Vers une simplification automatique de textes en français :

bilan des travaux du projet ALECTOR (ressources, approches et évaluations)

Núria Gala

MCF en Sciences du Langage Habilitée à Diriger des Recherches

Mots & Machines #4, Université de Bretagne, 25/03/2022

UMR 7309 • CNRS • AMU PAROLE ET LANGAGE

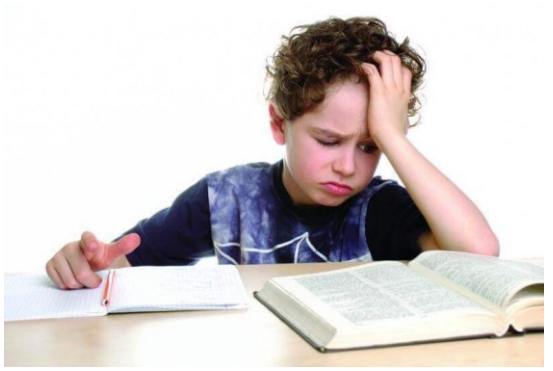
Laboratoire Parole et Langage 50 1972 2022

Aix*Marseille université Socialement engagée

ILCB Institute of Language Communication and the Brain

PROJET FINANCÉ PAR L'ANR - PROJECT FUNDED BY THE ANR

Enjeux



En fin de CE1 (Évaluation Repères CE1 2022) :

→ 15 à 20 % des élèves sont en difficulté

→ résultats en baisse par rapport à des années précédentes dans :

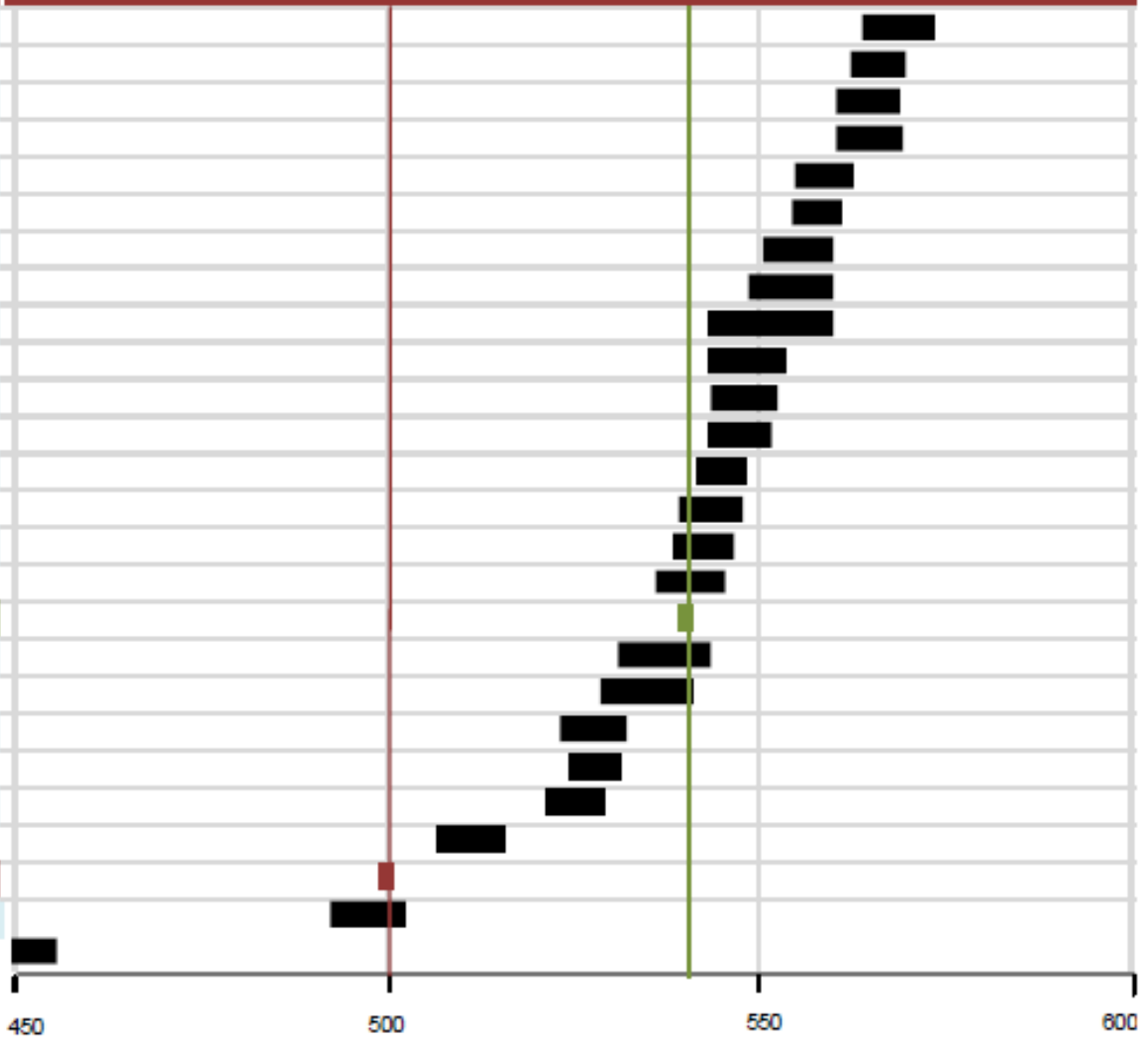
- la lecture de textes et de mots
- la compréhension de textes et de phrases

En fin d'école primaire (Rapport CEDRE 2015) :

→ 10 % sont des faibles 'compreneurs' et des lecteurs en difficulté (compréhension locale, pas en mesure de réaliser un portrait global du texte)

