Noisy and Unbalanced Multimodal Document Classification: Textbook Exercises as a Use Case

Élise Lincker, Camille Guinaudeau, Olivier Pons, Jérôme Dupire, Céline Hudelot, Vincent Mousseau, Isabelle Barbet, Caroline Huron

Noisy and Unbalanced Multimodal Document Classification: Textbook Exercises as a Use Case

Élise Lincker
Cedric, CNAM
Paris, France
elise.lincker@lecnam.net

Camille Guinaudeau
JFLI, CNRS, NII
Tokyo, Japan
University Paris-Saclay
Gif-sur-Yvette, France
guinaudeau@nii.ac.jp

Olivier Pons
Cedric, CNAM
Paris, France
olivier.pons@lecnam.net

Jérôme Dupire
Cedric, CNAM
Paris, France
jerome.dupire@lecnam.net

Céline Hudelot
MICS, CentraleSupélec, University Paris-Saclay
Gif-sur-Yvette, France
celine.hudelot@centralesupelec.fr

Vincent Mousseau
MICS, CentraleSupélec, University Paris-Saclay
Gif-sur-Yvette, France
vincent.mousseau@centralesupelec.fr

Isabelle Barbet
Cedric, CNAM
Paris, France
isabelle.barbet@lecnam.net

Caroline Huron
SEED, Inserm, University Paris Cité
Paris, France
Learning Planet Institute
Paris, France
caroline.huron@cri-paris.org

ABSTRACT

In order to foster inclusive education, automatic systems that can adapt textbooks to make them accessible to children with Developmental Coordination Disorder (DCD) are necessary. In this context, we propose a task to classify exercises according to their DCD adaptation type. We introduce a challenging exercise dataset extracted from French textbooks, with two major difficulties: limited and unbalanced, noisy data. To set a baseline on the dataset, we use state-of-the-art models combined through early and late fusion techniques to take advantage of text and vision/layout modalities. Our approach achieves an overall accuracy of 0.802. However, the experiments show the difficulty of the task, especially for minority classes, where the accuracy drops to 0.583.

CCS CONCEPTS

• **Applied computing** → **Education**; *Interactive learning environments*; • **Computing methodologies** → **Natural language processing**; **Neural networks**.

KEYWORDS

multimodal document classification, noisy data, textbook adaptation, unbalanced data

CBMI 2023, September 20–22, 2023, Orléans, France

ACM ISBN 979-8-4007-0912-8/23/09…$15.00
https://doi.org/10.1145/3617233.3617239

ACM Reference Format:
Élise Lincker, Camille Guinaudeau, Olivier Pons, Jérôme Dupire, Céline Hudelot, Vincent Mousseau, Isabelle Barbet, and Caroline Huron. 2023. Noisy and Unbalanced Multimodal Document Classification: Textbook Exercises as a Use Case. In *20th International Conference on Content-based Multimedia Indexing (CBMI 2023), September 20–22, 2023, Orléans, France*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3617233.3617239

1 INTRODUCTION

Dyspraxia, called Developmental Coordination Disorder (DCD) in international classifications, affects 5% of children. This neurodevelopmental disorder is characterized as an impairment in motor coordination which affects academic and daily tasks. For instance, children with DCD might have difficulties with dressing, using cutlery, or riding a bike. They also do not automate the handwriting process and consistently depend on letter tracing when writing. At school, children with DCD face a cognitive dual task that hinders their ability to focus on more conceptual tasks while handwriting. In addition, their eye movement disorders may prevent them from reading a text if its presentation is not adapted to make it more accessible. To ensure inclusive education for children with DCD, textbooks used in class should take into account their handwriting and gaze organization difficulties and should be adapted in digital format in order to avoid handwriting, but without changing the content of the activities and their instructional intent.

Some non-profit organizations have started to produce adapted digital textbooks for children with DCD through manual transformations. Figure 1 shows an example of an exercise of the "Choix multiples" class (*Multiple-choice* in English) and its adaptation, allowing children with DCD to complete the sentence by *clicking* on the correct answer, avoiding the use of handwriting. Unfortunately, given the great diversity of textbooks and their frequent renewal

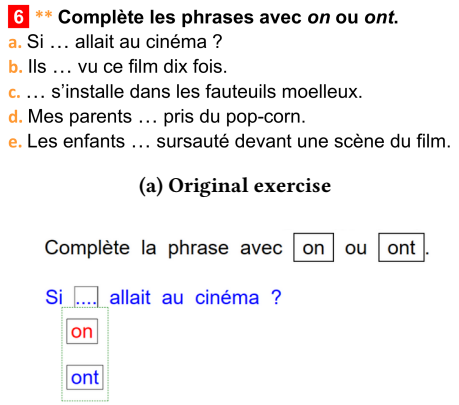

Figure 1: Fill-in-the-blank with multiple-choice options exercise and its adaptation. *Complete the sentences using "on" or "ont".*

due to changes in the curriculum, manual adaptation alone is not sufficient to meet the needs effectively.

