Layout- and Activity-based Textbook Modeling for Automatic PDF Textbook Extraction

Élise Lincker[1], Olivier Pons[1], Camille Guinaudeau[2,3], Isabelle Barbet[1], Jérôme Dupire[1], Céline Hudelot[4], Vincent Mousseau[4] and Caroline Huron[5,6]

[1]*Cedric, CNAM, Paris, France*
[2]*JFLI, CNRS, NII, Tokyo, Japan*
[3]*University Paris-Saclay, Gif-sur-Yvette, France*
[4]*MICS, CentraleSupélec, University Paris-Saclay, Gif-sur-Yvette, France*
[5]*SEED, Inserm, University Paris Cité, Paris, France*
[6]*Learning Planet Institute, Paris, France*

Abstract

Ensuring accessible textbooks for children with disabilities is essential for inclusive education. However, providing native accessibility for educational content remains a challenge. In the mean time, existing educational materials need to be adapted, for example by providing interactive versions to overcome difficulties caused by disabilities. In this context, our project aims to automatically adapt PDF textbooks to make them accessible to children with disabilities. The first step towards this adaptation involves extracting and structuring the content of textbooks. In this paper, we introduce textbook models, propose an automated extraction pipeline, and conduct preliminary experiments. Our textbook models are based on the various activities involved and provide layout and semantic information. They enable normalized and structured representations of educational content at both document and page levels, facilitating the automatic extraction process and the conversion to popular formats such as TEI and DocBook. In order to automatically extract PDF textbooks structure, our experiments, using a state-of-the-art multimodal transformer for a token classification task, demonstrate promising results. However, these experiments also highlight the difficulty of the task, especially cross-textbook collection generalization. Finally, we discuss the extraction pipeline and the directions of future work.

Keywords

textbook adaptation, inclusive education, modeling textbooks, modeling textbook pages, textbook extraction, interactive textbooks, digital textbooks, PDF processing

1. Introduction

The use of e-learning environment and e-textbooks is growing in higher education, yet paper textbooks remain prevalent in elementary and secondary schools in France. Despite the

availability of digital textbooks, the majority of them are not natively accessible due to their fixed layout and lack of reflowability, which prevents the adjustment of the page layout (font size, line spacing, word spacing, letter spacing, etc.). To ensure inclusive education, there is a pressing need to create accessible textbooks that allow children with disabilities to participate in classroom activities. Inclusive textbooks should take into account students' difficulties, while preserving the content of the activities and their instructional intent.

Some non-profit organizations have started to produce adapted digital textbooks by doing all the transformations manually. For example, the association *Le Cartable Fantastique*[1] provides *the Fantastiques Exercices*, a collection of French exercises and their interactive version adapted for children with Developmental Coordination Disorder (DCD). This neurodevelopmental disorder is defined as an impairment in motor coordination which interferes with academic achievement and daily life activities. At school, children with DCD struggle with handwriting. More specifically, they do not automate the handwriting process and continue to pay attention to letter tracing through their school life. In addition, their eye movement disorders may impede their ability to read text that is not presented in an accessible format. Hence, for children with DCD to succeed at school, textbooks must address their difficulties with handwriting and gaze organization. Figure 1 shows an example of a "Fill-in the blank" exercise and its adaptation, allowing children with DCD to complete the sentence by *clicking* on the correct answer, avoiding the use of handwriting.

6 ** **Complète les phrases avec *on* ou *ont*.**
a. Si … allait au cinéma ?
b. Ils … vu ce film dix fois.
c. … s'installe dans les fauteuils moelleux.
d. Mes parents … pris du pop-corn.
e. Les enfants … sursauté devant une scène du film.

(a) Original exercise

Complète la phrase avec on ou ont.
Si .... allait au cinéma ?
on
ont

(b) Adapted exercise

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

Unfortunately, the large variety of textbooks and frequent renewal due to changes in the curriculum make it challenging to adapt them manually. In this context, the MALIN project (*MAnuels scoLaires INclusifs*, French for *Inclusive textbooks*) aims to automatically adapt PDF textbooks for children with DCD or visual impairment, and in the long term, for other disabilities. Adapted textbooks facilitate inclusive participation in class, providing students with disabilities with the same educational content as their classmates. We do not aim to enrich textbooks with additional content or provide personalized assistance to students, but only modify the mode of interaction by generating alternative outputs or enabling connection to external tools (e.g. text-to-speech synthesis or braille displays for blind and low vision readers). Since there are no structured versions of textbooks that contain sufficient semantic information, we must start from textbooks in PDF format. Hence, the first and fundamental step towards our goal is to extract the textbooks' structure and content. This work presents our proposed approach for this extraction pipeline and concentrates on textbook modeling. We examined a large set of

---

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

French language study and mathematics textbooks used in elementary classroom to create a template of a textbook, and introduce two complementary textbook models: one that represents the textbook as a whole, and another that operates at the page level, which is essential for the extraction process. This paper also comes as a position paper on the automatic extraction task ahead. In particular, we have already conducted preliminary experiments on the token classification task, using hybrid transformer architecture.

Our main contributions are: (i) a model for structuring textbooks and textbook pages; (ii) a general approach for PDF textbooks extraction, (iii) preliminary token classification experiments.