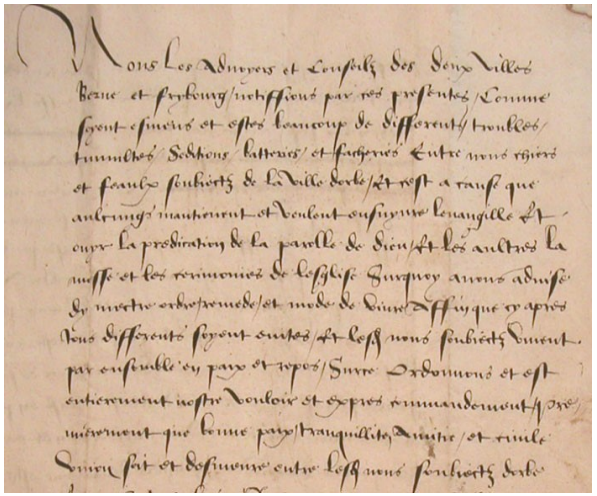
PIRLS 2016 : évaluation internationale des élèves de 7-8 ans en compréhension de l'écrit

Pays	Scores moyens
Irlande	567
Finlande	566
Pologne	565
Irlande du Nord	565
Angleterre	559
Lettonie	558
Suède	555
Hongrie	554
Bulgarie	552
Lituanie	548
Italie	548
Danemark	547
Pays-Bas	545
Rép. Tchèque	543
Slovénie	542
Autriche	541
Moyenne européenne	*540*
Allemagne	537
Slovaquie	535
Portugal	528
Espagne	528
Belgique (FL)	525
France	**511**
Centre de l'échelle PIRLS	*500*
Belgique (FR)	498
Malte	452



© N. Gala

Capacité de lecture



Aspects typographiques,
lexique (mots, locutions),
syntaxe (structures de phrase),
organisation des idées...

Connaissances, âge, expériences, langue maternelle, vécu, motivation, attitude, persévérance...

© N. Gala

Simplification de textes

Transformation du texte tout en conservant son contenu (Siddhartan 2014).

Réduction de la complexité pour un public cible (Gala et al. 2018), p. ex. enfants dyslexiques et faibles lecteurs.

Objectif : améliorer le décodage et la compréhension

→ entraînement à la lecture (« béquille »)

→ susciter le plaisir de la lecture !

Projet ALECTOR

Projet ALECTOR

2017 - 2021

https://alectorsite.wordpress.com

UMR 7309 • CNRS • AMU
PAROLE ET LANGAGE

LIS
LABORATOIRE D'INFORMATIQUE & SYSTÈMES

Laboratoire de Psychologie Cognitive

lilpa
linguistique, langues, parole

Limsi

CENTAL
centre de traitement automatique du langage

Núria Gala

Firas Hmida

Mokhtar B. Billami

Solange Lâm

Carlos Ramisch

Johannes Ziegler

Ludivine Javourey

Stéphane Dufau

Amalia Todirascu

Delphine Bernhard

Jean Paul Meyers

Rodrigo Souza Wilkens

Anne Laure Ligozat

Thierry Hamon

Thomas François

Anaïs Tack

Adeline Müller

Questions de recherche

1) Peut-on augmenter la compréhension et la fluidité de la lecture en simplifiant les textes ?

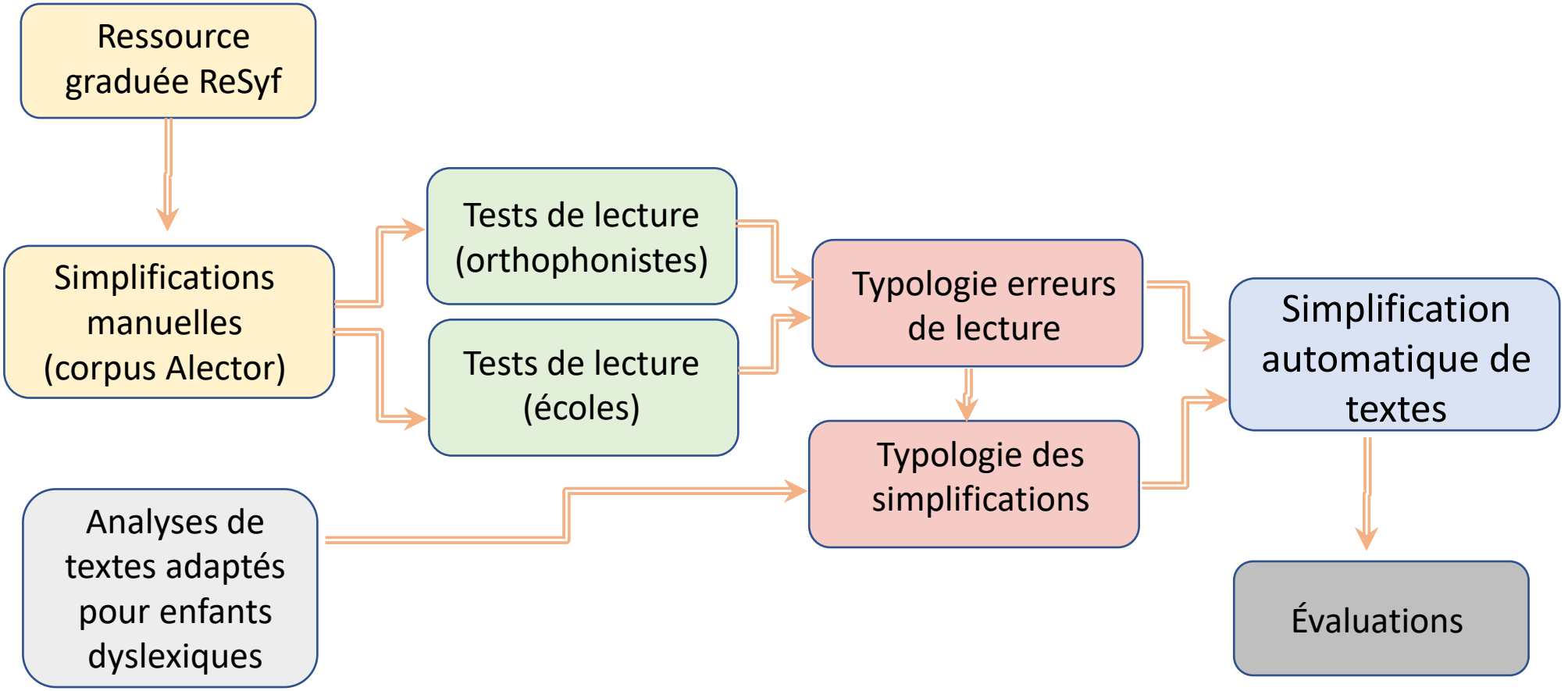
Hypothèse : oui !

