Evaluation de la lisibilité des textes biomédicaux selon le profil du lecteur

Anya NAIT DJOUDI[1]
(1) Aix-Marseille Université, CNRS, LIS, Marseille, France
anya-amel.NAIT-DJOUDI@univ-amu.fr

RÉSUMÉ

La lisibilité des textes biomédicaux est perçue différemment selon le profil du lecteur, ce qui est amplifié par la complexité intrinsèque de ces documents et par l'inégale littératie en santé au sein de la population. Bien que 72% des internautes consultent des informations médicales en ligne, une part significative rencontre des difficultés de compréhension. Pour garantir l'accessibilité des textes à un public varié, l'évaluation de la lisibilité est donc essentielle. Or, les formules de lisibilité classiques, conçues pour des textes généraux, ne tiennent pas compte de cette diversité, soulignant la nécessité d'adapter les outils d'évaluation aux besoins spécifiques des textes biomédicaux et à l'hétérogénéité des lecteurs. Pour répondre à ce besoin, nous avons développé une méthode d'évaluation automatique de la lisibilité, adaptée à trois profils de lecteurs (adultes experts/non-experts, enfants). Cette méthode s'appuie sur un corpus biomédical bilingue de 20 008 documents (11 154 en anglais, 8 854 en français), que nous avons constitué et rendons accessible librement. Elle utilise une architecture hybride combinant embeddings de transformers et caractéristiques linguistiques, atteignant un score F1 macro-moyen de 0,987. Cette approche ouvre des perspectives pour l'évaluation fine de la lisibilité, la personnalisation de la recherche d'information, et la validation de la lisibilité des résumés générés automatiquement.

ABSTRACT

Assessing the readability of biomedical texts according to the reader's profile

The readability of biomedical texts is perceived differently depending on the reader's profile, which is amplified by the intrinsic complexity of these documents and the uneven health literacy of the population. Although 72% of Internet users consult medical information online, a significant proportion still encounter comprehension difficulties, so assessing readability is essential to ensure that texts are accessible to a diverse audience. Conventional readability formulas, designed for general texts, do not take this diversity into account, underlining the need to adapt evaluation tools to the specific needs of biomedical texts and the heterogeneity of readers. To meet this need, we have developed an automatic readability evaluation method, adapted to three reader profiles (expert/non-expert adults, children). This method is based on a bilingual biomedical corpus of 20 008 documents (11 154 in English, 8 854 in French), which we have built up and are making freely available. It uses a hybrid architecture combining transformer embeddings and linguistic features, achieving a macro-average F1 score of 0.987. This approach opens up new prospects for fine-tuning readability, personalizing information retrieval and validating the readability of automatically-generated summaries.

MOTS-CLÉS : lisibilité, accésibilité, transformers, caractéristiques linguistiques, texte biomédical.

KEYWORDS: readability, accessibility, transformers, linguistic features, biomedical text.

1 Introduction

À l'ère du numérique, internet est devenu une source d'informations biomédicales largement utilisée. Selon une enquête du Pew Research Center, 72% des utilisateurs recherchent des informations sur la santé en ligne (Fox & Duggan, 2013). Toutefois, la littératie en santé varie considérablement au sein de la population. (Magnani *et al.*, 2018) ont ainsi révélé que 36% des adultes américains présentent

une littératie en santé de niveau basique ou inférieur, mettant en évidence la difficulté d'accès à l'information médicale pour une proportion significative d'usagers. Cette hétérogénéité, qui s'étend des professionnels de santé aux individus sans formation médicale, implique la nécessité d'une communication adaptée à des profils de lecteurs aux connaissances en santé variées (Carter *et al.*, 2024). Or, les documents biomédicaux sont perçus comme particulièrement complexes (Cheng Sheang *et al.*, 2022), ce qui constitue un obstacle à leur compréhension pour une large part des lecteurs profanes.

Dans ce contexte, l'évaluation de la lisibilité s'impose comme un levier essentiel pour garantir l'accessibilité des textes à un public varié. L'évaluation automatique de la lisibilité vise à estimer la facilité avec laquelle un texte peut être lu et compris. Elle revêt une importance particulière dans le domaine biomédical, où la clarté et la compréhensibilité des informations sont directement liées à des enjeux de santé publique.

Cependant, les méthodes d'évaluation de la lisibilité actuellement disponibles reposent encore majoritairement sur des métriques traditionnelles, initialement conçues pour des textes généraux, et peu adaptées à la complexité du discours des écrits biomédicaux. Cette inadéquation concerne tant les ressources en langue anglaise qu'en langue française. Bien que certaines formules de lisibilité aient été développées spécifiquement pour les textes biomédicaux en anglais, leur usage reste limitée non seulement en raison du manque d'API accessibles, mais aussi parce que les mesures traditionnelles telles que FKGL (Kincaid *et al.*, 1975) et SMOG (Mc Laughlin, 1969) offrent une mise en œuvre plus simple, une reconnaissance plus large et des références établies, ce qui conduit de nombreux chercheurs à utiliser par défaut ces outils familiers malgré leurs limites dans des contextes spécialisés.

Par ailleurs, un défi important réside dans le manque de ressources annotées pour l'évaluation de la lisibilité biomédicale, en particulier en français. Alors que l'anglais bénéficie de métriques spécialisées (Kim *et al.*, 2007; Proulx *et al.*, 2013; Leroy *et al.*, 2008) et de corpus plus étoffés, les travaux en français demeurent limités. De plus, les bases de données existantes catégorisent souvent les textes de manière binaire (adulte expert/non-expert), une distinction qui ne permet pas de refléter avec suffisamment de précision la diversité des profils de lecteurs. Cette limitation constitue un frein à la mise en place d'outils de recommandation ou de simplification réellement efficaces.

Pour pallier ces insuffisances, nous proposons une méthode d'évaluation automatique de la lisibilité des textes biomédicaux fondée sur le profil du lecteur. Nous avons ainsi constitué un corpus biomédical structuré en français et en anglais, intégrant trois niveaux de lisibilité : adulte expert, adulte non-expert et enfant.

