Document généré le 8 mai 2024 20:35

Mesure et évaluation en éducation

ALSI : un nouvel outil d'analyse automatisée de la complexité linguistique pour le français québécois

Guillaume Loignon

Volume 44, numéro 3, 2021

L'évaluation des compétences langagières : enjeux et perspectives

Réception : 12 octobre 2021
Version finale : 16 mars 2022
Acceptation : 18 mai 2022

URI : https://id.erudit.org/iderudit/1093065ar
DOI : https://doi.org/10.7202/1093065ar

Aller au sommaire du numéro

Éditeur(s)

ADMEE-Canada

ISSN

0823-3993 (imprimé)
2368-2000 (numérique)

Découvrir la revue

Citer cet article

Loignon, G. (2021). ALSI : un nouvel outil d'analyse automatisée de la complexité linguistique pour le français québécois. *Mesure et évaluation en éducation, 44*(3), 29–57. https://doi.org/10.7202/1093065ar

Résumé de l'article

Estimer la complexité linguistique est un aspect important de la mesure et de l'évaluation de l'éducation qui peut servir, par exemple, à contrôler la variance indésirable attribuable à la langue ou à fournir aux élèves des textes propices à l'apprentissage. Des techniques de traitement automatique des langues permettent d'extraire différents attributs (features) qui reflètent la complexité du vocabulaire et de la structure des phrases. Dans cet article, nous présentons un nouvel outil appelé ALSI (Analyseur Lexico-Syntaxique Intégré). Nous résumons le fonctionnement de l'outil et présentons les types d'attributs qu'il peut extraire. Nous appliquons ensuite ALSI à 600 textes utilisés dans les écoles primaires et secondaires du Québec et analysons les corrélations entre les attributs et le niveau scolaire associé au texte. Les résultats montrent le potentiel d'ALSI pour la modélisation de la complexité des textes français.

Tous droits réservés © ADMEE-Canada - Université Laval, 2022

Ce document est protégé par la loi sur le droit d'auteur. L'utilisation des services d'Érudit (y compris la reproduction) est assujettie à sa politique d'utilisation que vous pouvez consulter en ligne.

https://apropos.erudit.org/fr/usagers/politique-dutilisation/

Cet article est diffusé et préservé par Érudit.

Érudit est un consortium interuniversitaire sans but lucratif composé de l'Université de Montréal, l'Université Laval et l'Université du Québec à Montréal. Il a pour mission la promotion et la valorisation de la recherche.

https://www.erudit.org/fr/

ALSI : un nouvel outil d'analyse automatisée de la complexité linguistique pour le français québécois

Guillaume Loignon
Université du Québec à Montréal

Mots-clés : analyse de corpus, traitement automatique du langage naturel, lisibilité, psycholinguistique, français

Estimer la complexité linguistique est un aspect important de la mesure et de l'évaluation de l'éducation qui peut servir, par exemple, à *contrôler la variance indésirable attribuable à la langue ou* à *fournir aux élèves des textes propices à l'apprentissage. Des techniques de traitement automatique des langues permettent d'extraire différents attributs (*features*) qui reflètent la complexité du vocabulaire et de la structure des phrases. Dans cet article, nous présentons un nouvel outil appelé ALSI (Analyseur Lexico-Syntaxique Intégré). Nous résumons le fonctionnement de l'outil et présentons les types d'attributs qu'il peut extraire. Nous appliquons ensuite ALSI à 600 textes utilisés dans les écoles primaires et secondaires du Québec et analysons les corrélations entre les attributs et le niveau scolaire associé au texte. Les résultats montrent le potentiel d'ALSI pour la modélisation de la complexité des textes français.*

Key words: corpus analysis, natural language processing, readability, psycholinguistics, French

Estimating language complexity is an important aspect of educational measurement and assessment that can be used, for instance, to control unwanted variance due to language, or to provide students with texts that are conducive to learning. Automatic language processing techniques can be used to extract various linguistic features that reflect the complexity of vocabulary and sentence structure. In this paper, we present a new tool called ILSA (Integrated Lexico-Syntactic Analyzer), which we developed for research and educational applications. We summarize how the tool works and present the types of attributes it can extract. We then apply ALSI to 600 texts used in Quebec elementary and secondary schools and analyze the correlations between the attributes and the school grade associated with the text. The results show the potential of ALSI for modeling the complexity of French texts.

