Paying attention to the words: explaining readability prediction for French as a foreign language

Rodrigo Wilkens, Patrick Watrin, Thomas François
CENTAL, IL&C, University of Louvain, Belgium
{rodrigo.wilkens, patrick.watrin, thomas.francois}@uclouvain.be

Abstract

Automatic text Readability Assessment (ARA) has been seen as a way of helping people with reading difficulties. Recent advancements in Natural Language Processing have shifted ARA from linguistic-based models to more precise black-box models. However, this shift has weakened the alignment between ARA models and the reading literature, potentially leading to inaccurate predictions based on unintended factors. In this paper, we investigate the explainability of ARA models, inspecting the relationship between attention mechanism scores, ARA features, and CEFR level predictions made by the model. We propose a method for identifying features associated with the predictions made by a model through the use of the attention mechanism. Exploring three feature families (i.e., psycho-linguistic, word frequency and graded lexicon), we associated features with the model's attention heads. Finally, while not fully explanatory of the model's performance, the correlations of these associations surpass those between features and text readability levels.

Keywords: readability, model explainability, linguistic features, attention maps

1. Introduction

A significant proportion of the population suffers from poor reading skills in their everyday life (Schleicher, 2019, 2022). According to the results of international surveys on reading abilities like PISA (Schleicher, 2019), approximately 20% of 15-year-old students are ranked as poor readers. This highlights the widespread nature of reading difficulties among young individuals globally and reminds us of the importance of improving literacy skills and assisting those struggling with reading difficulties. Poor reading skills may make day-to-day life difficult, e.g., restricting access to medical information (Friedman and Hoffman-Goetz, 2006) or complicating administrative tasks (Kimble, 1992). Automatic Readability Assessment (ARA) has long been seen as a means of combating these difficulties, for example, by automating recommendations of texts suited to a specific audience to support reading practice and the development of reading skills (Pera and Ng, 2014; Sare et al., 2020).

Research on readability assessment traces back to the 1920s' when Lively and Pressey (1923) used statistical models for predicting the reading difficulty of texts.[1] These models are commonly named readability formulas. At the time, readability formulas were computed by hand and designed as a trade-off between reliability and minimization of effort (e.g., (Flesch, 1948; Dale and Chall, 1948). Later, the first automatized formulas appeared, such as the Automated Readability Index (Smith and Senter, 1967). In addition, readability formulas incorporate features (Bormuth, 1966; Coleman and Liau, 1975; Kintsch and Vipond, 1979).

With the advent of the 21st century, the use of Natural Language Processing (NLP) techniques enabled researchers to capture complex textual features automatically, and sophisticated Machine Learning (ML) algorithms allowed them to combine them better through feature engineering (see François and Miltsakaki, 2012; Crossley and McNamara, 2012; Collins-Thompson, 2014; Vajjala, 2021). These models rely on linguistic features exploiting knowledge about the reading process from cognitive psychology (Chall and Dale, 1995), offering insights on how textual characteristics affect readers (Javourey-Drevet et al., 2022). For instance, Collins-Thompson and Callan (2005) showed that taking into account word distributions across grade levels within a multinomial Naïve Bayes classifier outperforms classic readability formulas such as (Flesch, 1948). Schwarm and Ostendorf (2005) captured several syntactic features based on parsing trees, whereas Pitler and Nenkova (2008) designed various semantic and discourse features for capturing properties of lexical chains and discourse relations. In addition, the relatively good interpretability of features allows them to be included in tools that help writers simplify a text by analyzing the reading difficulties of the text (François et al., 2020).

Current ARA work relies on distributed representations of texts (i.e. embeddings) (Cha et al., 2017; Filighera et al., 2019) and Deep Learning (DL) (Nadeem and Ostendorf, 2018; Azpiazu and Pera, 2019; Martinc et al., 2021), yielding improve-

[1] Readability should not be confused with Text Simplification that aims to modify a text, making it simpler (Saggion, 2017).

3rd Workshop on Tools and Resources for People with REAding DIfficulties (READI), pages 102–115
20-25 May, 2024May 20th, 2024.

ment over linguistic feature-based systems (e.g., Deutsch et al. (2020); Martinc et al. (2021) for English and Yancey et al. (2021a) for French). Consequently, DL has become the standard in ARA. Contrary to feature-based approaches, the interpretability needs to be improved.

That being said, researchers have been making progress in developing methods to provide explanations for DL models, thus making them more transparent (see Danilevsky et al., 2020; Liang et al., 2021; Sun et al., 2021; Saleem et al., 2022). These methods can provide *global* explanations – i.e., an "overall understanding of deep neural networks model features and each of the learned components such as weights and structures providing" (Liang et al., 2021, 1) – or *local* explanations that try to understand how the model makes a decision based on individual observations. In this paper, we will be concerned with the second class of methods, including saliency maps, explanation generation, probing, and attention scores. Attention scores have been a popular interpretation technique. However, it is subject to some criticisms[2]. Nevertheless, the association between attention head, model's predictions and the linguistic features remains an open question.

In this work, we aim to narrow this gap by identifying if the scores from a attention head in a fine-tuned transformer model for readability are related to ARA features. Our work concentrated on French as a Foreign Language (FFL) readability, using the Common European Framework of Reference for Languages (CEFR) scale (Council of Europe, 2001). Specifically, our objective in this paper is to inspected whether the scores assigned to the tokens by the attention mechanism may relate the ARA features and the CEFR level predictions made by the model. In this work, we focus on the attention mechanism of the transformer model (i.e., self-attention) since it is one of the main keys to the high performance of these models. The main contributions of this work are two. A method for identifying features associated with the prediction made by a model through the attention mechanism. This allows the generation of an explanation of the model's decision from the point of view of linguistic features, which enables a justification of the predicted level to the model's user. The second is contribution consists of the identification that filtering by attention seems to magnify the correlation between feature and text level.

The structure of this paper is as follows. In Section 2, we introduce the standard modeling approach for ARA and discuss related interpretability approaches. Section 3 outlines the features, corpus, and model utilized in this study, accompanied by a detailed description of the proposed method. Our findings, including an analysis of the features related to model's prediction and a feature-based description of model's decision process, are presented in Section 4. Finally, we offer concluding remarks and suggest avenues for future research in Section 5.

[2]See Bibal et al. (2022).

