Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Prédiction de la complexité lexicale : Une étude comparative entre ChatGPT et un modèle dédié à cette tâche.

Abdelhak Kelious[1] Mathieu Constant[1] Christophe Coeur[2]
(1) Université de Lorraine et CNRS/ATILF, France
(2) Consultant, France
(1) (abdelhak.kelious,mathieu.constant)@univ-lorraine.fr,
christophe.coeur@gmail.com

RÉSUMÉ

Cette étude s'intéresse à la prédiction de la complexité lexicale. Nous explorons des méthodes d'apprentissage profond afin d'évaluer la complexité d'un mot en se basant sur son contexte. Plus spécifiquement, nous examinons comment utiliser des modèles de langue pré-entraînés pour encoder le mot cible et son contexte, en les combinant avec des caractéristiques supplémentaires basées sur la fréquence. Notre approche obtient de meilleurs résultats que les meilleurs systèmes de SemEval-2021 (Shardlow *et al.*, 2021) sur cette tâche. Enfin, nous menons une étude comparative avec ChatGPT afin d'évaluer son potentiel pour prédire la complexité lexicale en comparaison avec un modèle dédié à cette tâche.

ABSTRACT

Lexical Complexity Prediction : a Comparative Study Between ChatGPT and a Dedicated Model for this Task

This study focuses on lexical complexity prediction. We explore deep learning methods to assess the complexity of a word based on its context. Specifically, we investigate how to use pre-trained language models to encode both the sentence and the target word, and then fine-tune them by combining them with additional frequency-based features. Our approach outperforms the best systems in SemEval-2021 (Shardlow *et al.*, 2021). Finally, we conduct a comparative study with ChatGPT to assess its potential for predicting lexical complexity compared to a model dedicated to this task.

MOTS-CLÉS : Traitement du langage naturel, Prédiction de la complexité lexicale, Modèles linguistiques, ChatGPT.

KEYWORDS: Natural language processing, Lexical complexity prediction, Language models, ChatGPT.

1 Introduction

Comprendre le langage est un défi qui mobilise de nombreuses compétences linguistiques, dont la maîtrise de la grammaire, l'enrichissement du vocabulaire et la production de discours cohérents. La complexité lexicale, influencée par la fréquence et le contexte d'utilisation des mots, joue un rôle crucial dans l'apprentissage des langues, affectant la rapidité et l'aisance avec lesquelles les apprenants maîtrisent de nouvelles compétences. Les difficultés de compréhension surgissent notamment quand des mots apparaissent dans des contextes peu familiers, poussant le lecteur à mal interpréter, ignorer

31ème Conférence sur le Traitement Automatique des Langues Naturelles,
volume 1 : articles longs et prises de position, pages 617–629.

ou perdre le fil du texte. Le traitement automatique des langues offre des outils pour identifier ces mots complexes.

La prédiction de la complexité lexicale constitue un champ de recherche dédié à estimer la difficulté des mots dans un contexte spécifique. Cette difficulté est mesurée sur une échelle continue entre 0 et 1. Les études dans ce domaine ont exploité des caractéristiques lexicales comme la fréquence des mots ou la longueur des phrases (Zampieri *et al.*, 2016) et des modèles de langue avancés tels que BERT (Devlin *et al.*, 2019). Cette tâche a des applications pratiques notables, surtout en didactique des langues et en compréhension de texte (Alfter, 2021).

Notre recherche s'inscrit dans la continuité des efforts récents, notamment ceux de SemEval-2021, et propose une méthode associant des modèles de langue pré-entraînés comme DeBERTa (He *et al.*, 2023) à des analyses de fréquence textuelle pour prédire la complexité des mots. Par ailleurs, nous examinons la capacité des modèles de langue de grande taille (LLMs) génératifs, tels que ChatGPT, à réaliser cette tâche par rapport à un modèle spécifiquement appris pour la prédiction de la complexité lexicale.

Les principales contributions de cet article sont les suivantes :

— Un nouveau modèle de prédiction de complexité lexicale entraîné sur le jeu de données Complex 2.0 (Shardlow *et al.*, 2021), combinant des modèles de langue pré-entraînés avec des caractéristiques de fréquence. Ce modèle obtient des performances dépassant celles de la compétition SemEval 2021 avec ces même données.
— Une évaluation comparative de la capacité de ChatGPT pour notre tâche. Les résultats montrent qu'il n'est pas performant pour prédire un score de complexité, mais il présente une certaine capacité à classer les contextes selon leur complexité, en particulier lorsque les contextes sont clairement difficiles ou clairement faciles.

2 Travaux connexes

2.1 Identification des mots complexes : méthodes et ensembles de données

Il existe plusieurs raisons d'évaluer l'identification des mots complexes dans une phrase (contexte), et cela fait l'objet de recherches depuis plusieurs années. Par exemple, cela a été exploré dans le contexte de la simplification lexicale, qui consiste à remplacer automatiquement les mots complexes par des alternatives plus simples (Shardlow, 2013).