2. Related work

2.1. Digital textbook envisioning

There are numerous ways to interact with digital textbooks. The simplest digital textbook includes the content of the paper textbook and can be flipped through like a traditional e-book. More advanced versions are enriched with additional multimedia material or internal functionalities such as hyperlinks. Researchers envision the future of textbooks as interactive learning environments rather than traditional books and promote adaptive, personalized and collaborative learning. Ou et al. [1] propose a pedagogical framework for designing and developing intelligent textbooks. Their framework is based on 5 key components: learners, text content, visual content, assessment and AI technologies, and integrates 4 learning strategies: multimedia learning, adaptive learning, personalized learning and collaborative learning. A similar vision [2] takes advantage of Adaptive Classroom Environment (ACE) and Adaptive Learning Recommandation System (ALRS) to encourage active dialog centered on structures activities. These learning environments could be natively accessible, or at least more easily adaptable. However, the implementation of such initiatives appears unlikely to happen soon in France, since the production process of publishers relies on paper textbooks. One issue arises from the fact that their digital versions are not accessible to children who have difficulties accessing visual information (blind, visually impaired, visual and motor coordination disorders) or writing (DCD, motor disabilities, autism specific disorders, attention disorders) because their formats are incompatible with assistive technology tools [3, 4]. These formats do not allow students with disabilities to access information, process content, or perform educational tasks effectively, efficiently and satisfactorily [5].

2.2. Textbook modeling

Textbook modeling is a fundamental step common to all textbook segmentation research. Three widely accessible markup schemes cover a range of applications, including textbooks: HTML, DocBook[2], and the Text Encoding Initiative Guidelines (TEI)[3]. Publishers and authors customize or combine existing schemes or create their own, tailoring them to their specific needs and objectives [6]. Hence, a basic but comprehensive textbook model [7] consistent with the TEI standards has been formalized in order to develop an ontology for the textbook research

---

[2]https://docbook.org/
[3]https://tei-c.org/

discipline. Many papers in the field of textbook digitization and extraction tend to adopt a generic structure (headings, sections, body text, paragraphs, etc.), whereas we aim to describe all the types of instructional activities present in textbooks. In this direction, conceptual guides for the elaboration of textbooks, such as [8], propose a standard structure and provide a range of elements to be considered to develop "a good textbook". The first pages of some textbooks also supply information on how to successfully use them, including details on structure, different types of activities, and helpful notes for the pupil or the supervising adult. All these various aspects should be taken into account when modeling textbooks.

2.3. Automatic PDF document structure extraction

2.3.1. Visually-rich document understanding

Emerging deep learning approaches have demonstrated significant potential in addressing natural language processing (NLP) tasks, particularly those related to visually-rich document understanding (VrDU). Most rely on the Transformer architecture, with its self-attention mechanism [9], extended to multimodal data. VrDU models involve combining textual, spatial and visual features to interpret scanned documents, PDFs and web pages. Thus, LayoutLM [10, 11, 12] is built upon BERT's architecture [13] and incorporates additional 2-D position and visual embeddings along with text embeddings. BROS [14] uses relative instead of absolute positions between blocks, and DocFormer [15] introduces a multimodal cross-attention mechanism enabling the exchange of information across modalities. TILT [16] relies on the T5 [17] architecture and provides additional contextualized image embeddings at the input. However, most NLP models are pre-trained and fine-tuned on English documents. While French pre-trained large language models CamemBERT [18] and FlauBERT [19] have been very efficient in many NLP tasks, no VrDU French model has been released. To tackle this issue, LiLT [20] allows to plug-and-play any pre-trained RoBERTa-like model with a layout module and thus leverage layout features for languages other than English. Besides, those models obtain state-of-the-art results on several downstream VrDU tasks, such as form and receipt understanding, respectively on FUNDS [21] and SROIE [12] benchmark datasets. Some studies focus on more complex document layout. For example, Najem-Meyer et al. [22] compares text-only, visual and multimodal models as well as 3 annotation standards for historical commentaries layout analysis. To the best of our knowledge, there has been no specific research conducted on the application of deep learning models, such as VrDU, to the analysis of textbook pages.

2.3.2. Automatic textbook processing

Research on textbook digitization, extraction and automatic analysis has been limited. Similar to MALIN, the Intextbooks [23] system transforms a PDF textbook into an interactive digital version, based on formal structure and hierarchy modeling. However, it targets university textbooks, which are very different from elementary and secondary school textbooks. They differ both in their format (linear, sober, no double columns, no illustrative images) and in their content (more conceptual knowledge than training exercises). Moreover, the ultimate objective of Intextbooks is to integrate smart interactive content by building enriched knowledge graphs [24, 25, 26, 27], while we aim at improving accessibility. Another method relies on the

identification and cutting of target areas followed by OCR [28] has been proposed for non-PDF electronic textbooks, where the text may not be easily extractable. This approach was specifically developed for textbooks, as OCR is effective for standard texts but not for documents with complex layout. Both this work and Intextbooks' extraction step rely on rules, using layout and table of content analysis, font styles, coordinates and distances, for example. Currently, rule-based approaches prevail over machine learning approaches for textbook extraction.

2.3.3. Analogous approaches: PDF processing

Despite limited research on textbook extraction, multiple techniques have been proposed for document segmentation or specific content identification in various PDF document types. This includes books and scientific papers. Many tools also employ rule-based methods. For instance, SEB [29] extracts books at both page level and document level, and [30] performs PDF-to-ePUB conversion. Both structures aim to provide a more comfortable reading experience by enabling a reflowed reading mode. More specifically applied to scholarly papers, hybrid methods were proposed to extract and organize documents, using layout and style rules as well as statistical machine learning algorithms [31]. Part of Semantic Scholar[4], the Semantic Scholar Open Data Platform [32] provides resources for scientific literature and the Semantic Reader Project [33] intends to create an intelligent, interactive and accessible reading interface. Their pipeline combines multiple PDF parsing tools, VILA [34], LayoutParser [35], as well as their own libraries[5]. Other works aim to identify specific objects contained in scientific paper, such as metadata of algorithms [36] or mathematical statements and results [37]. Experiments with the extraction leverage style-based rules, computer vision and NLP techniques.