2. Related Work

As this paper combines different research lines, this section first explores the work investigating readability features, identifying informative features for ARA and focusing on those that are explored in this paper (Section 2.1). In Section 2.2, we examine the current literature to predict text readability, focusing on their model's architectures. Finally, in Section 2.3, we discuss frameworks for explaining models.

2.1. Linguistic Features for ARA

There exists a plethora of linguistic features for readability (e.g., 484 are described by Kyle and Crossley (2015), 154 by Chen and Meurers (2016), 380 by Kyle (2016), 16 by Crossley et al. (2016), 400 by Okinina et al. (2020) and 427 by Wilkens et al. (2022)). These may be grouped in different ways. For example, François and Fairon (2012) grouped them by level of information (i.e., lexical, syntactic, semantic and specific) and Wilkens et al. (2022) grouped them by families (e.g., word length, lexical frequency, graded lexicons and lexical norms). From those, our work focuses on lexical norms, lexical frequency and graded lexicons.

Psycho-linguistics explores the relationship between the human mind and language (Field, 2003), where psycho-linguistics norms (or lexical norms) describe how human beings process and understand language and words. These norms are also associated with the reading comprehension of young readers (Crossley et al., 2017; Beinborn et al., 2014), and their scores have been associated with writing quality and development (Sadoski et al., 1995; Crossley et al., 2019; Crossley, 2020). The most commonly explored psycholinguistic norms in readability research are age of acquisition (AoA), subjective frequency (or familiarity), and concreteness (sometimes conflated with imageability).

Age of acquisition refers to the average age at which individuals acquire a particular word in their vocabulary. This norm is related to readability because earlier acquired words tend to be easier to recognize and understand (Juhasz, 2005). As regards subjective frequency, it measures the perceived frequency of words as a result of individual's experience (i.e. reading experience, oral input, etc.). Initially identified by Solomon and

Postman (1952), the familiarity effect explains that more familiar words to a given reader tend to be processed more quickly and accurately (Balota et al., 2004). Gernsbacher (1984) showed that (1) frequency effects coexists with familiarity effects and (2) word familiarity is fairly stable from one individual to another, at least for high and median frequency items, which justified building lists of familiar words. In ARA, texts containing predominantly familiar words are generally easier to read and comprehend. The last lexical norms we focus on is word concretness. Neuroscientists have found that concrete and abstract words are processed differently in the brain, and that concreteness gives an advantage in recognition and recall tasks due to their higher degree of imageability (Jessen et al., 2000; Steacy and Compton, 2019).

Lexical frequency strongly predicts lexical complexity and readability (Rayner and Duffy, 1986). Howes and Solomon (1951) first identified the frequency effect, which was subsequently confirmed by numerous studies in psychology (Monsell, 1991; O'Regan and Jacobs, 1992). This effect corresponds to a more frequent word being recognized more quickly. At the text level, a higher reading speed puts less demand on memory resources, which can be allocated to higher-level processes related to comprehension. This explains why word frequency also indirectly affects the comprehension rate of a text (Crossley et al., 2008).

Finally, commonly used for foreign language teaching, graded lexical resources relate a vocabulary to a proficiency scale, assigning each word of the vocabulary to a given proficiency level, at which the word is considered known by most learners of this level. It can be built based on expert perceptions, such as the reference level descriptors for the CEFR (Beacco et al., 2008; Capel, 2010), or derived from an annotated corpus, as in the CEFRLex project (François et al., 2014). Graded lexicons have been already used in ARA as a way to help readability models to encode readers' expected knowledge (Xia et al., 2016; Yancey et al., 2021a).

2.2. ARA models

Recent literature on ARA has consistently demonstrated the superiority of DL methods over conventional feature engineering approaches. Martinc et al. (2021) compared these methods across multiple manually labeled English and Slovenian corpora, concluding that deep neural networks are effective for both supervised and unsupervised readability prediction tasks. However, they noted that the choice of architecture depends on the dataset. Similarly, Deutsch et al. (2020) evaluated various models including conventional machine learning (ML) methods (e.g., SVMs, Linear Models, Logistic Regression), Convolutional Neural Networks, Transformers, and Hierarchical Attention Networks, and also found that the optimal architecture varies depending on the corpus being tested. However, achieving superior performance with DL models in readability assessment requires fine-tuning the model; otherwise, its performance would be comparable to that of a feature-based model (Imperial, 2021).

Targeting French as foreign language readability, Yancey et al. (2021b) compared linguistic, cognitive and pedagogical features and deep learning models. Despite their efforts, non fine-tuned transformers model (i.e., CamemBERT (Martin et al., 2020)) failed to surpass the baseline model by François and Fairon (2012). However, fine-tuning CamemBERT led to a significant improvement, outperforming the previous state-of-the-art model for French.

2.3. Model Explainability

We begin this section by distinguishing interpretability (or comprehensibility) from explainability, to avoid the confusion existing in the literature (Rudin et al., 2022; Broniatowski et al., 2021). In this work, we follow the definitions outlined by Broniatowski et al. (2021): an *interpretable* model offers only the essential information required to make significant decisions, ensuring that the information provided is justified based on the system's functional objectives, while an *explainable* model elucidates the intricate mechanisms by which a particular implementation produced a specific output, without considering the significance of that output to the decision-maker. Our work thus falls under explainability.

In the context of explainability, Rogers et al. (2020) review several papers investigating how BERT encode linguistic information (e.g, represent phrase-structures (Reif et al., 2019), dependency relations (Jawahar et al., 2019), semantic roles (Kovaleva et al., 2019), and lexical semantics (Garí Soler and Apidianaki, 2020). Most studies on linguistic information in transformers uses the probing (or probing-like) method, thus training a classifier ("probe") to map LLM-states to linguistic target labels (Tenney et al., 2019; Niu et al., 2022). Although this allows inferring the linguistic knowledge of a model, this method does not tell us whether the model actually uses information associated with these features in a given prediction.

Alternatively, Clark et al. (2019) proposed methods to analyze the attention mechanisms of pretrained models. They found that certain attention heads process information in such a way that corresponds well to linguistic notions of syntax and coreference. They also demonstrated that a substantial amount of BERT's attention focuses on a

limited number of tokens (e.g., the special token *[SEP]*). Indeed, the inspection of attention heads and attention weights assigned to words is a common method applied in explanatory visualization systems such as Vig (2019); Braşoveanu and Andonie (2020).

