Book Complexity Level Assignment in French and Portuguese

Jorge Baptista[1,2], **David Antunes**[2], **Wafa Aissa**[3], **Julien Zakhia Doueihi**[3]
Trang Pham Tran Hanh[3], **Eugénio Ribeiro**[2,4], **Thomas François**[3], **Raquel Amaro**[5]
[1]U. Algarve, Portugal; [2]INESC-ID Lisboa, Portugal; [3]UCLouvain, Belgium;
[4]Iscte-IUL, Portugal; [5]U. NOVA Lisboa, Portugal
{jorge.baptista, david.f.l.antunes, eugenio.ribeiro}@inesc-id.pt;
{wafa.aissa, julien.zakhia, tran.pham, thomas.francois}@uclouvain.be;
raquelamaro@fcsh.unl.pt

Abstract

Selecting reading materials suitable for adults with low literacy levels is a challenging task in Adult Learning (AL) contexts, particularly when dealing with full books or long texts, as textual complexity may be difficult to infer from the analysis of a small set of extracted samples. This paper presents experiments on estimating the complexity level of full-length books in Portuguese and French, with the aim of identifying the most effective sampling and result aggregation procedure. The procedure is supported by a natural language processing and machine-learning-based system for the automatic assignment of textual complexity levels. Complexity levels follow a scale specifically devised for AL audience, aligned with internationally recognised scales, ranging from L1 (*Very Easy*) to L4 (*More Complex*). The ultimate goal is to guide the expansion of a publicly accessible database of book titles with reliable complexity information, particularly benefiting key stakeholders in AL, such as students and teachers, as well as librarians and publishers concerned with literacy promotion.

Keywords: Text Complexity Assignment, Full Books, Sampling, Adult Learning, French, Portuguese

1. Introduction and Motivation

Selecting reading materials that are appropriate for adults with low literacy skills remains a central challenge in Adult Learning (AL) contexts. This challenge becomes particularly acute when the unit of analysis is not a short passage but a full book or long-form text, where internal heterogeneity in lexical, syntactic, and discourse-level properties makes global readability estimation non-trivial (Li et al., 2024). In practice, librarians, educators, and publishers often need to make decisions about the suitability of books for specific learner populations without access to complete texts, relying instead on partial excerpts or limited samples, and personal intuition.

Within the iRead4Skills project[1] (iR4S), textual complexity is operationalised through a four-level scale (L1–L4) specifically designed to address low-literacy adult learners' needs (Amaro et al., 2025; Monteiro et al., 2023), and aligned with CEFR and PIACC-inspired notions of linguistic difficulty (Council of Europe, 2020; OECD, 2013a,b, 2021). While CEFR primarily targets second-language learners, its descriptors have been adapted in iR4S to address functional literacy in adult native populations. Nevertheless, the extent to which these levels generalise across different populations remains an open question.

Whereas classic readability research explicitly addressed sampling procedures for long texts, most recent natural language processing (NLP)-based approaches demonstrate strong performance at the short-text or passage level (Collins-Thompson and Callan, 2005; Ribeiro et al., 2024b,a). The methodological and empirical implications of extrapolating such passage-level models to full-book assessment, however, remain comparatively underexplored. A naive aggregation of passage-level predictions can lead to biased or unstable estimates, particularly for long texts that exhibit substantial variation in difficulty across chapters or sections.

While recent large language models can process longer contexts, their effective input length remains constrained relative to full-length books, particularly for texts exceeding tens of thousands of tokens. Moreover, such approaches entail high computational costs and reduced interpretability when applied in operational settings such as Adult Learning. In contrast, sampling-based strategies offer a lightweight, transparent, and controllable alternative, compatible with existing readability models and realistic deployment constraints (e.g., partial access to texts, limited processing resources).

This paper addresses this gap by investigating principled strategies for estimating the global complexity level of full-length books in Portuguese and French based on a limited number of textual samples and on the iR4S complexity analysis (see Appendix C). Our objective is twofold: first, we aim to identify sampling and aggregation strategies that yield robust and interpretable book-level complexity estimates under realistic usage constraints, such as partial text availability and limited compu-

[1] https://iread4skills.com/

Proceedings of the Joint Workshop on Readability and Text Simplification (READIxTSAR) @ LREC 2026, pages 12–25
11 May 2026.

tational resources; second, we seek to empirically validate these strategies against human-annotated gold standards, on the one hand, and against full-text analysis, on the other, assessing their reliability and practical adequacy for deployment in AL and literacy promotion scenarios.

The work reported here builds directly on earlier methodological discussions within iR4S concerning lightweight yet informed sampling protocols for long texts, and is explicitly inspired by recent advances in long-document readability modelling. In particular, we adapt the conceptual framework proposed by Li et al. (2024), which treats book-level readability as an emergent property arising from the distribution of difficulty across multiple textual segments rather than from isolated passages. By grounding our experiments in this perspective, we aim to contribute empirical evidence and concrete methodological guidance for book-level readability assessment in diversified settings.

2. Related Work: Readability Beyond Passages

Automatic readability assessment has a long tradition in educational research and Natural Language Processing. Early approaches relied on handcrafted readability formulas—such as Flesch (Flesch, 1948), Dale–Chall (Dale and Chall, 1948; Chall and Dale, 1995), and SMOG (McLaughlin, 1969)—which operationalised textual difficulty through a small set of shallow surface-level proxies, including sentence length, syllable-based word length, or predefined lexical lists. While these formulas remain influential in educational practice, particularly in institutional and pedagogical settings, their limited linguistic coverage, reliance on linear models, and reduced sensitivity to discourse, syntactic, and semantic variation restrict their applicability across genres, populations, and languages (Benjamin, 2012; Collins-Thompson, 2014).

These limitations are especially salient in Adult Learning (AL) contexts, where textual complexity must align with functional literacy goals and proficiency scales, such as CEFR. As a result, research has progressively shifted towards data-driven, non-linear machine learning approaches (Feng et al., 2010), capable of modelling richer lexical, syntactic, and semantic representations of text difficulty.

Subsequent work reframed readability assessment as a supervised classification or regression task, exploiting a wider range of linguistically-informed features. These include lexical frequency and diversity measures, morphosyntactic complexity indicators, and discourse-level cues, often combined within statistical or kernel-based classifiers (Schwarm and Ostendorf, 2005; Petersen and Ostendorf, 2009; Sheehan et al., 2010; François and Fairon, 2012). More recent studies have incorporated neural representations, either as standalone predictors or in hybrid architectures that integrate pre-trained embeddings with handcrafted linguistic features (Deutsch et al., 2020; Lee et al., 2021; Wilkens et al., 2024). These approaches consistently report strong performance at the passage or short-text level.

However, the vast majority of readability research remains fundamentally passage-oriented. Large-scale readability corpora typically consist of short, decontextualised excerpts, carefully controlled for length and often sampled from pedagogical materials (Crossley et al., 2023). While such design choices are methodologically sound for passage-level prediction, they implicitly assume textual homogeneity and fail to account for the internal variation that characterises long documents and full books.

The limitations of passage-based approaches for long texts have long been acknowledged in applied settings. Educational and psychometric guidelines already recommend sampling multiple excerpts from different parts of a book and aggregating their scores to improve reliability (Allan et al., 2005; Stenner et al., 2006). These practices recognise positional effects and within-text variation, but they predate contemporary NLP models and lack empirical validation under modern computational paradigms.