3. Approach: using layout and conceptual textbook models for automatic textbook extraction

We propose activity-based textbook models with a mix of layout and conceptual features, and a PDF textbook extraction pipeline built upon these models. Our models result from the observation of dozens of elementary school textbooks widely used in French educational institutions, and were inspired by existing models and guidelines for textbook creation.

3.1. Textbook modeling: section and activity inventory

Most French language study and mathematics textbooks share a basic structure. Textbooks are divided into sub-disciplines (respectively grammar, conjugation, etc., and arithmetic, geometry, etc.), then into learning themes containing several chapters. The bulk of the textbook to be adapted for MALIN is found in the chapter pages. In the majority of textbooks, chapters are structured around a double-page spread and include the following content blocks: the chapter title, an introductory activity to the chapter, a lesson, and a series of exercises. An exercise is then divided into several parts: it always has an instruction, often a statement, and it may

---

[4]https://semanticscholar.org

[5]ScienceParse https://github.com/allenai/science-parse, PaperMage https://github.com/allenai/papermage

contain: a number, a title, examples, hints and illustrations, as well as additional indicators such as the level of difficulty. The introductory activity can assume diverse variations (exploratory activity, revision activity, etc.), typically comprising a single statement and several sub-exercises. In some books, opening activities introduce each theme. Depending on the publisher and the collection, other information may be included, such as the skills that are involved in completing an exercise, indicative activity headings, chapter numbers, a reminder of the discipline or the theme in which the chapter is located, various indicators of the modality or interdisciplinarity of an activity, etc. Revision pages are often added at the end of a theme or sub-discipline, with a summary of the learning content from a set of chapters as well as application and integration exercises. Finally, additional tool pages may be found at the beginning and end of the textbook: foreword, directions for use, pedagogical approach, preface, table of contents, index, glossary, bibliography, acknowledgements, pedagogical resources, or other various appendices.

3.2. Textbook modeling: XML model formalization

From this inventory of sections, we have developed two models: a whole textbook model and a textbook page model. We created our own Document Type Descriptors (DTDs) to formalize the structure in XML format. This new model precisely aligns with the abstract syntax tree of the document, encompassing all the necessary information for our project.

The first model captures the encapsulation of blocks as described in Section 3.1. Any additional element that is not part of the main content of a block is represented using an "indicator" tag.

We use tokens as the smallest linguistic unit and group them into text segments. In fact, textbooks have a distinct language compared to regular texts. In language study textbooks, tokens are sometimes bound morphemes; in mathematics textbooks, they can be operators, symbols or isolated letters. Besides, text segments usually correspond to grammatical sentences (starting with a capital letter and ending with a strong punctuation mark), bloc titles or labels. However, according to the nature of the activities, we may find ungrammatical and asemantic sequences. Some examples are fill-in-the-blank words (*"c...bat"*), sentences (*"Manon a perdu ... chat."*) or operations (*"4 × ... = 8"*), multiple-choice choices (*"(son/sont)"*), concatenated words (*"cirageâgégéantenfant"*), scattered blocks of text (*"est une fleur", "la tulipe"*), list numbers (*"a.","b."*), etc. As a result, referring to segments instead of sentences is more appropriate to describe such units of text. For the automation process, the use of text segments also allows to easily infer roles through typography, whereas this is not sufficiently discriminating at the token level. Moreover, due to the short length of the text blocks in the analyzed grade-level textbooks, the concept of paragraph is not relevant to our project.

The scheme is also extended to lists and tables. In addition, two or more lists can be linked, for instance in an exercise where the instruction asks to match items from different lists to each other (e.g., *"Match each subject with its predicate"*). As needed, we can further refine our model by incorporating sub-elements (e.g., choices in multiple-choice questions) and additional semantic and morpho-syntactic attributes. This refinement is possible despite the fixed general XML structure and layout attributes.

To ensure consistency with previous research on textbook modeling and to guarantee long-term usability, our pivot format can be converted to conform popular formats such as DocBook

and TEI. Table 1 shows the correspondence between our elements and their equivalents according to the TEI Guidelines. The documents in the appendix depict the exercise featured in Figure 1, in our format (Listing 1), converted to TEI (Listing 2) and to DocBook (Listing 3).

