Recommandation d'âge pour des textes

Alexis Blandin[1] Gwénolé Lecorvé[1] Delphine Battistelli[2] Aline Étienne[2]
(1) Univ Rennes, CNRS, IRISA, 6, rue de Kerampont, 22300 Lannion, France
(2) Univ. Paris-Nanterre, CNRS, MoDyCo, 200, avenue de la République, 92001 Nanterre, France
{alexis.blandin, gwenole.lecorve}@irisa.fr,
{delphine.battistelli, aline.etienne}@parisnanterre.fr

RÉSUMÉ

Cet article étudie une première tentative pour prédire une recommandation d'âge estimant à partir de quand un enfant pourrait comprendre un texte donné. À ce titre, nous présentons d'abord des descripteurs issus de divers domaines scientifiques, puis proposons différentes architectures de réseaux de neurones et les comparons sur un ensemble de données textuelles en français, dédiées à des publics jeune ou adulte. Pour contourner la faible quantité de données de ce type, nous étudions l'idée de prédire les âges au niveau de la phrase. Les expériences montrent que cette hypothèse, quoique forte, conduit d'ores et déjà à de bons résultats, meilleurs que ceux fournis par des experts psycholinguistes, y compris lorsque les phrases isolées sont remplacées par textes complets.

ABSTRACT

Age recommendation for texts.

This paper studies a first attempt to predict an age recommendation from which a text can be understood by a child. As such, we first exhibits features derived from various scientific domains, then propose different architectures of neural network and compare them on a dataset of French texts dedicated to young or adult audiences. To circumvent the lack of data, we study the idea to predict ages at the sentence level. The experiments show that this strong assumption leads to good results yet, better than those provided by psycholinguists, even when shifting isolated sentences to full texts.

MOTS-CLÉS : recommandation d'âge, enfants, psycholinguistique, réseaux de neurones.

KEYWORDS: age recommendation, children, psycholinguistics, neural networks.

1 Introduction

La façon dont un individu comprend un texte dépend à la fois des caractéristiques du texte et des capacités de l'individu. Par exemple, les capacités à se souvenir d'informations, à positionner un événement dans un scénario, analyser la structure d'une phrase, comprendre un mot ou simplement à lire sont autant de facteurs qui peuvent évoluer d'un individu à l'autre. C'est particulièrement vrai en fonction de l'âge puisque, pendant l'enfance, les capacités cognitives, linguistiques et culturelles évoluent beaucoup. Dans ce contexte, cet article vise à prédire automatiquement des recommandations d'âge pour des textes en français dans le but de maximiser leur compréhension par des enfants.

En tant que tâche, la recommandation d'âge peut être globalement affiliée à celle d'analyse de la lisibilité d'un texte, au sens de la prédiction de la difficulté de lecture d'un texte pour une population spécifique (François, 2015), par exemple, la lecture d'un texte par une personne non-native, ou celle d'un formulaire par des clients. Cependant, la lisibilité du texte est centrée sur l'activité de lecture,

alors que notre travail cherche à intégrer une dimension lié à la compréhension du langage. Cela signifie que nous ne nous limitons pas aux personnes en capacité de lire et considérons également des textes transmis oralement (par exemple, une histoire racontée). En tant que tâche nouvelle, ce document porte les contributions suivantes : (*i*) nous listons un ensemble de descripteurs issus de divers domaines et potentiellement pertinents pour notre tâche ; (*ii*) nous étudions différentes façons de formaliser la prédiction de l'âge comme un problème de régression ; (*iii*) nous répondons au besoin en données annotées en collectant des textes dédiés aux enfants et en exploitant les recommandations fournies par les auteurs ou éditeurs ; (*iv*) nous testons l'hypothèse selon laquelle toutes les phrases d'un même texte partagent la même recommandation d'âge. Bien que linguistiquement contestable, les expériences montrent qu'il s'agit d'une hypothèse de travail concluante ; (*v*) enfin, nous étudions l'acceptabilité des erreurs produites par nos modèles en les comparant avec des prédictions faites par des experts psycholinguistes.