Recent advances in long-document modelling have focused primarily on architectural solutions for processing long sequences, such as sparse-attention transformers and sequence compression mechanisms (Beltagy et al., 2020; Zaheer et al., 2020). While these models extend the maximum input length, they do not directly address the conceptual mismatch between passage-level difficulty signals and book-level readability, nor do they encode explicit notions of difficulty distribution within a text.

A decisive step towards bridging this gap is provided by Li et al. (2024), who explicitly argue that book-level readability cannot be reliably inferred from isolated passages or naive truncation strategies. Through extensive experiments, they demonstrate that direct transfer from passage-level models leads to systematic bias, typically overestimating difficulty for long texts. Their work introduces the notion of books as compositions of multiple "difficulty fragments", whose distribution and aggregation are informative for global readability estimation. Crucially, they show that aggregation strategies and sampling density play a more decisive role than raw model capacity when moving from passages to books.

The present work adopts this perspective while departing from purely model-centric solutions.

Rather than proposing a new end-to-end neural architecture for long texts, we focus on sampling-aware and aggregation-aware strategies that are compatible with realistic usage scenarios, namely those involving the use of the iR4S system by trainers in AL contexts, as well as by librarians and publishers with limited time, to classify books according to their level of complexity. In particular, we build on established multi-sample readability practices (Allan et al., 2005; Stenner et al., 2006), recent corpus-based insights on controlled excerpt selection (Crossley et al., 2023), and the distributional view of difficulty advanced by Li et al. (2024). This allows us to investigate how lightweight sampling regimes and simple aggregation operators can yield robust and interpretable book-level complexity estimates across languages, even under constrained access to full textual data.

3. Automatic Book-Level Complexity Estimation

3.1. Corpus and Sampling Procedure

Two corpora of full-length texts were compiled for the experiments, one per language. The Portuguese corpus was constructed following a two-step procedure. The iR4S corpus (Pintard et al., 2024) contains short excerpts from longer texts that had previously been assigned a complexity level. In order to enable full-text analysis, the complete digital versions of the corresponding source texts (e.g., full books or documents) were retrieved for each excerpt and incorporated into the new corpus.

Second, additional texts from publicly available repositories were included to ensure balance across complexity levels. The overall classification of the texts was validated by experts familiar with the iR4S classification framework (Levels L1–L4) and specifically trained for this level-assignment task. The final corpus comprises 16 texts, with four texts per level, totalling 943,169 words. Corpus details are provided in Appendix A.

For the French corpus, we initially retrieved open-access books from the ABU repository[2], but preliminary experiments revealed that such corpus includes nearly only books from the 'More Complex' level (L4). As we need a fair representation of the four levels of the scale used, we then selected 29 simplified readers published for French as a foreign language. These books are either simplified version of classic literary work (e.g., *Carmen*, *Germinal*) or original stories adapted for learners. According to their CEFR levels, these texts can be classified as L1 ('Very Easy'), roughly corresponding to A1, L2 ('Easy') roughly corresponding to A2, and L3 ('Plain'), roughly corresponding to B2. (cf.Table 7 in Appendix B for their name and other details). Together with the previously collected titles, the compiled corpus covers all iR4S complexity levels. Selected books were scanned with optical character recognition tools and manually revised. Both corpora consist of full-text books classified according to the iR4S levels. Due to copyright constraints, full texts of the corpora cannot be distributed. However, sampling procedures, model descriptions, and aggregated annotations are documented in sufficient detail in Appendices A and B to ensure methodological reproducibility.

While the two corpora (Portuguese and French) are not strictly parallel in composition, particularly due to the inclusion of simplified readers in French to ensure level balance, both datasets are aligned with the same iR4S complexity scale and validated by expert annotation. The goal of the study is not strict cross-lingual comparison but rather the evaluation of sampling and aggregation strategies across two typologically different yet methodologically compatible settings.

All texts were manually cleaned to remove paratextual material (e.g., headers, footers, tables, figures and captions, indices, footnotes, and references).

A dedicated sampling tool was developed to automatically segment text files into excerpts of approximately equal length. The tool allows fine-grained control over the minimum and maximum size of each excerpt, while preserving paragraph boundaries and preventing paragraph splits. It also enables users to specify the number of excerpts to be extracted. The selected excerpts are drawn from different sections of the text and are evenly distributed throughout the document.

For the purposes of our experiments, the sampling tool first segments each text into as many valid excerpts as possible. It then selects a fixed set of 10 samples per text. These samples are drawn from evenly distributed *loci* within the book (beginning, 10%, 20%, . . . , end). Each sample contains between 250 and 300 words (approximately one printed page) and preserves paragraph boundaries. Ideally, all selected texts would have had a minimum length of 5,000 words. Two texts in the Portuguese corpus were shorter; from these, only 2 and 8 samples, respectively, could be extracted in compliance with the constraints described above.

3.2. Experiment 1: Sampling-Based Automatic Estimation

The first experiment examines the estimation of overall book complexity based exclusively on automatically predicted complexity levels assigned to sampled text segments. It is conducted for both Portuguese and French and reflects a realistic de-

[2] http://abu.cnam.fr/

ployment scenario in which no manual expert annotations are available for the full books.

Each book is segmented into fixed-length excerpts according to the sampling protocol described in Subsection 3.1. Segment-level complexity is then automatically estimated using the iR4S complexity assessment model for each language (see Appendix C), resulting in a distribution of predicted complexity levels across the book.

To derive a single book-level complexity estimate, we evaluate several aggregation strategies, including measures of central tendency (mean and median), extremal values (minimum and maximum), frequency-based aggregation, and quantile-based approaches (25th and 75th percentiles).

The primary objective of this experiment is to assess how well a reduced number of samples or excerpts approximates the complexity profile obtained from a denser representation of the book, based on *all* excerpts into which it can be segmented.

Reduced sampling configurations (ranging from 1 to 10 excerpts) are then compared against this reference profile using correlation and error-based metrics, including Mean Absolute Error (MAE) and Spearman's rank correlation coefficient. This allows us to quantify how closely the reduced samples approximate the estimated full-book complexity level. For each sampling configuration, excerpts are selected from evenly distributed positions throughout the text (e.g., $S = 5$ corresponds to samples drawn at 20%, 40%, 60%, 80%, and 100% of the text length).

Table 1 presents the results of this experiment for Portuguese texts, while Table 2 reports the corresponding results for French texts.

For the Portuguese books, Table 1 presents the approximation performance of reduced excerpt samples across aggregation strategies and sample sizes, while Figure 1 in Appendix C illustrates the corresponding trends. Overall, performance improves as the number of sampled excerpts (S) increases, although the progression is not strictly monotonic across all aggregation strategies.

Overall, MAE decreases with increasing S, although convergence behaviour differs across aggregation strategies. Central-tendency estimators (mean, median, mode) stabilise more quickly, yielding consistently low error from moderate sampling levels onward ($S \geq 5$). In particular, the mean reaches minimal error at $S = 7$ (MAE = 0.0000) and remains stable thereafter. In contrast, extreme-value strategies (min, max) exhibit higher variability and less consistent improvement, while quantile estimators (Quantile (Q25), Quantile (Q75)) show intermediate behaviour, with higher error at small S followed by gradual stabilisation.

