De l'image à la représentation structurée : analyse et modélisation des manuels scolaires

Mohamed-Amine Lasheb [1], **Olivier Pons** [1],
Mohammed Bekkouche [2], **Isabelle Barbet** [1], **Caroline Huron** [3]

1. *Laboratoire Cedric, Conservatoire national des Arts et Métiers*
   292 Rue Saint-Martin, 75003 Paris, France
   firstname.lastname@lecnam.net
2. *LabRI-SBA, Ecole Supérieure en Informatique*
   22000 Sidi Bel Abbes, Algeria
   m.bekkouche@esi-sba.dz
3. *SEED, Inserm, Paris Cité*
   Paris, France
   firstname.lastname@cri-paris.org

RÉSUMÉ. *Le projet ANR MALIN vise à améliorer l'accessibilité des manuels scolaires numériques pour les élèves en situation de handicap en automatisant l'ensemble du processus d'adaptation. Une étape cruciale de ce processus est l'extraction et la structuration du contenu (leçons, illustrations, exercices, etc.). Cet article présente une solution basée sur des techniques de vision par ordinateur et d'apprentissage profond et compare l'efficacité de différents modèles.*

ABSTRACT. *The ANR MALIN project aims to improve the accessibility of digital textbooks for students with disabilities by automating the entire adaptation process. A crucial step in this process is the extraction and structuring of content (lessons, illustrations, exercises, , etc.). This paper presents a solution based on computer vision and deep learning techniques and compares the effectiveness of different models.*

MOTS-CLÉS : *éducation inclusive, accessibilité, extraction de contenu, structuration, modèles de vision par ordinateur.*

KEYWORDS: *inclusive education, accessibility, content extraction, structuring, computer vision models.*

1. Introduction

L'inclusion des enfants en situation de handicap dans les écoles et établissements scolaires ordinaires a été posée par la loi (lois du 11 février 2005 et du 8 juillet 2013) France (2005, 2013), ce qui a permis d'augmenter le nombre d'enfants en situation de handicap inscrits dans leur école de référence. Cependant, sa mise en place n'est pas simple. Un point d'achoppement concerne notamment les manuels scolaires, très utilisés en classe, mais qui, même dans leur version numérique (lorsqu'elles existent), sont très rarement accessibles.

Des adaptations sont généralement faites à la main par des associations et des organismes spécialisés, mais la diversité des manuels et leur renouvellement fréquent rendent ces adaptations lentes, coûteuses et peu nombreuses.

Sur la scène internationale, de nombreux pays ont introduit des obligations légales d'accessibilité minimale pour leurs livres scolaires. À l'échelle mondiale, c'est l'objet de l'initiative "Accessible Digital Textbooks", portée par l'UNICEF [1].

L'objectif de rendre accessibles les manuels scolaires en automatisant le processus de transposition, puis de permettre l'évaluation et l'amélioration des adaptations via la mise à disposition d'une plateforme d'adaptation, est donc un défi sociétal majeur.

L'objectif du projet ANR MALIN, dans lequel s'inscrit notre travail, est de répondre à ce défi. Du fait des collaborations avec l'association Le Cartable Fantastique [2] et l'INJA [3], le focus est principalement mis sur la dyspraxie et la déficience visuelle, mais le projet vise à se généraliser à tout type de handicap.

La Figure 1 montre des adaptations réalisées pour des élèves dyspraxiques. Elles visent à minimiser les tâches trop coûteuses pour eux, notamment l'écriture.

Une fois adaptés, les manuels peuvent être mis à disposition des publics en situation de handicap.

La structure complexe des manuels scolaires, illustrée dans la Figure 2, complique sensiblement leur adaptation. L'extraction et la structuration des contenus sont une première étape cruciale vers une automatisation des adaptations visant à rendre les manuels accessibles.

Ce travail explore l'utilisation de la vision par ordinateur pour automatiser l'extraction des contenus, permettant ainsi une représentation structurée des éléments des manuels.

1. https://www.accessibletextbooksforall.org/
2. https://www.cartablefantastique.fr/
3. Institut National des Jeunes Aveugles

