Prédiction de recommandations d'âge pour l'accès à des enfants à des textes

Alexis Blandin, Gwénolé Lécorvé, Delphine Battistelli

Rapport de recherche

Prédiction de recommandations d'âge pour l'accès à des enfants à des textes

Alexis Blandin[1], Gwénolé Lecorvé[1], Delphine Battistelli[2]

[1] Univ Rennes, CNRS, IRISA
[2] Université Paris-Nanterre, CNRS, MoDyCo

Travail réalisé avec le soutien du projet ANR TREMoLo (ANR-16-CE23-0019).

Abstract: La compréhension d'un texte par un individu est conditionnée par l'adéquation des caractéristiques de ce texte par rapport aux capacités et aux connaissances de l'individu. Dans le cas d'un enfant, il est donc intéressant de déterminer en quoi son âge influe sur sa compréhension d'un texte. Des travaux psycholinguistiques ont étudié ce problème de près afin d'établir dans quelles mesures un texte serait, ou non, destiné à un enfant. En parallèle à cela, les avancées en traitement automatique des langues offrent de nouvelles possibilités pour étudier les informations issues de textes . Ce rapport présente donc une manière d'utiliser ces techniques pour déterminer une recommandation d'âge pour un texte destiné à des enfants.

Contents

1 Introduction

La manière dont un individu comprend un texte est mal connue car elle fait appel aussi bien à des caractéristiques du dit texte qu'aux capacités du lecteur. Le problème se pose d'autant plus pour un enfant, pour qui les capacités de lecture évoluent à mesure qu'il grandit. C'est dans cette optique que nous présentons notre travail, dont le but est d'automatiser des recommandations d'âges, pour des textes pour enfants.
Pour cela nous avons dans un premier temps cherché plusieurs critères psycholinguistiques exploitables pour une telle prédiction. Puis nous avons modélisé ces critères à l'aide de techniques de traitement automatique des langues (TAL), afin d'extraire des textes plusieurs informations. Enfin, nous avons appris différents modèles prédictifs sur ces jeux de données, avec pour objectif de prédire l'âge associé à ces textes.
Ces prédictions pourraient alors servir en tant qu'indicateurs pour des auteurs ou des éditeurs qui cherchent à cibler un public d'une certaine tranche d'âge, et les différentes performances en fonction des critères peuvent aussi offrir un retour à des travaux en psycholinguistique.
Ainsi, nous allons dans un premier temps décrire l'état de l'art sur le sujet (section 2), à travers les travaux en psycholinguistique. Puis, à partir de ses travaux nous présenterons les différents critères d'apprentissage que nous avons retenus pour notre modèle (section 3). Ensuite nous verrons de quelle manière nous avons traité les données à notre disposition (section 4), et enfin nous présenterons les architectures de ces modèles, ainsi que les résultats obtenus (section 5).

2 État de l'art

Le sujet étudié ici étant assez neuf, il faut au préalable explorer les pistes possibles pour entamer ce travail. Pour cela, cet état de l'art présente les pistes offertes par la psycholinguistique pour modéliser l'évolution de la compréhension d'un texte par un enfant, ainsi que les différents moyens de calculer la lisibilité via des formules de lisibilité, et enfin envisager les techniques de linguistique computationelle sur des problèmes similaires au nôtre.

2.1 Travaux psycholinguistiques

Un enfant ne perçoit pas le monde de la même manière qu'un adulte et cela pour plusieurs raisons. D'une part son cerveau est encore en cours de développement et, d'autre part, il est en plein processus d'apprentissage de sa langue maternelle. C'est pour cela que nous aborderons ici la question de la gestion par l'enfant de sa mémoire à court terme, très sollicitée lors des activités de langage et de l'acquisition des repères temporels dans la langue, qui donne lieu parfois à des interprétations erronées ou naïves de l'enfant. Enfin nous verrons comment ces considérations peuvent être mises en application.

2.1.1 Apport de la psychologie du développement

La mémoire à court terme se développe fortement entre deux et huit ans. Cette mémoire influe principalement sur la compréhension du langage ou sur la réalisation de tâches complexes. En effet, selon Susan E. Gathercole [12], la mémoire qui est responsable de la mémorisation des informations verbales est la mémoire phonologique à court terme. Cette mémoire est située dans l'hémisphère gauche, et bien que son fonctionnement soit bien connu chez l'adulte, il est plus difficile de le comprendre chez l'enfant. De plus, le cerveau d'un enfant croit très rapidement dans les premières années de vie, cela influe donc aussi sur le développement de cette zone de mémoire. C'est vers l'âge de quatre ans que l'activité du cerveau atteint son apogée, avec une activité équivalente à 150% de celle d'un adulte [12]. Or cette période coïncide avec une étape clé de l'acquisition du langage. Cette mémoire phonologique à court terme intervient dans plusieurs aspects : le stockage des informations acoustiques, l'analyse et la mémorisation des informations phonologiques, l'ordonnancement temporel du langage, le mécanisme de répétition d'un mot pour conserver dans cette mémoire un mot pendant une longue période, la récupération d'informations mémorisées et enfin l'acquisition des liens avec les informations sémantiques ou syntaxiques associées au mot entendu.

Ces différentes capacités se développent à mesure que l'enfant grandit. Les aspects phonologiques ou encore la longueur d'une phrase semblent donc êtres des critères intéressants à étudier, car ces éléments font directement intervenir la mémoire à court terme phonologique. L'immaturité du cerveau de l'enfant a aussi d'autre conséquences. En particulier, Valérie Tartas [20] présente comment l'enfant parvient à acquérir les notions temporelles. Le manque de maîtrise de ces notions, qui permettent de se repérer soi-même dans le temps, ainsi que d'ordonner des événements dans le temps, se ressent alors dans le langage de l'enfant. L'usage des temps verbaux ou la mention des heures et minutes pour repérer une action dans le temps ne s'acquièrent pas simultanément, et il alors possible d'observer quels âges sont liés à quelle utilisation de ces marqueurs temporels. La figure 1 [20] présente comment un enfant manipule ces marqueurs durant trois phases de son développement.

Périodes de l'enfance	Construits temporels	Exemples de conduites temporelles
Bébé (0-1 an 1/2)	Perception des rythmes	Activités rythmiques (succion…)
	1ères attitudes temporelles Permanence de l'objet / de soi Représentation d'événements du temps proche : début de l'ordre temporel	Gestes temporels : attendre, désirer… Imitation différée de séquences d'actions familières
Petite enfance (1an 1/2 -4/5 ans)	Développement des marques temporelles dans le langage	Présent, passé, futur Adverbes temporels
	Soi étendu temporellement *Script :* ordre temporel d'événements du temps proche au temps lointain	Reconnaissance de soi à différents âges, début des récits autobiographiques Petites narrations quotidiennes coconstruites puis élaborées par l'enfant Systèmes de repérage avant/après : les événements comme repères pour d'autres Planifier des actions
Enfance (5/6 ans- 10/11 ans)	Temps conventionnel : ordre, récurrence	Les jours, les heures et les mois comme repères
	Calculs temporels : durée, vitesse Temps historique	Comparer des durées, des âges, résoudre des problèmes de vitesse, de durée de trajet Situations d'enseignement-apprentissage de l'histoire : apprendre de nouvelles unités : siècles Comparer le présent et le passé lointain pour le comprendre Début de compréhension du temps comme une construction humaine…

Figure 1: Tableau présentant les différentes phases de l'acquisition des marqueurs temporels chez l'enfant [20]

Ces informations sur le développement de l'enfant, ne nous renseignent que sur la manière dont un enfant s'exprime, et non sur son aptitude à comprendre un texte. Cependant, on peut garder à l'esprit pour la suite, que l'usage de temps verbaux complexes ou de connecteurs logiques dans un texte est d'autant plus faible que l'enfant est jeune [22]. Mesurer le nombre ou la densité de ces connecteurs dans un texte pourrait être alors un indicateur intéressant pour déterminer un âge auquel ce texte est destiné.

2.1.2 Modèle d'apprentissage de la lecture

Si la modélisation psycholinguistique de l'apprentissage de la lecture chez l'enfant est une bonne base pour pouvoir envisager une modélisation informatique, on dénombre un nombre important de modèles possibles. Parmi ces modèles, celui de Frith [21], semble être l'un des plus solides. Il est régulièrement cité dans diverses sources, et est considéré comme

un jalon parmi les autres modèles [8]. Ce modèle s'appuie sur des stades, en estimant que l'acquisition de la lecture se fait par trois stades principaux : le stade logographique, le stade alphabétique, et le stade orthographique.

Le premier, le stade logographique, qui apparaît entre 5 et 6 ans, est également dénommé stratégie visuelle ou stratégie globale. Il consiste à reconnaître le dessin d'un mot plutôt qu'à le déchiffrer. Le mot est alors interprété comme un logogramme, de la même manière qu'un adulte reconnaît une marque en voyant son logo. L'enfant comprend alors le sens d'un mot seulement par sa forme, car il a appris par cœur la signification de ce mot. La procédure de reconnaissance s'apparente alors à celle de reconnaissance d'image et cette reconnaissance serait alors d'autant plus facile que le mot serait inscrit dans un contexte. Par exemple, le mot "bonbon" est plus facilement reconnaissable sur une boîte de confiseries [7]. Progressivement, certains aspects saillants du mot permettent à l'enfant de caractériser le mot, cela peut être sa longueur, l'occurrence ou la forme de certaines lettres comme le "b" de "bonbon". L'identification du mot ne devient plus strictement idéographique mais passe par une reconnaissance des aspects graphiques du mot. Cette reconnaissance n'est pas encore comparable à de la lecture. Elle ne fait pas intervenir les aspects phonétiques du mot. Cela explique des erreurs communes lors des premières phases de l'apprentissage de la lecture, comme la confusion entre deux mots graphiquement proches mais phonétiquement différents comme "bonbon" et "bouton", ou encore du fait du contexte, qui peut pousser un enfant à substituer un mot en un mot de sens proche (par exemple : "voiture" en "auto"). Cette première phase est une introduction au monde de l'écrit et apparaît de manière spontanée mais elle est très limitée et ne permet pas de mémoriser plus d'une centaine de mots. Toutefois, elle permet à l'enfant de passer à une autre étape plus conventionnelle de l'apprentissage de l'écrit, le stade alphabétique.

Ce deuxième stade, le stade alphabétique consiste en une stratégie d'apprentissages de nouveaux mots. Cette stratégie consiste à décomposer un mot en unités graphique plus simples, des graphèmes, et de convertir ces unités graphiques en phonèmes, des unités sonores permettant la distinction de sens [7]. Le stade alphabétique permet de manière systématique de décrypter des mots connus ou inconnus. On peut alors qualifier cette méthode de décryptage de "générative". Ce processus permet donc d'associer des éléments de l'écrit à des éléments phonétiques. Cependant cette association n'est pas biunivoque c'est-à-dire qu'un même graphème peut avoir plusieurs phonèmes associés, et à l'inverse un même phonème peut être associé à plusieurs graphèmes. De même que les phonèmes sont organisés en morphèmes (plus petites unités porteuses de sens oral), les graphèmes sont organisés en morphogrammes. On peut les distinguer en deux catégories, les morphogrammes lexicaux, qui marquent l'appartenance à une famille de mots, comme les mots "laits", "laiterie" et "laitage" reliés par le morphogramme "lait". Il y a aussi les morphogrammes grammaticaux qui marquent des variations de nombre ou de genre comme la terminaison "-ent" à la fin des verbes conjuguées à la troisième personne du pluriel. Toutefois, il existe des différences notables entre les morphèmes grammaticaux (oraux), et les morphogrammes grammaticaux (écrits). En effet, à l'oral en français contemporain, beaucoup de marques de genre et de nombre ne sont plus sensibles. Parfois, des morphogrammes grammaticaux comme "c'est" ou

"ses" et des morphogrammes lexicaux comme "sein" et "saint" n'ont de différences qu'à l'écrit. Ces morphogrammes particuliers forment des logogrammes et sont traités alors comme des logos. En permettant la distinction entre deux homophones, ces logogrammes jouent un rôle important dans la construction du sens de l'écrit. De plus, les signes de ponctuation participent aussi à la compréhension du langage écrit : de manière syntaxique par des blancs entre les mots ou des apostrophes, de manière sémantique avec, par exemple, un point d'interrogation permettant de distinguer une phrase affirmative d'une phrase interrogative, mais aussi de manière prosodique, avec des virgules marquant les groupes de souffle, ou le rythme d'une phrase, qui ne sont pas porteuses de sens comme les ponctuations fortes. Cependant cette approche alphabétique se révèle laborieuse pour des lecteurs apprentis, car elle rend le cap d'identification de 250 à 400 mots par minutes, synonyme d'une bonne compréhension de texte, trop dur à atteindre [7]. C'est pourquoi il devient nécessaire pour l'enfant d'automatiser cette procédure pour ensuite pallier ses insuffisances via une autre stratégie, la stratégie orthographique.