Dans cet article, la section 2 dresse un panorama des travaux connexes en psycholinguistique et en traitement automatique des langues (TAL), puis la section 3 présente les descripteurs retenus dans notre travail. La section 4 présente les différentes approches proposées et la section 5 l'ensemble de données sur lesquelles elles sont étudiées. Enfin, la section 6 présente les résultats.

2 États de l'art

Les études en psycholinguistique mettent en avant différents facteurs majeurs de l'évolution de la compréhension du langage. Tout d'abord, la mémoire phonologique à court terme, qui se développe fortement entre 2 et 8 ans (Gathercole, 1999), joue un rôle essentiel dans le stockage et la restitution de l'information. L'acquisition des notions temporelles est également cruciale car elle permet à un enfant de se situer dans le temps, ainsi que d'ordonner chronologiquement des événements (Tartas, 2010; Hickmann, 2012). Ainsi, la compréhension de notions comme les jours de la semaine, la vitesse, une date très ancienne ou certains connecteurs ou adverbes temporels plus ou moins complexes varie en fonction de l'âge (Vion & Colas, 1999). Les émotions sont également rapportées comme contributives à l'établissement et au maintien de la cohérence des faits dans un texte (Mouw *et al.*, 2019). Cependant, leur repérage progresse avec le temps, s'appuyant initialement sur le lexique, puis progressivement sur des suggestions d'ordre culturel (Blanc, 2010). De même, celles reconnues le plus tôt sont centrées sur l'enfant (la peur, la joie, etc.), puis intègrent progressivement une dimension sociale (la culpabilité, l'empathie, etc.) (Davidson, 2006).

Dès 5-6 ans, les études en apprentissage de la lecture apportent d'autres éléments d'analyse. En particulier, le modèle historique de Frith (Frith, 1985) fait valoir que la lecture est acquise en trois étapes principales à travers lesquelles l'enfant passe d'un décodage des mots par la reconnaissance de symboles globaux, puis de graphèmes et, enfin, de morphèmes. Par ailleurs, d'autres travaux notent que l'intonation lors de la lecture d'un texte – induite par le lexique, la ponctuation, la syntaxe, etc. – influe sur la perception d'un texte et que cette intonation évolue avec l'âge (Aguert *et al.*, 2009).

Enfin, des approches calculatoires existent depuis longtemps pour lier la lisibilité d'un texte à un niveau d'étude. Historiquement, celles-ci se fondent sur les complexités lexicale et syntaxique, à l'image de l'indice Flesch-Kincaid (Flesch, 1948), ou de la formule Dale-Chall qui considère en outre la notion de mots « difficiles » (Dale & Chall, 1948). Plus récemment et plus généralement en TAL, des travaux sur la simplification du texte pour les enfants (De Belder & Moens, 2010; Gala *et al.*, 2018) ou sur l'acquisition du français comme langue étrangère (François & Fairon, 2012) se rapprochent des nôtres. En particulier, (François & Fairon, 2012) propose de prédire des niveaux de lisibilité en utilisant des approches par apprentissage automatique et 46 critères linguistiques

mêlant lexique, syntaxe et sémantique. Notre travail s'en distingue de multiples manières. Outre des méthodes d'apprentissage revisitées, la différence principale tient dans le fait que nous intègrons des informations liées à certaines dimensions développementales et cognitives (temporalité, émotions) pour prédire un âge, là où (François & Fairon, 2012) se limite à des éléments linguistiques pour prédire un niveau de compétence dans une langue (A1, C2...).

3 Descripteurs

Nous considérons 10 dimensions linguistiques qui reflètent l'état de l'art précédent. Pour chaque dimension, nous avons cherché à maximiser le nombre d'informations pouvant être extraites automatiquement. Lorsque les informations linguistiques portent au niveau des mots, les descripteurs sont calculés comme la moyenne et l'écart-type des valeurs par mot. Finalement, pour chaque énoncé, le vecteur de descripteurs globaux est composé de 606 valeurs réelles. Le détail des informations extraites est donné ci-dessous.

