Alector: A Parallel Corpus of Simplified French Texts with Alignments of Misreadings by Poor and Dyslexic Readers

Núria Gala, Anaïs Tack, Ludivine Javourey-Drevet, Thomas François, Johannes C. Ziegler

Alector: A Parallel Corpus of Simplified French Texts with Alignments of Misreadings by Poor and Dyslexic Readers

Núria Gala[1], Anaïs Tack[2,3,4], Ludivine Javourey-Drevet[5,6], Thomas François[2], Johannes C. Ziegler[5]
[1]Aix Marseille Univ, Laboratoire Parole et Langage, LPL CNRS (UMR 7309), France
[2]CENTAL, UCLouvain [3]ITEC, imec research group at KU Leuven [4]F.R.S.-FNRS Research Fellow, Belgium
[5]Aix Marseille Univ, Laboratoire de Psychologie Cognitive, LPC CNRS (UMR 7290), France
[6]Aix Marseille Univ, Apprentissage, Didactique, Évaluation, Formation (EA 4671), France
{nuria.gala,ludivine.javourey, johannes.ziegler}@univ-amu.fr
{anais.tack,thomas.francois}@uclouvain.be

Abstract

In this paper, we present a new parallel corpus addressed to researchers, teachers, and speech therapists interested in text simplification as a means of alleviating difficulties in children learning to read. The corpus is composed of excerpts drawn from 79 authentic literary (tales, stories) and scientific (documentary) texts commonly used in French schools for children aged between 7 to 9 years old. The excerpts were manually simplified at the lexical, morpho-syntactic, and discourse levels in order to propose a parallel corpus for reading tests and for the development of automatic text simplification tools. A sample of 21 poor-reading and dyslexic children with an average reading delay of 2.5 years read a portion of the corpus. The transcripts of readings errors were integrated into the corpus with the goal of identifying lexical difficulty in the target population. By means of statistical testing, we provide evidence that the manual simplifications significantly reduced reading errors, highlighting that the words targeted for simplification were not only well-chosen but also substituted with substantially easier alternatives. The entire corpus is available for consultation through a web interface and available on demand for research purposes.

Keywords: Parallel corpora, text simplification, readability, linguistic complexity, misreading, poor-readers, dyslexia

1. Introduction

Reading is a complex cognitive task. Since reading comprehension is necessary for all school learning activities, poor reading and comprehension skills compromise children's academic and professional success. Typical readers also tends to progress quickly in reading because, as the process becomes more and more automatized, they enter a virtuous circle in which good reading comprehension skills boosts word identification and vice-versa (Stanovich et al., 1986; Stanovich, 2009). On the contrary, a child facing difficulties will tend to read less and therefore will not enter this virtuous circle. His/her reading difficulties will increase as the grade level becomes more demanding in terms of reading speed and comprehension (Tunmer and Hoover, 2019).

Given that reading comprehension skills of French-speaking students have decreased over recent years (Mullis et al., 2017), we have decided to address this issue in the framework on the Alector project[1]. Our aim was to develop and to test resources that make it possible to propose simplified texts to children facing problems in reading. For these children, text simplification might be a powerful and possibly the only way to leverage document accessibility. The idea is not to impoverish written language, but to propose simplified versions of a given text that convey the exact same meaning. The main assumption is that the simplification of a text will allow children with reading difficulties to eventually get through a text and thus discover the pleasure of reading through understanding what they actually read. This will allow them to enter the above mentioned virtuous circle, whereby word recognition and decoding skills are trained through reading more. The promise of this enterprise is that training children on simpler texts will lower their give-up threshold and improve their decoding, word recognition and comprehension skills, which ultimately would allow them to move on to more complex texts.

In order to test our hypothesis on text simplification and readability, we compiled a corpus of 183 texts (including 79 authentic texts), which was tested in schools during a three-year study. In this paper, we describe the corpus, its possibilities of use, and its availability. The resource is mainly addressed to a community of professionals interested in helping French-speaking learners who struggle with learning to read. It could also be of interest for research, i.e. for developing and training automatic text simplification systems.

The paper is organized as follows. In Section 2., we give an overview of related work (currently available simplified corpora and annotated corpora with errors). In Section 3., we specify how the corpus was created and provide quantitative details about it. Section 4. describes how a sub-part of the corpus was annotated with reading errors from poor and dyslexic readers.

2. Related work

The use of corpora is essential in many domains for different purposes. For reading, there are a number of standardized reading tests such as the International Reading Tests (IReST) (Vital-Durand, 2011) which exists in a variety of languages. However, standardized or specifically annotated corpora (i.e., with errors) are very costly to build and not al-

[1] https://alectorsite.wordpress.com/

ways available. In this section, we first report on resources for text simplification that are similar to ours in the sense that they include parallel original and simplified texts. Second, we discuss previous resources having reading errors annotated.