Nos contributions s'articulent autour de trois axes majeurs :

1. Construction d'un corpus biomédical bilingue et structure :
   — constitution d'un ensemble de 20 008 documents (8 854 en français, 11 154 en anglais ;
   — classification des textes selon trois catégories de lecteurs (adultes experts, adultes non-experts, enfants), offrant une granularité plus fine que les approches binaires existantes.
2. Analyse linguistique approfondie :
   — identification des caractéristiques linguistiques distinctives des textes destinés à chaque profil de lecteur ;
   — mise en évidence des spécificités de lisibilité entre le français et l'anglais dans un contexte biomédical.
3. Modélisation et optimisation de la prédiction de la lisibilité :
   — expérimentation de modèles d'apprentissage automatique traditionnels (XGBoost, Random Forest) et de modèles de type transformers spécialisés (BERT (Devlin *et al.*, 2019) ; CamemBERT (Martin *et al.*, 2020) ; BioBERT (Lee *et al.*, 2020) ; DrBERT (Labrak *et al.*, 2023) ;
   — développement d'un modèle hybride pour l'évaluation de la lisibilité des textes biomédicaux en anglais.

Notre travail offre des perspectives pour :

1. la simplification de textes : en identifiant en amont les contenus nécessitant une réécriture adaptée à un lectorat spécifique ;
2. la personnalisation des résultats de recherche d'informations : en filtrant les textes disponibles en ligne selon leur lisibilité et le profil de l'utilisateur ;
3. l'évaluation automatique de résumés biomédicaux générés : en vérifiant leur accessibilité pour des publics variés.

En fournissant une évaluation fine de la lisibilité tenant compte du profil du lecteur, notre travail contribue à améliorer l'accès équitable à l'information biomédicale, dans un environnement où celle-ci est de plus en plus consultée en autonomie par des utilisateurs aux niveaux de littératie en santé hétérogènes.

2 Etat de l'art

2.1 Approches générales d'évaluation de la lisibilité

2.1.1 Métriques de lisibilité traditionnelles

Les mesures traditionnelles telles que SMOG, FKGL, FRE (Flesch, 1948) et GFI (Gunning, 1952) évaluent la lisibilité sur la base de caractéristiques de surface telles que la longueur des mots et des phrases. Parmi ces formules, FRE a notamment été adaptée à la langue française (Kandel & Moles, 1958). Ces formules, bien que largement utilisées, présentent des limites importantes. (Uluslu, 2023) remet en question l'application des formules de lisibilité anglaises à d'autres langues, soulignant notamment que leur utilisation sur des textes turcs à tendence à surestimer la lisibilité (Akgül, 2019, 2024). Par ailleurs, les corrélations entre les formules de lisibilité et la difficulté perçue par les utilisateurs sont très faibles (Zheng & Yu, 2017; Carter *et al.*, 2024), notamment dans le domaine biomédical où, selon (Zheng & Yu, 2017), les prédictions sur les dossiers de santé électroniques ne reflètent pas leur réelle difficulté.

2.1.2 Approches linguistiques avancées et modèles d'apprentissage

Les approches modernes d'évaluation automatique de la lisibilité (ARA) intègrent une ingénierie de caractéristiques linguistiques avancées (syntaxiques, sémantiques, discursives, etc.), définies notamment à partir de taxonomies comme celle de (Collins-Thompson, 2014). Ces caractéristiques alimentent divers modèles d'apprentissage automatique. Par exemple, pour le turc (Uluslu, 2023) (classification en 3 niveaux sur un corpus de magazines scientifiques), cinq catégories de caractéristiques ont été utilisées : traditionnelles, syntaxiques, lexico-sémantiques, morphologiques et discursives. Les modèles testés ont obtenu des scores F1 de 85,1% pour Random Forest (RF), 82,3% pour SVM, 78,4% pour la régression logistique et 83,7% pour XGBoost. Pour le basque (Gonzalez-Dios *et al.*, 2014) (classification binaire sur des articles scientifiques), les caractéristiques étaient réparties en six groupes : globales, lexicales, morphologiques, morpho-syntaxiques, syntaxiques et pragmatiques. Les performances respectives des modèles étaient de 89,5% pour SVM (avec SMO), 86,7% pour RF, 84,2% pour J48, 82,5% pour IBk et 72,3% pour Naive Bayes (NB). Ces résultats soulignent que l'efficacité de l'ARA dépend étroitement de la qualité des caractéristiques linguistiques, des outils TAL disponibles et du choix de modèle avec des scores F1 variant ici entre 72% et 89,5%.

2.1.3 Approches basées sur l'apprentissage profond

Ces méthodes utilisent des architectures de réseaux neuronaux, (Azpiazu & Pera, 2019; Deutsch *et al.*, 2020; Qiu *et al.*, 2021) et les transformeurs (ex. BERT et ses variantes), pour traiter directement le texte et prédire son niveau de lisibilité. L'affinage (fine-tuning) de modèles linguistiques préentraînés sur des corpus de lisibilité annotés a permis d'atteindre des performances à l'état de l'art pour l'évaluation de la lisibilité dans plusieurs langues.

2.1.4 Approches hybrides

Ces approches combinent les avantages des méthodes basées sur des caractéristiques linguistiques et des modèles d'apprentissage profond (Uluslu, 2023). Elles peuvent fusionner caractéristiques linguistiques et embeddings, ou utiliser les prédictions neuronales pour alimenter des modèles classiques (et vice-versa). Les modèles hybrides visent à tirer parti à la fois de la connaissance linguistique explicite et de la compréhension implicite capturée par les réseaux neuronaux, conduisant souvent à des améliorations de performance significatives. Deux illustrations récentes : (Imperial, 2021) combine des caractéristiques linguistiques (syntaxiques, morphologiques, sémantiques) avec des embeddings BERT dans un vecteur fusionné, utilisé pour entraîner des modèles traditionnels (SVM, RF, LR), améliorant les performances en philippin sur un corpus d'histoires pour enfants.(Lee *et al.*, 2021) concatènent les prédictions probabilistes (soft labels) de modèles Transformers (BERT, RoBERTa (Liu *et al.*, 2019), BART (Lewis *et al.*, 2019), XLNet (Yang *et al.*, 2019)) avec des caractéristiques linguistiques. Le vecteur combiné alimente des modèles non neuronaux (LR, SVM, RF, XGBoost), avec de bonnes performances, notamment sur de petits jeux de données comme WeeBit (Vajjala & Meurers, 2012), OneStopEnglish (Vajjala & Lučić, 2018) et Cambridge (Xia *et al.*, 2019) .

