Modélisation de la lisibilité en français pour les personnes en situation d'illettrisme

Wafa Aissa[1, *], Thibault Bañeras-Roux[1, *] Elodie Vanzeveren[1] Lingyun Gao[1]
Alice Pintard[1] Rodrigo Wilkens[2] Thomas François[1]
(1) CENTAL, IL&C, UCLouvain, Belgique
(2) University of Exeter, Royaume-Uni
wafa.aissa@uclouvain.be, thibault.roux@uclouvain.be,
elodie.vanzeveren@uclouvain.be, lingyun.gao@uclouvain.be,
alice.pintard@gmail.com, r.wilkens@exeter.ac.uk,
thomas.francois@uclouvain.be
* Ces auteurs ont contribué de manière égale à cet article.

RÉSUMÉ

Nous présentons une nouvelle formule de lisibilité en français spécifiquement conçue pour les personnes en situation d'illettrisme. À cette fin, nous avons construit un corpus de 461 textes annotés selon une échelle de difficulté spécialisée à ce public. Dans un second temps, nous avons systématiquement comparé les principales approches en lisibilité, incluant l'apprentissage automatique reposant sur des variables linguistiques, le fine-tuning de CamemBERT, une approche hybride combinant CamemBERT et des variables linguistiques et des modèles de langue génératifs (LLMs). Une analyse approfondie de ces modèles et de leurs performances est menée afin d'évaluer leur applicabilité dans des contextes réels.

ABSTRACT

Modeling French Readability for Adults with Low Literacy

We present a new French readability formula specifically designed for people with low literacy. To this end, we built a corpus of 461 texts annotated according to a specialized difficulty scale tailored to this audience. We then conducted a systematic comparison of the main approaches to readability, including machine learning based on linguistic variables, CamemBERT fine-tuning, a hybrid approach combining CamemBERT and linguistic variables, and generative large language models (LLMs). An in-depth analysis of these models and their performance is carried out in order to assess their applicability in real-life contexts.

MOTS-CLÉS : lisibilité, illettrisme, français, TAL, modèles hybrides, modèles génératifs.

KEYWORDS: readability, illiteracy, French, NLP, hybrid models, generative AI.

ARTICLE : **Accepté à** CORIA-TALN 2025.

1 Introduction

L'illettrisme est défini par l'Agence de lutte contre l'illetrisme (ANLCI) comme la situation de personnes « qui, bien qu'ayant été scolarisées, ne parviennent pas à lire et à comprendre un texte portant sur des situations de leur vie quotidienne » (Besse *et al.*, 2009, 31). Ce déficit en lecture entraîne de nombreuses conséquences dommageables, tant pour les individus que pour les sociétés. Ainsi, les personnes en situation d'illettrisme rencontrent davantage de problèmes de santé (Berkman *et al.*, 2011), voient leur espérance de vie réduite (Messias, 2003) et sont payées 30 à 42% de moins en moyenne (Lal, 2015), notamment en conséquence d'une moins bonne insertion professionnelle. Au niveau des sociétés, il est estimé que l'impact de l'illettrisme sur le PIB des pays développés est de 2% (Steward, 2023).

Face à ces constats, les gouvernements ont depuis longtemps pris des mesures pour réduire l'illettrisme, notamment au travers de divers programmes d'aide à l'apprentissage de la lecture (ex. ANLCI en France). Un exemple de contenu efficace pour ces programmes est « Alpha Plus » (Russeler *et al.*, 2012), qui a produit de meilleurs résultats sur un groupe expérimental par rapport à un groupe contrôle suivant une formation plus classique. Parallèlement, depuis la grande dépression des années 1930 et la volonté de mieux outiller les nombreux travailleurs sans emploi via des formations à la lecture, des formules de lisibilité ont été développées par divers chercheurs pour les personnes en situation d'illettrisme (DuBay, 2004). Ce type d'outil, qui évalue automatiquement la difficulté de lecture de textes, permet d'associer efficacement et rapidement textes et lecteurs dans le cadre de formations. Néanmoins, pour le français, comme nous le discuterons plus en détail à la section 2, il n'existe actuellement aucune formule spécialisée pour ce public.

En l'absence de formules dédiées, il est courant de réutiliser des formules développées pour un public différent, ce qui est loin d'être optimal, comme François (2011) l'a discuté pour le cas du français langue étrangère. C'est pourquoi, nous proposons, dans cet article, la première formule de lisibilité pour le français spécifiquement dédiée aux personnes en situation d'illettrisme. Cette formule peut être considérée comme spécialisée en ce sens qu'elle a été entraînée sur un corpus de textes calibrés spécifiquement par des formateurs experts du public des personnes en situation d'illettrisme. De plus, elle utilise une échelle spécifiquement conçue pour ces personnes. Notre formule présente aussi l'intérêt de s'appuyer sur les dernières avancées technologiques en lisibilité, à savoir les modèles hybrides. Une seconde contribution de cet article est le corpus d'entraînement en lui-même, constitué de 461 textes annotés sur une double échelle (ordinale, à 4 niveaux et discrète, à 20 valeurs), qui sera rendu disponible [1].

Dans la suite de cet article, nous développerons les travaux en lisibilité, en mettant l'accent sur les recherches en français et celles portant sur un public de personnes en situation d'illettrisme à la section 2. Ensuite, nous présenterons le jeu de données utilisé pour notre étude à la section 3, avant de présenter les différents modèles que nous avons explorés pour notre formule de lisibilité à la section 4. Enfin, nous évaluerons ces différents modèles à la section 5 avant de conclure.

1. https://github.com/tfrancoiscental/iread4skills_readability_corpus_fr

2 État de l'art

Parmi les toutes premières études en lisibilité, nous comptons des formules de lisibilité visant explicitement les jeunes adultes faibles lecteurs ou en situation d'illettrisme (Dale & Tyler, 1934; Gray & Leary, 1935). Toutefois, par la suite, le focus se déplaça vers des formules visant les adultes en général (Flesch, 1948; Gunning, 1952) ou les écoliers (Dale & Chall, 1948). Aujourd'hui, diverses études exploitent les formules de lisibilité pour mesurer la difficulté de lecture de différents documents - par ex. les textes médicaux (Wilson, 2009; Mcinnes & Haglund, 2011) ou les contrats (Arbel, 2024) - pour les personnes en situation d'illettrisme, mais elles ne recourent qu'à des formules classiques comme celle de Flesch.