2.1. Parallel Corpora for Text Simplification

Researchers in supervised text simplification have used English Wikipedia – Simple English Wikipedia (EW–SEW) sentence-aligned corpora for training the systems (Štajner and Nisioi, 2018). Zhu et al. (2010) were the first to investigate this way of collecting simpler versions of texts. Soon after, Coster and Kauchak (2011) built a corpus of 137K aligned sentence pairs and computed transformations to compare original to simplified sentences (rewordings, deletions, reorders, mergers and splits). This approach received much attention from researchers working on text simplification in English, until it got criticized by Xu et al. (2015). More recently, (Xu et al., 2015) have advocated for the use of the Newsela corpus (Newsela, 2016) in automatic text simplification and demonstrated its value. This corpus was initially developed by editors and intended for learners of English. The data-set (version 2016-01-29.1) is composed of 10,787 news articles in English: 1,911 articles in their original form and in 4 equivalent versions rewritten by humans to suit different reading levels. Having different versions of an original article offers a great potential to study linguistic transformations, which explains why the automatic text simplification community was eager to use this resource. However, the way levels were defined - using the Lexile formula (Stenner, 1996) - should be subject to caution. To the best of our knowledge, there is no equivalent corpus for other languages offering the possibility to align sentences at different levels of reading complexity.

2.2. Corpora of Reading Errors

Apart from the use of parallel simplified corpora as a gold standard for text simplification, a number of studies have resorted to empirical measures of cognitive difficulty when reading in a native or foreign language including, but not exhaustively, eye-tracking data on readers suffering from autism spectrum disorder (Yaneva et al., 2016; Štajner et al., 2017) or from dyslexia (Bingel et al., 2018), as well as subjective annotations of difficult words (Paetzold and Specia, 2016; Tack et al., 2016; Yimam et al., 2018).

As far as reading errors are concerned, the situation can be considered unsatisfactory given that we find a limited availability of resources of this type. For dyslexia, some examples of corpora with writing errors have been compiled, for instance that of (Pedler, 2007), a dataset of productions of dyslexic English readers (3,134 words and 363 errors), or Dyslist (Rello et al., 2014), a corpus of 83 texts written in Spanish by dyslexic children (with 887 misspelled words). Such corpora can be used as source of knowledge to study different aspects of dyslexia. They can also be used to develop tools such as spellcheckers and games, and for screening with applications for readers (Rauschenberger et al., 2019), e.g. Dytective (Rello et al., 2016) a web-based game with different stages to detect dyslexia with machine learning prediction models.

Nb ORIG	Nb SIMP	Level	Type of corpus	
10	15	IReST	1 LIT	9 SCI
25	45	CE1	15 LIT	10 SCI
24	24	CE2	14 LIT	10 SCI
20	20	CM1	10 LIT	10 SCI
79	104		40	39

Table 1: Distribution of the original versions (LITerary and SCIentific texts) across primary school levels: CE1 (second grade), CE2 (third grade), CM1 (fourth grade).

Nb	Type of Simplification
79	Lexical, morpho-syntactic, discourse
15	Lexical simplification only
10	Syntactic simplification only
104	Total simplified versions

Table 2: Type and number of simplified versions in the corpus.

To the best of our knowledge, there is currently no available data set for French with reading errors of poor readers and dyslexics.

3. Alector Corpus: Simplified Texts in French

The Alector corpus is a collection of 79 original literary (tales, stories) and scientific (documentary) texts along with their simplified equivalents (see Table 1). The texts were chosen among a variety of materials available for students in French primary schools[2]. We targeted second to fourth grades, i.e. beginning readers (in the French educational system, this corresponds to *cours élémentaire 1* and *2* (CE1 and CE2), and *cours moyen* 1 (CM1), as described in Table 1). We also included the French version of the IReST corpus (Vital-Durand, 2011), a set of 10 standardized texts usually used for assessment of reading performances.

Our focus was on literary and scientific genres. While literary texts reflect the world view and the sensitivity of its author with a language that emphasizes the aesthetics, the rhythm, etc. (they highlight the poetic or expressive function of the language), scientific (documentary) texts aim at explaining or describing a scientific or technological causality, they are descriptive and explanatory with a logical structure based on scientific reasoning.

All the 79 original texts underwent simplifications at four linguistic levels, namely lexical, morphological, syntactic, and discursive (see Section 3.1.). In addition, 5 IReST and 15 CE1 texts were also simplified only at the lexical level, whereas 10 CE1 texts were simplified using only syntactic strategies. As a result, the total amount of simplified texts is 104, as detailed in Table 2.

[2]In more detail, we selected extracts from books in the Antoon Krings' collection "Drôles de petites bêtes", from the "J'aime Lire" collection or from Goscinny Sempé's book "Le Petit Nicolas". Scientific extracts were selected from Wapiti, Bibliothèque de Travail Junior (BTJ) or Images DOC, which are standard scientific magazines in French for young readers.