2) Quel est l'impact des simplifications sur un public normo-lecteur ? Et sur un public en difficulté (dyslexiques et faibles lecteurs) ?

3) Qu'est-ce qu'on simplifie et comment ?

Lexique, syntaxe, éléments du discours (pronoms anaphoriques).
Manuellement → (semi-)automatiquement

Méthodologie



Ressource

ReSyf

(Billami et al. 2018)

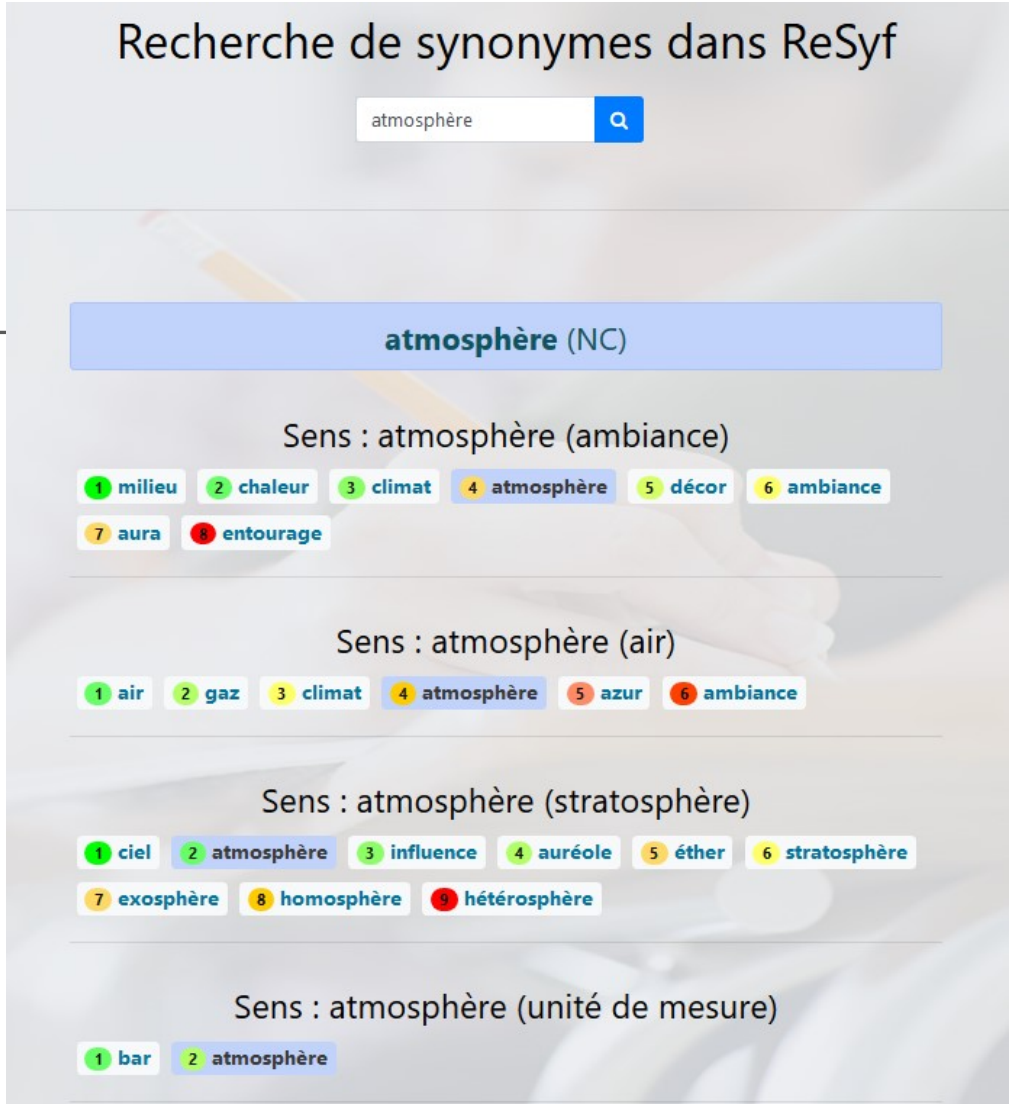
Sur liste de JeuxDeMots
(Lafourcade 2007)

57.589 entrées

Ranking des synonymes
(François et al. 2016)

Désambiguïsation sémantique
(Billami 2018)

Interface : Dorian Ricci et Brayan Delmée
(CENTAL 2017), Nader Janhaoui (LIF 2015)



https://cental.uclouvain.be/resyf/

© N. Gala

Analyse

Analyses de corpus (versions dys)

Analyse d'un corpus de 9 paires de contes (versions originales et leurs équivalents destinés à un public dyslexique).

https://methodolodys.ch/lecture-comprehension/

Identification et typologie des simplifications : lexicales, morpho-syntaxiques, syntaxiques, discursives (substitutions, suppressions, transformations).