While modeling the entire textbook seems to be sufficient, the content extraction task is performed at the page level. Therefore, we define a different scheme to model each textbook page separately. In this second model, each token and segment tags will also be assigned position and style attributes, reflecting the layout and formatting of the textbook page. In addition, the nesting *discipline > learning theme > chapter > activity blocks within the chapter* is not straightforward. If chapter, theme or discipline titles appear on the page, these titles (along with potential indicators) constitute a block on their own, separate from the activity blocks on the page. As shown in Figure 2, this method enhances the visual representation of the blocks on the page and each text segment can easily be associated with a role (chapter title, lesson heading, lesson content, exercise number, exercise instruction, etc.) for the successful completion of the automatic extraction task.
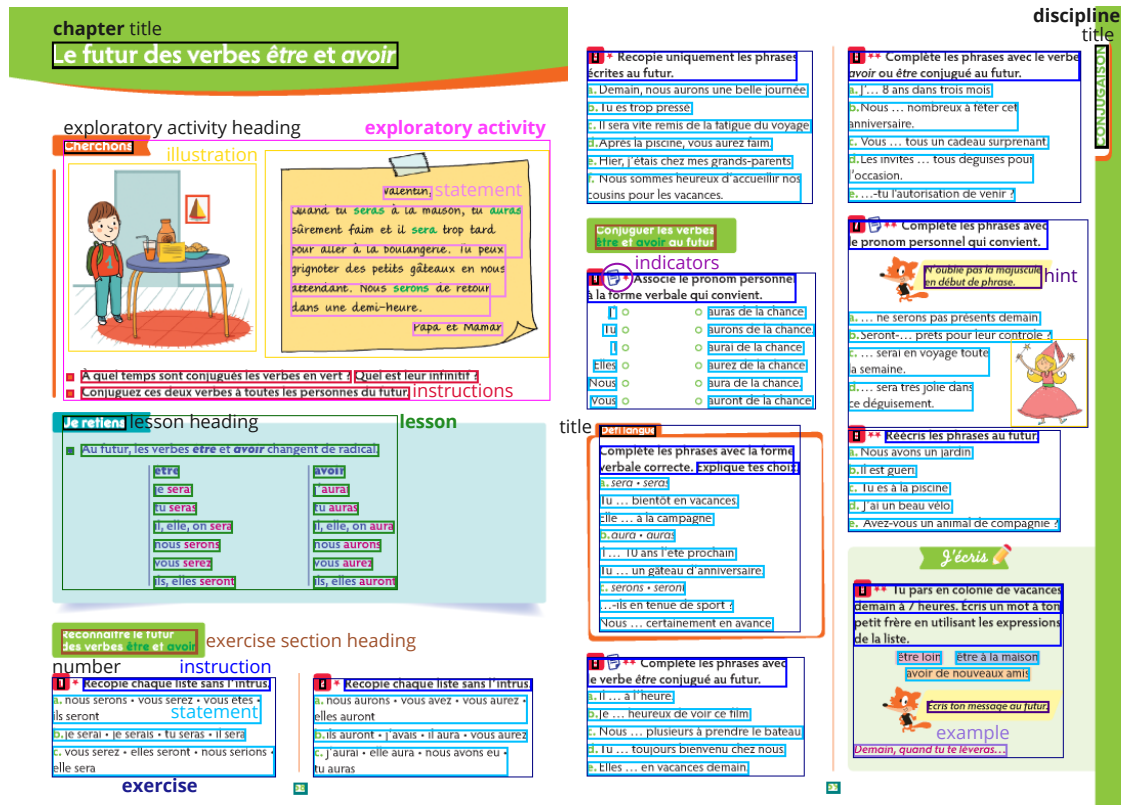

Figure 2: Structural visualization of a second grade French textbook chapter using our model.[6]
Source: [38]

---

[6]Note that some text segments (e.g. heading "J'écris" above the last exercise in Figure 2) are integrated into the background images and will require OCR along with our PDF extraction process.

Table 1
Alignment of our textbook model with the TEI guidelines.
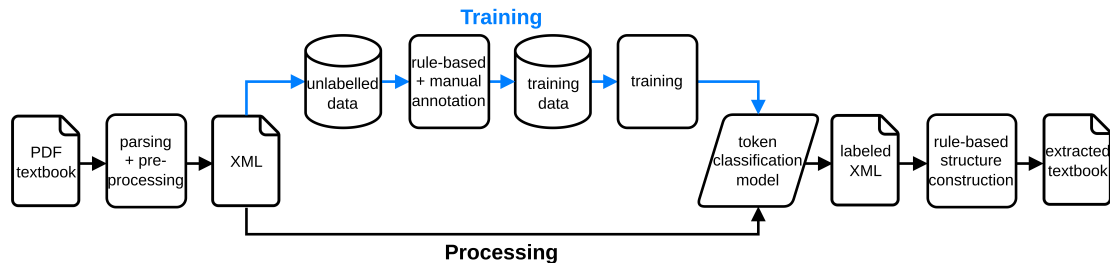A unique identifier (@id) and positional attributes (*@xmin, @ymin, @xmax, @ymax*) are assigned to all elements.
Attributes in italics correspond to layout and style characteristics added for the page extraction model.

Label	Element(s)	Attributes	TEI element(s)	TEI attributes
pagebreak	<pb/>		<pb/>	@n="[next page number]" @facs="[facsimile image of the page]"
linebreak	No linebreak element, lines can be identified with tokens' position attributes		<lb/>	
textbook division (discipline, learning theme, chapter, section, activity, lesson, subcomponent of activity)	<discipline> <theme> <chapter> <intro> <lessonsection> <exsection> <exsubsection> <ex> <instruction> <statement> <example> <hint>	@type: <intro type="open|explo|revision..."> <lesson type="vocabulary">	<div>	@type="discipline|theme|chapter|intro|lesson|exercises|exercise|instruction|statement|example|hint" @subtype="[subtype of activity]" @n="[theme, chapter or exercise number]"
bloc/section/activity heading	<heading> <title>		<head>	
page or textbook division number	<num>		No corresponding element, converted to attribute @n (<pb/>, <div>)	
indicator	<indicator>	@type="modality|difficulty|revision..."	<note>	@type
image	<img>	*@rotation* @filename="[file path]"	<figure> <graphic/>	@url="[file path]"
text segment	<seg>	*@font*	<seg>	
token	<token>	*@font* @sep="space|tab|crlf|empty" *@spnext= "[spacing to the next character]"*	<w> or <pc> or <number>	@join <pc join="left"> if preceding token element @sep="empty")
table	<table>		<table>	
table header	<th>		No corresponding element, converted to an attribute: <cell role="label">	
table caption	<caption>		<head>	
table row	<tr>		<row>	
table cell	<td>	@cols @rows	<cell>	@cols @rows
list	<list>		<list>	@rend="numbered|lettered|bulleted|inline"
list item	<item>		<item>	@n="[item number]"
bullet	<bullet>		No corresponding element, converted to an attribute: <item n="a.">	
list item separator	<sep>		No corresponding element, converted to an attribute: <list rend="inline bulleted">	
block of related lists	<listsection>		<list>	

3.3. Automatic textbook extraction

Due to the large quantity and diversity of textbook collections, an approach relying solely on rules would not be appropriate for our purposes because it would require extensive manual annotation. Our approach builds upon the model described in Section 3.2 and integrates both rule-based and deep learning methods. Figure 3 depicts the pipeline we are developing to convert a PDF textbook to its structured version.

