Amélioration de la lisibilité de textes via l'utilisation de LLM

Baptiste Ramonda Isabelle Ferrane Julien Pinquier
IRIT, CNRS, UT, Toulouse, France
{Baptiste.Ramonda, Isabelle.Ferrane, Julien.Pinquier}@irit.fr

RÉSUMÉ

La lisibilité d'un texte est essentielle pour garantir un accès équitable à l'information. Cet article propose une méthodologie visant à simplifier des textes complexes tout en préservant leur sens. Un indice global de lisibilité a été défini en combinant plusieurs scores normalisés. Ensuite, une chaîne de traitement automatique, basée sur l'API de Gemini (LLM de Google), a généré des versions simplifiées des textes. Les résultats montrent une amélioration significative de la lisibilité, selon l'indice global et les critères spécifiques. Pour vérifier la conservation des idées clés, des résumés ont été extraits des versions initiales et simplifiées. Une mesure de la distance sémantique confirme que les concepts essentiels sont préservés. Cette approche prouve qu'il est possible d'automatiser efficacement la simplification textuelle tout en maintenant la cohérence et la pertinence des contenus, améliorant ainsi l'accessibilité de l'information.

ABSTRACT

Improving text readability using LLM.

Text readability is essential to ensure equitable access to information. This article proposes a methodology for simplifying complex texts while preserving their meaning. A global readability index was defined by combining several standardized scores. Then, a pipeline, based on the Gemini API (LLM from Google), generated simplified versions of the texts. The results show a significant improvement in readability, depending on the overall index and the specific criteria. To check the retention of key ideas, summaries were extracted from the initial and simplified versions. A measure of semantic distance confirms that essential concepts are preserved. This approach proves that it is possible to automate textual simplification efficiently while maintaining the coherence and relevance of content, thus improving the accessibility of information.

MOTS-CLÉS : Lisibilité de textes, Simplification de textes, LLM, Sémantique.

KEYWORDS: Readability, Text simplification, LLM, Semantics.

1 Introduction

La lisibilité, concept fondamental dans l'analyse textuelle, se définit comme la capacité à mesurer la difficulté d'un texte en s'appuyant sur divers critères linguistiques, tels que la longueur des phrases, la complexité lexicale et la structure syntaxique (Flesch, 1948). Bien que largement étudié, ce domaine demeure en perpétuelle transformation, car la langue et les pratiques rédactionnelles évoluent en réponse aux changements sociétaux et culturels.

Aujourd'hui, de nombreux indices de lisibilité sont reconnus comme des outils fiables pour évaluer la complexité d'un texte. Parmi les plus courants figurent le SMOG (Simple Measure of Gobbledy-

gook) (Mc Laughlin, 1969), l'indice de lisibilité de Flesch (Flesch, 1948), l'indice Dale-Chall (dans sa version initiale (Dale & Chall, 1948) et révisée (Chall & Dale, 1995)), l'indice Coleman-Liau (Coleman & Liau, 1975), l'indice Gunning Fog (Gunning, 1952), ainsi que l'Automated Readability Index (ARI) (Smith & Senter, 1967). Ces modèles offrent des évaluations globales et trouvent des applications dans de nombreux domaines tels que l'éducation (Dubay, 2004), la santé publique (Wang *et al.*, 2013) ou l'industrie (Du Toit, 2017).

Parallèlement, des indices spécialisés ont été développés pour répondre à des besoins contextuels spécifiques. Par exemple, l'indice Linsear Write (O'hayre, 1966) a été conçu pour évaluer la lisibilité des manuels techniques destinés à l'armée de l'air américaine, tandis que l'indice SPACHE (Spache, 1953) cible principalement les textes destinés aux jeunes lecteurs. Toutefois, l'étude que nous présentons ici se concentre davantage sur les indices généralistes évoqués, dans l'objectif de définir un indice global de lisibilité. L'utilisation d'indices polyvalents, applicables à plusieurs domaines, constitue une base solide essentielle pour atteindre cet objectif.

Cet article abordera dans un premier temps la création d'un indice global de lisibilité, puis la réalisation de deux corpus dans le cadre de notre étude. Grâce à ces corpus, composés de textes complexes et de leurs équivalents simplifiés, nous avons obtenu un ensemble de résultats relatifs à la lisibilité globale des textes et à la similarité sémantique entre version complexe et simplifiée. L'objectif de cette étude est d'évaluer si la simplification de textes complexes améliore leur lisibilité tout en préservant une sémantique proche, c'est-à-dire en conservant leur sens et leurs concepts principaux.

2 Indices de lisibilité

2.1 Principaux indices de lisibilité

Cette section présente les principaux indices de lisibilité généralistes mentionnés dans l'introduction, avant de présenter la construction d'un indice global de lisibilité normalisé entre 0 et 10 utilisant ces mêmes indices.