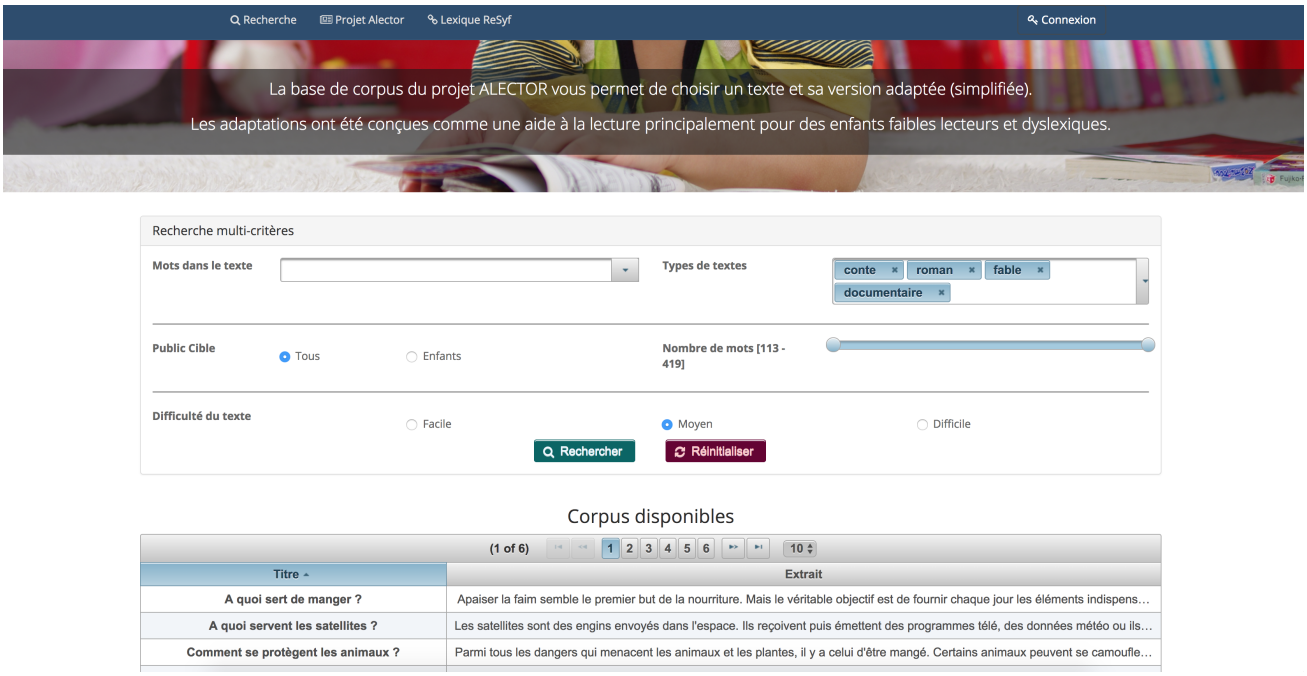

Figure 1: Interface for searching a corpus.

3.1. Manual Simplifications

Text simplification is defined as a reduction of the complexity of a text while preserving its original content (Saggion, 2017). By doing this, the text may be more easily read and understood by a target audience (in our work, we focus on the needs of poor-readers and dyslexic children).

From an initial set of 79 original texts, 104 different simplified versions were manually carried out by a group of researchers in educational sciences, cognitive psychology, linguistics and speech therapy. The objectives were twofold: (a) to keep the simplified text as close as possible to the original version; (b) to consider only linguistic transformations that could be later implemented in an automatic text simplification system (e.g. lexical and coreference chains substitutions, paraphrasing). The resulting simplified versions, along with their originals, were tested in five classes during current reading activities (children read the texts without knowing that they were reading adapted versions). On the other hand, simplified versions will be used as reference (gold-standard) to train and to evaluate a text simplification system. To our knowledge, despite Vikidia free adaptations of Wikipedia, our corpus is the first one to propose simplified versions of original texts in French.

Linguistically speaking, manual simplifications were done at four levels or dimensions (Gala et al., 2018): lexical (lexical substitutions), morphological (grammatical changes), syntactic (phrase structure adaptations), discourse (anaphora resolution). Guidelines were established following recommendations in the literature and after studying edited material for children (with and without dyslexia) and for illiterate adults (i.e. belgian collection *La Traversée*[3]).

a. LEXICAL REPLACEMENTS were performed using Manulex[4] (Lété et al., 2004), ReSyf[5] (Billami et al., 2018) and Lexique3[6] (New et al., 2001), three available lexical resources in French that contain indications of the presence of a word in a school level (Manulex), reading difficulty grades (ReSyf) and word frequencies in standard oral and written French (Lexique 3). We took into account specifications already defined in Gala and Ziegler (2016) and François et al. (2016). For instance, long and less frequent words, with irregular graphemes and complex syllable structures were modified by simpler synonyms: e.g. *volumineux* (unwieldy) by *gros* (big) and *ressemblaient* (seemed) by *étaient* (were).