Diving deeper into the specifics of the Transformer architecture, it is important to note that not all attention heads are equally important, and some of them can be pruned with marginal performance degradation (Hao et al., 2021). Moreover, it is unclear what relationship exists between attention weights and model outputs (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019; Bibal et al., 2022). Therefore, the association between attention, prediction and the linguistic properties of the model remains an open question.

The only other existing work that focuses specifically on explainability of readability models, to the best of our knowledge, is Imperial and Ong (2021). Using ELI5[3], they analyzed the weights that classic ML models assign to the features that are part of the model's input vector. The explanation is an interpretation of the features based on their meaning and models' weights.[4]

3. Methodology

Given our goal of identifying how ARA features could explain the predictions of a transformer model fine-tuned for ARA, our starting point is to fine-tune such a model. In this work, we follow the methodology described by Yancey et al. (2021a) for fine-tuning CamemBERT (Martin et al., 2020).[5] Then, we use this model to study the association between ARA features and the tokens on which the model's attention mechanism focuses on. CamemBERT is a model based on the RoBERTa architecture, so it is made up of 12 layers, each with 12 heads of attention. As in all transformers, each attention head uses an attention mechanism to assign weights to the tokens and multiplies these weights by the embeddings of the tokens, thus weighting them. This process is carried out when multiplying *value* by the *softmax* (i.e., a matrix of words by words where values indicate the attention score) in Equation 1. The results of these weightings are concatenated and fed the next layer. The result of this process passes from one layer to the next until, in the last layer, it is sent to an Multi-layer Perceptron (MLP) which performs the classification.

$$attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

The method explored in this paper relies exclusively on linguistic features (see Section 3.1) and on the attention scores that the model assigns to each token. To identify the attention score of each token, we use the attention heads from the last encoding layer since these are the closest to the classification layer. Thus, we obtained 12 attention scores for each token, each one corresponding to a different head from CamemBERT.

It should be noted that the information produced by an attention head is a matrix of tokens by tokens produced by a self-attention mechanism. The values of this attention matrix indicate the weight of attention to be given to all tokens when another is processed. This mechanism is the core element for creating contextual embeddings in the transformer's architecture. Since an attention matrix indicates weights for all tokens, identifying which tokens receive the most attention is an important question. A simple answer would be to use the n biggest values. However, this method always indicates the same number of tokens. As the model may concentrate the attention scores on a few tokens, which often are punctuation marks, we follow Clark et al. (2019) by considering that a token receives significant score attention only if it is greater than the scores assigned to the punctuation marks and special tokens. In this way, we can distinguish the tokens that receive attention from the others for each attention head. For example, given the output of *softmax* illustrated in Figure 1, our method analyzes row by row, selecting the tokens that have an attention score higher than the highest attention score between <*s*>, </*s*> and punctuation. Therefore, for the token *vous*, in the second row, the selected tokens are *vous*, *étudier*, *un*, *pays*, *european* and *pas*. Next, in our method, we annotate the select tokens with linguistic features (see Section 3.1). In this way, given a feature f, we weight the token by the feature value.[6] For example, lets consider f as word length, the tokens selected in the previous example would therefore be $f(vous) = 4$, $f(\text{étudier}) = 7$, $f(un) = 2$, and so on.

[3] https://eli5.readthedocs.io/en/latest/overview.html

[4] The main difference between Imperial and Ong (2021)'s work and ours is the type of model used. While we focus on one type of transformer, Imperial and Ong (2021) focuses on classic ML models.

[5] Note that we explore CamemBERT in this work, but the proposed methodology could be applied in any transformer encoder such as BERT (Devlin et al., 2018) or RoBERTa (Liu et al., 2019).

[6] The annotation process consists of a tokenization normalization step, due to the fact that the tokenizer used by the transformer model is different from the one used by the lexical resources in which the linguistic features are stored.

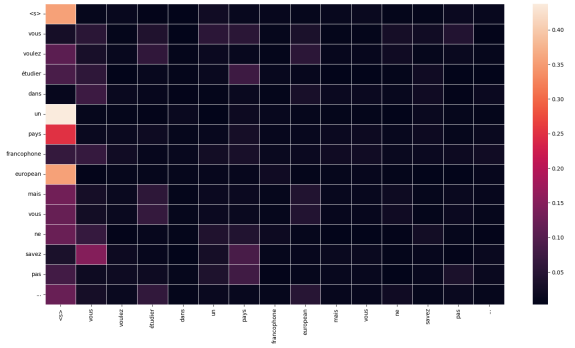

Figure 1: Example of matrix from *softmax*

Then, we use the Equation 2 to calculate the Spearman correlation (ρ) between all tokens that receive attention and the level predicted by the model.[7] More precisely, this correlation is computed based on the predicted CEFR level (l) of a text and the average token score ($score$; see Equation 3), which is the average, for each selected token, of the value of linguistic feature (f) corresponding to the token ($f(token)$) weighted by the attention score assigned to it (α). Similarly, we calculate the correlation for tokens whose attention score were lower than the threshold. In other words, we measure the correlation between the features and the difficulty levels based on the words either considered important to the model or not.

$$\rho = corr(average(score(token)), l) \quad (2)$$

$$score(token) = \alpha(token) \times f(token) \quad (3)$$

As the final step of our analysis, we investigate whether some attention heads tend to specialize towards specific features. We attribute a feature to a specific attention head when the correlation between the feature and the predicted level is higher in the group of tokens selected by the attention threshold than in the group of non-selected tokens.

3.1. Linguistic Features

We explored three families of linguistic features: psycho-linguistic norms, frequency score and graded lexicon. These are widely used in readability studies, as outlined in Section 2. For the annotation of features associated with these families, we used the FABRA toolkit (Wilkens et al., 2022), thus obtaining 19 features:

psycho-linguistic norms: age of acquisition (*AoA*), word *concreteness*, and word *subjective* frequency (also know as subjective word familiarity). These scores are based on (Ferrand et al., 2008; Alario and Ferrand, 1999) for AoA, (Desrochers and Thompson, 2009; Ferrand et al., 2008; Bonin et al., 2003; Desrochers and Bergeron, 2000) for subjective frequency, and (Bonin et al., 2018, 2011; Desrochers and Thompson, 2009; Bonin et al., 2003; Desrochers and Bergeron, 2000) for concreteness.

frequency score: word frequency and word frequency band. The latter identifies to which frequency band each word belongs, based on its rank in a reference frequency list. So, as opposed to the word's frequency, we consider the value of the associated band in this feature (e.g., 1000 for the 1000 most frequent words and 2000 for words with a frequency between 1000 and 2000). Since this feature could also be considered as a proportion of words belonging to a frequency band, we chose to use this feature in two ways: the value of the frequency band and the proportion of a band in the text. For the latter, the proportion of each band is named *freq. band*$_{\text{"band value"}}$ (e.g., *freq. band*$_{1000}$).

graded lexicon: proportion of words at one of the 6 CEFR levels (between A1 to C2). These features are named *word level*$_{\text{"CEFR level"}}$. In this work, we use FLELex (François et al., 2014) are reference for the expected CEFR level of a word.

