Une chaîne de traitements pour prédire et appréhender la complexité des textes pour enfants d'un point de vue linguistique et psycho-linguistique.

Delphine Battistelli, Aline Etienne, Rashedur Rahman, Charles Teissèdre, Gwénolé Lecorvé

Une chaîne de traitements pour appréhender la complexité des textes pour enfants d'un point de vue linguistique et psycho-linguistique

Delphine Battistelli[1] Aline Etienne[1] Rashedur Rahman[2] Charles Teissèdre[3]
Gwénolé Lecorvé[4]
(1) Univ. Paris-Nanterre, CNRS, MoDyCo, 200, avenue de la République, 92001 Nanterre, France
(2) Univ Rennes, CNRS, IRISA, 6, rue de Kerampont, 22300 Lannion, France
(3) Synapse Développement, 7, boulevard de la Gare, 31500 Toulouse, France
(4) Orange Innovation, 2, avenue Pierre Marzin, 22307 Lannion, France
delphine.battistelli@parisnanterre.fr, aline.etienne@parisnanterre.fr,
md-rashedur.rahman@irisa.fr, charles.teissedre@synapse-fr.com,
gwenole.lecorve@orange.com

RÉSUMÉ

Nos travaux abordent la question de la mesure de la complexité d'un texte vis-à-vis d'une cible de lecteurs, les enfants en âge de lire, au travers de la mise en place d'une chaîne de traitements. Cette chaîne vise à extraire des descripteurs linguistiques, principalement issus de travaux en psycholinguistique et de travaux sur la lisibilité, mobilisables pour appréhender la complexité d'un texte. En l'appliquant sur un corpus de textes de fiction, elle permet d'étudier des corrélations entre certains descripteurs linguistiques et les tranches d'âges associées aux textes par les éditeurs. L'analyse de ces corrélations tend à valider la pertinence de la catégorisation en âges par les éditeurs. Elle justifie ainsi la mobilisation d'un tel corpus pour entraîner à partir des âges éditeurs un modèle de prédiction de l'âge cible d'un texte.

ABSTRACT

A Processing Chain to Explain the Complexity of Texts for Children From a Linguistic and Psycho-linguistic Point of View

Our work addresses the issue of measuring the complexity of a text with respect to a target group of readers, namely children of reading age, through the implementation of a processing chain. This chain aims at extracting linguistic descriptors, mainly from psycholinguistics and readability studies, that can be used to understand and describe the complexity of a text. By applying it to a corpus of fiction texts, it allows us to study the correlations between various linguistic descriptors and the age ranges associated with the texts by the editors. The analysis of these correlations tends to validate the relevance of the age categorization by the editors. It thus justifies the mobilization of such a corpus to train a prediction model to recommend a text to the children of a target age on the basis of the editors' ages.

MOTS-CLÉS : complexité d'un texte, âge, descripteurs linguistiques, étapes développementales.

KEYWORDS: text complexity, age, linguistic descriptors, developmental stages.

1 Introduction

La mesure de la complexité d'un texte vis-à-vis d'une cible de lecteurs, les enfants en âge de lire, revêt un caractère éminemment pluridisciplinaire, à la croisée de la linguistique, de l'informatique et de la psycholinguistique. Le présent travail prend part au projet ANR TextToKids [1] consacré à la compréhension de textes par les enfants sans trouble du développement, approximativement âgés de 6 à 14 ans, et au sein duquel la question de la complexité d'un texte et des critères pour en décider occupe donc une place centrale.

Nous appréhendons ici cette question comme l'attribution à un texte d'un âge (ou d'une tranche d'âge) en dessous duquel (de laquelle) le texte posera des difficultés de compréhension. Il s'agit ainsi de pouvoir discriminer des seuils de complexité correspondant à des étapes développementales (et donc à des classes d'âge). Nous nous appuyons pour cela sur des travaux psycholinguistiques étudiant la compréhension de textes et de caractéristiques linguistiques précises (par ex. phrases relatives, informations temporelles calendaires, etc.) par les enfants et sur des travaux de lisibilité. De ces travaux, nous tirons un ensemble de descripteurs linguistiques (phonétiques, morphologiques, syntaxiques, sémantiques, discursifs) auxquels des seuils de difficulté sont associés (par ex. les phrases relatives en QUE sont difficiles pour les enfants de moins de 10 ans). Ces descripteurs interviennent dans la description de textes et la mesure de leur complexité.