The Spearman's ρ results mirror these patterns. Mean and median aggregation maintain relatively high and stable rank correlations across sampling levels, indicating preserved book ordering under reduced sampling. Quantile estimators improve steadily with larger S, whereas extreme-value methods display greater fluctuation in ranking consistency.

Paired t-tests on per-book differences confirm these patterns. Central-tendency estimators show no significant deviation from the baseline at any sampling level (mean: $p = 0.43$ at $S = 1$, $p = 0.97$ at $S = 6$, all $p > 0.16$ overall; median and mode similarly non-significant across all S). In contrast, the maximum aggregator is significant for all sampling levels ($S = 1$–10; all $p \leq 0.041$), and the minimum is significant for every $S \neq 8$ (all $p < 0.05$). Quantile estimators exhibit limited early instability: Q25 is significant at $S = 1$ ($p = 0.029$) but not thereafter, while Q75 is significant at $S = 2$ ($p = 0.009$) and non-significant for larger S.

Taken together, the results demonstrate that central-tendency aggregation is highly robust to reduced sampling in Portuguese. Moderate sampling levels ($\approx$ 6–8) are sufficient to achieve low error, stable ranking, and no statistically significant deviation from the full-book baseline, whereas extreme-value methods remain comparatively unstable and systematically biased.

For the French books, Table 2 reports the book-difficulty approximation performance of reduced excerpt samples across aggregation strategies and sample sizes (Figure 2 illustrates the corresponding trends). We observe that performance improves consistently as the number of sampled excerpts (S) increases. For central-tendency aggregators, MAE decreases markedly between $S = 1$ and $S = 5$ and stabilises from $S \geq 7$, while Spearman's ρ exceeds 0.95 from $S = 7$ onward, indicating strong ranking consistency with the baseline.

Paired t-tests on per-book differences were conducted to assess whether reduced sampling introduces systematic deviation from the full-excerpt baseline. Results reveal a sample-size effect for quantile and extreme-value aggregators, whereas central-tendency estimators (mean, median, mode) remain stable across sampling levels. Mean aggregation shows no significant deviation at any S (e.g., $p = 0.43$ at $S = 1$, $p = 0.97$ at $S = 6$), and similar patterns are observed for median and mode. In contrast, Q25 is significant for $S = 1$–3 ($p = 0.0006$ at $S = 1$) before stabilising, while Q75 remains significant up to $S = 5$ ($p = 0.0040$ at $S = 1$). Extreme-value methods are the least stable: min is significant from $S = 1$ to $S = 9$ ($p < 0.01$), and max is significant for $S = 1$–7 and again at $S = 10$ ($p = 0.0433$). Overall, increasing S reduces systematic deviation for robust aggregation strategies, but extreme-value approaches remain persistently biased.

Table 1: Approximation of Portuguese full-book complexity using reduced samples. S denotes the number of samples. Performance is measured using Mean Absolute Error (MAE ↓) and Spearman's rank correlation coefficient (ρ ↑). Best values in each column are marked in bold. Best values in each row are marked in italic.

	Mean		Max		Min		Median		Mode		Q25		Q75	
S	MAE ↓	ρ ↑	MAE ↓	ρ ↑	MAE ↓	ρ ↑	MAE ↓	ρ ↑	MAE ↓	ρ ↑	MAE ↓	ρ ↑	MAE ↓	ρ ↑
1	*0.3125*	*0.6873*	0.7500	0.4927	0.8125	0.4849	0.3750	0.6407	0.4375	0.5526	0.5625	0.4277	*0.3125*	*0.6873*
2	0.5000	0.5032	0.5625	*0.7037*	0.5000	0.6551	0.3125	0.6184	*0.2500*	0.7032	*0.2500*	0.6990	0.3750	*0.7037*
3	0.3125	0.6407	0.5000	0.7062	0.4375	0.5952	0.3125	0.6793	0.2500	0.7868	0.3125	0.6587	*0.1250*	*0.8615*
4	*0.1250*	0.8281	0.4375	0.7174	0.3750	0.6708	*0.1250*	0.8502	0.2500	0.6859	*0.1250*	*0.8783*	0.2500	0.7075
5	*0.1250*	0.7570	0.5625	0.7037	0.3750	0.6708	0.2500	0.6762	0.2500	0.7032	0.1250	0.8783	*0.1250*	*0.9074*
6	0.3125	0.6407	0.4375	0.7174	0.3750	0.6708	0.3125	0.6793	0.2500	0.7868	0.3125	0.5928	*0.1250*	*0.9033*
7	***0.0000***	***1.0000***	**0.2500**	**0.7998**	0.2500	0.7266	**0.0625**	**0.9952**	**0.0625**	**0.9923**	0.1875	0.7838	0.1250	**0.9902**
8	*0.0625*	0.8953	0.4375	0.7174	**0.1250**	**0.8478**	**0.0625**	***0.9952***	0.1875	0.7238	***0.0625***	**0.9747**	***0.0625***	0.9172
9	0.1875	0.7649	**0.2500**	0.7304	0.2500	0.7266	0.1875	0.7863	0.2500	0.6859	0.3125	0.5928	*0.1250*	*0.9172*
10	0.1875	0.7146	**0.2500**	0.7304	0.3125	0.6929	0.1875	0.7503	0.1875	0.7764	*0.1250*	*0.8783*	0.1875	0.8128

Table 2: Approximation of French full-book complexity using reduced samples. S denotes the number of samples. Performance is measured using Mean Absolute Error (MAE ↓) and Spearman's rank correlation coefficient (ρ ↑). Best values in each column are marked in bold. Best values in each row are marked in italic.

	Mean		Max		Min		Median		Mode		Q25		Q75	
S	MAE ↓	ρ ↑	MAE ↓	ρ ↑	MAE ↓	ρ ↑	MAE ↓	ρ ↑	MAE ↓	ρ ↑	MAE ↓	ρ ↑	MAE ↓	ρ ↑
1	0.2787	*0.8396*	0.4828	0.5847	0.7931	0.3970	*0.1379*	0.8333	*0.1379*	0.8333	0.3276	0.7352	0.2414	0.7701
2	0.2378	*0.8519*	0.3103	0.6708	0.7241	0.4181	0.1897	0.8260	0.2069	0.7723	0.4052	0.7033	*0.1466*	0.8425
3	0.1521	*0.9146*	0.3103	0.5787	0.5172	0.2940	*0.1379*	0.8487	*0.1379*	0.8487	0.2241	0.8138	0.1897	0.8368
4	0.1907	0.8978	0.2414	0.6920	0.4828	0.5670	0.1897	0.8334	0.2069	0.8044	0.2672	0.7810	*0.1034*	*0.9104*
5	0.1671	*0.9172*	0.2414	0.7335	0.2759	0.6002	*0.1379*	0.8261	*0.1379*	0.8261	0.2241	0.7246	0.2414	0.7824
6	*0.1109*	*0.9428*	0.2069	0.7273	0.2759	0.5287	0.1379	0.8678	0.1379	0.8341	0.1810	0.8619	0.1293	0.8488
7	0.0825	*0.9718*	0.1379	0.8051	0.3103	0.4862	***0.0345***	**0.9595**	***0.0345***	**0.9595**	0.1379	**0.9396**	**0.0517**	**0.9459**
8	0.1011	*0.9645*	**0.1034**	**0.8482**	0.2069	0.6772	0.1379	0.8816	0.1724	0.8218	0.1897	0.8923	*0.0603*	0.9271
9	*0.0912*	*0.9630*	**0.1034**	**0.8482**	0.2414	0.5666	0.1034	0.8880	0.1034	0.8880	0.1379	0.8790	0.1034	0.8974
10	**0.0699**	***0.9856***	0.1379	0.8051	**0.1034**	**0.7614**	0.0517	0.9571	***0.0345***	**0.9595**	**0.1034**	0.8852	0.0603	0.9057

Overall, $S = 7$ represents a practical trade-off between computational efficiency and approximation quality, as error is low, rank consistency is high, and systematic deviations are substantially reduced for robust aggregation methods, with only marginal gains observed beyond this point.

