Classification automatique de données déséquilibrées et bruitées : application aux exercices de manuels scolaires

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

Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Classification automatique de données déséquilibrées et bruitées : application aux exercices de manuels scolaires

Elise Lincker[1] Camille Guinaudeau[2,3] Olivier Pons[1] Jérôme Dupire[1]
Céline Hudelot[4] Vincent Mousseau[4] Isabelle Barbet[1] Caroline Huron[5,6]
(1) Cedric, CNAM, Paris, France
(2) Japanese French Laboratory for Informatics, CNRS, NII, Tokyo, Japon
(3) Université Paris-Saclay, Orsay, France
(4) MICS, CentraleSupélec, Université Paris-Saclay, Orsay, France
(5) SEED, Inserm, Université Paris Cité, Paris, France
(6) Learning Planet Institute, Paris, France
elise.lincker@lecnam.net, guinaudeau@nii.ac.jp, olivier.pons@lecnam.net, jerome.dupire@lecnam.net, celine.hudelot@centralesupelec.fr, vincent.mousseau@centralesupelec.fr, isabelle.barbet@lecnam.net, caroline.huron@cri-paris.org

RÉSUMÉ

Pour faciliter l'inclusion scolaire, il est indispensable de pouvoir adapter de manière automatique les manuels scolaires afin de les rendre accessibles aux enfants dyspraxiques. Dans ce contexte, nous proposons une tâche de classification des exercices selon leur type d'adaptation à la dyspraxie. Nous introduisons un corpus d'exercices extraits de manuels de français de niveau élémentaire, qui soulève certains défis de par sa petite taille et son contenu déséquilibré et bruité. Afin de tirer profit des modalités textuelles, structurelles et visuelles présentes dans nos données, nous combinons des modèles état de l'art par des stratégies de fusion précoce et tardive. Notre approche atteint une exactitude globale de 0.802. Toutefois, les expériences témoignent de la difficulté de la tâche, particulièrement pour les classes minoritaires, pour lesquelles l'exactitude tombe à 0.583.

ABSTRACT

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

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's adaptation. We introduce an exercises dataset automatically extracted from French textbooks, with two major difficulties : a small size and an unbalanced and 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.

MOTS-CLÉS : adaptation de manuels scolaires, classification multimodale, données bruitées, données déséquilibrées.

KEYWORDS: textbook adaptation, multimodal document classification, noisy data, unbalanced data.

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 4 : articles déjà soumis ou acceptés en conférence internationale, pages 121–130.

1 Introduction et travaux connexes

La dyspraxie est un trouble développemental de la coordination affectant 5% des enfants, qui interfère avec la réussite scolaire et les activités de la vie quotidienne. Plus précisément, les enfants dyspraxiques n'automatisent pas l'écriture manuscrite et leurs troubles des mouvements oculaires peuvent les empêcher de lire un texte si sa présentation n'est pas adaptée. Ainsi, pour que les enfants dyspraxiques réussissent à l'école, les manuels utilisés en classe doivent tenir compte de leurs difficultés d'écriture et d'organisation du regard. Une adaptation au format numérique permet de contourner le déficit d'écriture manuscrite sans modifier le contenu des exercices et leur objectif pédagogique. La Figure 1 montre un exemple d'exercice de type « *choix multiple* » et son adaptation, permettant aux enfants de compléter la phrase en *cliquant* sur la bonne réponse. Des associations commencent à produire de tels manuels numériques adaptés, en effectuant toutes les transformations à la main. Malheureusement, étant donné la grande diversité des collections et le renouvellement des programmes d'enseignement, ces adaptations artisanales ne permettent pas de répondre aux besoins. D'autre part, les manuels scolaires sont peu explorés en Traitement Automatique du Langage Naturel (TALN), et la plupart des études existantes portent sur l'analyse du contenu linguistique (Green, 2019; Lucy *et al.*, 2020) ou la génération de questions (Ch & Saha, 2022; Ghosh, 2022), généralement dans des manuels de niveau universitaire. À notre connaissance, aucune ne traite des tâches de classification ou de formatage du contenu.