La tâche d'identification des mots complexes (CWI : Complex Word Identification) a été étudiée en tant que tâche d'annotation de séquences, prenant en compte le contexte (phrase) dans lequel le mot apparaît (Gooding & Kochmar, 2019). Cette tâche a particulièrement été mise en avant lors de la compétition SemEval en 2021 *Lexical Complexity Prediction* (Shardlow *et al.*, 2021), ce qui a permis de mettre en place un jeu de données dédié de référence ainsi que de standardiser des métriques d'évaluation. Cette compétition a mis en évidence l'impact de l'utilisation de modèles de langue pré-entraînés pour la CWI, ainsi que l'utilisation de diverses techniques d'affinage (fine-tuning). Par exemple, l'un des meilleurs systèmes, JUST BLUE (Bani Yaseen *et al.*, 2021), combine les modèles BERT et RoBERTa (Liu *et al.*, 2019). Les modèles ont été affinés séparément pour prédire les scores de complexité, le score final étant leur moyenne. Un autre exemple est le système DeepBlueAI (Pan *et al.*, 2021) qui intègre des modèles de langue pré-entraînés affinés avec des techniques telles que le

pseudo-étiquetage, l'augmentation de données, les modèles d'entraînement empilés et l'utilisation d'une couche de *dropout* multi-échantillon.

Il existe divers jeux de données pour l'identification des mots complexes. Tout d'abord, CWI-2016 (Paetzold & Specia, 2016) est un jeu de données pour l'analyse de la complexité des mots en anglais en contexte, annoté par des locuteurs non natifs. Ce jeu de données a deux versions : une première version donnant pour chaque instance (mot-cible dans un contexte donné) les étiquettes binaires des 20 annotateurs ou annotatrices (1 si elles ou ils la jugeaient complexe, 0 sinon) ; une deuxième version attribue une seule étiquette à chaque instance, 1 si au moins un des 20 annotateurs ou annotatrices la jugeait complexe, sinon 0. CWI-2018 (Yimam *et al.*, 2018) est un jeu de données pour l'analyse de la complexité des mots dans plusieurs langues (anglais, espagnol, allemand et français uniquement pour les tests pour ce dernier), annoté par des locuteurs natifs et non natifs. Le jeu de données fournit deux types d'étiquetage : une étiquette binaire (0 ou 1) si l'instance est considérée comme facile ou difficile par les annotateurs ; une valeur réelle entre 0 et 1 indiquant la proportion d'annotateurs trouvant l'instance difficile. Dans cet article, nous utilisons un jeu de données plus récent Complex 2.0 (Shardlow *et al.*, 2022). Ces données ont été annotées manuellement en anglais en utilisant une échelle de Likert à 5 points pour indiquer le degré de complexité. Chaque instance (mot-cible dans un contexte donné) est ainsi annotée par une valeur réelle étant la moyenne des scores likert donnés par les différents annotateurs ou annotatrices normalisée entre 0 et 1. Complex 2.0 représente une amélioration significative par rapport à la version précédente, Complex 1.0 (Shardlow *et al.*, 2020) avec, notamment, un nombre d'annotateurs supplémentaires. Ce jeu de données offre ainsi un réglage plus précis pour identifier la complexité des mots.

La tâche de la prédiction de la complexité lexicale s'apparente d'une certaine manière à la tâche de prédiction automatique de la lisibilité d'un texte qui utilise des techniques proches. Les textes peuvent être simplifiés en fonction des scores de lisibilité, rendant l'information plus accessible (Wilkens *et al.*, 2022) (Wilkens *et al.*, 2024). Des recherches récentes indiquent que l'emploi de modèles de langue avancés, tels que ceux basés sur des *transformers*, combiné à l'application de caractéristiques linguistiques spécifiquement choisies, permet d'affiner l'évaluation de la lisibilité (Lee *et al.*, 2021). Cependant, bien que ces caractéristiques linguistiques puissent améliorer les résultats sur des échantillons de taille réduite, leur impact reste variable sur les modèles d'apprentissage profond plus sophistiqués, ne se traduisant pas systématiquement par un avantage marqué (Deutsch *et al.*, 2020).

2.2 ChatGPT

Avec l'avènement des modèles génératifs, ChatGPT notamment, il est difficile d'approcher notre étude sans inclure une comparaison par rapport à ce type de modèle. Des études récentes ont démontré le potentiel prometteur de ChatGPT [1] pour diverses tâches d'annotation (Dai *et al.*, 2023; Kuzman *et al.*, 2023) et de classification de texte (Liu *et al.*, 2023; Amin *et al.*, 2023; Zhang *et al.*, 2022). Parmi les tâches qui ont suscité l'intérêt de la communauté scientifique, il y a l'augmentation de données, et l'une des techniques d'augmentation de données utilisées par ChatGPT est la paraphrase. En reformulant le texte d'entrée de diverses manières, le modèle peut générer un ensemble plus diversifié d'exemples pour l'entraînement. Cela aide le modèle à saisir le sens sous-jacent du texte plutôt que de simplement mémoriser des phrases ou des motifs spécifiques. Par exemple, AugGPT (Dai *et al.*, 2023) reformule les phrases d'entraînement en plusieurs échantillons similaires mais sémantiquement différents, améliorant ainsi les performances du modèle. L'étude montre que AugGPT

