Applied Human Factors and Ergonomics (AHFE2025), Vol. 199, 2025, 2166–2175
https://doi.org/10.54941/ahfe1007028

Combining Large Language Models With Linguistic Features for the Readability Complexity Assessment of Texts

Diego Palma and Christian Soto

University of Concepción, Concepción, Chile

ABSTRACT

Readability is central to learning, yet traditional text complexity measures based on surface features such as word frequency and sentence length fail to capture deeper dimensions of cohesion, coherence, and reader knowledge. This paper proposes a hybrid framework for text complexity assessment that integrates novel linguistic features with large language models (LLMs). We introduce discourse and semantic based features that approximate cohesion and coherence through lexical distribution, segmentation heuristics, and sentence embeddings. To account for background knowledge, we incorporate a fine-tuned LLM as an external assessor, yielding a hybrid model that combines structural and semantic features with knowledge-sensitive judgments. We further contribute a new corpus of 656 Spanish educational texts, annotated with grade-level labels. Experiments compare three approaches: a fine-tuned LLM, a model using only linguistic features, and the hybrid model. Results show that the LLM alone performs poorly (accuracy = 0.18), the linguistic model achieves higher accuracy (0.61), and the hybrid model outperforms both (0.75). Feature analysis highlights the predictive value of measures such as KL divergence, lexical diversity, semantic distances, and givenness. This work advances readability assessment by demonstrating that hybrid approaches more accurately and theoretically capture text complexity, bridging computational linguistics with educational practice.

Keywords: Natural language processing, Intelligent systems, NLP, Artificial intelligence, Computational linguistics, Large language models, LLM

INTRODUCTION

Readability is an important skill for students, as most of the knowledge that they acquire over time, is based on text. Texts usually present ideas with the goal of being understandable so the readers can make further interpretations and deductions. Moreover, assessments over a person's life are mostly presented as tests based on texts, even in oral assessments, the material from which the person prepares for such an assessment is based on text. There are different types of texts depending on the context and the target reader, and thus, different text complexities.

Text complexity plays an important role, as can be the difference in how the target reader digests the information from it. For a text to be readable, there are multiple dimensions, and all of them can be condensed into cohesion

and coherence. However, evaluating how cohesive and coherent a text is is a subjective task. For example, two different readers might assess a text differently based on complexity. To approximate cohesion and coherence, multiple works have been done in the literature. A simple example would be lexical diversity, that could be approximated as the ratio between the unique words used in a text and the total words of the text. However, surface proxies for cohesion and coherence are not sufficient. There are other dimensions that play a role in coherent text, such as the semantic relationship between the text utterances. As an example, the text "Diego went to the pharmacy and got his prescriptions" would be more coherent than "Diego went to the pharmacy and got his plane". Naturally, there is also a context component that needs to be considered.

In education contexts, the texts are usually self-contained, as the goal is to be able to improve readability skills of students. Therefore, the text will provide the context needed to understand it, and thus, all the features that make the text readable are contained within the text. Nevertheless, there is common knowledge that also plays a role when reading a text, and this common knowledge is embedded in the reader and not the text. Thus, we also need an approach to capture this variable, at least in an approximated fashion. In the literature, a plethora of cohesion and coherence have been studied. However, there is still a missing component on how to connect all the pieces to make a text readable.

In this work we propose a computational linguistics approach to measure text complexity. The contributions of this work is threefold:

We developed a new set of linguistic features for text complexity assessment.
To capture common knowledge we developed a hybrid approach that combines an LLM model and linguistic features for text complexity classification.
For reproducibility purposes, we also developed a CORPUS of texts and their level of complexity.

We show that the combination of our new features with an LLM, outperforms both the LLM alone and a model based on linguistic traits.

RELATED WORK

A text is an autonomous linguistic product designed to fulfill a communicative intent and composed of smaller units (Albaladejo, 1983). Consequently, a text possesses an internal structure that gives it meaning, enabling readers to interpret it as a coherent whole. The structure of a text is not limited to the superficial organization of words and sentences but also encompasses deeper relationships among its components (Van Dijk, 1985), which ensure coherence and comprehensibility for the reader. In this regard, there are various levels that constitute the structure of a text, which support each other. The microstructure pertains to the internal organization of sentences and paragraphs. The macrostructure involves the arrangement of primary and secondary ideas on a global scale, forming a coherent

message. Finally, the superstructure refers to the conventions and typical organizational forms of different text types (Van Dijk, 1997).