Texte original (C. Perrault / extrait) : 481 mots / 2648 signes	*Texte simplifié (J.-C. Marguerite / extrait) : 301 mots / 1582 signes*
Mais lorsque l'argent fut dépensé,	Quand il ne reste plus d'argent il ne reste plus rien à manger.
ils retombèrent dans leur premier chagrin	Le mari et la femme refusent de voir leurs enfants mourir de faim.
et résolurent de les perdre encore	Ils décident de les perdre dans la forêt.
et pour ne pas manquer leur coup, de les mener bien plus loin que la première fois.	
Ils ne purent parler de cela si secrètement	Un soir ils en parlent à voix basse pour ne pas être entendus.
qu'ils ne fussent entendus par le petit Poucet,	Mais le petit Poucet les a entendus.
qui fit son compte de sortir d'affaire comme il avait déjà fait ;	
mais quoiqu'il se fût levé de bon matin pour aller ramasser des petits cailloux,	Le lendemain matin le petit Poucet se lève le premier pour remplir ses poches de petits cailloux blancs
il ne put en venir à bout,	
car il trouva la porte de la maison fermée à double tour.	Mais la porte est fermée à clé.
Il ne savait que faire, lorsque	Il se demande comment faire pour retrouver le chemin de la maison.
la Bûcheronne leur ayant donné à chacun un morceau de pain pour leur déjeuner,	
	Le petit Poucet a une idée.
il songea qu'il pourrait se servir de son pain au lieu de cailloux en le jetant par miettes le long des chemins où ils passeraient ;	Il pense à laisser tomber de ses poches des miettes de son pain comme il l'a fait avec les cailloux.
il le serra donc dans sa poche.	
Le Père et la Mère les menèrent dans l'endroit de la Forêt le plus épais et le plus obscur,	Quand ils sont tous loin dans la forêt le père dit à ses sept garçons
et dès qu'ils y furent,	d'aller encore plus loin pour ramasser des petites branches.
ils gagnèrent un faux-fuyant	Quand les sept garçons sont plus loin leurs parents retournent chez eux en abandonnant leurs enfants dans la forêt

Ressource

Textes manuellement simplifiés pour des tests de lecture

Création d'un corpus de textes parallèles : version originale et version manuellement simplifiée (Gala et al., 2020a)

79 corpus originaux, textes narratifs et documentaires scientifiques, niveaux CE1, CE2 et CM1 (7 à 9 ans).

Simplifications lexicales, syntaxiques et discursives.

Moyenne nombre occurrences / texte	ORIG	SIMPL
Littéraires	339	271
Scientifiques	313	239

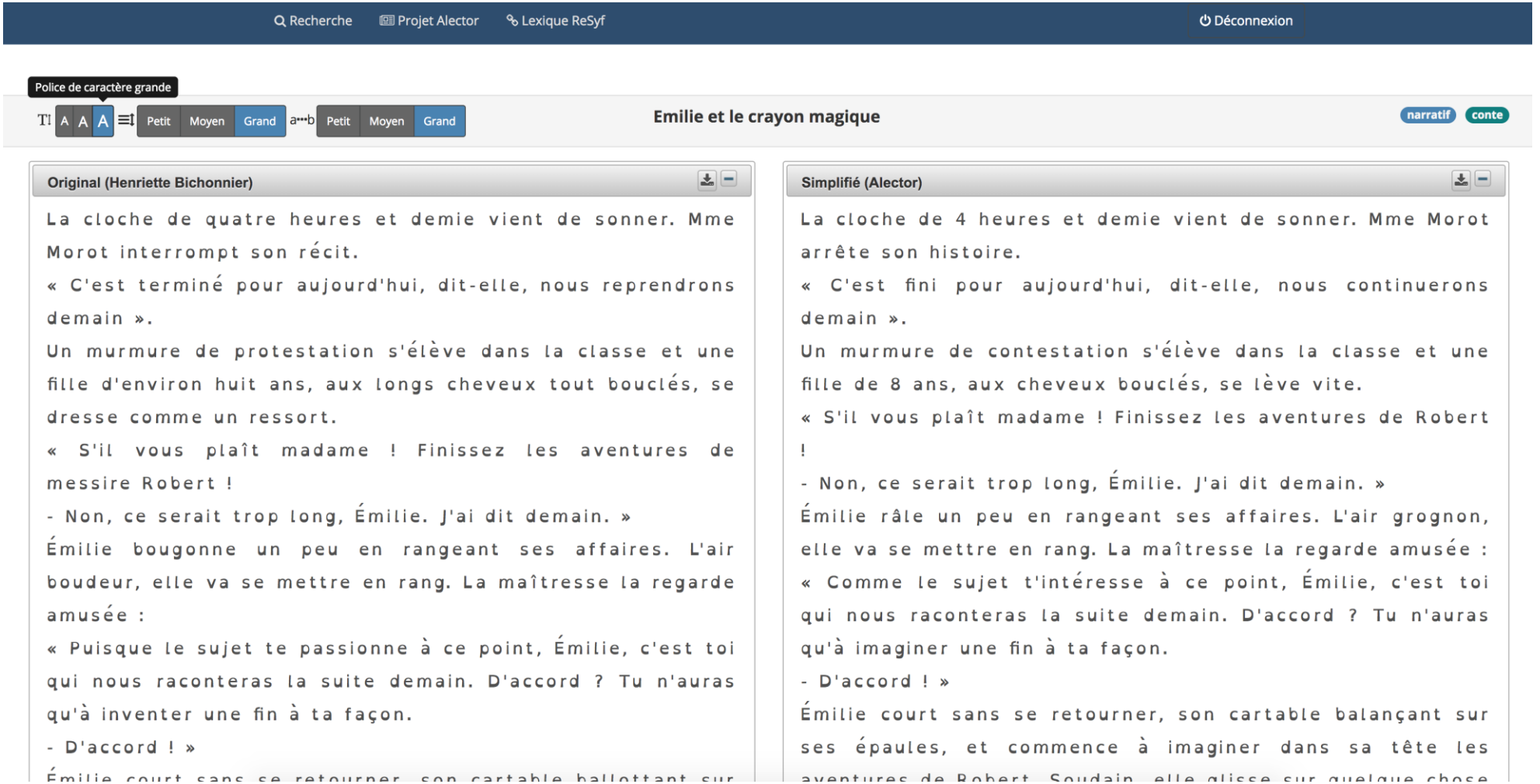
Quel que soit le temps, il rapportait toujours au village quantité de beaux et rares poissons.
Par tous les temps, il ramenait au village beaucoup de poissons.