b. MORPHOLOGICAL SIMPLIFICATIONS dealt with complex verb forms replacements: e.g., *elle peut devenir* ('she can become') was replaced by *elle devient* ('she becomes'); replacements in a same morphological family by more frequent equivalents in the same family, e.g., *construction* replaced by the infinitive form *construire* ('to build'), and diminutive morphemes deletion, e.g., *maisonnette* ('little house') became *maison* ('house').

c. SYNTACTIC TRANSFORMATIONS included changes on the sentence structures, such as deletion of subordinates and of complex modifiers, transformation of passive voice to active voice or of negative sentences to positive, etc. For instance, the original construction *C'est le vent qui apporte la pluie* ('it is the wind that brings the rain') becomes *Le vent apporte la pluie*

[3] http://www.lire-et-ecrire.be/latraversee

[4] http://www.manulex.org/

[5] https://cental.uclouvain.be/resyf/

[6] http://www.lexique.org/

Figure 2: Interface of visualisation with medium font size, big interlinear spacing and medium intercharacter spacing.

('the wind brings the rain'), the later keeps the standard subject-verb-object (SVO) order.

d. DISCOURSE SIMPLIFICATION mainly dealt with replacements of pronouns with the referenced expression, e.g., *il y plongea la main* ('he thrusted his hand') by *il mit la main* ('he put his hand') and the pronoun *y* is replaced by the full referent *dans son cartable* ('in his schoolbag').

Details on all types of transformations can be found in the simplification guidelines, soon available on the webpage of the Alector project (see Footnote 1).

3.2. Corpus Analysis

A statistical analysis of the Alector corpus was carried out. The corpus contains a total of 52,704 tokens, with a global average number of 288 tokens/text, distributed as showed in Table 3. When looking in more details within the different components of the corpus – defined by genre (LIT *vs* SCI), condition (original *vs* simplified) and levels (CE1, CE2, CM1) –, several interesting facts appear. First, simplified texts are shorter on average (275 words/text) than original texts (306 words/text). This is the case for every grade level as can be seen when one compares the two total columns in Table 3. Second, it is also obvious that literary texts are on average longer than scientific documents not only for original versions (339 *vs* 271 words/text), but also for their simplified equivalents (313 *vs* 239 words/text). Finally, we also observed that texts become longer on average as their grade level increases, both for original and for simplified versions.

In Table 4, we can notice that scientific documents, although shorter, tend to have longer sentences either in the original (16.1 *vs* 15 words/sentence) or in the simplified version (13.5 *vs* 11.2 words/sentence). This difference between genre is more noteworthy as the grade level increases. As regards the grade level in general, original versions of texts do not seem to vary much as regards to sentence length, whereas a slight trend can be seen in simplified versions. However, Table 4 clearly shows that the simplification process has reduced sentence length (15.5 *vs* 12.4 words/sentence).

We also report the proportion of nouns in the text, which roughly corresponds to the conceptual density of texts. Without surprise, scientific documents include a higher density of nouns, as they refer to various notions and realities. There also seems to occur an increasing proportion of nouns as the grade level rises, with the notable exception of the IResT literary text. Finally, it is interesting to note that the proportion of nouns does not decrease as a result of the simplification process (23% *vs* 22.6%). Most likely, this can be explained by the attempt to make explicit some of the pronominal anaphora, using the referent instead of the pronoun. In sum, it is clear that the simplified documents have different properties from the original texts.

3.3. Availability

The Alector corpus is available through a web interface. A text can be searched from a keyword, or by introducing some predefined labels on the kind of document (narrative, documentary, etc.). The number of words in the text or its difficulty to be read by the students can also be selected (Figure 1).

The "text difficulty" score (easy, medium, difficult) is based on the reading and comprehension tests of 164 students. We have used a normalized average of two factors, namely the reading speed and the comprehension scores obtained from each student.

The result of the search through the web interface is a list

Level	Original texts			Simplified texts		
	LIT	**SCI**	**Total**	**LIT**	**SCI**	**Total**
IResT	1 (135)	9 (141)	10 (140)	1 (115)	9 (131)	10 (129)
CE1	15 (282)	10 (243)	25 (266)	25 (276)	25 (211))	50 (244)
CE2	14 (363)	10 (303)	24 (338)	14 (340)	10 (288)	24 (318)
CM1	10 (413)	10 (385)	20 (399)	10 (388)	10 (358)	20 (373)
Total	40 (339)	39 (271)	79 (306)	50 (313)	54 (239)	104 (275)

Table 3: Number of texts per level, with the average number of words for each component.