Indice Flesch Reading Ease Développé par Rudolf Flesch (Flesch, 1948), cet indice évalue la lisibilité d'un texte à l'aide de la formule suivante :

$$\text{Score}_{\text{IFRE}} = 206,835 - 1,015 \times \overline{M_{phrase}} - 84,6 \times \overline{S_{mot}} \tag{1}$$

où :
- $\overline{M_{phrase}}$ est la longueur moyenne des phrases (nombre moyen de mots par phrase),
- $\overline{S_{mot}}$ est le nombre moyen de syllabes par mot.

Un score élevé indique une meilleure lisibilité. Cet indice est couramment utilisé pour évaluer des documents destinés à un large public. Cependant, il tend à sous-estimer la complexité syntaxique et ne prend pas en compte l'aspect sémantique.

SMOG (Simple Measure of Gobbledygook) Le SMOG (Mc Laughlin, 1969) est utilisé pour estimer le niveau scolaire requis afin de comprendre un texte. Il se base sur le comptage des mots de trois syllabes ou plus dans un sous-ensemble de phrases issues du texte.

Sa formule est donnée par :

$$\text{Score}_{\text{SMOG}} = 1,0430 \times \sqrt{30 \times \frac{P}{N}} + 3,1291 \tag{2}$$

où :

— P est le nombre de mots polysyllabiques (mots de trois syllabes ou plus),
— N est le nombre de phrases dans le sous-ensemble considéré.

Cet indice est principalement employé en santé publique pour évaluer si un texte est compréhensible par un large public. Cependant, il peut surestimer la difficulté des textes contenant un vocabulaire technique couramment utilisé dans certains domaines.

Indice Dale-Chall Cet indice repose sur une liste de mots considérés comme familiers (Dale & Chall, 1948) pour la plupart des enfants américains de 4ème année, ce qui équivaut à un niveau primaire dans le système scolaire français. La présence de mots hors de cette liste augmente la difficulté du texte. Sa formule est la suivante :

$$\text{Score}_{\text{IDC}} = 0,1579 \times H + 0,0496 \times \overline{M_{phrase}} \tag{3}$$

où :

— H est le pourcentage de mots difficiles,
— $\overline{M_{phrase}}$ est la longueur moyenne des phrases (nombre moyen de mots par phrase).

Cet indice est particulièrement adapté à l'éducation et permet d'identifier les textes potentiellement trop complexes pour un jeune public. Toutefois, sa pertinence est réduite pour les textes destinés aux adultes ou aux spécialistes. Il est important de noter que la liste des mots considérés comme familiers a été mise à jour dans une nouvelle version de l'indice Dale Chall (Chall & Dale, 1995). C'est d'ailleurs cette liste que nous utiliserons dans la suite de cette étude.

Indice Coleman-Liau Cet indice fait intervenir d'autres niveaux de granularité, car il repose sur le nombre moyen de lettres par mot et le nombre moyen de phrases par texte (Coleman & Liau, 1975). Il est défini par :

$$\text{Score}_{\text{ICL}} = 0,0588 \times \overline{L_{100}} - 0,296 \times \overline{N_{100}} - 15,8 \tag{4}$$

où :

— $\overline{L_{100}}$ est le nombre moyen de lettres pour 100 mots,
— $\overline{N_{100}}$ est le nombre moyen de phrases pour 100 mots.

Sa méthode de calcul, facilement automatisable, car ne nécessitant pas un comptage des syllabes, le rend particulièrement utile pour l'analyse de grands corpus numériques. Toutefois, l'absence de prise en compte des syllabes peut biaiser l'évaluation de la difficulté phonologique des mots.

Indice Gunning Fog Couramment utilisé dans la presse et le milieu professionnel, cet indice mesure la complexité d'un texte en prenant en compte la longueur des phrases et la proportion de mots complexes (Gunning, 1952) :

$$\text{Score}_{\text{IGF}} = 0,4 \times \left(\overline{M_{phrase}} + 100 \times \frac{C}{N} \right) \tag{5}$$

où :

— $\overline{M_{phrase}}$ est la longueur moyenne des phrases (nombre moyen de mots par phrase),
— C est le nombre de mots complexes (au moins trois syllabes),
— N est le nombre total de mots.

Cet indice est utilisé pour ajuster la lisibilité des documents destinés au grand public. Cependant, il peut surestimer la difficulté des textes contenant un vocabulaire spécialisé, mais courant pour une audience cible.

Automated Readability Index (ARI) Tout comme l'indice Coleman-Liau, l'ARI (Smith & Senter, 1967) repose sur le nombre moyen de lettres par mot et le nombre moyen de mots par phrase. Il est défini par la formule suivante :

$$\text{Score}_{\text{ARI}} = 4,71 \times \overline{L_{mot}} + 0,5 \times \overline{M_{phrase}} - 21,43 \tag{6}$$

où :
— $\overline{L_{mot}}$ est le nombre moyen de lettres par mot,
— $\overline{M_{phrase}}$ est la longueur moyenne des phrases (nombre moyen de mots par phrase).