Interface d'accès disponible en ligne, textes intégrés dans un eBook pour des tests de lecture dans des écoles.

© N. Gala

Ressource

Corpus Alector



https://corpusalector.huma-num.fr

Interface : Solange Lâm (2018)

Tests de lecture (orthophonistes)

Brunel et Combes (2015)
Nandiegou et Reboul (2018)

Application tablette : Mokhtar B. Billami

Étude 1 (ponctuelle)

2 études (2015 et 2018)

10 + 20 enfants dyslexiques (moyenne 10 ans)

Retard de lecture estimé à 2 ans et 9 mois (moyenne)

Tests de lecture et compréhension

Lecture à voix haute

Textes scientifiques (doc) et littéraires (narratifs)

Versions originales et simplifiées

© N. Gala

Tests de lecture (orthophonistes)

Nandiegou et Reboul, 2018

Analyses

Figure 22 : Relation entre les pourcentages moyens de mots complexes ou simplifiés mal lus

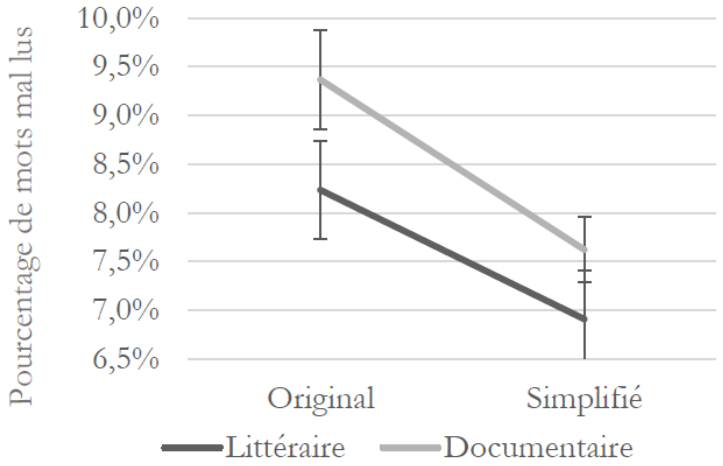


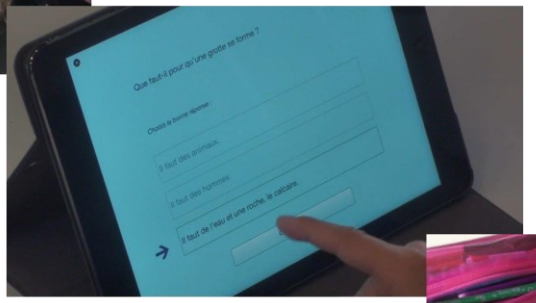
Figure 24 : Pourcentage de mots mal lus en fonction du type et de la nature du texte

Tests de lecture (écoles)

Javourey-Drevet (2021) ; Javourey-Drevet et al., 2022

Application HIBOU : Stéphane Dufau

Étude 2 (en longitudinal)



3 ans (2017-2019)

160 – 170 enfants/an

2e- 4e année, 7-10 ans

Tests de lecture et compréhension

Tablette numérique (lecture silencieuse)

Textes scientifiques (doc) et littéraires (narratifs)

Versions originales et simplifiées

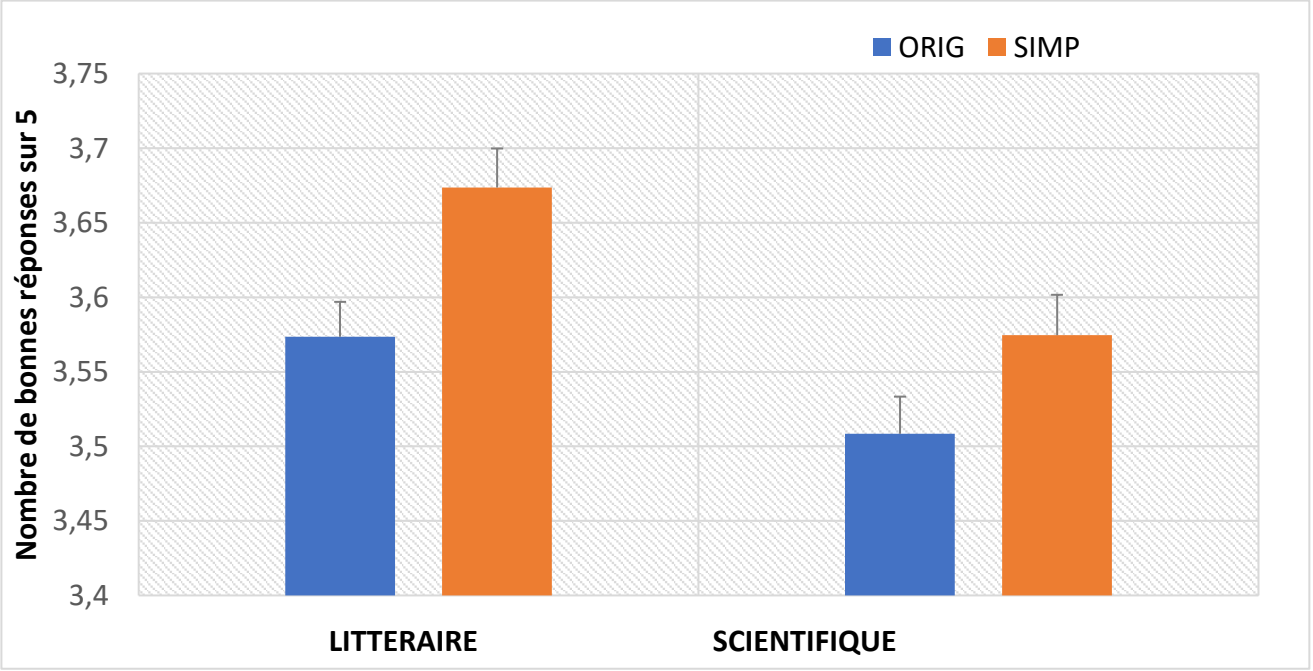
© N. Gala

Tests de lecture (écoles)