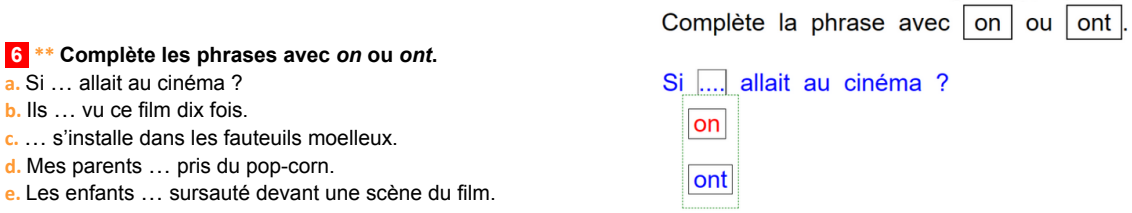

FIGURE 1 – Exercice à trous de type choix multiple et son adaptation.

Dans ce contexte, nous proposons une première étape vers l'automatisation de l'adaptation des manuels scolaires, avec la classification des exercices en fonction du type d'adaptation à la dyspraxie. Nous construisons un corpus d'exercices de manuels scolaires de français de niveau élémentaire annotés manuellement avec des étiquettes de type d'adaptation. Ce jeu de données reflète les difficultés de la tâche. Tout d'abord, le jeu de données est non seulement déséquilibré, certains types d'adaptation étant beaucoup plus fréquents que d'autres, mais aussi bruité, car il peut contenir des phrases agrammaticales ou incomplètes ainsi que des erreurs d'extraction. Ensuite, la classification concerne l'objectif pédagogique des exercices, qui peut être porté de manières très différentes. Enfin, les droits de propriété intellectuelle restreignent l'accès à un nombre limité de manuels scolaires, et conduisent par conséquent à un ensemble de données relativement petit. Notre approche de classification des exercices repose sur des stratégies de fusion précoce et tardive, afin de prendre en compte les informations sémantiques ainsi que structurelles et visuelles des documents.

Deux modèles de langue français basés sur RoBERTa (Liu *et al.*, 2019) – CamemBERT (Martin *et al.*, 2020), entraîné sur la partie française d'OSCAR (Suárez *et al.*, 2019) et FlauBERT (Le *et al.*, 2020), entraîné sur 24 corpus de styles variés collectés sur Internet – présentent des résultats similaires pour la classification. Cependant, ces modèles entraînés sur du texte sans erreur peuvent être impactés négativement par les erreurs présentes dans notre corpus, comme l'ont montré les études de Huang & Chen (2020) et Jiang *et al.* (2021). Alors que de nombreuses approches de classification reposent sur

le texte comme modalité unique, certaines études récentes se concentrent sur l'analyse de la mise en page. Ainsi, LayoutLM (Xu *et al.*, 2020) reprend l'architecture de BERT en ajoutant des plongements visuels et de position 2-D. Deux versions améliorées (Xu *et al.*, 2021b; Huang *et al.*, 2022) ont été récemment proposées, ainsi qu'une extension multilingue (Xu *et al.*, 2021a). BROS (Hong *et al.*, 2022) propose une méthode d'encodage spatial utilisant les positions relatives entre les blocs et DocFormer (Appalaraju *et al.*, 2021) introduit un mécanisme d'attention multimodal permettant le partage d'informations entre les modalités. Enfin, TILT (Powalski *et al.*, 2021) combine des caractéristiques convolutionnelles avec l'architecture T5 (Raffel *et al.*, 2020). Cependant, la plupart des modèles sont pré-entraînés et fine-tunés sur des documents monolingues, généralement en anglais. Pour dépasser cette limitation, LiLT (Wang *et al.*, 2022) permet d'imbriquer n'importe quel modèle pré-entraîné de type RoBERTa dans un module de mise en page. Par ailleurs, ces modèles sont entraînés sur des pages entières correctement formatées. Bien qu'ils puissent être efficaces pour la compréhension de pages de manuels, nous cherchons ici à catégoriser des exercices : des documents courts et très semblables.

Dans cet article, nos principales contributions sont : (i) l'introduction d'une nouvelle tâche de classification, pour l'adaptation automatique des exercices de manuels scolaires pour les enfants dyspraxiques ; (ii) un cadre de classification multimodale pour la tâche de classification des exercices de manuels scolaires ; (iii) des expériences avec différentes architectures multimodales, y compris le modèle LiLT (Wang *et al.*, 2022) récemment proposé.