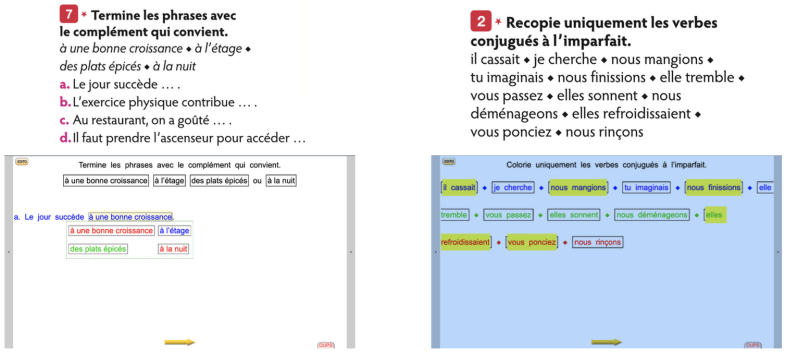


FIGURE 1 – Exemples d'adaptations d'exercices

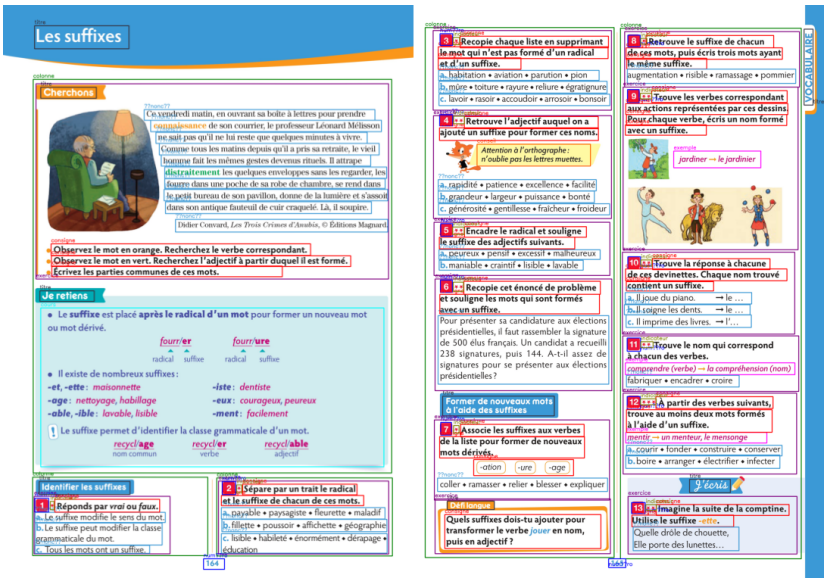


FIGURE 2 – Exemple de structure de mise en page d'un manuel scolaire. Source: Magnard (2019)

2. État de l'Art

Pour transformer les manuels scolaires en supports accessibles et interactifs, l'équipe du projet MALIN a proposé un pipeline complet, illustré dans la Figure 3. Ce pipeline décrit les étapes nécessaires à la conversion d'un fichier PDF, qu'il soit natif ou scanné,

en un manuel adapté au format HTML. On remarque que les deux dernières étapes du diagramme sont représentées de manière plus estompée : cela reflète le fait que ce travail se concentre principalement sur les deux premières, à savoir la collecte de données et la modélisation de la mise en page.

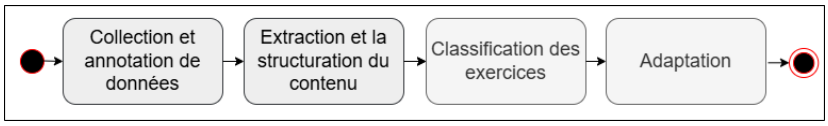


FIGURE 3 – Diagramme d'activité du processus global d'adaptation.

Ce processus débute par l'entrée d'un fichier PDF et aboutit à la création d'un manuel interactif. Les étapes clés incluent :

– **L'extraction et la structuration du contenu** des manuels scolaires. La sortie de cette étape est le point d'entrée de toutes les autres. Elle repose sur des modèles sémantiques formalisés dans Lincker, Pons *et al*. (2023). Ces formalismes s'appuient sur des DTD et des schémas XML ou JSON, et peuvent être traduits vers les normes TEI ((Rahtz *et al*., 2004; Stahn *et al*., 2016)) et DocBook (Walsh, Hamilton (2010)).

Après la constitution d'un premier corpus par des méthodes basées sur des règles et des approches statistiques, s'appuyant principalement sur les caractéristiques des polices et la position dans le document, un corpus de base a pu être établi, corrigé et étendu manuellement, bien qu'il ne puisse être diffusé publiquement en raison de restrictions liées aux droits d'auteur.

En utilisant ce corpus pour l'entraînement, des méthodes de TAL, utilisant des LLM basés sur des transformers et des transformers multimodaux (BERT Devlin *et al*. (2019), LayoutLM Xu *et al*. (2020), ViLa J. Lin *et al*. (2024)), ont été proposées dans Lincker, Pons *et al*. (2023).

– **La classification des exercices:** elle représente une autre étape essentielle. Les adaptations jouent un rôle clé dans ce processus. L'objectif est de classer chaque exercice en fonction du type d'adaptation le plus approprié. Pour ce faire, Lincker, Guinaudeau *et al*. (2023) s'appuient sur des modèles de langage pré-entraînés Martin *et al*. (2020); Le *et al*. (2020), et tire parti d'architectures multimodales Xu *et al*. (2022); Wang *et al*. (2022). Pour donner un aperçu concret du périmètre de la classification, voici les principales catégories d'exercices rencontrées : identification, classement, QCM, transformation, production, remise en ordre, oral, association, dictée et justification. Ces classes varient en fonction de l'unité linguistique mobilisée (mot, phrase, lettre, etc.) et du mode d'interaction (écriture, coche, échange, etc.).