Si nous nous restreignons aux publications proposant de nouveaux modèles de lisibilité dédiés aux personnes en situation d'illettrisme, nous pouvons identifier des recherches menées sur le portugais (Aluisio *et al.*, 2010), l'italien (Dell'Orletta *et al.*, 2011) ou l'allemand (Weiss *et al.*, 2018). Cette dernière est particulièrement intéressante, car les auteurs y proposent une formule reposant sur l'échelle de difficulté « Alpha », spécifiquement dédiée aux personnes en situation d'illettrisme (Riekmann & Grotlüschen, 2011). Par ailleurs, cette formule a ensuite été intégrée dans un moteur de recherche spécialisé pour les personnes en situation d'illettrisme (Dittrich *et al.*, 2019) ciblant les niveaux « Alpha » 3 à 6. Néanmoins, généralement, les études sur ce sujet sont rares, comme le confirme l'article de synthèse de Collins-Thompson (2014), qui liste les différents publics ciblés en lisibilité, mais ne mentionne pas les personnes en situation d'illettrisme. De même, il n'en existe aucune pour le français, à notre connaissance.

Si nous regardons les jeux de données qui ont été rendus disponibles en lisibilité pour soutenir le développement de nouvelles formules, le diagnostic se confirme, puisqu'aucun ne cible les personnes en situation d'illettrisme. Les corpus Weakly Reader (Schwarm & Ostendorf, 2005), WeeBit (Vajjala & Meurers, 2012), Newsela ou encore CLEAR (Crossley *et al.*, 2023) ciblent les écoliers, tandis que OneStopEnglish (Vajjala & Lučić, 2018) est dédié aux apprenants de l'anglais. Pour le français, on ne compte que les trois jeux de données proposés par (Hernandez *et al.*, 2022), qui ciblent à nouveau le contexte scolaire. On voit donc bien le manque de ressources pour le contexte de l'illettrisme.

Beaucoup de travaux récents, y compris en français, se focalisent plutôt sur la dimension algorithmique des modèles de lisibilité. Avant 2017, l'approche dominante reposait sur des modèles d'apprentissage automatique qui cherchaient à identifier les caractéristiques textuelles les plus prédictives de la difficulté textuelle et à les combiner au mieux (Schwarm & Ostendorf, 2005; Feng *et al.*, 2010; Vajjala & Meurers, 2012). Pour le français, citons parmi cette veine les travaux de François & Fairon (2012) et de Dascalu (2014). Ensuite, les recherches en lisibilité ont également été dynamisées par l'émergence des représentations distribuées (Cha *et al.*, 2017; Filighera *et al.*, 2019) et de l'apprentissage profond (Nadeem & Ostendorf, 2018; Azpiazu & Pera, 2019; Martinc *et al.*, 2021). En français, Blandin *et al.* (2020) exploitent des modèles profonds à propagation avant pour effectuer des recommandations d'âge de lecture pour les enfants ; Yancey *et al.* (2021) a proposé une formule de lisibilité basée sur BERT pour le français langue étrangère et Van Ngo & Parmentier (2023) ont exploré la relation entre la difficulté d'un texte et des phrases qui le constituent.

Plus récemment, nous distinguons deux tendances dans les recherches en lisibilité. D'une part, celles qui visent à intégrer des caractéristiques textuelles typiques du début du siècle au sein d'architectures d'apprentissage profond, aussi appelées approches hybrides (Qin *et al.*, 2020; Deutsch *et al.*, 2020; Liu & Lee, 2023), dont Wilkens *et al.* (2024) est représentatif pour le français. D'autre part, les

grands modèles de langue génératifs permettent désormais d'effectuer des évaluations de lisibilité sans disposer d'un modèle pré-entraîné, comme le suggèrent les résultats encourageants de Jamet *et al.* (2024).

3 Jeu de données

3.1 Collecte et annotation du corpus

La première étape dans la conception d'une formule de lisibilité est de collecter un corpus de textes dont la difficulté a été étalonnée pour le public cible. Pour ce faire, nous avons collecté 461 textes représentatifs de 11 types de communication différents [2] collectés principalement auprès de formateurs actifs avec des personnes en situation d'illettrisme, mais aussi sur le web. Ces textes sont relativement courts (de 25 à 608 tokens), car adaptés à un public d'illettrés. La distribution de la longueur des textes est reprise à l'Annexe A. Une fois collectés, les textes ont subi une procédure d'annotation de leur difficulté selon une échelle spécifiquement développée pour les personnes en situation d'illettrisme et qui comporte quatre niveaux : « Très Facile », « Facile », « Accessible » et « +Complexe ». Ces niveaux ont été définis comme suit et validés par un panel de formateurs experts :

— **Très Facile** : Textes entièrement ou presque entièrement compris par l'ensemble des lecteurs, y compris ceux ayant un très faible niveau de scolarisation (environ jusqu'à la sixième année d'éducation) et une expérience de lecture quasi inexistante. Ils sont très courts et traitent de sujets simples, avec un vocabulaire basique.
— **Facile** : Textes entièrement ou presque entièrement compris par des personnes ayant un faible niveau de scolarisation (c'est-à-dire ayant terminé l'école primaire, mais ne dépassant pas la neuvième année d'éducation) et une expérience de lecture limitée. Il s'agit de textes courts, pouvant notamment présenter des concepts abstraits et des figures de style communes.
— **Accessible** : Textes compréhensibles dès la première lecture par des individus ayant achevé la neuvième année d'éducation et possédant une expérience de lecture fonctionnelle à moyenne. Il s'agit de textes plus longs, pouvant présenter des concepts plus variés, des structures syntaxiques plus complexes et des verbes irréguliers s'ils sont très fréquents dans la langue.
— **+Complexe** : Textes nécessitant une lecture plus attentive et une certaine maîtrise linguistique pour être pleinement compris. Ils s'adressent à des lecteurs ayant un niveau de scolarisation plus avancé et une expérience de lecture plus approfondie. Ce niveau regroupe tous les éléments plus complexes que ceux décrits pour les catégories précédentes.

L'annotation a été réalisée par 15 professionnelles en lien avec l'illettrisme et notamment des formatrices [3]. La procédure d'annotation a été la suivante : un guide d'annotation reprenant un descriptif de ces 4 niveaux accompagnés d'exemples a d'abord été préparé et revu jusqu'à ce que chaque annotatrice en approuve le contenu. Ce guide finalisé, nous avons lancé une première phase d'annotation en utilisant le logiciel Qualtrics [4]. Des sessions de formation ont été proposées aux annotatrices afin de s'assurer de leur prise en main du site. Les retours à l'issue de cette première phase ont mené à des modifications de notre protocole d'annotation dont la plus marquante est l'ajout

2. Communication personnelle, professionnelle, commerciale, académique, politique, légal, religieuse, sur les réseaux sociaux, ainsi que des livres de fiction, des livres non fictionnels et des livres didactiques.

3. Tous nos annotateurs étaient des femmes, bien que le genre n'ait pas été un critère de sélection. Elles ont fait l'objet d'une rémunération.