2.2 Lisibilité des textes biomédicaux

Les textes devant être évalués en termes de lisibilité sont caractérisés principalement par deux dimensions : la langue dans laquelle ils sont rédigés et leur thématique. Ces deux dimensions influencent considérablement les approches à adopter pour une évaluation précise de la lisibilité.

Pour l'anglais, plusieurs modèles spécifiques ont été développés pour évaluer la lisibilité des textes biomédicaux. (Kim *et al.*, 2007) a introduit le Distance Score, qui compare les documents à des références "faciles" et "difficiles" à travers diverses caractéristiques linguistiques. (Leroy *et al.*, 2008) ont combiné des scores de lisibilité, une analyse linguistique et des évaluations utilisateurs pour mesurer la complexité des informations de santé. Le système ReDE (Proulx *et al.*, 2013) analyse les textes pour identifier les concepts difficiles et proposer des simplifications. Plus récemment, (Devaraj *et al.*, 2021) ont utilisé des métriques basées sur les modèles de langage masqués pour classifier les résumés techniques et les résumés en langage simple des revues médicales, tandis que (Crossley *et al.*, 2020) ont développé le modèle MoTeR-P, qui utilise 85 variables linguistiques pour évaluer la difficulté des messages médicaux. Malgré ces formules de lisibilité spécailités, les formules traditionnelles demeurent largement utilisées dans la pratique, probablement en raison de leur facilité d'implémentation et d'utilisation.

Pour le français, les ressources sont beaucoup plus limitées. Le laboratoire CENTAL a développé des outils comme AMesure[1] et FABRA[2] (François *et al.*, 2014; Wilkens *et al.*, 2022). Ces outils sont principalement conçus pour évaluer la lisibilité de textes, notamment scolaires et administratifs, AMesure ciblant plus spécifiquement les textes administratifs. FABRA, quant à lui, a fait l'objet de tests sur des textes destinés à des locuteurs natifs et à des apprenants du français langue étrangère (FLE), selon différents niveaux de difficulté. Toutefois, ils restent inadaptés aux particularités des textes biomédicaux.

Les lacunes pour le français sont multiples : l'absence de corpus annotés pour les textes biomédicaux en français, le manque de modèles spécifiques pour évaluer la lisibilité des textes médicaux, l'insuffisance d'études comparatives entre différents niveaux de lecteurs (experts, non-experts, enfants), la difficulté d'adapter les approches développées pour l'anglais aux spécificités linguistiques du français, ou l'absence d'évaluation de l'impact des terminologies médicales françaises sur la lisibilité. Ces lacunes souligent l'importance de créer des ressources et modèles adaptés à l'évaluation de la lisibilité des textes biomédicaux en français, en tenant compte de ses spécificités linguistiques.Notre étude s'intéresse particulièrement aux textes biomédicaux en français, en utilisant l'anglais comme langue étalon en raison de la richesse des ressources et des recherches disponibles dans cette langue.

1. https://cental.uclouvain.be/amesure/
2. https://cental.uclouvain.be/fabra/

3 Méthodologie

3.1 Corpus

Dans le but d'analyser la lisibilité des textes en fonction de l'expertise et de l'âge des lecteurs, nous avons constitué deux corpus, l'un en français et l'autre en anglais (Tableau 1). Le corpus français combine le corpus *Cochrane* en français[1] (7 619 textes : 3 810 pour les experts, 3 809 pour le grand public) et le corpus *WikipediaVikidia*[2] (Grabar & Cardon, 2018) (1 235 textes : 660 pour les adultes, 575 simplifiés pour les enfants). Le corpus *CochraneWikiviki* qui en résulte contient 8 854 textes, classés en *adultes experts* (3 810 textes d'experts Cochrane), *adultes non experts* (4 469 textes Cochrane grand public et textes originaux de Wikipédia[3], et *enfants* (575 textes Vikidia simplifiés[4]). De même, le corpus anglais, *CochranePlabaSJK*, comprend 11 154 textes provenant de trois sources : *Cochrane*[5] *(Guo* et al.*, 2021) en anglais* (8 918 textes), *Plaba*[6] (Attal *et al.*, 2023) (1 668 textes), et *SJK*[7] (Stefanou *et al.*, 2024) (568 textes). Suivant la même structure que le corpus français, les textes ont été classés en *adultes experts* (Cochrane et textes d'experts Plaba), *adultes non-experts* (textes Cochrane grand public et résumés SJK originaux[8]), et *enfants* (textes SJK simplifiés). La correspondance entre les étiquettes originales et finales est résumée dans leTable 1. Notre ensemble de données est accessible ici[9]

Corpus	Source	Nombre de textes	Etiquette initiale	Etiquette finale
CochraneWikiviki (Français)	Cochrane FR	3 810	Textes experts	Adultes experts
	Cochrane FR	3 809	Textes grand public	Adultes non-experts
	Wikipedia	660	Textes originaux	Adultes non-experts
	Vikidia	575	Textes simplifiés	Enfants
Total CochraneWikiviki		**8 854**	—	**Adultes experts (3 810), Adultes non-experts (4 469), Enfants (575)**
CochranePlabaSJK (Anglais)	Cochrane EN	4 459	Textes experts	Adultes experts
	Cochrane EN	4 459	Textes simplifiés	Adultes non-experts
	Plaba	749	Textes experts	Adultes experts
	Plaba	919	Textes simplifiés	Adultes non-experts
	SJK	284	Résumés scientifiques	Adultes non-experts
	SJK	284	Versions simplifiées	Enfants
Total CochranePlabaSJK		**11 154**	—	**Adultes experts (5 208), Adultes non-experts (5 662), Enfants (284)**

TABLE 1 – Résumé de la constitution des corpus avec étiquettes intiales et finales

Le tableau 2 révèle des schémas linguistiques distincts selon les groupes d'utilisateurs et les corpus. Dans *CochraneWikiwiki*, les textes non-experts sont les plus longs tandis que le contenu expert est plus concis. Cette variabilité est probablement due à la combinaison des résumés Cochrane et des articles plus longs de Wikipédia. Les textes pour enfants, en revanche, sont plus courts. Le corpus *CochranePlabaSJK* présente des longueurs de texte plus cohérentes, en particulier pour les documents destinés aux adultes ce qui reflète sa composition de résumés au format similaire. Les textes pour enfants de ce corpus sont particulièrement courts. Cela confirme que la complexité linguistique des documents d'information sur la santé varie en fonction du public cible et de la source.