Cette stratégie se différencie de la stratégie alphabétique par le fait que la mémoire récupère instantanément les données phonologiques des mots écrits. En effet, il semble que le lecteur adulte active systématiquement en lecture silencieuse les codes phonologiques des mots écrits identifiés. Toutefois, cette activation est automatisée et d'une rapidité suffisante pour la compréhension d'un texte lu, contrairement au cas de l'apprenti lecteur pour qui cette étape est moins automatisée. La stratégique orthographique se distingue aussi de la méthode alphabétique par les unités de bases considérées. Si, pour cette dernière, il s'agit de graphèmes (unités non signifiantes), la méthode orthographique tend à considérer des morphèmes (unités porteuses de sens) comme unité de base. Ainsi, alors que la méthode alphabétique consiste à convertir des graphèmes en phonèmes, puis à assembler ces phonèmes, la méthode orthographique, elle, décompose un mot directement en unité phonologiques porteuses de sens. Par exemple, le mot « danseur » se décompose avec la méthode alphabétique en les graphèmes <d>.<an>.<s>.<eu>.<r> pour être transcrits dans le domaine oral en les phonèmes /d.ã.s.œ.ʁ/, et enfin les morphèmes "dans- -eur", tandis que la méthode orthographique, identifie directement les morphèmes "dans- -eur" à partir de l'écriture du mot "danseur". Cette stratégie se distingue aussi de la méthode logographique du fait qu'elle fait intervenir un traitement linguistique de l'information, et non plus seulement de la reconnaissance visuelle.

Ce modèle permet, comme nous l'avons vu, de définir clairement des stades d'évolution de l'apprentissage de la lecture chez l'enfant, et chacune de ces étapes permet de définir un aspect du langage à considérer. Une exploitation possible de ce modèle serait de considérer qu'un texte présentant des mots avec une graphie reconnaissable favorise la phase logographique, la lisibilité des symboles la phase alphabétique, et la complexité d'un mot la phase orthographique. Mais si ce modèle est efficace, il n'est pas dénué de défauts. En effet, les modèles les plus récents tendent à réfuter une évolution discrète de l'apprentissage de la lecture, y préférant une évolution interactive, permettant de considérer les connaissances linguistiques préalables à l'apprentissage de la lecture, ce qui n'est pas le cas avec les modèles en séquences comme celui vu précédemment [8]. Nous ne nous attarderons pas sur ces

aspects qui cherchent à savoir quels sont les processus mis en œuvre lors de l'apprentissage de la lecture, d'une part parce qu'aucun modèle ne fait véritablement l'unanimité, et d'autre part parce que notre objectif est d'arriver à une application technique visant à prédire l'âge d'un enfant auquel un texte est destiné, et que bon nombre de ces débats paraissent alors superflus. En revanche, on peut noter que d'autres éléments en dehors du texte peuvent influer sur la compréhension du texte par un enfant. En effet, l'intonation lors de la lecture d'un texte peut influer sur la perception d'un texte [2]. Plus un enfant sera âgé, plus il aura tendance à interpréter un texte selon l'intonation donnée plutôt que par le contexte lu. Ainsi, un jeune enfant (5 à 7 ans) aura tendance à prendre un texte au premier degré. Plus tard (vers 9 ans), il se met à considérer la composante intonative de l'énoncé et, à mesure qu'il grandit, à accorder de plus en plus d'importance au contexte prosodique dans lequel il lui est transmis, jusqu'à faire comme un adulte et considérer en priorité l'intonation. De même, on peut voir dans tout ce qui entoure le texte des indicateurs pour déterminer à qui est destiné le texte. Ainsi, il est possible d'associer à des pages web un âge auquel elles sont destinées, en considérant que plus une page est destiné aux enfants plus elle renvoie à d'autres pages pour enfant, il suffit alors d'initialiser l'algorithme avec un ensemble de pages considérées comme destinées aux enfants, pour référencer tout un ensemble de pages par récurrence [15]. Cet algorithme, nommé AgeRank, en référence à l'algorithme plus générale nommé PageRank, permet alors de dire dans quelle mesure un texte est destiné à tel ou tel âge. Par extension, les textes sur ces pages web peuvent donc être considérés comme destinés au même âge, ou à un âge proche.

2.2 Formules de lisibilité

Les formules de lisibilité ont pour objectif d'offrir une solution simple à un problème complexe, à savoir déterminer de manière systématique la complexité d'un texte. Les premières formules ont été réalisées pour des langues anglo-saxonnes, et afin de déterminer un niveau d'étude associé à un texte, en se basant sur la complexité des mots et des phrases. Plus tard, des travaux francophones ont cherché à les adapter, mais tout en gardant un recul sur la capacité de ces formules à effectivement prédire la lisibilité d'un texte. Nous aborderons successivement les unes et les autres.

2.2.1 Formules de lisibilité anglo-saxonnes

Les formules de lisibilité anglo-saxonnes sont très nombreuses, nous ne nous attarderons que sur les plus significatives, qui ont été citées dans des travaux récents comme base de travail.L'une de ces formules permet de calculer, l'indice de Flesh [17] comme suit :

$$I_{Flesh} = 206,835 - 1,015 \times (\frac{total_{mots}}{total_{phrases}}) - 84,6 \times (\frac{total_{syllabes}}{total_{mots}})$$

Avec $total_{mots}$, $total_{phrases}$ et $total_{syllabes}$ désignant respectivement, le nombre total de mots, de phrases et de syllabes dans le texte étudié. On obtient alors un indice qui correspond à un niveau scolaire, comme le montre le tableau 1 [17]

Indice de Flesh-Kincaid	Niveau d'étude associé
100-90	5^{th} grade level (CM2)
90-80	6^{th} grade level ($6^{ème}$)
80-70	7^{th} grade level ($5^{ème}$)
70-60	$8^{th} - 9^{th}$ grade level ($4^{ème}$-$3^{ème}$)
60-50	$10^{th} - 12^{th}$ grade level (de la seconde à la terminale)
50-30	College (Université)
30-00	College graduates (Diplômés d'université)

Table 1: Niveaux scolaires associés à l'indice de Flesh-Kincaid, pour des textes en anglais.

Une autre formule fréquemment citée est celle de Dale-Chall [18], qui se calcule de la manière suivante:

$$I_{Dale-Chall} = 0,1579(\frac{mots_{difficiles}}{total_{mots}} \times 100) + 0,0496(\frac{total_{mots}}{total_{phrases}})$$

Si pour Flesh, les valeurs issues du textes sont pondérées pour obtenir un indice entre 1 et 10, pour la formule de Dale-Chall, l'indice se situe entre 5 et 10, ce qui explique les constantes utilisées. Pour ce dernier, le terme $mots_{difficiles}$ désigne un corpus de mots désignés comme difficiles. On remarque que pour ces deux formules, les valeurs extraites du textes sont très similaires. On retrouve deux quotients, un représentant la complexité des phrases, (critère syntaxique), et un autre la complexité des mots (critère lexical). Si ces formules sont destinées à être appliquées à la langue anglaise, elles peuvent être adaptées à la langue française.

2.2.2 Application au français et limites

Les formules de lisibilité pour l'anglais, datant du milieu du XXème siècle, ont, pour plusieurs raisons, mis du temps à être adaptée pour le français. C'est pourquoi, lorsque la première des formules de lisibilité, celle de Flesh, a été adaptée pour le français, en en modifiant les coefficients, on prenait déjà conscience des limites d'une approche par formules [18]. Si cette approche, permet de mettre l'accent sur des critères prépondérant dans la compréhension d'un texte (complexité lexicale, complexité syntaxique), elle est trop grossière pour modéliser fidèlement la compréhension d'un texte. Si au XXème siècle cette approche était limitée par la capacité de calcul, les techniques actuelles d'apprentissage automatique permettent de prendre en compte beaucoup plus de critères, et ainsi construire des modèles bien plus fins et fidèles.

2.3 Problème connexes de linguistique computationelle

Les études en TAL (Traitement Automatique des Langues) sur la prédiction automatique de recommandation d'âge, n'existent pas. Cependant, il existe des problèmes similaires, dont les méthodologies offrent des perspectives intéressantes pour notre étude.

2.3.1 Simplification de textes, à destination d'enfants

Notre problème concerne les textes pour enfants. Ainsi il nous est apparu pertinent de voir comment ces textes pouvaient être mis en lien avec des travaux plus généraux sur la lisibilité. Ces travaux tendent à considérer l'importance prépondérante des informations lexicales et syntaxiques dans la compréhension. Ainsi, lorsque De Bolder et Moens [6] cherchent à simplifier un texte à destination des enfants, ils proposent de se pencher sur la simplification lexicale et la simplification syntaxique en priorité. Pour la simplification lexicale, leur méthode consiste à remplacer un mot par un de ses synonymes. Ce synonyme est sélectionné selon sa fréquence, indiquée dans un dictionnaire de mots, parmi tout les candidats possibles. Plus la fréquence d'un de ces mots est élevée, plus il est considéré comme simple. Le plus simple des candidats étant retenu pour ce substituer en tant que synonyme. Une limite de cette méthode est qu'elle donne lieu parfois à des remplacements peu judicieux qui aboutissent à des phrases étranges. Cela serait dû à ce que certains mots complexes ont un sens unique et propre, les rendant irremplaçables dans certains contextes. Quant à la simplification syntaxique, les auteurs considèrent plusieurs cas de structures, qui peuvent être simplifiées. On peut illustrer cela par le tableau 2.

Structure syntaxique	Avant simplification	Après simplification
Apposition	John Smith, a New York taxi driver, won the lottery. (John Smith, un conducteur de taxi New Yorkais, a gagné à la lotterie.)	John Smith is a New York taxi driver. John Smith won the lottery. (John Smith est un conducteur de taxi New Yorkais. John Smith a gagné à la lotterie.)
Proposition relative	The mayor, who recently got a divorce, is getting married again. (Le maire, qui a récemment divorcé, se marie à nouveau.)	The mayor recently got a divorce. The mayor is getting married again. (Le maire a récemment divorcé. Le maire se marrie à nouveau.)
Subordination prefixe	Although it is raining, the sun is shining. (Bien qu'il pleuve, le soleil brille.)	It is raining. But the sun is shining. (Il pleut. Mais le soleil brille.)
Séparation	Tom ate a bagel, and drank his coffee. (Tom mangea un bagel, et but un café.)	Tom ate a bagel. Tom drank his coffee. (Tom mangea un bagel. Tom but un café.)

Table 2: Tableau présentant les différentes structures syntaxiques simplifiables, et leur résultat après simplification

Parmi toutes les possibilités de simplification d'une phrase, seule la plus pertinente est considérée. Cependant, on observe que toutes ces méthodes visent à réduire le nombre de mots par phrase. Cela s'explique par la méthode d'évaluation utilisée. Elle consiste à calculer l'indice scolaire de Flesh-Kincaid [17], qui détermine le niveau scolaire de lisibilité d'un texte. Utiliser cette formule, incite à ne considérer que les critères du nombre de mots par phrases (pour mesurer la complexité syntaxique) et du nombre de syllabes par mots (pour mesurer la complexité lexicale). Outre le fait que l'on puisse qualifier ces critères de trop limités pour mesurer ces complexités, cela nous invite aussi à chercher d'autres critères pour la compréhension des textes par des enfants.

2.3.2 Français langue étrangère

Si on considère des types de critères différents il est aussi intéressant de voir à quel point ils influent sur la compréhension. Ainsi, on peut s'intéresser à l'article de Thomas François et Cedrick Fairon [11], qui cherche à établir une formule de prédiction de lisibilité pour le FLE (Français Langue Étrangère). Le français en tant que langue étrangère est ici intéressant, car il s'agit de lecteurs peu expérimantés de la langue française. L'article liste 46 critères différents, à partir desquels plusieurs modèles sont définis pour prédire les niveaux de lisibilité

du FLE. Ces différents niveaux sont considérés comme des classes, et les modèles prennent la forme de différents classifieurs : SVM, régression linéaire multimodale (RLM) et ordinale (RLO). Les auteurs comparent alors les modèles entre eux, ainsi qu'avec un classifieur aléatoire. Il en résulte que les méthodes les plus complexes sont les plus efficaces, même si le taux de bonne classification tant à plafonner en dessous des 80%. Si les modèles les plus complexes sont certes plus efficaces, ils ne le sont pas assez pour justifier l'utilisation de 46 critères de classification. Certains critères sont donc superflus. Alors, il est intéressant de distinguer les critères les plus utiles pour cette classification.

Outre l'utilité technique de déterminer les critères les plus pertinents, cela permettrait aussi d'offrir à la recherche en psychologie, une validation ou une invalidation de certains modèles. Les performances des différents types d'information sont décrites dans le tableau 3.

	Famille de critères considérée seulement	Toutes les autres familles	Famille de critères considérée seulement	Toutes les autres familles
	Precision	Adj.Precision	Precision	Adj.Precision
Lexical	40,5	75,6	41,1	73,5
Syntaxique	39,3	69,5	43,2	78,4
Sémantique	28,8	61,5	47,8	79,2
FLE	24,9	58,5	47,8	79,6

Table 3: Précision et précision adjacente de la prédiction de classification des différents modèles appris selon certaines informations issues du texte [11]

Comme le montre le tableau 3, l'aspect lexical semble alors être celui qui impacte le plus l'efficacité de la prédiction, ce qui coïncide avec l'idée que plus un texte contient un vocabulaire complexe, plus il est difficile à lire. Ensuite viennent les aspects syntaxiques et sémantiques, déjà exploité par des travaux précédents sur la lisibilité[1], et enfin viennent les critères spécifiques à l'étude de français langue étrangère.