1. Malgré un potentiel intéressant, ChatGPT a de nombreuses limites (Ray, 2023) que nous évoquerons au fil de l'article.

améliore significativement les performances du modèle. Huang *et al.* (2023) comparent ChatGPT aux annotateurs humains, révélant que ChatGPT peut détecter efficacement les tweets haineux avec une précision de 80%. Les désaccords restants sont généralement sujets à débat, mais les résultats montrent que les évaluations humaines tendent à soutenir les classifications de ChatGPT. L'étude montre aussi que ChatGPT produit des explications en langage naturel comparables à celles des humains en termes d'informativité et de clarté.

3 Un modèle dédié à l'identification des mots complexes

3.1 Données

Nous avons utilisé le jeu de données "CompLex 2.0" (Shardlow *et al.*, 2022) lors des phases d'entraînement et de test. Le corpus comprend des évaluations humaines de la complexité lexicale pour un ensemble de textes anglais, utilisant une échelle de Likert à 5 points, provenant de sources telles que Wikipedia, des livres éducatifs et des articles de journaux, couvrant divers sujets. Les textes ont été annotés par des évaluateurs humains qui ont évalué la complexité lexicale du mot cible dans son contexte (phrase) à l'aide de l'échelle de Likert. Chaque instance a été annotée plusieurs fois, et la moyenne de ces annotations a été prise comme score pour chaque instance de données. Ce score est normalisé en une valeur continue entre 0 et 1. La taille des données d'entraînement est de 7662 et de 917 pour le test.

3.2 Modèle

Dans notre étude, nous avons développé un modèle de réseau neuronal pour prédire la complexité des mots en utilisant des plongements lexicaux des mots basés sur des *transformers (encodeur)* et des caractéristiques telles que la fréquence du mot cible et les mots qui composent la phrase. Pour affiner les prédictions, nous avons intégré la loi de Zipf via la bibliothèque *wordfreq* (Speer, 2022), exploitant la fréquence d'occurrence des mots dans plus de 40 langues. Nous avons défini plusieurs caractéristiques d'entrée : F1 (le score Zipf de la fréquence des mots), F2 (le score Zipf moyen dans une phrase), F3 (la différence entre le score Zipf du mot cible et le score moyen), F4 (le nombre de mots avec un score Zipf supérieur au mot cible) et F5 (une valeur binaire indiquant si le mot cible est considéré comme rare avec un score inférieur ou égal 3). Ces caractéristiques sont combinées à des couches cachées pour capter la non-linéarité des données.

La formule de prédiction de complexité s'exprime comme suit dans notre modèle :

$$\hat{y} = f(W_h \cdot \sigma(W_e \cdot E + W_f \cdot F + b_e) + b_h)$$

où :

- E représente les plongements lexicaux des mots basés sur des *transformers*,
- F est le vecteur des caractéristiques d'entrée $[F_1, F_2, F_3, F_4, F_5]$,
- W_e et W_f sont les poids appliqués respectivement aux plongements lexicaux et aux caractéristiques d'entrée.
- b_e et b_h sont les biais pour les couches d'entrée et cachées, respectivement.
- σ est une fonction d'activation non-linéaire (ReLU) , appliquée à la combinaison linéaire des plongements lexicaux et des caractéristiques d'entrée.