Cet indice est fréquemment utilisé dans l'industrie pour classifier rapidement des documents. Cependant, il présente les mêmes biais que l'indice Coleman-Liau, notamment en raison de la non-prise en compte de la structure syllabique des mots.

2.2 Création d'un indice de lisibilité global

Chacun des indices de lisibilité présentés précédemment fournit une évaluation partielle de la complexité d'un texte en se basant sur des critères spécifiques et des niveaux de granularité différents. Toutefois, ces critères individuels peuvent introduire des biais méthodologiques :
— Flesch favorise de courts textes simples sans considération pour la structure syntaxique,
— SMOG et le Gunning Fog surestiment la difficulté des textes contenant des termes techniques,
— Dale-Chall repose sur une liste de mots préétablie, limitant ainsi son applicabilité,
— Coleman-Liau et ARI, basés uniquement sur le nombre de lettres par mot, ne prennent pas en compte les variations phonétiques et syllabiques.

Afin d'atténuer ces limitations et obtenir une évaluation plus robuste de la lisibilité, nous proposons la construction d'un Indice Global de Lisibilité (IGL). Cet indice est élaboré à partir d'une combinaison normalisée des scores des différents indices mentionnés précédemment. La normalisation est réalisée via une transformation linéaire, définie comme suit :

$$IGL = \frac{1}{N_i} \sum_{i=1}^{N_i} \frac{I_i - I_{i,\min}}{I_{i,\max} - I_{i,\min}} \times 10 \tag{7}$$

où :
— I_i représente le score de l'indice i,
— $I_{i,\min}$ et $I_{i,\max}$ correspondent respectivement aux valeurs minimales et maximales théoriques de l'indice i ,
— N_i désigne le nombre total d'indices intégrés.

Grâce à cette fusion, plusieurs biais peuvent être atténués. Une meilleure représentativité de la complexité textuelle est assurée par l'intégration d'indices prenant en compte le nombre de syllabes

(Flesch, SMOG, Gunning Fog) et ceux basés sur la longueur lexicale (Coleman-Liau, ARI). Une dimension cognitive liée à la familiarité lexicale est introduite par l'inclusion de l'indice Dale-Chall. De plus, la moyenne arithmétique appliquée permet de lisser les valeurs extrêmes et de réduire l'impact des spécificités propres à chaque indice. Nous faisons donc l'hypothèse dans cette étude que l'IGL constitue une métrique synthétique et équilibrée, mieux adaptée à l'analyse automatique de la lisibilité textuelle globale.

3 Corpus : création et simplification

Un défi rencontré dans le cadre de cette étude a été l'absence de corpus existants de textes annotés comme « difficiles ». Pour répondre à cette limitation, un corpus original de 100 textes complexes a été créé à l'aide d'une chaîne de traitement développée en s'appuyant sur le grand modèle de langage (Large Language Model, LLM) Gemini de Google[1]. Le modèle utilisé dans cette étude est Gemini 1.5 Flash[2], choisi pour sa capacité à générer des textes complexes avec une grande efficacité, tout en restant accessible gratuitement pour un volume de requêtes quotidien suffisant dans le cadre de nos expérimentations. Les principaux paramètres du modèle sont configurés ainsi : *température : [0–2], top-p : 0,95, top-k : 64, candidate-count : [1–8].*

Il est important de rappeler que dans cette étude, le concept de simplification désigne un processus de transformation automatique visant à rendre un texte plus accessible sur les plans lexical, syntaxique et discursif, tout en préservant son sens. Le LLM est ainsi sollicité pour reformuler les textes en abaissant les obstacles linguistiques, selon une logique d'adaptation plutôt que de réécriture libre.

La génération des textes complexes s'est appuyée sur une liste de 50 thématiques définies comme à la fois complexes et spécifiques, tout en étant suffisamment larges pour permettre la création de textes distincts, et ce, même si le même sujet est sélectionné plusieurs fois. Cette stratégie a été mise en œuvre afin de réduire les biais potentiels du modèle, qui, lorsqu'il est sollicité pour générer des textes complexes et précis sans contrainte de thématique, a tendance à produire des textes focalisés sur un nombre très limité de thèmes récurrents.

Le prompt utilisé pour générer les textes complexes via l'API Gemini est le suivant :

> *"Write a complex and specific text about the topic : {topic}. Use precise and advanced vocabulary. Ensure the text is detailed and incorporates technical or academic terms."*

L'usage de l'anglais pour les prompts, la génération et le traitement des textes par le LLM s'impose par sa prédominance dans le développement des modèles de langage, garantissant un accès optimal aux ressources, une meilleure performance et une diffusion plus large des résultats. La liste complète des sujets pouvant être tirés au sort par le modèle est disponible à l'adresse suivante : liste des 50 thématiques.