Nous présentons ici une chaîne de traitements dont le but est d'objectiver l'analyse de la complexité d'un texte destiné aux enfants jeunes lecteurs, en extrayant des descripteurs linguistiques identifiés en psycholinguistique et en lisibilité. En appliquant la chaîne sur un corpus de 1130 textes de fiction, pour lesquels des recommandations d'âge sont indiquées par les éditeurs (par exemple, les ouvrages de la collection Chien Pourri sont recommandés à partir de 6-8 ans par L'École des loisirs), nous mettons en œuvre le protocole d'analyse de la complexité qu'offre la chaîne pour répondre à deux questions plus spécifiques :

1. Peut-on étayer la pertinence supposée des tranches d'âge fournies par les éditeurs de contenus-jeunesse, sachant qu'elles répondent notamment parfois à des critères commerciaux ?
2. Peut-on prédire automatiquement un âge (ou une tranche d'âge) minimal(e) à associer préférablement à la lecture d'un texte, à partir des âges éditeurs ?

Nous présentons tout d'abord en section 2 les travaux de lisibilité et de psycholinguistique sur lesquels nous prenons appui pour identifier des descripteurs linguistiques permettant d'appréhender la complexité des textes en fonction d'âge cible. La chaîne que nous avons développée pour extraire ces descripteurs est ensuite décrite en section 3. Elle constitue un cadre pour explorer la question de la complexité d'un texte en donnant la possibilité de mettre en œuvre diverses expérimentations informatiques à partir de classes de descripteurs linguistiques. En section 4, nous appliquons notre chaîne sur un corpus de textes de fiction catégorisés en âge par les éditeurs, afin de tester la pertinence des recommandations d'âge faites par les éditeurs. Plus précisément, nous étudions les corrélations entre les descripteurs linguistiques extraits par notre chaîne et les âges donnés par les éditeurs, et mettons ainsi en regard les âges éditeurs avec les normes développementales issues de travaux psycholinguistiques. Le corpus de texte fiction est ensuite mobilisé pour entraîner un modèle de prédiction de l'âge cible des textes présenté en section 5.

1. https://texttokids.irisa.fr/

2 Étudier la complexité linguistique d'un texte destiné aux enfants

2.1 Etudier la complexité linguistique d'un texte

En TAL, la thématique de la complexité d'un texte est à situer principalement du côté des travaux en lisibilité et en simplification de textes, les seconds s'appuyant volontiers sur les premiers. L'évaluation de la lisibilité peut en effet jouer un rôle autant dans l'évaluation d'un algorithme de simplification, que dans l'identification de ce qu'il faut simplifier. Dans la continuité des travaux pionniers décrits dans (Dale & Chall, 1948), ces travaux s'intéressent globalement à catégoriser des textes en niveaux de complexité en s'appuyant sur des descripteurs linguistiques et à proposer éventuellement des reformulations. Il y est ainsi courant de recourir à des descripteurs aisément calculables tels que la longueur moyenne d'une phrase (vue comme une approximation pour la complexité syntaxique), la longueur moyenne des mots en syllabes ou en caractères (vue comme une approximation de la complexité au niveau lexical) ou encore la présence de mots rares. En lisibilité comme en simplification (domaines qui ont tendance à fortement se recouvrir ces dernières années en TAL), les cas d'usage sont le plus souvent liés à l'apprentissage d'une langue seconde (François & Fairon, 2012), à la compréhension de textes en domaine spécialisé, comme par exemple le domaine médical (Cardon & Grabar, 2021) ou encore au contexte pédagogique de choix de textes adaptés à des enfants du niveau de l'école primaire, par ex. (Gala *et al.*, 2020; Imperial & Ong, 2021). Ce qui nous différencie en termes méthodologiques des travaux précédemment cités concerne principalement le choix de faire intervenir des descripteurs linguistiques clairement motivés sur le plan développemental selon des travaux en psycholinguistique sur la tranche des enfants jeunes lecteurs. À notre connaissance aucune démarche de ce type n'a en effet été proposée dans les domaines de la lisibilité et de la simplification de textes à destination d'enfants. Qui plus est, ces descripteurs sont pour certains de nature sémantique et jusqu'ici peu - voire pas - utilisés dans les chaînes de traitement comparables.