Plongements (1 descripteur de dimension 500) : plongement moyen des mots du texte [1].
Lexique (5 descripteurs) : log-prob. des mots en français (estimé sur un vaste corpus d'articles de journaux, romans, transcriptions...) ; nombre de mots/lemmes[2] différents p/r à la longueur du texte.
Graphie/typographie (6) : Score de confusion graphique des mots [3] ; longueur des mots ; ratio de caractères alphanumériques par mot ; ratio de ponctuations par mot.
Morphosyntaxe (7) : classes grammaticales (verbes, verbes d'état, noms, adjectifs, clitiques, adverbes)[2] ; mots-outils.
Temps verbaux (24) : diversité des temps verbaux ; proportions de 14 temps (simples et composés) ; modes ; systèmes temporels (passé, présent, futur).
Personne et forme verbale (5) : proportion des première/deuxième/troisième personnes ; proportion des formes singulier/pluriel.
Syntaxe (8) : mots par phrase ; distances moy. et max. entre un mot et ses dépendances syntaxiques[2] ; nombre de dépendances entrantes/sortantes par mot ; profondeur de l'arbre de dépendances.
Connecteurs logiques (16) : addition ; temps ; but ; cause ; comparaison ; concession ; conclusion ; condition ; conséquence ; énumération ; explicat. ; illustrat. ; justificat. ; opposit. ; restrict. ; exclusion [4].
Phonétique (9) : longueur de la phrase en phonèmes [5] ; nombre de phonèmes par mot ; diversité des phonèmes dans le texte / dans les mots ; scores d'ordinarité phonétique [6].
Sentiments/émotions (26) : Scores de subjectivité et de polarité [7] ; mots identifiés comme déclencheurs d'une parmi 24 émotions [8].

4 Modèles

Notre objectif est de prédire à partir de quel âge un texte d'entrée peut être compris. Précisons que nous ne tenons pas compte d'éventuels retards d'apprentissage. Dans ce problème de régression, l'une des questions majeures est de savoir si l'âge peut être considéré comme une valeur réelle unique

1. Skip-grammes (Fauconnier, 2015) appris sur le corpus FrWaC (Baroni *et al.*, 2009).
2. En utilisant Bonsai (Candito *et al.*, 2010).
3. Inspiré de (Geyer, 1977).
4. Les catégories et les connecteurs ont été établis par consensus de diverses sources.
5. En utilisant eSpeak : `http://espeak.sourceforge.net/index.html`
6. Calculé comme la probabilité moyenne de chaque phonème en français, comme indiqué dans (Gromer & Weiss, 1990).
7. Utilisation du classifieur de sentiments TextBlob
8. Les mots et les émotions sont issus d'un raffinement du dictionnaire EMOTAIX (Piolat & Bannour, 2009).

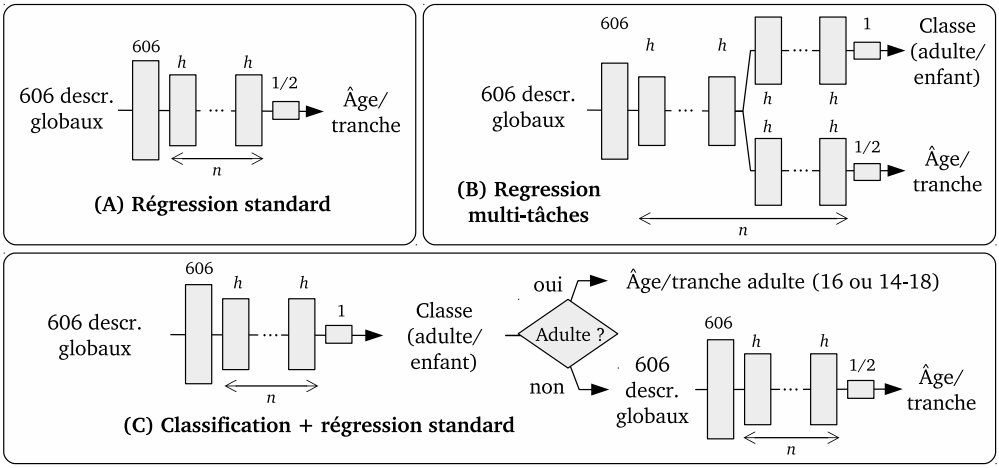

FIGURE 1 – Architectures des approches étudiées.