3.3. Experiment 2: Comparison with Human Gold Standard

The second experiment evaluates the correspondence between automatically estimated book-level complexity and a human-annotated gold standard. In this setting, expert annotators assigned complexity levels to sampled excerpts using the iR4S L1–L4 scale (Amaro et al., 2025; Monteiro et al., 2023), following established annotation guidelines.

The use of excerpt-based annotation reflects both practical constraints and real-world evaluation scenarios, where full-book assessment is typically performed through representative sampling.

Segment-level annotations were adjudicated to obtain a reliable reference, and inter-annotator agreement was measured to assess annotation consistency. Book-level complexity estimates were then derived by aggregating the annotated excerpt-level labels using the same set of aggregation strategies explored in Experiment 1. The strategy of estimating long-text complexity by aggregating excerpt–level complexity not only mimics the intended real-world scenario — where trainers/librarians assess a book's difficulty by reviewing a limited number of pages or passages–but also replicates the experimental setting of Experiment 1.

For Portuguese, inter-annotator agreement was measured on a subset of 37 excerpts (out of 150 total) that were independently labelled by two annotators. Because the complexity levels are ordinal (i.e., the ordering between classes is meaningful), we employed Cohen's weighted kappa coefficient, which accounts not only for agreement by chance but also for the degree of disagreement between ordered categories. The resulting kappa score of 0.733 indicates substantial agreement between an-

notators, according to commonly used interpretation scales. This level of agreement suggests that the annotation task is reasonably well-defined and consistently interpretable.

Performance is evaluated by comparing the automatically derived book-level estimates against the reference book-level labels inferred from human annotations. Metrics include accuracy, MAE, and agreement-oriented measures appropriate for ordinal scales. This experiment allows us to identify aggregation strategies that not only perform well in automatic settings but also align most closely with expert human judgment.

Although complexity levels form an ordinal scale, we report standard classification metrics (Accuracy, Precision, Recall, F1-score) to facilitate comparability with prior work in readability assessment. We complement these with error-based metrics (MAE) and agreement-oriented measures, which better capture the magnitude of deviations between predicted and reference levels. We acknowledge that future work could explore ordinal-specific evaluation measures.

The results of this comparison should provide critical validation for the proposed methodology and inform practical recommendations regarding sample size and aggregation functions for book-level complexity estimation. Again, the main objective of this experiment was to determine the most effective sampling strategy for estimating the global complexity level of a book. Since book-level complexity is inferred from annotations at the excerpt level, two key design choices were explored: (i) the size of the pool of excerpts sampled from each book, and (ii) the aggregation method used to combine the complexity annotations of these excerpts into a single book-level estimate.

To this end, we experimented with three different sample sizes (3, 5, and 10 evenly-distributed excerpts per book) and the same diverse set of aggregation strategies as before, including statistical operators (e.g., average, minimum, maximum, median), frequency-based aggregation, and quantile-based approaches. This setup allows us to assess how sensitive the final complexity estimation is to both the amount of available evidence and the way in which this evidence is summarised. Each configuration was evaluated using standard classification metrics (macro-averaged Accuracy, Precision, Recall, and F1-score), comparing the predicted global complexity level with the reference labels.

For the Portuguese books, a first observation from the results in Table 3 is that aggregation strategy has a stronger impact on performance than sample size. While increasing the number of excerpts from 3 to 10 sometimes leads to modest improvements, these gains are neither consistent nor systematic across aggregation methods. This suggests that simply sampling more excerpts does not necessarily yield a better estimate of global book complexity, and that the way excerpt-level information is combined is crucial.

Across all configurations, Accuracy values range between 0.25 and 0.50, indicating that the task remains challenging and sensitive to methodological choices. Precision, on the other hand, shows larger variability, particularly for aggregation methods that emphasise extreme values.

Aggregation methods based on minimum values (Min and Q25) consistently achieve the best overall performance. These methods reach the highest Accuracy (0.500) and Recall (0.500), and also obtain the best Precision scores (up to 0.711). This trend suggests that the least complex excerpts within a book are particularly informative for estimating its global complexity level. In other words, even a small number of simpler passages may significantly influence how the overall complexity of a book is perceived or categorised.

The Median aggregation strategy also performs competitively, especially for 5 samples, where it achieves an Accuracy of 0.500 and a strong F1-score of 0.490. This indicates that robust, central-tendency measures can provide a good balance between ignoring outliers and retaining representative information from the excerpt pool.

In contrast, Average aggregation yields moderate but stable results across different sample sizes. While it does not achieve the best scores on any metric, its performance is relatively insensitive to the number of excerpts considered. This stability may make it attractive in scenarios where robustness and simplicity are prioritised over peak performance.

Aggregation strategies based on maximum values consistently underperform. The Max strategy exhibits the lowest Accuracy, Precision, and F1-scores across almost all sample sizes. This suggests that focusing on the most complex excerpts leads to noisy or overly pessimistic estimates of global book complexity, likely because highly complex passages are not representative of the book as a whole.

The Mode strategy shows intermediate performance, with reasonable Precision but lower Recall and Accuracy, particularly as the sample size increases. As observed in Experiment 1, this may indicate that mode-based aggregation is sensitive to sampling variability and may struggle when the excerpt-level annotations are heterogeneous.

Finally, in this experiment, Q75 performs worse than its lower-quantile counterpart, reinforcing the suggestion that emphasising higher complexity excerpts does not yield reliable global estimates.

For the French books, an annotation task was carried out using the same guidelines as for the

Table 3: Results for different aggregation strategies and sample sizes of Portuguese books.

Aggregator	S	Accuracy	Precision	Recall	F1-score
Average	3	0.375	0.310	0.375	0.323
	5	0.438	0.357	0.438	0.379
	10	0.438	0.360	0.438	0.379
Max	3	0.250	0.161	0.250	0.195
	5	0.250	0.156	0.250	0.192
	10	0.312	0.179	0.312	0.227
Min	3	**0.500**	**0.711**	**0.500**	0.465
	5	0.438	0.558	0.438	0.398
	10	**0.500**	0.635	**0.500**	**0.493**
Median	3	0.438	0.573	0.438	0.430
	5	**0.500**	0.624	**0.500**	0.490
	10	0.438	0.569	0.438	0.419
Mode	3	0.375	0.538	0.375	0.389
	5	0.438	0.583	0.438	0.443
	10	0.312	0.487	0.312	0.322
Q25	3	**0.500**	**0.711**	**0.500**	0.465
	5	0.375	0.519	0.375	0.364
	10	**0.500**	**0.711**	**0.500**	0.465
Q75	3	0.438	0.573	0.438	0.430
	5	0.375	0.456	0.375	0.344
	10	0.375	0.485	0.375	0.343

Table 4: Results for different aggregation strategies and sample sizes of French books.