3.2. Corpus

A common way to build readability corpora is to collect textbooks and label each extracted text with the level of the textbook it comes from (e.g., Sato et al. (2008); Volodina et al. (2014)). In this work, we focus on French as a Foreign Language readability, using the CEFR scale (Council of Europe, 2001), which includes six levels: A1 (Breakthrough); A2 (Waystage); B1 (Threshold); B2 (Vantage); C1 (Effective Operational Proficiency) and C2 (Mastery). We used the same corpus as Yancey et al. (2021a), which is composed of 2.734 texts with a balanced distribution of texts in each of the target levels, as described in Table 1.

This corpus is build upon pedagogical materials published after 2001 indicate which CEFR level they are intended for. It was originally proposed by François and Fairon (2012) who creates an initial version of 1.793 texts. Later, Yancey et al. (2021a) expanded their collection into a larger and more diverse corpus extracted from 47 FFL textbooks published between 2001 and 2018. In this corpus, the level of a text is the level indicated in the textbook it was extracted from; with the exception of the C1

[7] We used the level predicted by the model because, in this study, we aim to explain the readability model and not the readability phenomenon.

and C2 levels that the authors have grouped into a single level.

Target	Texts	Words
A1	572	60,022
A2	574	83,294
B1	580	119,048
B2	442	130,877
C1 and C2	566	198,517
Total	2734	591,758

Table 1: Description of the corpus compiled by Yancey et al. (2021a)

4. Results

The first result to report in this paper is the performance of the readability prediction model. After training, the fine-tuned model achieved an accuracy of 0.57 and an F-score of 0.54 (0.74 for level Al, 0.53 fro A2, 0.48 for B1, 0.26 for B2, and 0.72 for C), estimated with a five-fold cross-validation. These results are similar to those reported by Yancey et al. (2021a). As the model is not the focus of this work, we are looking for a model close to the state of the art in terms of architecture and performance. This being achieved, this model can serve as the cornerstone for the results reported in the rest of this section.

4.1. Discrimination power

Before we start to study the applicability of the feature to explain the model, we assess their discrimination power. So, we computed the correlation between each of the 20 features studied and the target levels, as is usually done in ARA studies. Although these values are not connected with our model, they will serve as a reference. As can be seen at column "true label (0)" in Table 2, we found correlations ranging from -0.65 (*word level*$_{A1}$) to 0.55 (*word level*$_{C1}$) when relating the true readability level with the average feature value of all tokens in a text. These correlations confirms that some of our features are good predictor of the CEFR level of a text. In addition, in column "pred (1)", we also calculate the correlation between the predicted readability level with the average feature value, since our ultimate goal is to identify whether the model might be explained by the features. We observe tiny increases when comparing these correlations, which suggest that the approximation made by the model is closer from these features than these features are from the real readability level.

The model explainability analysis starts by considering the relationship between the features and the model's predictions. This is done without distinguishing the attention heads, meaning that we calculate the attention for each head, but we do not differentiate which head generates the association. We calculated the correlation between the level predicted by the model and each feature, but, this time, we removed the words that had a small attention score (see Section 3). These values can be seen in the selected words column (2) of Table 2, and the absolute difference between these correlations and the original correlations is in column "(1) - (2)". The latter shows an increase in correlation for all the features, except for *word level*$_{A1}$, which had a decrease of 0.23 in its correlation with the predicted level. This already allow us to identify that attention scores acts as a sort of filter that magnifies the correlation between ARA features and predictions, possibly by removing noise (i.e., word embeddings unnecessary for the classification).

Although this analysis already reveals an association between the features and the predictions, it does not indicate how the model measures the features (as they are not provided to the model). We, therefore, explored an alternative version of the correlation between the predicted levels and the values of the features in the list of selected words. In this version, we weighted the features' values by the attention score assigned by the model. These results are shown in column "selected words weighted by attention (3)" of Table 2. As can be seen, the weight of attention does not affect the intensity of the correlation for most of the features[8], except *AoA* (increase of 0.16 points), *concreteness* (0.24), *subjective frequency* (0.31) and *frequency band* (0.08). We therefore observed that the attention-based word filter has a greater impact than the combination of attention weights.

In order to complement the analysis of the correlation between the features and the readability levels, we also analyzed the impact of the predictive capacity of a simple machine learning model to identify the readability level of the text using only the words selected by the attention filter. The goal of this analysis is to identify how the reduction in the text length caused by the proposed filter would affect the performance of a classification model based purely on linguistic features. For that end, we compared the performance of Random Forest classifiers trained using all tokens in the document with RF classifiers using only the tokens selected by the proposed filter. Moreover, we also assess the impact of training the RF classifiers on the true labels and the transformer predictions. This allowed us to further confirm the relation existing between the linguistic variables and the predictions of transformer that are not leveraging any of these

[8] Absolute value of column "(2) - (3)" ≤ 0.05.