ou, comme le font souvent les auteurs et éditeurs, comme une tranche d'âges (un âge bas et un âge haut). Cette seconde modélisation reflète le fait que les enfants ne développent pas leurs compétences exactement au même âge et que les textes ont des irrégularités en terme de complexité. Dans ce travail, nous testons les deux modélisations. Par ailleurs, nous fixons une borne supérieure à nos recommandations, borne intuitivement associée à ce que nous qualifions de niveau « adulte ». Elle est fixée à 16 ans, ou 14-18 ans en terme de tranche. Cela coïncide avec la période du lycée en France. Pour un texte, l'évaluation de la tâche se fait en termes de différence absolue entre l'âge prévu et l'âge attendu (vérité terrain). Ainsi, à l'échelle d'un corpus, la métrique utilisée dans notre travail est l'erreur absolue moyenne, notée MAE pour *Mean Absolute Error*. Dans le cas des tranches d'âges, les deux bornes sont prédites, puis ramenés à leur moyenne. Les MAE sont ainsi reportés sur l'âge bas, l'âge haut et la moyenne des deux.

La figure 1 présente les 3 architectures de modèles neuronaux non-récurrents que nous étudions, toutes fondées sur un vecteur de 606 descripteurs globaux et produisant soit un âge ou une tranche d'âges recommandés. Le modèle A est un modèle de régression standard. Le modèle B est un modèle multi-tâches où la prédiction de l'âge est augmentée d'une classification binaire adulte/enfant. Enfin, le modèle C enchaîne un classificateur et un modèle de régression si la classe prédite est "enfants". L'idée est que la régression est inutile pour les textes considérés comme "adultes" car les âges associés sont fixes (16 et 14-18). Le modèle de régression est le même que A mais estimé sur l'ensemble d'apprentissage restreint aux seuls textes pour enfants.

La taille h et le nombre n des couches, ainsi que les fonctions d'activation sont réglées sur le modèle A et dupliqués pour les autres modèles. Le réglage de ces paramètres, ainsi que le nombre de couches spécifiques dans le modèle B, sont détaillés en section 6.

5 Données

Comme détaillé par la figure 2, nous avons collecté un ensemble de 631 textes, dont 541 sont destinés aux enfants de 0 à 14 ans, les 90 autres étant pour les adultes. Les textes pour enfants proviennent de