From the reader's perspective, there have been proposals of the existence of various levels of representation in the construction of textual meaning (Kintsch, 1988). The surface linguistic representation retains the words and grammatical relationships used by the author, emphasizing the lexical and syntactic elements in the text (Kintsch, 1998). The text base represents the semantic relationships between ideas, both locally and globally. Lastly, the situation model integrates the prior linguistic representations with the reader's knowledge. As a result, text comprehension produces not a purely linguistic entity but a mental representation evoked by a particular individual based on the text's content. Thus, the aforementioned theories converge on the idea that texts must be analysed from a multidimensional or multilevel perspective.

The study of linguistic features in texts has been developing for decades, establishing links to writing quality, development, and genre appropriateness. Early research employed manual analysis methods, such as rubrics and element coding matrices. While these approaches provided valuable insights into the interaction between linguistic features and various variables, they were impractical, difficult to replicate, and prone to human error (Crossley, 2020). Advances in computational linguistics and discourse processing have fortunately enabled the automation and enhancement of text processing mechanisms. In recent years, technologies have emerged to analyse texts using computational linguistic indices, bringing systematicity and objectivity to the field. Notable examples include Coh-Metrix for English (Graesser, 2011) and TRUNAJOD for Spanish (Palma, 2021).

Coh-Metrix analyses texts using over 200 measures of language, text, and readability, examining more than 50 types of cohesion relationships (Graesser, 2011). It generates numerous linguistic indices that provide information on lexical, syntactic, semantic, phonological, and cohesion variables. Coh-Metrix has been used in studies to detect significant differences between spoken and written English samples (Louwerse, 2004), identify individual authorship of texts (McCarthy, 2006), and examine linguistic traits characterizing good and poor writing (McNamara, 2001). It has also served as the foundation for developing tools such as TAACO (a software for analyzing textual cohesion at local, global, and textual levels) (Crossley, 2019) and TERA (an evaluator of text ease and readability), created to support educators in selecting appropriate texts for students (Jackson, 2016).

Traditionally, text complexity assessment methods have relied on shallow features extracted from a text to evaluate its quality. These features typically include word and sentence frequencies, counts of grammatical errors, lexical categories, and readability indices. Some more advanced indicators also consider the lexical diversity of the vocabulary used in the text. In this approach, text evaluation is often framed as a linear regression problem, where each shallow feature is assigned a weight to predict a text readability score. These weights are determined by analyzing samples of texts reviewed and assessed by multiple expert raters.

In contrast, semantic-based assessment approaches operate in how semantic relationships between text utterances will be correlated with the text difficulty. To estimate semantic relationships, these approaches represent texts using vector space models, capturing the most relevant words or terms in each text. Similarity between utterance vectors is then measured using metrics like the cosine similarity, or even other similarity measurements.

To improve these methods, latent semantic analysis (Dumais, 2004) has been applied to uncover hidden semantic relationships between textual elements while reducing the dimensionality of the data. LSA generates high-dimensional semantic vectors that represent the lexical and semantic knowledge of each term in an essay. These vectors can be compared to those of human-assessed essays using cosine similarity. However, a limitation of LSA is that it does not account for word order, meaning that sentences with different meanings may be treated as equivalent.

HYBRID LINGUISTIC LLM FOR TEXT COMPLEXITY

This research presents a novel method for text complexity assessment. We developed new semantic features for coherence and cohesion approximation, and combined them with classical shallow features (linguistic traits) and large language models (LLMs).

This approach builds on text assessment and uses ideas from automated essay scoring (Atkinson, 2025), and we incorporate what we will call a common knowledge assessor which contains knowledge embedded as an LLM. Moreover, we contribute to computational linguistics research by providing a new set of coherence and cohesion features which are built on top of transformer models. To evaluate the performance of our model, we conducted a comparative analysis on the different approaches, on a school text corpus labelled by human experts. The results yielded promising findings, establishing a high accuracy between the model's prediction and human assessments. The architecture of the solution is shown in Figure 1.