– **Les adaptations** : une fois les exercices classifiés, ils sont transformés en formats interactifs HTML afin de faciliter l'interaction avec les élèves. Pour chaque classe d'exercice, un type d'adaptation spécifique est défini, en tenant compte à la fois de la nature de l'activité (QCM, dictée, association, etc.) et des capacités motrices, visuelles ou cognitives des élèves.

Par exemple, dans un exercice où l'élève doit « souligner le verbe », l'adaptation consistera à proposer une interaction par clic sur le mot correspondant, plutôt qu'une écriture manuscrite difficilement accessible pour certains élèves. De même, un exercice de classement pourra être adapté en glisser-déposer, tandis qu'un QCM utilisera des cases à cocher élargies.

Ces adaptations interactives prennent également la forme d'ajustements visuels (taille des caractères, espacement) ou de simplification du langage, et visent à maintenir les objectifs pédagogiques tout en respectant les capacités spécifiques des enfants.

Par ailleurs, une plateforme de vérification et un processus d'évaluation des exercices adaptés sont abordés dans (Pacini *et al*., 2023).

3. Problématiques et hypothèses

Dans ce papier, nous nous concentrons sur l'extraction déjà réalisée, ses défis, et notre nouvelle approche d'extraction avec la vision par ordinateur.

Les méthodes d'extraction et de structuration précédemment développées par notre équipe, notamment dans les travaux de Lincker, Guinaudeau *et al*. (2023), reposent principalement sur des modèles de langue ou des architectures multimodales. Toutefois, ces approches présentent certaines limitations, en particulier une difficulté de généralisation entre différentes collections de manuels scolaires, en raison des variations de mise en page (voir Figure 4).

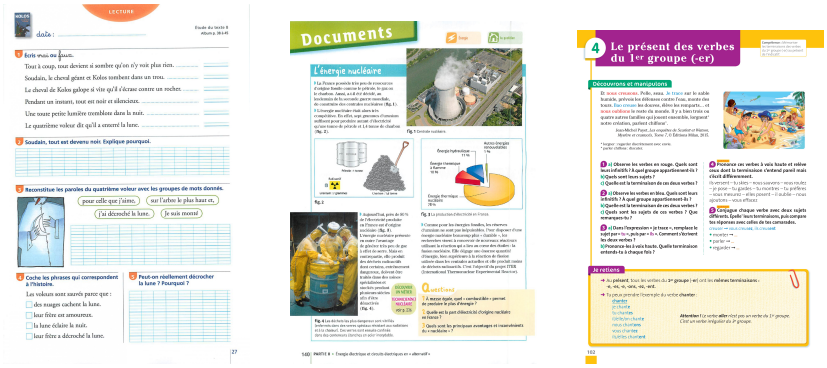
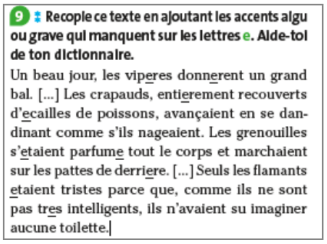


FIGURE 4 – Des pages provenant de différents manuels scolaires

Par ailleurs, l'annotation manuelle demeure une tâche longue et fastidieuse, et l'association précise entre les éléments visuels (tableaux, encadrés, pictogrammes, etc.) et leur contenu textuel reste complexe. De plus, la majorité des modèles existants sont principalement optimisés pour la langue anglaise, ce qui constitue un obstacle supplémentaire dans le contexte francophone.

Les exercices personnalisés avec des objectifs différents incluent des cas particuliers comme les fautes d'orthographe et de grammaire volontaires dans certains exercices (voir Figure 5). Ces fautes sont souvent introduites pour tester la capacité des élèves à identifier et corriger des erreurs, ajoutant une couche de complexité pour les modèles de langue et les systèmes de traitement de texte. En effet, ces derniers peuvent soit ne pas bien interpréter le texte, ce qui entraîne une classification et une structuration incorrectes, soit ne pas reconnaître ces erreurs comme intentionnelles et les corriger automatiquement à l'aide de leur dictionnaire intégré, altérant ainsi la nature pédagogique de l'exercice.