2.2 Décrire la complexité d'un texte à destination des enfants

La description de la complexité d'un texte à destination des enfants est à aborder par le prisme de l'évolution des compétences linguistiques de compréhension. De ce point de vue, les études psycholinguistiques mettent en lumière des marqueurs linguistiques représentant des points de difficultés potentielles pour les enfants. Elles donnent de plus une indication globale des âges auxquels ces difficultés peuvent se présenter. (Tartas, 2001, 2010) indique par exemple que, jusqu'à 9 ans environ, les enfants rencontrent des difficultés à manipuler les notions calendaires comme les jours, les mois ou encore les saisons. (Davidson, 2006) et (Blanc & Quenette, 2017) dégagent quant à eux une norme développementale concernant la maîtrise des émotions. Plus specifiquement, ils observent une meilleure compréhension des émotions de base (ex. joie, colère, etc.) que des émotions complexes (ex. embarras, culpabilité, etc.) lorsque l'émotion est directement désignée par un terme du lexique émotionnel (ex. joyeux, honteux, etc.), et ce jusqu'à 10-11 ans.

Les grandes tendances développementales identifiées sont bien sûr à pondérer du fait de la difficulté à effectuer des associations strictes entre un marqueur linguistique et un âge auquel ce marqueur est compris. Cette difficulté s'explique en premier lieu par les différences inter-individuelles de développement des compétences linguistiques. En second lieu, des différences du point de vue des matériaux linguistiques (énoncés isolés, textes entiers, etc.) et des procédés d'évaluation (réponse orale à une question, choix d'une image, etc.) employés pour mesurer les capacités de compréhension

chez les enfants engendrent parfois des estimations divergentes des âges auxquels certains marqueurs linguistiques sont maîtrisés (van den Broek, 1997). Malgré ces réserves, il reste que les descripteurs psycholinguistiques (tels que l'expression du repérage temporel ou l'expression des émotions par exemple) sont utiles pour échelonner la complexité des textes à destination des enfants du point de vue d'âges frontières, et non plus du simple point de vue d'une opposition entre simple et complexe, comme dans (Gala *et al.*, 2018).

3 Description de la chaîne d'extraction de descripteurs

La chaîne d'extraction que nous présentons ici permet d'extraire des descripteurs linguistiques en lien avec la complexité des textes, exploités ensuite pour mesurer des corrélations entre tranches d'âge telles que fournies par les éditeurs et descripteurs linguistiques.

Bien que largement remanié, le noyau originel de la chaîne de traitements est constitué d'un ensemble de scripts d'extraction automatique de descripteurs développés dans le cadre des travaux de (Blandin *et al.*, 2020) qui visaient également à extraire des descripteurs linguistiques mobilisables dans l'analyse de la complexité des textes à destination des enfants. Notre chaîne s'appuie donc par exemple toujours sur un lexique émotionnel adapté de EMOTAIX (Piolat & Bannour, 2009) pour extraire des descripteurs liés à l'expression des émotions. Elle a été mise à jour à l'issue d'un travail d'analyse linguistique visant à affiner le choix des descripteurs extraits. La chaîne remaniée permet désormais d'extraire de nouveaux descripteurs, relevant :

— de l'analyse lexicale (proportion de mots appartenant aux différents niveaux de l'échelle Dubois-Buyse (Ters *et al.*, 1970)) ;
— de l'analyse de constructions syntaxiques particulières, réalisée à l'aide de patrons morphosyntaxiques et syntaxiques (proportion de tournures à la voix passive, proportion de subordonnées relatives typées selon la nature de leur introducteur) ;
— ou encore de l'analyse sémantique effectuée à partir de transducteurs (proportion de duratifs, d'adverbiaux de localisation temporelle (Teissèdre *et al.*, 2010), de métaphores (Blanchard *et al.*, 2001)).

Certains descripteurs de la chaîne initiale ont à l'inverse été écartés, en particulier des descripteurs qui paraissent insuffisamment motivés du point de vue linguistique ou psycho-linguistique (par exemple, la proportion de "stop words" ou la proportion de lemmes différents). Enfin, l'organisation de certains descripteurs et les listes servant à leur repérage ont été mises à jour et étoffées (connecteurs de type argumentatif et concessif, connecteurs temporels, connecteurs explicatifs et justificatifs).

La chaîne mise en œuvre mobilise la librairie open source STANZA, une chaîne neuronale permettant entre autres d'effectuer une analyse morphosyntaxique et syntaxique des textes(Qi *et al.*, 2020). Le choix de cette librairie est motivé par deux principaux éléments : elle permet d'effectuer une analyse des textes dans une architecture de traitement homogène et également d'ajouter des attributs à chaque niveau d'analyse à différentes échelles (token, mot, phrase et document).

Les traitements opérés dans le cadre de nos travaux s'appliquent à deux niveaux d'analyse : au niveau du document (par exemple la proportion de verbes au présent dans l'ensemble d'un texte) et au niveau de la phrase (par exemple la proportion de mots appartenant au vocabulaire de l'échelle Dubois-Buyse répartis par niveaux scolaires).