Figure 3: MALIN structure extraction pipeline

Each textbook is first parsed to an XML file in ALTO format by pdfalto[7] coupled with MuPDF[8]. This combination of OpenSource tools enables the extraction of *words* along with their font style and spatial coordinates, as well as images, in a well-organized structure. The extracted words are tokenized and grouped into text segments using rules on font sizes and styles, character types (numbers, symbols, punctuation marks) and spacing between tokens and characters.

To reconstruct the textbook structure, segments must be labeled according to their role. In order to reduce the manual annotation workload, we utilize an annotation interface designed for MALIN. This web-based interface is supported by a TypeScript and Node.js back-end. The core idea is to map the XML ALTO file to an HTML format, enabling visual representation and annotation. Firstly, the annotator manually tags text segments with roles with just a few clicks. Segments are then semi-automatically labeled based on their dominant font. Secondly, the labeled segments are organized into higher-level categories that reflect the document's structure (e.g. lesson, exercise, etc.). This organization process leverages geometric features, font types, font sizes, spacing and text patterns. At both stages, the results of the automatic annotation are visually presented in HTML format, allowing for easy examination and potential corrections. This ensures the accuracy of the annotation process.

Once we have collected and processed enough textbook pages, we constitute a dataset to train and evaluate deep learning models to achieve this annotation task automatically. Textbook page structure extraction can be achieved through a token classification task. Preliminary experiments conducted on a few French textbook pages are described in Section 3.4. New pre-processed pages will then directly pass through the extractor.

Upon token classification, textbook pages are formatted into our desired structure. Sections are built using geometric features and logical sequencing. For example, an exercise number following a statement indicates the beginning of a new exercise section. The process also

---

[7]https://github.com/kermitt2/pdfalto
[8]https://github.com/ArtifexSoftware/mupdf

involves identifying lists and tables, and filtering images in the PDF file. Activity illustrations are matched with the corresponding sections using geometric features, while images used for aesthetic purposes are removed.

After textbook pages are structured, and possibly merged to our document-scale model or its TEI or DocBook equivalent, the resulting data can be used for various artificial intelligence applications in education. For our adaptation purposes and depending on the disability, we would then be able to produce digital textbooks with a custom layout and interactive adaptations.

3.4. Preliminary token classification experiments

These preliminary experiments correspond to the training phase of the extraction pipeline depicted in Figure 3. This step involves training a deep-learning model on a token classification task to automatically predict the role of each token in the document, enabling the reconstruction of the document structure.

3.4.1. Experimental setup

We constructed a dataset of textbook pages extracted from 1 elementary grade French textbook in PDF format. This constitutes a total of 167 pages, which are then split into 3 subdatasets: training (70%), validation (10%) and test (20%). For evaluation purposes only, we selected an additional 30 pages from a second textbook of the same collection, and 30 pages from a third textbook of a different collection. Each token is annotated with a coarse-grained page region label among: *discipline, chapter, heading, introductory activity, lesson, exercise, page number*. Future experiments will go further by introducing fine-grained labels. Table 2 lists the equivalences between coarse- and fine-grained page region classes.

Table 2
Coarse- and fine-grained classes used for layout annotation and token classification

Coarse	Fine
discipline	discipline_title, discipline_indicator
theme	theme_title, theme_number, theme_indicator
chapter	chapter_title, chapter_number, chapter_indicator
heading	heading_introductory activity, heading_lesson, heading_exercises
introductory activity	introductory activity_title, introductory activity_statement
lesson	lesson
exercise	exercise_number, exercise_title, exercise_indicator, exercise_instruction, exercise_statement, exercise_example, exercise_hint
page number	page number

In our first research on the classification of French textbook exercises according to their adaptation to DCD with multimodal transformers [39, 40], we demonstrated the importance of layout and vision modalities along with French educational language in textbook understanding. We therefore take advantage of recently introduced LiLT, combined with CamemBERT prior

fine-tuned on textbooks and reading materials[9], to obtain a LayoutLM-like model for educational French. We use the BASE architecture for both pre-trained models. Fine-tuning on the token classification task is completed at 10-15 epochs due to early stopping, with a batch size of 8. The initial learning rate is set to 1e-5. We use Adam optimizer and cross-entropy loss. Results on the test set are obtained with the fine-tuned model performing the best on the validation set.

Considering the supported limits of the models, we set the maximum input length to 512. However, 60% of the pages are longer. These pages are encoded in 2 overlapping segments: once by truncating the end of the document to the maximum length, and a second time by truncating the beginning. Inspired by the sliding-window approach [42], this solution allows to cover all the document, while maximizing the window size which also maximizes the context. During evaluation, predictions are generated for each segment and aligned to the entire textbook page. If the overlapping section between segments results in different predictions[10], the section is re-encoded with additional context tokens on both the left and right sides and passed through the model. The three predictions for the overlapping part are merged using a majority vote to obtain a single prediction. This approach ensures that the model's predictions are accurately consolidated for the entire page, even when there are variations in the overlapping segment.

3.4.2. Results and discussion

We report the accuracy and macro-F-measure of the token classification task in Table 3. The evaluation was performed on 3 textbooks, from both familiar and unfamiliar collections.