Une fois ces 100 textes complexes créés, il fallait les simplifier pour obtenir la deuxième partie du corpus permettant de réaliser les études de lisibilité et de sémantique. La simplification passe là encore par la chaîne de traitement créée, mais cette fois-ci plusieurs variantes de prompt ont été tentées, avec des résultats radicalement différents.

1. https://gemini.google.com/
2. https://ai.google.dev/gemini-api/docs/models/gemini

3.1 Méthode 1 : simplification par itérations successives

La première méthode consiste à simplifier le texte par itérations successives. L'objectif est de donner un prompt simple à l'API, tel que :

> *"Simplify this text while retaining the main concepts : {text}"*

Puis, il s'agit d'étudier l'évolution de la lisibilité à chaque itération pour identifier à partir de quel moment les textes deviennent véritablement plus accessibles. Un extrait de texte complexe généré et son équivalent après simplification en 5 itérations avec la méthode 1 sont donnés ci-après :

> Texte complexe : *"The remarkable phenomenon of superconductivity, characterized by zero electrical resistivity and the Meissner effect (complete expulsion of magnetic fields), arises from a complex interplay of electron-phonon interactions, electron correlations, and band structure effects within a material below a critical temperature, Tc. Understanding the physics necessitates a departure from the simplistic Drude model and a dive into the realm of many-body quantum mechanics."*
>
> Texte simplifié selon la méthode 1 : *"Superconductivity happens when a material loses electrical resistance and repels magnetic fields (Meissner effect) below a critical temperature, Tc, due to quantum interactions between electrons and atomic vibrations (phonons). Cooper pairs—electron pairs bound by phonons—play a key role by overcoming repulsion. This interaction and the material's structure determine Tc."*

3.2 Méthode 2 : simplification pour de jeunes enfants

Une seconde approche consiste à demander à l'API Gemini de simplifier le texte en l'expliquant comme s'il s'adressait à un jeune enfant.

Le prompt utilisé pour la simplification des 100 textes complexes du corpus est le suivant :

> *"Simplify the following text as if explaining it to a young child, keeping the main concepts and important information : {ComplexText}"*

En utilisant cette formulation, il est intéressant d'observer que la quasi-totalité des textes simplifiés respecte la même structure, et les concepts difficiles sont désormais expliqués via des métaphores ou des représentations imagées pouvant parler à un enfant. Voici un extrait du même texte complexe cité précédemment, mais simplifié avec la méthode 2 :

> Texte simplifié selon la méthode 2 :
> *"Imagine electricity flowing through a wire. Usually, it bumps into things and slows down, like running through a crowded room. But some special materials, when they're super cold, become superconductors ! The electricity flows perfectly, without bumping into anything, like running in an empty hallway. That's because tiny particles called electrons team up in pairs, and they move together smoothly."*

La Table 1 présente une analyse quantitative des deux corpus obtenus à l'aide des méthodes décrites précédemment. Les caractéristiques mesurées montrent une grande similarité entre les deux corpus, offrant ainsi une base solide et comparable pour les analyses de lisibilité et de sémantique, développées dans la suite de cet article.

TABLE 1 – Résumé de notre corpus.

	Méthode 1	Méthode 2
Nombre total de fichiers	200	200
Nombre total de mots	50 185	65 896
Nombre total de phrases	2 256	3 406
Moyenne de caractères par mot	6,7	6,1
Moyenne de mots par phrase	22,2	19,3

4 Expérimentations

La section suivante décrit les expérimentations réalisées afin d'évaluer l'efficacité de notre approche en termes de lisibilité et de conservation des aspects sémantiques. Nous présentons tout d'abord une analyse quantitative de l'évolution de la lisibilité après simplification, avant d'examiner la conservation des concepts essentiels à travers une étude sémantique.

4.1 Étude de lisibilité

Une fois le corpus constitué, l'étape suivante a consisté à analyser l'évolution de la lisibilité entre les textes complexes générés dans un premier temps et leurs versions simplifiées. Cette analyse repose sur l'IGL défini dans la section 2.2. Comme évoqué précédemment, cette approche permet d'atténuer les biais propres à chaque indice pris séparément et d'offrir ainsi une évaluation plus équilibrée et représentative de la difficulté perçue.

Les résultats de cette expérimentation, obtenus après calcul systématique de l'indice global de lisibilité pour chaque texte sont présentés dans les figures 1 et 2.