Palavras-chave: atributos de texto, análise de corpus, processamento automático da linguagem natural, legibilidade, francês

Estimar a complexidade linguística é um aspeto importante da medição e da avaliação educacional que pode ser usado, por exemplo, para controlar a variação indesejada devido à linguagem ou para fornecer aos alunos textos que conduzam à aprendizagem. As técnicas de processamento automático de linguagem permitem extrair diferentes atributos (features) que refletem a complexidade do vocabulário e a estrutura das frases. Neste artigo, apresentamos uma nova ferramenta chamada ALSI (Analisador Léxico-Sintético Integrado). Resumimos o funcionamento da ferramenta e apresentamos os tipos de atributos que ela pode extrair. Em seguida, aplicamos o ALSI a 600 textos usados em escolas primárias e secundárias no Québec e analisamos as correlações entre os atributos e o ano letivo associado ao texto. Os resultados mostram o potencial do ALSI para a modelização da complexidade dos textos em francês.

Introduction

ALSI, pour Analyseur lexico-syntaxique intégré, est un outil automatisé de traitement du langage naturel qui extrait un ensemble d'attributs caractérisant la complexité intrinsèque du texte. Nous avons créé ALSI pour répondre à certains besoins dans le domaine de la mesure et de l'évaluation en éducation. Par exemple, un outil d'analyse linguistique peut aider à sélectionner des textes appropriés selon l'âge des élèves et les objectifs pédagogiques. Un analyseur similaire, SATO-Calibrage (Daoust et al., 1996), est disponible actuellement sur le Web, mais date des années 1990 et n'a pu profiter des innovations théoriques et méthodologiques concernant les sources de difficultés du texte et leur mesure automatisée. ALSI s'appuie sur des avancées techniques et théoriques plus récentes, comme la base de données de l'Échelle québécoise de l'orthographe lexicale (ÉQOL) (Stanké et al., 2019), Manulex (Lété, 2004), de même que les travaux entourant l'outil anglo-saxon Coh-Metrix (McNamara & Graesser, 2011). Le présent article poursuit deux objectifs : d'abord présenter ALSI, son contexte théorique et ses fonctions, ensuite effectuer un premier essai de validation en analysant 600 textes utilisés aux niveaux primaire et secondaire au Québec.

La complexité linguistique en mesure et évaluation de l'éducation

En phase avec la théorie de la charge cognitive (Clevinger, 2014), on peut se représenter la complexité du texte comme émergeant de facteurs intrinsèques et extrinsèques. La complexité intrinsèque au texte est celle qui peut être ramenée à ses caractéristiques mesurables, appelées attributs (en anglais, *features*). La longueur des phrases est un exemple classique d'attribut du texte (Flesch, 1948; Szmrecsányi, 2004). La complexité extrinsèque dépend d'un ensemble de facteurs qui ne peuvent se mesurer à partir du texte, dont les caractéristiques du lecteur, l'intention de lecture, la situation, l'aide fournie au lecteur, etc. De manière similaire, Zakaluk et Samuels (1988) parlent de facteurs « en dehors de la tête » et « dans la tête ». Nous proposons en ce sens l'analogie d'un parcours à obstacles

dont la difficulté résulte à la fois des caractéristiques du parcours (attributs linguistiques) et de l'athlète (la personne lisant le texte). Modéliser la complexité du texte représente un défi important puisqu'il faut, en s'appuyant sur des mesures faites à partir du texte, émettre des hypothèses quant à ce qui serait susceptible d'augmenter la charge cognitive du lecteur.