2 Préparation du corpus

Le corpus est construit à partir de 3 manuels scolaires de français - étude de la langue, de niveau élémentaire (CE1 et CE2), au format PDF. Dans un premier temps, chaque manuel est converti en un document XML au format ALTO par pdfalto [1] couplé à MuPDF [2]. Cette combinaison d'outils OpenSource permet d'extraire les mots avec leur style de police et leurs coordonnées spatiales, ainsi que les images, dans une structure XML. Les mots extraits sont tokénisés et regroupés en segments grâce à des règles sur la taille et le style des polices, les types de caractères (chiffres, symboles, signes de ponctuation) et l'espacement entre les tokens ou les caractères. Une interface d'annotation conçue spécifiquement permet ensuite de réorganiser les segments dans une structure de manuel scolaire. Les segments sont étiquetés en rôles (*titre de chapitre*, *leçon*, *numéro d'exercice*, *consigne*, etc.) de manière semi-automatique sur la base de leur police dominante, puis les blocs d'activité sont reconstitués en utilisant des caractéristiques géométriques et une logique dans l'enchaînement. Par exemple, un numéro d'exercice précédé par un énoncé indique le début d'un nouvel exercice. Cette étape d'extraction aboutit à 2748 exercices, divisés en plusieurs parties : un numéro ou un nom, toujours une consigne, souvent un énoncé, et parfois aussi un exemple ou un conseil.

Toutefois, la complexité de la mise en page des manuels rend la tâche d'extraction difficile : les pages peuvent comporter des tableaux, des listes, des illustrations ou des blocs de texte épars, qui peuvent introduire du bruit dans les données. D'autre part, les notions de *phrase* et *token* ont été étendues en adéquation avec la nature du corpus. Si la plupart des consignes sont grammaticalement et sémantiquement correctes, les énoncés peuvent contenir des mots ou phrases à trous (« c...bat », « Manon a perdu ... chat. »), des choix de type choix multiple (« (son/sont) »), des suites de mots

1. https://github.com/kermitt2/pdfalto
2. https://github.com/ArtifexSoftware/mupdf

concaténés (« cirageâgégéantenfantfantômetomate »), des portions de phrases dispersées (« est une fleur », « la tulipe »), des numéros de liste (« a. », « b. »), etc. Par conséquent, il est plus approprié de faire référence à des segments de texte plutôt qu'à des phrases. De plus, un quart des segments est composé de 1 à 5 *tokens*, tandis que les segments les plus longs comptent jusqu'à 65 *tokens*. Avec 1 à 10 segments et 5 à 91 *tokens* par exercice, la longueur des documents est tout aussi variable mais reste courte par rapport aux jeux de données de référence.

Les exercices extraits sont manuellement annotés par 2 experts de la dyspraxie en 33 catégories. Ces catégories correspondent aux types d'adaptation à la dyspraxie et reflètent l'objectif pédagogique de l'exercice ainsi que le processus d'interaction impliqué dans sa résolution. Par exemple, la Figure 1 illustre un exercice de la classe *choix multiple*, et la Figure 2 en annexe contient 5 exercices d'autres classes. Résultent de cette annotation 2567 exercices qui n'ont qu'une seule étiquette, 146 exercices appartenant à plusieurs classes et 36 exercices à retirer des manuels adaptés, car ils demandent une compétence directement perturbée par le handicap. Pour ce travail, nous ne traitons que les exercices avec une seule étiquette et excluons la classe la moins représentée qui ne contient qu'un seul exercice. Ainsi, notre jeu de données final est composé de 2566 exercices étiquetés avec 32 catégories. Le jeu de données est très déséquilibré : 2 classes sur 32 comptent plus de 300 exercices, tandis que 11 classes en comptent moins de 20. Les 21 classes les plus fournies représentent 95% du corpus.

Le corpus est divisé en 3 sous-ensembles : apprentissage (70%), validation (10%) et test (20%). La proportion des classes d'exercices par manuel et le ratio entre les classes sont conservés. Les numéros et les noms des exercices sont supprimés. Si un exercice contient un exemple ou un conseil, ceux-ci sont concaténés avec la consigne. Le texte est normalisé en minuscules et tous les caractères d'espacement sont réduits à un espace.