Ces résultats semblent alors assez intuitifs. En revanche, pour Pitler et Nenkova [19], les informations comme le nombre de mots par phrases, ou le nombre de caractères par mots ne semblent pas être de bons critères. Pourtant ce sont ces mêmes critères qui sont au cœur de la formule de Flesh, qui sert encore de base pour la simplification de textes comme par exemple dans les travaux de Bolder et Moens [6]. De plus, des travaux de Pitler et Nenkova, aucun critère clé n'émerge comme étant prépondérant, et c'est plus la combinaison de divers aspects qui permet une réelle performance, allant jusqu'à une prédiction de 88,88% lorsque tous sont utilisés. Cependant il faut rappeler que cet article [19] étudiait un modèle de prédiction de la lisibilité de la langue anglaise, pour des adultes, en considérant notamment le jugement de la qualité d'un article, ce qui nous éloigne de notre problème initial. Il en est de même pour

[1] Voir la partie sur les formules de lisibilité

Feng, Elhadad et Huenerfauth [10], qui étudient des adultes atteints de déficiences, ce qui est un autre problème bien spécifique. Il apparaît selon eux que les résultats sur les textes à destination des enfants sont loin des résultats pour d'autres populations ayant des capacités de lecture déficientes.

Ainsi, des travaux sur le sujet ont permis de défricher le problème et de cibler certains critères à prendre en compte en priorité. Notre problème, qui est la prédiction d'âge pour des textes en français à destination d'enfants, est très spécifique, ce qui d'une part rend les jeux de données pour l'apprentissage rares ou peu connus, et, d'autre part, nous empêche de prendre les résultats des travaux connexes pour argent comptant, surtout dans l'élimination de certains critères de prédiction. La linguistique computationnelle permet donc de concilier deux exigences du problème, le manque de modèle psycholinguistique pouvant faire consensus au sein de la commuanuté scientifique et qui peut être considéré comme fiable, ainsi que la complexité calculatoire du problème, qui demande à prendre en compte de plus en plus de paramètres à mesure que les modèles s'affinent. Les résultats obtenus par des travaux utilisant ces méthodes de linguistique computationelle sont encourageants et invitent à aller dans le même sens. Ainsi, en considérant notre problème comme un problème similaire, il convient de suivre la méthode employée par les auteurs de ces articles. Cependant, nous privilégierons directement les réseaux de neurones plutôt que des modèles de classifieurs plus simples (machine à support vectoriel, régression linéaire, ...), car d'une part ils sont plus performants, et d'autres part, ils permettent des architectures plus souples.

2.3.3 Création de critères basés sur des aspects cognitifs

Tout d'abord, il convient de justifier le fait d'avoir une approche guidée par des travaux en psychologie pour établir de nouveaux descripteurs. C'est à ce titre que l'on peut s'intéresser à l'article de Feng, Elhadad et Huenerfauth [10]. Cet article a pour but d'établir des critères de prédiction de lisibilité, fondés sur une approche cognitive, pour des textes destinés à des adultes ayant des déficiences mentales. Ces critères sont par la suite utilisés par des algorithmes d'apprentissage automatique, pour créer des modèles prédictifs de la lisibilité de ces textes, pour des adultes déficients. Les auteurs de l'article prennent leurs distances avec les formules de lisibilité classique, car selon eux, elles sont certes utiles mais peu représentatives de la complexité d'un texte. Ils proposent donc une liste de critères de lisibilité pour créer de nouveaux modèles, plus fidèles aux spécificités cognitives de la population-cible. Ces critères sont ou bien issus des recherches précédentes dans le domaine, ou bien nouveaux et établis en fonction des problématiques cognitives. Par exemple, les entités d'un texte sont prises en compte. Le terme d'entité désigne ici les différents objets ou protagonistes dans un texte. Pour des adultes déficients, le nombre d' «entités» dans un texte est corrélé positivement avec la difficulté qu'ont ces personnes à lire le texte. C'est pourquoi, les auteurs ajoutent les critères de densité d'entités dans un texte, du nombre d'entités par phrases et du nombres de mentions de ces entités (*via* des pronoms, des articles, etc). Les auteurs travaillent sur quatre corpus utilisés pour les jeux d'apprentissage et les tests des modèles. Parmi ces quatre corpus, trois sont des corpus déjà existants. Le premier regroupe des articles simplifiés pour enfants et leurs versions originales, le second est un ensemble de nouvelles locales, un thème

populaire chez la population cible, et le troisième contient des articles destinés à des élèves d'écoles primaires. Pour avoir un modèle plus en lien avec leur problème, les auteurs ont créé un quatrième corpus, composé dune vingtaines d'articles édités spécifiquement pour des adultes ayant des déficiences. Les modèles d'apprentissage du taux de lisibilité sont alors entraînés avec 80% des textes totaux, puis testés sur les 20% restants. On peut classer tous les critères en deux catégories : d'une part, ceux inspirés des travaux précédents sur le sujet, et d'autre part, ceux créés selon une approche cognitive. À partir de ces deux catégories de critères, on peut construire plusieurs modèles et les comparer pour évaluer les performances des différents critères. Les résultats de l'évaluation des modèles sont résumés par le tableau 4.

Modèle de lisibilité	Erreur moyenne de prédiction
Indice de Flesh-Kincaid	2,569
Critères basiques seulement	0,6032
Critères cognitifs seulement	0,6110
Tous les critères	0,5650

Table 4: Présentation de l'erreur moyenne de prédiction des modèles obtenus à partir du groupement de critères en fonction de leur conception

Ces résultats semblent corroborer la thèse initiale de l'auteur, qui était que les critères fondés sur une approche cognitive du problème de lisibilité augmentaient l'efficacité des résultats. Cependant les auteurs précisent aussi que ce modèle est très sensible au corpus utilisé pour l'apprentissage. En effet, lorsque seul le corpus des textes pour enfant est utilisé pour l'apprentissage et que le test est réalisé sur les nouvelles locales, il apparaît que les modèles ne prédisent plus un résultat satisfaisant. Cet article éclaire donc notre problème en nous invitant à choisir les critères d'évaluation des modèles à partir de travaux sur les sciences cognitives de la population cible. De plus, la population cible est composée d'individus ayant une pratique de la lecture faible ou approximative, de ce fait, la méthode employée par les auteurs semble être solide, mais doit toutefois être adaptée à la population-cible, surtout pour le choix des textes du corpus.

3 Choix des descripteurs

L'état de l'art nous a montré les nombreuses facettes de la psycholinguistique qui contribue à la compréhension d'un texte pour des enfants. Néanmoins, cet état de l'art nous a également permis de constater que les travaux en TAL exploitent que trop ces facettes. Ainsi, cette section liste les différents descripteurs que nous avons retenu, avec pour chacun d'eux une justification psycholinguistique associée.

3.1 Lexique et syntaxe

Lorsque l'on pense à un texte pour enfant, on a tendance à imaginer un texte avec des phrases simples et peu ou pas de mots très compliqués. Autrement dit, plus les informations lexicales et syntaxiques sont complexes, moins on estime que le texte est destiné à un enfant. Et en effet, comme nous l'avons vu, ce sont ces informations qui sont le plus souvent utilisées pour déterminer la lisibilité d'un texte [18] ou juger la simplification de textes [6]. Pour extraire ces informations d'un texte, nous ne pouvons cependant pas être satisfait des anciennes formules de lisibilité[18], car, comme nous l'avons vu dans l'état de l'art, ces formules ne modélisent que très grossièrement les informations lexicales et syntaxiques. Pour cela on peut utiliser des outils plus récents d'analyse de textes, comme le POS-Tagging qui permet d'avoir un étiquetage morphosyntaxique de la phrase donnée. Dans notre cas nous utiliserons le POS-Tagging de Berkeley adapté au français[4]. Selon cet article les performances des différents parsers sont similaires, nous avons donc choisi celui utilisé par l'équipe ALPAGE de l'INRIA[2]. Cet outil a été appris sur un corpus arboré de phrases annotées selon les rôles morphosyntaxiques, ainsi que que les dépendances syntaxiques de chacun des mots, de plus, ce corpus est très utilisé dans le traitement automatique de langue en français [1]. Cette annotation prend la forme d'un arbre, comme on peut le voir dans la figure 2.

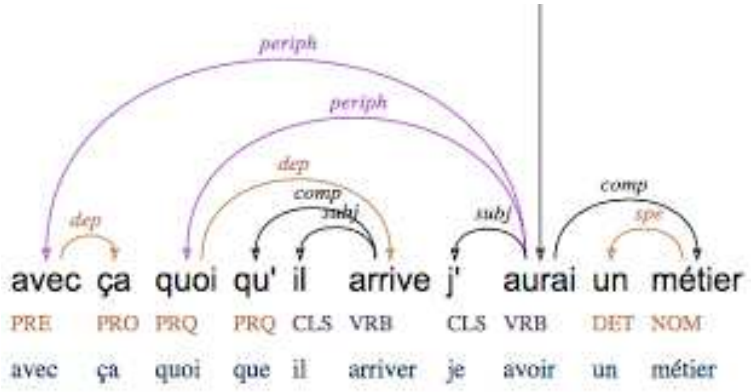


Figure 2: Exemple d'une annotation arborée d'une phrase, présentant les rôles morphosyntaxiques des mots ainsi que leurs dépendances syntaxiques

Grâce à cet outil nous pouvons créer plusieurs descripteurs. On privilégiera les quantités relatives à la tailles de la phrase, plutôt que des quantités absolues, comme c'est le cas pour les travaux sur le français langue étrangère [11]. Afin d'avoir d'éviter que les informations importantes soient trop diluées dans l'ensemble des mots, on peut écarter les mots "accesoires", ou stopwords, tels que "le", "de", ou "à", qui rallongent la taille de la phrase, sans apporter énormément d'informations. Pour discerner ces mots vides, on peut se baser sur une liste prédéfinie. Celle utilisée ici à été conçue par nous mêmes, mais on peut en trouver des similaires un peu partout[3].

[2] `http://alpage.inria.fr/statgram/frdep/fr_stat_dep_bky.html`
[3] `https://www.ranks.nl/stopwords/french`

Une autre exploitation possible de cet étiquetage morphosyntaxique, est l'usage des inter-dépendances syntaxiques au sein d'un même phrase. Ces dépendances forment alors un arbre, dont chacun des noeuds est un mot. Dans cet arbre les fils dépendant syntaxiquement de leur père, comme c'est le cas dans l'exemple donné en figure 2. Étudier la forme de cet arbre (nombre moyen de fils par noeud, profondeur, etc) permet d'avoir des informations utiles pour modéliser la complexité syntaxique. Ainsi, grâce à ce travail sur les lemmes et les dépendances syntaxiques grâce au POS-Tagging, on peut construire les descripteurs suivants :

- Pourcentage de verbes/noms/adjectifs/clitiques/stop words;
- Nombre de fonctions grammaticales différentes dans la phrase;
- Profondeur de l'arbre de dépendance syntaxique;

À ces descripteurs globaux (sur l'ensemble de la phrase), on peut ajouter des descripteurs dits locaux (sur chacun des mots):

- Nombre de dépendances pour un mot donné;
- Nombre de mots dépendants d'un mot donné;
- Moyenne des distances entre un mot donné et ses dépendances;
- Maximum des distances entre un mot donné et ses dépendances;

De plus, pour continuer à modéliser la complexité lexicale, on peut ajouter les probabilité d'apparitions d'un mot donné dans la langue. Cette valeur est celle d'une probabilité de présence dans la langue obtenue à partir d'un vocabulaire de 64000 mots, provenant de corpus de textes variés (journaux, parole transcrite, émissions de radios, etc).

3.2 Temporalité et causalité

Nous avons vu que l'utilisation chez l'enfant de repères temporels évolue en fonction de son âge, selon sa maîtrise de la langue [20]. Si ces indications concernent le langage utilisé par l'enfant à l'oral, alors que nous étudions le langage écrit destiné aux enfants, nous pouvons toutefois tenter d'ajouter ces critères à nos modèles et en déduire des descripteurs.

