# TOWARDS SENTENCE-LEVEL TEXT READABILITY ASSESSMENT FOR FRENCH

Duy Van Ngo, Yannick Parmentier

van-duy.ngo@loria.fr, yannick.parmentier@loria.fr

**Abstract**

In this paper, we report on some experiments aimed at exploring the relation between document-level and sentence-level readability assessment for French. These were run on an open-source tailored corpus, which was automatically created by aggregating various sources from children's literature. We also report on sentence readability scores obtained when applying both traditional arithmetic approaches and state-of-the-art deep learning techniques. Results show a relatively strong correlation between document and sentence-level readability, suggesting ways to reduce the cost of building annotated sentence-level readability datasets.

# Introduction

Text readability assessment can be defined as the ability to automatically estimate the difficulty for someone to understand a given text. This language understanding task can, in some contexts, take the form of a classification task that matches a given text with a label of complexity. Fine-tune CamemBERT model to assess the readability level of a text. This temporarily reduces the assessment into a classification/ranking problem. There exists a common bottleneck in machine learning-based text readability assessment that lies in the scarcity of annotated resources for languages other than English such as French, thus the motivation. The main contribution of this work is the compilation of an open corpus for French, which has been preprocessed to remove noisy data and subsequently used to perform sentence-level automatic readability assessment with BERT architectures, giving results in line with those obtained at the document level [1]. [1]

# Methodology

## Corpus Construction

The target corpus is designated to be the consolidation of contents from French books available on the StoryWeaver[2] website under a creative commons licence. The website lists 1257 children stories in French that belong to 5 readability levels:

| Level | Word count | Other descriptions |
|---|---|---|
| 0 | $< 50$ | Familiar words, word rep. |
| 1 | $50 - 250$ | Easy words, word rep. |
| 2 | $250 - 600$ | Simple concepts |
| 3 | $600 - 1500$ | Longer sentences |
| 4 | $> 1500$ | Long & nuanced stories |

**Table 1:** StoryWeaver level description

To back up the claim that simpler texts tend to be more repetitive, we computed repetition rates for each of these levels. Results are given in Table 2 below.

| Level | #uniq. lemmas | #tokens | Rep. rate(%) |
|---|---|---|---|
| 0 | 852 | 4,076 | 20.90 |
| 1 | 4,919 | 63,255 | 7.78 |
| 2 | 8,776 | 157,372 | 5.58 |
| 3 | 10,318 | 192,137 | 5.37 |
| 4 | 9,014 | 128,440 | 7.02 |

**Table 2:** Repetition rate of unique lemmas

The books are filtered by readability levels before their information such as ID, level title, author, level, and translator (optional) are stored. The preprocessing then takes place to remove irrelevant information (such as cover page, credits pages, page numbers, etc.), then each document results in a string that contains targeted contents. The table below shows a few statistics regarding the compiled corpus:

| Level | #documents | #sentences | #tokens |
|---|---|---|---|
| All | 1,228 | 52,168 | 545,280 |
| 0 | 84 | 700 | 4,076 |
| 1 | 424 | 7,903 | 63,255 |
| 2 | 421 | 16,672 | 157,372 |
| 3 | 215 | 16,748 | 192,137 |
| 4 | 84 | 10,145 | 128,440 |

**Table 3:** Corpus level-based $x$-counts

[1]This work was financially supported by the French Scientific Research Center (CNRS) within the GramEx project.

[2]https://storyweaver.org.in/en/

# Results

**ARA with finetuned LLMs** To examine the distinctiveness of documents from different readability levels from a LLM perspective, we consider fine-tuning and evaluating CamemBERT models [2] with the corpus we obtained. We conduct experiments under two scenarios, attempting to decipher the correlation between document-level and sentence-level readability (keeping in mind that the distinctiveness of classes is a key factor). Due to the insignificant volume of data compared to other classes, the documents with level 0 are ignored.