Agg.	S	Accuracy	Precision	Recall	F1-score
Mean	3	0.724	0.663	0.662	0.628
	5	0.655	0.523	0.454	0.484
	10	0.759	0.678	0.680	0.651
Max	3	0.793	**0.807**	0.728	0.755
	5	0.724	0.758	0.789	0.759
	10	0.621	0.711	0.644	0.643
Min	3	0.690	0.529	0.462	0.484
	5	0.793	0.549	0.527	0.536
	10	0.759	0.794	0.752	0.748
Median	3	0.724	0.663	0.662	0.628
	5	0.655	0.523	0.454	0.484
	10	0.759	0.678	0.680	0.651
Mode	3	0.759	0.683	0.833	0.651
	5	0.690	0.541	0.458	0.495
	10	0.793	0.696	0.850	0.678
Q25	3	0.586	0.497	0.399	0.442
	5	0.793	0.592	0.533	0.561
	10	0.793	0.575	0.525	0.547
Q75	3	**0.828**	0.781	**0.874**	**0.806**
	5	0.621	0.567	0.599	0.572
	10	0.724	0.763	0.708	0.652

Portuguese team. Inter-rater agreement was measured on a subset of 40 excerpts (out of 290 total). The resulting quadratic kappa score of 0.606 indicates a fair degree of agreement between the annotators, which then conducted a full annotation of the remaining excerpts. Finally, the annotators discussed a Gold Standard which would be used as a baseline to assess the model's performance.

Table 4 compares different aggregation strategies for combining excerpt-level predictions into book-level readability labels across varying numbers of sampled excerpts ($S \in \{3, 5, 10\}$). Overall, Q75 with $S = 3$ achieves the best performance in terms of Accuracy (0.828), Recall (0.874), and F1-score (0.806), while Max with $S = 3$ yields the highest Precision (0.807). These results suggest that upper-quantile–based aggregation strategies are particularly effective at capturing overall book-level difficulty. Interestingly, increasing the number of excerpts does not consistently improve performance. In several cases (e.g., Q75 and Max), performance decreases when moving from $S = 3$ to larger sample sizes. This indicates that a small number of informative excerpts may be sufficient to characterise the overall readability of a book, and that aggregating too many segments may introduce noise rather than additional signal. In contrast, central tendency measures such as Average and Median exhibit stable but consistently lower performance compared to upper-bound–oriented strategies (Max, Q75). This pattern suggests that book-level readability may be driven more strongly by the most difficult segments rather than by average difficulty. Overall, these findings indicate that readability at the book level behaves as a *peak-driven phenomenon*, where the most complex portions disproportionately influence global difficulty judgments.

3.4. Discussion

Regarding sampling-based automatic estimation, the results for both languages show that moderate sampling levels ($S = 6$-8) are sufficient to achieve low error, stable ranking, and no statistically significant deviation from the full-book baseline. In contrast, extreme-value methods remain comparatively unstable and systematically biased. Overall, $S = 7$ represents a practical trade-off between computational efficiency and approximation quality: error rates are low, rank consistency is high, and systematic deviations are substantially reduced for robust aggregation methods, with only marginal improvements observed beyond this point. Taken together, these findings demonstrate that central-tendency aggregation is highly robust under reduced sampling conditions.

The comparison with the human gold standard provides further insights for practical recommendations regarding both sample size and aggregation functions. In both languages, central-tendency estimators yield moderate yet stable results across different sample sizes, making them attractive for the practical scenarios considered. However, extreme-value estimators achieve the best performance in specific cases: in French, the Q75 with three samples yields the lowest MAE and the highest Spearman correlation; in Portuguese, Min and the Q25 suggest that book-level complexity may be shaped

by its most extreme sections rather than by passages that reflect the general level of the text. Nevertheless, given the level of inter-annotator agreement observed, these results may also reflect systematic tendencies among annotators to over- or underestimate the complexity level of the samples.

The two experiments provide complementary perspectives. Experiment 1 evaluates approximation to an automatic full-book profile and favours central-tendency aggregation with moderate sampling ($\approx$ 7 excerpts). In contrast, Experiment 2 evaluates alignment with human judgments, where quantile- and extreme-based strategies may capture perceptual difficulty more effectively. The final recommendation should, therefore, be understood as a practical compromise between stability, interpretability, and alignment with human perception.

Importantly, the proposed sampling framework is not intended as a substitute for full-text processing when such analysis is feasible, but rather as a robust approximation strategy under realistic constraints. These include limited access to full texts, computational cost considerations, and the need for interpretable and reproducible procedures in educational and library settings.

3.5. Book-Complexity Profiling

In light of the results discussed above, and consistent with the empirical finding that a limited number of well-distributed excerpts suffices to approximate global book complexity, we propose a lightweight, user-oriented sampling protocol for the real-world deployment of a database of book titles with automatically assigned complexity levels.[3]

Users are first asked to indicate the total number of pages of the text to be classified. If the text exceeds seven pages, the system requests seven evenly distributed samples; if it contains seven pages or fewer (e.g., a short article), users are asked to provide one sample per page. In all cases, each sample should contain approximately 250–300 words of running text, excluding paratextual material (e.g., table of contents, headers, footers, notes, images, captions, or references). Users are instructed to avoid truncating paragraphs and to indicate the page number of each excerpt. Samples should be drawn from the initial, middle, and final sections of the text to ensure even distribution and representativeness. Each excerpt is analysed automatically and assigned a predicted complexity level, which is stored together with its relative position in the text.

Once all required samples have been processed, the system computes the overall book-level complexity by averaging the excerpt-level predictions. It then presents both the individual sample classifications (the "book profile") and the aggregated result. Users may accept the proposed classification—thereby creating or updating the corresponding database entry, repeat the procedure with alternative samples, or terminate the process. The protocol is intentionally simple and resource-efficient, reflecting the empirical evidence that reliable book-level estimates can be obtained without exhaustive sampling.

[3] https://db.iread4skills.com/

4. Conclusion

This paper investigated principled strategies for estimating the global complexity level of full-length books in Portuguese and French under realistic sampling constraints. Departing from purely passage-oriented readability assessment, we examined how different sampling densities and aggregation operators affect the robustness of book-level complexity estimation.

Across both languages, results show that reliable global estimates can be obtained from a limited number of evenly distributed excerpts. Increasing the number of samples beyond a moderate threshold yields diminishing returns, indicating that lightweight sampling protocols are both feasible and methodologically sound. However, aggregation strategy plays a decisive role. While central-tendency measures (mean, median) provide stable approximations for automatic settings, aggregation functions that emphasise distributional extremes or upper quantiles prove particularly informative when aligning with human judgments, especially in French.

The findings support a distributional view of book-level readability, whereby global complexity emerges from the interplay between representative and peak difficulty segments. Importantly, the experiments demonstrate that computationally efficient, sampling-aware procedures can approximate full-text analysis without requiring complete textual access, making them suitable for deployment in Adult Learning scenarios and large-scale database construction.