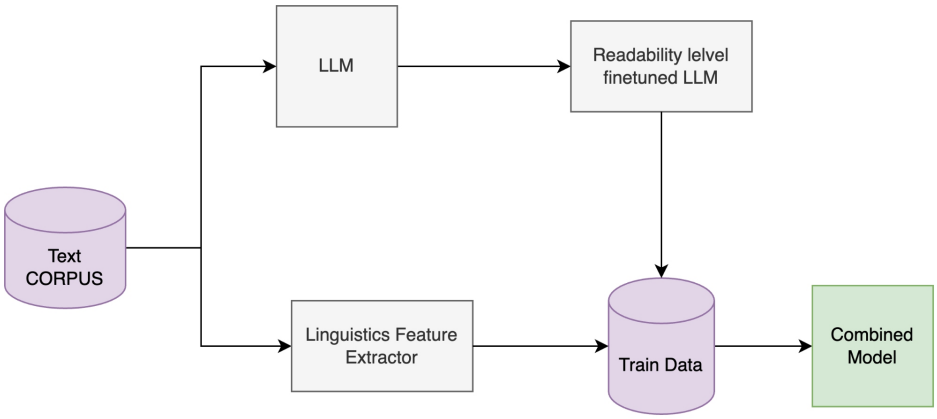

Figure 1: Architecture of the LLM-hybrid readability assessor.

Surface Proxies Features

The surface proxies (i.e. features that can be computed purely from the text) were grouped into three categories: readability, lexical diversity and grammar. The readability indices aim to determine how easy it is to read the text. The lexical diversity is a measurement of vocabulary diversity in the text. Grammar indices establish how efficient is the use of linguistic resources across the text (Palma, 2018).

A drawback of these features is that they only use shallow information in the text, such as word counts, syllable count, POS tag distribution, among others. They do not use semantics or any other trait that is needed for coherence and cohesion.

We use classical readability features that can be extracted directly from the text. These are shallow features that are approximations to text complexity. As the set of features we use has a high cardinality, we will just mention the most relevant ones:

- TTR of lemmas: We first lemmatize the words to a canonical form, and then we compute:

$$TTR = \frac{Unique(words)}{Count(words)}$$

- We also compute the TTR to do the same for functional words, as an approximation of cohesion. The computation is similar, except we only consider functional words.
- Verb and lexical overlap between sentences.
- Concreteness, valence and context availability (Guasch, 2016).
- Verb and auxiliary density.

For the computation we rely on TRUNAJOD, which implements most of text complexity features existing in the literature.

Discourse Patterns Based Features

Discourse pattern features aim to compute the local coherence of a text, measuring shifts in the discourse. We use an entity grid (Barzilay, 2008) based approach, in which we represent the text in a grid where rows are sentences and the columns are entities (e.g. noun phrases). Each cell contains the information of what role the entity took in a given sentence, for example subject, object or not present.

Then we compute transitions of entity types, for example {`so`} would be that in a sentence the entity took the role of a subject and in the next sentence it took the role of an object. The rationale for using these features is that the distribution of entities on coherent texts have certain regularities in the topology of the grid, whereas non coherent texts would have topologies that differ from these regularities.

While discourse patterns have been successful in measuring text coherence, these features still only use surface information and are only based on

distribution of topics, entities, but not in the meaning, which is required to identify how coherent a text is.

	GOVERNMENT	ALLEGATION	TURKEY	CONSTITUTION	SECULAR	REPUBLIC	MAJORITY	MILITARY	LEADER	PRESIDENT	DEMIREL	VIRTUE	PARTY	FRINGE	BUSINESS	EU
s_1	S	X	–	–	–	–	–	–	–	–	–	–	–	–	–	–
s_2	–	–	X	S	X	O	X	–	–	–	–	–	–	–	–	–
s_3	–	–	–	–	X	–	–	X	S	X	S	X	O	X	–	–
s_4	–	–	–	–	–	–	–	–	S	–	–	S	–	–	X	O

Figure 2: Example of an entity grid (Barzilay, 2008).

Text Segmentation

For the text segmentation we split the text into chunks based on two hyper-parameters, an overlap o, and a window w. The window size is the text segment size in words, the overlap is the amount of words that the end of a segment overlaps with the beginning of the next segment. This heuristic is to extract utterances, to avoid biases introduced by long sentences. Then, each text segment is converted into a vector, using embeddings from a pre-trained language model based on BERT for the Spanish language (Cañete, 2023).