Automatic adaptation of textbooks for children with DCD is a brave new task. We have developed a pipeline towards this automation, and in this paper, we specifically concentrate on classifying exercises based on their DCD adaptation type. To achieve this, we construct a dataset of French textbook exercises that have been manually annotated with adaptation type labels. This dataset poses several challenges, which reflect the difficulties of the task. First, the dataset is largely unbalanced, certain types of adaptation being much more frequent than others, and noisy as it may contain agrammatical or incomplete sentences, as well as errors stemming from the extraction process. Second, the classification objective concerns the educational intent of the exercises, which can be carried in very different ways. Finally, because of intellectual property issues, we have access to a limited number of textbooks, resulting in a relatively small dataset.

In this context, we tackle the task of exercise classification relying on semantic and layout information, merged through late and early fusion techniques. To this end, we rely on semantic dense representation, via the French language model CamemBERT [27] and visually-rich document understanding techniques. These approaches [35, 38], largely developed these last years for the understanding of articles, forms, letters or invoices, can be applied to our data. However, textbook exercises are often shorter and diverse in terms of content and layout; not only may some sentences be agrammatical, but also the layout and elements within the statement may vary: tables, lists, illustrations or scattered blocks of text.

Our main contributions are: (i) a new classification task, for textbook exercises automatic adaptation for children with DCD; (ii) a multimodal classification framework for the textbook exercises classification task; (iii) experiments with different multimodal architectures, including recently proposed LiLT [35].

The rest of the paper is organized as follows. Section 2 presents related work on textbook processing, textual and multimodal classification. Section 3 gives details about the constructed dataset, while Section 4 describes the architectures proposed for multimodal classification. In Section 5, we present the experimental results, analyzed in Section 6. Future research directions for automatic adaptation of textbook exercises for children with DCD are discussed in Section 7.

2 RELATED WORK

Literature on automatic Natural Language Processing (NLP) methods dedicated to textbooks is quite limited; existing studies focus on linguistic content analysis [9, 26] or question generation [3, 7]. If these approaches rely on content understanding and use various data representation techniques that can be beneficial for our classification objective, none of them deals with classification or content formatting tasks. However, some work has focused on the production of lexical resources that can be used in the context of textbook classification or representation. For example, Manulex [21] is a database that provides grade-level word frequency lists computed from 54 French elementary school readers. Two other resources have been introduced for text simplification in the context of accessible reading for dyslexic children: ReSyf [2], a lexical resource of monolingual synonyms ranked according to their difficulty to be read and understood by native learners of French, and Alector [6], a parallel corpus of 79 elementary grade reading texts that have been simplified at the lexical, syntactic, and discourse levels. More particularly related to inclusive education, the association *Le Cartable Fantastique*[1] provides *the Fantastiques Exercices*, a collection of French exercises and their interactive version adapted for children with DCD.

Document classification, on the contrary, is a very active field of research which has seen its results greatly improved by recent deep learning approaches. The Transformer model, in particular, with its self-attention mechanism [34] has proved to be very successful in many tasks including classification. A number of pre-trained language models have thus been developed, such as ELMO [29], ULMFiT [13], OpenAI GPT [31], and BERT [5]. Though many NLP resources are made from and for English content, some studies have released pre-trained language models for other languages. Multilingual models are also available, but their performance is not comparable to that of specific-language pre-trained models. For French, two models based on RoBERTa [25] are trained and optimized: CamemBERT [27], trained on the French part of OSCAR [33] and FlauBERT [20], trained on 24 corpora of various styles collected from the internet. These state-of-the-art French pre-trained language models have been fine-tuned on various data for text classification tasks, such as tweets classification [19] or clinical notes classification [4]. Finally, studies on automatic speech transcriptions [14] and digitized texts with optical character recognition (OCR) [18] analyzed the impact of noisy inputs on contextualized word embeddings. Indeed, erroneous data, such as ungrammatical sentences found in many exercise statements, may cause a decline in performance due to model resilience. Experiments showed that classifiers built with fine-tuned language models perform better than those built with pre-trained language models.