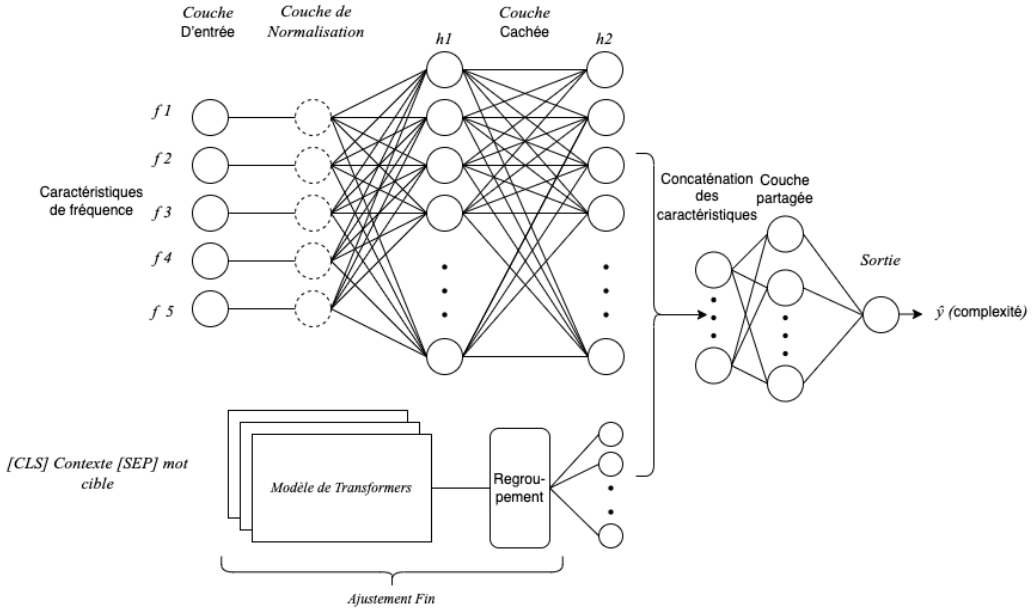

FIGURE 1 – L'architecture globale pour prédire les scores de complexité

— W_h sont les poids de la couche cachée.
— f est la fonction linéaire d'activation de la couche de sortie.
— $\hat{y}$ correspond à la valeur de prédiction (valeur réelle entre 0 et 1).

3.3 Évaluation

Comme le montre la figure 1, en combinant les différentes caractéristiques, nous avons utilisé le modèle pour faire des prédictions sur la complexité des mots. Nous avons mené nos expériences pour quatre modèles de plongements de mots : bert-base, DeBERTa-base-v3, ainsi que leurs versions large et multilingues. Pour évaluer les résultats, nous utilisons les mêmes métriques de corrélation que celles utilisées dans Shardlow *et al.* (2021) : Pearson, Spearman et le score R^2. Nous avons conservé les mêmes hyperparamètres pour chaque entraînement, incluant un taux d'apprentissage de 5e-5, une taille de lot de 4 et une longueur de séquence maximale de 300. Le tableau 1 montre les résultats de notre approche en utilisant les modèles de langue cités. Nous incluons également le meilleur score obtenu lors de la tâche partagée SemEval-2021 (Shardlow *et al.*, 2021). Le modèle de base fournie par les organisateurs de la tâche est reproduite en utilisant la log-fréquence du Google Web1T (Evert, 2010) et la régression linéaire. Nous constatons une amélioration significative des performances en utilisant le modèle de langue DeBERTa-large-v3, atteignant un score R^2 de **0.65**, supérieur aux performances obtenus dans la compétition de SemEval 2021. Il est à noter que la comparaison n'est pas totalement juste par rapport aux systèmes de SemEval 2021 puisque nous avons utilisé des modèles de langue plus récents (DeBERTa).

Afin de comprendre l'impact des modèles de langue pré-entraînés et des caractéristiques de fréquence dans chacun des modèles, nous avons mené une étude d'ablation pour mesurer l'importance de chaque composant. La figure 2 est une représentation graphique qui montre la performance des modèles pour prédire la complexité lexicale. Chaque colonne représente un modèle différent, avec différentes couleurs pour distinguer les types de modèles. La barre bleue représente un modèle

Modèles	Pearson	Spearman	R^2
Deberta-v3-large	**0,81**	**0,74**	**0,65**
Deberta-v3-base	0,79	0,74	0,62
Le score le plus élevé dans SemEval 2021 (Bani Yaseen et al., 2021)	*0,78*	-	*0,61*
mDeberta-v3-base	0,75	0,70	0,57
bert-base-cased	0,74	0,70	0,55
bert-base-multilingue	0,67	0,64	0,45
Baseline de fréquence fournie par (Shardlow *et al.*, 2021)	0,52	-	0,27

TABLE 1 – Résultats avec différents modèles de langues

qui utilise uniquement les caractéristiques de fréquence, la barre rouge représente un modèle basé uniquement sur des modèles de langue, tandis que la verte utilise à la fois les caractéristiques de fréquence et les modèles de langue. Les scores de corrélation de Pearson sont utilisés pour évaluer la performance des modèles en comparant leurs prédictions aux données de référence. Les résultats indiquent qu'ajouter des caractéristiques de fréquence aux modèles de langue améliore légèrement la prédiction de la complexité lexicale. En d'autres termes, les modèles qui utilisent à la fois des caractéristiques de fréquence et des modèles de langue sont meilleurs pour prédire la complexité lexicale que les modèles qui utilisent seulement l'un ou l'autre. On notera que notre modèle basé uniquement sur les caractéristiques de fréquence obtient de meilleurs résultats que la baseline dans le tableau 1 avec un score de corrélation de Pearson de 0,61.