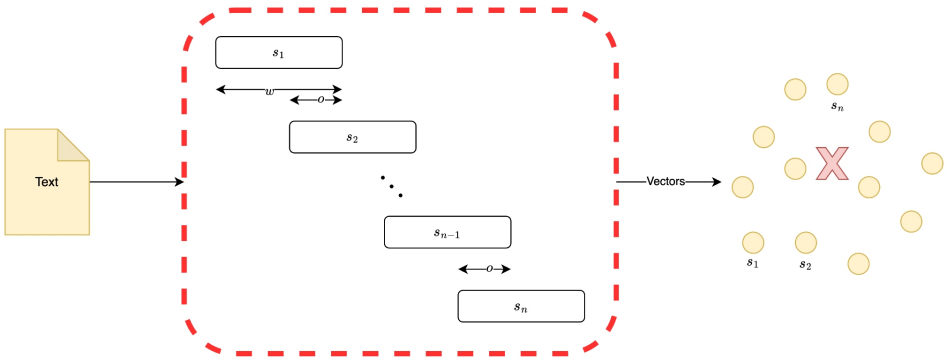

Figure 3: Text segmentation process to compute coherence features.

Coherence Features

For the development of the new readability features we use text segments in a vector space (Zupanc, 2017). In our case, however, we extend the set of features and also we use a language model BETO to get the vectorial

representations, this adds semantic relatedness between words inside the text as well as text segment, instead of using a term frequency approach, which has the disadvantage of considering tokens as a bag of word and approximating meaning to word/token overlap.

The givenness of a text is defined as the new information that is added to the text across text segments, for example sentences, utterances, paragraphs, etc. To compute the givenness, we segment the text with the approach described in the text segmentation, and then we pass the text to a trained sentence transformer, from which we extract sentence embeddings, which are used to represent the texts in a vector space so we can apply mathematical operations on them.

We compute two types of givenness, based on different mathematical operations:

1. Sequential Givenness
2. Projection Givenness

For the sequential givenness, given a sentence, we create a vector space based on all the previous text segments, then we compute an orthogonal vector to this space. For the current text segment, we get its embedding representation and compute the similarity to the orthogonal vector. We use this similarity as the amount of new information added to the text, i.e. givenness.

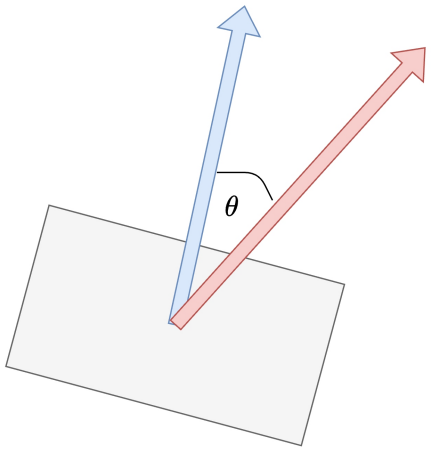

Figure 4: Vector projections of two text segments.

Then, we compute the average, the max and the min givenness. For the projection givenness, we use all the embeddings from the sentences previous to the current sentence. Then, using least squares we compute:

$$Ax = b$$

In this case A is the set of embeddings previous to the current sentence and b is the current sentence. Then we compute the the projection of the new sentence embeddings in the hyperplane formed by the previous vectors, this

vector is $b_w = Ax$, and then we compute the opposite vector $b_{wo} = b - bw$. Then we compute the similarity to these projections $N = sim(b, b_{wo})$ and $G = sim(b, b_w)$ and we can compute a measurement of the new information based on the similarity to the vertical projection by computing $Givenness = N/(N + G)$.

For the lexical diversity, we consider the distribution of words (tokens) in the text, and compare this distribution to a uniform distribution. Then, as a measure of the lexical diversity, we compute the Kullback–Leibler divergence between these two distributions:

$$LD = D_{KL}(T, U) = \sum_{x \in X} T(x) \log \left(\frac{T(x)}{U(x)} \right)$$

Given text segments, we define the distance from one segment to its closest text segment as r_i. We can define a relative distance between segments which measures how the observed distribution differs from the mean of the nearest neighbors. This is an approximation of how fast an idea is developed throughout the text. Considering N text segments, we compute:

$$R = \frac{2\sqrt{N} \sum_{i=1}^{N} r_i}{N}$$

Combination With LLMs

Since we have text and group levels of complexity, we can fine-tune a LLM so it can classify texts based on the level. We fine-tuned the GPT4o model for this task. Once the LLM is fine-tuned, we can pass a new text, and get the classification from it.

On the linguistics side, we can train a machine learning model based on the linguistics features described in the previous sections. Then, we consider the LLM as an extra judge of the text classification, and add their assessment to the feature group.