L'analyse de la complexité linguistique a de multiples applications dans le domaine de l'éducation, notamment pour la sélection de textes et de manuels favorisant l'apprentissage en fonction des caractéristiques des élèves (Graesser et al., 2004). Il s'agit d'un aspect peu abordé, mais important de la démarche de conception des tests (Lane et al., 2015; McNamara et al., 2012 Visone, 2009). Contrôler les attributs linguistiques de l'item permet d'atténuer la variance indésirable attribuable à la langue. La variance indésirable (*construct irrelevant variance*) est le degré d'influence sur les scores de processus étrangers à l'objectif d'un test. Selon les *Standards*, la difficulté linguistique de l'item est l'une des sources potentielles de variance indésirable qu'il faut contrôler lorsque c'est possible (Joint Committee on Standards for Educational and Psychological Testing, 2014; Lane et al., 2015). L'influence de la langue sur la réponse à l'item a été démontrée par plusieurs travaux. Par exemple, des études réalisées en contexte suédois (Persson, 2016), sud-africain (Dempster & Reddy, 2007) et américain (Martiniello, 2009) ont révélé la présence de biais linguistiques dans des tests standardisés de mathématiques.

Les aspects linguistiques de l'évaluation ne sont pas uniquement une source de variance indésirable. Leur influence peut être *désirable* lorsque la langue fait partie, ou ne peut être séparée, de la compétence évaluée (Avenia-Tapper & Llosa, 2015). Par exemple, des études de traitement automatique des langues résumées par Crossley (2020) ont montré une association statistique entre le score attribué à la qualité de l'écriture en anglais langue seconde et certains attributs linguistiques portant sur la complexité des phrases. Ce type d'études soutient l'idée que le traitement automatique des langues peut aider à mesurer la complexité linguistique.

Mesurer la complexité linguistique

La complexité du texte en langue anglaise a depuis longtemps été mesurée par des formules de lisibilité s'appuyant sur des attributs dits « de surface » (Benjamin, 2012; Feng et al., 2010), typiquement la longueur moyenne du mot et de la phrase. La situation est similaire du côté francophone : quelques formules de lisibilité conçues pour l'anglais ont été adaptées pour la langue française, d'autres créées spécifiquement pour le français

(Mesnager, 1989). L'usage intensif des attributs de surface a été grandement critiqué, principalement parce que ceux-ci tiennent peu compte d'éléments de complexité découlant du caractère subjectif de la lecture (Boyer, 1992). Les comptes-rendus historiques sur la modélisation de la complexité linguistique concluent assez unanimement que l'utilisation d'attributs de surface n'est pas suffisante pour mesurer correctement la complexité linguistique, et proposent plutôt de s'orienter vers des attributs théorisés en psycholinguistique (Boyer, 1992; François, 2015; Kintsch & Vipond, 2014; McNamara et al., 2012; Zakaluk & Samuels, 1988). C'est dans cette perspective que nous avons créé l'analyseur linguistique présenté dans cette étude.

Pourquoi créer un nouvel outil?

ALSI, pour analyseur lexico-syntaxique intégré, est un outil de traitement automatique du langage naturel créé dans l'objectif de modéliser la complexité du texte en français utilisé en enseignement primaire et secondaire. Des outils ont déjà été proposés dans des visées similaires; nous en résumons les caractéristiques. Développée dans les années 1990, la plateforme québécoise d'analyse textuelle SATO-Calibrage (Daoust et al., 1996) est toujours disponible en ligne. SATO-Calibrage extrait des attributs relativement simples, comparativement aux outils anglo-saxons tels Coh-Metrix (Grasser et al. 2011), que nous décrivons dans les sections suivantes de cet article. *DMesure* et *Amesure* s'appuient sur des travaux de linguistique computationnelle (François, 2009; François & Fairon, 2012; François & Miltsakaki, 2012). *Dmesure* classifie des textes en français langue seconde selon les six niveaux du Cadre européen commun de référence. *Amesure* se spécialise dans l'estimation de la lisibilité de documents en français des affaires, ce qui en réduit l'intérêt en éducation primaire et secondaire. *ReaderBench* a été conçu dans une approche similaire à *Dmesure* pour analyser du texte en plusieurs langues, dont le français (Dascalu et al., 2013) et produit un grand nombre d'attributs linguistiques. *Dmesure* et *ReaderBench* n'étaient cependant plus disponibles au moment de publier le présent article, motivant la création d'un nouvel analyseur de texte en français répondant à des besoins actuels.