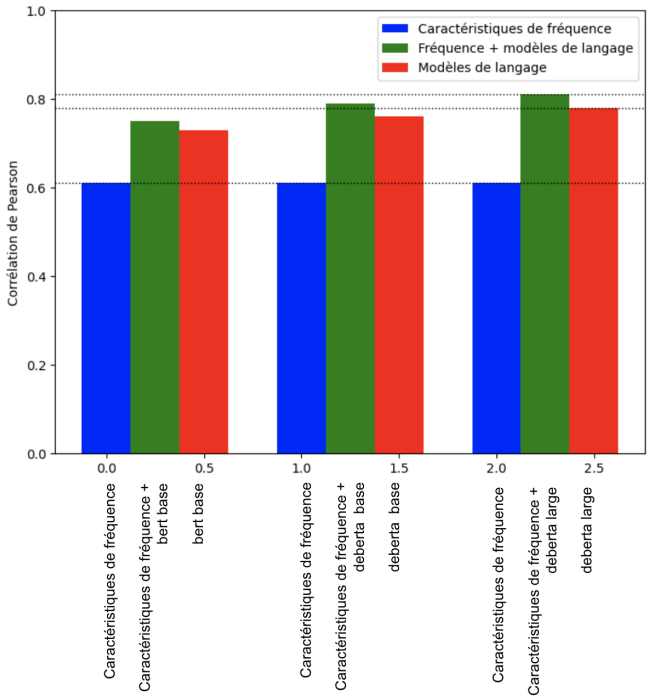

FIGURE 2 – Comparaison des valeurs de corrélation de Pearson pour les différents modèles.

4 Évaluation comparative avec ChatGPT

L'un des objectifs de cet article vise à comparer ChatGPT avec un modèle dédié pour mesurer la complexité lexicale, afin de voir si ChatGPT représente une alternative viable. Pour cette expérience, nous avons utilisé ChatGPT Turbo 3.5 (le 4 octobre 2023) via l'API fournie par OpenAI. Soulignons néanmoins une limite importante de notre expérience : ChatGPT est régulièrement mis à jour, ce qui pose des problèmes de reproductibilité (Chen *et al.*, 2023). Par ailleurs, il est difficile d'interpréter en profondeur les résultats obtenus puisque ChatGPT fonctionne comme une "boîte noire". Nous nous limiterons donc à des observations de surface des résultats.

4.1 Méthodologie de comparaison

Notre idée initiale était de fournir à ChatGPT une instance en entrée (un mot-cible et son contexte d'apparition) et de lui demander d'évaluer sa complexité, sur une échelle de 0 à 1. Cependant, nous avons rapidement constaté des résultats très médiocres, le score de corrélation de Pearson entre les évaluations humaines et ChatGPT étant de l'ordre de 0,034. Comme ChatGPT n'a pas été spécifiquement entraîné pour la tâche de prédiction de la complexité lexicale sur notre ensemble de données spécifique, une telle comparaison directe est quelque peu injuste. De plus, ChatGPT n'a pas accès à des informations complètes sur la façon dont les humains ont évalué ces données. Même avec une consigne sophistiquée, la tâche reste très complexe. Nous devons donc rendre l'évaluation plus équitable. Au lieu d'évaluer la capacité de ChatGPT à prédire un score de complexité pour une instance donnée (mot cible + contexte), nous évaluons sa capacité à comparer deux instances selon leur complexité, et ainsi à classer un ensemble d'instances selon celle-ci. En d'autres termes, nous éliminons la nécessité de prédire des scores entre 0 et 1, permettant à ChatGPT de se concentrer uniquement sur l'évaluation de l'ordre relatif de complexité parmi les instances. Pour ce faire, nous avons utilisé l'algorithme de tri à bulles pour trier une liste d'instances, où la comparaison entre deux instances est effectuée par ChatGPT [2]. Pour chaque paire d'instances à comparer, nous avons utilisé le *prompt* suivant :

> """
>
> I give you two sentences, evaluate the complexity of the target word in quotes based on its context, and return only the sentence or the target word that is simpler to understand. The output format should be as follows :
>
> {'simplest sentence' : sentence}
>
> The two sentences are :
>
> sentence 1
>
> sentence 2
>
> """

Ce prompt a été produit par essais-erreurs successifs [3]. La première difficulté a été de créer un prompt

2. Notons que pour un ensemble d'instances de taille importante, cet algorithme peut avoir un impact écologique non négligeable du fait d'un appel potentiel à ChatGPT pour chaque paire d'instances.

3. Le "prompt engineering" ne repose pas sur une méthodologie systématique, ce qui est une limite claire de cette approche.

qui produit de manière stable la sortie désirée. Le paramètre de température a été fixé à 0 pour éviter de favoriser la créativité pouvant entraîner des effets indésirables (Peng *et al.*, 2023).

Pour rendre la sortie pleinement comparable avec nos données de référence et notre approche décrite dans la section 1, la liste des instances est triée selon leurs scores de référence et leurs scores prédits. Pour l'évaluation, nous utilisons le *Tau de Kendall* (également connu sous le nom de *coefficient de corrélation de rang de Kendall*), qui est une mesure statistique utilisée pour quantifier la similarité entre deux classements. Il évalue la correspondance ou l'accord entre les classements du même ensemble d'éléments dans deux listes différentes. Le Tau de Kendall est fréquemment utilisé lorsqu'on travaille avec des données ordinales, où le classement ou l'ordre des éléments est pertinent. Nous emploierons cette métrique lors de la comparaison avec ChatGPT.