Features		Correlation				Difference		
		entire corpus		selected words				
		true label (0)	pred (1)	(2)	wgt att (3)	(0) - (1)	(1) - (2)	(2) - (3)
AoA		0.31	0.33	0.36	-0.52	0.02	0.03	0.16
Concreteness		-0.31	-0.34	-0.39	-0.63	0.03	0.05	0.24
Subjective F.		-0.15	-0.17	-0.27	-0.58	0.02	0.10	0.31
Word Freq.		0.23	0.26	0.39	-0.34	0.03	0.13	-0.05
Freq. Band		0.34	0.37	0.39	-0.47	0.03	0.02	0.08
Freq. Band	1000	-0.40	-0.45	0.47	0.45	0.05	0.02	-0.02
	2000	0.26	0.31	0.54	0.58	0.05	0.23	0.04
	3000	0.18	0.20	0.54	0.53	0.02	0.34	-0.01
	4000	0.24	0.28	0.55	0.53	0.04	0.27	-0.02
	5000	0.15	0.16	0.53	-0.05	0.01	0.37	-0.05
	6000	0.20	0.21	0.51	0.46	0.01	0.30	-0.05
	7000	0.24	0.24	0.51	0.46	0.00	0.27	-0.05
	8000	0.27	0.27	0.50	0.45	0.00	0.23	-0.05
Word Level	A1	-0.65	-0.73	0.41	0.46	0.08	-0.32	0.05
	A2	0.25	0.28	0.54	0.51	0.03	0.26	-0.03
	B1	0.27	0.32	0.58	0.58	0.05	0.26	0.00
	B2	0.16	0.17	0.51	0.45	0.01	0.34	-0.06
	C1	0.55	0.60	0.66	0.70	0.05	0.06	0.04
	C2	0.38	0.44	0.63	0.63	0.06	0.19	0.00

Table 2: Correlation between features and CEFR target levels of documents. The last two columns indicate the absolute difference between the correlations of the other three columns.

Target	Attention Filter	F1	Acc
true label	no	0.43	0.45
true label	yes	0.41	0.43
prediction	no	0.48	0.51
prediction	yes	0.47	0.51

Table 3: The ability of a feature to predict the target

features.

As can be seen in Table 3, the result of the predictive capacity shows a reduction of 0.02 of F1 and 0.01 of accuracy when using the word filter for predicting the document readability level and 0.01 of F1 and accuracy when predicting the transformer predictions. These results point out that the reduction of a considerable part of the words in the documents does not strongly impact the model's performance, suggesting that the filter is removing possible duplicated or unnecessary words. In other words, the filter allows us to train models with similar performance with less input. However, it is essential to note that this experiment aims to assess whether the selected words can still be used for the task, not to propose an explanation of the transformer model.

4.2. Features and Attention heads

Moving on in our study, we compared the attention head level. This analysis found that psycholinguistic features tend to be associated with the same attention heads. Similarly, the features related to frequency tend to be grouped in the same way. Following the same behavior but with fewer associated heads, the graded lexicon features tend to be found in the same attention heads.

4.2.1. Base Method

The association between attention heads and features is shown in Table 4. In this table, we can see that several heads are related to at least one feature of the three families of features. However, some heads are associated with several features from the same family. Furthermore, some of them are associated with more than one family. For example, *Head 5* is associated with *psycho-linguistic* and *frequency* features, *Head 9* with *graded lexicon* and *frequency* features, and *Head 7* is associated with all three groups of features. Considering the perspective of features, the psycho-linguistics features are related with, on average, 6.5 attention heads, 2.8 for frequency features, and 2.5 for graded lexicon. In addition, *psycho-linguistics* features are also associated with *Head 4*, *7* and *10*, and the *frequency* features are also associated with *Head 2* and *3*. In general, these results are in line with those of Clark et al. (2019), where it was identified that only a few heads are related to the model's decision.

	Psycholinguistic	Frequency	Graded lexicon	Count
Head 1	-	-	-	0
Head 2	subj.Freq. (-0.49)	freq. $band_{6000}$ (0.44) freq. $band_{8000}$ (0.45)	**word $level_{B2}$** (0.45)	6
Head 3	subj.Freq. (-0.51)	freq. $band_{6000}$ (0.43) freq. $band_{8000}$ (0.44)	-	5
Head 4	**aoa** (-0.49) **concreteness** (-0.58) **subj.Freq.** (-0.54)	freq. $band_{2000}$ (0.58)	word $level_{C1}$ (0.7)	6
Head 5	**aoa** (-0.52) **concreteness** (-0.58) **subj.Freq.** (-0.54)	**freq. band** (-0.42) **wordFreq** (-0.34) freq. $band_{2000}$ (0.56)	word $level_{C1}$ (0.69)	9
Head 6	subj.Freq. (-0.52)	freq. $band_{3000}$ (0.53)	word $level_{C1}$ (0.7)	4
Head 7	aoa (-0.51) **concreteness** (-0.57) **subj.Freq.** (-0.56)	**freq. band** (-0.47) freq. $band_{1000}$ (0.45) freq. $band_{3000}$ (0.51) freq. $band_{5000}$ (0.48) **freq. $band_{6000}$** (0.46) freq. $band_{8000}$ (0.45)	word $level_{B2}$ (0.45) word $level_{C2}$ (0.63)	14
Head 8	-	freq. band (-0.43) **freq. $band_{6000}$** (0.46)	**word $level_{A1}$** (0.46) word $level_{B2}$ (0.44)	6
Head 9	concreteness (-0.54) **subj.Freq.** (-0.53)	**freq. $band_{3000}$** (0.53)	word $level_{B1}$ (0.58) word $level_{C1}$ (0.69)	6
Head 10	**aoa** (-0.51) **concreteness** (-0.63) **subj.Freq.** (-0.58)	freq. $band_{2000}$ (0.58)	word $level_{C1}$ (0.69)	6
Head 11	**aoa** (-0.51) **concreteness** (-0.59) **subj.Freq.** (-0.53)	freq. $band_{2000}$ (0.57)	-	5
Head 12	-	-	-	0

Table 4: Association between attention heads and features. The number in brackets indicates the correlation between the predicted CEFR level and feature weighted by attention score for each attention head. Items in bold are those selected with a threshold of 0.02.

4.2.2. Acceptance threshold

The results we have presented so far rely on the assumption that a feature is related to an attention head if the correlation between the feature and the level predicted is higher in the group of words selected based on attention scores. In order to better understand the method explored in this paper, we relaxed this assumption. To do this, we defined a simple acceptance threshold based on the difference in correlation between the groups of words (selected v. non-selected). When this threshold is set to zero, the results described above in this section (with 67 associations between features and heads) are obtained, while no association is observed when it is set to 0.14. The other values explored in this threshold show 53 heads selected for 0.01, 35 for 0.02, 25 for 0.03, 19 for 0.04, 19 for 0.05, 16 for 0.06, 14 for 0,07, 8 for 0,08, 6 for 0,09, 4 for 0,10, 2 for 0,11, 2 for 0,12, and 1 for 0,13. This trend towards a reduction in the method's selectivity should be considered in light of the range of the correlation values. These have an average value of 0.43. Thus, the 0.1 limit range explored accounts for 23% of the correlation range available for exploration. Taking a closer look at the distribution of the distance between the absolute correlation values of selected and non-selected words, we see a median of 0.05 (variance of 0.004, Q1 of 0.02, Q3 of 0.09 and max of 0.32).