[1] https://www.cartablefantastique.fr/

While many classification approaches rely on text as a single modality, some recent studies focus on Visual Document Understanding (VDU). By encoding visual and layout information, Text-Image-Layout transformers have demonstrated promising performance on downstream tasks, including structured and visually rich document classification. LayoutLM [36] modified the BERT architecture by adding 2-D position and visual embeddings along with text embeddings. Two upgraded versions [15, 38] have been recently proposed, as well as a multilingual extension [37]. BROS [12] proposes a new spatial encoding method, using relative positions between blocks, instead of the absolute 2–D positions used in most previous works. DocFormer [1] also relies on a BERT-like masked language model and introduces a multimodal cross-attention mechanism, allowing information to be shared across modalities. Lastly, TILT [30] uses the T5 [32] architecture along with convolutional features. However, most models are pre-trained and fine-tuned on single-language documents, typically English. To address this issue, LiLT [35] allows to plug-and-play any pre-trained RoBERTa-like model with a layout module and thus benefit from the pre-training of document layout structure while being applicable on any language. Besides, most applications of Text-Image-Layout models involve whole pages and well-formatted documents: receipts in the CORD [28] and SROIE [16] datasets or forms in FUNSD [17] for example. Though they may perform well on textbook pages understanding, we focus here on short individual exercises that share many similarities with each other.

3 DATASET

In this section, we describe our dataset construction process, based on 3 elementary grade French textbooks in PDF format. Unfortunately, we are unable to release our dataset due to intellectual property issues, but we plan to share other scholarly data publicly in the future.

3.1 Text structure extraction

Each textbook is first parsed to an XML file in ALTO format using pdfalto[2] and MuPDF[3] tools. This approach allows for the extraction of *words* in a structured and organized representation of the content, while also providing layout and font style information. The extracted words are initially grouped into text segments based on rules involving font types, font sizes, character spacing and character types such as numbers, symbols and punctuation marks. These text segments are then semi-automatically organized into sections based on their dominant font. We use an annotation interface designed for MALIN to reorganize the segments into a textbook structure, enabling manual tagging of font styles with roles (chapter title, lesson, exercise, etc.). We define 6 subsections under the exercise section: number, title, instruction, statement, example and hint. The layout complexity makes the extraction task more challenging: pages may contain tables, lists, illustrations or scattered text blocks, which are harder to parse and may induce noise in the data. This extraction step results in 2748 exercises.

[2]https://github.com/kermitt2/pdfalto
[3]https://github.com/ArtifexSoftware/mupdf

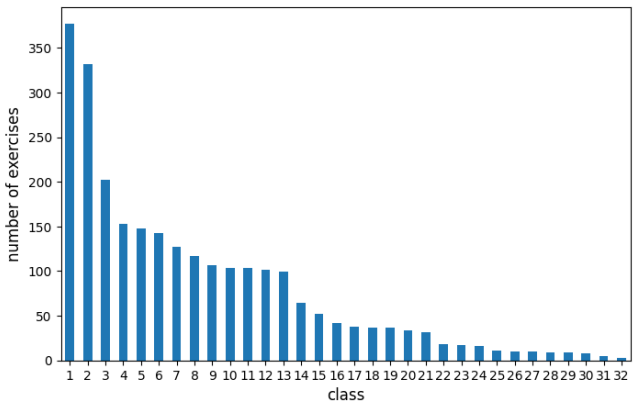

Figure 2: Distribution of exercises by class.

3.2 Exercise annotation and preprocessing

Extracted exercises are then manually annotated by two DCD experts. 33 classes are defined by experts, reflecting the combination of the learning objective of the exercise and the interaction process involved in its resolution (e.g. *multiple-choice*, *sentence transformation*, *letter/word/sentence ticking*, *written expression* etc.). For example, Figure 1 illustrates an exercise from the *multiple-choice* class and Figure 7 in the appendix contains 5 exercises from other classes. From this annotation, 2567 exercises are single-labeled, 146 are multi-labeled (with an average of 2.2 classes), and 36 exercises must be either manually adapted or removed from adapted textbooks. In this work, we focus on single-labeled exercises and exclude the least represented class that contains only one exercise. Thus, we end up with a dataset of 2566 exercises labeled with 32 classes.

Figure 2 shows the unequal distribution between the classes. The dataset is strongly unbalanced: 2 classes out of 32 have over 300 exercises, whereas 11 classes have less than 20. The 21 most populated classes represent 95% of the dataset.

Exercises are split into 3 subdatasets: training (70%), validation (10%) and test (20%). Proportions between classes in subsets and textbooks are maintained. Exercise numbers and titles are removed. If there are any examples and hints, they are concatenated with the instruction, which forms the first part of the input. The second part is the exercise statement. The text is normalized to lower case and all whitespace characters are reduced to one space.

3.3 Dataset characteristics

To better fit the data, we expand the definition of *sentence* and *token*. If most of the instructions are grammatically and semantically correct, statements may contain single words scattered around or ungrammatical word sequences, as shown in Figure 1. Therefore, statements' sentences can contain meaningless character sequences such as fill-in-the-blank words (*"c…bat"*), sentences (*"Manon a perdu … chat."*), multiple-choice choices (*"(son/sont)"*), concatenated words (*"irageâgégéantenfant"*), scattered blocks of text (*"est une fleur"*, *"la tulipe"*), list numbers (*"a."*, *"b."*), etc.