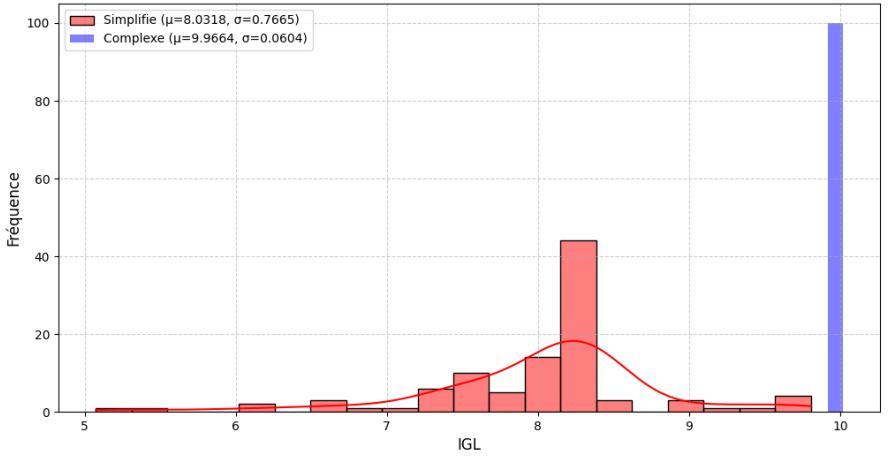

FIGURE 1 – Indice global de lisibilité avec la méthode 1 : 100 textes simplifiés sur cinq itérations.

Le premier graphique illustre la méthode 1, consistant à simplifier 100 textes complexes sur cinq itérations. L'itération initiale (en bleu) affiche un score constant de 10, correspondant au texte généré pour avoir une complexité élevée, considérée comme maximale. Le résultat au bout de la 5ème itération de simplification (en rouge sur le graphe) montre une réduction des scores, avec une moyenne autour de 8. Cependant, ces scores restent relativement élevés, indiquant une lisibilité encore limitée et rendant les textes peu accessibles à un large public. Le choix de cinq itérations repose sur une convergence des scores et la stabilité des résultats observés dès ce stade, indiquant que des itérations supplémentaires n'auraient eu qu'un impact marginal sur la simplification.

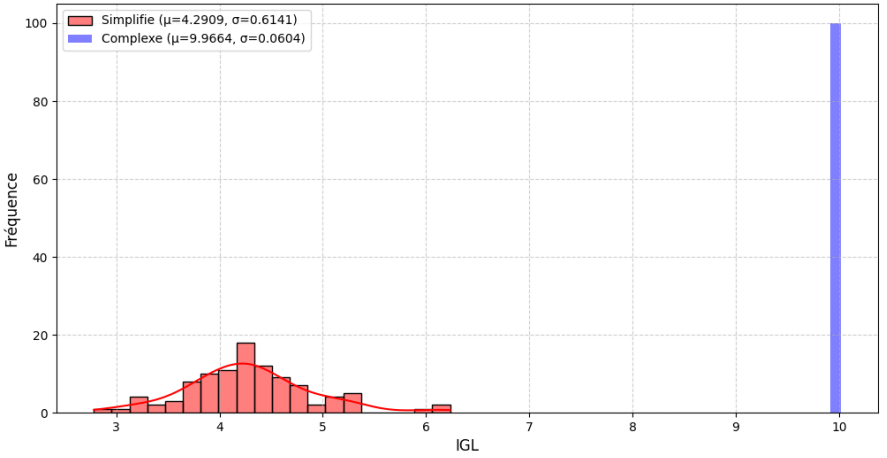

FIGURE 2 – Indice global de lisibilité avec la méthode 2 : 100 textes simplifiés sur une itération

Sur le second graphique, on retrouve le même sous-ensemble du corpus utilisé pour la méthode 1, composé de 100 textes jugés très difficiles, comme en témoigne leur score maximal de difficulté (valeur 10). Ce choix de 100 textes vise à trouver un équilibre entre un nombre suffisant de fichiers et un temps de calcul raisonnable, l'objectif étant ici davantage de démontrer un concept que d'entraîner réellement un modèle sur un grand jeu de données. Après une première itération du modèle visant à simplifier ces textes pour les rendre compréhensibles par de jeunes enfants, leur niveau de difficulté diminue de manière significative. La note moyenne chute alors à un peu plus de 4, soit une réduction de plus de 50% par rapport au score initial.

L'expérimentation visait à comparer l'efficacité des méthodes de simplification. La méthode 1 réduit progressivement la difficulté des textes, passant de 10 à environ 8 après cinq itérations, tandis que la méthode 2 atteint une réduction de plus de 50% en une seule itération. Cette dernière s'avère donc plus efficace pour améliorer rapidement et significativement la lisibilité. De plus, la méthode 1 nécessitant plusieurs itérations, elle engendre un coût de calcul et un temps de traitement plus élevés, alors que la méthode 2 offre un gain immédiat en lisibilité dès la première application.