4.2.3. Base Method with Acceptance threshold

We therefore revisited the association between the attention heads and the features, setting a threshold of 0.02. These values are in bold in Table 4.

The application of the threshold allows us to see a clearer picture of the data. It can be seen that *psycho-linguistic* family is the one most associated with the attention heads, contrary to the previous perspective marked by a similar presence of all types of features. In fact, *psycholinguistic* features are most related with 6 heads (*Heads 4*, *5*, *7*, *9*, *10* and *11*). Surprisingly, the features of family *graded lexicon*, which represent features most associated with the task the model was fine-tuned for, were not associated with most of the heads. They were

only associated with *Heads 2* and *8*. For *Head 2*, the feature identified was *word level*$_{B2}$, which had the lowest correlation with the corpus of features in its family. Finally, the *frequency* family, previously the most relevant feature, now is associated with 4 heads. However, it only has few relevant features per head, in contrast to family *psycho-linguistic* where there are several features associated with the same head. In this family, the most relevant features were *Frequency Band*$_{6000}$, which indicates words of medium complexity, and *Frequency Band*$_{3000}$ which indicates easy words.

5. Conclusion

The field of ARA has evolved a lot recently due to recent advances in NLP: it has shifted from models based on theoretically-grounded linguistic features to more accurate black-box DL models. As a consequence, the relationship between readability models and the literature about the cognitive processes involved in reading has been weakened. Thus, it could be possible for a model to identify the expected level of a text, but for the wrong reasons.

Aiming to narrow the gap opened by the widespread use of black-box models, we proposed a method to investigate whether the transformer architecture, when fine-tuned on the readability task, is sensitive to word characteristics that traditional readability features were capturing. We also explore whether attention heads might get specialized to some ARA features. For that, we correlated the level of the predictions made by the model with the ARA features on tokens to which the model is paying attention.

In our finding, we identified that the filtering of word information by the attention layer seems to magnify the correlation between features and the predicted text level. In addition, we were able to identify that attention heads are more sensitive to some linguistic features than others, and describe those which are associated to most of the ARA features explored in this work. Despite being able to identify a relationship between attention heads and linguistic features, these do not explain 100% of the model's behavior as well as the ARA features cannot fully explain the readability level in the corpus. This might indicate that the method is not capable of indicating the feature precisely, but rather something more interesting: the readability effect that the feature seeks to approximate.

As future work, we foreseen the extension of the proposed method to include other than lexical features, such as grammatical or discursive properties. We could also reproduce the analysis to the other layers of the transformer, as it is expected than some layers might be more sensitive to some kind of information than others. Finally, it would be necessary to assess our method on other corpora and using more diverse transformer architectures in order to assess its robustness.

6. Acknowledgements

Part of this research is supported by the European Commission (Project: iRead4Skills, Grant number: 1010094837. Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. Computational resources have been provided by the Consortium des Équipements de Calcul Intensif (CÉCI), funded by the Fonds de la Recherche Scientifique de Belgique (F.R.S.-FNRS) under Grant No. 2.5020.11 and by the Walloon Region.

7. Bibliographical References

Ion Madrazo Azpiazu and Maria Soledad Pera. 2019. Multiattentive recurrent neural network architecture for multilingual readability assessment. *Transactions of the Association for Computational Linguistics*, 7:421–436.

David A Balota, Michael J Cortese, Susan D Sergent-Marshall, Daniel H Spieler, and Melvin J Yap. 2004. Visual word recognition of single-syllable words. *Journal of experimental psychology: General*, 133(2):283.

J.-C. Beacco, S. Lepage, R. Porquier, and P. Riba. 2008. *Niveau A2 pour le français: Un référentiel*. Didier.

Lisa Beinborn, Torsten Zesch, and Iryna Gurevych. 2014. Readability for foreign language learning: The importance of cognates. *ITL-International Journal of Applied Linguistics*, 165(2):136–162.

Adrien Bibal, Rémi Cardon, David Alfter, Rodrigo Wilkens, Xiaoou Wang, Thomas François, and Patrick Watrin. 2022. Is attention explanation? an introduction to the debate. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3889–3900.

J.R. Bormuth. 1966. Readability: A new approach. *Reading research quarterly*, 1(3):79–132.

Adrian MP Braşoveanu and Răzvan Andonie. 2020. Visualizing transformers for nlp: a brief survey. In *2020 24th International Conference*

Information Visualisation (IV), pages 270–279. IEEE.

David A Broniatowski et al. 2021. Psychological foundations of explainability and interpretability in artificial intelligence. *NIST, Tech. Rep*.

A. Capel. 2010. A1-b2 vocabulary: Insights and issues arising from the english profile wordlists project. *English Profile Journal*, 1(1):1–11.

M. Cha, Y. Gwon, and H.T. Kung. 2017. Language modeling by clustering with word embeddings for text readability assessment. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 2003–2006. ACM.

J.S. Chall and E. Dale. 1995. *Readability Revisited: The New Dale-Chall Readability Formula*. Brookline Books, Cambridge.

Xiaobin Chen and Detmar Meurers. 2016. Ctap: A web-based tool supporting automatic complexity analysis. In *Proceedings of the workshop on computational linguistics for linguistic complexity (CL4LC)*, pages 113–119.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*.

M. Coleman and T.L. Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283–284.

K. Collins-Thompson and J. Callan. 2005. Predicting reading difficulty with statistical language models. *Journal of the American Society for Information Science and Technology*, 56(13):1448–1462.

Kevyn Collins-Thompson. 2014. Computational assessment of text readability: A survey of current and future research. *International Journal of Applied Linguistics*, 165(2):97–135.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Press Syndicate of the University of Cambridge.

S. Crossley, J. Greenfield, and D. McNamara. 2008. Assessing text readability using cognitively based indices. *TESOL Quarterly*, 42(3):475–493.

Scott Crossley. 2020. Linguistic features in writing quality and development: An overview. *Journal of Writing Research*, 11(3).

Scott A Crossley, Kristopher Kyle, and Danielle S McNamara. 2016. The tool for the automatic analysis of text cohesion (taaco): Automatic assessment of local, global, and text cohesion. *Behavior research methods*, 48(4):1227–1237.