1. https://www.cochranelibrary.com
2. Sous-ensemble du corpus CLEAR : http://natalia.grabar.free.fr/CLEAR/clear-res.php
3. https://fr.wikipedia.org
4. https://fr.vikidia.org
5. https://github.com/qiuweipku/Plain_language_summarization/tree/main/CDSR_data
6. https://bionlp.nlm.nih.gov/plaba2023/
7. https://sciencejournalforkids.org
8. https://github.com/loukritia19/science-journal-for-kids-data
9. BioReadMatch corpus https://github.com/Anyantd/BioReadMatch-Corpus

Etiquette	MoyM	ETM	MoyP	ETP	NbDoc
Corpus CochraneWikiwiki					
adulte_ex	756,73	280,09	34,62	9,42	3810
adulte_gp	910,71	2355,60	35,63	88,51	4469
enfant	382,89	449,86	17,52	18,73	575
Corpus CochranePlabaSJK					
adulte_ex	406,73	180,10	13,94	5,60	5208
adulte_gp	249,00	122,44	10,25	5,13	5662
enfant	161,94	34,70	9,71	2,81	284

TABLE 2 – Résultats des statistiques descriptives pour les sous-corpus *CochraneWikiwiki* et *CochranePlabaSJK*

*MoyM/MoyP = Moyenne des mots/phrases par document ; ETM/ETP = Écart-type des mots/phrases ; NbDoc = Nombre total de documents.

Etiquette	FRE	FKGL	TTR
adulte_ex	54,65	14,08	0,40
adulte_gp	53,98	14,19	0,49
enfant	70,48	10,35	0,51

TABLE 3 – Scores de lisibilité pour le corpus français *CochraneWikiviki*

Etiquette	FRE	FKGL	TTR
adulte_ex	44,61	10,86	0,43
adulte_gp	42,72	12,56	0,51
enfant	65,83	8,00	0,57

TABLE 4 – Scores de lisibilité pour le corpus anglais *CochranePlabaSJK*

* FRE = Flesch Reading Ease ; FKGL = Flesch-Kincaid Grade Level ; TTR = Type-Token Ratio

Nous avons également effectué une analyse préliminaire des trois niveaux de lecture du corpus et, comme le montre les tableaux 3, 4, les formules de lisibilité standard ([FRE ; 1-100, plus élevé = plus facile] et [FKGL ; niveau scolaire américain]) ne permettent pas de distinguer de manière fiable les textes d'adultes experts de ceux qui ne le sont pas dans des contextes biomédicaux. En anglais, les passages experts semblent même légèrement plus lisibles que les passages profanes (FRE 44,61 contre 42,72 ; FKGL 10,86 contre 12,56), avec une tendance similaire en français, mais les deux catégories d'adultes restent dans la bande « difficile » (niveau collégial pour l'anglais ; assez difficile, classes 10-12, pour le français). Parce que ces mesures dépendent uniquement de la longueur des phrases et du nombre de syllabes, elles interprètent à tort la précision technique comme de la simplicité et pénalisent les formulations explicatives plus longues typiques des résumés non spécialisés. En revanche, les textes pour enfants sont sans ambiguïté plus faciles (FRE 65,83 en anglais ; 70,48 en français ; FKGL 8,00 et 10,35 respectivement). En outre, le rapport Type-Token augmente entre les experts, les non-experts et les enfants (0,43 / 0,51 / 0,57 en anglais), traduisant une plus grande variété lexicale dans les textes simplifiés. Cela pourrait s'expliquer par l'ajout d'explications visant à rendre le contenu plus accessible. Si les scores de lisibilité distinguent bien les textes pour enfants de ceux pour adultes, ils peinent à différencier les documents pour adultes experts et non-experts.

3.2 Les caractéristiques associées à la lisibilité

En utilisant LFTK [10] (Lee & Lee, 2023) , un système multilingue pour l'extraction de caractéristiques linguistiques, nous avons extrait 141 caractéristiques des textes sur nos deux corpus : *CochraneWikiviki* et *CochranePlabaSJK*. Ces caractéristiques couvrent plusieurs domaines linguistiques, classés en *lexico-sémantique, syntaxe, discours et métriques de surface*. Chaque domaine se concentre sur des propriétés linguistiques spécifiques

— *Lexico-Sémantique* : capture les attributs liés au choix des mots et au sens, en se concentrant sur la diversité et la richesse du vocabulaire.

10. LFTK : https://github.com/brucewlee/lftk

— *Syntaxe* : examine la disposition et la structure des mots et des phrases, y compris les catégories grammaticales et la structure des phrases.
— *Discours* : traite des relations de haut niveau entre les mots et les phrases, en se concentrant sur la cohésion et le flux d'idées dans un texte.
— *Surface* : englobe les caractéristiques de surface, qui ne sont souvent pas liées à des propriétés linguistiques spécifiques, mais qui saisissent les caractéristiques générales du texte, telles que la longueur des mots et des phrases ou le nombre de syllabes.

En plus des caractéristiques linguistiques, nous avons évalué le corpus anglais en utilisant le score proposé dans (Devaraj *et al.*, 2021), qui est basé sur un modèle de langage masqué (MLM). Plus précisément, nous avons utilisé SciBERT[11] pour estimer le degré d'alignement d'un texte sur les modèles linguistiques capturés par le modèle. Une valeur élevée indique que le texte est bien aligné sur le modèle, ce qui suggère qu'il est proche du style et du domaine sur lesquels SciBERT a été entrainé, tandis qu'une valeur plus faible peut signaler une divergence linguistique ou conceptuelle. Nous avons appelé ce score Masked Random Token-based Text Complexity (MRTTC), en accord avec (Luo *et al.*, 2022). Cette évaluation n'a été appliquée qu'au corpus anglais, SciBERT n'étant pas un modèle multilingue.