4. https://www.qualtrics.com

d'une échelle de Likert à 5 degrés à chaque niveau. Celle-ci permet d'obtenir une analyse plus fine et de rassurer les annotatrices en cas d'hésitation entre deux catégories. Cette échelle de Likert à cinq points visait à préciser où le texte évalué se situe au sein du niveau. De cette façon, chaque texte se voit caractérisé par l'une des 4 classes ci-dessus ainsi que par un score de 1 à 20, obtenu en additionnant la valeur de l'échelle de Likert d (de 1 à 5) à la classe G convertie de la façon suivante : Très Facile (0), Facile (5), Accessible (10) et +Complexe (15). Chaque texte a été évalué par au moins trois annotatrices. À partir de ces annotations multiples, nous avons établi des valeurs de référence pour les deux échelles de scores (le score de 1 à 20 et sa conversion en nos quatre niveaux de difficulté) en prenant simplement la moyenne des trois annotations [5].

3.2 Analyses des données

La table 1 présente le nombre de textes par niveau de difficulté à l'issue de la campagne d'annotation. Le jeu de données présente malheureusement des catégories relativement déséquilibrées, puisqu'environ 85 % des textes du corpus ont été attribués aux catégories « Facile » et « Accessible ».

Très Facile	Facile	Accessible	+Complexe
19	212	198	32

TABLE 1 – Répartition des textes par catégories de difficulté

La figure 1 illustre, quant à elle, la distribution des scores sur 20 attribués à chaque texte par les annotatrices, au sein de chaque catégorie de difficulté. Si chaque catégorie présente le même nombre de points, les répartitions en leur sein peuvent varier. Ainsi, les textes « Très Facile » ont généralement les notes les plus élevées de leur catégorie, le troisième et quatrième quartile se confondant. Les autres catégories semblent présenter des scores plus homogènes.

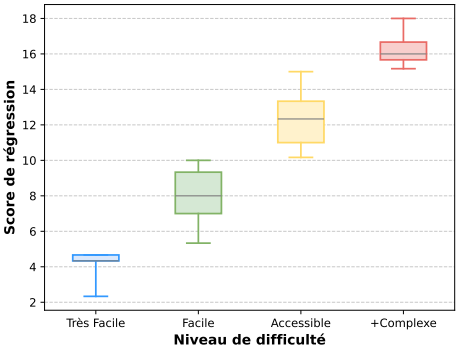

FIGURE 1 – Répartition des scores de difficulté attribué aux textes par catégorie de difficulté

Les scores d'accord inter-annotateurs ont ensuite été calculés. Signalons que comme chaque annotatrice était libre de choisir le nombre d'annotations qu'elle souhaitait effectuer, certaines n'ont annoté

5. Dans le jeu de données, nous proposons également une seconde manière de calculer ces valeurs de référence en considérant l'étiquette la plus représentée, la moyenne étant utilisée pour gérer les cas où toutes les étiquettes étaient différentes

	Kappa	Spearman
Annot 1 vs. réf.	0,28 ± 0,12	0,63 ± 0,13
Annot 2 vs. réf.	0,38 ± 0,15	0,69 ± 0,13
Annot 3 vs. réf.	0,54 ± 0,11	0,69 ± 0,15

TABLE 2 – Accord inter-annotatrices selon le Kappa Quadratique Pondéré et corrélation de Spearman pour les scores sur 20.

qu'un seul batch, tandis que d'autres en ont annoté jusqu'à vingt-huit. Afin de pouvoir calculer des scores d'accord entre annotatrices sur une base comparable, nous avons regroupé les annotations individuelles en trois ensembles que nous appelons super-annotatrices. Les trois super-annotatrices ainsi constituées rassemblent respectivement les annotations de 1, 5 et 10 annotatrices. Ces trois super-annotatrices comptent respectivement une, cinq et dix annotatrices.

L'accord inter-annotateur a été calculé à la fois pour les super-annotatrices par série de textes afin d'en extraire une moyenne et un écart-type. La table 2 présente les scores d'accord inter-annotatrices, que l'on calcule comme le Kappa quadratique pondéré (KQP) ou la corrélation de Spearman entre chaque super-annotatrice et la valeur moyenne. Ainsi, pour les scores sur 20 points, nous avons calculé le KQP ainsi que la corrélation de Spearman sur nos séries. Pour le KQP, les scores sont situés entre 0,28 et 0,54 avec une moyenne par annotatrice et par série à 0,33. Ces résultats sont certes faibles, mais assez proches du KQP obtenu sur une tâche proche à SemEval 2012 (Specia *et al.*, 2012). Pour le second, les scores se situent entre 0,63 et 0,69. La p-value a été calculée et n'excède pas les 0,02.

4 Méthodologie de conception de la formule

Dans cette étude, nous avons exploré diverses approches pour l'évaluation de la lisibilité des textes : des modèles d'apprentissage automatique classiques, des techniques d'apprentissage profond – un affinage de CamemBERT (Martin *et al.*, 2020), mais aussi un modèle hybride, comme illustré dans la Figure 2 –, ainsi que des modèles de langage génératifs (via des techniques de prompting).

Signalons également que comme nous disposons de deux échelles de difficulté (1-4 et 1-20), nous avons exploré la tâche d'évaluation automatique de la lisibilité à la fois comme un problème de classification et de régression. Si la classification permet d'assigner des niveaux plus fonctionnels sur le terrain aux textes, la régression favorise une évaluation plus nuancée et plus granulaire, tout en atténuant les limitations liées à la rigidité des frontières de classe.

Les performances des modèles ont été estimées au moyen d'une validation croisée à 5 échantillons stratifiés. À chaque itération, 60 % des données étaient utilisées pour l'entraînement, 20% pour la recherche des hyperparamètres, et les 20 % restants pour le test.

4.1 Modèles d'apprentissage automatique

La première étape de l'utilisation des algorithmes classiques d'apprentissage automatique pour la classification de la difficulté des textes consiste à transformer les textes en variables numériques. Pour

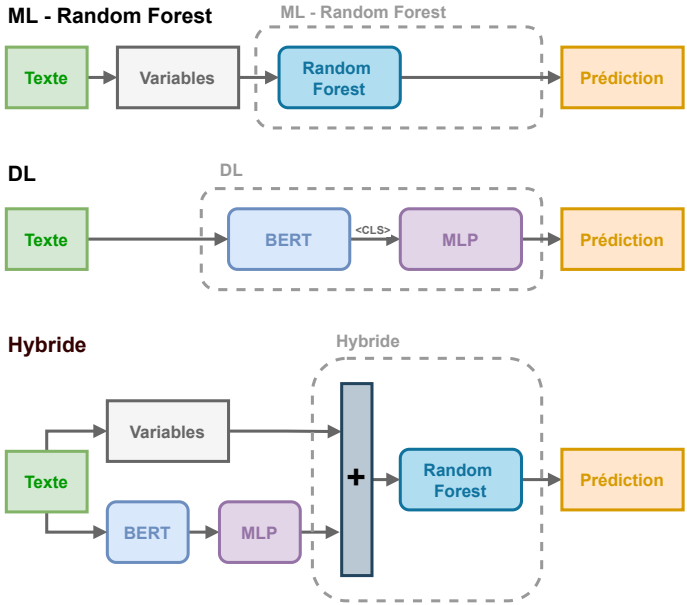

FIGURE 2 – Architecture des systèmes de prédiction de lisibilité.

ce faire, nous avons utilisé l'outil FABRA (Wilkens *et al.*, 2022) [6]. Cet outil propose de nombreux descripteurs linguistiques pertinents pour l'évaluation de la lisibilité à la fois au niveau du document, de la phrase et du mot. Par ailleurs, FABRA propose, pour chacun de ces descripteurs, jusqu'à 20 aggrégateurs statistiques différents (ex. moyenne, médiane, mode, 80P, etc.).