Scott A Crossley and Danielle S McNamara. 2012. Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading*, 35(2):115–135.

Scott A Crossley, Stephen Skalicky, and Mihai Dascalu. 2019. Moving beyond classic readability formulas: New methods and new models. *Journal of Research in Reading*, 42(3-4):541–561.

Scott A Crossley, Stephen Skalicky, Mihai Dascalu, Danielle S McNamara, and Kristopher Kyle. 2017. Predicting text comprehension, processing, and familiarity in adult readers: New approaches to readability formulas. *Discourse Processes*, 54(5-6):340–359.

E. Dale and J.S. Chall. 1948. A formula for predicting readability. *Educational research bulletin*, 27(1):11–28.

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A survey of the state of explainable ai for natural language processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 447–459.

Tovly Deutsch, Masoud Jasbi, and Stuart Shieber. 2020. Linguistic features for readability assessment. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–17, Seattle, WA, USA → Online. Association for Computational Linguistics.

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

John Field. 2003. *Psycholinguistics: A resource book for students*. Psychology Press.

Anna Filighera, Tim Steuer, and Christoph Rensing. 2019. Automatic text difficulty estimation using embeddings and neural networks. In *European Conference on Technology Enhanced Learning*, pages 335–348. Springer.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

T. François. 2015. When readability meets computational linguistics: a new paradigm in readability. *Revue française de linguistique appliquée*, 20(2):79–97.

T. François, N. Gala, P. Watrin, and C. Fairon. 2014. FLELex: a graded lexical resource for French foreign learners. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 3766–3773.

T. François and E. Miltsakaki. 2012. Do NLP and machine learning improve traditional readability formulas? In *Proceedings of the 2012 Workshop on Predicting and improving text readability for target reader populations (PITR2012)*.

Thomas François. 2011. *Les apports du traitement automatique du langage à la lisibilité du français langue étrangère*. Ph.D. thesis, Ph. D. thesis, Université Catholique de Louvain. Thesis Supervisors: Cédrick ….

Thomas François and Cédrick Fairon. 2012. An "ai readability" formula for french as a foreign language. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 466–477.

Thomas François, Adeline Müller, Eva Rolin, and Magali Norré. 2020. Amesure: a web platform to assist the clear writing of administrative texts. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 1–7.

Daniela B Friedman and Laurie Hoffman-Goetz. 2006. A systematic review of readability and comprehension instruments used for print and web-based cancer information. *Health Education & Behavior*, 33(3):352–373.

Aina Garí Soler and Marianna Apidianaki. 2020. BERT knows Punta Cana is not just beautiful, it's gorgeous: Ranking scalar adjectives with contextualised representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7371–7385, Online. Association for Computational Linguistics.

M.A. Gernsbacher. 1984. Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology: General*, 113(2):256–281.

Yaru Hao, Li Dong, Furu Wei, and Ke Xu. 2021. Self-attention attribution: Interpreting information interactions inside transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12963–12971.

D. Howes and R. Solomon. 1951. Visual duration threshold as a function of word probability. *Journal of Experimental Psychology*, 41(40):1–4.

Joseph Marvin Imperial. 2021. Knowledge-rich bert embeddings for readability assessment. *arXiv preprint arXiv:2106.07935*.

Joseph Marvin Imperial and Ethel Ong. 2021. Under the microscope: Interpreting readability assessment models for filipino. *arXiv preprint arXiv:2110.00157*.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. *arXiv preprint arXiv:1902.10186*.

Ludivine Javourey-Drevet, Stéphane Dufau, Thomas François, Núria Gala, Jacques Ginestié, and Johannes C Ziegler. 2022. Simplification of literary and scientific texts to improve reading fluency and comprehension in beginning readers of french. *Applied Psycholinguistics*, 43(2):485–512.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

F. Jessen, R. Heun, M. Erb, D.-O. Granath, U. Klose, A. Papassotiropoulos, and W. Grodd. 2000. The concreteness effect: Evidence for dual coding and context availability. *Brain and Language*, 74:103–112.

Barbara J Juhasz. 2005. Age-of-acquisition effects in word and picture identification. *Psychological bulletin*, 131(5):684.

J. Kimble. 1992. Plain english: A charter for clear writing. *TM Cooley L. Rev.*, 9:1.

W. Kintsch and D. Vipond. 1979. Reading comprehension and readability in educational practice and psychological theory. In L.G. Nilsson, editor, *Perspectives on Memory Research*, pages 329–365. Lawrence Erlbaum, Hillsdale, NJ.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of BERT. 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 4365–4374, Hong Kong, China. Association for Computational Linguistics.

Kristopher Kyle. 2016. *Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication*. Ph.D. thesis, Georgia State University, Atlanta, Georgia.

Kristopher Kyle and Scott A Crossley. 2015. Automatically assessing lexical sophistication: Indices, tools, findings, and application. *Tesol Quarterly*, 49(4):757–786.

Yu Liang, Siguang Li, Chungang Yan, Maozhen Li, and Changjun Jiang. 2021. Explaining the black-box model: A survey of local interpretation methods for deep neural networks. *Neurocomputing*, 419:168–182.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

B.A. Lively and S.L. Pressey. 1923. A method for measuring the "vocabulary burden" of textbooks. *Educational Administration and Supervision*, 9:389–398.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Laurent Romary, Éric Villemonte de La Clergerie, Djamé Seddah, Benoît Sagot, et al. 2020. Camembert: a tasty french language model. In *ACL 2020-58th Annual Meeting of the Association for Computational Linguistics*.

Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. 2021. Supervised and unsupervised neural approaches to text readability. *Computational Linguistics*, 47(1):141–179.

S. Monsell. 1991. The nature and locus of word frequency effects in reading. In D. Besner and G. Humphreys, editors, *Basic processes in reading : Visual word recognition*, pages 148–197. Lawrence Erlbaum Associates Inc., Hillsdale, NJ.

Farah Nadeem and Mari Ostendorf. 2018. Estimating linguistic complexity for science texts. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, pages 45–55.

Jingcheng Niu, Wenjie Lu, and Gerald Penn. 2022. Does bert rediscover a classical nlp pipeline? In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3143–3153.

Nadezda Okinina, Jennifer-Carmen Frey, and Zarah Weiss. 2020. Ctap for italian: Integrating components for the analysis of italian into a multilingual linguistic complexity analysis tool. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 7123–7131.