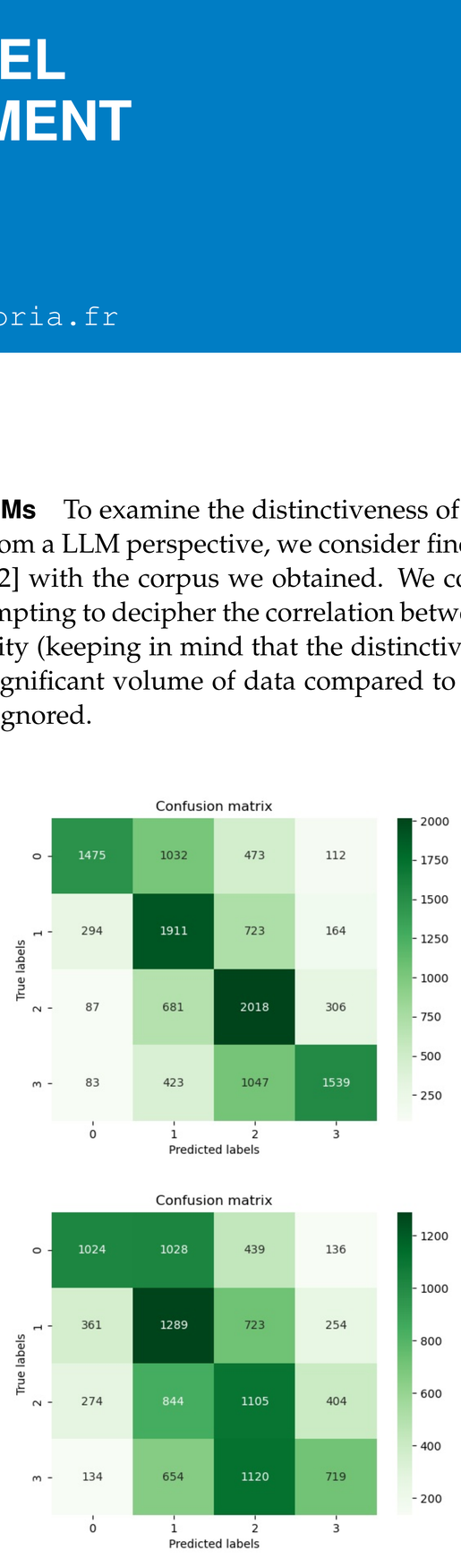


**Figure 1:** Classification results for two scenarios

| Class | General | | | Disjoint Sets | | |
|---|---|---|---|---|---|---|
| | Precision | Recall | F1 score | Precision | Recall | F1 score |
| 1 | 76.07 | 47.70 | 58.64 | 57.11 | 38.98 | 46.33 |
| 2 | 47.22 | 61.80 | 53.54 | 33.79 | 49.07 | 40.02 |
| 3 | 47.36 | 65.27 | 54.89 | 32.62 | 42.06 | 36.75 |
| 4 | 72.56 | 49.77 | 59.04 | 47.52 | 27.37 | 34.73 |

**Table 4:** Traditional scores for LLM classifier's results, general scenario

Experiments and error analysis show that the model is less likely to make disastrous (aka polemical) mistakes. On top of that, the models' result aligns with the intuition that from a document, each sentence should have a level in line with that of the entire document, and the divergence becomes less and less as the levels differ.

# Conclusion

In this paper, we presented a freely available corpus for French sentence-level text readability, which was automatically extracted from online resources and evaluated against state-of-the art deep-learning techniques. Results show some correlation between document-level and sentence-level readability assessment, which suggests that extending training corpora could be done by considering labelled documents, thus saving annotation costs.

# References

[1] N. Hernandez, N. Oulbaz, and T. Faine. Open corpora and toolkit for assessing text readability in french. In *Proceedings of the 2nd Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI) within the 13th Language Resources and Evaluation Conference*, pages 54–61, 2022.

[2] L. Martin, B. Muller, P. J. Ortiz Suárez, Y. Dupont, L. Romary, É. de la Clergerie, D. Seddah, and B. Sagot. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online, July 2020. Association for Computational Linguistics.