Dès lors, nous disposons d'un nombre très conséquent de descripteurs. Afin d'optimiser les performances du modèle tout en réduisant sa complexité, nous avons procédé à une sélection automatique de variables en appliquant l'algorithme **minimum Redundancy Maximum Relevance** (**mRmR**) (Ding & Peng, 2003) qui permet d'identifier les sous-ensembles de caractéristiques les plus significatifs. Les nombres de descripteurs testés vont de 10 à 500.

Au niveau des algorithmes, nous avons utilisé la librairie Scikit-Learn (Kramer & Kramer, 2016) pour entraîner des machines à vecteurs de support (**SVM**), des arbres de décision (**DT**) et des forêts d'arbres décisionnels (**RF**) pour la tâche de classification, ainsi que leurs équivalents pour la tâche de régression. Nous avons utilisé une recherche par grille (*grid search*) pour explorer l'ensemble des configurations d'hyperparamètres pour nos différents modèles. Les hyperparamètres pour chacun de nos modèles sont décrits à l'annexe B. Pour la classification, la fonction de coût pondère les erreurs en tenant compte de la distribution des classes, afin de compenser le déséquilibre lié à la sous-représentation de certaines classes par rapport à d'autres.

6. La liste des variables est disponible à l'adresse `https://cental.uclouvain.be/fabra`.

4.2 Modèles d'apprentissage profond

Suivant la dominance de l'apprentissage profond en TAL, nous avons aussi logiquement exploré une stratégie basée sur l'affinage de modèles transformeurs pré-entraînés de type BERT, en utilisant **CamemBERT** (Martin *et al.*, 2020) ainsi qu'une variante améliorée, **CamemBERT-v2** (Antoun *et al.*, 2024). Afin d'adapter le modèle pré-entraîné à notre tâche spécifique, nous avons ajouté une couche entièrement connectée en sortie, qui a été entraînée sur notre jeu d'entraînement. Ce choix nous permet d'exploiter les représentations riches extraites par les couches internes du transformeur tout en spécialisant le modèle pour la prédiction de la lisibilité des textes.

L'optimiseur Adam a été employé, et l'ensemble des paramètres du modèle a été affiné sans congeler les couches intermédiaires du modèle de langue. Afin de prévenir le sur-apprentissage et de réduire la durée d'entraînement, nous avons appliqué l'arrêt précoce avec une patience de 10 époques. Une recherche par grille exhaustive des meilleurs hyperparamètres a été réalisée, dont les détails sont présentés à l'annexe B.

Signalons que, pour la classification, nous avons utilisé une entropie croisée pondérée en raison du déséquilibre des classes de notre jeu de données. Cette pondération, basée sur les distributions inverses de chaque classe, permet au modèle de mieux prendre en compte les classes moins fréquentes. Pour la régression, nous avons utilisé l'erreur quadratique moyenne pour optimiser les modèles.

4.3 Modèles hybrides

Pour les modèles hybrides, nous nous inspirons de l'architecture Soft-Label (SO) proposée par (Wilkens *et al.*, 2024), qui a montré des résultats particulièrement performants lors d'expériences sur quatre jeux de données différents. Ce modèle Soft-Label (SO) concatène des variables linguistiques avec la sortie softmax d'un modèle d'apprentissage profond, les utilisant ensuite comme entrée dans un modèle d'apprentissage automatique pour prédire la difficulté d'un texte.

En s'appuyant sur les résultats obtenus lors des expérimentations des modèles d'apprentissage automatique et des modèles d'apprentissage profond rapportés dans le Tableau 3, nous avons utilisé les variables linguistiques sélectionnées par l'algorithme mRMR concaténées avec la sortie du modèle **CamemBERT-v2** affiné. Ces résultats sont ensuite utilisés comme entrée à un modèle de forêt d'arbres décisionnels (**RF**) pour prédire la difficulté du texte. La stratégie d'entraînement et d'évaluation suit l'approche décrite dans la Section 4.1.

4.4 Modèles de langue génératifs

Finalement, nous avons mené des expérimentations visant à évaluer l'efficacité des modèles de langue génératifs en situation d'apprentissage zéro coup (zero-shot) et d'apprentissage en quelques coups (few-shot) pour la tâche de lisibilité. Nous avons utilisé le modèle **DeepSeek-R1 70B**, capable de générer un raisonnement intermédiaire avant de produire sa prédiction ainsi que les modèles **Mistral-large 123B** et **GPT-4.1**. Pour formuler nos requêtes, nous avons réutilisé le prompt proposé dans une étude précédente (Jamet *et al.*, 2024), lui demandant de générer des scores CECR, avec chaîne de pensée, à partir desquels nous avons établi une correspondance avec nos annotations (*Très Facile* : *A1*, *Facile* : *A2*, *Accessible* : *B1*, *+Complexe* : *B2* et plus). Les prompts utilisés sont présentés dans les annexes E et F.

Dans l'expérience en quelques coups, nous avons sélectionné un exemple représentatif pour chaque niveau de difficulté (*c.-à-d.* 4 exemples, voir en annexe 6), en nous assurant qu'ils fassent consensus parmi nos annotatrices. Afin de mesurer la variabilité des performances, nous avons divisé le jeu de données en cinq plis et calculé, sur chacun d'eux, l'exactitude, l'exactitude adjacente [7] et le macro-F1 moyen, accompagnés de leurs écarts-types.

5 Évaluation des systèmes de lisibilité

5.1 Méthodologie d'évaluation

Les performances des différents modèles entraînés sur notre jeu de données ont été évaluées selon plusieurs métriques. Pour les modèles de classification, nous avons utilisé l'exactitude, l'exactitude adjacente, le macro-F1, et l'erreur quadratique moyenne (MSE) pour les systèmes entraînés pour la régression. Les résultats de ces évaluations sont présentés dans le tableau 3.

Pour comparer équitablement les performances des modèles de classification et de régression, nous avons converti les scores continus prédits par les modèles de régression en classes de lisibilité correspondant à celles utilisées pour la classification. Cette conversion repose sur les seuils suivants : **Très Facile** si score ≤ 5 ; **Facile** si $5 <$ score ≤ 10 ; **Accessible** si $10 <$ score ≤ 15 ; **+Complexe** si score > 15. Cela permet d'évaluer les modèles de régression à l'aide des mêmes métriques que pour la classification, et ainsi de comparer les deux approches sur une base commune.