J. O'Regan and A. Jacobs. 1992. Optimal viewing position effect in word recognition : A challenge to current theory. *Journal of Experimental Psychology : Human Perception and Performance*, 18(1):185–197.

Maria Soledad Pera and Yiu-Kai Ng. 2014. Automating readers' advisory to make book recommendations for k-12 readers. In *Proceedings of the 8th ACM Conference on Recommender Systems*, pages 9–16.

E. Pitler and A. Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 186–195.

Keith Rayner and Susan A Duffy. 1986. Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & cognition*, 14(3):191–201.

Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B Viegas, Andy Coenen, Adam Pearce, and Been Kim. 2019. Visualizing and measuring the geometry of bert. *Advances in Neural Information Processing Systems*, 32.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. 2022. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistic Surveys*, 16:1–85.

Mark Sadoski, Ernest T Goetz, and Enrique Avila. 1995. Concreteness effects in text recall: Dual coding or context availability? *Reading Research Quarterly*, pages 278–288.

Horacio Saggion. 2017. Automatic text simplification. *Synthesis Lectures on Human Language Technologies*, 10(1):1–137.

Rabia Saleem, Bo Yuan, Fatih Kurugollu, Ashiq Anjum, and Lu Liu. 2022. Explaining deep neural networks: A survey on the global interpretation methods. *Neurocomputing*.

Antony Sare, Aesha Patel, Pankti Kothari, Abhishek Kumar, Nitin Patel, and Pratik A Shukla. 2020. Readability assessment of internet-based patient education materials related to treatment options for benign prostatic hyperplasia. *Academic Radiology*, 27(11):1549–1554.

Satoshi Sato, Suguru Matsuyoshi, and Yohsuke Kondoh. 2008. Automatic Assessment of Japanese Text Readability Based on a Textbook Corpus. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, Marrakech, Morocco.

Andreas Schleicher. 2019. Pisa 2018: Insights and interpretations. *OECD Publishing*.

Andreas Schleicher. 2022. How the european schools compare internationally pisa for schools 2022. *OECD Publishing*.

Sarah E Schwarm and Mari Ostendorf. 2005. Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 523–530.

E.A. Smith and R.J. Senter. 1967. Automated Readability Index. Technical report, AMRL-TR-66-220, Aerospace Medical Research Laboratories, Wright-Patterson Airforce Base, OH.

R.L. Solomon and L. Postman. 1952. Frequency of usage as a determinant of recognition thresholds for words. *Journal of Experimental Psychology*, 43(3):195–201.

L.M. Steacy and D.L. Compton. 2019. Examining the role of imageability and regularity in word reading accuracy and learning efficiency among first and second graders at risk for reading disabilities. *Journal of Experimental Child Psychology*, 178:226–250.

Xiaofei Sun, Diyi Yang, Xiaoya Li, Tianwei Zhang, Yuxian Meng, Han Qiu, Guoyin Wang, Eduard Hovy, and Jiwei Li. 2021. Interpreting deep learning models in natural language processing: A review. *arXiv preprint arXiv:2110.10470*.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline. *arXiv preprint arXiv:1905.05950*.

Sowmya Vajjala. 2021. Trends, limitations and open challenges in automatic readability assessment research. *arXiv preprint arXiv:2105.00973*.

Jesse Vig. 2019. A multiscale visualization of attention in the transformer model. *arXiv preprint arXiv:1906.05714*.

Elena Volodina, Ildikó Pilán, Stian Rødven Eide, and Hannes Heidarsson. 2014. You get what you annotate: a pedagogically annotated corpus of coursebooks for Swedish as a Second Language. In *Proceedings of the third workshop on NLP for computer-assisted language learning*, pages 128–144.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. *arXiv preprint arXiv:1908.04626*.

Rodrigo Wilkens, David Alfter, Xiaoou Wang, Alice Pintard, Anaïs Tack, Kevin P Yancey, and Thomas François. 2022. Fabra: French aggregator-based readability assessment toolkit. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1217–1233.

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. Text readability assessment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22.

Kevin Yancey, Alice Pintard, and Thomas Francois. 2021a. Investigating readability of french as a foreign language with deep learning and cognitive and pedagogical features. *Lingue e Linguaggio*, 2021(2):229–258.

Kevin Yancey, Alice Pintard, and Thomas Francois. 2021b. Investigating readability of french as a foreign language with deep learning and cognitive and pedagogical features. *Lingue e linguaggio*, 20(2):229–258.

8. Language Resource References

Alario, F-Xavier and Ferrand, Ludovic. 1999. *A set of 400 pictures standardized for French: Norms for name agreement, image agreement, familiarity, visual complexity, image variability, and age of acquisition*. Springer.

Bonin, Patrick and Méot, Alain and Aubert, Louis-F and Malardier, Nathalie and Niedenthal, Paula and Capelle-Toczek, M-C. 2003. *Normes de concrétude, de valeur d'imagerie, de fréquence subjective et de valence émotionnelle pour 866 mots*. Persée-Portail des revues scientifiques en SHS.

Bonin, Patrick and Méot, Alain and Bugaiska, Aurélia. 2018. *Concreteness norms for 1,659 French words: Relationships with other psycholinguistic variables and word recognition times*. Springer.

Bonin, Patrick and Méot, Alain and Ferrand, Ludovic and Roux, Sébastien. 2011. *L'imageabilité: normes et relations avec d'autres variables psycholinguistiques*. NecPlus.

Desrochers, Alain and Bergeron, Mylène. 2000. *Valeurs de fréquence subjective et d'imagerie pour un échantillon de 1,916 substantifs de la langue franĉaise.* Canadian Psychological Association.

Desrochers, Alain and Thompson, Glenn L. 2009. *Subjective frequency and imageability ratings for 3,600 French nouns*. Springer.

Ferrand, Ludovic and Bonin, Patrick and Méot, Alain and Augustinova, Maria and New, Boris and Pallier, Christophe and Brysbaert, Marc. 2008. *Age-of-acquisition and subjective frequency estimates for all generally known monosyllabic French words and their relation with other psycholinguistic variables*. Springer.

François, T. and Fairon, C. 2012. *An "AI readability" formula for French as a foreign language*.

François, Thomas and Gala, Núria and Watrin, Patrick and Fairon, Cédrick. 2014. *FLELex: a graded lexical resource for French foreign learners.*