3.3 Approches de modélisation

Pour prédire la lisibilité des textes biomédicaux en fonction du profil du lecteur (adultes experts, adultes non-experts ou enfants), nous avons expérimenté les approches ci-dessous :

3.3.1 Modèles d'apprentissage automatique avec caractéristiques linguistiques

Nous avons défini deux configurations de référence (baselines), construites à l'aide de caractéristiques linguistiques extraites avec LFTK (détails en section 3.2). En l'absence de baseline établie pour la lisibilité biomédicale en français, et afin d'assurer une approche méthodologique uniforme et translinguistique comparable à l'anglais (malgré les travaux existants pour cette dernière), nous avons développé nos propres baselines en nous inspirant des travaux de (Uluslu, 2023) pour l'évaluation de la lisibilité des textes trucs. La première, *TB* : basée sur des caractéristiques de surface (longueur des phrases, des mots, etc.) inspirées des formules de lisibilité classiques. La seconde, *MB* : combinaison des caractéristiques de surface avec des indices lexico-sémantiques, syntaxiques et discursifs (141 au total). Nous avons ensuite entraîné quatre modèles d'apprentissage automatique (Forêts aléatoires, XGBoost, régression logistique et SVM) sur ces ensembles de caractéristiques, en utilisant une validation croisée répétée 10 fois et une optimisation des hyperparamètres via RandomizedSearchCV. Les performances ont été évaluées à l'aide de l'exactitude, de la précision, du rappel et du score F1 macro-moyen.

3.3.2 Approche basée sur les modèles de language

Nous avons utilisé quatre modèles pré-entraînés basés sur les transformeurs : CamemBERT[12], DrBERT[13] (pour le français), BERT[14], et BioBERT[15] (pour l'anglais). Nous avons affiné ces derniers en utilisant 80% des données pour l'entraînement et 20% pour les tests, en réalisant une validation croisée par entraînement multiple avec différentes initialisations et taux d'apprentissage (1e-5, 2e-5, 5e-5). Chaque modèle a été entraîné pour 3 époques par exécution. L'optimisation a été réalisée avec AdamW et un planificateur linéaire.

3.3.3 Approche hybride

Pour évaluer la lisibilité du corpus anglais, nous avons développé une approche hybride combinant caractéristiques linguistiques, scores de perplexité et embeddings générés par BioBERT. Comme

11. SciBERT : https://huggingface.co/allenai/scibert_scivocab_uncased
12. camembert-base : https://huggingface.co/almanach/camembert-base
13. DrBERT-7GB : https://huggingface.co/Dr-BERT/DrBERT-7GB
14. bert-base-uncased : https://huggingface.co/docs/transformers/model_doc/bert
15. biobert-v1.1 : https://huggingface.co/dmis-lab/biobert-v1.1

illustré dans la Figure 1, nous avons extrait 141 caractéristiques linguistiques, dont les 10 plus informatives ont été sélectionnées via un modèle XGBoost. En parallèle, nous avons calculé le score MRTTC, basé sur la perplexité issue de SciBERT (Beltagy *et al.*, 2019), et généré des embeddings BioBERT. Chaque entrée : caractéristiques, score MRTTC, embeddings, a été normalisée séparément, puis concaténée en un vecteur de 779 dimensions, utilisé comme entrée d'un perceptron multicouche (MLP) composé d'une couche cachée dense (256 neurones, ReLU), suivie d'une normalisation par lots et d'un dropout (taux = 0,5). La prédiction de la lisibilité est effectuée via une couche de sortie softmax. Contrairement aux approches des sections 3.3.1 et 3.3.2, utilisées pour les corpus français et anglais, cette méthode hybride a été conçue uniquement pour l'anglais, SciBERT et BioBERT n'étant pas multilingues. Nous avons également évalué plusieurs variantes hybrides : SBioEmb-PPL et XBioEmb-PPL, associant embeddings BioBERT non affinés, caractéristiques linguistiques et scores de perplexité, ainsi que SBioSL-PPL et XBioSL-PPL, intégrant des soft-labels au lieu d'étiquettes dures. Les résultats sont présentés dans le tableau 10.

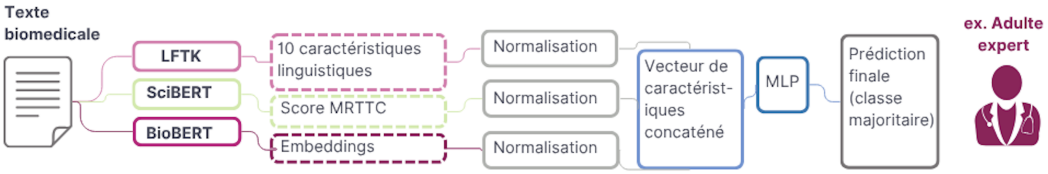

FIGURE 1 – Pipeline pour l'évaluation de la lisibilité des textes biomédicaux (détails en section 3.3.3)

4 Résultats

4.1 Performance des classificateurs ML traditionnels

Sur les deux corpus (français et anglais), les modèles d'apprentissage automatique entrainés sur les 141 caractéristiques linguistiques ont obtenu des performances élevées et plutôt similaires. Sur le corpus français (Tableau 5), XGBoost s'est montré légèrement supérieur (macro-F1 = 0,946), tandis que sur le corpus anglais (Tableau 7), le SVM a obtenu les meilleurs résultats (macro-F1 = 0,891). Ces performances confirment la pertinence des caractéristiques sélectionnées. L'analyse par ablation met en évidence l'importance des caractéristiques lexico-sémantiques, qui améliorent nettement les résultats dans les deux langues (Tableau 6) , (Tableau 8) : le score macro-F1 passe de 0,850 à 0,940 pour le français, et de 0,776 à 0,891 pour l'anglais. On note cependant que les caractéristiques de surface ("SUR") sont plus prédictives en français qu'en anglais, ce qui suggère une dépendance plus forte aux indices structurels dans les textes français, tandis que la lisibilité en anglais semble reposer davantage sur des aspects lexico-sémantiques.

