−0.21	-	-	-	-	-	-
C_dep_root	−0.24	-	−0.33	-	-	-	-	-	-	-	-	-	−0.31	-
C_dep_xcomp	-	-	-	-	-	-	−0.26	-	-	-	−0.28	-	-	-
C_lexical_dens	-	-	−0.23	-	−0.21	-	-	-	-	-	−0.27	-	-	-
C_max_link	-	-	-	-	-	-	-	-	-	-	-	-	0.25	-
C_n_prep_chain	0.23	-	-	-	-	-	-	-	-	-	-	-	0.25	-
C_n_tok	0.24	-	0.33	-	-	-	-	-	-	-	-	-	0.31	-
C_princ_prop	-	-	-	-	-	-	-	-	-	-	0.23	−0.21	-	-
C_sub_2	-	-	-	-	-	-	-	-	-	-	−0.20	-	-	-
C_sub_4	-	-	-	-	-	-	-	-	-	-	−0.24	-	-	-
C_sub_post	-	-	-	-	-	-	-	-	-	-	−0.28	-	-	-
C_sub_prop	-	-	-	-	-	-	-	-	-	-	−0.29	0.20	-	-
C_tok_sent	0.24	-	0.33	-	-	-	-	-	-	-	-	-	0.31	-
C_upos_ADJ	−0.21	-	−0.25	-	−0.22	-	-	-	-	-	−0.26	-	-	-
C_upos_AUX	−0.24	-	-	-	−0.27	-	-	-	−0.23	-	−0.32	-	−0.23	-
C_upos_DET	-	-	−0.28	-	−0.21	-	-	-	-	-	-	-	-	-
C_upos_NUM	0.30	-	0.41	-	0.31	-	-	-	0.28	-	0.39	-	0.33	-
C_upos_PRON	−0.21	-	-	-	−0.21	-	-	-	-	-	−0.31	-	-	-
C_upos_PUNCT	-	-	-	-	-	-	-	−0.21	-	-	-	-	-	-
C_upos_SYM	0.26	-	0.30	-	-	-	-	-	-	-	0.34	-	0.24	-
C_upos_VERB	-	-	-	-	-	-	-	-	-	-	−0.24	-	-	-
C_verb_+	-	-	-	-	-	-	0.22	-	-	-	-	-	-	-
C_verb_edge_1	-	-	-	-	-	-	-	-	−0.28	-	-	0.20	-	-
C_verb_edge_2	-	-	-	-	-	-	-	-	-	-	−0.26	-	-	-
C_verb_form_Fin	-	-	-	-	-	-	-	-	0.24	-	-	-	-	-
C_verb_form_Inf	-	-	-	-	-	-	-	-	-	-	−0.27	-	-	-
C_verb_head_sent	-	-	-	-	−0.23	-	-	-	-	-	−0.28	-	-	-
C_verb_Ind	-	-	-	-	-	-	-	-	0.21	-	-	-	-	-
C_verb_root_perc	-	-	-	-	-	-	-	-	-	-	-	−0.22	-	-
E_aux_Pres	−0.27	-	−0.21	-	-	-	-	-	-	-	-	-	-	-
E_avg_link	−0.29	−0.23	-	-	-	-	-	-	-	-	-	-	-	-
E_avg_max_depth	-	-	-	-	-	-	0.30	-	-	-	-	-	-	-
E_avg_max_link	−0.23	−0.25	-	-	-	-	-	-	-	-	-	-	-	-
E_avg_sub_chain	−0.21	-	-	-	-	-	-	-	-	-	−0.26	-	-	-
E_avg_tok_clause	-	-	-	-	-	-	-	-	-	-	0.25	-	-	-
E_avg_verb_edge	−0.21	-	-	-	-	-	-	-	-	-	-	-	-	-
E_dep_advmod	−0.24	-	−0.20	-	-	-	-	-	-	-	-	-	-	-
E_dep_aux	-	−0.22	-	-	-	-	-	-	-	-	-	-	-	-
E_dep_case	-	-	-	-	-	-	0.20	-	-	-	-	-	-	-
E_dep_ccomp	-	-	-	-	-	-	-	-	-	-	−0.31	-	-	-
E_dep_nummod	-	-	-	-	-	-	0.28	-	-	-	0.33	-	0.22	-
E_dep_parataxis	-	−0.22	-	-	-	-	-	-	-	-	-	-	-	-
E_dep_root	0.21	0.21	-	-	-	-	-	-	-	-	-	-	-	-
E_dep_xcomp	-	−0.21	−0.23	-	-	-	-	-	-	-	-	-	-	-
E_lexical_dens	-	-	-	-	-	-	−0.25	-	-	-	−0.22	-	-	-
E_max_link	−0.23	−0.25	-	-	-	-	-	-	-	-	-	-	-	-
E_n_tok	−0.21	−0.21	-	-	-	-	-	-	-	-	-	-	-	-
E_prep_1	-	-	-	-	−0.22	-	-	-	-	-	-	-	-	-
E_prep_2	-	−0.20	-	-	-	-	0.20	-	-	-	-	-	-	-
E_sub_post	-	-	-	-	-	-	-	-	-	-	−0.22	-	-	-
E_sub_pre	-	-	−0.22	-	-	-	-	-	-	-	-	-	-	-
E_sub_prop	−0.23	-	-	-	-	-	-	-	-	-	-	-	-	-