Nous divisons notre évaluation en deux expériences. Dans la première expérience, pour chaque mot cible, nous trions la liste des instances où le mot apparaît, l'objectif de cette expérience est d'évaluer la capacité de ChatGPT à prédire la complexité d'un même mot dans différents contextes. Dans la deuxième expérience, nous trions des listes d'instances choisies au hasard en variant la taille de l'échantillon, l'objectif cherche à déterminer si ChatGPT est capable de distinguer les niveaux de difficulté lorsque des mots différents sont utilisés dans des contextes variés.

4.1.1 Tri des contextes par mot cible

Dans la première expérience, pour chaque mot cible, la liste des instances où le mot apparaît est triée. Sur 917 instances, nous ne conservons que celles dont le mot-cible est le mot-cible d'au moins deux instances, réduisant l'ensemble à 685 instances. Les mots-cibles avec un seul contexte sont exclus pour éviter un biais dans la comparaison, car elles entraîneraient un taux de correspondance de 100%.

L'évaluation pour chaque mot cible de notre modèle et de ChatGPT repose sur le calcul du tau de Kendall, qui compare les classements obtenus à partir de notre modèle ou de ChatGPT avec ceux basés sur des annotations humaines. Le score global d'un système est dérivé de la moyenne des scores de tau de Kendall pour l'ensemble des mots cibles du jeu de données. Cette méthode révèle que, même sans entraînement spécifique sur le jeu de données, ChatGPT montre des performances supérieures à notre approche (cf. Table 2). Notons néanmoins que les performances de ChatGPT et de notre approche ont tendance à se rapprocher quand le nombre d'instances par mot-cible augmente.

Modèles	Score de Tau de Kendall
ChatGPT	**0.61**
Notre approche	0.52

TABLE 2 – Résultats des scores basés sur les classements.

4.1.2 Tri des instances échantillonnées

Dans cette partie, nous évaluons le classement des instances échantillonnées aléatoirement dans le jeu de données de test. Cela signifie que les listes considérées peuvent inclure des contextes avec divers mots cibles. L'idée est d'évaluer la capacité des différents systèmes à gérer la complexité des mots en général, indépendamment d'un mot cible donné. Dans notre expérience, nous varions la taille de l'échantillon pour évaluer son impact sur les performances de classement. Nous allons réaliser les

expériences pour 5 tailles d'échantillon (n=4, 5, 8, 10, 20). Pour chaque taille d'échantillon, nous effectuons 10 tirages aléatoires séparés pour obtenir des mesures plus robustes et améliorer notre évaluation. Nous calculons la moyenne et l'écart type de ces dix tirages pour obtenir une estimation plus précise.

n	Moyenne		Écart Type	
	ChatGPT	Notre modèle	ChatGPT	Notre modèle
4	0.145	**0.491**	0.566	0.391
5	0.011	**0.152**	0.391	0.288
8	0.062	**0.376**	0.275	0.281
10	-0.078	**0.153**	0.236	0.324
20	-0.079	**0.415**	0.065	0.095

TABLE 3 – Moyennes et écarts-types pour différentes tailles d'échantillons (100% aléatoire)

La Table 3 montre que notre modèle obtient systématiquement de meilleures performances par rapport à ChatGPT quelle que soit la taille de l'échantillon n. Par exemple, avec $n = 20$, notre modèle atteint un score de Kendall moyen de 0,415, tandis que ChatGPT obtient -0,079, ce qui indique une différence significative. De plus, à mesure que la taille de l'échantillon augmente, l'écart-type diminue principalement en raison de la réduction de la variabilité résultant d'un plus grand nombre d'instances, ce qui rend la moyenne une estimation plus précise de la tendance centrale.

Pourquoi cette tâche semble-t-elle être difficile pour ChatGPT ? L'une des hypothèses que nous voulons vérifier est que ChatGPT pourrait avoir des difficultés à distinguer les degrés de complexité lorsqu'ils sont proches. Pour tester cette hypothèse, nous continuerons à échantillonner aléatoirement des exemples, mais 50% des exemples auront un niveau de complexité facile (degré <0,25), et les 50% restants auront un niveau de complexité difficile (degré > 0,5). Cela créera une distinction claire entre les exemples en fonction de leur complexité.

n	Moyenne		Écart Type	
	ChatGPT	Notre modèle	ChatGPT	Notre modèle
4	**0.655**	0.425	0.261	0.321
5	**0.6**	0.567	0.249	0.3
8	**0.483**	0.412	0.154	0.142
10	**0.504**	0.429	0.116	0.099
20	**0.495**	0.432	0.077	0.091

TABLE 4 – Moyennes et écarts-types pour différentes tailles d'échantillons (aléatoire mais 50% contextes facile, 50% très difficile)

La Table 4 et la figure 3 montrent que les scores sont meilleurs avec ChatGPT dans ce scénario. ChatGPT se comporte de manière plus satisfaisante lorsque l'écart entre les niveaux de complexité des contextes est plus élevé, ce qui indique qu'il peut plus facilement différencier entre les contextes ayant des niveaux de complexité clairement distincts. Les valeurs d'écart-type deviennent également plus petites lorsqu'on les compare avec les résultats de la Table 3, ce qui indique que l'échantillon est plus représentatif et que les valeurs sont plus proches de la moyenne par rapport à lorsque les données sont échantillonnées de manière aléatoire à 100%.