Modèle	Acc.	Prec.	Rec.	Macro avg f1
XGBoost	0,957	0,937	0,955	**0,946**
RF	0,947	0,926	0,951	0,938
SVM	0,954	0,930	0,953	0,941
LR	0,955	0,932	0,958	0,944

TABLE 5 – Performance des méthodes ML sur le corpus français

Domaines	Modèle	Macro avg f1
SUR	XGBoost	0,850
+LXSM	XGBoost	0,940
+SYN	XGBoost	0,947
+DISC	XGBoost	**0,946**

TABLE 6 – Performance par combinaison de domaines linguistiques sur le corpus français

Modèle	Acc.	Prec.	Rec.	Macro avg f1
XGBoost	0,871	0,896	0,883	0,889
RF	0,852	0,897	0,841	0,864
SVM	**0,879**	**0,903**	0,883	**0,891**
LR	0,874	0,887	0,891	0,889

TABLE 7 – Perfomance sur des méthodes ML sur le corpus anglais

Domaines	Modèle	Macro avg f1
SUR	SVM	0,776
+LXSM	SVM	0,888
+SYN	SVM	0,891
+DISC	SVM	**0,891**

TABLE 8 – Performance par combinaison de domaines linguistiques sur le corpus anglais

4.2 Performance des modèles basés sur les transformers

Les modèles de langage affinés surpassent nettement les approches d'apprentissage machine traditionnelles avec des gains de +4,1% en français (FT-DrBERT : 0,985 vs MB : 0,946) (Tableau 9) et jusqu'à +9,6% en anglais (FT-BioBERT : 0,987 vs MB : 0,891). (Tableau 10). Les modèles spécialisés biomédicaux (DrBERT, BioBERT) surpassent légèrement leurs homologues généralistes (CamemBERT, BERT). La comparaison inter-linguistique révèle des performances globalement équivalentes entre le français et l'anglais pour les modèles de langage affinés($F_1 \approx 0{,}98 \pm \Delta F$), contrastant avec un écart plus important pour les approches MB traditionnelles (0,946 vs 0,891), indiquant une meilleure robustesse des transformeurs face aux spécificités linguistiques.

Modèle	Acc.	Prec.	Rec.	F1
TB	0,864	0,864	0,864	0,850
MB	0,957	0,957	0,957	0,946
FT-DrBERT	0,995	0,995	0,995	**0,985**
FT-CamBERT	0,995	0,995	0,995	<u>0,984</u>

TABLE 9 – Résultats des différentes approches selon les métriques d'évaluation sur le corpus français.

TB / MB : Ligne de base traditionnelle / moderne, définie dans la section 3.3.1.
FT-DrBERT / FT-CamBERT : Modèles DrBERT et CamBERT affinés sur le corpus français.
Note : **gras** = meilleur, <u>souligné</u> = deuxième meilleur.

Modèle	Acc.	Prec.	Rec.	F1
TB	0,797	0,797	0,797	0,797
MB	0,879	0,903	0,883	0,891
FT-BioBERT	0,982	0,982	0,982	**0,987**
FT-BERT	0,970	0,970	0,970	0,979
SBioEmb-PPL	0,948	0,948	0,948	0,956
XBioEmb-PPL	0,948	0,948	0,948	0,958
SBioSL-PPL	0,855	0,855	0,855	0,862
XBioSL-PPL	0,870	0,870	0,870	0,887
BioReadMatch (notre modèle)	0,990	0,984	0,987	**0,987**

TABLE 10 – Résultats des différentes approches selon les métriques d'évaluation sur le corpus anglais.

TB / MB : Référence (baseline) traditionnelle / moderne, telle que définie dans la section 3.3.1.
FT-BioBERT / FT-BERT : Modèles BioBERT / BERT affinés.
Modèles hybrides : Classificateurs MLP entraînés sur les 10 caractéristiques linguistiques les plus pertinentes (via SVM ou XGB), les scores MRTTC (perplexité), et les embeddings BioBERT (Emb) ou les soft-labels (SL).
BioReadMatch : Version améliorée de XBioEmb-PPL avec un BioBERT affiné sur notre corpus.
Note : **Gras** = meilleur, souligné = deuxième meilleur.

5 Discussion

5.1 Interprétation du modèle

L'analyse par permutation des caractéristiques(feature importance) du corpus anglais (Figures 2 et 3) montre que SVM et XGBoost partagent plusieurs attributs tels que (a-char-pw, a-num-pw, n-upunct, a-num-ps, t-stopword). Nous avons privilégié les 10 meilleures caractéristiques XGBoost pour BioReadMatch, choix validé par les performances supérieures des approches hybrides : SBioSL-PPL et XBioSL-PPL atteignent respectivement 0,862 et 0,887 en macro-F1 avec des soft-labels, tandis que SBioEmb-PPL et XBioEmb-PPL obtiennent 0,956 et 0,958 avec des hard-labels (Tableau 10). XGBoost étant le modèle le plus performant sur le corpus français (Tableau 5), nous avons sélectionné les 10 meilleures caractéristiques (Figure 4) en appliquant la méthode de permutation des caractéristiques sur le modèle XGBoost.

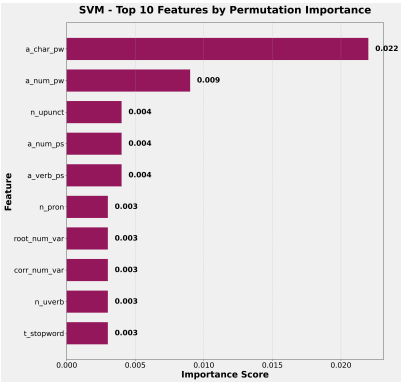

FIGURE 2 – Caractéristiques linguistiques les plus importantes pour SVM sur le corpus anglais[1]

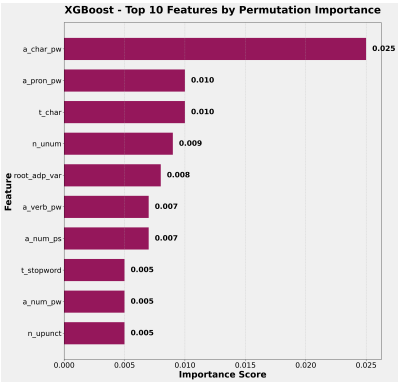

FIGURE 3 – Caractéristiques linguistiques les plus importantes pour XGBoost sur le corpus anglais

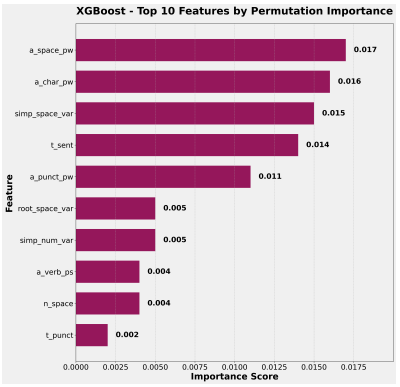

FIGURE 4 – Caractéristiques linguistiques les plus importantes pour XGBoost sur le corpus français