5.2 Performances des systèmes

Au niveau des systèmes d'apprentissage automatique classiques, nous observons une tendance d'une légère supériorité du modèle RF. Toutefois, aucune différence importante n'a été observée au niveau du macro-F1 par rapport aux autres modèles, ce qui suggère que son amélioration en exactitude ne se traduit pas nécessairement par une meilleure prise en compte des classes sous-représentées.

En effet, les modèles d'apprentissage automatique présentent une importante différence entre l'exactitude et le macro-F1. Cette disparité reflète une tendance des modèles à prédire les classes majoritaires au détriment des classes moins fréquentes. Une pondération des erreurs sur ces classes a permis de réduire cet écart, qui subsiste néanmoins en comparaison avec les autres architectures.

Que cela soit pour la classification ou la régression, nous observons que les modèles d'apprentissage profond obtiennent des performances meilleures que les modèles ML. En revanche, bien que nos résultats suggèrent une dégradation du macro-F1, nous n'observons pas de différences importantes avec les systèmes hybrides sauf pour l'erreur quadratique moyenne qui semble indiquer que les modèles BERT obtiennent de meilleures performances.

Le tableau 4 présente les performances de classification des LLMs génératifs en fonction de la langue de la requête et de l'apprentissage à zéro ou quelques coups. Les résultats indiquent que l'apprentissage en quelques coups, qui repose sur des exemples, génère des performances globalement supérieures à celles obtenues avec l'apprentissage en zéro coup. Il apporte aussi plus de stabilités

7. L'exactitude adjacente considère une prédiction comme correcte si elle correspond à la classe réelle ou à une classe voisine sur l'échelle ordonnée. (ex. : prédire "Très facile" au lieu de "Facile")

	Exactitude	Exactitude Adj.	Macro-F1	EQM
Classification				
ML - SVM (500)	55,84 ± 4,26	97,83 ± 1,95	47,54 ± 6,15	
ML - DT (300)	54,10 ± 3,28	94,81 ± 0,81	43,84 ± 4,98	
ML - RF (400)	62,77 ± 4,18	98,05 ± 1,26	47,78 ± 7,60	
DL - CamemBERT	64,04 ± 9,97	98,71 ± 1,77	**60,36 ± 8,23**	
DL - CamemBERT-v2	64,26 ± 5,67	99,17 ± 0,91	60,05 ± 6,01	
Hybride - RF (300)	**67,32 ± 4,08**	**99,14 ± 0,81**	56,26 ± 9,17	
Régression				
ML - SVR (500)	39,60 ± 5,39	93,06 ± 3,61	22,63 ± 1,96	4,94 ± 1,07
ML - DT (50)	38,75 ± 5,29	88,10 ± 2,24	22,13 ± 3,15	7,22 ± 1,55
ML - RF (500)	40,89 ± 6,28	91,77 ± 3,10	22,96 ± 4,71	4,70 ± 0,73
DL - CamemBERT	**70,77 ± 5,48**	**100,00 ± 0,00**	**59,63 ± 2,55**	3,87 ± 0,72
DL - CamemBERT-v2	68,38 ± 5,70	**100,00 ± 0,00**	47,52 ± 8,36	**3,78 ± 0,75**
Hybride - RF (300)	64,28 ± 6,55	99,57 ± 0,53	36,50 ± 5,21	4,88 ± 0,75

TABLE 3 – Comparaison des performances des métriques (précision, exactitude adjacente, macro-F1 et EQM le cas échéant) pour les modèles de classification et de régression de la lisibilité.

	Exactitude	Exactitude Adj.	Macro-F1
Mistral-large-FR-zero-shot	31.46 ± 5.48	91.33 ± 2.83	26.72 ± 4.24
Mistral-large-FR-few-shot	56.61 ± 3.03	**98.48 ± 0.87**	43.86 ± 3.57
Mistral-large-EN-zero-shot	31.47 ± 8.33	90.02 ± 3.85	22.78 ± 5.34
Mistral-large-EN-few-shot	**58.13 ± 3.36**	**98.48 ± 1.11**	48.60 ± 4.08
GPT-4.1-FR-zero-shot	30.15 ± 6.98	84.60 ± 4.00	35.55 ± 9.34
GPT-4.1-FR-few-shot	46.64 ± 5.08	96.10 ± 1.62	43.02 ± 7.82
GPT-4.1-EN-zero-shot	30.18 ± 11.37	91.76 ± 3.78	27.01 ± 6.04
GPT-4.1-EN-few-shot	48.38 ± 5.61	96.75 ± 1.36	44.69 ± 6.84
DeepSeek-FR-zero-shot	41,75 ± 5,14	91,77 ± 2,14	38,41 ± 3,14
DeepSeek-FR-few-shot	53,89 ± 1,14	96,53 ± 1,87	47,06 ± 3,55
DeepSeek-EN-zero-shot	33,31 ± 8,95	87,87 ± 4,04	29,64 ± 8,22
DeepSeek-EN-few-shot	52,80 ± 3,25	94,81 ± 1,26	**48,95 ± 0,49**

TABLE 4 – Comparaison des performances des métriques (exactitude, exactitude adjacente, macro-F1) des LLMs génératifs pour la classification des niveaux de difficulté.

dans les performances des LLMs. Nous avons évalué deux langues d'instruction dans les prompts donnés aux LLM, le français (FR) et l'anglais (EN), afin d'analyser l'impact de la langue d'interaction sur les performances du modèle. Selon le Macro-F1, l'anglais serait plus adapté en quelques coups, alors que le français serait plus adapté en zéro coups pour l'ensemble des modèles testés. Concernant les modèles utilisés, DeepSeek affiche de meilleures performances en termes de Macro-F1 sur l'ensemble des stratégies d'apprentissage contextuel. Toutefois, cet avantage ne se retrouve pas lorsqu'on considère les mesures d'exactitude.

Par ailleurs, bien que les modèles de langue génératifs affichent des performances inférieures aux modèles DL, ils parviennent néanmoins à des résultats comparables en termes de macro-F1 sans avoir été spécifiquement entraînés pour cette tâche. Cette faculté à généraliser sans apprentissage supervisé direct souligne leur capacité, bien qu'ils ne surpassent pas encore les modèles spécialisés.

6 Conclusion et perspectives

Cette étude a permis d'évaluer différents systèmes de classification de la difficulté des textes en français pour les personnes en situation d'illettrisme, en mettant en lumière les performances des approches basées sur l'apprentissage automatique, l'apprentissage profond, les méthodes hybrides, et les modèles de langue génératifs. Les résultats ont montré une tendance des modèles hybrides et d'apprentissage profond à obtenir de meilleures performances en matière d'exactitude et de robustesse. Malgré leurs performances dans de nombreuses tâches, les LLMs génératifs semblent obtenir de moins bonnes performances, confirmant les conclusions de précédentes études (Jamet *et al.*, 2024). Il serait intéressant toutefois de poursuivre une analyse en affinant un LLM génératifs sur une tâche de lisibilité.