Les descripteurs que nous considérons pour représenter ces différents repères temporels sont:

- Les pourcentages d'utilisation des différents temps verbaux au sein d'une phrase;
- Les pourcentages d'utilisation des modes de conjugaison utilisés (infinitif, indicatif, subjonctif);
- Les pourcentages d'utilisation des systèmes temporels (passé, présent, futur);

En plus de ces repères temporels fondés sur la conjugaison des verbes, nous nous sommes intéressés à ceux fondés sur les connecteurs logiques. En effet, le repérage de l'enfant dans le temps par un enfant, peut se faire aussi en structurant plusieurs événements dans une phrase, et en se positionnant dans une chaîne de causalité. Il est alors possible de supposer que, plus un texte utilise de connecteurs logiques, et plus ils sont diversifiés, plus ce texte s'adresse, à un lecteur comprenant la langue. Ainsi, à partir de plusieurs listes de connecteurs logiques, on peut calculer la proportion de ceux-ci dans une phrase par rapport à l'ensemble des mots. Nous avons ainsi retenu 16 listes prédéfinies de connecteurs logiques selon leur types: addition, but, cause, comparaison, concession, conclusion, condition, conséquence, énumération, explication, illustration, justification, opposition, restriction, exclusion et temps. Ces listes sont obtenues en croisant plusieurs sources venant d'Internet, aucune ne faisant vraiment consensus.

3.3 Graphie et Phonétique

Comme nous l'avons vu précédemment dans l'état de l'art, le modèle de Frith d'apprentissage de la lecture [21], l'apprentissage de la lecture passe d'abord par une reconnaissance de la graphie, et suppose aussi une maîtrise de langage oral. Il est alors intéressant d'observer des descripteurs caractérisant cette complexité du texte, qu'elle soit graphique ou phonétique. Ainsi, pour chacun des mots, nous calculons un score de confusion, à mesure que les lettres contiguës d'un mot se confondent, cette confusion étant déterminée à partir d'une matrice de confusion des vingt-six lettres de l'alphabet en annexe 1. En notant M la matrice de confusion, alors le score de confusion d'un mot composé de n lettres notées l, se calcule comme suit :

$$Score_{confusion}(\text{mot}) = \sum_{i=0}^{n-1} M_{l_i,l_{i+1}}$$

À cela nous avons aussi pris en compte la longueur du mot, ainsi que la moyenne du pourcentage d'apparition des lettres qui composent le mot dans la langue française. Cette fréquence des lettres étant obtenue à partir du corpus WikipédiaFR2008 obtenue par le laboratoire CLLE-ERSS, à partir des articles de Wikipedia en français[4].

D'une manière similaire, on peut construire des descripteurs locaux sur la phonologie, mais pour cela il faut au préalable obtenir la transcription phonétique du mot. Cela est fait grâce au logiciel eSpeak[5]. On peut alors obtenir la diversité des phonèmes du mot, une moyenne des fréquence des phonèmes selon l'occurrence de ces phonèmes dans la langue [9], ou encore le nombre de phonèmes dans le mot. De plus, la prononciation d'une phrase en français diffère celle de la prononciation mots à mots, du fait des liaisons. Ainsi nous avons aussi créés les mêmes descripteurs que ceux décrits précédemment, mais sur la prononciation de l'ensemble de la phrase, et non plus seulement de la prononciation mot à mot.

[4] `http://redac.univ-tlse2.fr/corpus/wikipedia.html`.

[5] `https://doc.ubuntu-fr.org/espeak`.

3.4 Sémantique

Jusqu'ici nous n'avons fait qu'explorer en peu plus profondément les complexité lexicales et syntaxiques du texte. Or, lorsque l'on cherche un texte pour enfant, on s'attend surtout à trouver un champ sémantique assez défini, avec des thèmes enfantins. En effet, les mots 'Mangée" et "Pangée" ont une prononciation et une graphie proches, mais ils sont sémantiquement très distincts. Or il est assez difficile d'obtenir une modélisation directe de la sémantique d'un mot ou d'une phrase. Pour cela nous nous sommes appuyés sur des embedding de mots.

L'embedding de mots consiste en une représentation d'un objet (ici un mot) dans un espace de grande dimension, ici de dimension 500. Les mots sémantiquement proches sont alors proches dans cet espace vectoriel. Le mot "Mangée" est alors proche, dans cet espace, des mots en rapport avec la nourriture, tandis que "Pangée" se trouve plus proche de mots en rapport avec la géologie. L'eMebedding que nous avons utilisé est celui développé par Jean-Philippe Fauconnier, appris sur 1.6 milliards de mots[6], utilisé dans des travaux sur des sujets similaires au nôtre [16]. Un mot issu d'un des textes que nous étudions a alors de fortes chances d'avoir sa représentation vectorielle dans cet espace. Le vecteur de dimension 500 qui est alors associé au mot peut être alors considéré comme comme une modélisation de son information sémantique.

3.5 Sentiments

Enfin, lorsque l'on parle de textes pour enfants, la transmission d'émotions semble elle aussi importante. En effet, les très jeunes enfants, ne lisent pas eux-mêmes le texte mais ce sont leurs parents ou leurs proches qui oralisent leur lecture. Pour autant le texte leur est bien destiné et l'accent est souvent mis sur le lien émotionnel entre les enfants et leurs parents pour faciliter la compréhension et la mémorisation du texte par l'enfant [3]. En outre, plus l'enfant grandit, plus il est en mesure d'appréhender des émotions complexes [5].

Pour analyser ces sentiments, on utilise généralement un corpus d'avis de spectateurs, en français, sur des films [14]. A partir de ce corpus, il est possible, grâce au module TextBlob de Python[7], d'obtenir un score de polarité entre -1 et 1 et un score de subjectivité entre 0 et 1 pour chaque phrase, représentant respectivement si la phrase est positive ou négative et si son discours est plutôt objectif ou subjectif. Cependant, si cette méthode est souvent utilisée pour des modèles d'analyse d'opinion, elle n'est pas suffisante pour étudier l'ensemble du spectre émotionnel. Pour cela, nous avons utilisé un lexique de mots annotés où une émotion est associée à chaque mot. De là, nous avons construit pour chacun des mots un vecteur one-hot de taille 15, ou chacune des coordonnées correspond à une émotion, cette coordonnée vaut 1 si le label associé au mot dans le lexique est bien l'émotion associée à la coordonnée, et 0 sinon.

[6] Ces mots sont issus du corpus FrWac construit à partir du Web en prenant les noms de domaine terminant par ".fr".

[7] On utilise ici une version adaptée au français : `https://github.com/sloria/textblob-fr`

3.6 Ajustements

Le fait de devoir travailler à l'échelon de la phrase, nous pousse aussi à revoir certains descripteurs. En effet, pour garder les informations extraites par des descripteurs locaux il nous a fallu "globaliser" ces informations. Pour cela, nous avons créé des descripteurs globaux en utilisant des outils statistiques sur les descripteurs locaux d'une même phrase. Ainsi, nous avons gardé la moyenne et la variance des descripteurs sur les mots d'une même phrase comme descripteurs de cette phrase. De la même manière, les descripteurs qui caractérisent les mots selon un vecteur (Embedding, Émotions), sont globalisés, en prenant le barycentre des vecteurs des mots d'une même phrase. Une liste exhaustive des descripteurs dans leur forme finale est laissée en annexe 3.

4 Données utilisées

Maintenant que nous avons défini les critères à extraire de nos données, en suivant des études portant sur la psycholinguistique, nous pouvons nous intéresser aux données à traiter. Ainsi, nous allons voir ici le corpus utilisé et comment nous l'avons traité.

4.1 Corpus utilisé

Le corpus que nous allons utiliser est composé de textes pour enfants de 0 à 14 ans, qui peuvent être des histoires ou des articles de journaux et pour chacun de ces textes, une tranche d'âge est associée. Cette tranche d'âge a été établie de par les auteurs ou les éditeurs du texte source. Ainsi par exemple, notre corpus comporte des textes du journal *WAKOU*. Ce journal étant destiné aux enfants de 7 à 10 ans, alors on peut en déduire que les textes qui le composent sont destinés à ces mêmes enfants de 7 à 12 ans. Toutefois nous n'avons gardé que les textes de ces sources, qui sont bien souvent accompagnées d'illustrations pouvant aider à la compréhension du texte.

À ces textes pour enfants, nous avons ajouté des textes destinés aux adultes. Ces textes pour adultes sont issues de sources très diverse (romans, blogs, articles de journaux), et recouvrent trois niveaux de langues (courant, familier, soutenu). La diversité de ces sources et nivaux de langues, représentés par quinze textes chacun, permet d'au mieux modéliser un langage adulte moyen.

Ce corpus a été recueilli par des techniques diverses, dont par OCR[8]. Cela implique qu'il peut y avoir du bruit issu de ces techniques au sein de ces textes. Cependant, ce corpus ne comprend que 230 textes, ce qui est insuffisant pour travailler à cet échelon de données. De plus la majorité des outils de traitement automatique des langues, à l'heure actuelle, s'applique à des phrases. Ainsi, nous avons décidé de réduire l'échelle de travail à l'échelon de la phrase pour augmenter le nombre d'exemples à observer et ainsi pouvoir appliquer des méthodes d'apprentissage automatique via des réseaux de neurones. Or, si l'on peut avoir une expertise sur les tranches d'âges quand il s'agit de textes, expertise fournie grâce

[8]Optical Caracter Recognition (Reconnaissance Optique des Symboles)

aux éditeurs de ces textes, il n'est pas possible d'avoir une annotation similaire phrases par phrases. Ainsi, nous avons fait le postulat que, si une phrase est issue d'un texte avec une tranche donnée, la phrase sera annotée avec la même tranche d'âge. De plus, afin d'avoir des données indépendantes les unes des autres, nous avons considérés que deux phrases d'un même texte sont indépendantes entre elles.

Si pour les textes pour enfant, une annotation de la tranche d'âge est proposée, nous avons décidé d'attribuer aux textes pour adulte la tranche d'âge 14-100 de manière arbitraire.

Cela donne au final un corpus de plus de 9400 phrases, dont un tiers est issue de textes pour adulte. Bien que cela reste assez peu il est toujours possible d'utiliser des techniques d'apprentissages automatiques utilisant des réseaux de neurones simples.

4.2 Distribution des données

Les tranches d'âges associées aux phrases sont diverses et ont tendance à se chevaucher. Ce chevauchement rend une classification par tranche d'âge obsolète, la tranche d'âge 3-5 ans et 3-6 ans seraient ainsi deux classes différentes, alors que leurs textes s'adressent en majorité aux mêmes enfants. Ainsi, il est plus pratique pour notre problème de parler d'âge minimum et d'âge maximum que d'une tranche d'âge, ces deux âges pouvant par la suite être l'objectif d'une tâche de régression. Pour un âge donné, la figure en annexe 2 présente le nombre de phrases dont la tranche d'âge recouvre un âge donné.

On remarque alors que les âges faibles, de 0 à 3 ans sont très peu représentés, ceux de 3 à 10 assez peu, ceux de 12 à 14 ans le sont beaucoup, et ceux au-delà de 14 ans, c'est à dire uniquement les textes destinés aux adultes, représentent un tiers du jeu total de données et sont donc assez bien représentés. En raison de ces déséquilibres et manques partiels, nous pouvons nous attendre à avoir des modèles moins performants pour prédire les âges faibles.

Par la suite nous allons subdiviser aléatoirement l'ensemble des phrases en trois jeux de données. Un jeu d'entraînement (70% des phrases), sur lequel les modèles testés apprennent, un jeu de développement (15% des phrases) grâce auquel les modèles testés peuvent converger en évitant un sur-apprentissage sur les données d'entraînement et, enfin, un jeu de test (15% des phrases) qui nous permet de tester les modèles une fois appris et d'analyser les résultats. Cette subdivision aléatoire conserve cependant bien les proportions du corpus d'origine, en conservant la même proportion de chaque tranche d'âge dans chacun des trois jeux de données.

5 Architectures testées et Résultats

Dans cette section, nous étudions et comparons différentes modélisations possibles, nous détaillons leur implémentation et discutons leurs résultats. A l'issue de cette analyse globale, nous terminons par un examen plus détaillé des différents descripteurs.

5.1 Modélisations

Afin de résoudre la prédiction d'âge à partir d'un texte, il faut choisir quel type de modèle de prédiction automatisée est à retenir. Nous avons choisi de nous concentrer sur les réseaux de neurones, car ils sont connus pour avoir une bonne performance et permettent divers types d'architectures, tout en étant applicables sur un corpus relativement faible. Nous n'utiliserons ici que des réseaux de neurones dits *feed-forward*, c'est à dire non récurrents, plus aisées à mettre en oeuvre.

La performance des modèles sera évaluée à l'aide d'une erreur moyenne absolue, entre l'âge prédit et l'âge réel associé à la phrase, que l'on notera MAE[9]. Ainsi une MAE de 2,5 signifie que, en moyenne, le modèle en question prédit un âge de plus ou moins deux ans et six mois par rapport à l'âge réel associé au texte. Pour les modèles prédisant à la fois l'âge minimum et l'âge maximum associé à une phrase, on parlera du couple (MAEmin, MAEmax), où MAEmin désigne l'erreur moyenne sur la prédiction de l'âge minimum, et MAEmax sur l'âge maximum.

5.1.1 Modèles naïfs

Tout d'abord, nous allons décrire les modèles les plus simples. Une méthode naïve de prédiction, serait de prédire pour chacune des phrases du jeu de test, la moyenne des âges minimum des phases données en apprentissage. Par cette méthode, on obtient pour ce modèle une MAEmin d'environ 3,5. Ainsi cette valeur de 3,5 peut être vue comme une valeur plancher, discriminant tout modèle qui serait moins performant que ce modèle naïf.