(Javourey-Drevet 2021 ; Javourey-Drevet et al., 2022)

Résultats

Nombre de réponses correctes (test de compréhension) en fonction du type de texte

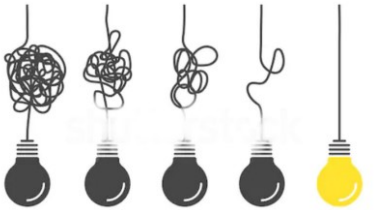


EFFET de la simplification	F = 8,139	P = 0,005
EFFET du type de texte	F = 7,883	P = 0,006

© N. Gala

Typologie

Guide de transformations



(Gala et al. 2020b)

Recommandations pour des transformations de textes en français afin d'améliorer leur lisibilité et leur compréhension.

- Typographie
- Lexique
- Morphologie
- Syntaxe
- Discours

3.3 R09.L3. Syllabes avec une structure CV ou V

On privilégiera les syllabes simples et fréquentes (CV-V) au détriment de structures plus complexes (CVCC, CCVC, CYVC, CYV, CCV, VCC, CVY, VC, YV, VCCC, CCYC, CCVC).

Une structure complexe est par définition plus longue (par exemple, double consonne). Les temps de lecture sont inférieurs et les erreurs moindres dans des phrases ayant des unités lexicales qui comportent des syllabes simples et fréquentes :

Difficulté	Caractéristiques, type de syllabe	Syllabe
1	Simple, fréquente	CV, V
2	Complexe, fréquente	CVC
3	Complexe, moins fréquente	CVCC, CCVC, CYVC, CYV, CCV, VCC, CVY
4	Complexe, rare	VC, YV, VCCC, CCYC, CCVC

Figure 5. Typologie de syllabes.

Original. Je suis sûre que tu peux gagner ce concours. (CV-CVC)

Simplifié. Je suis sûre que tu peux gagner ce défi. (CV-CV)

https://alectorsite.files.wordpress.com/2020/11/guidelines-linguistiques_alector_final.pdf

© N. Gala

Vers une simplification automatique de textes en français

Domaine de plus en plus présent dans le TAL (techniques du résumé et de la traduction automatique, lisibilité automatique).

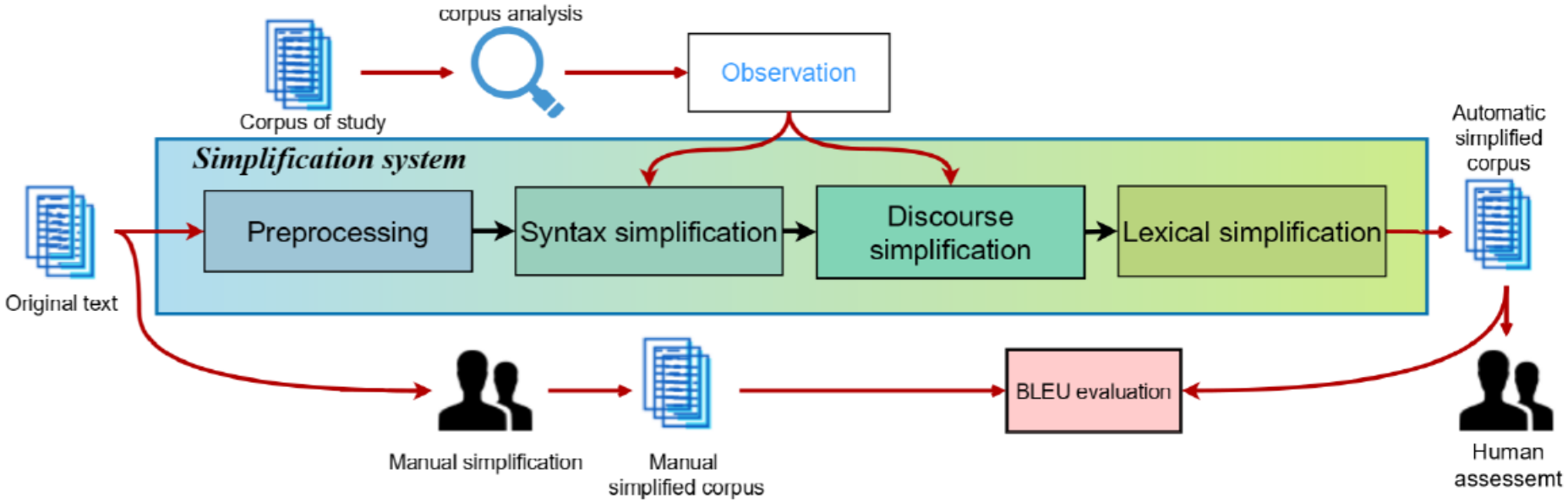
Travaux principalement sur le lexique (Bott et al. 2012) et la syntaxe (Chandrasekar et al. 1996) mais aussi le discours (Todirascu et al. 2016). Peu de travaux pour le français.

Applications : aphasie (Carrol et al. 1999), dyslexie (Rello et al. 2013), illettrisme (Candido et al. 2009) , domaine de la terminologie médicale (Cardon 2018), etc.