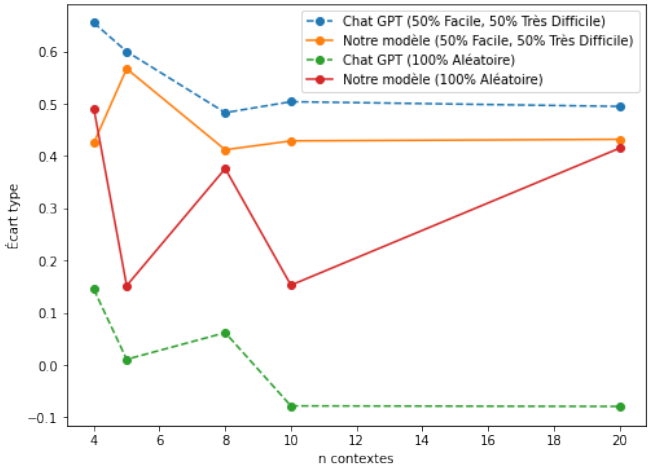

FIGURE 3 – Scores moyens en fonction de la taille de l'échantillon.

5 Conclusion

Dans cet article, nous avons proposé une méthode pour prédire la complexité lexicale. Cette méthode repose sur l'utilisation de modèles de langue pré-entraînés et de caractéristiques de fréquence. Nous avons montré que ChatGPT a une certaine capacité à comparer et classer des instances ayant le même mot-cible en fonction de leur complexité lexicale, notamment lorsque le nombre d'instances est limité. Dans de tels cas, il obtient de meilleures performances que notre modèle. Cependant, cette tâche devient plus difficile pour ChatGPT à mesure que le nombre d'instances comparées augmente, en particulier lors de l'introduction de différents mots-cibles dans des contextes variés. En effet, ChatGPT rencontre des difficultés lorsque la complexité entre les instances est très similaire. Dans de telles situations, il semble préférable d'utiliser un modèle spécifiquement entraîné pour cette tâche. De plus, un tel modèle peut produire un degré de complexité plus précis.

Pour les prochaines étapes de recherche, nous envisageons d'explorer davantage le potentiel du modèle dans des environnements multilingues. Cette exploration impliquera l'utilisation de méthodes d'apprentissage par transfert pour étendre la capacité de notre approche à prédire la complexité lexicale dans différentes langues. Nous envisageons également d'explorer l'utilisation d'autres modèles open source, en plus de ChatGPT, pour des questions de reproductibilité. Par ailleurs, nous prévoyons également d'explorer des techniques d'annotation et d'augmentation des données pour améliorer les performances de prédiction de la complexité lexicale. L'annotation supplémentaire des données pourrait aider à mieux capturer les nuances de la complexité lexicale dans différents contextes, tout en aidant à accroître la diversité des exemples disponibles pour l'entraînement du modèle, améliorant ainsi sa capacité à généraliser à de nouveaux contextes.

Références

ALFTER D. (2021). *Exploring natural language processing for single-word and multi-word lexical complexity from a second language learner perspective*.

AMIN M. M., CAMBRIA E. & SCHULLER B. W. (2023). Will affective computing emerge from foundation models and general ai ? a first evaluation on chatgpt. *arXiv preprint arXiv :2303.03186*.

BANI YASEEN T., ISMAIL Q., AL-OMARI S., AL-SOBH E. & ABDULLAH M. (2021). JUST-BLUE at SemEval-2021 task 1 : Predicting lexical complexity using BERT and RoBERTa pre-trained language models. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, p. 661–666, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.semeval-1.85.

BENAMARA F., HATOUT N., MULLER P. & OZDOWSKA S., Éds. (2007). *Actes de TALN 2007 (Traitement automatique des langues naturelles)*, Toulouse. ATALA, IRIT.

CHEN L., ZAHARIA M. & ZOU J. (2023). How is chatgpt's behavior changing over time ? *arXiv preprint arXiv :2307.09009*.

DAI H., LIU Z., LIAO W., HUANG X., CAO Y., WU Z., ZHAO L., XU S., LIU W., LIU N., LI S., ZHU D., CAI H., SUN L., LI Q., SHEN D., LIU T. & LI X. (2023). Auggpt : Leveraging chatgpt for text data augmentation.

DEUTSCH T., JASBI M. & SHIEBER S. (2020). Linguistic features for readability assessment. *arXiv preprint arXiv :2006.00377*.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). Bert : Pre-training of deep bidirectional transformers for language understanding.

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

EVERT S. (2010). Google web 1t 5-grams made easy (but not for the computer). In *Proceedings of the NAACL HLT 2010 Sixth Web as Corpus Workshop*, p. 32–40.