Figure 3 presents the token length, the number of tokens per sentence, the number of tokens per exercise, and the number of

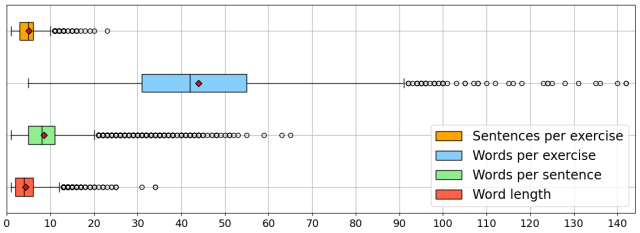

Figure 3: Exercise, sentence and word length distribution on the whole dataset.

sentences per exercise in our dataset. The average length of an exercise is 5 sentences and 44 tokens, but there is a significant variability. Most documents are 1-10 sentences and 5-91 tokens long. The dataset is, therefore, composed of short documents compared to the reference datasets[4]. Moreover, depending on the type of exercise, the word and sentence lengths are highly variable. A quarter of the sentences has between 1 and 5 tokens, while the longest sentences contain up to 65 tokens.

4 METHODOLOGY

Following the extraction of the textbook content, we aim to categorize the extracted exercises according to their type of adaptation to DCD. We propose different approaches using three pre-trained models fine-tuned for document classification task.

4.1 Single model approaches

First, we use CamemBERT. To further enhance the model, we fine-tune its masked language model on the following educational texts: lessons and activities from 4 textbooks (the 3 textbooks we used to build our dataset, apart from the exercises of the validation and test subsets, and a 4th unannotated textbook), 1293 *Fantastiques Exercices*, and the 79 original reading texts from Alector, of about 300 words each. For training, the model is fed with the concatenation of instruction and statement, separated by the special token <sep>[5].

To learn from other modalities, we use LayoutLMv2, the second version of LayoutLM. We feed textual, positional and visual features into the transformer. The model is pre-trained on IIT-CDIP [22] and fine-tuned for classification on its RVL-CDIP [11] subset, composed of scanned document images such as letters, forms, invoices, advertisements, etc. Most documents use English as their primary language, but IIT-CDIP contains a few documents in other languages, including French. If LayoutLMv2 processes text, layout and vision embeddings, the model is designed for English documents and may not fully benefit from text features.

4.2 Early and late fusion approaches

As classes in our dataset contain a high variability in terms of semantic and layout, we assume that French language as well as layout and vision modalities are relevant. In order to take advantage of each of these modalities, two fusion approaches are implemented.

A first solution is to experiment score-level fusion of CamemBERT and LayoutLMv2 classifiers. It consists in predicting the final label by considering the scores of the involved classifiers. We use Min-Max Normalization to transform the scores so that they are between 0 and 1 while preserving the relationships among the original predictions values. Then, we combine the normalized scores with common late fusion rules – *Average* and *Maximum* –, and the final prediction is the class with the highest merged score.

In the second solution, we take advantage of the recently introduced model LiLT [35] allowing to plug-and-play any pre-trained RoBERTa-like model with a layout module. The model, pre-trained on IIT-CDIP, is combined to our version of CamemBERT, fine-tuned on textbooks and reading materials, to obtain a LayoutLM-like model for educational French. Eventually, we apply a Majority Vote fusion of CamemBERT, LayoutLMv2 and LiLT[CamemBERT] at the decision-level with LiLT[CamemBERT] as the default classifier.

4.3 Additional unbalanced data approaches

To cope with data imbalance, two techniques can be used to solve unbalanced multiclass classification problems: configuration of the loss function and sampling. Other data augmentation approaches to address data imbalance are discussed in Section 6, *Discussion*.

Concerning the loss function, cross-entropy loss is widely used for classification tasks, either on balanced or unbalanced datasets. The focal loss [24] was introduced as a dynamically weighted loss function suitable for class-imbalanced data in binary classification tasks. It was then extended to multiclass problems and showed promising results. Increasing focal loss γ parameter allows to put more focus on misclassified examples. We experimented this function with different γ values, ranging from 0 (which corresponds to weighted cross-entropy loss) to 5.

Regarding sampling approaches, undersampling our dataset, which is not only unbalanced but also small, would result in a loss of information. However, we took inspiration from ensemble learning methods to build multiple undersampled subdatasets by distributing the exercises of the majority classes, then fuse the outputs of the different models trained on these subdatasets.

4.4 Setups

We use the BASE architecture for each model. In our final experiments, we fine-tuned the models for 30 to 40 epochs with a batch size of 16. We use Adam optimizer, inverse frequency weighted cross-entropy loss, and the initial learning rate is selected from 1e-5 to 1e-4. Input max length is set to 256. Results on the test set are obtained with the fine-tuned models performing the best on the validation set.