Texte sans accents

Absence d'espaces entre les mots

FIGURE 5 – Exemples d'exercices contenant des fautes volontaires

Extraire une consigne, un exemple ou un conseil est plus complexe que d'extraire simplement un titre ou un tableau, comme le faisaient la plupart des modèles entraînés sur des jeux de données tels que PubLayNet (Zhong *et al.*, 2019) ou DocBank (Li *et al.*, 2020), qui ciblent principalement des classes visuellement identifiables comme les titres, tableaux ou numéros de page. Les consignes et les exemples sont souvent intégrés dans des paragraphes plus larges et ne suivent pas toujours une structure typographique claire, rendant leur identification plus difficile. Un autre point bloquant est que ces modèles sont malheureusement efficaces uniquement avec les PDF natifs.

Pour surmonter ces limitations, il devient évident que l'analyse des images et des structures visuelles offre un moyen efficace de reconnaître divers éléments présents sur une page. Par exemple, lorsqu'un exercice contient deux listes verticales, il est facile et rapide d'identifier qu'il s'agit d'un exercice de type "liaison entre les choix".

Il apparaît également que, pour certaines catégories d'exercices, l'application de modèles de vision par ordinateur peut améliorer de manière significative la classification. En effet, de nombreux exercices présentent des éléments visuels caractéristiques — telles que des listes, des couleurs, des cases à cocher, des flèches ou des zones à remplir — qui permettent d'identifier leur type plus efficacement que par le seul traitement du texte. Ces indices visuels offrent une aide précieuse pour détecter la structure et la logique de l'exercice, en particulier lorsqu'ils suivent un format graphique récurrent.

Pour évaluer les performances de la vision par ordinateur, en particulier dans l'extraction et la structuration du contenu, nous avons testé plusieurs modèles tels que

LayoutParser (Shen *et al*., 2021), Detectron2 (Merz *et al*., 2023) et YOLO (Redmon *et al*., 2016). Ce dernier s'est avéré plus efficace pour notre cas d'utilisation, comme montré dans les résultats du (Tab.1), et a démontré une meilleure capacité à gérer les variations de mise en page et les éléments spécifiques aux manuels scolaires, notamment pour les documents scannés.

La suite de ce papier montre comment les méthodes utilisant la vision par ordinateur permettent de surmonter les limitations des approches existantes, notamment en ce qui concerne la gestion des documents scannés de qualité variable.

4. Méthodologie

Notre jeu de données est composé de 22 manuels scolaires français de l'école élémentaire, couvrant l'apprentissage de la langue, les mathématiques, les sciences et l'histoire-géographie. Ces manuels sont répartis équitablement en deux catégories : 11 manuels scannés et 11 manuels au format PDF natif. Cependant, notre concentration s'est principalement portée sur les manuels de langue.

Pour les PDF natifs, deux méthodes complémentaires d'annotation ont été mises en œuvre. Une première partie a été réalisée par des spécialistes de l'adaptation scolaire, notamment les membres de l'association Le Cartable Fantastique, via une plateforme développée en interne. Cette plateforme convertit les PDF en HTML, puis extrait automatiquement des éléments structurants à l'aide de règles basées sur les polices et la mise en page, après quelques annotations manuelles initiales des experts. Les annotations ainsi produites ont ensuite été converties au format *Labelme*[4] après transformation des pages en images.

Les pages restantes, non couvertes par cette plateforme, ainsi que les 11 manuels scannés, ont été directement convertés en images, puis annotés manuellement via l'outil *Labelme*. Chaque élément (consigne, titre, énoncé, etc.) a été encadré par une boîte rectangulaire. Pour les documents scannés, un prétraitement a été nécessaire (suppression des marges, correction des ombres, séparation des pages) afin d'améliorer la clarté avant l'annotation.

L'outil Labelme a ensuite été utilisé dans une boucle de rétroaction : après un certain taux d'annotation et un premier entraînement préliminaire (par exemple, dans l'annotation des exercices, nous avons entraîné le modèle de détection d'exercices), les prédictions du modèle ont été rouvertes dans Labelme pour correction. Cela nous a permis de construire progressivement un corpus annoté de grande taille, sans devoir tout annoter manuellement dès le départ.

Afin d'éviter le surapprentissage, 40 à 50 pages par manuel ont été sélectionnées pour l'entraînement, garantissant une diversité structurelle tout en limitant l'influence

4. https://labelme.io/

des formats spécifiques. La Figure (6) montre la répartition des annotations dans notre jeu de données.

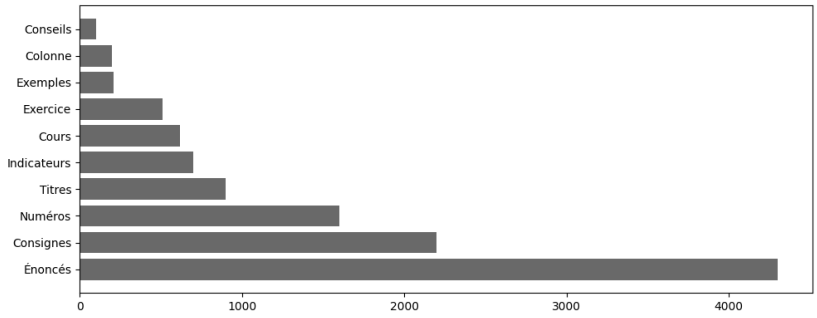


FIGURE 6 – Répartition des annotations dans notre jeu de données.