5.2 Corrélation des descripteurs

L'analyse des corrélations de Spearman Table 11 appliquée aux caractéristiques linguistiques clés pour XGBoost dans le corpus français révèle des différences structurelles marquées entre les textes destinés aux experts et ceux plus accessibles. Les textes spécialisés se caractérisent par une structure syntaxique dense : un nombre élevé de phrases (t_sent, -0.563), une forte densité de ponctuation (t_punct, -0.528), et une grande variation des espaces ($simp_space_var$, -0.511 ; $root_space_var$, -0.497). La ponctuation par mot (a_punct_pw, -0.330) renforce cette complexité syntaxique. À l'inverse, la moyenne de verbes par phrase (a_verb_ps, 0.207) montre que les textes plus accessibles tendent vers une narration plus fluide. D'autres variables, comme la variation numérique ($simp_num_var$, 0.116) ou le nombre moyen de caractères par mot (a_char_pw, 0.025), présentent des corrélations faibles, suggérant un impact limité sur la perception de la complexité en français.

Par ailleurs, l'analyse des corrélations de Spearman Table 12 permet de mieux comprendre les caractéristiques linguistiques distinctives des textes à différents niveaux de lisibilité dans un corpus anglais. Les textes les plus accessibles se distinguent par une richesse verbale et pronomiale particulièrement riche, avec une fréquence élevée de verbes par mot (a_verb_pw, corrélation de 0,549) et de pronoms par mot (a_pron_pw, 0,454).La variation des adpositions ($root_adp_var$, 0,307) suggère également une structure plus fluide, facilitant la compréhension, notamment pour un public non expert dans le

15. Liste des caractéristiques linguistiques https://docs.google.com/spreadsheets/d/1uXtQ1ah0OL9cmHp2Hey0QcHb4bifJcQFLvYlVIAWWwQ/edit?gid=693915416#gid=693915416

domaine biomédical.À l'inverse, les textes plus complexes, en particulier dans les corpus Cochrane et Plaba présentent une densité numérique marquée, reflet de la rigueur scientifique. On observe une corrélation négative avec le nombre de nombres uniques (n_unum, -0,627), de chiffres par mot (a_num_pw, -0,593) et de nombres par phrase (a_num_ps, -0,605). Ces éléments soulignent une volonté de précision : quantification des résultats, description des protocoles, ou communication de données statistiques. La variation ajustée des nombres ($corr_num_var$, -0,614), la densité en ponctuations uniques (n_upunct, -0,486) et le nombre total de caractères (t_char, -0,474) confirment une structure textuelle plus dense et techniquement exigeante, typique des publications scientifiques.

Ainsi, l'analyse comparative met en évidence n_upunct, t_punct et a_char_pw comme descripteurs clés de la complexité textuelle dans les corpus anglais et français, reflétant une structuration syntaxique et informationnelle étroitement liée au degré de sophistication des textes scientifiques.

Feature	CS
t_sent	-0,563
t_punct	-0,528
simp_space_var	-0,511
root_space_var	-0,497
a_punct_pw	-0,330
a_verb_ps	0,207
n_space	-0,118
simp_num_var	0,116
a_space_pw	-0,082
a_char_pw	0,025

TABLE 11 – Corrélation de Spearman (CS) - corpus français

Feature	CS
n_unum	-0,627
a_num_ps	-0,605
a_num_pw	-0,593
a_verb_pw	0,549
n_upunct	-0,486
t_char	-0,474
a_pron_pw	0,454
root_adp_var	0,307
t_stopword	-0,296
a_char_pw	0,199

TABLE 12 – Corrélation de Spearman (CS) - corpus anglais

6 Conclusion

En conclusion, nous avons introduit un corpus biomédical bilingue de 20 008 documents, spécifiquement conçu pour l'évaluation de la lisibilité selon le profil du lecteur (adultes experts/non-experts, enfants). Nos analyses montrent que l'intégration de caractéristiques linguistiques avancées améliore significativement la prédiction du niveau de lisibilité. Les modèles de type Transformer, tels que BioBERT, DrBERT et CamemBERT, atteignent des performances (macro-F1 jusqu'à 0.987), mais souffrent d'un manque d'explicabilité. Pour répondre à cette limitation, nous avons proposé BioReadMatch, une architecture hybride combinant embeddings et indices linguistiques interprétables. Ce modèle offre un bon compromis entre performance et transparence, atteignant lui aussi un macro-F1 de 0.987. Ce travail ouvre des perspectives pour la simplification du langage biomédical, la personnalisation de la recherche d'information, et l'évaluation automatique de l'accessibilité des résumés automatiques des textes médicaux.

Références

AKGÜL Y. (2019). The accessibility, usability, quality and readability of turkish state and local government websites an exploratory study. *International Journal of Electronic Government Research (IJEGR)*, **15**(1), 62–81.

AKGÜL Y. (2024). Evaluating the performance of websites from a public value, usability, and readability perspectives : a review of turkish national government websites. *Universal Access in the Information Society*, **23**(2), 975–990.

ATTAL K., ONDOV B. & DEMNER-FUSHMAN D. (2023). A dataset for plain language adaptation of biomedical abstracts. *Scientific Data*, **10**(1), 8.

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.

BELTAGY I., LO K. & COHAN A. (2019). Scibert : A pretrained language model for scientific text. *arXiv preprint arXiv :1903.10676*.

CARTER B., NAYAK K. & VEMBENIL I. (2024). The accuracy of readability formulas in health content : a systematic review. *J Health Commun*, **9**, 9002.

CHENG SHEANG K., KOPTIENT A., GRABAR N. & SAGGION H. (2022). Identification of complex words and passages in medical documents in French. In Y. ESTÈVE, T. JIMÉNEZ, T. PARCOLLET & M. ZANON BOITO, Éds., *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, p. 116–125, Avignon, France : ATALA.

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

CROSSLEY S. A., BALYAN R., LIU J., KARTER A. J., MCNAMARA D. & SCHILLINGER D. (2020). Predicting the readability of physicians' secure messages to improve health communication using novel linguistic features : Findings from the eclippse study. *Journal of communication in healthcare*, **13**(4), 344–356.