De plus, si l'on essaye un modèle très simple, n'observant que l'âge minimum des phrases et ayant une unique couche cachée de 100 neurones, on obtient un modèle ayant une MAEmin de 2,71, ce qui est déjà bien mieux que notre modèle naïf.

Pour la suite, nous ne nous intéresserons qu'à des modèles cherchant à prédire par régression, à la fois l'âge minimum et l'âge maximum associé à une phrase, afin d'être plus fidèle à notre objectif initial qui est de prédire une tranche d'âge.

5.1.2 Sélection des hyperparamètres

Puisque nous ne nous intéressons uniquement aux réseaux de neurones dits *feed-forward*, il convient de choisir les bons hyperparamètres associés. De tels réseaux de neurones denses se caractérisent par deux hyperparamètres principaux, le nombre de couches cachées du réseau et le nombre de neurones composant ses couches. Il existe bien d'autres hyperparamètres à définir comme les fonctions d'activation ou la méthode d'optimisation, mais ceux-ci influant très peu sur les performances des modèles appris, nous avons préféré les laisser constants, avec une fonction d'activation sigmoïde, une méthode d'optimisation dite "adam". La fonction de loss, quant à elle, est calculée en fonction de la MSE[10], qui permet de prendre en compte les erreurs quadratiques lors de l'apprentissage. Entre les différents modèles présentés ici,

[9] pour *Mean Average Error*
[10] pour *Mean Squared Error*

nous ne faisons que les hyperparamètres plus importants, que sont le nombre de couches et le nombres de neurones par couches.

À la vue du problème étudié et de la taille de nos données, nous avons tester des réseaux de 1, 2, 4 et 8 couches, et dont ces couches peuvent être composées de 100, 200, ou 500 neurones chacune. Les résultats des tests de ces modèles sont données par la figure 3

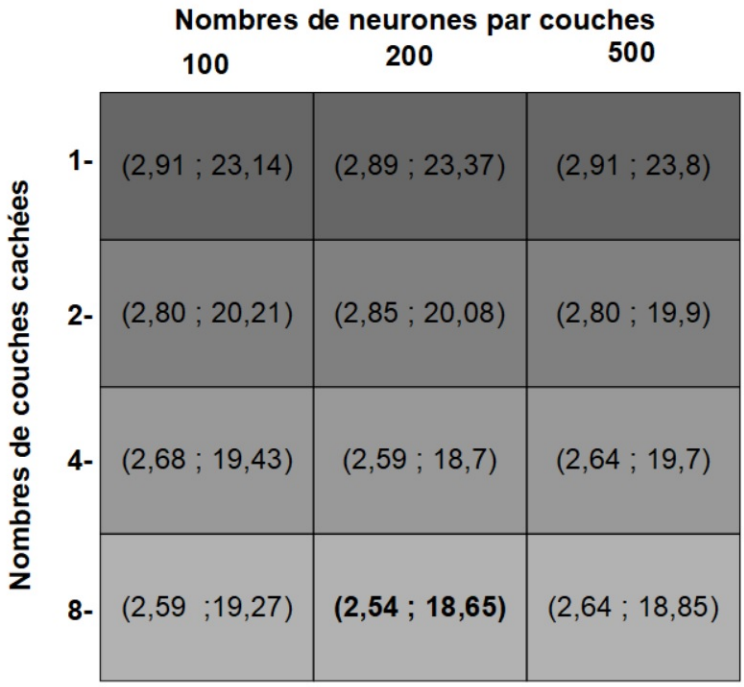


Figure 3: Heatmap présentant les couples (MAEmin ; MAEmax) des différents modèles selon leur nombre de couches cachées et du nombre de neurones au sein de ces couches

On remarque alors que la performance du modèle augmente à mesure que l'on augmente le nombre de couches cachées, mais tend à se dégrader si ces couches comportent trop de neurones. En effet, la meilleure combinaison d'hyperparamètres semble être celle avec 8 couches de 200 neurones . C'est donc cette combinaison d'hyperparamètres que nous utiliserons pour la suite, dans nos tâches de régression, quelque soit les architectures employées.

5.2 Architectures

Si notre problème semble bien défini, il existe plusieurs manière de l'aborder, ainsi nous présenterons ici les différentes architectures envisagées.

5.2.1 Architecture par régression

Ici nous utilisons des modèles très similaires à ceux décrits précédemment. Ces modèles utilisent une méthode de régression pour prédire l'âge minimum et l'âge maximum, en se

basant sur les caractéristiques extraites du la phrase. Ces caractéristiques prennent la forme d'un vecteur de dimension 621, qui est alors le vecteur d'entrée. Cette architecture que nous nommerons "RegOnly", peut être schématisée par la figure 4

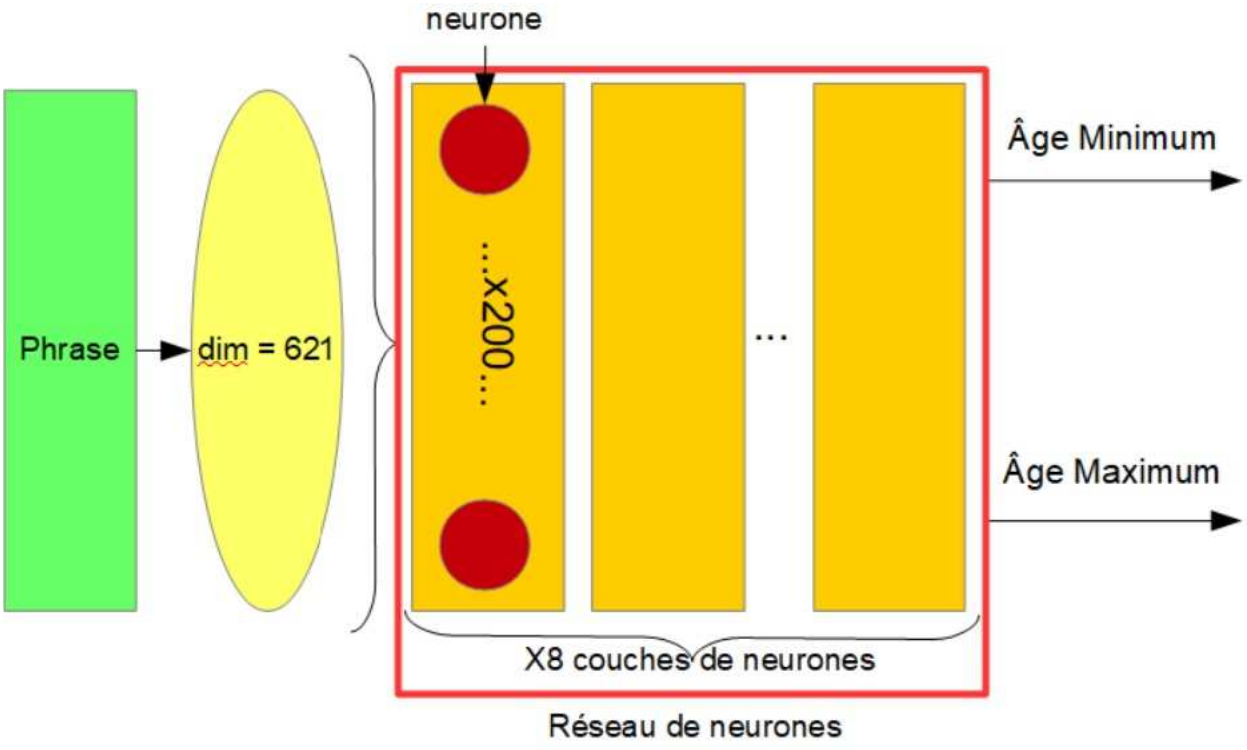


Figure 4: Diagramme représentant l'architecture utilisée pour la régression simple

Cependant, nous avons décidé arbitrairement d'attribuer aux textes pour adultes la tranche d'âge 14-100, de la même manière on peut dire que si un texte est un texte pour adulte, alors il a un âge minimum de 14, et un âge maximum de 100. Cette distinction permet alors d'éviter des erreurs absolues trop importantes sur ces textes pour adultes. Ainsi, outre les âges minimum et maximum pour chaque phrase, on peut essayer par ces modèles de déterminer si cette phrase est pour adulte ou non.

5.2.2 Architecture multi-tâche

Afin d'utiliser ces méthodes de régression pour prédire si un texte est destiné aux adultes ou non, on peut rajouter une nouvelle variable, qui vaut 1 si la phrase est issue d'un texte pour adulte et 0 sinon.

On peut alors réaliser un travail de régression sur plusieurs variables. Cette méthode d'apprentissage multitâche ou MTL[11], prédit donc trois informations, l'âge minimum, l'âge

[11] pour *Multi Task Learning*

maximum, et une variable déterminant si un texte est pour adulte ou non. À partir de cette dernière variable, si sa valeur est strictement supérieure à 0.5, on estime que la phrase en question est destinée aux adultes, et donc on change les valeurs prédites d'âge minimum et maximum par respectivement 14 et 100. A l'inverse, si cette valeur est inférieure ou égale à 0.5, on estime que la phrase est issue d'un texte pour enfant et on conserve les valeurs prédites. Ce modèle est schématisé par la figure 5.

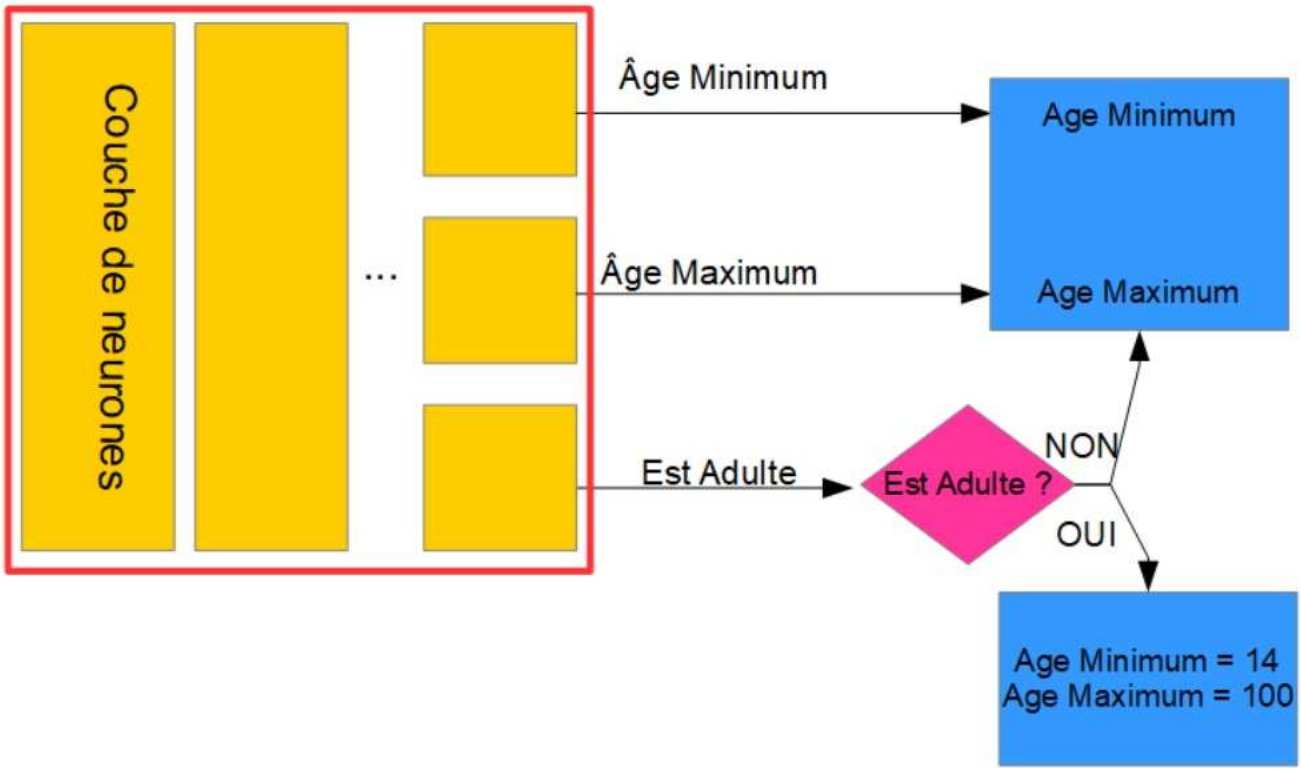


Figure 5: Diagramme représentant l'architecture utilisée pour la régression multitâche

Cependant, si cette distinction semble utile, on peut toutefois hésiter à utiliser le même réseau pour prédire à la fois les recommandations d'âges, et la variable déterminant si une phrase est issue d'un texte pour adulte. On peut alors choisir d'utiliser un réseau de neurones uniquement pour déterminer si une phrase est destinée aux adultes via une tâche de classification, puis réaliser la régression a posteriori.