Le processus de modélisation des pages (extraction et structuration), illustré dans le diagramme d'activité comme deuxième étape (Figure 7), commence par la détection des exercices, où les zones pertinentes sont localisées sous forme de régions rectangulaires. Ensuite, la détection du format de page permet de distinguer les mises en page à une ou plusieurs colonnes. La détection des éléments de la page (tels que les consignes, énoncés, titres, indicateurs, exemples, etc.) est réalisée en fonction du format de la mise en page. Une fois ces éléments identifiés, une vérification est effectuée pour s'assurer qu'ils appartiennent bien à un exercice spécifique.

L'extraction de texte est ensuite réalisée à l'aide de *PDFAlto* [5] pour les fichiers PDF natifs, et de *RapidOCR* (RapidAI, 2022) pour les PDF scannés. Le texte extrait est alors associé aux éléments précédemment détectés, permettant une structuration cohérente et précise des contenus.

Enfin, les résultats sont organisés dans un fichier JSON structuré, selon un schéma prédéfini conforme aux spécifications de Attouche *et al.* (2024). Cette standardisation garantit la cohérence, l'interopérabilité et l'intégration aisée dans les systèmes éducatifs. Un exemple de sortie est illustré dans la figure 8, correspondant à l'analyse de l'exercice de la figure 1.

Une fois les exercices correctement extraits et structurés, la classification devient une étape plus simple et plus efficace. Dans le cadre du projet MALIN, 45 classes principales d'exercices ont été identifiées. Grâce à notre modèle, qui assure une extraction fiable des exercices, l'annotation manuelle, qu'elle soit dédiée à la classification ou à la vérification des classifications automatiques issues de Lincker, Guinaudeau *et al.* (2023), a été considérablement simplifiée : il suffit désormais de cliquer sur un exercice et de choisir une classe parmi les options proposées, comme illustré en Fig. 9.

5. https://github.com/kermitt2/pdfalto

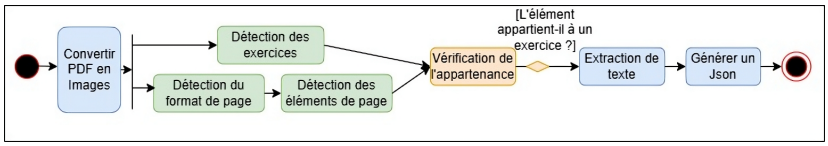


FIGURE 7 – Diagramme d'activité du processus d'extraction des exercices.


```
{ "manuel": "Outils pour le Français CE2 (2019)","isbn": "978-2-210-50538-4",
  "pages": [ { "id": "165",
      "exercices": [
        { "numéro": 5,
          "consigne": "Encadre le radical et souligne le suffixe des adjectifs suivants.",
          "énoncé": [ "a. peureux • pensif • excessif • malheureux",
           "b. maniable • craintif • lisible • lavable" ] }, ... ] }, ... ] }
```


FIGURE 8 – Sortie d'extraction et structuration.

Auparavant, cette tâche nécessitait soit une délimitation manuelle des exercices dans l'image, soit la copie intégrale du contenu textuel.

Ce processus garantit une extraction précise et adaptable du texte, couvrant tous les formats de manuels.

5. Résultats et Discussion

Nous débutons notre pipeline avec les modèles de reconnaissance du layout de la page et de localisation des exercices, en utilisant YOLOv10x sur des images de 736px avec un batch de 16 et l'optimiseur AdamW (Loshchilov, Hutter (2019)). Cette variante d'Adam (Adaptive Moment Estimation) inclut une correction de la régularisation par poids (weight decay), améliorant ainsi la stabilité et la convergence du modèle. Un taux d'apprentissage de 0,0001 avec une décroissance cosinus a également été appliqué. Les performances sont évaluées avec un seuil de confiance standard de 0,50 en vision par ordinateur. Cette configuration a permis d'atteindre une précision élevée pour les deux tâches principales : 97,8 % pour les exercices et 98 % pour le format de page. Cette étape assure une continuité fluide vers les suivantes.

Une fois le layout de la page et les exercices détectés avec précision, l'étape suivante consiste à identifier les différents éléments constitutifs des exercices.

La Table 1 compare les performances des trois modèles testés pour la détection des éléments : Detectron2, LayoutParser et YOLOv10x.