3 Méthodologie

Notre objectif consiste à catégoriser les exercices extraits en fonction de leur type d'adaptation à la dyspraxie. Dans un premier temps, nous utilisons CamemBERT. Afin d'adapter le modèle au domaine scolaire, son modèle de langue masqué est fine-tuné sur les textes suivants : les leçons et exercices tirés de 4 manuels de français (les 3 manuels utilisés pour construire notre corpus, hormis les exercices présents dans les sous-corpus de validation et de test, et un 4ème manuel non annoté) ; 1293 *Fantastiques Exercices* de l'association *Le Cartable Fantastique* [3], qui fournit une collection d'exercices accompagnés de leur version interactive adaptée aux enfants dyspraxiques ; les textes de lecture originaux d'Alector (Gala *et al.*, 2020), corpus parallèle de 79 textes de lecture simplifiés au niveau lexical, syntaxique et discursif. Pour la phase d'apprentissage, la consigne et l'énoncé sont concaténés (séparés par le token spécial `<sep>`) avant d'être donnés au modèle. En vue d'exploiter d'autres modalités, nous utilisons LayoutLMv2 (Xu *et al.*, 2021b), qui prend en entrée des plongements textuels, positionnels et visuels. LayoutLMv2 est pré-entraîné sur IIT-CDIP (Lewis *et al.*, 2006) et fine-tuné pour la classification sur son sous-ensemble RVL-CDIP (Harley *et al.*, 2015), composé d'images de documents scannés tels que des lettres ou des formulaires. La plupart des documents utilisent l'anglais comme langue principale, mais IIT-CDIP contient quelques documents dans d'autres langues, dont le français. Si LayoutLMv2 traite des plongements de 3 modalités différentes, il est conçu pour du texte en anglais et peut ne pas bénéficier pleinement des caractéristiques textuelles de notre corpus.

3. https://www.cartablefantastique.fr/

Les 32 classes définies présentant une grande variabilité tant sémantique que structurelle, comme le montrent les exemples de la Figure annexe 2, nous pensons que le texte en français ainsi que la mise en page et l'image sont pertinents pour notre objectif de classification. Des approches de fusion sont mises en œuvre afin d'exploiter chacune de ces modalités. La première solution consiste à appliquer une fusion tardive au niveau des scores des classifieurs CamemBERT et LayoutLMv2. Les scores sont normalisés entre 0 et 1 avec la normalisation $Min - Max$, qui préserve les relations entre les valeurs d'origine, puis fusionnés avec les stratégies de fusion classiques *Moyenne* et *Maximum*. Dans la deuxième solution, nous tirons parti du modèle LiLT (Wang *et al.*, 2022), qui permet d'associer n'importe quel modèle pré-entraîné de type RoBERTa avec un module pré-entraîné sur la structure. Ce module, pré-entraîné sur IIT-CDIP, est combiné à notre version de CamemBERT fine-tunée sur des manuels scolaires et des textes de lecture, afin d'obtenir un modèle de type LayoutLM pour le français scolaire. Finalement, nous appliquons un vote majoritaire sur les prédictions de CamemBERT, LayoutLMv2 et LiLT[CamemBERT], LiLT[CamemBERT] étant le classifieur par défaut.

Enfin, pour faire face au déséquilibre des données, nous envisageons la configuration de la fonction de perte et les méthodes d'échantillonage, dites de *sampling*. Pour les tâches de classification sur des données équilibrées ou déséquilibrées, l'entropie croisée est la fonction de perte la plus largement utilisée. Dans un contexte de classification binaire de données déséquilibrées, la fonction de perte focale (Lin *et al.*, 2017) a été introduite comme une amélioration de l'entropie croisée classique, où les exemples faciles sont dynamiquement sous-pondérés. La perte focale a ensuite été étendue aux problèmes multi-classes et a montré des résultats prometteurs. L'augmentation du paramètre de focalisation γ permet de contrôler le poids des exemples faciles et de focaliser l'attention sur les exemples mal classés. Nous avons expérimenté cette fonction avec différentes valeurs pour γ, allant de 0 (ce qui correspond à l'entropie croisée pondérée) à 5. En ce qui concerne les approches de sampling et compte tenu de la taille de notre corpus, l'*undersampling* entraînerait une perte d'informations. Cependant, nous nous sommes inspirés des techniques d'apprentissage ensembliste pour construire plusieurs sous-ensembles sous-échantillonnés en y répartissant les exercices des classes majoritaires, puis en fusionnant les sorties des différents modèles formés sur ces sous-ensembles sous-échantillonnés.