5.2.3 Architecture par double réseau de neurones

Pour pouvoir déterminer en amont si une phrase est issue d'un texte pour enfants ou d'un texte pour adultes, on peut utiliser un réseau de neurones annexe, réalisant une tâche de classification. Ce réseau prend en entrée exactement le même vecteur que les réseaux évoqués précédemment, mais est bien plus simple. En effet il ne comporte que 5 couches de 100 neurones chacune, mais cela semble suffisant d'autant plus qu'augmenter au diminuer un de

ses hyperparamètres ne change pas significativement les performances du modèle, mais tend à augmenter son temps d'apprentissage.

À la suite de cette classification, pour toutes les phrases classées comme *adulte*, on assigne la tranche d'âge 14-100, et pour toutes celles classées comme *enfant*, on applique les méthodes de régression évoquées précédemment. Puisqu'on suppose que tous les textes offerts à ce modèle sont censés être des textes pour enfant, on apprend ce modèle uniquement sur les textes pour enfant, ce qui permet d'avoir un modèle plus précis. L'ensemble de cette architecture, que nous nommerons Bi-Model car elle fait intervenir deux réseaux distincts, est schématisé par la figure 6.

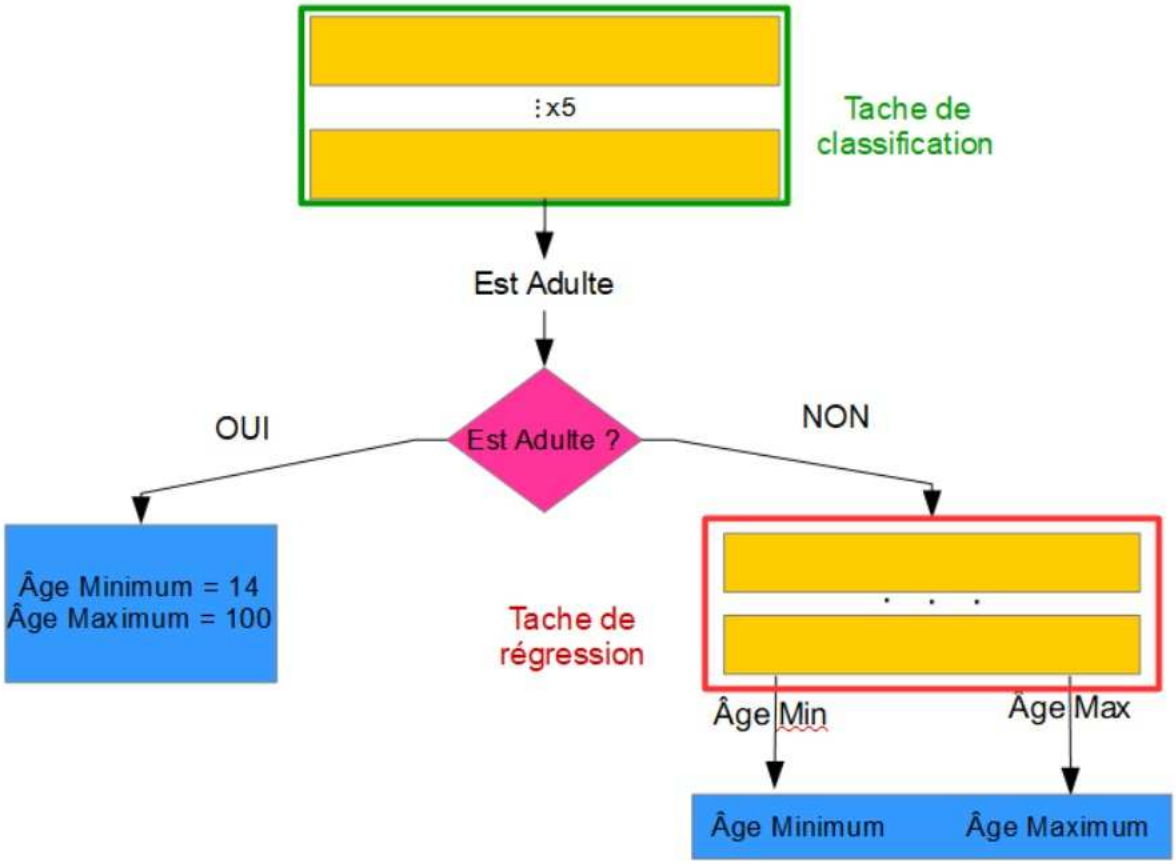


Figure 6: Diagramme représentant l'architecture faisant intervenir une tâche de classification, puis de régression

5.3 Résultats

Si on peut chercher à travers ces différentes architectures une manière de plus en plus précise de prédire un âge minimum et un âge maximum, on peut aussi chercher à déterminer quelles sont les catégories de descripteurs les plus influentes.

5.3.1 Performances des architectures

Une fois appris sur les mêmes jeux d'apprentissage et de développement, on peut comparer les trois architectures précédemment décrites. On obtient alors les résultats du tableau 5.

Architecture	MAEmin	MAEmax
RegOnly	2,54	18.65
MTL	2,44	19,3
Bi-Model	1,61	18,3

Table 5: Erreurs moyennes absolues (MAE) de prédiction, sur les âges minimums et maximums pour les différentes architectures

On peut remarquer plusieurs choses. Tout d'abord, la meilleure des architectures est celle nommée BiModel, qui fait intervenir deux réseaux de neurones indépendants. En effet, elle prédit un âge pour chaque phrase, qui est en moyenne, juste à 1.61 ans près. En revanche, on remarque que les prédictions pour les âges maximum sont bien plus grossières. En effet, les modèles prédisent juste à une vingtaine d'années près. Cela peut s'expliquer par la présence de textes pour adulte ayant des âges maximaux de 100, beaucoup plus élevés que les textes pour enfants, qui ont un âge maximum de 14 ans. Cependant, cette grande erreur sur l'âge maximum est moins préjudiciable qu'une erreur sur l'âge minimum, car, pour conseiller un texte à un enfant, on privilégie surtout l'âge minimum, l'âge maximum n'ayant qu'une valeur indicative.

Enfin, il convient de regarder les prédictions erronées, lorsque cette classification est réalisée, pour essayer de comprendre la raison de leurs erreurs ainsi que d'essayer de voir par quels moyens ces réseaux de neurones de classifications peuvent être améliorés.

5.3.2 Analyse des prédictions erronées

Voici quelques exemples de prédictions fausses réalisées par le modèle appris avec tous les descripteurs selon l'architecture BiModel :

- "La matière première du courage c'est le temps" .

Cette phrase est issue d'un texte pour adulte. Cependant, le modèle l'a considérée comme un texte pour enfants, et a associé à cette phrase la tranche d'âge de 7 à 12 ans. Si on imagine bien cette phrase être tirée d'une œuvre pour adultes, de par son ton sérieux, il est difficile de dire en quoi elle ne serait pas adaptée à un enfant entre 7 et 12 ans. Ainsi, si, d'après nos critères, le modèle a tort, il est difficile cependant d'y voir une erreur dans son apprentissage. Cela nous invite remettre en cause notre hypothèse qui dit que l'annotation de la phrase suit l'annotation du texte dont elle est issue. De la même manière on peut s'intéresser à cette phrase :

- "Vendredi soir , on a aussi parlé de kamikazes."

Cette phrase, d'un article du *P'tit Libé* est issue d'un texte recommandé aux enfants de 7 à 12 ans, et pourtant le modèle l'a classée comme une phrase issue d'un texte pour adulte. Cela peut s'expliquer par la présence du mot "kamikazes" qui est peu commun aussi bien phonétiquement que graphologiquement, et dont la sémantique est très éloignée de celle de textes pour enfants. Le fait qu'une telle phrase soit dans un journal pour enfant est compréhensible, mais, hors de son contexte elle semble intuitivement plus adressée aux adultes. Ainsi, là encore, si on estime que le modèle a tort de par notre hypothèse sur les annotations des phrases, son apprentissage semble pour autant relativement correct. Ces fausses prédictions, qui semblent plus dues à une approximation de notre part qu'à une réelle erreur de modèle sont assez courantes. Bien sûr, il reste des prédictions erronées qui ne trouvent pas d'explications de ce genre, mais elle sont assez rares. On peut aussi s'intéresser sur les âges les plus touchés par ces prédictions erronées. Ainsi la figure 7, nous présente les âges correspondant aux tranches d'âges des phrases mal classées.

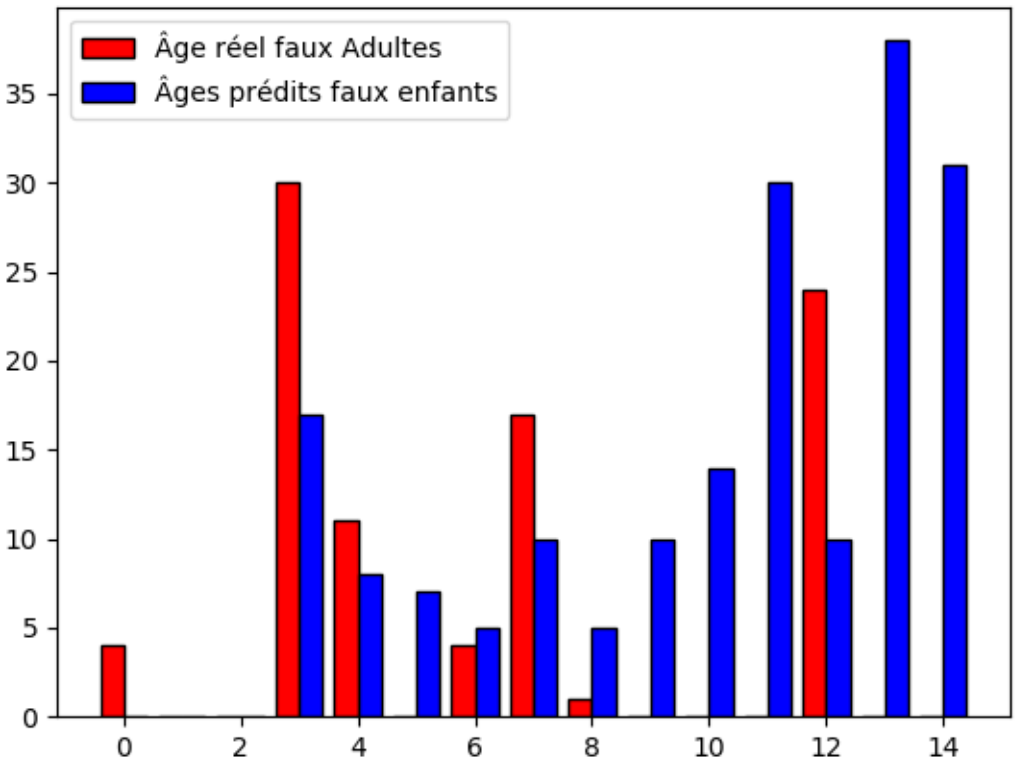


Figure 7: Histogramme présentant les âges recouverts par les tranche d'âges réels des textes faussement prédits comme adulte (en rouge) et par les tranches d'âges prédites des phrases faussement prédites comme étant destinées à des enfants (en bleu)

Tout d'abord, on peut voir que ces erreurs de classification se font surtout par des erreurs sur des phrases pour adulte, qui sont prédites comme destinées aux enfants. Toutefois, de par notre architecture, ce type d'erreur de classification peut être compensé par la tâche de régression réalisée *a posteriori*. En effet on remarque que les âges recouverts par ces fausses prédictions, représentés sur la figure 7 par les colonnes bleues, sont des âges autour de 11, 13 ou 14 ans, soit la borne maximale des âges pour enfants. Dans une moindre mesure, on peut faire cette même remarque aux colonnes rouges, qui représente les âges réels des phrases prédites comme destinées aux adultes, alors qu'elle venait de textes pour enfants. Toutefois dans ce cas là on remarque aussi que les phrases destinées aux enfants les plus jeunes sont alors les plus susceptibles d'être confondus avec des textes pour adulte. Cela est sûrement dû à la rareté de ce type de données pour l'apprentissage des modèles, comme nous l'avons exposé en section 4.

5.4 Retour sur la sélection de descripteurs

S'il est possible d'avoir une indication sur la prédiction d'âge proprement dite, on peut aussi chercher à déterminer quels sont les critères les plus influents sur la performance du modèle. Ainsi, en prenant l'architecture la plus performante, mais en enlevant certaines informations extraites des phrases pour l'apprentissage, puis en observant l'évolution des performances nous pouvons estimer quelles sont les informations les plus pertinentes.

5.4.1 Performance en fonction des descripteurs

Pour cela nous utilisons les différentes catégories de descripteurs définies en section 3. Pour chacune de ces catégories on entraîne alors un modèle uniquement sur cette catégorie et un autre avec toutes les catégories sauf celle-ci. À partir de l'architecture BiModel, qui est la plus performante, on obtient alors des résultats contenus dans le tableau 6.