Detectron2 utilise un modèle GeneralizedRCNN avec un backbone ResNet-101 FPN (He *et al.* (2015) T.-Y. Lin *et al.* (2017)), un optimiseur SGD (Ruder (2017)) avec un taux d'apprentissage initial de 0.0025, et un entraînement sur des images de

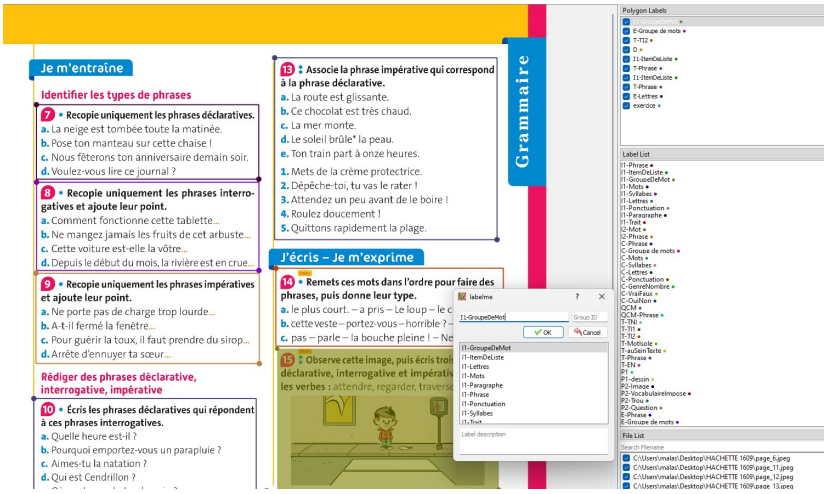


FIGURE 9 – Illustration de la méthode d'annotation pour la classification.

tailles variées (640 à 960 pixels) avec des augmentations de données incluant des flips horizontaux.

LayoutParser utilise également un modèle GeneralizedRCNN avec un backbone ResNet-101 FPN, un optimiseur SGD avec un taux d'apprentissage initial de 0.0025, et un entraînement sur des images de tailles variées (640 à 960 pixels) avec des augmentations de données incluant des flips horizontaux (Yang *et al*. (2023)).

YOLOv10x, avec les paramètres mentionnés précédemment.

Ce dernier, grâce à son approche en une seule passe, se distingue nettement en optimisant simultanément vitesse et précision. Avec un **AP50** de 78,1 % (précision moyenne à un seuil de recouvrement de 50 %) et un **AP (IoU 0.50:0.95)** de 55,6 % (moyenne des précisions sur plusieurs seuils, de 0,50 à 0,95), il surpasse largement les autres modèles. LayoutParser améliore légèrement les résultats de Detectron2, mais reste en retrait par rapport à YOLOv10x, notamment en raison de difficultés à générer correctement les boîtes englobantes des classes détectées. Detectron2 seul montre des performances plus limitées et inférieures aux autres modèles.

L'analyse par catégorie révèle que YOLOv10x excelle dans l'identification de la majorité des éléments structurants des documents, atteignant 92,2 % pour les numéros, 97,7 % pour les titres et 92,3 % pour les conseils. Cependant, ses performances sont plus faibles pour les "exemples" (53,8 %) en raison d'un nombre limité d'annotations (<100, comme montré dans Fig. 6), et pour les "cours" (41,8 %), car peu de pages contenant des cours ont été annotées, l'objectif principal étant l'extraction des exercices.

TABLEAU 1 – Performance Comparative des Modèles d'Analyse de Mise en Page

Métrique / Classe	detectron2	LayoutParser	YOLOv10x
AP50 (IoU 0.50)	0.396	0.540	**0.781**
AP (IoU 0.50:0.95)	0.187	0.262	**0.556**
consigne	0.227	0.235	**0.811**
énoncé	0.083	0.098	**0.837**
numéro	0.459	0.461	**0.922**
titre	0.248	0.298	**0.977**
indicateur	0.385	0.359	**0.824**
cours	0.072	0.063	**0.418**
exemple	0.017	0.108	**0.538**
conseil	0.001	0.475	**0.923**

Note : Les entraînements ont été réalisés pour la même durée et avec le même matériel.

Certaines erreurs observées sont principalement dues aux annotations, qui combinent des annotations générées automatiquement et manuelles, présentant des différences au niveau des marges et du padding des boîtes englobantes. Les annotations générées automatiquement sont généralement plus précises, tandis que les annotations manuelles introduisent parfois des marges additionnelles aléatoires.

6. Conclusion

Les contributions du projet résident dans la démonstration du potentiel de la vision par ordinateur pour comprendre des documents complexes comme les manuels scolaires, ainsi que dans l'obtention de résultats prometteurs pour l'extraction et la structuration de contenu multimodal, en particulier pour les documents numérisés.

Cette étude valide l'efficacité de la vision par ordinateur pour l'extraction et la structuration des contenus scolaires, avec une précision de 97,8 % pour les exercices et de 98 % pour le format de page. Les éléments clés, comme les numéros (92,2 %), les titres (97,7 %) et les conseils (92,3 %), sont bien identifiés. Nos futurs travaux viseront à améliorer les classes moins performantes afin d'assurer une accessibilité encore plus inclusive.