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

DEVARAJ A., WALLACE B. C., MARSHALL I. J. & LI J. J. (2021). Paragraph-level simplification of medical texts. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, volume 2021, p. 4972.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). Bert : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics : human language technologies, volume 1 (long and short papers)*, p. 4171–4186.

FLESCH R. (1948). A new readability yardstick. *Journal of applied psychology*, **32**(3), 221.

FOX S. & DUGGAN M. (2013). Health online 2013. *Health*, **2013**, 1–55.

FRANÇOIS T., BROUWERS L., NAETS H. & FAIRON C. (2014). Amesure : a readability formula for administrative texts (amesure : une plateforme de lisibilité pour les textes administratifs)[in french]. In *Proceedings of TALN 2014 (Volume 2 : Short Papers)*, p. 467–472.

GONZALEZ-DIOS I., ARANZABE M. J., DE ILARRAZA A. D. & SALABERRI H. (2014). Simple or complex ? assessing the readability of basque texts. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics : Technical papers*, p. 334–344.

GRABAR N. & CARDON R. (2018). Clear-simple corpus for medical french. In *ATA*.

GUNNING R. (1952). The technique of clear writing. *(No Title)*.

GUO Y., QIU W., WANG Y. & COHEN T. (2021). Automated lay language summarization of biomedical scientific reviews. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, p. 160–168.

IMPERIAL J. M. (2021). Bert embeddings for automatic readability assessment. *arXiv preprint arXiv :2106.07935*.

KANDEL L. & MOLES A. (1958). Application de l'indice de flesch à la langue française. *Cahiers Etudes de Radio-Télévision*, **19**(1958), 253–274.

KIM H., GORYACHEV S., ROSEMBLAT G., BROWNE A., KESELMAN A. & ZENG-TREITLER Q. (2007). Beyond surface characteristics : a new health text-specific readability measurement. In *AMIA Annual Symposium Proceedings*, volume 2007, p. 418.

KINCAID J. P., FISHBURNE JR R. P., ROGERS R. L. & CHISSOM B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

LABRAK Y., BAZOGE A., DUFOUR R., ROUVIER M., MORIN E., DAILLE B. & GOURRAUD P.-A. (2023). DrBERT : A robust pre-trained model in French for biomedical and clinical domains. In A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, Éds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 16207–16221, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.896.

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*.

LEE B. W. & LEE J. (2023). LFTK : Handcrafted features in computational linguistics. In E. KOCHMAR, J. BURSTEIN, A. HORBACH, R. LAARMANN-QUANTE, N. MADNANI, A. TACK, V. YANEVA, Z. YUAN & T. ZESCH, Éds., *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, p. 1–19, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.bea-1.1.

LEE J., YOON W., KIM S., KIM D., KIM S., SO C. H. & KANG J. (2020). Biobert : a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, **36**(4), 1234–1240.

LEROY G., HELMREICH S., COWIE J. R., MILLER T. & ZHENG W. (2008). Evaluating online health information : Beyond readability formulas. In *AMIA Annual Symposium Proceedings*, volume 2008, p. 394.

LEWIS M., LIU Y., GOYAL N., GHAZVININEJAD M., MOHAMED A., LEVY O., STOYANOV V. & ZETTLEMOYER L. (2019). Bart : Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv :1910.13461*.

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*.

LUO Z., XIE Q. & ANANIADOU S. (2022). Readability controllable biomedical document summarization. *arXiv preprint arXiv :2210.04705*.

MAGNANI J. W., MUJAHID M. S., ARONOW H. D., CENÉ C. W., DICKSON V. V., HAVRANEK E., MORGENSTERN L. B., PAASCHE-ORLOW M. K., POLLAK A., WILLEY J. Z. *et al.* (2018). Health literacy and cardiovascular disease : fundamental relevance to primary and secondary prevention : a scientific statement from the american heart association. *Circulation*, **138**(2), e48–e74.

MARTIN L., MULLER B., ORTIZ SUÁREZ P. J., DUPONT Y., ROMARY L., DE LA CLERGERIE É., SEDDAH D. & SAGOT B. (2020). CamemBERT : a tasty French language model. In D. JURAFSKY, J. CHAI, N. SCHLUTER & J. TETREAULT, Éds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 7203–7219, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.645.

MC LAUGHLIN G. H. (1969). Smog grading-a new readability formula. *Journal of reading*, **12**(8), 639–646.

PROULX J., KANDULA S., HILL B. & ZENG-TREITLER Q. (2013). Creating consumer friendly health content : Implementing and testing a readability diagnosis and enhancement tool. In *2013 46th Hawaii International Conference on System Sciences*, p. 2445–2453 : IEEE.

QIU X., CHEN Y., CHEN H., NIE J.-Y., SHEN Y. & LU D. (2021). Learning syntactic dense embedding with correlation graph for automatic readability assessment. In C. ZONG, F. XIA, W. LI & R. NAVIGLI, Éds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 3013–3025, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.acl-long.235.

STEFANOU L., PASSALI T. & TSOUMAKAS G. (2024). AUTH at BioLaySumm 2024 : Bringing scientific content to kids. In D. DEMNER-FUSHMAN, S. ANANIADOU, M. MIWA, K. ROBERTS & J. TSUJII, Éds., *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, p. 793–803, Bangkok, Thailand : Association for Computational Linguistics. DOI : 10.18653/v1/2024.bionlp-1.73.

ULUSLU A. Y. (2023). Exploring hybrid linguistic features for Turkish text readability. In M. ABBAS & A. A. FREIHAT, Éds., *Proceedings of the 6th International Conference on Natural Language and Speech Processing (ICNLSP 2023)*, p. 223–232, Online : Association for Computational Linguistics.

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.

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.

XIA M., KOCHMAR E. & BRISCOE T. (2019). Text readability assessment for second language learners. *arXiv preprint arXiv :1906.07580*.

YANG Z., DAI Z., YANG Y., CARBONELL J., SALAKHUTDINOV R. R. & LE Q. V. (2019). Xlnet : Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, **32**.

ZHENG J. & YU H. (2017). Readability formulas and user perceptions of electronic health records difficulty : a corpus study. *Journal of medical Internet research*, **19**(3), e59.