5 RESULTS

Table 1 shows the results for the exercise classification. Taking into account the challenges posed by the dataset, every method achieves satisfactory results. Best performance is reached by LiLT encapsulating CamemBERT followed with a latter late fusion of all 3 models: the accuracy comes out to be 0.802. Overall, the 3 models are complementary.

For majority classes, LayoutLMv2 performs almost as good as CamemBERT, even though it does not capture semantic information

[4]For example, the FUNSD documents are 158 words long on average.

[5]We also experimented dual and siamese network architectures to process the instructions and statements separately by feeding two different inputs to the model. These are not presented in this paper as a single model performs better.

Model	Modalities	Language	Accuracy			Macro F1
			Total	Maj.	Min.	
Majority Class Baseline			0.147			0.008
CamemBERT	text-only	French	0.775	0.788	0.500	0.663
LayoutLMv2	text + vision + layout	English	0.708	0.722	0.250	0.487
CamemBERT + LayoutLMv2 (Max Fusion)	text + vision + layout	French / English	0.767	0.784	0.417	0.627
CamemBERT + LayoutLMv2 (Avg Fusion)	text + vision + layout	French / English	0.782	0.796	0.500	0.664
LiLT[CamemBERT]	text + layout	French	0.786	0.796	**0.583**	0.696
CamemBERT + LayoutLMv2 + LiLT[CamemBERT]	text + vision + layout	French / English	**0.802**	**0.813**	**0.583**	**0.714**

Table 1: Performance comparison of classifiers. Accuracy scores are provided for the entire test dataset (Total), the 21 majority classes (Maj.) and the 11 minority classes (Min.).

Model	Acc.	Macro F1
CamemBERT	0.747	0.653
CamemBERT + FT	**0.775**	**0.663**
CamemBERT + FT + focal loss	0.772	**0.663**
CamemBERT + FT + undersampling	0.730	0.653

Table 2: Performance comparison of CamemBERT-based classifiers. (FT = fine-tuning on an educational corpus, undersampling = fusion of multiple models trained on undersampled subdatasets)

Model	Accuracy	
	Known collection	Unseen collection
CamemBERT	0.763	0.655
LayoutLMv2	0.702	0.371
LiLT	**0.778**	**0.775**

Table 3: Classification performance comparison on different textbooks: intra- vs. inter-collection generalization.

	B^-L^-	B^-L^+	B^+L^-	B^+L^+
# of exercises	79	74	36	321
LiLT	16 (20%)	61 (82%)	20 (56%)	304 (95%)
Max Fusion	0	54 (73%)	16 (44%)	321 (100%)
Avg Fusion	0	59 (80%)	19 (53%)	321 (100%)

Table 4: Classification comparison of the 3 fusion strategies on exercises correctly ($^+$) and incorrectly ($^-$) classified by CamemBERT (B) and LayoutLMv2 (L).

as well as French models. This confirms that layout and vision modalities must not be overlooked. However, LayoutLMv2 leads to very poor performance for under-represented classes.

The increase in accuracy with the late and early fusion strategies is statistically significant compared to LayoutLMv2. In comparison with CamemBERT text-only classifier, both early fusion through LiLT and average late fusion strategy slightly improve the global accuracy. If the improvement does not seem significant, detailed scores on minority classes reveal a larger gap between simple CamemBERT and LiLT[CamemBERT] classifiers.

The results of complementary methods applied to CamemBERT classifiers are shown in table 2. Fine-tuning the masked language model on an educational corpus increases the accuracy from 0.747 to 0.775. This improvement was found to be statistically significant with a p-value of 0.015 using a t-test. However, techniques applied to cope with data imbalance were not successful. Although focal loss is claimed to be an improved version of weighted cross-entropy loss, it is not effective on our dataset. Depending on the γ settings, it leads to statistically equal or worse scores than using weighted cross-entropy loss. Moreover, undersampling a very small dataset is not efficient since valuable information is lost. At best, we reach an accuracy of 0.730 using undersampled subdatasets.

Lastly, additional experiments were conducted to evaluate the generalizability of the models intra- and inter-textbook collection. Two textbooks from the same collection were used to train the models, while the third textbook, from a different collection, was used for evaluation purposes only. Results, presented in Table 3, suggest that the generalization capacity of the models is greater for text features compared to layout features. Indeed, LayoutLMv2 generalizes poorly across collections and requires a larger amount of data than CamemBERT to achieve satisfactory results. Combination of both text and layout features through LiLT continues to outperform single-model approaches and demonstrates good generalization capabilities on separate collections.

6 DISCUSSION

Misclassification do not relate to specific classes. In fact, multiple classes share similar textual content and/or layout, which makes classification more challenging. Errors typically occur when an exercise of a given class shares traits with exercises from another class, like length, layout or semantic properties.