© N. Gala

HECTOR: premier prototype de simplificateur en français

Système hybride (Todirascu et al., 2022 –soumis) : règles pour les simplifications syntaxiques et discursives et apprentissage automatique (CamemBERT, SVM) pour les simplifications lexicales afin de mieux gérer la polysémie et les contextes d'usage (FrenLyS (Rolin et al., 2021))



Évaluation

Premiers résultats

BLEU : nombre de transformations

BLEU	Textes ORIG-SIMP CE1 (2e)	Textes ORIG-SIMP CM1 (4e)
Phrases LIT	**0,64**	0,60
Phrases SCIENT	**0,78**	0,51

Évaluation humaine : accord inter-annotateurs (α de Krippendorf) sur une échelle de Likert 1-5 pour :

- Grammaticalité
- Sens préservé
- Simplicité

Transformations	Syntaxiques	Discursives	Lexicales
Grammaticalité	**0,74**	0,63	0,46
Sens préservé	0,58	0,26	0,45
Simplicité	0,48	0,29	0,37

Exemples d'erreurs Gramm.
Le salsola kali, plante de la steppe, était réduit en cendres par les bédouins (…)
**Les bédouins le salsola kali, plante de la steppe, réduit en cendres.*

Exemples d'erreurs Sens
(…) les chats l'ont griffé.
(…) les chats l'ont déchiré.

Exemples d'erreurs Simplicité (et sens)
Il lui demanda où elle allait
La mère-grand demanda où la mère-grand allait

© N. Gala

Conclusion :

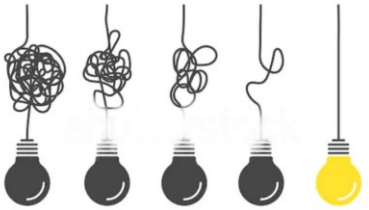
résultats du projet (méthodologie)

Tests de lecture (orthophonistes 2015 et 2018 et écoles 2017 à 2019).

Analyse de corpus DYS.

Texte original (C. Perrault / extrait) : 481 mots / 2648 signes	*Texte simplifié (J.-C. Marguerite / extrait) : 301 mots / 1582 signes*
Mais lorsque l'argent fut dépensé,	Quand il ne reste plus d'argent il ne reste plus rien à manger.
ils retombèrent dans leur premier chagrin	Le mari et la femme refusent de voir leurs enfants mourir de faim.
et résolurent de les perdre encore	Ils décident de les perdre dans la forêt.
et pour ne pas manquer leur coup, de les mener bien plus loin que la première fois.	
Ils ne purent parler de cela si secrètement	Un soir ils en parlent à voix basse pour ne pas être entendus.
qu'ils ne fussent entendus par le petit Poucet.	Mais le petit Poucet les a entendus.
qui fit son compte de sortir d'affaire comme il avait déjà fait ;	
mais quoiqu'il se fût levé de bon matin pour aller ramasser des petits cailloux,	Le lendemain matin le petit Poucet se lève le premier pour remplir ses poches de petits cailloux blancs

Recommandations pour des transformations de textes en français afin d'améliorer leur lisibilité et leur compréhension.

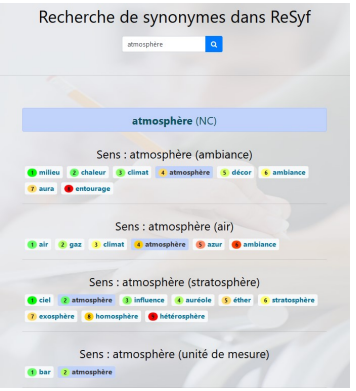


© N. Gala

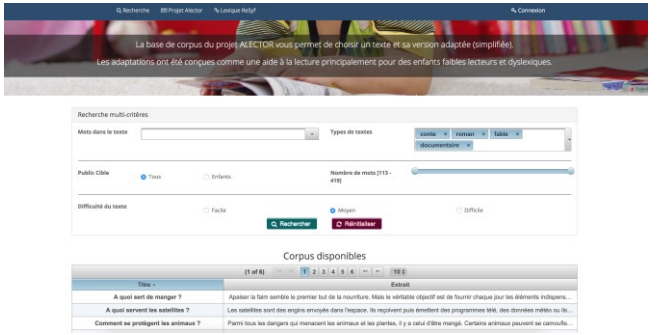
Conclusion :

résultats du projet (ressources)

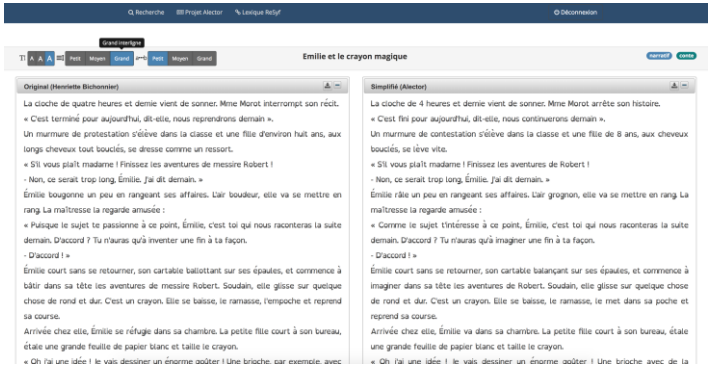
Ressource lexicale ReSyf (2018)



Corpus parallèle Alector (2020)



Pipeline complet : simplificateur Hector (2022)



Perspectives et mot de la fin

- Travaux en cours pour proposer des simplifications automatiques : analyse des erreurs d'HECTOR, amélioration des procédures (ordre des mots, expressions polylexicales).

- Enrichissement du corpus ALECTOR (nouveaux textes) et portage d'HIBOU dans un format standard EPUB accessible en ligne (plateforme ISIaccess de la société ISI).

La simplification de textes s'avère une solution possible pour l'aide à la lecture des enfants en difficulté (amélioration de la vitesse de lecture sans aucune perte en compréhension du texte lu, réduction significative du nombre d'erreurs de lecture).