La présente étude

L'objectif général de cette étude est de présenter un nouvel outil d'analyse de la complexité linguistique et d'énoncer en sa faveur un argumentaire de validité (Loye, 2018) en deux parties. La première partie est une vue d'ensemble de l'outil ALSI, qui résume son fonctionnement général. Elle décrit les types d'attributs extraits et les procédures utilisées pour les

extraire. Nous nous appuyons sur des travaux de psycholinguistique et de linguistique computationnelle pour expliquer ce qui relie ces attributs à la complexité linguistique. La deuxième partie explique l'utilisation d'ALSI sur un corpus de 600 textes. Nous identifions des attributs ayant un potentiel intéressant pour estimer la difficulté de textes, exprimée sur l'échelle des 11 années scolaires du système primaire et secondaire québécois.

L'outil ALSI

Fonctionnement général d'ALSI

ALSI est un outil de traitement automatique du langage naturel spécialisé dans l'extraction d'attributs caractérisant la complexité linguistique des textes français. Le texte est d'abord décodé, puis transformé en une liste de mots (*tokens*) annotés[1]. Les annotations incluent le lemme (forme canonique du mot), la partie du discours ou classe de mot (nom, verbe, adjectif, etc.), les relations hiérarchiques entre les mots et des informations périphériques (temps verbaux, genre, nombre, etc.). D'autres annotations sont ajoutées par croisement avec des bases de données spécialisées que nous décrivons plus loin. Le résultat, illustré à la Figure 1, constitue une matrice dont chaque ligne représente un mot et chaque colonne une information ou une mesure portant sur le mot.

Des opérations sur la matrice de mots produisent ensuite divers attributs linguistiques au niveau de la phrase et de l'ensemble du texte. Par exemple, le nombre de mots divisé par le nombre de phrases donne un attribut linguistique : la longueur moyenne des phrases du texte. De même, analyser la matrice des mots permet d'identifier lesquels sont des verbes conjugués; en divisant leur nombre par le nombre de phrases du texte, on obtient un attribut indiquant le nombre moyen de verbes conjugués par phrase. Nous détaillons dans ce qui suit les types d'attributs extraits par ALSI ainsi que leurs bases théoriques et leurs procédures d'extraction.

Typologie des attributs extraits par ALSI

Les attributs extraits par ALSI s'inscrivent dans une typologie simple ayant pour but de regrouper les attributs en catégories cohérentes reposant sur des caractéristiques similaires du texte, tout en exprimant une vision

1. Le *token*, ou jeton, est la plus petite unité linguistique extraite par l'analyseur; pour simplifier, nous employons le terme « mot » dans la suite de l'article.

Figure 1
Exemple d'analyse automatique d'un extrait de texte

Il a perdu son oncle, il y a quelques mois. **J'ai couru pour ne pas manquer le départ.** Cette hâte, cette course, c'est à cause de tout cela sans doute, ajouté aux cahots, à l'odeur de l'essence, à la réverbération de la route et du ciel, que je me suis assoupi. (…)

#	Token	Lemme	Partie du discours	Fréquence selon ÉQOL	Nombre de caractères	…
1	J'	il	PRON	74,2	2	
2	ai	avoir	AUX	72,1	2	
3	couru	courir	VERB	53,6	5	
4	pour	pour	ADP	79,7	4	
5	ne	ne	ADV	74,3	2	
6	pas	pas	ADV	76,3	3	
7	manquer	manquer	VERB	54,2	7	
8	le	le	DET	85,1	2	
9	départ	départ	NOUN	61,4	6	
10	.	.	PUNCT	--	--	
…						