La subjectivité des annotatrices peut entraîner des désaccords inter-annotatrices, mais les résultats des modèles démontrent que la moyenne des annotations permet d'obtenir des données plus cohérentes et exploitables pour la classification de la difficulté des textes. Une étude comparative plus approfondie entre les annotatrices, en particulier pour normaliser les annotations des cas où des divergences importantes existent, pourrait enrichir cette analyse. L'exploration des différences entre annotatrices pourrait aussi offrir des pistes pour mieux comprendre les critères de jugement utilisés dans l'évaluation de la lisibilité pour les personnes en situation d'illettrisme.

Remerciements

Cette recherche est soutenue par la Commission européenne (Projet : iRead4Skills, Numéro de subvention : 1010094837. [Sujet : HORIZONCL2- 2022-TRANSFORMATIONS-01-07, DOI :10.3030/101094837])) et le projet PDR n°T.0080.23, intitulé CMesure, soutenu par le FNRS (Fonds national de la Recherche Scientifique). Les points de vue et opinions exprimés sont ceux des auteurs et ne reflètent pas nécessairement ceux de l'Union européenne ou de l'Agence exécutive de la recherche européenne. Ni l'Union européenne ni l'autorité chargée de l'octroi des subventions ni le FNRS ne peuvent en être tenues pour responsables.

Nous remercions l'ANLCI, *Le Français pour adultes*, les centres *Savoirs pour réussir Paris* et *Poinfor* ainsi que toutes les annotatrices pour leur précieuse contribution au projet.

Références

ALUISIO S., SPECIA L., GASPERIN C. & SCARTON C. (2010). Readability assessment for text simplification. In *Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, p. 1–9, Los Angeles.

ANTOUN W., KULUMBA F., TOUCHENT R., DE LA CLERGERIE É., SAGOT B. & SEDDAH D. (2024). Camembert 2.0 : A smarter french language model aged to perfection. *arXiv preprint arXiv :2411.08868*.

ARBEL Y. A. (2024). The readability of contracts : Big data analysis. *Journal of Empirical Legal Studies*, **21**(4), 927–978.

ARTSTEIN R. & POESIO M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, **34**(4), 555–596. DOI : 10.1162/coli.07-034-R2.

AZPIAZU I. M. & PERA M. S. (2019). Multiattentive recurrent neural network architecture for multilingual readability assessment. *Transactions of the Association for Computational Linguistics*, **7**, 421–436.

BERKMAN N. D., SHERIDAN S. L., DONAHUE K. E., HALPERN D. J. & CROTTY K. (2011). Low health literacy and health outcomes : an updated systematic review. *Annals of internal medicine*, **155**(2), 97–107.

BESSE J.-M., LUIS M.-H., BOUCHUT A.-L. & MARTINEZ F. (2009). La mesure des compétences en traitement de l'écrit chez des adultes en grande difficulté. *Économie et statistique*, **424**(1), 31–48.

BLANDIN A., LECORVÉ G., BATTISTELLI D. & ETIENNE A. (2020). Recommandation d'âge pour des textes (age recommendation for texts). In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 2 : Traitement Automatique des Langues Naturelles*, p. 164–171.

CHA M., GWON Y. & KUNG H. (2017). Language modeling by clustering with word embeddings for text readability assessment. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, p. 2003–2006 : ACM.

COLLINS-THOMPSON K. (2014). Computational assessment of text readability : A survey of current and future research. *International Journal of Applied Linguistics*, **165**(2), 97–135.

CROSSLEY S., HEINTZ A., CHOI J. S., BATCHELOR J., KARIMI M. & MALATINSZKY A. (2023). A large-scaled corpus for assessing text readability. *Behavior Research Methods*, **55**(2), 491–507.

DALE E. & CHALL J. (1948). A formula for predicting readability. *Educational research bulletin*, **27**(1), 11–28.

DALE E. & TYLER R. (1934). A study of the factors influencing the difficulty of reading materials for adults of limited reading ability. *The Library Quarterly*, **4**, 384–412.

DASCALU M. (2014). Readerbench (2)-individual assessment through reading strategies and textual complexity. In *Analyzing Discourse and Text Complexity for Learning and Collaborating*, p. 161–188. Springer.

DELL'ORLETTA F., MONTEMAGNI S. & VENTURI G. (2011). Read-it : Assessing readability of italian texts with a view to text simplification. In *Proceedings of the second workshop on speech and language processing for assistive technologies*, p. 73–83.

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

DING C. & PENG H. (2003). Minimum redundancy feature selection from microarray gene expression data. In *Computational Systems Bioinformatics. CSB2003. Proceedings of the 2003 IEEE Bioinformatics Conference. CSB2003*, p. 523–528. DOI : 10.1109/CSB.2003.1227396.

DITTRICH S., WEISS Z., SCHRÖTER H. & MEURERS D. (2019). Integrating large-scale web data and curated corpus data in a search engine supporting german literacy education. In *Proceedings of the 8th workshop on NLP for computer assisted language learning*, p. 41–56.

DUBAY W. H. (2004). The principles of readability. *Online submission*.

FENG L., JANSCHE M., HUENERFAUTH M. & ELHADAD N. (2010). A Comparison of Features for Automatic Readability Assessment. In *COLING 2010 : Poster Volume*, p. 276–284.

FILIGHERA A., STEUER T. & RENSING C. (2019). Automatic text difficulty estimation using embeddings and neural networks. In *European Conference on Technology Enhanced Learning*, p. 335–348 : Springer.

FLESCH R. (1948). A new readability yardstick. *Journal of Applied Psychology*, **32**(3), 221–233.

FRANÇOIS T. (2011). *Les apports du traitement automatique du langage à la lisibilité du français langue étrangère*. Thèse de doctorat, Université Catholique de Louvain. Thesis Supervisors : Cédrick Fairon and Anne Catherine Simon.

FRANÇOIS T. & FAIRON C. (2012). An "AI readability" formula for French as a foreign language. In J. TSUJII, J. HENDERSON & M. PAŞCA, Éds., *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, p. 466–477, Jeju Island, Korea : Association for Computational Linguistics.

GRAY W. & LEARY B. (1935). *What makes a book readable*. Chicago : Illinois : University of Chicago Press.

GUNNING R. (1952). *The technique of clear writing*. New York : McGraw-Hill.

HERNANDEZ N., OULBAZ N. & FAINE T. (2022). Open corpora and toolkit for assessing text readability in french. In *Proceedings of the 2nd Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI) within the 13th Language Resources and Evaluation Conference*, p. 54–61.