Nos modèles utilisent l'architecture BASE et la longueur des séquences d'entrée est fixée à 256. Pour les expériences finales, la taille du batch est paramétrée à 16, le nombre d'époques entre 30 et 40 et le taux d'apprentissage initial entre 1e-5 et 1e-4. Nous utilisons l'optimiseur Adam et la fonction de perte d'entropie croisée pondérée par les effectifs des classes. Les résultats sur le corpus de test sont obtenus avec les modèles fine-tunés qui donnent les meilleurs résultats sur le corpus de validation.

4 Résultats et discussion

Le Tableau 1 présente les scores d'exactitude et de macro F-mesure pour la classification des exercices. La meilleure performance est atteinte par LiLT combiné à CamemBERT suivi d'une fusion tardive des 3 modèles : l'exactitude est alors de 0.802, ce qui indique que les 3 modèles sont complémentaires. Pour la majorité des classes, LayoutLMv2 est presque aussi performant que CamemBERT, bien qu'il capture moins bien les informations sémantiques qu'un modèle français. Cela met en évidence la pertinence des modalités structurelles et visuelles. Ce n'est toutefois pas suffisant, les performances de LayoutLM étant très faibles pour les classes sous-représentées. L'augmentation des scores avec les stratégies de fusion tardive et précoce est statistiquement significative par rapport à LayoutLMv2.

Modèle	Exactitude			Macro F1
	Total	Maj.	Min.	Total
Baseline Classe Majoritaire	0.147			0.008
CamemBERT^{T}	0.775	0.788	0.500	0.663
LayoutLMv2^{T+L+I}	0.708	0.722	0.250	0.487
CamemBERT + LayoutLMv2 (Fusion Maximum)	0.767	0.784	0.417	0.627
CamemBERT + LayoutLMv2 (Fusion Moyenne)	0.782	0.796	0.500	0.664
LiLT^{T+L}[CamemBERT]	0.786	0.796	**0.583**	0.696
CamemBERT + LayoutLMv2 + LiLT[CamemBERT]	**0.802**	**0.813**	**0.583**	**0.714**

TABLE 1 – Résultats sur l'ensemble des données de test (Total), les 21 classes majoritaires (Maj.) et les 11 classes minoritaires (Min.). Pour chaque modèle, on rappelle s'il a été pré-entraîné sur le texte (T), la mise en page (L) ou l'image (I).

	$\mathbf{B^-L^-}$	$\mathbf{B^-L^+}$	$\mathbf{B^+L^-}$	$\mathbf{B^+L^+}$
# exercices	79	74	36	321
LiLT	16 (20%)	61 (82%)	20 (56%)	304 (95%)
Fusion Maximum	0	54 (73%)	16 (44%)	321 (100%)
Fusion Moyenne	0	59 (80%)	19 (53%)	321 (100%)

TABLE 2 – Comparaison des classifications pour les 3 stratégies de fusion sur les exercices correctement ($^+$) et incorrectement ($^-$) classifiés par CamemBERT (B) et LayoutLMv2 (L).

Par rapport à CamemBERT, la fusion via LiLT et la fusion tardive *Moyenne* améliorent légèrement l'exactitude globale. Si cette amélioration ne semble pas significative, les scores sur les classes minoritaires révèlent un écart plus important entre les classifieurs CamemBERT et LiLT[CamemBERT]. Les méthodes de fusion surpassent les modèles individuels, soulignant ainsi l'importance des trois modalités pour le traitement des données issues de manuels scolaires. La combinaison du français scolaire et de la mise en page est particulièrement efficace avec LiLT. En outre, des expériences complémentaires sur CamemBERT confirment l'impact positif du fine-tuning du modèle de langue masqué sur un corpus scolaire, l'exactitude augmentant de 0,747 à 0,775.