continued on next page

continued from previous page

Features	**Length 10**		**Length 15**		**Length 20**		**Length 25**		**Length 30**		**Length 35**		**All sents**	
	judg	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*
E_tok_sent	−0.21	−0.21	-	-	-	-	-	-	-	-	-	-	-	-
E_upos_ADV	-	-	−0.23	-	-	0.22	-	-	-	-	-	-	-	-
E_upos_NUM	-	-	-	-	-	-	0.33	-	-	-	0.34	-	0.22	-
E_upos_PART	-	-	-	-	-	-	-	-	-	-	−0.23	-	-	-
E_upos_PRON	−0.22	-	-	-	-	-	-	-	-	-	-	-	-	-
E_upos_SYM	-	-	-	-	-	-	0.28	-	-	-	0.30	-	-	-
E_upos_VERB	-	-	-	-	-	-	-	-	-	-	−0.21	-	-	-
E_verb_edge_3	-	-	-	-	-	-	-	−0.22	-	-	-	0.21	-	-
E_verb_edge_4	−0.30	-	-	-	-	-	-	-	-	-	-	-	-	-
E_verb_form_Fin	-	-	-	-	-	-	-	-	0.21	-	-	-	-	-
E_verb_form_Inf	-	-	-	-	-	-	-	-	-	-	−0.22	-	-	-
E_verb_head_sent	−0.24	-	-	-	-	-	-	-	-	-	-	-	-	-
E_verb_Past	-	0.20	-	-	-	-	0.23	-	-	-	-	-	-	-
E_verb_Pres	−0.20	−0.30	-	-	-	-	-	-	-	-	-	-	-	-
E_verb_Sing+3	−0.20	−0.21	-	-	-	-	-	-	-	-	-	-	-	-

Table 7: Values of correlation for statistically significant (p-value<0.05) linguistic features with $\rho \geq 0.20$ that correlate with either the average judgment of complexity or the complexity standard deviation. For the *center context window*, for all sentences and for sentences divided according to their length.