JAMET H., MANDERLIER M., SHRESTHA Y. R. & VLACHOS M. (2024). Evaluation and simplification of text difficulty using llms in the context of recommending texts in french to facilitate language learning. In *Proceedings of the 18th ACM Conference on Recommender Systems*, p. 987–992.

KRAMER O. & KRAMER O. (2016). Scikit-learn. *Machine learning for evolution strategies*, p. 45–53.

LAL B. S. (2015). The economic and social cost of illiteracy : an overview. *International Journal of Advance Research and Innovative Ideas in Education*, **1**(5), 663–670.

LIU F. & LEE J. S. (2023). Hybrid models for sentence readability assessment. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, p. 448–454.

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

MARTINC M., POLLAK S. & ROBNIK-ŠIKONJA M. (2021). Supervised and unsupervised neural approaches to text readability. *Computational Linguistics*, **47**(1), 141–179.

MCINNES N. & HAGLUND B. J. (2011). Readability of online health information : implications for health literacy. *Informatics for health and social care*, **36**(4), 173–189.

MESSIAS E. (2003). Income inequality, illiteracy rate, and life expectancy in brazil. *American Journal of Public Health*, **93**(8), 1294–1296.

NADEEM F. & OSTENDORF M. (2018). Estimating linguistic complexity for science texts. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, p. 45–55.

QIN Q., HU W. & LIU B. (2020). Feature projection for improved text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 8161–8171.

RIEKMANN W. & GROTLÜSCHEN A. (2011). Konservative entscheidungen : Größenordnung des funktionalen analphabetismus in deutschland. *REPORT-Zeitschrift für Weiterbildungsforschung*, (3), 24–35.

RUSSELER J., MENKHAUS K., AULBERT-SIEPELMEYER A., GERTH I. & BOLTZMANN M. (2012). " alpha plus" : An innovative training program for reading and writing education of functionally illiterate adults. *Creative Education*, **3**(3), 357–361.

SCHWARM S. & OSTENDORF M. (2005). Reading level assessment using support vector machines and statistical language models. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, p. 523–530.

SHANNON C. E. (2001). A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review*, **5**(1), 3–55.

SPECIA L., JAUHAR S. K. & MIHALCEA R. (2012). Semeval-2012 task 1 : English lexical simplification. In ** SEM 2012 : The First Joint Conference on Lexical and Computational Semantics–Volume 1 : Proceedings of the main conference and the shared task, and Volume 2 : Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, p. 347–355.

STEWARD J. (2023). The economic & social cost of illiteracy : A snapshot of illiteracy in a global context.

TACK A., FRANÇOIS T., ROEKHAUT S. & FAIRON C. (2017). Human and automated CEFR-based grading of short answers. In J. TETREAULT, J. BURSTEIN, C. LEACOCK & H. YANNAKOUDAKIS, Éds., *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, p. 169–179, Copenhagen, Denmark : Association for Computational Linguistics. DOI : 10.18653/v1/W17-5018.

VAJJALA S. & LUČIĆ I. (2018). Onestopenglish corpus : A new corpus for automatic readability assessment and text simplification. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, p. 297–304.

VAJJALA S. & MEURERS D. (2012). On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the seventh workshop on building educational applications using NLP*, p. 163–173.

VAN NGO D. & PARMENTIER Y. (2023). Towards sentence-level text readability assessment for french. In *Second Workshop on Text Simplification, Accessibility and Readability (TSAR@ RANLP2023)*.

WEISS Z., DITTRICH S. & MEURERS D. (2018). A linguistically-informed search engine to identifiy reading material for functional illiteracy classes. In *Proceedings of the 7th workshop on NLP for Computer Assisted Language Learning*, p. 79–90.

WILKENS R., ALFTER D., WANG X., PINTARD A., TACK A., YANCEY K. P. & FRANÇOIS T. (2022). Fabra : French aggregator-based readability assessment toolkit. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 1217–1233.

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.

WILSON M. (2009). Readability and patient education materials used for low-income populations. *Clinical Nurse Specialist*, **23**(1), 33–40.

YANCEY K., PINTARD A. & FRANÇOIS T. (2021). Investigating readability of french as a foreign language with deep learning and cognitive and pedagogical features. *Lingue e Linguaggio*, **2021**(2), 229–258. DOI : 10.1418/102814.

A Distribution de la longueur des textes du corpus

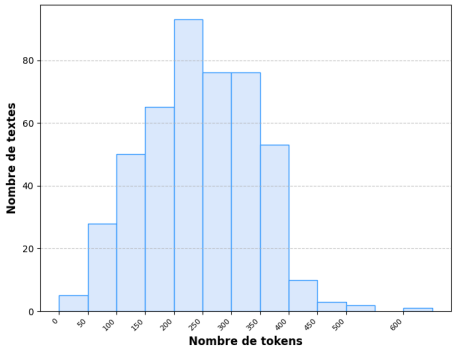

FIGURE 3 – Répartition de la taille des textes du corpus selon le nombre de tokens

B Liste des hyperparamètres des modèles

Pour les modèles d'apprentissage automatique, voici les hyperparamètres par tâche et par modèle :

Classification :

- **SVM** : Nous avons évalué trois types de noyau (`linear`, `rbf`, `sigmoid`). Pour le paramètre de régularisation `C`, nous avons testé les valeurs suivantes : 0,001, 0,005, 0,01, 0,05, 0,1, 0,5, 1, 5, 10, 50, 100.
- **DT :** Parmi les fonctions objectifs, nous avons testé `gini`, `entropy`, et `log_loss`. Pour la profondeur maximale de l'arbre, nous avons utilisé des valeurs de 1 à 8. Le nombre minimum d'échantillons requis pour diviser un nœud interne a été sélectionné parmi 2, 3, 5, 10, 15 et 20.
- **RF :** Le nombre d'arbres dans la forêt pouvait prendre des valeurs entre 10, 20, 30, 40, 50, 100, et 200. Pour la fonction objective, la profondeur maximale de l'arbre et le nombre minimum d'échantillons, nous avons exploré les mêmes valeurs que pour les arbres de décisions.

Régression :

- **SVR** : Nous avons évalué trois types de noyau (`linear`, `rbf`, `sigmoid`). Pour le paramètre de régularisation `C`, nous avons testé les valeurs suivantes : 0,001, 0,01, 0,1, 1, 10, 100.

— **DTR** : Nous avons testé différentes fonctions objectifs : `squared_error`, `friedman_mse`, `absolute_error` et `poisson`. La profondeur maximale de l'arbre a été fixée à 1, 5, 7 ou 8, et le nombre minimum d'échantillons requis pour diviser un nœud interne à 2, 5, 10 ou 20.
— **RFR** : Le nombre d'arbres dans la forêt (`n_estimators`) pouvait être de 10, 30, 50 ou 200. Les autres paramètres (`criterion`, `max_depth`, `min_samples_split`) ont été explorés avec les mêmes valeurs que pour les arbres de décision.