La matrice de corrélation (voir figure 3) confirme la pertinence des indices de lisibilité utilisés. Elle met en évidence des corrélations très fortes entre les différentes métriques, avec des coefficients très proches de 1. Même si certaines paires d'indices sont représentées en bleu, indiquant une corrélation légèrement plus faible que les autres, elles restent très élevées (> 0,98). Cette forte cohérence entre

les indices justifie leur utilisation conjointe pour évaluer l'impact des méthodes de simplification et renforce le choix de la méthode 2 comme approche privilégiée. Bien que fortement corrélés, ces indices capturent des aspects complémentaires de la lisibilité, permettant ainsi une analyse plus fine et robuste des effets des différentes approches.

Il convient cependant de s'assurer que la simplification ne se fait pas au détriment du contenu. Les concepts clés et les idées principales véhiculés par le texte original doivent être préservés dans la version simplifiée afin d'éviter toute altération du discours initial.

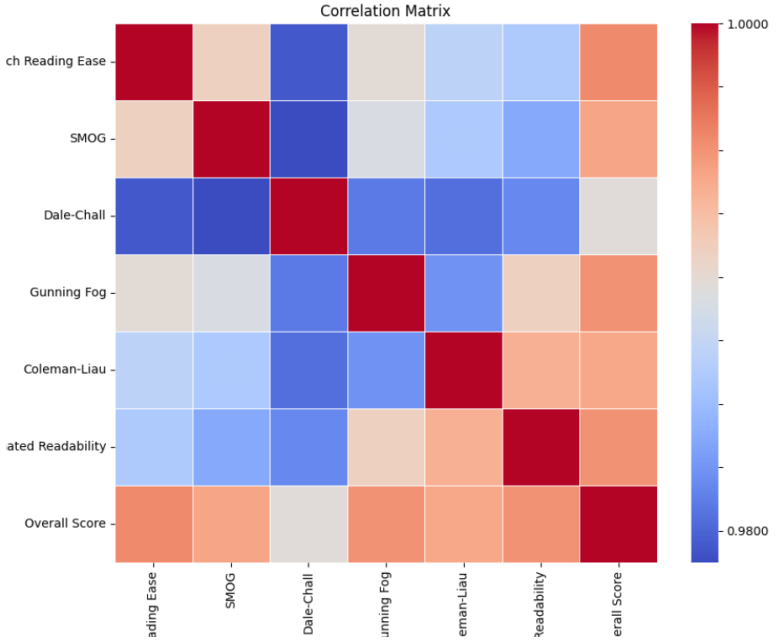

FIGURE 3 – Matrice de corrélation entre les différents indices de lisibilité avec la méthode 2.

4.2 Étude sémantique

Pour assurer la conservation des aspects sémantiques entre un texte complexe et sa version simplifiée, l'API Gemini a été utilisée pour extraire, à partir de chaque fichier, un résumé de trois phrases représentant les concepts et sujets principaux abordés. Le prompt précis envoyé à l'API était le suivant :

> *"Summarize the following text in 3 sentences keeping the main concepts and ideas from it : {text}"*

Les résumés extraits ont ensuite été convertis en vecteurs denses (Mikolov *et al.*, 2013) à l'aide des représentations contextuelles générées par le modèle Sentence-BERT (SBERT, *"all-mpnet-base-v2"*), une variante optimisée de BERT pour la représentation de phrases (Reimers & Gurevych, 2019)[3].

3. https://sbert.net/

Chaque résumé a été segmenté en phrases, puis ces dernières ont été encodées en représentations vectorielles. Cette étape permet de capturer les relations sémantiques entre les phrases, facilitant ainsi les analyses ultérieures, telles que la comparaison de proximité sémantique entre les résumés ou la détection de similarités thématiques.

Pour évaluer la proximité sémantique entre les résumés des textes complexes et ceux des textes simplifiés, une mesure de similarité cosinus a été appliquée (Salton *et al.*, 1975). Cette métrique calcule le cosinus de l'angle entre deux vecteurs dans un espace vectoriel, fournissant ainsi une mesure normalisée de la similarité, avec des valeurs comprises entre -1 et 1 (1 indiquant une similarité maximale). Concrètement, pour chaque segment issu du texte complexe, le système identifie le segment le plus proche dans le texte simplifié en fonction de cette mesure. Cela permet d'établir un appariement optimal entre les deux résumés, mettant en évidence les correspondances sémantiques les plus fortes. Cependant, il est important de noter qu'en pratique, les valeurs négatives sont très rares. En théorie, une similarité cosinus de -1 indiquerait que deux segments ont des significations opposées. Pourtant, si nous testons ce phénomène sur des exemples simples comme *"I am happy."* et *"I am sad."*, nous obtenons des scores avoisinants 0,3 ou 0,4, plutôt que des valeurs négatives. Cette divergence s'explique principalement par un biais contextuel : bien que les phrases expriment des émotions opposées, elles appartiennent au même champ lexical et apparaissent souvent dans des contextes similaires. Le modèle, influencé par cette proximité contextuelle, perçoit alors une certaine similarité, même lorsque le sens des phrases est contrasté.