Cependant, l'application de la perte focale pour faire face au déséquilibre des données s'avère inefficace. Selon le paramétrage de γ, elle conduit à des scores égaux ou inférieurs à ceux obtenus avec l'entropie croisée pondérée. Par ailleurs, le sous-échantillonnage d'un très petit ensemble de données n'est pas efficace car des informations sont perdues. Au mieux, nous obtenons une précision de 0,730 en utilisant des sous-ensembles sous-échantillonnés.

Enfin, des expériences supplémentaires ont été menées pour évaluer la généralisation des modèles intra- et inter-collection de manuels scolaires. De nouveaux modèles ont été entraînés sur des exercices issus de deux manuels de même collection, tandis que le troisième manuel, d'une collection différente, a été utilisé à des fins d'évaluation uniquement. Les résultats révèlent que la capacité de généralisation des modèles est plus élevée à partir des caractéristiques textuelles que positionnelles. En effet, LayoutLMv2 ne parvient pas à généraliser efficacement entre les collections et requiert une quantité de données plus importante que CamemBERT pour obtenir des résultats satisfaisants. La fusion

précoce avec LiLT continue de surpasser les approches à modèle unique et démontre de bonnes capacités de généralisation sur des collections distinctes.

Le Tableau 2 permet la comparaison des méthodes de fusion. Les 3 méthodes peuvent améliorer les prédictions. Bien que LiLT ne saisisse pas la totalité des exercices correctement prédits par CamemBERT et LayoutLMv2 individuellement, il corrige 20% des exercices mal classés par les deux classifieurs. Les scores obtenus avec les stratégies de fusion tardive démontrent que CamemBERT est davantage confiant [4] et fiable que LayoutLMv2. En revanche, LayoutLMv2 gère mieux les exercices où la mise en page prévaut sur le contenu sémantique. Il catégorise correctement 36 exercices que CamemBERT n'a pas su catégoriser, et environ la moitié de ces prédictions sont conservées par les stratégies de fusion *Moyenne* et *Maximum*. Enfin, les classes minoritaires sont les plus difficiles à traiter. Les scores détaillés au niveau des classes soulignent les difficultés posées par le déséquilibre des données. Cela reste un problème délicat qui, pour notre objectif de classification, nécessite une augmentation des données. Cette tâche se révèle complexe en raison de la quantité de données et de la spécificité du langage des manuels, en particulier dans les consignes.

5 Conclusion

Pour favoriser l'inclusion scolaire et avec un objectif à long terme d'adapter automatiquement des manuels scolaires complets pour les rendre accessibles aux enfants dyspraxiques, nous avons introduit dans cet article une nouvelle tâche : la classification des exercices de manuels scolaires en fonction de leur type d'adaptation. Nous avons mené une étude comparative de méthodes de classification neuronales sur notre propre jeu de données composé de 2566 exercices de manuels de français.

Nous avons proposé différentes approches basées sur trois modèles pré-entraînés ayant fait leurs preuves sur de nombreuses tâches de TALN et des corpus de référence. Nous avons cherché à exploiter différentes modalités et avons finalement obtenu un score d'exactitude de 0,802 en utilisant des méthodes de fusion. Les expériences ont démontré l'importance de la mise en page et de l'image en plus du texte dans la compréhension des manuels.

Afin d'améliorer ces résultats prometteurs, nos travaux futurs se concentreront sur l'étape d'extraction et le nettoyage des données. Par ailleurs, si l'exactitude globale est encourageante, les résultats pour les classes minoritaires doivent encore être améliorés. Nous prévoyons de générer artificiellement des exercices pour résoudre les problèmes de petite taille et de déséquilibre entre les classes.

Remerciements

Les auteurs remercient les relecteurs anonymes pour leurs commentaires constructifs, ainsi que Guillaume Faure pour le développement de la plateforme d'extraction et d'annotation. Ce travail a été réalisé dans le cadre du projet MALIN (MAnuels scoLaires INclusifs) sous le financement ANR-21-CE38-0014. Il a bénéficié d'un accès au cluster de calcul Lab-IA.