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

Textbook	Accuracy	Macro-F1
Majority Class Baseline	0.5779	
Known collection, know textbook	0.9934	0.9867
Known collection, unseen textbook	**0.9963**	**0.9873**
Unseen collection	0.8654	0.6168

Our model outperforms the majority class baseline on all test sets. The intra-collection performance is very high. Since layout is the same for textbooks of the same collection, these results highlight the significance of layout features for the model.

The performance scores are lower for the 3rd textbook. Upon closer examination of the predictions and comparing the pages with those from the textbook used for training, it becomes evident that the errors are primarily due to layout differences. Specifically, chapters in the training collection are typically structured as shown in the example in Figure 2. However, in the new textbook, exercises can be located before the corresponding lesson, which is consistently at the bottom of the page. Besides, the distinction between introductory activities

---

[9]CamemBERT-BASE's masked language model is fine-tuned on the following educational texts: pages from 4 French textbooks (apart from the pages of the validation and test subsets), 1293 *Fantastiques Exercices*, and the 79 original reading texts from the parallel corpus Alector [41].

[10]Among the overlapping segments, prediction differences occur in 1/3 of the pages and impact an average of 5% of the tokens within those segments.

exercises and actual exercises within a chapter is not clearly defined. As a result, some parts of lessons are wrongly predicted as exercises, and exercises within exploratory activities are occasionally misclassified as lessons or exercises. With the aim of developing a system with a high generalizability, training and evaluation sets must comprise more books from various collections.

These first experiments do not fully reflect the complexity of the task, since we use coarse-grained labels. For adaptation purposes, it will be necessary to provide a more detailed structure for each activity. Nevertheless, the results obtained with coarse-grained labels already point to a range of adaptation levels, depending on the nature of the blocks identified (e.g. lesson vs. exercise). On one hand, lesson adaptation for children with DCD or any other dyslexia-related disorder is already achievable, as it involves implementing standard accessibility modifications such as adjusting font, size, spacing and colors to enhance readability. On the other hand, activities that necessitate a shift in the mode of interaction require more in-depth extraction, and further processing to accurately identify[11] and apply this shift. Regarding the extraction task and given the results obtained in this paper, we can consider a 2-step token classification. Another limitation is that some components do not appear in all textbook collections. For example, pages comprising our dataset do not explicitly indicate the learning theme (this information is exclusive to the index), whereas some textbooks mention it on each chapter page. Finally, our experiments cover only French language study textbooks. Textbooks of different subjects may present certain specificities not only in layout but also in content. When applying our models to numerous existing collections, it is imperative to account for these variations in layout and semantics and ensure sufficient generalizability.

4. Conclusion

In this paper, we introduce our approach for automatic textbook structure and content extraction using activity-based textbook models. Our models not only provide layout and conceptual information, but also support the representation of an entire textbook according to widespread standards. We also report preliminary results on the token classification task to automatically identify all the components of a textbook page. These results are promising but reflect the difficulty of the task: generalization to various collections, whose content and layout are very different. Future work will address the progression of this automatic extraction task, using new multimodal transformer-based methods and going deeper into fine-grained labeling. Eventually, we will cover the implementation of the whole pipeline to convert a PDF textbook into a structured version according to our models.

Acknowledgments

---

[11]Our previous work introduces a classification task that aims to classify French language study exercises based on their adaptation type for children with DCD. [39, 40]

References

[1] C. Ou, A. Goel, D. Joyner, Towards a Pedagogical Framework for Designing and Developing iTextbooks, in: Proceedings of the 4th International Workshop on Inteligent Textbooks, 23rd International Conference on Artificial Intelligence in Education, 2022.

[2] S. Ritter, J. Fisher, A. Lewis, S. B. Finocchi, B. Hausmann, S. Fancsali, What's a Textbook? Envisioning the 21st Century K-12 Text., in: Proceedings of the 1st Workshop on Inteligent Textbooks, 20th International Conference on Artificial Intelligence in Education, 2019.

[3] L. Castillan, J. Lemarié, M. Mojahid, Numérique, handicap visuel et accessibilité des apprentissages. Contenus pédagogiques numériques: quelle accessibilité pour les élèves présentant une déficience visuelle?, Éducation & Formation (2018).

[4] L. Castillan, J. Lemarié, M. Mojahid, L'accessibilité des manuels scolaires numériques: l'exemple suédois, entre édition adaptée et édition inclusive, La nouvelle revue-Éducation et société inclusives (2019).

[5] L. R. Ketterlin-Geller, G. Tindal, Embedded technology: Current and future practices for increasing accessibility for all students, Journal of special education technology 22 (2007).

[6] S. Rahtz, N. Walsh, L. Burnard, A unified model for text markup: TEI, DocBook, and beyond, Proceedings of XML Europe (2004).

[7] L.-L. Stahn, S. Hennicke, E. W. De Luca, Using TEI for textbook research, in: Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH), 2016.

[8] F.-M. Gérard, X. Roegiers, Des manuels scolaires pour apprendre: concevoir, évaluer, utiliser, De Boeck Supérieur, 2009.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is All you Need, in: Proceedings of the 21st Conference on Neural Information Processing Systems, 2017.

[10] Y. Xu, M. Li, L. Cui, S. Huang, F. Wei, M. Zhou, 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, 2020.

[11] Y. Xu, Y. Xu, T. Lv, L. Cui, F. Wei, G. Wang, Y. Lu, D. Florencio, C. Zhang, W. Che, M. Zhang, L. Zhou, 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, 2021.

[12] Y. Huang, T. Lv, L. Cui, Y. Lu, F. Wei, LayoutLMv3: Pre-training for document ai with unified text and image masking, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022.

[13] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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, 2019.

[14] T. Hong, D. Kim, M. Ji, W. Hwang, D. Nam, S. Park, 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, 2022.