Table 4 compares the classifiers' predictions. From this table, it can be seen that the 3 fusion methods can improve the predictions. Although LiLT does not catch all exercises correctly predicted by CamemBERT and LayoutLMv2 individually, 20% of the exercises misclassified by both classifiers are corrected by LiLT.

Gains achieved by the late fusion strategies demonstrate that CamemBERT is more confident[6] and reliable than LayoutLMv2. Besides, even if LayoutLMv2 performs slightly worse than CamemBERT, it handles better exercises where the layout prevails over the semantic content. Indeed, it correctly categorizes 36 exercises that CamemBERT was unable to categorize, and about half of them are fixed by maximum and average fusion techniques. For instance, the exercise shown in figure 4 is labeled "Associe" (*combine* in English) as the pupil must click to associate the statement items. Most of "Associe" documents are correctly labeled by all classifiers because of the layout and the use of the verb *"associer"* in the instruction. However, when the verb *"associer"* is replaced by a synonym *"réunir"*, as depicted in Figure 4, the text-based classifiers are unlikely to identify the main feature they learned for this class.

On the contrary, for other exercises, textual content and semantics can carry more significance than layout. Consider the label "Classe" (*classify* in English) as an example. While most of exercises' statements contain a table and a list of items to be classified in that table (Figure 7c in the appendix), some exceptions lack a table (Figure 5). CamemBERT is thus better suited for this exercise as it likely identifies the semantics of the instruction using the *"classe"* keyword, whereas LayoutLMv2 makes a wrong prediction due to the absence of a table feature.

A final example is the exercise depicted in Figure 6, which was misclassified by both CamemBERT and LayoutLMv2 classifiers. However, it was rectified by merging the two modalities using LiLT. The misclassification is likely due to the presence of ellipses and rectangular frames. Such features are commonly used in exercises where students are required to fill in the blanks, without being given predetermined choices.

Finally, for 8 out of the 11 minority classes, precision achieves a perfect score of 1, indicating accurate recognition when these classes are identified. However, the variability in recall highlights the lack of sufficient training data. Also, 3 classes consistently remain unrecognized, resulting in both precision and recall scores of 0. Data imbalance remains a tricky problem which, for our classification purpose, requires data augmentation and a deeper understanding of the dataset for effective classification. Additionally, specific errors can be partially resolved by leveraging predefined rules to identify certain classes that are easily recognizable, such as *"True or false"* or *"Find/hide the odd one out"* exercises, by using the keywords *"vrai ou faux"* and *"intrus"*, respectively. We also attempted to generate artificial data through deep learning and lexical substitution, utilizing the educational resources ReSyf and Manulex. Unfortunately, the resulting automatically generated exercises were not sufficiently satisfactory and looked unnatural due to limited data availability and the specificity of language in textbooks, especially in the instructions. Regarding layout generation, it is indeed valuable exploring more advanced techniques such as generative adversarial networks (GANs) [8] and self-attention-based Transformer models. Notably, LayoutGAN [23] and LayoutTransformer [10] have demonstrated their effectiveness in document layout generation.

[6]The difference between the highest score and the next highest score is larger with CamemBERT compared to LayoutLMv2.

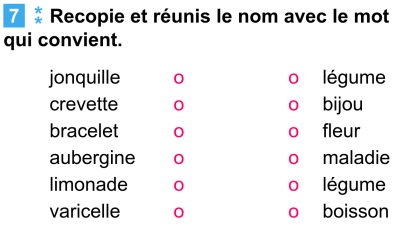

Figure 4: Example of an "Associe" exercise where layout prevails. *Copy and match the noun with the appropriate label.*

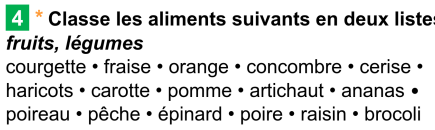

Figure 5: "Classe" exercise where text prevails. *Categorize the following foods into two lists: fruits, vegetables.*

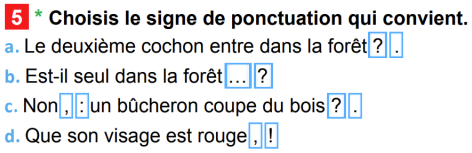

Figure 6: Example of a "Choix multiples" exercise. *Choose the appropriate punctuation mark.*

7 CONCLUSION

With a longer-term goal of automatically adapting full textbooks to make them accessible to children with DCD, we introduced in this paper a new education inclusive task: classification of textbook exercises according to their DCD adaptation type. We conducted a comparative study of document classification transformer methods on our own dataset composed of 2566 French textbook exercises.

We proposed different approaches based on three pre-trained models that achieved state-of-the-art performance on downstream tasks and reference datasets. We aimed to take advantage of different modalities and finally achieved an accuracy of 0.802 using fusion methods. The experiments demonstrated the importance of layout and vision modalities along with French educational language in textbook understanding.