Catégorie	Uniquement	Toutes sauf
Dépendances	(4,04 ; 31,2)	(2,55 ; 29,3)
Personnes Nombres	(4,05 ; 28,9)	(2,6 ; 29,4)
Graphie	(3,89 ; 31,3)	(2,68 ; 29,4)
Occurence	(4,30 ; 31,5)	(2,59 ; 28,6)
Phonétique	(3,27 ; 30,7)	(2,47 ; 28,9)
Embeddings	(1,70 ; 18,9)	(2,44 ; 28,6)
Temps Verbaux	(4,31 ; 31,6)	(2,58 ; 29,6)
Types de Lemmes	(3,95 ; 31,2)	(2,39 ; 27,8)
Types de Connecteurs	(4,14 ; 31,5)	(2,51 ; 28,6)
Sentiments	(4,20 ; 31,4)	(2,45 ; 28,7)

Table 6: Présentation des erreurs moyennes absolues de prédiction, sur les âges minimum et maximum, des modèles appris selon une unique catégorie de descripteurs ou selon toutes ces catégories sauf une

Les résultats obtenus par ces expériences peuvent ainsi nous offrir un angle nouveau sur la manière dont on détermine l'âge associé à un texte. Ainsi, on peut discuter des critères qui semblent dicter ce choix. De plus, notre architecture faisant intervenir une classification entre les phrases venant de textes pour adultes et celles issues de textes pour enfants, on peut s'intéresser aux erreurs de classification, pour essayer de déterminer en quoi notre modèle est perfectible.

En nous intéressant aux résultats fournis par le tableau 6 nous pouvons remarquer que tous les modèles ont à peu près les mêmes performances, à l'exception du modèle appris uniquement sur les embeddings, qui a de bien meilleurs performances, même si elles restent moindres que celle du modèle cumulant toutes les informations. On peut supposer que ce critère d'embeddings, censé modéliser l'information sémantique dans une phrase est prédominant, ce qui pourrait signifier que pour une bonne recommandation de texte à un enfant le sens des mots est primordial. Toutefois, on peut aussi remarquer que ce critère est celui qui est de plus grande dimension. En effet, sur les 621 coordonnées que composent notre vecteur d'entrée du modèle, 500 sont dues aux embeddings. Ainsi la performance de ce modèle uniquement appris avec les embeddings pourrait être dues simplement à une plus grande quantité d'informations. Or, on n'observe pas les mêmes performances sur les modèles appris sur une quantité d'information aussi grande ou supérieure.

Si on peut estimer que l'ensemble des embeddings offre une information précieuse à nos modèles, on ne peut pas vraiment conclure sur sa réelle nécessité ou sur l'absence d'utilité des autres catégories. Pour pouvoir le faire, il nous faudrait plus de données afin d'avoir des jeux d'apprentissage et de tests plus conséquents et donc des différences plus significatives entre les performances des modèles, ainsi que des tests statistiques significatifs.

De plus, notre architecture nécessite que la tâche de classification soit suffisamment discriminante pour écarter des textes pour adultes du travail de régression. Or, de par la forte présence de ceux-ci (un tiers du corpus), il semble que certains de ces réseaux de classification

ne convergent pas et tendent à systématiquement classer tous les textes comme étant pour enfants, rendant la tâche de classification en amont de la régression peu ou pas utile dans certains cas. Il est alors difficile de dire si ce problème de convergence est dû à un manque de données, à certaines conditions d'initialisation ou bien bel et bien du fait des informations données en apprentissage.

Le meilleur moyen de réduire l'erreur moyenne absolue de notre modèle serait d'avoir un corpus de phrases annotées une par une avec une expertise humaine, ce qui est trop fastidieux et chronophage et n'a donc pas pu être mis en place dans le cadre de ce stage. De plus ce corpus, bien que composé de 10 000 phrases, semble est trop petit pour avoir des conclusions franches et comporte de plus quelques coquilles à cause de son extraction par OCR. Ainsi le meilleur moyen d'obtenir de meilleur résultats serait d'avoir un corpus plus important et si possible, de meilleure qualité que celui-ci. Toutefois on peut mettre en lumière l'importance des embeddings dans la prédiction d'âge, éléments absents des sujets connexes vus en section 2.

6 Conclusion

Dans ce rapport, nous avons décrit notre travail sur une problématique de TAL peu étudiée, qui est la prédiction de recommandation d'âge. En effet, si des travaux similaires peuvent exister sur le français comme langue étrangère, ou sur de la simplification de texte destiné à des enfants, on ne trouve pas dans la littérature de travaux de TAL sur ce sujet précis. De ce fait, il a été nécessaire de chercher au sein de travaux de psycholinguistiques divers moyens de déterminer en quoi un texte pouvait être recommandé à un enfant. À partir de ces travaux, nous avons établis une liste d'informations à extraire d'un texte, pour ensuite les donner à un ou plusieurs réseaux de neurones chargés d'apprendre à en déduire une recommandation d'âge.

Les performances de prédictions de ces réseaux sont correctes, en moyenne juste à 1,6 ans près. Si ce n'est pas assez pour se substituer complètement à une expertise humaine ces modèles restent une bonne indication pour déterminer si une phrase est destinée au public voulu par son auteur ou un éditeur. De plus, les différences de performances de ces modèles en fonction des informations qui leur sont données en entrée, nous invite à considérer l'apport de la sémantique et du champ lexical comme prépondérant dans la recommandation d'âge. Or, cet aspect était peu souvent mis en avant par les travaux en psycholinguistique, à l'inverse des informations lexicales et syntaxiques.

Parmi les améliorations possibles, il serait souhaitable de constituer un corpus plus vaste, notamment pour expérimenter des modèles plus complexes et également d'écarts statistiques plus significatifs entre les différentes approches, RegOnly MTL et BiModel. À cela on peut aussi souhaiter d'ajouter des annotations humaines phrases par phrases pour notre corpus, afin d'avoir une annotation un peu plus cohérente qu'ici, où l'hypothèse qu'une phrase est destinée aux personnes du même âge que le texte dont elle est issue, est une source d'erreurs de prédictions.

References

[1] L. Clément Abeillé, A. and F. Toussenel. Building a treebank for french. pages 165–187, 2003.

[2] Marc Aguert, Josie Bernicot, and Virginie Laval. Prosodie et compréhension des énoncés chez les enfants de 5 à 9 ans. *Enfance*, 2009:341, 09 2009.

[3] Nathalie Blanc and Quenette. Guy. La production d'inférences émotionnelles entre 8 et 10 ans : quelle méthodologie pour quels résultats ? *Enfance*, 4:503–511, 2017.

[4] Marie Candito, Joakim Nivre, Pascal Denis, and Enrique Henestroza Anguiano. Benchmarking of statistical dependency parsers for french. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 108–116. Association for Computational Linguistics, 2010.

[5] Denise Davidson. The role of basic, self-conscious and self-conscious evaluative emotions in children's memory and understanding of emotion. *Motivation and Emotion*, 30:232–242, 09 2006.

[6] Jan De Belder and Marie-Francine Moens. Text simplification for children. 01 2010.

[7] Marc Delahaie. *L'évolution du langage de l'enfant De la difficulté au trouble*, volume 2009. INPES, 07 2010.

[8] Elisabeth Demont and Jean Gombert. L'apprentissage de la lecture : évolution des procédures et apprentissage implicite. *Enfance*, 56, 01 2004.

[9] D.Gromer and M.Weiss. *Lire, tome 1 : apprendre à lire - Armand Colin*. 1990.

[10] Lijun Feng, Noémie Elhadad, and Matt Huenerfauth. Cognitively motivated features for readability assessment. In *EACL*, 2009.

[11] Thomas François and Cédrick Fairon. An "ai readability" formula for french as a foreign language. In *EMNLP-CoNLL*, 2012.

[12] Susan Gathercole. Cognitive approaches to the development of short-term memory. *Trends in cognitive sciences*, 3:410–419, 12 1999.

[13] L. H. Geyer. Recognition and confusion of the lowercase alphabet. *Perception & Psychophysics*, 22(5):487–490, Sep 1977.

[14] Hatem Ghorbel and David Jacot. *Sentiment Analysis of French Movie Reviews*, volume 361, pages 97–108. 01 1970.

[15] Karl Gyllstrom and Marie-Francine Moens. Wisdom of the ages: Toward delivering the children's web with the link-based AgeRank algorithm. pages 159–168, 01 2010.

[16] Firas Hmida, Mokhtar Boumedyen Billami, Thomas François, and Nuria Gala. Assisted lexical simplification for french native children with reading difficulties. In *The Workshop of Automatic Text Adaptation, 11th International Conference on Natural Language Generation*, 2018.

[17] Robert P. Jr; Rogers Richard L.; Kincaid, J. Peter; Fishburne and Brad S Chissom. "derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel". *Institute for Simulation and Training*, 02 1975.

[18] Jean Mesnager. Communication et langages, issn 0336-1500, n 79. pages 18–38, 01 1989.

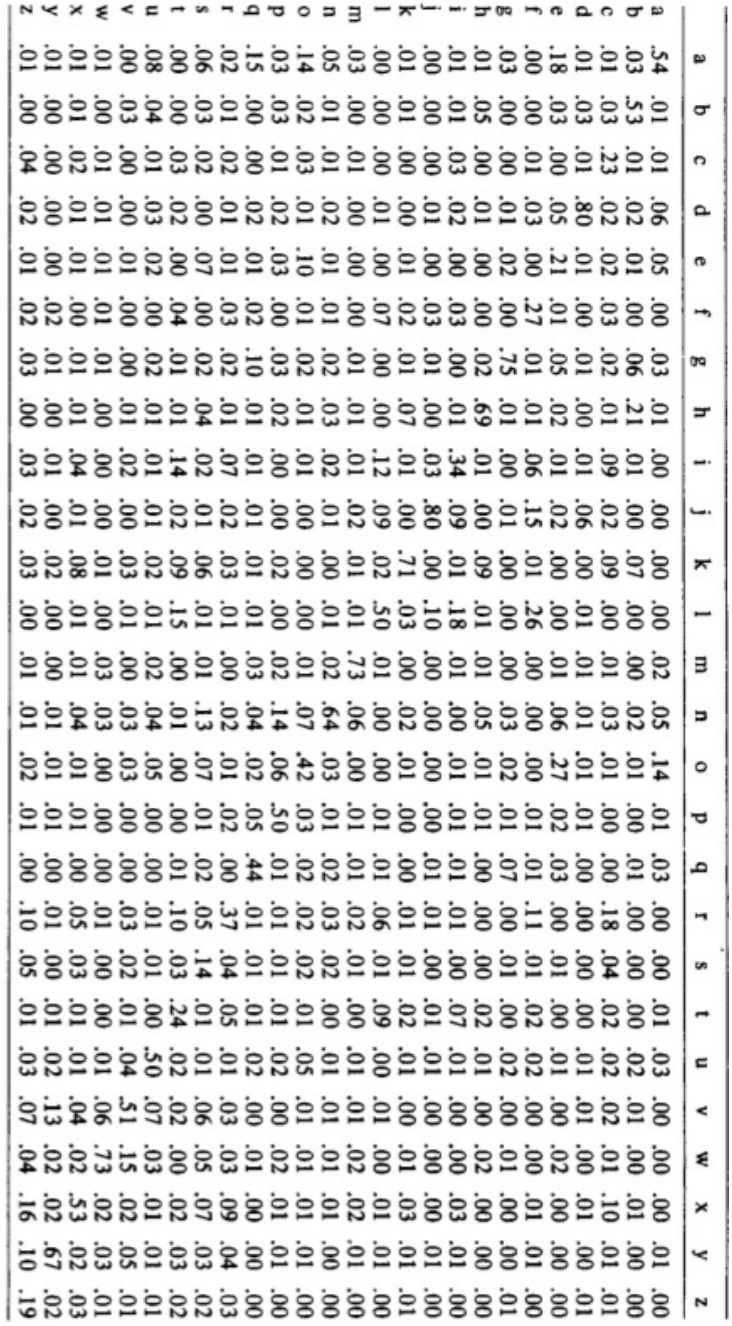
[19] Emily Pitler and Ani Nenkova. Revisiting readability: A unified framework for predicting text quality. In *Proceeding of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2008.

[20] Valérie Tartas. Le développement de notions temporelles par l'enfant. *Développements*, 4:17–26, 2010.

[21] U.Frith. Beneath the surface of developmental dyslexia. *In K. E. Patterson, J. C. Marshall, & M. Colthear (Eds.),Surface Dyslexia: Neuropsychological and Cognitive Studies of Phonological Reading*, pages 301–330, 1985.

[22] Monique Vion and Annie Colas. L'emploi des connecteurs en français : contraintes cognitives et développement des compétences narratives (le cas de la narration de séquences arbitraires d'événements). In *Conference of the International Association for the Study of Child Language*, pages 632–651, 1999.