Level	Original texts			Simplified texts		
	LIT	**SCI**	**Total**	**LIT**	**SCI**	**Total**
IResT	16.9 / 28.1%	16.2 / 21.6%	16.3 / 22.3%	14.4 / 27.8%	13.7 21.8%	10 / 22.4%
CE1	15.9 / 19.8 %	15 / 24.7%	15.6 / 21.7%	10.2 / 20.2%	12.9 / 24.9 %	11.6 / 22.5 %
CE2	14.9 / 20.2 %	15.5 / 26.4%	15.2 / 22.8%	12.4 / 20.8%	13.8 / 27%	12.9 / 23.4 %
CM1	13.5 / 22%	17.8 / 25.4%	15.6 / 23.7%	11.9 / 22.6%	14.4 / 26.3%	13.2 / 24.4 %
Total	15 / 20.7%	16.1 / 24.6%	15.5 / 22.6%	11.2 / 21 %	13.5 / 25 %	12.4 / 23%

Table 4: Average number of words per sentence and percentage of nouns per level.

of texts corresponding to the selected criteria. When the user clicks on the beginning of the text s/he has access to the parallel corpora (Figure 2).

Following the guidelines developed in the project and based on existing guidelines for other languages, e.g. Spanish (Rello, 2014) or English (British Dyslexia Association[7]), the interface enables different possibilities of visualization in terms of font (Open Dyslexic), interline spacing and intercharacter spacing (Zorzi et al., 2012).

4. A Subcorpus with Alignments of Misreadings

From the parallel corpus, twenty texts (i.e., ten authentic texts with their simplified version) were selected for use in speech therapy interventions with reading-impaired children. Based on the transcripts of the read-aloud interventions (Section 4.1.), we aligned and aggregated all errors made by each subject with the correct word in the base text (Section 4.2.). As a result, we were able to identify the probability of misreading a word in an authentic and a simplified text in the sample of dyslexic children. Using this error probability as a measure of lexical difficulty, we found evidence that the manual lexical simplifications had a significant alleviating effect on reading difficulties (Section 4.3.).

4.1. Tests with Dyslexic Readers

Reading tests were conducted in a sample (N=21) of French-speaking children aged between 9 and 12 attending mainstream schools. The subjects had a reading delay of two and a half years on average. During the experiment, the participants were asked to read aloud 10 different texts, including 5 original and 5 simplified texts drawn from both the literary and scientific genres. The texts were relatively short, counting 296 (literary) and 124 (scientific) tokens on average. All texts were presented on a digital tablet, displaying the sentences one by one. The children had to click to move to the next sentence. After each text, the participants were asked to answer a reading comprehension test in a multiple-choice format (read by the therapist to avoid a reading bias when evaluating comprehension). The experiments were run in private practices of speech-language therapists in Marseille (France), between November 2017 and March 2018. The read-aloud data were recorded and transcribed by students majoring in speech therapy (Nandiegou and Reboul, 2018). They manually encoded reading errors using ad-hoc guidelines.

4.2. Aligning Reading Errors

In order to identify the reading errors in the original and simplified texts, we aggregated the read-aloud transcripts of all participating children. First, because the transcripts had not been encoded with specialized annotation or transcription software and had not been linked to their original text, we needed to find a way to align each one of the 210 transcripts (i.e., ten transcripts per participant) with the original text. The alignment was done on the level of the word, associating each word as it was read aloud with the word as it appeared in the text. To this end, a modified version of the Needleman and Wunsch (1970) sequence alignment algorithm was used. The simplicity of the algorithm did not seem problematic given that the read-aloud transcripts did not contain any major syntactic modifications. The modified version of the algorithm aligned two words by constructing a similarity matrix for each pair of sentences in the original text (o) and the read-aloud transcript (r), represented as a sequence of tokens. The similarity scores s between two tokens w_o and w_r were computed by the integration of the edit distance and the length of a word (1). The modified algorithm thus privileged an alignment of words and non-words that displayed a higher formal proximity

[7] https://www.bdadyslexia.org.uk/advice/employers/creating-a-dyslexia-friendly-workplace/dyslexia-friendly-style-guide

Original version (10 readings)											
TEXT	Voilà	que	tu	t'	**agenouilles**	devant	ce	tas	de	neige.
MISREADINGS		*qui*			***agenou*** ***agenoui*** ***agenouiller*** ***agenouillies*** ***angenouillies*** ***aquenoulés***		*cette* *ça*			
ERROR COUNT	0	1	0	0	**7**	0	2	0	0	0
ERROR PROB.	0.0	0.1	0.0	0.0	**0.7**	0.0	0.2	0.0	0.0	0.0
GLOSS	'Now'		'you are'		**'kneeling'**	'in front of'	'this'	'pile'	'of'	'snow'.
Simplified version (11 readings)											
TEXT	Voilà	que	tu	te	**mets**	**à**	**genoux**	devant	ce	tas	de	neige.
MISREADINGS				*me*					*ces* *te*			
ERROR COUNT	0	0	0	1	**0**	**0**	**0**	0	4	0	0	0
ERROR PROB.	0.0	0.0	0.0	0.09	**0.0**	**0.0**	**0.0**	0.0	0.36	0.0	0.0	0.0
GLOSS	'Now'		'you are'		**'kneeling'**			'in front of'	'this'	'pile'	'of'	'snow'.

Table 5: Example of alignment with reading errors on a lexically simplified text with substitutions (literary text 3).