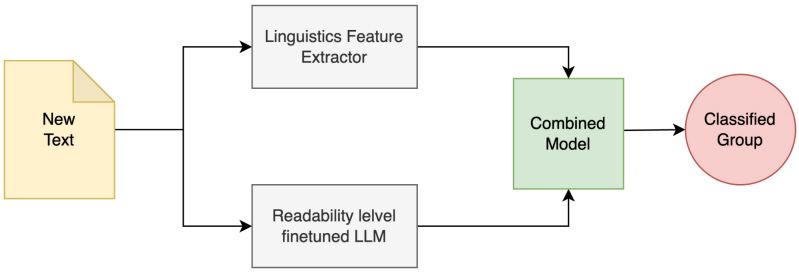

Figure 5: Inference of readability complexity on a new text.

RESULTS

We collected data from different grade levels in schools. In particular, we considered texts from the educational system in Chile. All the texts had a grade level assignment. Data and code are available in: https://github.com/dpalmasan/paper_trunajod_llm

Table 1: Sample human systems integration test parameters (Folds et al., 2008).

Grouped Levels	Number of Texts
1_2_3B	208
4_5_6B	270
7_8B	178

Table 2: Comparison of the different models.

Model	Accuracy
LLM (GPT4o)	0.18
Linguistics Features Model	0.61
Hybrid Model (LLM + linguistics features)	0.75

CONCLUSION

In this paper, a novel method for the text complexity assessment was proposed. The approach combines semantic based approximations to cohesion and coherence with a common knowledge base, in order to categorize the texts into different complexity levels. Unlike the state of the art, we enhanced readability features using a pre-trained model (based on BERT) for text segment representation in a vector space. Experiments showed that the proposed linguistic features are strong predictors of the text complexity. Moreover, when combining these features with a rater (LLM) fine-tuned for the text complexity assessment, we can boost the model's precision.

ACKNOWLEDGMENT

This work was supported by the National Fund for Scientific and Technological Development (FONDECYT ANID), under the Project No. 1231433: "Metacognición de la lectura y de la escritura: dimensiones teóricas y aplicadas." We gratefully acknowledge this funding, which made the development and completion of this research possible.

REFERENCES

Albaladejo, T., & García, A. (1983). La lingüística del texto (pp. 217–262). Alhambra.

Atkinson, J. and Palma, D., 2025. An LLM-based hybrid approach for enhanced automated essay scoring. Scientific Reports, 15(1), p. 14551.

Barzilay, R. and Lapata, M., 2008. Modelling local coherence: An entity-based approach. Computational Linguistics, 34(1), pp. 1–34.

Cañete, J., Chaperon, G., Fuentes, R., Ho, J. H., Kang, H. and Pérez, J., 2023. Spanish pre-trained Bert model and evaluation data. arXiv preprint arXiv:2308.02976.

Crossley, S. A. (2020). Linguistic features in writing quality and development: An overview. Journal of Writing Research, 11(3), 415–443.

Crossley, S. A., Kyle, K. and Dascalu, M., 2019. The Tool for the Automatic Analysis of Cohesion 2.0: Integrating semantic similarity and text overlap. Behaviour research methods, 51(1), pp. 14–27.

Dumais, S. T., 2004. Latent semantic analysis. Annual Review of Information Science and Technology (ARIST), 38, pp. 189–230.

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. Educational researcher, 40(5), 223–234.

Guasch, M., Ferré, P. and Fraga, I., 2016. Spanish norms for affective and lexico-semantic variables for 1, 400 words. Behaviour Research Methods, 48(4), pp. 1358–1369.

Jackson, G. T., Allen, L. K. and McNamara, D. S., 2016. Common core TERA: Text ease and readability assessor. In Adaptive educational technologies for literacy instruction (pp. 49–68). Routledge.

Louwerse, M. M. (2004). Un modelo conciso de cohesión en el texto y coherencia en la comprensión. Revista signos, 37(56), 41–58.

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

Palma, D. A., Soto, C., Veliz, M., Karelovic, B. and Riffo, B., 2021. TRUNAJOD: A text complexity library to enhance natural language processing. Journal of Open Source Software, 6(60), p. 3153.

Palma, D. & Atkinson, J. Coherence-based automatic essay assessment. IEEE Intelligent Systems 33, 26–36 (2018).

Van Dijk, T. A. (1985). Handbook of Discourse Analysis, Vol. 2: Dimensions of Discourse. London: Academic Press.

Van Dijk, T. A. (1997). Discourse as Structure and Process. London: SAGE Publications.

Zupanc, K. and Bosnić, Z., 2017. Automated essay evaluation with semantic analysis. Knowledge-Based Systems, 120, pp. 118–132.