La Table 1 présente l'ensemble des classes de descripteurs (une classe peut contenir un ou plusieurs descripteurs) correspondant à différents niveaux d'analyse de la langue que nous avons choisis

d'intégrer dans notre chaîne avec les indications dont nous disposons en termes de classes d'âge. Ces classes de descripteurs sont issus des domaines de la lisibilité ou de la simplification (noté L), notamment (François & Fairon, 2012; Gala *et al.*, 2018; Elguendouze, 2020; De Belder & Moens, 2010), et de la psycholinguistique (noté P), notamment (Lecocq, 1998; Hickmann *et al.*, 1993; Hickmann, 2012; Tamine & Bonnet, 1982; Bassano, 1985; Tartas, 2001, 2010; Blanc & Quenette, 2017; Davidson, 2006). Nous leur avons adjoint un petit ensemble de classes de descripteurs à caractère exploratoire (noté E). On notera une part importante accordée à des classes de descripteurs de nature sémantique (30 sur un total 66 classes de descripteurs) et principalement issues de la littérature psycholinguistique.

Niveau	**Domaine**	**Indication**	**Nb classes descripteurs**	**Exemples**
Phonétique	L	pas d'indic.	2	longueur du mots en phonèmes
Morphologie	L	pas d'indic.	1	fréquence, dans la langue, des morphèmes composant les mots
	P	< 9 ans	1	marques du pluriel
Morphosyntaxe	E	pas d'indic.	2	nombre de temps verbaux différents
	L	pas d'indic.	7	proportions de parties du discours, portions de verbes fléchis aux 1ère, 2e et 3e personnes
Lexique	E	du CP au lycée	1	nombre de mots d'un niveau scolaire donné selon l'échelle Dubois-Buyse
	L	pas d'indic.	3	diversité des lemmes
	P	< 10 ans	1	adjectifs ordinaux
Syntaxe	E	pas d'indic.	3	nombre moyen de dépendants pour un mot
	L	pas d'indic.	2	longueur des phrases
	P	<9 ans	7	relatives en QUI, structures passives, superlatifs d'infériorité
		< 10 ans	4	relatives en DONT, relatives en gérondif
		<12+ ans	1	relatives en QUE
Sémantique	E	pas d'indic.	2	entités nommées
	L	pas d'indic.	3	liens logiques explicites ou non
	P	<9 ans	13	informations temporelles de type heure, jour, mois, informations temporelles cycliques, mots polysémiques
		< 10 ans	5	émotions complexes, émotions suggérées par une situation
		<12+ ans	7	métaphores, subordonnées hypothétiques

TABLE 1 – Descripteurs linguistiques de complexité

4 Expérimentations sur un corpus de fictions

4.1 Des textes étiquetés en classes d'âge par les éditeurs

De plus en plus de contenus textuels présentés comme adaptés à de jeunes publics, selon des critères parfois intuitifs, sont disponibles. Citons par exemple en France les encyclopédies Wikimi et Vikidia, le journal Albert ou les nombreuses collections d'ouvrages de fiction. À ces textes, les éditeurs associent des recommandations d'âge, par exemple la tranche 6-8 ans pour la collection Chien Pourri ou la tranche 9-14 ans pour le journal Albert. Cette offre de plus en plus étoffée et diversifiée renouvelle en partie la méthodologie scientifique pour aborder la question de la mesure de la complexité des textes : peut-on en effet tirer parti de cette grande quantité de corpus de textes, catégorisés en classes d'âge par les éditeurs, et des techniques du TAL pour espérer mettre au jour des caractéristiques linguistiques propres à décrire un texte (ou une portion de texte) en termes d'adéquation à une tranche d'âge donnée ? C'est la voie explorée ici.

Dans (Rahman *et al.*, 2020), les auteurs avaient déjà fait le choix d'un corpus catégorisé en tranches

	Nb textes	Nb phrases	Tranches d'âges éditeur
Train	746	41 672	[0-2], [0-3], [3-5], [5-7], [6-8], [7-11], [8-10], [8-11], [8-13], [10-12], [10-13], [10-14], [11-13], [12-14], [14-18]
Dev	196	10 888	[0-3], [3-5], [4-6], [6-8], [8-11], [10-12], [14-18]
Test	188	11 096	[0-3], [2-4], [3-5], [4-6], [6-8], [7-11], [8-11], [10-12], [10-14], [14-18]