A limitation of the present study lies in the relatively small number of books per language, which constrains the statistical generalisation of the findings and calls for further validation on larger and more diverse corpora.

Future work should further investigate cross-linguistic generalisation, annotation biases, and the interaction between sampling design and model architecture, as well as explore alternative sampling strategies, including randomised selection across the document, and the interaction between sampling design and book length. These aspects were beyond the scope of the present study but consti-

tute important directions for extending the proposed framework. Nonetheless, the present study provides empirical evidence and practical guidelines for moving beyond passages towards scalable and interpretable book-level readability assessment.

5. Acknowledgements

This work was supported by the European Commission (Project: iRead4Skills, Grant number: 1010094837, Topic: HORIZON-CL2-2022-TRANSFORMATIONS-01-07, DOI: 10.3030/101094837), and by Portuguese national funds through FCT (References: UID/50021/2025 (DOI: 10.54499/UID/50021/2025), UID/PRR/50021/2025 (DOI: 10.54499/UID/PRR/50021/2025) and UID/03213/2025 - CLUNL, DOI: 10.54499/UID/03213/2025). We gratefully acknowledge the contribution of the human annotators of the Portuguese corpus: S. Barbosa, R. Monteiro, I. Müller, M. Moutinho, and S. Reis.

5.1. Ethical considerations

The iRead4Skills system[4] ensures a proportionate, legally sound, and socially beneficial use of technology in support of literacy and the European publishing ecosystem. It does not infringe copyright or prejudice the normal exploitation of works.

The system performs automated analysis of texts solely for the purpose of assessing text complexity, involving at most temporary and technically necessary acts of reproduction. The submission of titles to the iR4S database is carried out through a back-office interface accessible only to authorised publishers, librarians, and project team members. Submitted text excerpts are never displayed or made publicly accessible and are immediately discarded after processing. Extracted text is processed in a stateless computational environment, held only transiently in working memory for textual and linguistic analysis, and automatically deleted upon completion of the process. No text, excerpts, images, or other expressive elements of the work are stored, indexed, cataloged, or transmitted. The only persistent output is a non-expressive metadata label indicating the degree of complexity associated with the excerpts and the corresponding title.

Under the law[5], acts of reproduction that are transient or incidental, that constitute an integral and essential part of a technological process, and that have no independent economic significance fall outside the scope of the reproduction right.

Non-textual information—such as bibliographic metadata, the assigned complexity level, and the identifier of the contributing user—may be retained for the purposes of tracking, auditing, and analysis.

Users may, at any time, elect not to include a given title and its associated complexity assessment in the database or request their deletion.

6. Lay summary

Choosing books for adults with low reading skills is not easy. This is especially true when we want to judge the difficulty of a whole book, not just a short text. A few short passages do not always show how easy or hard the full book really is.

In this paper, we study how to estimate the difficulty of full books in Portuguese and French. We test different ways of choosing short passages from a book and combining the results. We also use computer tools to analyse language and estimate how hard a text is to read.

The books are grouped into four difficulty levels, from very easy (L1) to more complex (L4). This scale was designed for adult learners and follows widely used international standards.

Our goal is to help build a public database of books with reliable difficulty information. This can help adult learners, teachers, librarians, and publishers choose reading materials that better match readers' needs.

7. Bibliographical References

Wafa Aissa, Raquel Amaro, David Antunes, Thibault Bañeras-Roux, Jorge Baptista, Alejandro Catala, Luís Correia, Thomas François, Marcos Garcia, Mario Izquierdo-Álvarez, Nuno Mamede, Vasco Martins, Miguel Neves, Eugénio Ribeiro, Sandra Rodriguez Rey, and Elodie Vanzeveren. 2025a. The iRead4Skills intelligent complexity analyzer. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 73–84, Suzhou, China. Association for Computational Linguistics.

Wafa Aissa, Thibault Bañeras-Roux, Elodie Vanzeveren, Lingyun Gao, Rodrigo Wilkens, and Thomas François. 2025b. Assessing French readability for adults with low literacy: A global and local perspective. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20517–20539, Suzhou, China. Association for Computational Linguistics.

[4] https://v1.iread4skills.com/

[5] Directive 2001/29/EC: https://eur-lex.europa.eu/eli/dir/2001/29/oj/eng)

Simon Allan, Marie McGhee, and Rob van Krieken. 2005. Using readability formulae for examination questions. Technical report, Qualifications and Curriculum Authority.

Raquel Amaro, Susana Correia, Ricardo Monteiro, Alice Pintard, Michell Moutinho, and Sílvia Barbosa. 2025. Framework of textual complexity for low-literacy adults: Levels and descriptors within the iread4skill project. *Langues & Parole*, 10:57–119.

I. Beltagy, M. E. Peters, and A. Cohan. 2020. Longformer: The long-document transformer. *arXiv: Computation and Language*, pages 1–17.

Rebekah G. Benjamin. 2012. Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review*, 24(1):63–88.

Jeanne S. Chall and Edgar Dale. 1995. Readability revisited: The new dale–chall readability formula. *Brookline Books*.

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. *ITL – International Journal of Applied Linguistics*, 165(2):97–135.

Council of Europe. 2020. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment – Companion Volume*. Council of Europe Publishing, Strasbourg.

Scott A. Crossley, Andrea Heintz, Jae Sung Choi, Jesse Batchelor, Mehdi Karimi, and Adam Malatinszky. 2023. A large-scaled corpus for assessing text readability. *Behavior Research Methods*, 55:491–507.

Edgar Dale and Jeanne S. Chall. 1948. A formula for predicting readability. *Educational Research Bulletin*, 27(1):11–20, 28.

Tovly Deutsch, Masoud Jasbi, and Stuart M 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.

Lijun Feng, Noémie Elhadad, and Matt Huenerfauth. 2010. Comparison of features for automatic readability assessment. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, pages 276–284.

Rudolf Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233.

T. François and C. Fairon. 2012. An "AI readability" formula for French as a foreign language. In *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing (EMNLP 2012)*, pages 466–477.

B. W. Lee, Y. S. Jang, and J. Lee. 2021. Pushing on text readability assessment: A transformer meets handcrafted linguistic features. In *EMNLP 2021*, pages 10669–10686.

Wenbiao Li, Rui Sun, Tianyi Zhang, and Yunfang Wu. 2024. Going beyond passages: Readability assessment for book-level long texts. In *Proceedings of the 23rd Chinese National Conference on Computational Linguistics (Volume 1: Main Conference)*, pages 1298–1309, Taiyuan, China. Chinese Information Processing Society of China.

G. Harry McLaughlin. 1969. Smog grading—a new readability formula. *Journal of Reading*, 12(8):639–646.

OECD. 2013a. *The Survey of Adult Skills: Reader's Companion*. OECD Publishing, Paris, France.

OECD. 2013b. Technical report of the survey of adult skills (piaac). Technical report, OECD Publishing, Paris, France.

OECD. 2021. The assessment frameworks for cycle 2 of the programme for the international assessment of adult competencies. Technical report, OECD Publishing, Paris.

Sarah E. Petersen and Mari Ostendorf. 2009. A machine learning approach to reading level assessment. *Computer Speech & Language*, 23(1):89–106.