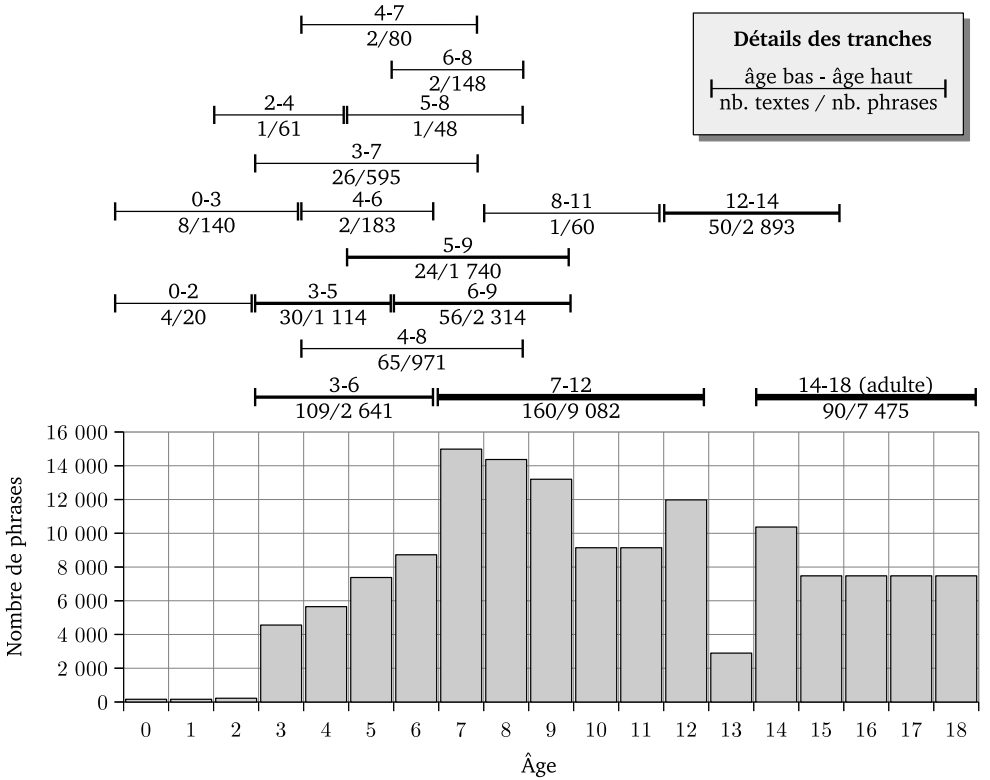

FIGURE 2 – Répartition des phrases et textes en tranches d'âges.

contes, romans, magazines et journaux [9]. Ces textes sont annotés avec les indications des éditeurs ou des auteurs sous la forme d'une tranche d'âge A-B et d'un âge $\frac{A+B}{2}$. Les textes pour adultes sont de genres similaires et sont d'un niveau difficile pour des enfants, par exemple des romans avec un langage soutenu, des articles Wikipedia et de journaux sur des sujets avancés (capitalisme, génétique, diplomatie…). Dans l'ensemble, la validité et l'homogénéité des annotations en tranche d'âge sont à considérer avec prudence car les recommandations des éditeurs et auteurs peuvent refléter des motivations autres que psycholinguistiques, par exemple commerciales ou thématiques.

Étant donné la faible quantité de données compte tenu du nombre de paramètres à régler dans nos modèles, nous avons décidé de décomposer les textes en phrases, chacune partageant la même annotation que son texte d'origine. Il s'agit d'une hypothèse forte et manifestement erronée car, par exemple, toutes les phrases d'un texte pour adulte ne sont pas nécessairement inaccessibles à des enfants. La complexité peut venir de certaines phrases spécifiques ou du raisonnement déroulé par leurs articulations réciproques.

Comme le montre la section 6, elle conduit cependant à de bons résultats en pratique. Le corpus ainsi obtenu est composé de 30K phrases et d'environ 446K mots, répartis en ensembles d'apprentissage, de développement et de test à raison de 60, 20 et 20 %. Une partie de l'ensemble test est composé de phrases provenant de 20 textes qui ne sont pas du tout vus dans les autres ensembles. Cette portion

9. Nous n'utilisons pas les encyclopédies dédiées aux enfants comme Wikimini (fr.wikimini.org) ou Vikidia (fr.vikidia.org) car les articles peuvent être écrits par des enfants. Cela introduirait un biais important car les capacités à écrire et à comprendre sont des capacités différentes.

	MAE Âge	Exac.	MAE Âge bas	MAE Âge haut	MAE Âge moyen	Exac.
Naïve	3,44	74,7%	3,59	3,30	3,44	74,7%
A	2,06	–	2,12	2,08	2,09	–
B	**2,01**	**84,7%**	**2,02**	**1,99**	**2,00**	**85,3%**
C	2,12	84,0%	2,11	2,09	2,09	84,2%

(a) Ensemble de développement

	MAE Âge	MAE Âge bas	MAE Âge haut	MAE Âge moyen
Naïve	3,67	3,74	3,61	3,67
A	**2,24**	**2,26**	**2,29**	**2,27**
B	2,26	2,27	2,30	**2,27**
C	2,31	2,29	2,33	2,30

(b) Ensemble de test

TABLE 1 – MAE et exactitude pour les approches naïve, A, B et C.

« vierge » a une distribution différente en terme de tranches d'âges. Sur l'ensemble d'apprentissage, l'âge moyen est d'environ $10,26$, tandis que la tranche d'âge moyenne est de $8,33$-$12,19$. Sur la portion vierge, ces valeurs sont différentes : $9,01$ et $7,54$-$10,48$.