Features	**Length 10**		**Length 15**		**Length 20**		**Length 25**		**Length 30**		**Length 35**		**All sents**	
	judg	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*
B_aux_Fin	-	-	-	-	−0.23	-	-	-	-	-	-	-	-	-
B_aux_Ind	-	-	-	-	−0.21	-	-	-	-	-	-	-	-	-
B_avg_link	-	-	-	-	−0.25	-	-	-	-	-	-	-	-	-
B_avg_max_link	-	-	-	-	−0.24	-	-	-	-	-	-	-	-	-
B_dep_acl	-	-	-	-	-	−0.23	-	-	-	-	-	-	-	-
B_dep_advcl	-	-	-	-	-	-	-	-	−0.20	-	-	-	-	-
B_dep_case	-	-	-	-	−0.21	-	-	-	-	-	-	-	-	-
B_dep_ccomp	-	-	-	-	-	-	-	-	-	-	−0.21	-	-	-
B_dep_nmod:poss	-	-	-	-	-	-	-	−0.22	-	-	-	-	-	-
B_dep_obj	-	-	−0.25	-	-	-	-	-	-	-	-	-	-	-
B_dep_obl	-	-	-	-	−0.26	-	-	-	-	-	-	-	-	-
B_dep_xcomp	-	-	−0.21	-	-	-	-	-	-	-	-	-	-	-
B_max_link	-	-	-	-	−0.24	-	-	-	-	-	-	-	-	-
B_prep_3	-	-	-	-	-	-	-	−0.20	-	-	-	-	-	-
B_sub_1	-	-	-	-	−0.23	-	-	−0.25	-	-	-	-	-	-
B_subj_pre	-	-	-	-	−0.21	-	-	-	-	-	-	-	-	-
B_ttr	-	-	-	-	−0.25	-	−0.20	-	-	-	-	-	-	-
B_ttr_lemma	-	-	-	-	−0.22	-	−0.22	-	-	-	-	-	-	-
B_upos_ADP	-	-	-	-	−0.23	-	-	-	-	-	-	-	-	-
B_upos_AUX	-	-	-	-	−0.26	-	-	-	-	-	-	-	-	-
B_upos_NOUN	-	-	-	-	-	-	-	-	-	-	0.22	-	-	-
B_upos_SYM	0.23	-	-	-	-	-	-	-	-	-	-	-	-	-
B_upos_VERB	-	-	-	-	−0.23	-	-	-	−0.24	-	-	-	-	-
B_verb_head_sent	-	-	-	-	−0.21	-	-	-	-	-	-	-	-	-
B_verb_Part	-	-	-	-	−0.26	-	-	-	-	-	-	-	-	-
B_verb_root_perc	-	-	-	-	-	-	-	-	-	-	-	−0.20	-	-
C_aux_Fin	−0.21	-	−0.21	-	−0.26	-	-	-	-	-	-	-	-	-
C_char_tok	-	-	-	-	-	-	-	-	-	-	−0.20	-	-	-
C_dep_appos	-	-	-	-	0.26	-	-	-	-	-	-	-	-	-
C_dep_aux	-	-	−0.27	-	-	-	-	-	-	-	-	-	-	-
C_dep_case	0.22	-	-	-	-	-	-	-	-	-	-	-	-	-
C_dep_compound	-	-	0.22	-	0.22	-	-	-	-	-	-	-	-	-
C_dep_det	−0.21	-	-	-	-	-	-	-	-	-	-	-	-	-
C_dep_fixed	-	-	-	-	-	−0.21	-	-	-	-	-	-	-	-
C_dep_nmod	0.22	-	-	-	-	-	-	-	-	-	-	-	-	-
C_dep_nsubj	-	-	-	-	−0.20	-	-	-	-	-	-	-	-	-
C_dep_nummod	-	-	-	-	0.26	-	-	-	0.20	-	-	-	-	-
C_dep_obl	-	0.20	-	-	-	-	-	-	-	-	-	-	-	-
C_dep_obl:tmod	-	−0.25	-	-	-	-	-	-	-	-	-	-	-	-
C_dep_punct	-	-	-	-	0.23	-	-	-	-	-	-	-	-	-
C_sub_2	-	0.22	-	-	-	-	-	-	-	-	-	-	-	-
C_sub_post	-	0.27	-	-	-	-	-	-	-	-	-	-	-	-
C_sub_pre	−0.22	−0.23	-	-	-	-	-	-	-	-	-	-	-	-
C_sub_prop	-	0.22	-	-	-	-	-	-	-	-	-	-	-	-
C_subj_pre	-	-	-	-	−0.27	-	-	-	-	-	-	-	-	-
C_ttr	-	-	-	-	−0.24	-	-	-	-	-	-	0.23	-	-
C_ttr_lemma	-	-	0.22	-	−0.27	-	-	-	-	-	-	0.22	-	-
C_upos_AUX	−0.25	-	−0.21	-	−0.21	-	-	-	-	-	-	-	-	-
C_upos_DET	−0.23	-	-	-	-	-	-	-	−0.22	-	-	-	-	-
C_upos_NUM	-	-	-	-	0.26	-	-	-	0.21	-	-	-	-	-
C_upos_PRON	-	-	−0.21	-	-	-	-	-	-	-	-	-	-	-
C_upos_PROPN	-	-	0.25	-	-	-	-	-	-	-	-	-	-	-
C_upos_PUNCT	-	-	-	-	0.23	-	-	-	-	-	-	-	-	-
C_upos_SYM	-	-	-	-	-	-	-	-	-	-	0.26	-	-	-
C_verb_Past	-	-	0.28	-	-	-	-	-	-	-	0.23	-	-	-