Note. Décodage, lemmatisation et identification de la partie du discours avec la librairie *UDPipe* pour R (Straka et al., 2016). Fréquence et longueur des mots tirées de la base de données ÉQOL (Stanké et al., 2019).

nuancée de la complexité de ce dernier. Cette typologie est composée de deux dimensions : 1) la complexité lexicale, qui est associée aux mots du texte, et 2) la complexité syntaxique, qui est associée à l'agencement des mots en phrases et au rôle que jouent les mots dans la phrase. Ce choix est motivé par le fait que la complexité du texte est fréquemment définie comme l'intersection d'une composante lexicale et d'une composante syntaxique (Ravid, 2005), une division cohérente avec le cadre conceptuel *Simple View of Reading* (Gough & Tunmer, 1986) tout en étant en phase avec le choix d'attributs des plateformes d'analyse de langue anglaise ATOS (Milone, 2014) et Lexile (Smith et al., 1989). Tel que l'illustre le Tableau 1, les deux dimensions sont subdivisées en trois strates : 1) attributs de surface, 2) attributs dont l'extraction nécessite le recours à des bases de données lexicales ou à une procédure automatisée d'analyse syntaxique et 3) attributs qui qualifient la complexité linguistique de manière plus globale (par exemple, les mesures de cohésion).

Tableau 1
Typologie des attributs extraits par ALSI

	Lexique	**Syntaxe**
Strate 1 *Surface*	Mesures de longueur orthographique (nombre de caractères) ou syllabique (nombre de syllabes)	Longueur de la phrase, nombre de virgules
Strate 2 *Intermédiaire*	Fréquence du mot ou du lemme dans un lexique de référence ; âge d'exposition au mot dans le cursus scolaire	Présence de certains constituants de la phrase (ex. : verbes conjugués) ; présence et longueur de syntagmes d'intérêt (ex. : subordonnée relative), hiérarchisation de la phrase
Strate 3 *Globale*	Diversité lexicale ; cohésion lexicale	Cohésion syntaxique

Les attributs extraits par ALSI et considérés dans le présent article sont décrits au Tableau 4 présenté dans le matériel supplémentaire de l'article. Notons qu'ALSI emploie une nomenclature où le suffixe indique quelle était la fonction d'agrégation employée : *m* est une moyenne, *logm* est la moyenne des valeurs transformées sur une échelle logarithmique, *p* est une proportion, *90* est le 90^{e} percentile et *i* est un indice.

Annotation du corpus

Les textes à analyser prennent initialement la forme des fichiers en format .txt, chaque fichier contenant un texte. Le décodage et l'annotation du texte utilisent la librairie *UDPipe* pour le langage R, version 0.8.9 (R Core Team, 2022; Wijffels, 2022). La typologie des annotations est celle du cadre *Universal Dependency* (De Marneffe et al., 2014). L'annotation avec *UDPipe* requiert un modèle du texte de langue française préentrainé par technique d'apprentissage machine. Ce modèle est ce qui permet d'identifier la partie du discours (nom, verbe, etc.) et les relations syntaxiques entre les mots. Le modèle utilisé était *French-GSD 2.5* (Guillaume et al., 2019).