(e.g., *jardin* and *jadi*).

$$s = \begin{cases} +1, & \text{if } w_o = w_r \\ -(1+\delta_{\text{Levenshtein}}(w_o, w_r)), & \text{if } w_o \neq w_r \\ -(1+|w|), & \text{as gap penalty} \end{cases} \quad (1)$$

Second, we aggregated all per-participant alignments so as to obtain a list of all different variants read aloud for each word attested in the original text. Table 5 shows an example of sentence in the original corpus in which each word is aligned with observed reading errors. We consider a misreading when a word is either mispronounced by the use of another word or non-word (i.e., substitutions) or skipped (i.e., deletions).[8] For each word, we compute the total number of reading errors $C_{\text{misreading}}(w)$ and the probability of the word being misread $C_{\text{misreading}}(w)/C_{\text{reading}}(w)$.

4.3. Analysis

There are two ways in which the aggregated misreadings can be useful to further our understanding of lexical difficulty. On the one hand, the data can be used to identify difficult words in a bottom-up manner, making use of empirical evidence of reading errors as the basis for data-driven lexical simplification. However, we find that further work needs to be done to obtain a valid heuristic of difficulty on this data. From the two excerpts listed in Tables 5 and 6, we see that not all reading errors appear to be equally indicative of lexical difficulty. Indeed, we find many misreadings to be highly idiosyncratic, with only one or two participants mispronouncing the word. This trend is confirmed when we look at the distribution of the probability to misread a word in the original and simplified texts. Figure 3 shows that more than half of the tokens appearing in both versions were correctly read by all subjects, whereas only a small percentage of tokens were mispronounced by all subjects. Also, from the examples in Tables 5 and 6, we see that not all errors might be equally indicative of reading problems. While problems can be unquestionably perceived in errors where either non-existent words (e.g., *aquenoulés*) or existing words with radically different grammatical categories (e.g., *jardin* [noun] instead of *jadis* [adverb]) are produced, a number of errors are concerned with a slight change of form without radically changing the grammatical category (e.g., *qui* instead of *que*). In future work, we therefore aim to further investigate how both this individual and gradual nature of lexical difficulty can be correctly accounted for.

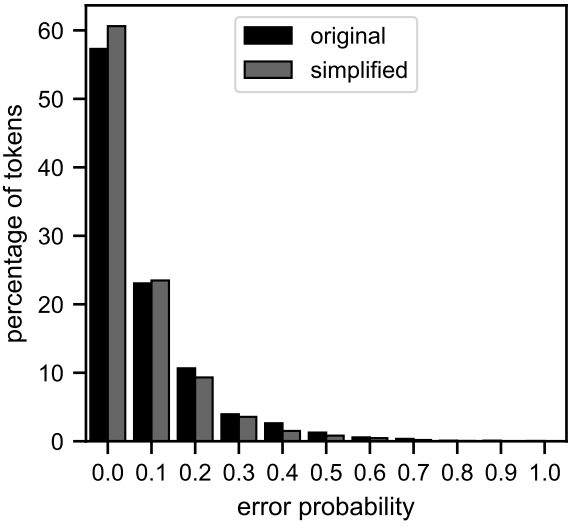

Figure 3: Distribution of the probability to misread a word token in the sample of poor-reading and dyslexic children

On the other hand, the data can also be used to validate the top-down identification of lexical difficulty as attested in the manually simplified corpus. From the examples in Tables 5 and 6, we see that the only words that were tar-

[8] For now, we do not consider insertions.

Original version (11 readings)												
TEXT	Il	y	avait	**jadis**	en	Irlande	un	homme	du	nom	de	Jack.
MISREADINGS			*avant* *était*	***jadi*** ***jardin*** ***jardis***		*Arlande*						
ERROR COUNT	0	0	2	**7**	0	1	0	0	0	0	0	0
ERROR PROBABILITY	0.0	0.0	0.18	**0.64**	0.0	0.09	0.0	0.0	0.0	0.0	0.0	0.0
GLOSS	'There'		'was'	**'once'**	'in'	'Ireland'	'a'	'man'		'named'		'Jack'.
Simplified version (10 readings)												
TEXT	Il	y	avait		en	Irlande	un	homme	du	nom	de	Jack.
MISREADINGS		∅			*un*	*rilande*			*de*			*Jean*
ERROR COUNT	0	1	0		1	1	0	0	2	0	0	1
ERROR PROBABILITY	0.0	0.1	0.0		0.1	0.1	0.0	0.0	0.2	0.0	0.0	0.1
GLOSS	'There'		'was'		'in'	'Ireland'	'a'	'man'		'named'		'Jack'.

Table 6: Example of alignment with reading errors on a lexically simplified text with deletions (literary text 4).

geted for simplification (i.e., *agenouiller* ['to kneel'] and *jadis* ['erstwhile']) are also the only words that we can consider to be difficult for the target reader population given the high error probability ($p = .70$ and $p = .64$, resp.). This leads us to believe that the manual simplifications were indeed necessary and that the substitutions were adequate.