TABLE 2 – Caractéristiques du corpus

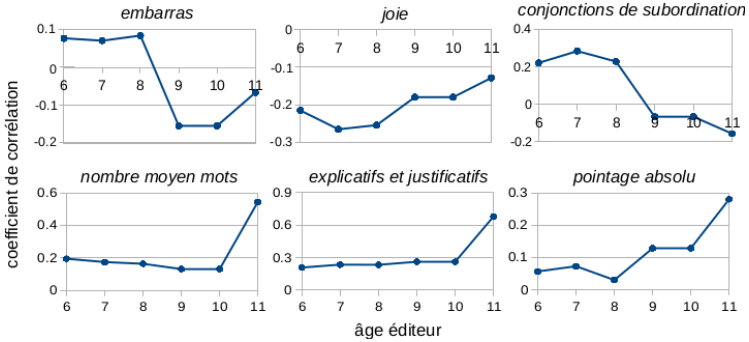

FIGURE 1 – Évolution de la corrélation entre plusieurs descripteurs et l'âge éditeur

d'âge telles que fournies par les éditeurs pour entraîner un modèle à même de prédire l'âge minimal adéquat pour comprendre un texte. Pour nos expérimentations, nous ne mobilisons qu'une partie de ce corpus en ne retenant que les œuvres de fiction, publiées par des éditeurs professionnels. Nous avons en effet fait le choix d'un corpus homogène en genre, cherchant par là à nous prémunir des effets de genre sur la lisibilité soulignés dans (Sheehan *et al.*, 2008) ou dans (Flor *et al.*, 2013).

Afin de limiter le biais sur la longueur des textes pour l'apprentissage des modèles de prédiction de l'âge cible (ex. association des textes longs à un âge élevé), les textes de plus de 10K caractères ont été découpés en segments d'environ 5000 caractères, en respectant les frontières de paragraphes. Le corpus se compose ainsi de 1130 textes de fiction (env. 70,8K phrases), correspondant soit à une portion d'ouvrage, soit à un ouvrage entier. Nous associons à chacun une tranche d'âge, selon les recommandations des éditeurs. 17 tranches d'âges, de [0-3] à [14-18], sont représentées. Le corpus a été partitionné de manière aléatoire (cf. Table 2), avec pour seule contrainte le fait que tous les textes issus d'un même ouvrage devaient se trouver dans le même ensemble. L'évaluation de la prédiction (Section 5) et les mesures de corrélation (Section 4.2) ont été effectuées sur l'ensemble de test.

4.2 Évolution des corrélations entre âge éditeur et descripteurs linguistiques

En observant la façon dont évoluent les corrélations au fil de l'âge entre âge éditeur et descripteurs linguistiques, il devient possible de dégager des pics de corrélation autour d'un âge, particulièrement saillants pour certains descripteurs. Ces pics de corrélation permettent de déterminer des "âges frontières" qui devraient correspondre, si les âges éditeurs sont cohérents, à des étapes développementales telles que dégagées par les travaux des psycho-linguistes.

De telles variations de corrélation apparaissent pour plusieurs descripteurs d'ordre sémantique. Par exemple, l'évolution des mesures de corrélation pour le descripteur *pointage absolu* (cf. Figure 1), qui correspond à la proportion d'adverbiaux temporels calendaires absolus (ex. *au cours du 19è siècle*) rapporté au nombre de mots du texte, présente un pic positif à partir des âges éditeur 9-10 ans. Cela indique que ce descripteur est opératoire pour identifier les textes considérés comme

adaptés aux enfants de 9 ans - et plus - par les éditeurs. Cette observation est en adéquation avec les savoirs psycholinguistiques concernant la compréhension des notions calendaires (cf. Section 2.2). L'évolution des corrélations pour le descripteur *explicatifs et justificatifs* (proportion de connecteurs de type *car*, *puisque*, *voilà*, *pourquoi*) comporte également un pic positif vers 10 ans. Des travaux en lisibilité recourent à ce descripteur pour traduire le fait qu'un lien logique non explicite est plus complexe qu'un lien logique explicite (Gala *et al.*, 2018). Une association plus forte entre ce descripteur et les jeunes âges est donc attendue mais le contraire est observé, ce qui reflète une difficulté potentielle de compréhension sur ce critère. Enfin, dans la Figure 1, les descripteurs *joie* et *embarras* renvoient à la proportion de termes appartenant au lexique de ces deux émotions (cf. Section 3). Ces deux descripteurs rendent compte de la norme développementale observée en psycholinguistique, selon laquelle les émotions de base comme la joie sont mieux comprises que les émotions complexes comme l'embarras, et ce jusqu'à 10-11 ans (cf. Section 2.2). L'évolution de la corrélation du descripteur *joie* montre une présence plus prononcée de cette émotion dans les textes pour les enfants plus jeunes (8 ans et moins), ce qui concorde avec les attentes de la norme développementale. Le descripteur *embarras* apparaît également plus corrélé avec les âges éditeurs inférieurs à 9 ans. Nous observons donc cette fois une divergence entre les recommandations éditeurs et celles pouvant être dégagées de la littérature psycholinguistique.