Analyse lexicale

L'analyse lexicale produit des attributs estimant la difficulté associée aux mots. Dans sa première version, ALSI s'appuie sur trois lexiques de référence : Manulex, ÉQOL, et la Liste orthographique du ministère de

l'Éducation du Québec. Manulex (Lété, 2004) contient environ 49 000 mots et a été compilé à partir de 54 manuels scolaires (niveaux scolaires CP à CM2 du système français) représentant environ deux millions de mots. ÉQOL (Stanké et al., 2019) est un lexique créé pour le système scolaire québécois et contient 16 652 mots tirés de manuels et d'ouvrages de littérature jeunesse dont le niveau va de la 1re à la 6e année du primaire. La Liste orthographique du ministère de l'Éducation du Québec est disponible via le projet *Franqus* de l'Université de Sherbrooke et contient 3 314 mots classifiés en six niveaux scolaires allant de la 1re à la 6e année du primaire ou 4921 mots après ajout des formes plurielles manquantes pour les noms communs.

Pour les attributs portant sur les fréquences d'occurrence, ALSI emploie les indices de fréquence standardisée (standard frequency index). Les attributs lexicaux des strates 1 et 2 sont produits à partir du lexique (liste de mots uniques) du texte, chaque lexème ne comptant alors qu'une fois[2]. Si un mot est absent de Manulex ou d'ÉQOL, la fréquence manquante est imputée à l'aide de la méthode d'estimation de fréquence de Good-Turing (pour une explication, voir Gale et Sampson, 1995).

ALSI estime en outre la diversité lexicale, qui est la tendance à employer un vocabulaire diversifié, les textes plus simples ayant davantage tendance à réutiliser les mêmes mots. Plusieurs formules existent pour ce faire (Fergadiotis et al., 2015); ALSI calcule le rapport type-jeton (*type-token ratio*) et l'indice de Maas (1972). Le rapport type-jeton estime la diversité lexicale en divisant le nombre de mots uniques par le nombre total de mots (longueur du texte). L'indice de Maas est une mesure similaire, calculée selon cette formule, T étant le nombre total de mots et U le nombre de mots uniques :

Indice de Maas

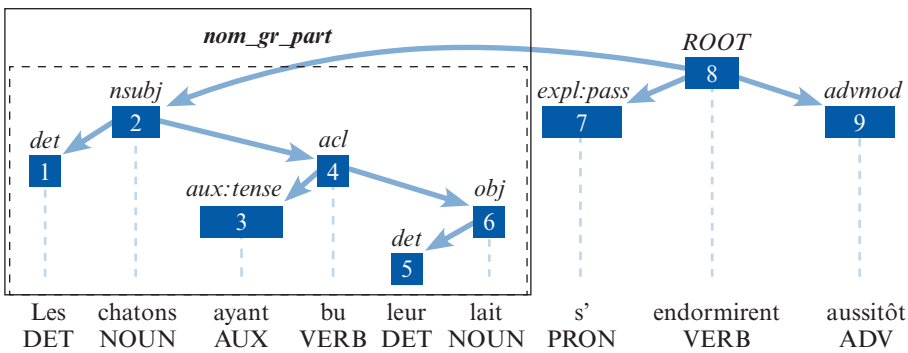
$$Mass^2 = \frac{logT - logU}{logT^2}$$

2. Les mots ayant les classes suivantes ne sont pas considérés : auxiliaires, noms propres, nombres, déterminants, symboles non alphanumériques.

Analyse syntaxique

Alors que les attributs portant sur la longueur des phrases sont calculés directement à partir de la liste annotée de mots (voir Figure 1), d'autres attributs syntaxiques requièrent des analyses supplémentaires. La hauteur (ou la profondeur) de la phrase, comptée en nombre de nœuds (*nodes*) est un indicateur de complexité syntaxique fréquemment employé (Sherstinova et al., 2020). Soit une phrase représentée comme un graphe hiérarchique, sa hauteur correspond au chemin le plus long reliant un mot à la racine de la phrase (Blache, 2010). ALSI utilise pour ce calcul l'arbre représentant les dépendances syntaxiques entre les mots. La Figure 2 présente un exemple d'arbre syntaxique dont la hauteur est 4, le chemin le plus long allant du mot *leur* à la racine de la phrase qui est *endormirent*.