6 Résultats

Tous les modèles sont entraînés sur l'ensemble d'entraînement en utilisant l'ensemble de développement pour éviter un surapprentissage. La fonction de coût est l'erreur quadratique moyenne pour la régression et l'entropie croisée binaire pour la classification. L'algorithme d'optimisation est Adam, avec 500 époques et une taille de lot de 256 phrases. Après avoir comparé les MAE sur l'ensemble de développement, il apparaît que les meilleurs résultats sont rapportés avec ReLU et $n = 6$ couches cachées de $h = 200$ unités, sans *dropout*. Pour le modèle B, le meilleur nombre de couches spécifiques est de 3 lorsque l'on considère des âges, de 4 pour des tranches d'âge.

Le tableau 1 présente les résultats des modèles A, B et C sur les ensembles de développement (a) et de test (b). Les colonnes de gauche montrent les résultats lorsqu'un âge unique est directement prédit, celle de droite quand une tranche d'âge est prédite. Dans ce second cas, les MAE pour chaque borne sont également signalés, ainsi que celle avec le barycentre de la tranche. Sur l'ensemble de développement, l'exactitude de la classification adultes/enfants est également fournie pour les modèles B et C. Tous les modèles sont comparés à l'approche naïve qui prédit constamment les valeurs moyennes observées dans l'ensemble d'apprentissage (*cf.* section 5). Dans l'ensemble, tous les modèles surpassent clairement cette approche naïve. Ensuite, bien que le modèle B fonctionne un peu mieux sur l'ensemble de développement, la différence avec le modèle A disparaît sur l'ensemble de test. Enfin, il semble que prédire des âges ou des tranches d'âges ne change pas grand chose.

Pour affiner ces résultats, nous avons fait annoter la portion vierge de l'ensemble de test par trois psycholinguistes spécialisés dans le développement des enfants. Ces annotations ont été effectuées soit sur l'ensemble des 20 textes concernés, soit sur 80 phrases isolées tirées aléatoirement. Le tableau 2.a donne les résultats, incluant ceux des experts pris individuellement ou après moyennage de leurs prévisions. Sur les phrases, il apparaît que les prédictions de notre modèle sont meilleures que celles des experts, même en essayant de trouver un consensus entre elles (dernière ligne). Lorsque plus de contexte est donné aux experts à travers des textes complets (b), leurs recommandations tendent à battre celles de notre modèle lorsque ses prédictions sont comptabilisées phrase par phrase (sauf pour un expert). Cependant, un simple calcul de moyenne de ces prédictions phrase par phrase amène à une amélioration substantielle (ligne "par texte") et à des prédictions meilleures que celles des experts. Outre les conclusions positives concernant notre approche de recommandation automatique,

	Âge	Âge bas	Âge haut	Âge moyen
Naïve	4,46	4,29	4,63	4,46
Modèle B	**2,70**	**2,53**	**2,65**	**2,57**
Expert 1	3,14	2,95	3,45	3,14
Expert 2	3,38	3,48	3,39	3,38
Expert 3	3,07	2,93	3,54	3,07
Moy. experts	2,88	2,86	3,05	2,88

(a) 80 phrases isolées

		Âge	Âge bas	Âge haut	Âge moyen
Naïve		4,57	4,32	4,83	4,57
Modèle B	Par phrase	3,13	3,07	3,35	3,18
	Par texte	**2,39**	**2,51**	**2,57**	**2,53**
Expert 1		2,60	2,60	2,80	2,60
Expert 2		3,50	3,80	3,30	3,50
Expert 3		2,70	2,90	2,60	2,70
Moy. experts		2,95	3,19	2,81	2,95

(b) 20 textes

TABLE 2 – Comparaisons entre notre approche et les experts sur une portion de l'ensemble de test.

ces résultats reflètent sans doute aussi un décalage quant à la notion de tranche d'âge entre les experts psycholinguistiques et les éditeurs ou auteurs. L'analyse des prédicitions montrent en effet que les experts tendent à utiliser une plage d'âges restreinte par rapport à celle autorisée (et utilisée par nos modèles), à savoir 4-13 ans *versus* 0-18 ans.