A Annexe 1 : Tableau présentant les probabilités de reconnaissance ou de confusion des lettres de l'alphabet



	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z
a	.54	.01	.01	.06	.05	.00	.03	.01	.00	.00	.00	.00	.02	.05	.14	.01	.03	.00	.00	.01	.03	.00	.00	.00	.01	.00
b	.03	.53	.01	.02	.01	.00	.06	.21	.01	.00	.07	.00	.00	.02	.01	.00	.01	.00	.00	.00	.02	.01	.00	.01	.00	.00
c	.01	.03	.23	.02	.02	.03	.02	.01	.09	.02	.09	.00	.01	.03	.01	.00	.00	.18	.04	.02	.02	.02	.01	.10	.01	.01
d	.01	.03	.01	.80	.01	.00	.01	.00	.01	.06	.00	.01	.01	.01	.01	.01	.00	.00	.00	.00	.01	.01	.00	.01	.00	.01
e	.18	.03	.00	.05	.21	.01	.05	.02	.01	.02	.00	.00	.01	.06	.27	.02	.03	.00	.01	.00	.01	.00	.02	.00	.00	.00
f	.00	.00	.00	.01	.03	.00	.27	.01	.01	.06	.15	.01	.26	.00	.00	.01	.01	.11	.01	.02	.02	.00	.00	.01	.01	.00
g	.03	.00	.00	.01	.02	.00	.75	.01	.00	.01	.00	.00	.00	.03	.02	.01	.07	.00	.01	.00	.02	.00	.00	.01	.00	.01
h	.01	.05	.00	.01	.00	.00	.02	.69	.01	.00	.09	.01	.01	.05	.01	.01	.00	.00	.00	.02	.01	.00	.02	.00	.00	.00
i	.01	.01	.03	.02	.00	.03	.00	.01	.34	.09	.01	.18	.01	.00	.01	.01	.00	.01	.00	.07	.01	.00	.00	.03	.01	.00
j	.00	.00	.00	.01	.00	.03	.01	.00	.03	.80	.00	.10	.00	.00	.00	.00	.01	.01	.00	.01	.01	.00	.00	.00	.01	.00
k	.01	.01	.00	.00	.01	.02	.01	.07	.01	.00	.71	.03	.00	.02	.01	.00	.01	.01	.00	.02	.01	.00	.01	.03	.01	.01
l	.00	.00	.00	.01	.00	.07	.00	.00	.12	.09	.02	.50	.01	.00	.00	.01	.01	.06	.01	.09	.00	.01	.00	.01	.01	.00
m	.03	.00	.01	.00	.00	.00	.01	.01	.01	.02	.01	.01	.73	.06	.00	.01	.01	.02	.01	.00	.01	.01	.02	.02	.01	.00
n	.05	.01	.01	.02	.01	.01	.02	.03	.02	.01	.00	.01	.02	.64	.03	.01	.02	.03	.02	.00	.01	.01	.01	.01	.00	.00
o	.14	.02	.03	.01	.10	.01	.02	.01	.01	.00	.00	.00	.01	.07	.42	.03	.02	.02	.02	.01	.05	.01	.01	.01	.01	.00
p	.03	.03	.01	.02	.03	.00	.03	.02	.00	.00	.02	.00	.02	.14	.06	.50	.01	.01	.01	.01	.02	.00	.02	.01	.01	.00
q	.15	.00	.00	.02	.00	.02	.00	.01	.01	.03	.01	.01	.03	.02	.05	.44	.01	.01	.01	.01	.02	.00	.01	.00	.00	.00
r	.02	.01	.02	.01	.01	.03	.02	.01	.07	.02	.03	.01	.00	.02	.01	.02	.00	.37	.04	.05	.01	.03	.03	.09	.04	.03
s	.06	.03	.02	.00	.07	.00	.02	.04	.02	.01	.06	.01	.01	.13	.07	.01	.02	.05	.14	.01	.01	.06	.05	.07	.03	.02
t	.00	.00	.03	.02	.00	.04	.01	.01	.14	.02	.09	.15	.00	.01	.00	.00	.01	.10	.03	.24	.02	.02	.00	.02	.03	.02
u	.08	.04	.01	.03	.02	.00	.02	.01	.01	.01	.02	.01	.02	.04	.05	.00	.00	.01	.01	.00	.50	.07	.03	.01	.01	.01
v	.00	.03	.00	.00	.01	.00	.00	.01	.02	.00	.03	.01	.00	.03	.03	.00	.00	.03	.02	.01	.04	.51	.15	.02	.05	.01
w	.01	.00	.01	.01	.01	.01	.01	.00	.00	.00	.01	.00	.03	.04	.00	.00	.00	.01	.00	.00	.01	.06	.73	.02	.03	.01
x	.01	.01	.02	.01	.01	.00	.01	.01	.04	.01	.08	.01	.01	.01	.01	.00	.00	.05	.03	.01	.01	.04	.02	.53	.02	.03
y	.01	.00	.00	.00	.00	.01	.00	.01	.04	.01	.01	.00	.01	.01	.01	.00	.00	.01	.00	.01	.02	.13	.02	.02	.67	.02
z	.01	.00	.04	.02	.01	.02	.03	.00	.03	.02	.03	.00	.01	.01	.00	.01	.00	.10	.05	.01	.03	.07	.04	.16	.10	.19

Figure 8: Tableau présentant les probabilités de reconnaissance ou de confusion des lettres de l'alphabet [13]

B Annexe 2 : Histogramme du nombre de phrases recouvrant les âges de 1 à 100.

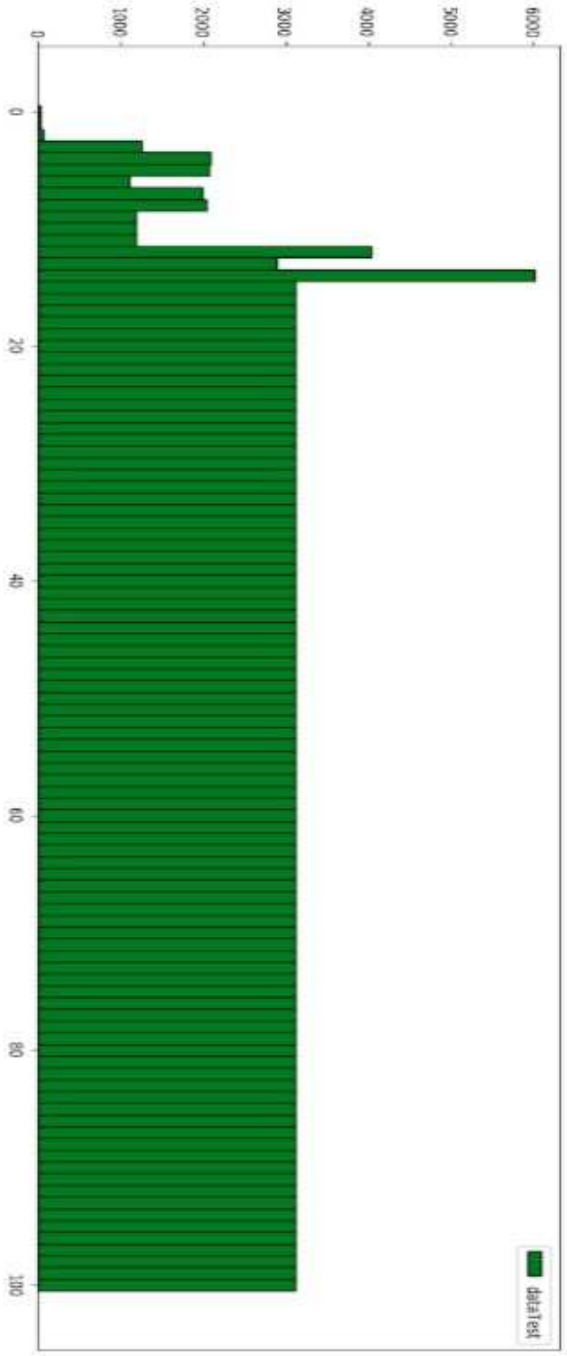


Figure 9: Histogramme du nombre de phrases recouvrant les âges de 1 à 100.

C Annexe 3 : Description de l'ensemble des descripteurs

Ici nous allons décrire l'ensemble des descripteurs retenus pour les modèles décrits dans ce rapport. Chaque liste correspond à une catégorie, avec entre parenthèses la dimension du vecteur retenu pour ces descripteurs.

METADONNÉES (4) :

- Titre;
- Âge miniumum;
- Âge maximum;
- Phrase;

DÉPENDANCE (7) :

- Profondeur de l'arbre de dépendance;
- Moyenne des distances (nombre de mot) entre un mot et ses dépendances;
- Maximum des distances (nombre de mot) entre un mot et ses dépendances;
- Moyenne de la distance entre un mot et ses pointeurs;
- Variance de la distance entre un mot et ses pointeurs;
- Nombre Moyen de dépendances (mots qui pointent vers un mot donné);
- Variance du nombre de dépendances;

PERSONNES ET NOMBRES (5) :

- Proportion de verbes conjugués à la 1ère personne;
- Proportion de verbes conjugués à la 2ème personne;
- Proportion de verbes conjugués à la 3ème personne;
- Proportion de verbes conjugués au singulier;
- Proportion de verbes conjugués au pluriel;

GRAPHIE (9) :

- Nombre de mots dans la phrase;
- Moyenne de la fréquence des mots de la phrase, calculée comme descripteur local ;
- Variance de la fréquence des mots de la phrase, calculée comme descripteur local ;

- Moyenne de la confusion des mots de la phrase, calculée comme descripteur local ;
- Variance de la confusion des mots de la phrase, calculée comme descripteur local ;
- Moyenne de la taille des mots de la phrase, calculée comme descripteur local ;
- Variance de la taille des mots de la phrase, calculée comme descripteur local ;
- Proportions du nombre de symboles (lettres et ponctuations) sur le nombres de mots ;
- Rapport, nombre de ponctuation sur le nombre de mots ;

OCCURRENCE (2) :

- Pour chaque mot de la phrase, on y associe une log probabilité de son utilisation dans la langue. À partir de ces log probabilités sur les mots d'une phrase, on calcul une moyenne et une variance.

PHONÉTIQUE (9):

- Moyenne et variance des scores de fréquences des phonèmes composant les mots de la phrase, ces scores basés sur les phonèmes les plus fréquents dans la langue;
- Moyenne et variance de la diversité des phonèmes composant les mots de la phrase;
- Moyenne et variance du nombre de phonèmes composant les mots de la phrase;
- Score de fréquence des phonèmes composant la prononciation de la phrase;
- Diversité des phonèmes dans la proposition de la phrase;
- Nombre de phonèmes dans la phrase;

EMBEDDING (500) :

- Moyenne des Embeddings composants la phrase, pour chacunes des 500 coordonnées.

TEMPS VERBAUX (24) :

- Diversité des temps verbaux utilisés;
- Proportions de 7 temps simples (présent, passé simple, futur, imparfait, subjonctif présent, conditionnel présent, infinitif) dans la phrase;
- Proportions de 7 temps composés (passé composé, passé antérieur, futur antérieur, plus que parfait, subjonctif passé, conditionnel passé, infinitif passé) dans la phrase;
- Diversité des Système temporels dans la phrase parmi les 3 possibles (passé, présent, futur);

- Proportion des verbes conjugués dans un des 3 systèmes temporelles (passé, présent, futur);
- Proportion des temps composés;
- Proportion des temps simples;
- Proportion des modes (infinitif, indicatif, subjonctif);

TYPES DE LEMMES (8) :

- Diversité des Lemmes dans la phrase;
- Proportion des Verbes dans la phrase;
- Proportion des Noms dans la phrase;
- Proportion des Adjectifs dans la phrase;
- Proportion de Clitiques dans la phrase;
- Proportion de StopWords dans la phrase;
- Proportion des adverbes temporels dans la phrase selon une liste prédéfinie;
- Proportion des verbes d'état dans la phrase selon une liste prédéfinie;

TYPES DE CONNECTEURS LOGIQUES (17) :

- Proportion des connecteurs logiques d'addition selon une liste prédéfinie;
- Proportion des connecteurs logiques de but selon une liste prédéfinie;
- Proportion des connecteurs logiques de cause selon une liste prédéfinie;
- Proportion des connecteurs logiques de comparaison selon une liste prédéfinie;
- Proportion des connecteurs logiques de concession selon une liste prédéfinie;
- Proportion des connecteurs logiques de conclusion selon une liste prédéfinie;
- Proportion des connecteurs logiques de condition selon une liste prédéfinie;
- Proportion des connecteurs logiques de conséquence selon une liste prédéfinie;
- Proportion des connecteurs logiques d'énumération selon une liste prédéfinie;
- Proportion des connecteurs logiques d'explication selon une liste prédéfinie;
- Proportion des connecteurs logiques d'illustration selon une liste prédéfinie;

- Proportion des connecteurs logiques de justification selon une liste prédéfinie;
- Proportion des connecteurs logiques d'opposition selon une liste prédéfinie;
- Proportion des connecteurs logiques de restriction selon une liste prédéfinie;
- Proportion des connecteurs logiques d'exclusion selon une liste prédéfinie;
- Proportion des connecteurs logiques de temps selon une liste prédéfinie;

SENTIMENT (26) :

- Score de subjectivité de la phrase;
- Score de polarité de la phrase;
- Proportion des 24 émotions dans la phrase parmi la liste suivante : neutre, admiration, amour, apaisement, audace, colère, comportement, culpabilité, dégoût, déplaisir, désir, embarras, empathie, fierté, impassibilité, inhumanité, jalousie, joie, mépris, non spécifiée, orgueil, peur, ressentiment, surprise et tristesse;