[15] S. Appalaraju, B. Jasani, B. U. Kota, Y. Xie, R. Manmatha, Docformer: End-to-end transformer for document understanding, in: Proceedings of the 18th IEEE International Conference on Computer Vision, 2021.

[16] R. Powalski, L. Borchmann, D. Jurkiewicz, T. Dwojak, M. Pietruszka, G. Palka, Going Full-TILT Boogie on Document Understanding with Text-Image-Layout Transformer, in: Proceedings of 16th International Conference on Document Analysis and Recognition, 2021.

[17] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, The Journal of Machine Learning Research (2020).

[18] L. Martin, B. Muller, P. J. Ortiz Suárez, Y. Dupont, L. Romary, E. de la Clergerie, D. Seddah, B. Sagot, CamemBERT: a Tasty French Language Model, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020.

[19] H. Le, L. Vial, J. Frej, V. Segonne, M. Coavoux, B. Lecouteux, A. Allauzen, B. Crabbé, L. Besacier, D. Schwab, FlauBERT: Unsupervised Language Model Pre-training for French, in: Proceedings of the 12th Language Resources and Evaluation Conference, 2020.

[20] J. Wang, L. Jin, K. Ding, 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, 2022.

[21] G. Jaume, H. K. Ekenel, J.-P. Thiran, Funsd: A dataset for form understanding in noisy scanned documents, in: Proceedings of the 15th International Conference on Document Analysis and Recognition Workshops, 2019.

[22] S. Najem-Meyer, M. Romanello, Page Layout Analysis of Text-heavy Historical Documents: a Comparison of Textual and Visual Approaches, in: Proceedings of the Computational Humanities Research Conference 2022, 2022.

[23] I. Alpizar-Chacon, M. van der Hart, Z. S. Wiersma, L. S. Theunissen, S. Sosnovsky, P. Brusilovsky, R. Baraniuk, A. Lan, Transformation of PDF textbooks into intelligent educational resources, in: Proceedings of the 2nd International Workshop on Inteligent Textbooks, 21st International Conference on Artificial Intelligence in Education, 2020.

[24] I. Alpizar-Chacon, S. Sosnovsky, Order out of chaos: Construction of knowledge models from pdf textbooks, in: Proceedings of the ACM Symposium on Document Engineering 2020, 2020.

[25] I. Alpizar-Chacon, S. Sosnovsky, Knowledge models from PDF textbooks, New Review of Hypermedia and Multimedia 27 (2021).

[26] I. Alpizar-Chacon, J. Barria-Pineda, K. Akhuseyinoglu, S. Sosnovsky, P. Brusilovsky, Integrating textbooks with smart interactive content for learning programming, in: Proceedings of the 3rd International Workshop on Inteligent Textbooks, 22nd International Conference on Artificial Intelligence in Education, 2021.

[27] I. Alpizar-Chacon, S. Sosnovsky, What's in an index: Extracting domain-specific knowledge graphs from textbooks, in: Proceedings of the ACM Web Conference 2022, 2022.

[28] Z.-M. Deng, M.-Y. Shi, C.-F. Li, Digitalization of Electronic Textbook Based on OPENCV, in: Proceedings of the International Conference on Machine Learning and Cybernetics, 2020.

[29] L. Gao, Z. Tang, X. Lin, Y. Liu, R. Qiu, Y. Wang, Structure extraction from PDF-based book

documents, in: Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries, 2011.

[30] S. Marinai, E. Marino, G. Soda, Conversion of PDF books in ePub format, in: Proceedings of the 11th International Conference on Document Analysis and Recognition, 2011.

[31] S. Tuarob, P. Mitra, C. L. Giles, A hybrid approach to discover semantic hierarchical sections in scholarly documents, in: Proceedings of the 13th International Conference on Document Analysis and Recognition, 2015.

[32] R. Kinney, C. Anastasiades, R. Authur, I. Beltagy, J. Bragg, et al., The Semantic Scholar Open Data Platform, arXiv preprint arXiv:2301.10140 (2023).

[33] K. Lo, J. C. Chang, A. Head, J. Bragg, A. X. Zhang, C. Trier, et al., The Semantic Reader Project: Augmenting Scholarly Documents through AI-Powered Interactive Reading Interfaces, arXiv preprint arXiv:2303.14334 (2023).

[34] Z. Shen, K. Lo, L. L. Wang, B. Kuehl, D. S. Weld, D. Downey, VILA: Improving Structured Content Extraction from Scientific PDFs Using Visual Layout Groups, Transactions of the Association for Computational Linguistics (2022).

[35] Z. Shen, R. Zhang, M. Dell, B. C. G. Lee, J. Carlson, W. Li, LayoutParser: A unified toolkit for deep learning based document image analysis, in: Proceedings of the 16th International Conference on Document Analysis and Recognition, 2021.

[36] I. Safder, S.-U. Hassan, A. Visvizi, T. Noraset, R. Nawaz, S. Tuarob, Deep learning-based extraction of algorithmic metadata in full-text scholarly documents, Information processing & management (2020).

[37] S. Mishra, L. Pluvinage, P. Senellart, Towards extraction of theorems and proofs in scholarly articles, in: Proceedings of the 21st ACM Symposium on Document Engineering, 2021.

[38] S. Aminta, A. Helbling, Outils pour le français CE1, Magnard, 2019. URL: https://www.magnard.fr/livre/9782210505377-outils-pour-le-francais-ce1-2019-manuel-eleve.

[39] E. Lincker, C. Guinaudeau, O. Pons, J. Dupire, C. Hudelot, V. Mousseau, I. Barbet, C. Huron, Classification automatique de données déséquilibrées et bruitées : application aux exercices de manuels scolaires, in: Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles, 2023.