Les variations de corrélations selon les âges éditeurs sont aussi observées sur des descripteurs relevant d'autres niveaux linguistiques. L'évolution de la corrélation du descripteur syntaxique *nombre moyen de mots* augmente fortement à partir de 10 ans. Ce descripteur, qui se rapporte au nombre moyen de mots par phrase dans le texte, est en lien avec un descripteur très fréquemment mobilisé en lisibilité : la longueur de la phrase, sachant qu'une phrase plus longue est jugée plus complexe (Elguendouze, 2020). Le pic positif de corrélation observé avec l'âge éditeur le plus élevé de notre corpus convient donc aux attentes pour ce descripteur. L'évolution de la corrélation du descripteur morphosyntaxique *conjonctions de subordination*, qui correspond au pourcentage des conjonctions sur le nombre total de mots, présente un pic positif aux alentours de 7 ans. Ce descripteur est inspiré de travaux de lisibilité qui mobilisent comme descripteur de complexité les proportions de différentes parties du discours (François & Fairon, 2012). Le pic de corrélation positive observé indique que ce critère permet de classer plus efficacement les textes autour des âges frontière 6-8.

Si quelques divergences sont observées, la catégorisation en âges proposée par les éditeurs semblent malgré tout globalement cohérente avec les attentes correspondant aux normes développementales issues de la littérature psycholinguistique. Il semble donc pertinent de s'appuyer sur un corpus de textes classés en âges éditeurs pour entraîner un modèle chargé de prédire l'âge adéquat pour comprendre un texte.

5 Modèle pour la prédiction d'un âge cible

Nous considérons ici la tâche de prédiction de l'âge adéquat pour comprendre un texte comme une tâche de régression visant à prédire deux valeurs réelles : les bornes de la tranche d'âge à associer au texte. Une tranche d'âge est alors définie par un intervalle $[x, y]$ de borne inférieure x et de borne supérieure y. Le modèle de prédiction de l'âge est entraîné sur l'ensemble du corpus décrit en Section 4.1, sans faire usage des descripteurs linguistiques extraits par la chaîne (cf. Section 3). Il prend en entrée une phrase et repose sur deux modèles distincts pour prédire la tranche d'âge : l'un pour prédire x et l'autre pour y. Pour prédire l'âge au niveau du texte, les prédictions au niveau de la phrase sont

agrégées entre elles, en calculant la moyenne des âges prédits pour chacune des phrases d'un texte.

Les modèles prédisant les bornes *x* et *y* utilisent le modèle pré-entraîné *CamemBERT* (Martin *et al.*, 2020) qui permet d'effectuer des tâches de régression à sortie unique. *CamemBERT* est un modèle de langue contextuel du français qui repose sur une architecture de type RoBERTa (Liu *et al.*, 2019), comprenant un codage positionnel et un masquage multi-têtes avec un mécanisme d'attention pour apprendre un modèle de représentation de la langue française. Ce modèle a appris de manière bidirectionnelle les relations contextuelles entre les mots d'une séquence textuelle et ce à chacune des positions des mots contrairement à l'apprentissage séquentiel des architectures de type RNN. Le modèle *CamemBERT* fournit également le vecteur contextuel d'un mot ou d'une phrase (moyenne des vecteurs-mots dans une phrase). Pour obtenir le résultat optimal et confirmer la stabilité de la prédiction de l'âge, les modèles *CamemBERT* sont ajustés avec plusieurs hyper-paramètres sur les ensembles de données d'apprentissage et de validation à l'aide de *Ktrain*. Les modèles sont calibrés avec les hypers-paramètres suivants : *learning_rate=1e-5, max_sentence_len=100, batch_size=32, epoch=3*. Une fois ces valeurs fixées, les modèles sont entraînés selon la méthode *fit_one_cycle* pour prédire respectivement les limites inférieure et supérieure d'une tranche d'âge.