L'automatisation de l'extraction des exercices a également simplifié la classification en réduisant la charge d'annotation. L'assignation des classes aux exercices est désormais plus rapide, préparant efficacement l'étape suivante du pipeline d'adaptation.

Les perspectives futures incluent l'amélioration des modèles et l'augmentation des données par la génération d'exercices et de manuels, ce qui, outre son utilité pédagogique directe, permettrait de remédier aux déséquilibres des classes.

Par ailleurs, les manuels traités sont essentiellement des manuels de français. Les premières expériences sur d'autres matières semblent montrer que les manuels de

mathématiques du primaire, étant également assez réguliers, devraient pouvoir être traités de la même manière (la géométrie posant néanmoins des problèmes spécifiques, non pour l'extraction, mais en termes d'adaptation). La structure des manuels d'histoire-géographie, de sciences physiques ou de sciences de la vie, beaucoup moins régulière, avec de nombreux schémas, documents et illustrations croisés, semble encore plus complexe.

En termes d'extraction et de structuration de documents complexes, une autre piste intéressante est l'extension des méthodes à d'autres ressources dont la structure reste assez proche de celle des manuels, notamment les revues de vulgarisation scientifique. Nous envisageons par exemple, dans cette perspective, d'explorer le fonds documentaire du CNUM [6].

Nous visons également d'autres approches, telles que les Approches End-to-End et les Modèles Multimodaux de Grande Taille (VLMs) comme GPT-4 (OpenAI *et al*., 2024), Qwen (Bai *et al*., 2023), LLaMA (Touvron *et al*., 2023) pour améliorer la précision et l'efficacité de l'extraction et de la structuration des contenus scolaires.

En conclusion, cette étude propose un pipeline d'analyse et de structuration de documents complexes. De plus, en contribuant significativement au processus d'adaptation automatisée des manuels scolaires, elle ouvre la voie à une éducation plus inclusive.

Remerciements

Ce travail a été soutenu par le projet ANR-21-CE38-0014 MALIN.

Bibliographie

Attouche L., Baazizi M.-A., Colazzo D., Ghelli G., Sartiani C., Scherzinger S. (2024, janvier). *Validation of modern json schema: Formalization and complexity* (vol. 8) n° POPL. New York, NY, USA, Association for Computing Machinery. Consulté sur https://doi.org/10.1145/3632891

Bai J., Bai S., Chu Y., Cui Z., Dang K., Deng X. *et al*. (2023). *Qwen technical report.* Consulté sur https://arxiv.org/abs/2309.16609

Devlin J., Chang M.-W., Lee K., Toutanova K. (2019). *Bert: Pre-training of deep bidirectional transformers for language understanding.* Consulté sur https://arxiv.org/abs/1810.04805

France. (2005). *Loi n° 2005-102 du 11 février 2005 pour l'égalité des droits et des chances, la participation et la citoyenneté des personnes handicapées.* Consulté sur https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000000809647 (Consulté le 2025-04-22)

France. (2013). *Loi n° 2013-595 du 8 juillet 2013 d'orientation et de programmation pour la refondation de l'école de la république.* Consulté sur https://www.legifrance.gouv.fr/dossierlegislatif/JORFDOLE000026973437/ (Consulté le 2025-04-22)

6. https://cnum.cnam.fr/

He K., Zhang X., Ren S., Sun J. (2015). *Deep residual learning for image recognition.* Consulté sur https://arxiv.org/abs/1512.03385

Le H., Vial L., Frej J., Segonne V., Coavoux M., Lecouteux B. *et al*. (2020). *Flaubert: Unsupervised language model pre-training for french.* Consulté sur https://arxiv.org/abs/1912.05372

Li M., Xu Y., Cui L., Huang S., Wei F., Li Z. *et al*. (2020). *Docbank: A benchmark dataset for document layout analysis.* Consulté sur https://arxiv.org/abs/2006.01038

Lin J., Yin H., Ping W., Lu Y., Molchanov P., Tao A. *et al*. (2024). *Vila: On pre-training for visual language models.* Consulté sur https://arxiv.org/abs/2312.07533

Lin T.-Y., Dollár P., Girshick R., He K., Hariharan B., Belongie S. (2017). *Feature pyramid networks for object detection.* Consulté sur https://arxiv.org/abs/1612.03144

Lincker E., Guinaudeau C., Pons O., Barbet I., Dupire J., Hudelot C. *et al*. (2023, juin). Classification automatique de données déséquilibrées et bruitées : application aux exercices de manuels scolaires. In C. Servan, A. Vilnat (Eds.), *Actes de CORIA-TALN 2023. Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 4 : articles déjà soumis ou acceptés en conférence internationale*, vol. 4, p. 121-130. Paris, France, ATALA. Consulté sur https://hal.science/hal-04130220