In the following two sections, we aim to further substantiate this claim with statistical[9] evidence. Our first hypothesis will be that the lexical simplifications can be considered necessary if the words in the authentic texts that were initially considered to be difficult for the target population (and hence targeted for lexical simplification) will be more often misread than words that were not targeted (Section 4.3.1.). Our second hypothesis will be that lexical simplifications can be considered correctly estimated if there are less misreadings on the substitute words than on the original words (Section 4.3.2.).

4.3.1. Misreadings of words in authentic texts targeted for simplification

To confirm our first hypothesis, we focus on the words attested in the authentic texts only and divide them into into three different categories, viz., words that were targeted for simplification with either substitution or deletion and words that were not targeted. We then examined whether the probability of misreading words belonging to either category was significantly different, which was confirmed by a Kruskal-Wallis test, $\chi^2(2) = 199$, $p < .001$, $\varepsilon^2 = 0.089$. Post-hoc Dwass-Steel-Critchlow-Fligner comparisons (Figure 7) showed that the largest difference was attested between words that were substituted ($Mdn = 0.18$) and words that were not targeted for simplification ($Mdn = 0.00$), $W = 19.6$, $p < .001$. A significant difference was also observed between words that were deleted ($Mdn = 0.090$) and words that were not targeted, but the effect was smaller, $W = 5.30$, $p < .001$. We therefore conclude that all words identified for simplification were well-targeted. However, the need for simplification was not equally strong, as can be evidenced by the significant difference between the number of misreadings on words that were substituted and deleted, $W = 6.16$, $p < .001$. While removing superfluous words was also necessary to reduce reading difficulties, substituting core but difficult content words seemed even more crucial. In our final analysis, we will therefore focus on these substituted words and have a look at the effect of the chosen simplifications.

[9]From Figure 3, it can be seen that the distribution of error probabilities violates the normality assumption, which is confirmed by a Shapiro-Wilk test, $W = 0.67$, $p < .001$. We will therefore use non-parametric tests for all statistical analyses.

			DSCF Comparisons (W)	
Group	N	Mdn	substituted	deleted
substituted	192	0.18		
deleted	80	0.09	6.16 ***	
not targeted	1972	0.00	19.6 ***	5.30 ***

*** $p < .001$

Table 7: Post-hoc comparisons of the probability of making reading errors on words in authentic texts that were targeted or not targeted for simplification

4.3.2. Misreadings on lexical substitutions

To confirm our second hypothesis, we focus on the words that were simplified by means of substitution and compare the expected decrease in number of errors made before and after simplification. A pairwise Friedman rank sum test showed that there were significantly fewer misreadings after simplification ($Mdn = 0.090$) than on the word in the original text ($Mdn = 0.18$), ${X^2}_F(1) = 40.6$, $p < .001$. Moreover, the effect on decreasing the probability of misreading the word in the simplified version was large (Kendall's $W = .527$). Consequently, we find that the words targeted for simplification were not only well-chosen, but also substituted with substantially easier alternatives, hence enhancing the readability of the texts for our targeted readers (i.e., dyslexic children).

5. Conclusion

In this paper, we presented the Alector corpus, a resource of parallel texts for French learners struggling with reading. The corpus is made of 79 original literary and scientific

texts that have been manually simplified at different linguistic levels. The corpus also contains a part of texts that have been annotated with reading errors by children with dyslexia. The data are available online and offer different criteria for searching and visualizing the texts. The entire corpus is available on demand for research purposes.
In the future, we plan to add the lexical resource ReSyf (Billami et al., 2018) to the corpus in order to highlight lexical difficulty and provide the reader with a list of graded synonyms, as advised by Rello (2014) for dyslexic learners. Moreover, we will continue to use the corpus in reading tests to extend our analysis of reading errors. Finally, the Alector corpus will be used as a gold standard for evaluating current developments in automatic text simplification for French.

Acknowledgements

This research was funded by the French *Agence Nationale pour la Recherche* (ANR), through the Alector project (ANR-16-CE28-0005). It also received the financial support of the Belgian National Fund for Scientific Research (F.R.S.-FNRS). We deeply thank Solange Lâm (Université de Bruxelles) and Carlos Ramisch (Aix Marseille Université, Laboratoire d'Informatique Fondamentale) for the current version of the interface. We also thank Aurore Brunel, Mathilde Combes, Marie Nandiegou and Stella Reboul, speech therapists that actively participated in the manual simplifications of parts of the texts and on the reading and comprehension tests with dyslexic children. We finally thank three anonymous reviewers for their thoughtful suggestions for improving the initial version of the paper.

References

Bibliographical References

Billami, M. B., François, T., and Gala, N. (2018). ReSyf: a French lexicon with ranked synonyms. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*, pages 2570–2581.

Bingel, J., Barrett, M., and Klerke, S. (2018). Predicting misreadings from gaze in children with reading difficulties. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 24–34, New Orleans, Louisiana, June. Association for Computational Linguistics.