Pour évaluer les performances du modèle de prédiction de l'âge, un âge moyen cible T = (*x*+*y*)/2 est dérivé des deux bornes de la tranche d'âge éditeur cible. Ensuite, nous mesurons l'erreur absolue (AE), soit la différence absolue entre l'âge éditeur T et l'âge prédit T'. L'erreur est égale à zéro lorsque T et T' sont identiques. Nous calculons l'erreur absolue moyenne (MAE) [2] pour évaluer globalement le modèle : sur le corpus de test, à l'échelle du texte, la MAE entre T et T' est de 2,02. Ceci signifie que l'écart entre les âges prédits et les âges éditeurs moyens est de 2 ans. Ce niveau d'erreur est comparable à celui de 2,09 observé dans (Blandin *et al.*, 2020) entre l'âge fourni par les éditeurs et l'âge prédit par leur modèle, entraîné sur un corpus composé de textes de trois genres différents (journalistique, encyclopédique et de fiction).

6 Conclusion

La chaîne de traitements que nous avons présentée ici, testable en ligne [3], extrait des descripteurs linguistiques issus de travaux en psycholinguistique et lisibilité/simplification, susceptibles d'être mobilisés pour décrire la complexité d'un texte à différents niveaux (lexique, syntaxe, sémantique, etc.). À travers l'application de la chaîne à un corpus de textes de fiction, nous avons pu observer des corrélations entre classes d'âge éditeurs et descripteurs de nature sémantique (descripteurs de temps, d'émotions et d'organisation discursive). En s'appuyant sur l'analyse de ces corrélations, nous pouvons identifier des critères linguistiques discriminants pour classer les textes selon les âges éditeurs (par exemple, le descripteur *pointage absolu* permet de distribuer les textes autour d'un âge frontière de 9 ans). La mise en regard de ces corrélations avec les normes développementales issues de la psycholinguistique (qui relèvent, par exemple, que les enfants de moins de 9 ans ont du mal à appréhender les notions temporelles absolues) permet d'étayer la pertinence des âges éditeurs. Nous proposons ainsi d'exploiter les corpus catégorisés en âge par les éditeurs pour entraîner des modèles de prédiction de l'âge cible d'un texte. À partir de notre chaîne d'extraction de descripteurs linguistiques et du modèle de prédiction d'un âge cible, il devient envisageable de procéder à une analyse de la complexité d'un ouvrage entier. Cela impliquerait de pouvoir quantifier les portions textuelles identifiées comme complexes pour un âge ou une tranche d'âge cible donné(e) et de préciser sur quels descripteurs linguistiques repose cette complexité.

2. La prédiction de l'âge étant considérée ici comme une tâche de régression et non de classification, nous ne pouvons pas évaluer le modèle à l'aide des mesures de Rappel et de Précision.

3. http://information.extraction.synapse-developpement.fr:83

Remerciements

Nous remercions l'Agence Nationale de la Recherche (ANR) qui a financé ce travail par l'intermédiaire du projet ANR TextToKids (AAPG 2019).

Références

BASSANO D. (1985). Procédures de traitement dans la compréhension d'énoncés modalisés chez l'enfant. *Revue française de pédagogie*, **70**(1), 35–40. Publisher : Persée - Portail des revues scientifiques en SHS, DOI : 10.3406/rfp.1985.1550.

BLANC N. & QUENETTE G. (2017). Valuating children's online emotional inferences : which methodology does generate the best results ? *Enfance*, **4**(4), 503–511.

BLANCHARD S., KORACH D., PENCREAC'H J. & VARONE M. (2001). *Le Robert et Nathan Vocabulaire*. Nathan.

BLANDIN A., LECORVÉ G., BATTISTELLI D. & ÉTIENNE A. (2020). Age Recommendation for Texts. In *Language Resources and Evaluation Conference (LREC)*, Marseille, France.

CARDON R. & GRABAR N. (2021). Simplification automatique de textes biomédicaux en français : lorsque des données précises de petite taille aident. In *Traitement Automatique des Langues Naturelles*, p. 275–277, Lille, France.

DALE E. & CHALL J. S. (1948). A formula for predicting readability. *Educational Research Bulletin*, **27**(1), 11–20. Publisher : Taylor & Francis, Ltd.

DAVIDSON D. (2006). The Role of Basic, Self-Conscious and Self-Conscious Evaluative Emotions in Children's Memory and Understanding of Emotion. *Motivation and Emotion*, **30**(3), 232–242. DOI : 10.1007/s11031-006-9037-6.

DE BELDER J. & MOENS M.-F. (2010). Text simplification for children. In *Prroceedings of the SIGIR workshop on accessible search systems*, p. 19–26 : ACM ; New York.

ELGUENDOUZE S. (2020). Simplification de textes : un état de l'art. In *6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles, RECITAL*, p. 96–109, Nancy, France.