GOODING S. & KOCHMAR E. (2019). Complex word identification as a sequence labelling task. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 1148–1153, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1109.

HE P., GAO J. & CHEN W. (2023). Debertav3 : Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing.

HUANG F., KWAK H. & AN J. (2023). Is ChatGPT better than human annotators ? potential and limitations of ChatGPT in explaining implicit hate speech. In *Companion Proceedings of the ACM Web Conference 2023* : ACM. DOI : 10.1145/3543873.3587368.

KUZMAN T., LJUBEŠIĆ N. & MOZETIČ I. (2023). Chatgpt : beginning of an end of manual annotation ? use case of automatic genre identification. *arXiv preprint arXiv :2303.03953*.

LAIGNELET M. & RIOULT F. (2009). Repérer automatiquement les segments obsolescents à l'aide d'indices sémantiques et discursifs. In A. NAZARENKO & T. POIBEAU, Éds., *Actes de TALN 2009 (Traitement automatique des langues naturelles)*, Senlis : ATALA LIPN.

LANGLAIS P. & PATRY A. (2007). Enrichissement d'un lexique bilingue par analogie. In (Benamara *et al.*, 2007), p. 101–110.

LEE B. W., JANG Y. S. & LEE J. H.-J. (2021). Pushing on text readability assessment : A transformer meets handcrafted linguistic features. *arXiv preprint arXiv :2109.12258*.

LIU Y., HAN T., MA S., ZHANG J., YANG Y., TIAN J., HE H., LI A., HE M., LIU Z. *et al.* (2023). Summary of chatgpt-related research and perspective towards the future of large language models. *Meta-Radiology*, p. 100017.

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.

PAETZOLD G. & SPECIA L. (2016). SemEval 2016 task 11 : Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, p. 560–569, San Diego, California : Association for Computational Linguistics. DOI : 10.18653/v1/S16-1085.

PAN C., SONG B., WANG S. & LUO Z. (2021). DeepBlueAI at SemEval-2021 task 1 : Lexical complexity prediction with a deep ensemble approach. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, p. 578–584, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.semeval-1.72.

PENG K., DING L., ZHONG Q., SHEN L., LIU X., ZHANG M., OUYANG Y. & TAO D. (2023). Towards making the most of chatgpt for machine translation. *arXiv preprint arXiv :2303.13780*.

RAY P. P. (2023). Chatgpt : A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*.

SERETAN V. & WEHRLI E. (2007). Collocation translation based on sentence alignment and parsing. In (Benamara *et al.*, 2007), p. 401–410.

SHARDLOW M. (2013). A comparison of techniques to automatically identify complex words. In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, p. 103–109, Sofia, Bulgaria : Association for Computational Linguistics.

SHARDLOW M., COOPER M. & ZAMPIERI M. (2020). CompLex — a new corpus for lexical complexity prediction from Likert Scale data. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, p. 57–62, Marseille, France : European Language Resources Association.

SHARDLOW M., EVANS R., PAETZOLD G. H. & ZAMPIERI M. (2021). SemEval-2021 task 1 : Lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, p. 1–16, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.semeval-1.1.

SHARDLOW M., EVANS R. & ZAMPIERI M. (2022). Predicting lexical complexity in english texts : the complex 2.0 dataset. *Language Resources and Evaluation*, **56**(4), 1153–1194. DOI : 10.1007/s10579-022-09588-2.

SPEER R. (2022). rspeer/wordfreq : v3.0. DOI : 10.5281/zenodo.7199437.

WILKENS R., ALFTER D., WANG X., PINTARD A., TACK A., YANCEY K. P. & FRANÇOIS T. (2022). FABRA : French aggregator-based readability assessment toolkit. In N. CALZOLARI, F. BÉCHET, P. BLACHE, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 1217–1233, Marseille, France : European Language Resources Association.

WILKENS R., WATRIN P., CARDON R., PINTARD A., GRIBOMONT I. & FRANÇOIS T. (2024). Exploring hybrid approaches to readability : experiments on the complementarity between linguistic features and transformers. In Y. GRAHAM & M. PURVER, Éds., *Findings of the Association for Computational Linguistics : EACL 2024*, p. 2316–2331, St. Julian's, Malta : Association for Computational Linguistics.

YIMAM S. M., BIEMANN C., MALMASI S., PAETZOLD G., SPECIA L., ŠTAJNER S., TACK A. & ZAMPIERI M. (2018). A report on the complex word identification shared task 2018. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, p. 66–78, New Orleans, Louisiana : Association for Computational Linguistics. DOI : 10.18653/v1/W18-0507.

ZAMPIERI M., TAN L. & VAN GENABITH J. (2016). MacSaar at SemEval-2016 task 11 : Zipfian and character features for ComplexWord identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, p. 1001–1005, San Diego, California : Association for Computational Linguistics. DOI : 10.18653/v1/S16-1155.

ZHANG B., DING D. & JING L. (2022). How would stance detection techniques evolve after the launch of chatgpt ? *arXiv preprint arXiv :2212.14548*.