4. La différence entre le score le plus élevé et le suivant est plus importante.

Références

APPALARAJU S., JASANI B., KOTA B. U., XIE Y. & MANMATHA R. (2021). Docformer : End-to-end transformer for document understanding. In *Proceedings of the 18th IEEE International Conference on Computer Vision*.

CH D. R. & SAHA S. K. (2022). Generation of multiple-choice questions from textbook contents of school-level subjects. *IEEE Transactions on Learning Technologies*.

DIAS G., Éd. (2015). *Actes de TALN 2015 (Traitement automatique des langues naturelles)*, Caen. ATALA, HULTECH.

GALA N., TACK A., JAVOUREY-DREVET L., FRANÇOIS T. & ZIEGLER J. C. (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*.

GHOSH K. (2022). Remediating textbook deficiencies by leveraging community question answers. *Education and Information Technologies*.

GREEN C. (2019). A multilevel description of textbook linguistic complexity across disciplines : Leveraging NLP to support disciplinary literacy. *Linguistics and Education*.

HARLEY A. W., UFKES A. & DERPANIS K. G. (2015). Evaluation of deep convolutional nets for document image classification and retrieval. In *Proceedings of the 13th International Conference on Document Analysis and Recognition*.

HONG T., KIM D., JI M., HWANG W., NAM D. & PARK S. (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*.

HUANG C.-W. & CHEN Y.-N. (2020). Learning ASR-robust contextualized embeddings for spoken language understanding. In *Proceedings of the 45th IEEE International Conference on Acoustics, Speech and Signal Processing*.

HUANG Y., LV T., CUI L., LU Y. & WEI F. (2022). LayoutLMv3 : Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia*.

JIANG M., HU Y., WORTHEY G., DUBNICEK R. C., UNDERWOOD T. & DOWNIE J. S. (2021). Impact of OCR quality on BERT embeddings in the domain classification of book excerpts. In *Proceedings of the Conference on Computational Humanities Research*.

LE H., VIAL L., FREJ J., SEGONNE V., COAVOUX M., LECOUTEUX B., ALLAUZEN A., CRABBÉ B., BESACIER L. & SCHWAB D. (2020). FlauBERT : Unsupervised language model pre-training for french. In *Proceedings of the 12th Language Resources and Evaluation Conference*.

LEWIS D., AGAM G., ARGAMON S., FRIEDER O., GROSSMAN D. & HEARD J. (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*.

LIN T.-Y., GOYAL P., GIRSHICK R., HE K. & DOLLÁR P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*.

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L. & STOYANOV V. (2019). RoBERTa : A robustly optimized BERT pretraining approach. *arXiv preprint arXiv :1907.11692*.

LUCY L., DEMSZKY D., BROMLEY P. & JURAFSKY D. (2020). Content analysis of textbooks via natural language processing : Findings on gender, race, and ethnicity in texas US history textbooks. *AERA Open*.

MARTIN L., MULLER B., ORTIZ SUÁREZ P. J., DUPONT Y., ROMARY L., DE LA CLERGERIE , SEDDAH D. & SAGOT B. (2020). CamemBERT : a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

POWALSKI R., BORCHMANN L., JURKIEWICZ D., DWOJAK T., PIETRUSZKA M. & PALKA G. (2021). Going full-TILT boogie on document understanding with text-image-layout transformer. In *Proceedings of 16th International Conference on Document Analysis and Recognition*.

RAFFEL C., SHAZEER N., ROBERTS A., LEE K., NARANG S., MATENA M., ZHOU Y., LI W. & LIU P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*.

SUÁREZ P. J. O., SAGOT B. & ROMARY L. (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*.

WANG J., JIN L. & DING K. (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*.

XU Y., LI M., CUI L., HUANG S., WEI F. & ZHOU M. (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*.

XU Y., LV T., CUI L., WANG G., LU Y., FLORENCIO D., ZHANG C. & WEI F. (2021a). LayoutXLM : Multimodal pre-training for multilingual visually-rich document understanding. *arXiv preprint arXiv :2104.08836*.

XU Y., XU Y., LV T., CUI L., WEI F., WANG G., LU Y., FLORENCIO D., ZHANG C., CHE W., ZHANG M. & ZHOU L. (2021b). 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*.