In order to improve these promising results, our future work will focus on the extraction step and data correction. Moreover, if the overall accuracy is encouraging, results for minority classes still need to be improved. For this, we plan to continue our preliminary experiments on data generation to address the small data size and imbalance problems.

ACKNOWLEDGMENTS

The authors thank the reviewers for their constructive comments and suggestions, as well as Guillaume Faure for developing the extraction and annotation interface. This work was supported by the French ANR funded project MALIN (ANR-21-CE38-0014).

REFERENCES

[1] Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R Manmatha. 2021. Docformer: End-to-end transformer for document understanding. In *Proceedings of the 18th IEEE International Conference on Computer Vision*.

[2] Mokhtar B. Billami, Thomas François, and Núria Gala. 2018. ReSyf: a French lexicon with ranked synonyms. In *Proceedings of the 27th International Conference on Computational Linguistics*.

[3] Dhawaleswar Rao Ch and Sujan Kumar Saha. 2022. Generation of Multiple-Choice Questions from Textbook Contents of School-Level Subjects. *IEEE Transactions on Learning Technologies* (2022).

[4] Gabrielle Chenais, Cédric Gil-Jardiné, Hélène Touchais, Eric Tellier, Xavier Combes, Loick Bourdois, Philippe Revel, and Emmanuel Lagarde. 2022. Development and Validation of Deep Learning Transformer Models for Building a Comprehensive and Real-time Trauma Observatory. In *preprint*.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

[6] Núria Gala, Anaïs Tack, Ludivine Javourey-Drevet, Thomas François, and Johannes C. Ziegler. 2020. Alector: A Parallel Corpus of Simplified French Texts with Alignments of Misreadings by Poor and Dyslexic Readers. In *Proceedings of the 12th Language Resources and Evaluation for Language Technologies*.

[7] Krishnendu Ghosh. 2022. Remediating textbook deficiencies by leveraging community question answers. *Education and Information Technologies* (2022).

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27 (2014).

[9] Clarence Green. 2019. A multilevel description of textbook linguistic complexity across disciplines: Leveraging NLP to support disciplinary literacy. *Linguistics and Education* (2019).

[10] Kamal Gupta, Justin Lazarow, Alessandro Achille, Larry S. Davis, Vijay Mahadevan, and Abhinav Shrivastava. 2021. LayoutTransformer: Layout Generation and Completion With Self-Attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

[11] Adam W Harley, Alex Ufkes, and Konstantinos G Derpanis. 2015. Evaluation of deep convolutional nets for document image classification and retrieval. In *Proceedings of the 13th International Conference on Document Analysis and Recognition*.

[12] Teakgyu Hong, Donghyun Kim, Mingi Ji, Wonseok Hwang, Daehyun Nam, and Sungrae Park. 2022. Bros: A pre-trained language model focusing on text and layout for better key information extraction from documents. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence*.

[13] Jeremy Howard and Sebastian Ruder. 2018. Universal Language Model Fine-tuning for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.

[14] Chao-Wei Huang and Yun-Nung Chen. 2020. Learning ASR-robust contextualized embeddings for spoken language understanding. In *Proceedings of the 45th IEEE International Conference on Acoustics, Speech and Signal Processing*.

[15] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. LayoutLMv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia*.

[16] Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and C. V. Jawahar. 2019. ICDAR2019 Competition on Scanned Receipt OCR and Information Extraction. In *Proceedings of the 15th International Conference on Document Analysis and Recognition*.

[17] Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. 2019. Funsd: A dataset for form understanding in noisy scanned documents. In *Proceedings of the 15th International Conference on Document Analysis and Recognition Workshops*.

[18] Ming Jiang, Yuerong Hu, Glen Worthey, Ryan C Dubnicek, Ted Underwood, and J Stephen Downie. 2021. Impact of OCR Quality on BERT Embeddings in the Domain Classification of Book Excerpts. In *Proceedings of the Conference on Computational Humanities Research*.

[19] Diego Kozlowski, Elisa Lannelongue, Frédéric Saudemont, Farah Benamara, Alda Mari, Véronique Moriceau, and Abdelmoumene Boumadane. 2020. A three-level classification of French tweets in ecological crises. *Information Processing & Management* (2020).

[20] Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. 2020. FlauBERT: Unsupervised Language Model Pre-training for French. In *Proceedings of the 12th Language Resources and Evaluation Conference*.

[21] Bernard Lété, Liliane Sprenger-Charolles, and Pascale Colé. 2004. MANULEX: A grade-level lexical database from French elementary school readers. *Behavior Research Methods, Instruments, & Computers* (2004).

[22] David Lewis, Gady Agam, Shlomo Argamon, Ophir Frieder, David Grossman, and Jefferson Heard. 2006. Building a test collection for complex document information processing. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*.