[40] E. Lincker, C. Guinaudeau, O. Pons, J. Dupire, C. Hudelot, V. Mousseau, I. Barbet, C. Huron, Noisy and Unbalanced Multimodal Document Classification: Textbook Exercises as a Use Case, in: Proceedings of the 20th International Conference on Content-based Multimedia Indexing (to appear), 2023.

[41] N. Gala, A. Tack, L. Javourey-Drevet, T. François, J. C. Ziegler, 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, 2020.

[42] Z. Wang, P. Ng, X. Ma, R. Nallapati, B. Xiang, Multi-passage bert: A globally normalized bert model for open-domain question answering, arXiv preprint arXiv:1908.08167 (2019).

Appendix

Documents depict the exercise featured in Figure 1 in our format, TEI and DocBook.
The representation has been simplified in Listing 1 by omitting position (@xmin, @ymin, @xmax, @ymax), spacing (@space, @spnext) and font style (@font) attributes. Unique id (@id/@ID) attributes are omitted in all documents.

Listing 1: Our format

```xml
<ex>
  <num>
    <seg>
      <token>6</token>
    </seg>
  </num>
  <indicator>
    <seg>
      <token>*</token>
      <token>*</token>
    </seg>
  </indicator>
  <instruction>
    <seg>
      <token>Complète</token>
      <token>les</token>
      <token>phrases</token>
      <token>avec</token>
      <token>on</token>
      <token>ou</token>
      <token>ont</token>
      <token>.</token>
    </seg>
  </instruction>
  <statement>
    <list>
      <item>
        <bullet>
          <token>a.</token>
        </bullet>
        <seg>
          <token>Si</token>
          <token>…</token>
          <token>allait</token>
          <token>au</token>
          <token>cinéma</token>
          <token>?</token>
        </seg>
      </item>
      <item>
        <bullet>
          <token>b.</token>
        </bullet>
        <seg>
          <token>Ils</token>
          <token>…</token>
          <token>vu</token>
          <token>ce</token>
          <token>film</token>
          <token>dix</token>
          <token>fois</token>
          <token>.</token>
        </seg>
      </item>
      <item>
        <bullet>
          <token>c.</token>
        </bullet>
        <seg>
          <token>…</token>
          <token>s'</token>
          <token>installe</token>
          <token>dans</token>
          <token>les</token>
          <token>fauteuils</token>
          <token>moelleux</token>
          <token>.</token>
        </seg>
      </item>
      <item>
        <bullet>
          <token>d.</token>
        </bullet>
        <seg>
          <token>Mes</token>
          <token>parents</token>
          <token>…</token>
          <token>pris</token>
          <token>du</token>
          <token>pop-corn</token>
          <token>.</token>
        </seg>
      </item>
      <item>
        <bullet>
          <token>e.</token>
        </bullet>
        <seg>
          <token>Les</token>
          <token>enfants</token>
          <token>…</token>
          <token>sursauté</token>
          <token>devant</token>
          <token>une</token>
          <token>scène</token>
          <token>de</token>
          <token>film</token>
          <token>.</token>
        </seg>
      </item>
    </list>
  </statement>
</ex>
```

Listing 2: TEI

```xml
<div type="exercise" n="6", difficulty="**">
  <div type="instruction">
    <seg>
      <w>Complète</w>
      <w>les</w>
      <w>phrases</w>
      <w>avec</w>
      <w>on</w>
      <w>ou</w>
      <w>ont</w>
      <pc join="left">.</pc>
    </seg>
  </div>
  <lb/>
  <div type="statement">
    <list rend="lettered">
      <item n="a.">
        <seg>
          <w>Si</w>
          <pc>…</pc>
          <w>allait</w>
          <w>au</w>
          <w>cinéma</w>
          <pc>?</pc>
        </seg>
      </item>
      <lb/>
      <item n="b.">
        <seg>
          <w>Ils</w>
          <pc>…</pc>
          <w>vu</w>
          <w>ce</w>
          <w>film</w>
          <w>dix</w>
          <w>fois</w>
          <pc join="left">.</pc>
        </seg>
      </item>
      <lb/>
      <item n="c.">
        <seg>
          <pc>…</pc>
          <w>s'</w>
          <w>installe</w>
          <w>dans</w>
          <w>les</w>
          <w>fauteuils</w>
          <w>moelleux</w>
          <pc join="left">.</pc>
        </seg>
      </item>
      <lb/>
      <item n="d.">
        <seg>
          <w>Mes</w>
          <w>parents</w>
          <pc>…</pc>
          <w>pris</w>
          <w>du</w>
          <w>pop-corn</w>
          <pc join="left">.</pc>
        </seg>
      </item>
      <lb/>
      <item n="e.">
        <seg>
          <w>Les</w>
          <w>enfants</w>
          <pc>…</pc>
          <w>sursauté</w>
          <w>devant</w>
          <w>une</w>
          <w>scène</w>
          <w>de</w>
          <w>film</w>
          <pc join="left">.</pc>
        </seg>
      </item>
    </list>
  </div>
</div>
```

Listing 3: DocBook

```xml
<section role="exercise">
  <section role="instruction">
    <para>Complète les phrases avec on ou ont.</para>
  </section>
  <section role="statement">
    <orderedlist numeration="loweralpha">
      <listitem>
        <para>Si … allait au cinéma ?</para>
      </listitem>
      <listitem>
        <para>Ils … vu ce film dix fois.</para>
      </listitem>
      <listitem>
        <para>… s'installe dans les fauteuils
              moelleux.</para>
      </listitem>
      <listitem>
        <para>Mes parents … pris du pop-corn.</para>
      </listitem>
      <listitem>
        <para>Les enfants … sursauté devant une scène
              de film.</para>
      </listitem>
    </statement>
  </section>
</section>
```