Annexes

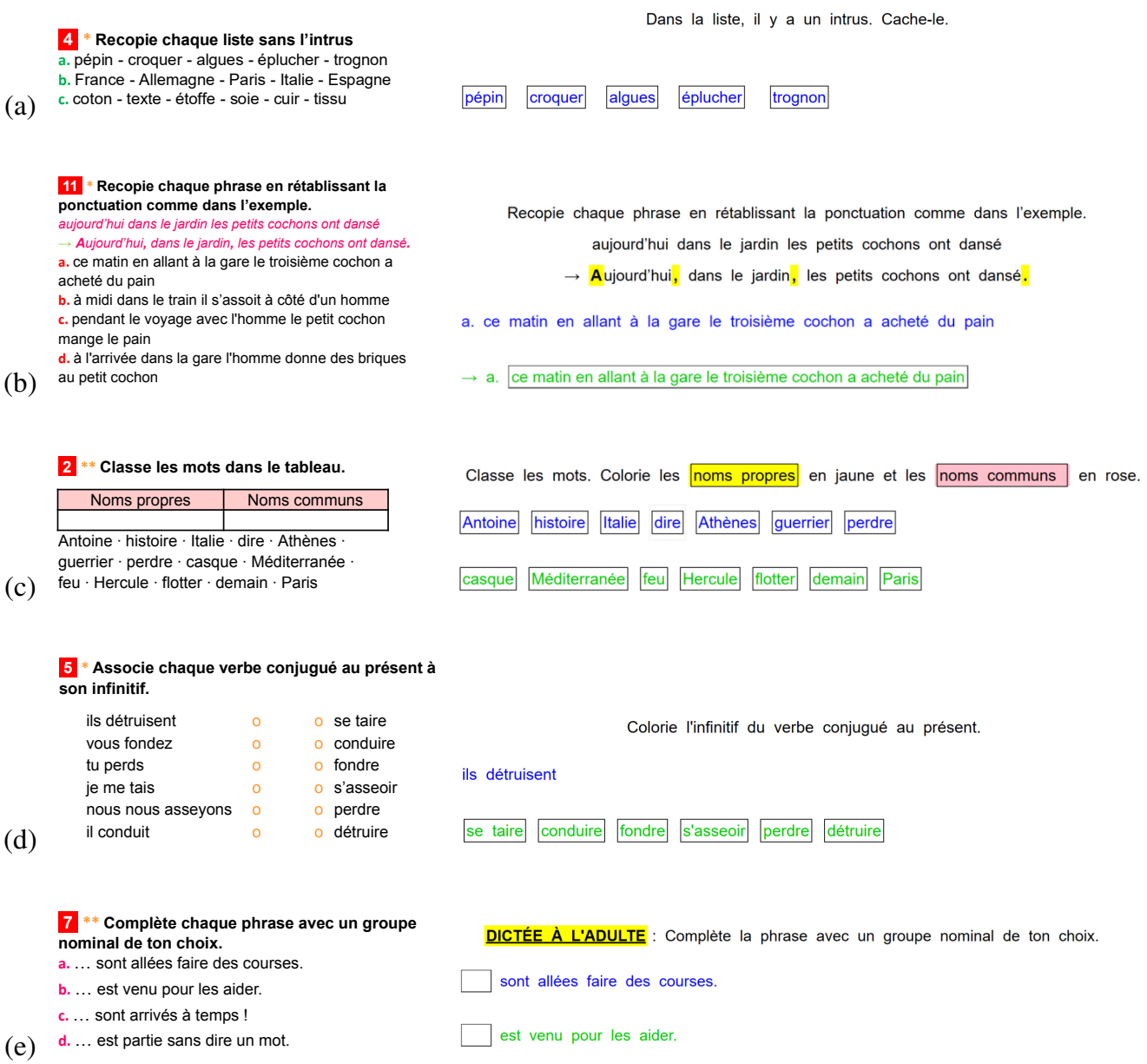

FIGURE 2 – Exemples d'exercices et exercices adaptés pour les classes (a) CacheIntrus, (b) EditePhrase, (c) Classe, (d) Associe, (e) RemplirAuClavier