Nécessité d'outils automatiques fiables pour la génération de textes simplifiés en français.

Publications liées au projet ALECTOR

Billami, M. B. (2018) Désambiguïsation sémantique dans le cadre de la simplification lexicale : contributions à un système d'aide à la lecture pour enfants dyslexiques et faibles lecteurs. Thèse de doctorat, Aix Marseille Université.

Billami, M. B., François, T. and Gala, N. (2018) ReSyf: a French lexicon with ranked synonyms. Proceedings of the *27th International Conference on Computational Linguistics* (COLING-2018), Santa Fe, New Mexico, USA, pp. 2570-2581.

Brunel, A. et Combes, M. (2015) Simplification de textes pour faciliter leur lisibilité et compréhension. Mémoire en vue de l'obtention du Certificat de Capacité en Orthophonie, Aix Marseille Université.

François, T. Billami, M. B. et Gala, N. (2016) Bleu, contusion, ecchymose : tri automatique de synonymes. Actes de la conférence en Traitement Automatique des Langues (TALN-2016), Paris.

Gala, N., Tack, A., Javourey-Drevet, L., François, T. and Ziegler, J. C. (2020) Alector: Parallel Corpus of Simplified French Texts with Alignments of Misreadings by Poor and Dyslexic Readers. In the *12th International Conference on Language Resources and Evaluation* (LREC), Marseille, France.

Gala, N., Todirascu, A., Bernhard, D., Wilkens, R. et Meyer, J.-P. (2020) Transformations syntaxiques pour une aide à l'apprentissage de la lecture : typologie, adéquation et corpus adaptés. *Congrès Mondial de Linguistique Française (CMLF 2020). Montpellier, France.*

Gala, N., François, T., Javourey-Drevet, L. et Ziegler, J.-C. (2018) La simplification de textes, une aide à l'apprentissage de la lecture. Dans *Langue Française «L'apprentissage de la lecture en français langue maternelle et seconde»*, Armand Colin, pp. 123-131.

Gala, N. and Ziegler, J. (2016) Reducing lexical complexity as a tool to increase text accessibility for children with dyslexia. *Proceedings of the workshop Computational Linguistics for Linguistic Complexity (CL4LC) at the 26th Int. Conference on Computational Linguistics (COLING-2016).* Osaka, Japon.

Javourey-Drevet, L., François, T., Gala, N., Dufau, S. et Ziegler, J.-C (2022) Simplification of literary and scientific texts to improve fluency and comprehension in beginning readers of French. *Applied Psycholinguistics*. Cambridge University Press (CUP), 2022, 1-28.

Nandiegoul, M. et Reboul, S. (2018) La simplification lexicale comme outil pour faciliter la lecture des enfants dyslexiques et faibles lecteurs. Mémoire en vue de l'obtention du Certificat de Capacité en Orthophonie, Aix Marseille Université.

Todirascu, A., Wilkens, R., Rolin, E., François, T., Bernhard, D. & Gala, N. (*soumis*) HECTOR: a Hybrid Text Simplification Tool for Raw Texts in French. Conférence internationale.

Autres références

Bott, S., Rello, L., Drndarevic, B. and Saggion, H. (2012) Can Spanish be Simpler? LexSiS: Lexical Simplification for Spanish. In proceedings of the *Conference on Computational Linguistics* (COLING 2012). Technical Papers, pp. 357-374. Mumbai, India.

Candido, A., Maziero, E., Gasperin, C., Pardo, T., Specia, L. & Aluisio, S. M. (2009) Supporting the Adaptation of Texts for Poor Literacy Readers: A Text Simplification Editor for Brazilian Portuguese. In the *Proceedings of the NAACL HLT Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 34–42, Boulder, Colorado.

Cardon, R. (2018) Approche lexicale de la simplification automatique de textes médicaux. Actes de la conférence *Traitement Automatique des Langues Naturelles* (TALN 2018), Rennes.

Carroll, J., Minnen G., Canning Y., Devlin, S. and Tait, J. (1998) Practical simplification of English newspaper text to assist aphasic readers. *In Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and assistive Technology*, pp. 7–10.

Cunningham, A. E. and Stanovich, K. E. (1998) What Reading does for the Mind. *Journal of Direct Instruction*, vol. 1, No. 2, pp. 137–149. Reprinted (2001) from The American Federation of Teachers. *American Educator*, vol. 22, No. 1–2, pp. 8–15.

Glavaš et Štajner, (2015) Simplifying lexical simplification: do we need simplified corpora? In Proceedings of the *53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (Volume 2: Short Papers), pp. 63–68.

Rello, L., Baeza-Yates, R. & Saggion, H. (2013) The impact of lexical simplification by verbal paraphrases for people with and without Dyslexia. *In Computational Linguistics and Intelligent Text Processing. Lecture Notes in Computer Science Volume 7817*, pp. 501-512.
Saggion, H. (2017) *Automatic Text Simplification. Synthesis Lectures on Human Language Technologies*, volume 10(1): pp. 1-137, California Morgan & Claypool Publishers.

Rolin, E., Langlois, Q., Watrin, P et François, T. (2021) FrenLyS: a tool for the automatic simplification of French general language texts. *Proceedings of Recent advances in Natural Language Processing (RANLP)*, 1196 – 1205.

Shardlow, M. (2014) A survey of automated text simplification. International Journal of Advanced Computer Science and Applications, 4(1): 58-70.

Siddharthan, A. (2014) A survey of research on text simplification. *ITL-International Journal of Applied Linguistics* 165, pp. 259-298.

Valdois, S. (2003) Les élèves en difficulté d'apprentissage de la lecture (2003) *Document envoyé au PIREF en vue de la conférence de consensus sur l'enseignement de la lecture à l'école primaire,* les 4 et 5 décembre 2003. www.bienlire.education.fr