Eugénio Ribeiro, Nuno Mamede, and Jorge Baptista. 2024a. Avaliação automática do nível de complexidade de textos em português europeu. *Linguamática*, 16(2):115–139.

Eugénio Ribeiro, Nuno Mamede, and Jorge Baptista. 2024b. Text readability assessment in european portuguese: A comparison of classification and regression approaches. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese (PROPOR 2024)*, pages 551–557.

Eugénio Ribeiro, David Antunes, Nuno Mamede, and Jorge Baptista. 2025. Exploring Few-Shot Approaches to Automatic Text Complexity Assessment in European Portuguese. *Journal of the Brazilian Computer Society*, 31(1):690–710.

Sarah 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, Ann Arbor, Michigan. Association for Computational Linguistics.

K. M. Sheehan, I. Kostin, Y. Futagi, and M. Flor. 2010. Generating automated text complexity classifications that are aligned with targeted text complexity standards. *ETS Research Report Series*, 2:1–44.

A. Jackson Stenner, Hal Burdick, Elizabeth E. Sanford, and David S. Burdick. 2006. How accurate are lexile text measures? *Journal of Applied Measurement*, 7(3):307–322.

Rodrigo Wilkens, Patrick Watrin, Rémi Cardon, Alice Pintard, Isabelle Gribomont, and Thomas François. 2024. Exploring hybrid approaches to readability: experiments on the complementarity between linguistic features and transformers. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2316–2331.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems*, volume 33, pages 17283–17297. Curran Associates, Inc.

8. Language Resource References

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Ricardo Monteiro, Raquel Amaro, Susana Correia, Alice Pintard, Roser Gauchola, Michell Moutinho, and Xavier Blanco Escoda. 2023. iRead4Skills - Complexity Levels. Technical report, Zenodo. Version 1.0.

Alice Pintard, Thomas François, Justine Nagant de Deuxchaisnes, Sílvia Barbosa, Maria Leonor Reis, Michell Moutinho, Ricardo Monteiro, Raquel Amaro, Susana Correia, Sandra Rodríguez Rey, Marcos Garcia González, Keran Mu, and Xavier Blanco Escoda. 2024. iRead4Skills Dataset 1: Corpora by Complexity Level for FR, PT and SP. Technical report, Zenodo. Version 2.1.

João Rodrigues, Luís Gomes, João Silva, António Branco, Rodrigo Santos, Henrique Lopes Cardoso, and Tomás Osório. 2023. Advancing Neural Encoding of Portuguese with Transformer Albertina PT-*. In *EPIA Conference on Artificial Intelligence*, pages 441–453. Springer.

A. Portuguese corpus

This appendix provides an overview of the Portuguese corpus (Table 5), reports the excerpt-level and aggregated complexity annotations for all extracted samples (Table 6), and presents the relationship between sampling density and approximation accuracy in terms of MAE and Spearman's ρ (Figure 1).

B. French corpus

This appendix provides an overview of the French corpus (Table 7), reports the excerpt-level and aggregated complexity annotations for all extracted samples (Table B), and presents the relationship between sampling density and approximation accuracy in terms of MAE and Spearman's ρ (Figure 2).

C. Technical details

The complexity assessment model used for Portuguese (Ribeiro et al., 2025) is a fine-tuned version of the smallest Albertina PT-PT model (Rodrigues et al., 2023), a transformer-based European Portuguese encoder. The model's classification strategy bases predictions on the weighted average of the probability distribution in contrast to simply selecting the most probable class.

For French passage readability assessment, we use the model proposed by Aissa et al. (2025a,b). The model is based on CamemBERT (Martin et al., 2020) and fine-tuned on a French corpus specifically designed for adults with low literacy levels.

Table 5: Portuguese corpus. Fragments from the texts with '*' have also been included in the iRead4Skills corpus (Pintard et al., 2024).

ID	Title	Level	#word	%
Text-01	A minha cidade é um livro	1	619	0,07
Text-02	Miguel e Sinatra	1*	3 205	0,34
Text-03	O paraiso sao os outros	1*	2 201	0,23
Text-04	O triunfo dos porcos	1	35 005	3,71
Text-05	Pageboy	2*	75 832	8,04
Text-06	O Principezinho	2	9 053	0,96
Text-07	Programa BE	2*	7 011	0,74
Text-08	Todos devemos ser feministas	2	11 129	1,18
Text-09	1984	3*	96 091	10,19
Text-10	Conto de fadas	3*	217 330	23,04
Text-11	Na sombra principe Harry	3*	162 634	17,24
Text-12	Programa ADN	3*	56 312	5,97
Text-13	Almoco de domingo	4*	58 697	6,22
Text-14	Mil anos de alegrias e tristezas	4*	120 766	12,80
Text-15	O remorso de Baltazar Serapião	4*	52 086	5,52
Text-16	Salario médio em Portugal	4*	35 198	3,73
		total	943 169	

Table 6: Excerpt-level and aggregated annotated complexity levels for each Portuguese text considering all samples extracted. M1 stands for Average, M2 stands for Max, M3 stands for Min, M4 stands for Median, M5 stands for Mode, M6 stands for Q25, and M7 stands for Q75.

	Excerpts										Overall Complexity (by metric)						
Book	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	M1	M2	M3	M4	M5	M6	M7
Text-01	1	2	X	X	X	X	X	X	X	X	2	2	1	1	1	1	1
Text-02	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
Text-03	1	3	3	2	2	2	2	2	X	X	2	3	1	2	2	3	3
Text-04	3	3	2	3	3	3	3	2	3	3	3	3	2	3	3	2	3
Text-05	3	2	3	3	2	3	3	3	2	2	3	3	2	3	3	2	3
Text-06	2	2	2	2	3	3	2	3	3	2	2	3	2	2	2	2	3
Text-07	1	2	3	3	3	3	2	2	3	2	2	3	1	2	3	2	3
Text-08	3	3	2	3	3	3	2	2	2	3	3	3	2	3	3	3	3
Text-09	3	3	3	3	3	4	3	4	3	3	3	4	3	3	3	3	3
Text-10	2	2	2	2	3	3	3	3	4	4	3	4	2	3	2	2	3
Text-11	2	3	3	3	3	2	3	2	3	3	3	3	2	3	3	2	3
Text-12	3	4	4	4	3	4	3	4	3	4	4	4	3	4	4	3	4
Text-13	3	2	4	3	3	2	2	3	2	2	3	4	2	2	2	2	3
Text-14	3	4	4	4	4	3	4	3	3	3	4	4	3	3	3	3	4
Text-15	3	4	3	3	3	4	3	4	3	3	3	4	3	3	3	3	3
Text-16	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4

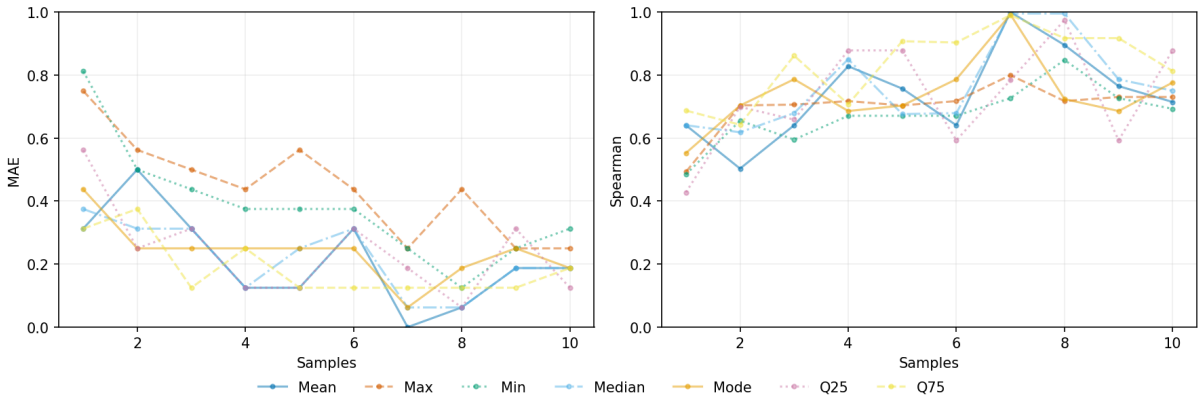

Figure 1: MAE and Spearman's ρ as a function of the number of sampled excerpts (S) for Portuguese books.