Coster, W. and Kauchak, D. (2011). Simple English Wikipedia: A New Text Simplification task. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, Oregon, USA.

François, T., Billami, M. B., Gala, N., and Bernhard, D. (2016). Bleu, contusion, ecchymose : tri automatique de synonymes en fonction de leur difficulté. In *Actes de la conférence Traitement Automatique des Langues Naturelles*, pages 15–28.

Gala, N. and Ziegler, J. (2016). Reducing lexical complexity as a tool to increase text accessibility for children with dyslexia. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC), COLING 2014*, pages 59–66.

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

Lété, B., Sprenger-Charolles, L., and Colé, P. (2004). Manulex: a grade level lexical database from French elementary-school readers. *Behavior Research Methods, Instruments and Computers*, 36:156–166.

Mullis, I. V. S., Martin, M. O., Foy, P., and Hooper, M. (2017). PIRLS 2016 International Results in Reading.

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

Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443 – 453.

New, B., Pallier, C., Ferrand, L., and Matos, R. (2001). Une base de données lexicales du français contemporain sur internet : Lexique^{TM}//a lexical database for contemporary French : Lexique. *L'Année psychologique*, 101:447–462.

Paetzold, G. and Specia, L. (2016). Semeval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569.

Pedler, J. (2007). *Computer Correction of Real-word Spelling Errors in Dyslexic Text.* Ph.D. thesis, Birkbeck College, London University.

Rauschenberger, M., Baeza–Yates, R., and Rello, L. (2019). Technologies for Dyslexia. In Yesilada Y. et al., editors, *Web Accessibility. Human–Computer Interaction Series*. Springer, London.

Rello, L., Baeza-Yates, R., and Llisterri, J. (2014). Dyslist: An annotated resource of dyslexic errors. In *Conference on Language Resources and Evaluation LREC-2014*, page 1289–1296, Reykjavick, Island.

Rello, L., Ballesteros, M., Abdullah, A., Serra, M., Sánchez, D., and Bigham, J. (2016). Dytective: diagnosing risk of dyslexia with a game. In *Pervasive health*, Cancun. ACM Press.

Rello, L. (2014). *DysWebxia : a text accessibility model for people with dyslexia*. Ph.D. thesis, Universitat Pompeu Fabra, Barcelona.

Saggion, H. (2017). *Automatic text simplification*, volume 10. Morgan & Claypool Publishers.

Štajner, S. and Nisioi, S. (2018). A detailed evaluation of neural sequence-to-sequence models for in-domain and cross-domain text simplification. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Štajner, S., Yaneva, V., Mitkov, R., and Ponzetto, S. P. (2017). Effects of Lexical Properties on Viewing Time per Word in Autistic and Neurotypical Readers. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 271–281,

Copenhagen, Denmark. Association for Computational Linguistics.

Stanovich, K. E., Nathan, R. G., and Vala-Rossi, M. (1986). Developmental changes in the cognitive correlates of reading ability and the developmental lag hypothesis. *Reading research quarterly*, x(x):267–283.

Stanovich, K. E. (2009). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Journal of education*, 189(1-2):23–25.

Stenner, A. (1996). Measuring reading comprehension with the lexile framework. In *Fourth North American Conference on Adolescent/Adult Literacy*.

Tack, A., François, T., Ligozat, A.-L., and Fairon, C. (2016). Evaluating Lexical Simplification and Vocabulary Knowledge for Learners of French: Possibilities of Using the FLELex Resource. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 230–236, Portorož, Slovenia, May. European Language Resources Association.

Tunmer, W. E. and Hoover, W. A. (2019). The cognitive foundations of learning to read: a framework for preventing and remediating reading difficulties. *Australian Journal of Learning Difficulties*, 24(1):75–93.

Xu, W., Callison-Burch, C., and Napoles, C. (2015). Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Yaneva, V., Temnikova, I., and Mitkov, R. (2016). A Corpus of Text Data and Gaze Fixations from Autistic and Non-Autistic Adults. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, May. European Language Resources Association (ELRA).

Yimam, S. M., Biemann, C., Malmasi, S., Paetzold, G., Specia, L., Štajner, S., Tack, A., and Zampieri, M. (2018). A Report on the Complex Word Identification Shared Task 2018. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 66–78, New Orleans, Louisiana, June. Association for Computational Linguistics.

Zhu, Z., Bernhard, D., and Gurevych, I. (2010). A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1353–1361. Association for Computational Linguistics.

Zorzi, M., Barbiero, C., Facoetti, A., Lonciari, I., Carrozzi, M., Montico, M., and Ziegler, J. C. (2012). Extra-large letter spacing improves reading in dyslexia. 109(28):11455–11459.

Language Resource References

Newsela. (2016). *Newsela Article Corpus 2016-01-29*.

Vital-Durand, F. (2011). *International Reading Speed Texts IResT (French version)*.