continued on next page

continued from previous page

Features	**Length 10**		**Length 15**		**Length 20**		**Length 25**		**Length 30**		**Length 35**		**All sents**	
	judg	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*
C_verb_Pres	-	-	−0.20	-	-	-	-	-	-	-	-	-	-	-
C_verb_root_perc	-	-	-	-	−0.28	-	-	-	-	-	-	-	-	-
E_aux_Fin	-	-	-	-	−0.23	-	-	-	-	-	-	-	-	-
E_aux_Inf	-	-	-	-	-	-	-	-	-	-	−0.25	-	-	-
E_aux_Pres	−0.20	-	-	-	-	-	-	-	−0.21	-	-	-	-	-
E_avg_link	-	-	-	-	-	-	-	-	-	-	-	-	0.24	-
E_avg_max_depth	0.21	-	0.22	-	-	-	-	-	-	-	-	-	0.27	-
E_avg_max_link	-	-	-	-	-	-	-	-	-	-	-	-	0.28	-
E_avg_sub_chain	-	-	-	-	-	-	-	-	-	-	−0.28	-	-	-
E_avg_tok_clause	-	-	-	-	-	-	-	-	0.20	-	-	-	-	-
E_avg_verb_edge	−0.28	−0.21	-	-	-	-	-	-	-	-	-	-	-	-
E_char_tok	-	-	-	-	-	-	−0.22	-	-	-	-	-	-	-
E_dep_acl:relcl	-	-	-	-	-	-	0.21	-	-	-	-	-	-	-
E_dep_advcl	-	-	-	-	-	-	-	-	−0.20	-	-	-	-	-
E_dep_advmod	-	-	−0.23	-	-	-	-	-	-	-	-	-	-	-
E_dep_amod	-	-	−0.23	-	-	-	-	-	-	-	-	-	-	-
E_dep_appos	0.28	-	-	-	-	-	-	-	-	0.23	-	-	-	-
E_dep_aux	-	-	-	-	-	-	-	-	-	-	−0.32	-	-	-
E_dep_compound	0.20	-	0.27	-	-	-	-	-	0.22	-	-	-	0.21	-
E_dep_det	-	-	−0.30	-	−0.33	-	-	-	-	-	-	-	-	-
E_dep_mark	-	-	-	-	-	-	-	-	-	-	−0.29	-	-	-
E_dep_nmod	0.20	-	-	-	-	-	-	-	-	-	-	-	-	-
E_dep_nsubj	-	-	-	-	-	-	-	-	-	-	-	-	−0.21	-
E_dep_nummod	-	-	-	-	0.27	-	0.23	-	0.21	-	0.25	-	0.22	-
E_dep_obj	-	−0.22	-	-	-	-	-	-	-	-	-	-	-	-
E_dep_obl	-	-	-	-	-	-	-	-	-	-	−0.27	-	-	-
E_dep_parataxis	-	-	-	-	-	-	0.22	-	-	-	-	-	-	-
E_dep_punct	-	-	-	-	0.22	-	-	-	-	-	-	-	-	-
E_dep_root	-	-	−0.33	-	-	-	-	-	-	-	-	-	−0.33	-
E_lexical_dens	-	-	-	-	-	-	−0.29	-	-	-	-	-	-	-
E_max_link	-	-	-	-	-	-	-	-	-	-	-	-	0.28	-