7 Conclusion et perspectives

Dans cet article, nous avons étudié la tâche originale de recommandation d'âge pour des textes. Plusieurs modélisations ont été proposées et les résultats ont été comparés aux recommandations de psycholinguistes. Ces résultats montrent que les prévisions de nos modèles sont meilleures que celles des experts. En outre, tout en s'appuyant sur une hypothèse forte selon laquelle toutes les phrases d'un texte peuvent être considérées comme toutes associées à une tranche d'âge unique, nos résultats sur l'agrégation des résultats au niveau des phrases sont clairement encourageants. Cela démontre la viabilité de l'approche et appelle des investigations supplémentaires.

Cependant, nous sommes conscients que ces résultats doivent être pris avec prudence car différents aspects portent de l'incertitude. En particulier, la précision et le bien-fondé scientifique des recommandations fournies par les éditeurs et les auteurs sont sans doute parfois discutables. Ensuite, considérer les erreurs absolues moyennes avec un âge cible est probablement trop dur. Il serait intéressant de comparer des tranches d'âges et non des âges dans l'évaluation. Enfin, il serait intéressant de corréler les résultats avec une campagne d'évaluation *in situ* auprès des enfants. Ceci est prévu dans les prochains mois.

Remerciements

Ce travail a bénéficié du soutien financiers des projets ANR TREMoLo et TextToKids.

Références

AGUERT M., BERNICOT J. & LAVAL V. (2009). Prosodie et compréhension des énoncés chez les enfants de 5 à 9 ans. *Enfance*, **3**.

BARONI M., BERNARDINI S., FERRARESI A. & ZANCHETTA E. (2009). The wacky wide web : a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, **43**(3).

BLANC N. (2010). La compréhension des contes entre 5 et 7 ans : Quelle représentation des informations émotionnelles ? *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, **64**(4).

CANDITO M., NIVRE J., DENIS P. & ANGUIANO E. H. (2010). Benchmarking of statistical dependency parsers for french. In *Proceedings of the International Conference on Computational Linguistics : Posters*, COLING '10 : Association for Computational Linguistics.

DALE E. & CHALL J. S. (1948). A formula for predicting readability : Instructions. *Educational research bulletin*, **27**.

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).

DE BELDER J. & MOENS M.-F. (2010). Text simplification for children. In *Proceedings of the SIGIR workshop on accessible search systems*.

FAUCONNIER J.-P. (2015). French word embeddings. http://fauconnier.github.io.

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

FRANÇOIS T. (2015). When readability meets computational linguistics : a new paradigm in readability. *Revue française de linguistique appliquée*, **20**(2).

FRANÇOIS T. & FAIRON C. (2012). An "AI readability" formula for french as a foreign language. In *Proceedings of EMNLP-CoNLL*.

FRITH U. (1985). Beneath the surface of developmental dyslexia. *In Surface Dyslexia : Neuropsychological and Cognitive Studies of Phonological Reading*.

GALA N., FRANCOIS T., JAVOUREY-DREVET L. & ZIEGLER J. C. (2018). Text simplification, a tool for learning to read. *Langue française*, (199).

GATHERCOLE S. (1999). Cognitive approaches to the development of short-term memory. *Trends in cognitive sciences*, **3**.

GEYER L. H. (1977). Recognition and confusion of the lowercase alphabet. *Perception & Psychophysics*, **22**(5).

GROMER D. & WEISS M. (1990). *Lire, tome 1 : apprendre à lire*. Armand Colin.

HICKMANN M. (2012). Diversité des langues et acquisition du langage : espace et temporalité chez l'enfant. *Langages*, (4).

MOUW J. M., VAN LEIJENHORST L., SAAB N., DANEL M. S. & VAN DEN BROEK P. (2019). Contributions of emotion understanding to narrative comprehension in children and adults. *European Journal of Developmental Psychology*, **16**(1).

PIOLAT A. & BANNOUR R. (2009). An example of text analysis software (emotaix-tropes) use : The influence of anxiety on expressive writing. *Current psychology letters. Behaviour, brain & cognition*, **25**(2).

TARTAS V. (2010). Le développement de notions temporelles par l'enfant. *Développements*, **4**.

VION M. & COLAS A. (1999). L'emploi des connecteurs en français : contraintes cognitives et développement des compétences narratives (le cas de la narration de séquences arbitraires d'événements). In *Proceedings of the Conference of the International Association for the Study of Child Language*.