Lincker E., Pons O., Guinaudeau C., Barbet I., Dupire J., Hudelot C. *et al*. (2023). Layout- and activity-based textbook modeling for automatic pdf textbook extraction. In *Proceedings of the intelligent textbooks workshop at the 24th international conference on artificial intelligence in education (aied)*, p. 37–53. CEUR Workshop Proceedings. Consulté sur https://hal.science/hal-04184895 (Available under a Creative Commons Attribution 4.0 International License)

Loshchilov I., Hutter F. (2019). *Decoupled weight decay regularization.* Consulté sur https://arxiv.org/abs/1711.05101

Magnard. (2019). *Outils pour le français ce2 (2019) - manuel élève*. Magnard. Consulté sur https://www.magnard.fr/livre/9782210505384-outils-pour-le-francais-ce2-2019-manuel-eleve (Consulté en avril 2025)

Martin L., Muller B., Ortiz Suárez P. J., Dupont Y., Romary L., Clergerie de la *et al*. (2020). Camembert: a tasty french language model. In *Proceedings of the 58th annual meeting of the association for computational linguistics.* Association for Computational Linguistics. Consulté sur http://dx.doi.org/10.18653/v1/2020.acl-main.645

Merz G., Liu Y., Burke C. J., Aleo P. D., Liu X., Carrasco Kind M. *et al*. (2023, septembre). *Detection, instance segmentation, and classification for astronomical surveys with deep learning (deepdisc): Detectron2 implementation and demonstration with hyper suprime-cam data* (vol. 526) n° 1. Oxford University Press (OUP). Consulté sur http://dx.doi.org/10.1093/mnras/stad2785

OpenAI, Achiam J., Adler S., Agarwal S., Ahmad L., Akkaya I. *et al*. (2024). Gpt-4 technical report. Consulté sur https://arxiv.org/abs/2303.08774

Pacini L., Dupire J., Barbet I., Pons O., Guinaudeau C., Mousseau V. *et al*. (2023). Textbook accessibility for children with dyspraxia and visual disabilities. In *17th international conference of the association for the advancement of assistive technology in europe (aaate 2023).*

Rahtz S., Walsh N., Burnard L. (2004). A unified model for text markup: Tei, docbook, and beyond. In *Proceedings of xml europe.*

RapidAI. (2022). *Rapidocr.* https://github.com/RapidAI/RapidOCR. (Disponible sur https://github.com/RapidAI/RapidOCR)

Redmon J., Divvala S., Girshick R., Farhadi A. (2016). *You only look once: Unified, real-time object detection.*

Ruder S. (2017). *An overview of gradient descent optimization algorithms.* Consulté sur https://arxiv.org/abs/1609.04747

Shen Z., Zhang R., Dell M., Lee B. C. G., Carlson J., Li W. (2021). *Layoutparser: A unified toolkit for deep learning based document image analysis.* Consulté sur https://arxiv.org/abs/2103.15348

Stahn L.-L., Hennicke S., De Luca E. W. (2016). Using tei for textbook research. In *Proceedings of the workshop on language technology resources and tools for digital humanities (lt4dh).*

Touvron H., Lavril T., Izacard G., Martinet X., Lachaux M.-A., Lacroix T. *et al*. (2023). *Llama: Open and efficient foundation language models.* (vol. abs/2302.13971). Consulté sur http://dblp.uni-trier.de/db/journals/corr/corr2302.html#abs-2302-13971

Walsh N., Hamilton R. L. (2010). *Docbook 5: The definitive guide: The official documentation for docbook*. " O'Reilly Media, Inc.".

Wang J., Jin L., Ding K. (2022). *Lilt: A simple yet effective language-independent layout transformer for structured document understanding.* Consulté sur https://arxiv.org/abs/2202.13669

Xu Y., Li M., Cui L., Huang S., Wei F., Zhou M. (2020, août). Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th acm sigkdd international conference on knowledge discovery amp; data mining*, p. 1192–1200. ACM. Consulté sur http://dx.doi.org/10.1145/3394486.3403172

Xu Y., Xu Y., Lv T., Cui L., Wei F., Wang G. *et al*. (2022). *Layoutlmv2: Multi-modal pre-training for visually-rich document understanding.* Consulté sur https://arxiv.org/abs/2012.14740

Yang S., Xiao W., Zhang M., Guo S., Zhao J., Shen F. (2023). *Image data augmentation for deep learning: A survey.* Consulté sur https://arxiv.org/abs/2204.08610

Zhong X., Tang J., Yepes A. J. (2019). *Publaynet: largest dataset ever for document layout analysis.* Consulté sur https://arxiv.org/abs/1908.07836