[23] Jianan Li, Jimei Yang, Aaron Hertzmann, Jianming Zhang, and Tingfa Xu. 2019. LayoutGAN: Generating graphic layouts with wireframe discriminators. *arXiv preprint arXiv:1901.06767* (2019).

[24] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*.

[25] 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* (2019).

[26] Li Lucy, Dorottya Demszky, Patricia Bromley, and Dan Jurafsky. 2020. Content analysis of textbooks via natural language processing: Findings on gender, race, and ethnicity in Texas US history textbooks. *AERA Open* (2020).

[27] 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*.

[28] Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. 2019. CORD: a consolidated receipt dataset for post-OCR parsing. In *Proceedings of the Workshop on Document Intelligence at NeurIPS*.

[29] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

[30] Rafal Powalski, Lukasz Borchmann, Dawid Jurkiewicz, Tomasz Dwojak, Michal Pietruszka, and Gabriela Palka. 2021. Going Full-TILT Boogie on Document Understanding with Text-Image-Layout Transformer. In *Proceedings of 16th International Conference on Document Analysis and Recognition*.

[31] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, and others. 2018. Improving language understanding by generative pre-training. (2018).

[32] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research* (2020).

[33] Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In *Proceedings of the 7th Workshop on the Challenges in the Management of Large Corpora*.

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Proceedings of the 21st Conference on Neural Information Processing Systems*.

[35] Jiapeng Wang, Lianwen Jin, and Kai Ding. 2022. LiLT: A Simple yet Effective Language-Independent Layout Transformer for Structured Document Understanding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

[36] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. LayoutLM: Pre-training of Text and Layout for Document Image Understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

[37] Yiheng Xu, Tengchao Lv, Lei Cui, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, and Furu Wei. 2021. LayoutXLM: Multimodal pre-training for multilingual visually-rich document understanding. *arXiv preprint arXiv:2104.08836* (2021).

[38] Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. 2021. LayoutLMv2: Multi-modal Pre-training for Visually-rich Document Understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*.

APPENDIX

(a)

4 * **Recopie chaque liste sans l'intrus**
a. pépin - croquer - algues - éplucher - trognon
b. France - Allemagne - Paris - Italie - Espagne
c. coton - texte - étoffe - soie - cuir - tissu

Dans la liste, il y a un intrus. Cache-le.

pépin | croquer | algues | éplucher | trognon

(b)

11 * **Recopie chaque phrase en rétablissant la ponctuation comme dans l'exemple.**
aujourd'hui dans le jardin les petits cochons ont dansé
→ Aujourd'hui, dans le jardin, les petits cochons ont dansé.
a. ce matin en allant à la gare le troisième cochon a acheté du pain
b. à midi dans le train il s'assoit à côté d'un homme
c. pendant le voyage avec l'homme le petit cochon mange le pain
d. à l'arrivée dans la gare l'homme donne des briques au petit cochon

Recopie chaque phrase en rétablissant la ponctuation comme dans l'exemple.
aujourd'hui dans le jardin les petits cochons ont dansé
→ Aujourd'hui, dans le jardin, les petits cochons ont dansé.

a. ce matin en allant à la gare le troisième cochon a acheté du pain

→ a. ce matin en allant à la gare le troisième cochon a acheté du pain

(c)

2 ** **Classe les mots dans le tableau.**

Noms propres	Noms communs

Antoine · histoire · Italie · dire · Athènes · guerrier · perdre · casque · Méditerranée · feu · Hercule · flotter · demain · Paris

Classe les mots. Colorie les noms propres en jaune et les noms communs en rose.

Antoine | histoire | Italie | dire | Athènes | guerrier | perdre

casque | Méditerranée | feu | Hercule | flotter | demain | Paris

(d)

5 * **Associe chaque verbe conjugué au présent à son infinitif.**

ils détruisent	o	o se taire
vous fondez	o	o conduire
tu perds	o	o fondre
je me tais	o	o s'asseoir
nous nous asseyons	o	o perdre
il conduit	o	o détruire

Colorie l'infinitif du verbe conjugué au présent.

ils détruisent

se taire | conduire | fondre | s'asseoir | perdre | détruire

(e)

7 ** **Complète chaque phrase avec un groupe nominal de ton choix.**
a. … sont allées faire des courses.
b. … est venu pour les aider.
c. … sont arrivés à temps !
d. … est partie sans dire un mot.

DICTÉE À L'ADULTE : Complète la phrase avec un groupe nominal de ton choix.

☐ sont allées faire des courses.

☐ est venu pour les aider.

Figure 7: Examples of original and adapted exercises for classes (a) "Cache intrus" *Hide the odd one out*, (b) "Transforme Phrase" *Edit the sentence*, (c) "Classe" *Classify the items*, (d) "Associe" *Match the items*, (e) "Remplis au clavier" *Fill in using the keyboard.*