Afin d'évaluer la robustesse du modèle et la pertinence des résultats obtenus, deux séries d'expérimentations ont été menées. Dans un premier temps, la similarité cosinus a été calculée pour des paires dites valides, définies comme les correspondances entre le résumé d'un texte complexe et celui du même texte simplifié. Dans un second temps, cette mesure a été appliquée à des paires de résumés sélectionnées aléatoirement afin de constituer un ensemble de contrôle. Étant donné l'absence de lien sémantique attendu entre ces paires, des scores de similarité faibles sont attendus, permettant ainsi de vérifier que le modèle discrimine correctement les relations sémantiques pertinentes.

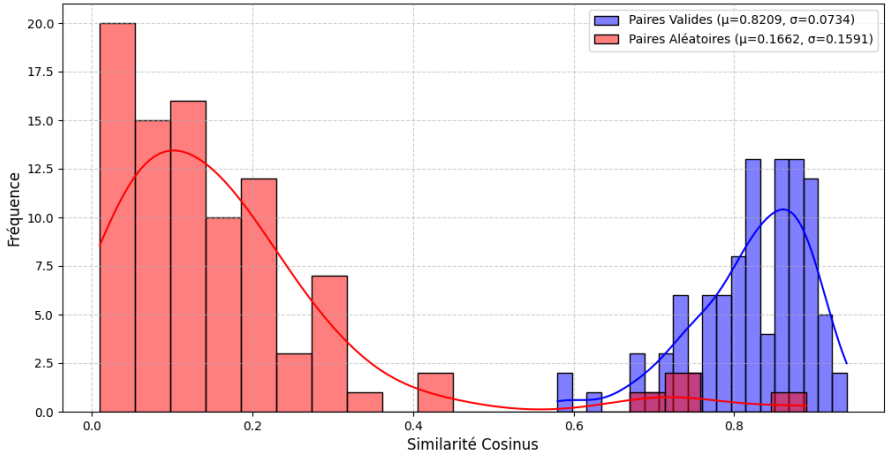

FIGURE 4 – Distances sémantiques entre les différents résumés de nos extraits.

La figure 4 illustre les résultats de l'évaluation de la similarité cosinus entre des résumés simplifiés et leurs versions complexes (PAIRED) ainsi qu'entre les paires de résumés sélectionnées aléatoirement (RANDOM). Les résultats mettent en évidence une similarité significativement plus élevée pour les paires associées ($\mu = 0,82$, $\sigma = 0,07$), traduisant une forte conservation de l'information sémantique entre les textes originaux et leurs versions simplifiées. À l'inverse, les paires aléatoires présentent une similarité beaucoup plus faible ($\mu = 0,15$, $\sigma = 0,16$), confirmant l'absence de relation sémantique entre des textes sans lien explicite. L'écart marqué entre ces distributions souligne l'efficacité du processus de simplification dans le maintien de la cohérence sémantique, tout en différenciant clairement les résumés des paires de textes indépendants.

5 Discussions

Ce travail s'inscrit dans le cadre d'un projet portant sur le concept d'écoutabilité, défini comme la capacité d'un discours oral à transmettre un maximum d'informations tout en minimisant la charge cognitive de l'auditeur (Cartier, 1952), c'est-à-dire en évitant qu'il ne décroche du discours. L'écoutabilité prend en compte de nombreuses composantes du discours, notamment : l'aspect psychologique, avec l'émotion suscitée et la charge cognitive ; l'aspect non verbal, incluant les gestes et les expressions faciales ; l'aspect socioculturel, où des facteurs tels que l'âge influencent la compréhension du discours ; l'aspect textuel, qui concerne la structure même du texte et le choix des mots ; et enfin, l'aspect oral, qui englobe la prosodie, l'intelligibilité, et d'autres éléments liés à la transmission vocale. Les travaux présentés dans cet article se concentrent spécifiquement sur la composante textuelle du discours, en explorant comment simplifier un texte afin de répondre aux objectifs de clarté et d'efficacité évoqués précédemment.

Il est toutefois important de souligner que les résultats obtenus dans cette étude sont principalement valables dans le cadre textuel, et qu'une transposition vers l'oral pourrait introduire de nouveaux paramètres (Harwood, 1955), susceptibles de modifier les conclusions obtenues. En effet, la dimension auditive du discours pourrait impacter la manière dont l'information est perçue et traitée par l'auditeur (Glenn *et al.*, 1995). Malgré cette limitation, cette première étude constitue une base solide pour nos recherches futures sur le concept d'écoutabilité.