Table 7: French corpus.

ID	Title	Publisher	#word	%
Text-01	15 jours pour réussir	Didier	7 701	3.18%
Text-02	5 Contes	Hachette	15 464	6.39%
Text-03	Carmen	Hachette	11 803	4.88%
Text-04	Carmen	De Boeck	7 577	3.13%
Text-05	Contes	Hachette	10 405	4.30%
Text-06	Lettres de mon moulin	De Boeck	4 783	1.98%
Text-07	Fantôme de l'Opéra	De Boeck	10 067	4.16%
Text-08	Germinal	Hachette	12 727	5.26%
Text-09	Julie est amoureuse	Hachette	3 696	1.53%
Text-10	La boîte en os	De Boeck	3 524	1.46%
Text-11	La disparition	Hachette	4 008	1.66%
Text-12	La fille qui vivait hors du temps	Didier	8 694	3.59%
Text-13	La nuit blanche de Zoé	Hachette	3 828	1.58%
Text-14	La tête d'un homme	Hachette	17 421	7.20%
Text-15	Lancelot	De Boeck	5 817	2.40%
Text-16	Le Roi Arthur	De Boeck	3 241	1.34%
Text-17	Le blog de Maia	Hachette	3 868	1.60%
Text-18	Le casque mystérieux	Didier	7 414	3.06%
Text-19	Le prisonnier du temps	Hachette	3 179	1.31%
Text-20	Le secret du vieil orme	De Boeck	8 159	3.37%
Text-21	Les Trois Mousquetaires	De Boeck	14 322	5.92%
Text-22	Les Misérables	Hachette	18 172	7.51%
Text-23	Lucas sur la route	Hachette	3 280	1.36%
Text-24	Maigret tend un piège	Hachette	18 595	7.68%
Text-25	Le mystère de la chambre jaune	De Boeck	12 213	5.05%
Text-26	Double assassinat dans la Rue Morgue	De Boeck	7 336	3.03%
Text-27	La Lettre volée	De Boeck	3 051	1.26%
Text-28	Tristan et Iseult	De Boeck	6 784	2.80%
Text-29	Une étrange disparition	De Boeck	4 885	2.02%
	Total		242 014	100.00%

Table 8: Excerpt-level and aggregated annotated complexity levels for each French books considering all samples extracted. M1 stands for Average, M2 stands for Max, M3 stands for Min, M4 stands for Median, M5 stands for Mode, M6 stands for Q25, and M7 stands for Q75.

	Excerpts										Overall Complexity (by metric)						
Book	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	M1	M2	M3	M4	M5	M6	M7
Text-01	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
Text-02	3	3	2	2	3	3	3	3	3	3	3	3	2	3	3	3	3
Text-03	3	3	3	3	3	3	2	2	2	3	3	3	2	3	3	2	3
Text-04	4	3	3	3	3	4	4	3	4	4	4	4	3	4	3	3	4
Text-05	2	3	2	2	2	2	2	2	2	2	2	3	2	2	2	2	2
Text-06	3	2	3	3	3	3	3	3	3	3	3	3	2	3	3	3	3
Text-07	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
Text-08	3	2	3	3	3	3	3	3	3	3	3	3	2	3	3	3	3
Text-09	3	2	3	2	2	2	2	2	2	2	2	3	2	2	2	2	2
Text-10	3	3	3	3	3	3	3	3	4	3	3	4	3	3	3	3	3
Text-11	2	2	2	2	3	2	2	2	2	3	2	3	2	2	2	2	2
Text-12	2	3	2	2	2	3	3	3	3	3	3	3	2	3	3	2	3
Text-13	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
Text-14	3	3	2	3	2	3	2	2	3	3	3	3	2	3	3	2	3
Text-15	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
Text-16	3	3	3	3	3	3	4	3	3	3	3	4	3	3	3	3	3
Text-17	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
Text-18	2	2	2	2	2	2	3	3	3	2	2	3	2	2	2	2	3
Text-19	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
Text-20	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
Text-21	3	4	3	3	3	3	4	4	3	3	3	4	3	3	3	3	4
Text-22	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
Text-23	3	2	2	3	3	3	3	3	2	2	3	3	2	3	3	2	3
Text-24	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
Text-25	4	3	3	3	4	4	3	3	3	3	3	4	3	3	3	3	4
Text-26	4	4	4	4	3	4	4	3	3	4	4	4	3	4	4	3	4
Text-27	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
Text-28	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
Text-29	2	2	2	2	2	3	2	3	3	3	2	3	2	2	2	2	3

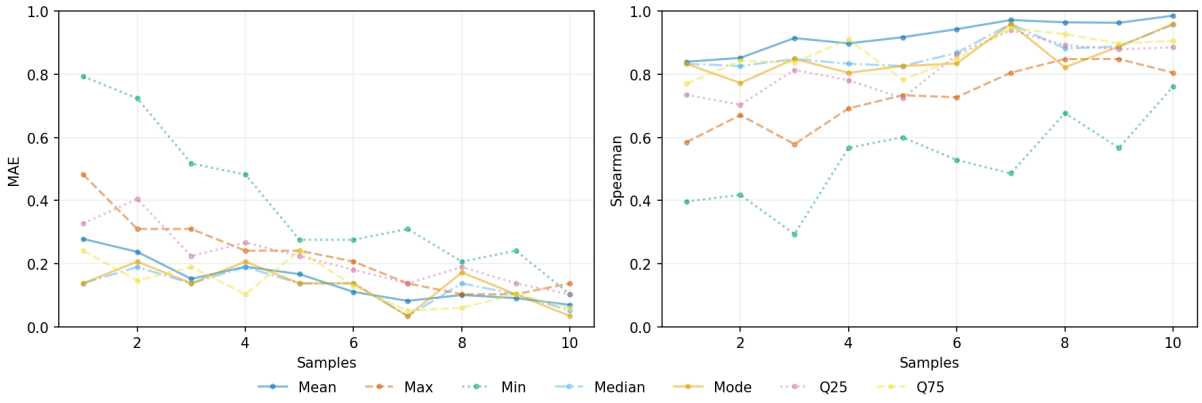

Figure 2: MAE and Spearman's ρ as a function of the number of sampled excerpts (S) for French books.