FLOR M., KLEBANOV B. B. & SHEEHAN K. (2013). Lexical tightness and text complexity. In *Proceedings of the Workshop on Natural Language Processing for Improving Textual Accessibility*, Atlanta, Georgia.

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

GALA N., FRANÇOIS T., JAVOUREY-DREVET L. & ZIEGLER J. C. (2018). La simplification de textes, une aide à l'apprentissage de la lecture. *Langue francaise*, **199**(3), 123–131.

GALA N., TODIRASCU A., BERNHARD D., WILKENS R. & MEYER J. P. (2020). Transformations syntaxiques pour une aide à l'apprentissage de la lecture : typologie, adéquation et corpus adaptés. In *7e Congrès Mondial de Linguistique Française*.

HICKMANN M. (2012). Diversité des langues et acquisition du langage : espace et temporalité chez l'enfant. *Langages*, **188**(4), 25–39. Bibliographie_available : 1 Cairndomain : www.cairn.info Cite Par_available : 1 Publisher : Armand Colin.

HICKMANN M., CHAMPAUD C. & BASSANO D. (1993). Pragmatics and metapragmatics in the development of epistemic modality : evidence from French children's reports of think-statements. *First Language*, **13**(39), 359–388. Publisher : SAGE Publications Ltd, DOI : 10.1177/014272379301303905.

IMPERIAL J. M. & ONG E. (2021). Diverse linguistic features for assessing reading difficulty of educational filipino texts. *ArXiv*, **abs/2108.00241**.

LECOCQ P. (1998). *L' É.co.s.se une épreuve de compréhension syntaxico-sémantique (manuel et épreuve) : Deux volumes*. Presses Univ. Septentrion. Google-Books-ID : y9tcOL8vv9kC.

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L. & STOYANOV V. (2019). Roberta : A robustly optimized bert pretraining approach. *arXiv preprint arXiv :1907.11692*.

MARTIN L., MULLER B., ORTIZ SUÁREZ P. J., DUPONT Y., ROMARY L., DE LA CLERGERIE É., SEDDAH D. & SAGOT B. (2020). CamemBERT : a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 7203–7219, Online : Association for Computational Linguistics.

PIOLAT A. & BANNOUR R. (2009). EMOTAIX : Un scénario de Tropes pour l'identification automatisée du lexique émotionnel et affectif. *L'Année Psychologique*, **109**(4), 655–698. Place : France Publisher : Editions NecPlus.

QI P., ZHANG Y., ZHANG Y., BOLTON J. & MANNING C. D. (2020). Stanza : A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics : System Demonstrations*.

RAHMAN R., LECORVÉ G., CHEVELU J., BÉCHET N., ÉTIENNE A. & BATTISTELLI D. (2020). Mama/Papa, Is this Text for Me ? In *International Conference on Computational Linguistics*, Virtuel, Spain.

SHEEHAN K., KOSTIN I. & FUTAGI Y. (2008). When do standard approaches for measuring vocabulary difficulty , syntactic complexity and referential cohesion yield biased estimates of text difficulty ? *The Elementary School Journal*, **115**(2), 184–209.

TAMINE J. & BONNET C. (1982). La compréhension des métaphores chez les enfants. Une hypothèse et quelques implications pédagogiques. *L'information grammaticale*, **14**(1), 17–22. Publisher : Persée - Portail des revues scientifiques en SHS, DOI : 10.3406/igram.1982.2352.

TARTAS V. (2001). The development of systems of conventional time : A study of the appropriation of temporal locations by four-to-ten-year old children. *European Journal of Psychology of Education*, **16**(2), 197–208. DOI : 10.1007/BF03173025.

TARTAS V. (2010). Le développement de notions temporelles par l'enfant. *Developpements*, **4**(1), 17–26.

TEISSÈDRE C., BATTISTELLI D. & MINEL J.-L. (2010). Resources for calendar expressions semantic tagging and temporal navigation through texts. In *Language Resources and Evaluation Conference (LREC)*, p. 3572–3577, Malta : European Language Resources Association (ELRA).

TERS F., REICHENBACH D., MAYER G. & MAYER G. (1970). *L'échelle Dubois-Buyse d'orthographe usuelle française*. H. Messeiller.

VAN DEN BROEK P. (1997). Discovering the cement of the universe : The development of event comprehension from childhood to adulthood. In *Developmental spans in event comprehension and representation : Bridging fictional and actual events*, p. 321–342. Hillsdale, NJ, US : Lawrence Erlbaum Associates, Inc.