Par ailleurs, plusieurs limites méthodologiques doivent être soulignées. Le pipeline proposé repose exclusivement sur le modèle Gemini, utilisé à la fois pour la génération initiale des textes et pour leur simplification. Cette dépendance à un unique LLM soulève plusieurs questions, notamment des biais potentiels liés à des phénomènes d'hallucination, à une homogénéité stylistique marquée, ou encore à une faible diversité dans les stratégies de réécriture mises en œuvre. L'utilisation répétée du même modèle pour différentes étapes du processus peut également engendrer des effets d'auto-renforcement ou de biais circulaire, dans la mesure où le système a tendance à reproduire et valider ses propres choix linguistiques sans confrontation externe. À terme, une évaluation comparative incluant d'autres modèles, en particulier open source (tels que LLaMa[4] ou Mistral[5]), apparaît nécessaire pour accroître la robustesse et la reproductibilité des résultats. Le recours à des modèles open source offrirait en outre un meilleur contrôle sur les paramètres de génération, une traçabilité accrue des décisions algorithmiques, et une personnalisation plus fine du processus de simplification.

En complément, une amélioration notable du pipeline consisterait à introduire une évaluation humaine,

4. https://www.llama.com/
5. https://mistral.ai/fr

afin de valider la lisibilité perçue et la préservation du sens par des lecteurs réels. Cette validation externe permettrait de confronter l'indice global de lisibilité (IGL) proposé à des jugements qualitatifs. Une analyse fine des erreurs serait également précieuse pour ajuster les prompts de simplification.

D'autres pistes d'amélioration incluent :

- l'intégration de métriques sémantiques plus fines (telles que BERTScore ou BLEURT), susceptibles de mieux capturer les nuances sémantiques et pragmatiques au-delà de la similarité vectorielle brute ;
- une analyse qualitative des transformations linguistiques opérées par les LLM (par exemple : substitution lexicale, simplification syntaxique, suppression d'informations redondantes) pour mieux comprendre les mécanismes de simplification ;
- l'exploration de stratégies de prompts plus variées, voire dynamiquement optimisées selon le type de texte ou la population cible ;
- l'application du pipeline à des corpus réels issus de contextes variés (textes journalistiques, administratifs, pédagogiques), plutôt qu'à des textes artificiellement générés ;
- l'adaptation de la pondération des différents constituants de l'IGL en fonction de leur impact sur le texte traité.

6 Conclusion

Cette étude vise à évaluer l'efficacité d'une approche automatisée basée sur de grands modèles de langage (LLM) pour simplifier des textes complexes, tout en en préservant le sens. En combinant un indice global de lisibilité avec une analyse de la distance sémantique, nous avons démontré que notre méthode réduit significativement la difficulté de lecture tout en maintenant une forte cohérence sémantique entre les textes d'origine et leurs versions simplifiées. Une validation complémentaire sur des textes réels, non générés par des LLM, permettrait d'en évaluer la robustesse.

Ces travaux ouvrent la voie à de nombreuses applications concrètes, notamment dans l'éducation, en facilitant l'accès à des contenus pédagogiques pour les élèves en difficulté ou les personnes apprenant une nouvelle langue. Ils peuvent également profiter aux personnes en situation de handicap cognitif en rendant l'information plus accessible, ainsi qu'aux institutions et entreprises cherchant à simplifier des documents administratifs ou juridiques pour un public non averti.

Références

CARTIER F. A. (1952). The social context of listenability research. *Journal of Communication*, **2**(1), 44–47.

CHALL J. S. & DALE E. (1995). Readability revisited : The new dale-chall readability formula. *(No Title)*.

COLEMAN M. & LIAU T. L. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, **60**(2), 283.

DALE E. & CHALL J. S. (1948). A formula for predicting readability : Instructions. *Educational research bulletin*, p. 37–54.

DU TOIT E. (2017). The readability of integrated reports. *Meditari Accountancy Research*, **25**(4), 629–653.

DUBAY W. (2004). The principles of readability. *CA*, **92627949**, 631–3309.

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

GLENN E. C., EMMERT P. & EMMERT V. (1995). A scale for measuring listenability : The factors that determine listening ease and difficulty. *International journal of listening*, **9**(1), 44–61.

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

HARWOOD K. A. (1955). I. listenability and readability. *Communications Monographs*, **22**(1), 49–53.

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

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv :1301.3781*.

O'HAYRE J. (1966). *Gobbledygook has gotta go*. US Department of the Interior, Bureau of Land Management.

REIMERS N. & GUREVYCH I. (2019). Sentence-bert : Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv :1908.10084*.

SALTON G., WONG A. & YANG C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, **18**(11), 613–620.

SMITH E. A. & SENTER R. (1967). Automated readability index. *Aerospace Medical Research Laboratories, Aerospace Medical Division, Air*, p. 1–14.

SPACHE G. (1953). A new readability formula for primary-grade reading materials. *The Elementary School Journal*, **53**(7), 410–413.

WANG L.-W., MILLER M. J., SCHMITT M. R. & WEN F. K. (2013). Assessing readability formula differences with written health information materials : application, results, and recommendations. *Research in Social and Administrative Pharmacy*, **9**(5), 503–516.