E_n_tok	-	-	0.33	-	-	-	-	-	-	-	-	-	0.33	-
E_obj_post	-	−0.23	-	-	-	-	-	-	-	-	-	-	-	-
E_sub_2	-	-	-	-	-	-	−0.21	-	-	-	−0.23	-	-	-
E_sub_post	-	-	-	-	-	-	-	-	-	-	−0.25	-	-	-
E_subj_pre	−0.32	−0.23	-	-	-	-	-	-	-	-	-	-	-	-
E_tok_sent	-	-	0.33	-	-	-	-	-	-	-	-	-	0.33	-
E_ttr	-	-	-	-	−0.22	-	−0.21	-	-	-	−0.23	-	−0.20	-
E_ttr_lemma	-	-	-	-	−0.22	-	-	-	-	-	−0.20	-	-	-
E_upos_ADV	−0.21	-	-	-	-	-	-	-	-	-	-	-	-	-
E_upos_AUX	-	-	-	-	−0.24	-	-	-	-	-	-	-	-	-
E_upos_DET	-	-	−0.30	-	−0.33	-	-	-	-	-	-	-	-	-
E_upos_NOUN	-	-	-	-	−0.25	-	-	-	-	-	-	-	-	-
E_upos_NUM	-	-	0.21	-	0.28	-	0.27	-	-	-	0.28	-	0.25	-
E_upos_PART	-	-	-	-	-	-	-	-	-	-	−0.23	-	-	-
E_upos_PRON	−0.22	-	-	-	-	-	-	-	−0.21	-	−0.24	-	-	-
E_upos_PROPN	-	-	-	-	-	-	-	-	-	0.24	-	-	-	-
E_upos_PUNCT	-	-	-	-	0.22	-	-	-	-	-	-	-	-	-
E_upos_SYM	-	−0.23	-	-	-	-	0.23	-	-	-	0.27	-	0.21	-
E_upos_VERB	-	-	-	-	-	-	-	-	−0.24	-	−0.25	-	-	-
E_verb_edge_2	-	-	-	-	-	-	-	-	-	-	−0.24	-	-	-
E_verb_edge_3	−0.24	-	-	-	-	-	-	-	-	-	-	-	-	-
E_verb_edge_6	-	-	-	-	-	−0.21	-	-	-	-	-	-	-	-
E_verb_Fin	-	-	-	-	0.22	-	-	-	-	-	-	-	-	-
E_verb_Ger	-	-	-	-	-	-	-	-	−0.25	-	-	-	-	-
E_verb_head_sent	-	-	-	-	−0.20	-	-	-	−0.20	-	-	-	-	-
E_verb_Inf	-	-	-	-	−0.23	-	-	-	-	-	−0.22	-	-	-
E_verb_Pres	−0.22	-	-	-	-	-	-	-	-	-	-	-	-	-
E_verb_root_perc	−0.20	-	-	-	-	-	-	-	-	-	-	-	-	-
E_verb_Sing+3	-	-	-	-	0.23	-	-	-	-	-	-	-	-	-

Table 8: Values of correlation for statistically significant (p-value< 0.05) linguistic features with $\rho \geq 0.20$ that correlate with either the average judgment of complexity or the complexity standard deviation. For the *end context window*, for all sentences and for sentences divided according to their length.