En ce qui concerne le modèle **CamemBERT** affiné, différentes valeurs ont été explorées pour les paramètres suivants : le taux d'apprentissage, testé avec des valeurs de 1e-5, 1e-4, 1e-3 ; la taille des batchs, avec des valeurs de 16, 32, 64 ; la décroissance du poids, explorée pour les valeurs 1e-5, 1e-4, 1e-3 ; et enfin, les valeurs de dropout 0.1, 0.3, 0.5 ont été testées pour atténuer le sur-apprentissage.

C Corrélation entre l'entropie des systèmes et le désaccord humain

Dans cette section, nous étudions la corrélation entre le désaccord des annotatrices et l'incertitude des modèles. L'entropie de Shannon (Shannon, 2001), calculée à partir des probabilités de sortie du modèle pour chaque texte, mesure l'incertitude des prédictions. Dans une précédente étude (Tack *et al.*, 2017), le désaccord entre les annotateurs pour un texte i se calcule en fonction du désaccord observé $D^{\alpha}_{o_i}$ concernant l'étiquette x donnée par l'annotateur c. Cette mesure est obtenue en décomposant la formule de Krippendorff pour le désaccord observé D^{α}_{o} (Artstein & Poesio, 2008), ce qui est égale à deux fois la variance empirique par texte s^2_i.

$$D^{\alpha}_{o_i} = \frac{1}{c(c-1)} \sum_{m=1}^{c} \sum_{n=1}^{c} \delta_{\text{interval}}(x_{ic_m}, x_{ic_n}) = 2s_i^2 \tag{1}$$

De cette manière, on peut calculer pour chaque texte un score de désaccord et un score d'entropie en fonction du système que l'on cherche à évaluer. On peut alors utiliser une corrélation de Pearson ou de Spearman pour étudier le lien entre l'incertitude humaine et des systèmes.

	Spearman	**Pearson**
ML - SVM (500)	-0,196*	-0,1777*
ML - DT (300)	0,0318	0,0292
ML - RF (400)	-0,1465*	-0,1489*
DL - CamemBERT-v2	-0,0461	-0,0314
Hybride - RF (300)	-0,1373*	-0,1298*

TABLE 5 – Résultats des Corrélations de Spearman et Pearson pour les modèles de classification. Les corrélations significatives sont indiqués par une astérisque (*).

Le tableau 5 présente les résultats des corrélations de Spearman et de Pearson pour différents modèles de classification. Les résultats indiquent que certains modèles ont une légère corrélation avec l'incertitude humaine. Par exemple, le modèle **SVM** présente une corrélation négative significative pour les deux tests (Spearman : -0.196), suggérant qu'une plus grande incertitude dans les prédictions

est associée à un plus grand désaccord entre les annotateurs. De la même manière, on observe ce comportement pour le modèle **RF**, et par extension le modèle hybride qui partage des traits architecturaux communs.

À l'inverse, les modèles **CamemBERT-v2** et **DT** ne présentent pas de corrélations significatives, ce qui indique que leur incertitude de prédiction ne semble pas avoir de lien direct avec les divergences d'opinions humaines.

D Textes pour l'apprentissage contextuelle des LLMs génératifs

Le coq est mort Le coq est mort, Le coq est mort, Le coq est mort. Il ne dira plus cocodi, cocoda Il ne dira plus cocodi, cocoda, cocodicodi, codicoda cocodicodi, codicoda	Très Facile
Inscription à la médiathèque BULLETIN D'INSCRIPTION Nom : ..... Prénom : ..... Date de Naissance :.....Sexe : F / M Adresse :..... Code postal :..... Ville : ..... Téléphone portable :..... Téléphone fixe :..... Email :..... L'email sera utilisé pour vous informer de la mise à disposition de vos réservations. J'autorise le réseau des médiathèq...	Facile
Quelqu'un de bien Debout devant ses illusions Une femme que plus rien ne dérange Détenue de son abandon Son ennui lui donne le change Que retient-elle de sa vie Qu'elle pourrait revoir en peinture Dans un joli cadre verni En évidence sur un mur Un mariage en Technicolor Un couple dans les tons pastel Assez d'argent sans trop d'efforts Pour deux trois folies mensuelles Elle a rêvé comme...	Accessible
Monsieur Charles Picqué, Bourgmestre de Saint-Gilles, Madame Martine Wille, Bourgmestre f. f. Madame Catherine François, Présidente, Monsieur Thierry Van Campenhout, Directeur, et l'équipe du Centre culturel Jacques Franck, LA TOPOGRAPHIE DU SIGNE Nicole Callebaut Peinture/Dessin L'œuvre picturale de Nicole Callebaut privilégie la suggestion. (...) Parfois dessins, photos et obj...	+Complexe

TABLE 6 – Exemples donnés en contexte des LLMs.

E Prompt zero-shot utilisé pour les LLMs génératifs

```
Vous êtes un expert linguistique spécialisé dans l'évaluation des niveaux de français selon le Cadre européen commun de référence pour les langues (CECR).
Votre tâche consiste à classer le texte français suivant dans l'un des niveaux du CECR : A1, A2, B1, B2, C1 ou C2.
Example :
Texte à classifier : "Bonjour, je m'appelle Jean. J'habite à Paris. J'aime jouer au football."
Le texte fourni est composé de phrases simples et courtes, utilisant des structures grammaticales de base et un vocabulaire élémentaire. Selon le Cadre européen commun de référence pour les langues (CECRL), le niveau A1 correspond à la capacité de comprendre et dútiliser des expressions familières et quotidiennes ainsi que des énoncés très simples visant à satisfaire des besoins concrets.
Niveau CECR : **A1**
Classifiez ce texte français : {text}
```

F Prompt few-shot utilisé pour les LLMs génératifs

```
Vous êtes un expert linguistique spécialisé dans l'évaluation des niveaux de français selon le Cadre européen commun de référence pour les langues (CECR).
Votre tâche consiste à classer le texte français suivant dans l'un des niveaux du CECR : A1, A2, B1, B2, C1 ou C2.
Example :
Texte à classifier : "Bonjour, je m'appelle Jean. J'habite à Paris. J'aime jouer au football."
Le texte fourni est composé de phrases simples et courtes, utilisant des structures grammaticales de base et un vocabulaire élémentaire. Selon le Cadre européen commun de référence pour les langues (CECR), le niveau A1 correspond à la capacité de comprendre et d'utiliser des expressions familières et quotidiennes ainsi que des énoncés très simples visant à satisfaire des besoins concrets.
Niveau CECR : **A1**
Classifiez ce texte français : {shot1}
{cot1} Niveau CECR : ** {classe2CECR[{value1}]} **
....
Classifiez ce texte français {shot4}
{cot4} Niveau CECR : ** {classe2CECR[{value4}]} **
Classifiez ce texte français : {text}
```