Figure 2
Représentation graphique d'une phrase



Note. L'encadré indique un groupe nominal complexe, dans ce cas un nom avec groupe participial, détecté à l'aide de la librairie *rsyntax* (Welbers et al., 2020). Voir De Marneffe et al. (2014) pour la liste des sigles. Figure produite avec *rsyntax*.

ALSI extrait en outre des attributs portant sur la fréquence ou sur la longueur de constituants syntaxiques comme le groupe verbal, détectés avec la librairie *rsyntax* pour R (Welbers et al., 2020)[3]. Dans cette première version d'ALSI, nous avons ciblé les groupes verbaux et les groupes

3. Une autre solution d'analyse syntaxique a été proposée récemment par la librairie fsca pour le langage R (Vandeweerd, 2021), mais nous n'avions pas encore eu l'occasion de la tester au moment de publier cet article.

nominaux complexes. Le groupe verbal (GV) est opérationnalisé comme un groupe de mots dominé par un verbe conjugué. Le groupe nominal complexe (GNC) est opérationnalisé dans ALSI comme un groupe de mots dominé par un nom, en incluant ses expansions. ALSI peut détecter les expansions suivantes : l'adjectif, le groupe participial (voir Figure 2), la subordonnée relative, le groupe prépositionnel et le groupe infinitif agissant comme sujet du verbe (p. ex. *Bien dormir est important*).

Mesures de cohésion

Une cohésion accrue entre les phrases signifie que les entités mentionnées dans une phrase ont une probabilité plus élevée d'être à nouveau abordées dans la phrase suivante, ce qui peut faciliter la lecture (Graesser et al., 2004; Kintsch & Van Dijk, 1978). ALSI produit deux mesures de cohésion lexicale : l'une compare tous les lemmes uniques des phrases adjacentes, l'autre compare uniquement les noms communs et les noms propres. La cohésion lexicale est alors estimée en calculant la similarité cosinus entre phrases adjacentes alors représentées comme des vecteurs de mots (pour une explication du calcul, voir Han et al. 2012). Cette technique est employée notamment par l'outil Coh-Metrix (Grasser et al., 2004).

Dans le but d'estimer la cohésion syntaxique, ALSI crée pour chaque phrase du texte un vecteur contenant trois attributs syntaxiques préalablement convertis en scores standardisés afin d'être sur la même échelle : la longueur de la phrase, la hauteur de l'arbre syntaxique et le nombre de groupes nominaux complexes. Ces attributs ont été choisis puisqu'ils étaient, dans nos essais préliminaires, les trois attributs syntaxiques les plus corrélés avec le niveau scolaire. La cohésion syntaxique est ensuite estimée en calculant la distance euclidienne entre les vecteurs des phrases adjacentes. La distance obtenue est convertie en mesure de cohésion (similitude) en faisant $1/(d + 1)$, où d est la distance.

Méthodologie

Survol de la méthodologie

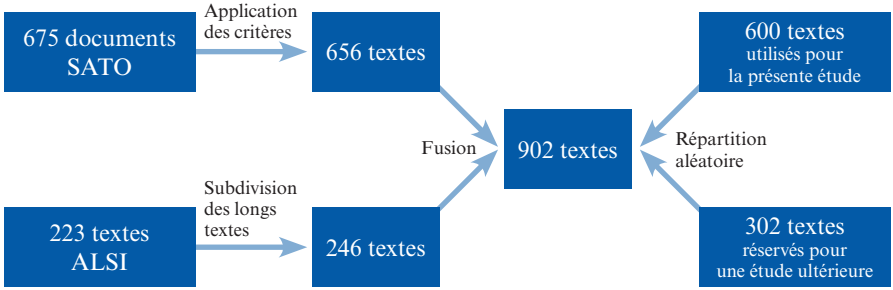
L'objectif des analyses était de tester la capacité de l'outil ALSI à extraire des attributs qui caractérisent la complexité linguistique de textes de langue française. Nous décrivons d'abord la composition du corpus de 600 textes que nous avons analysés à l'aide d'ALSI, puis la procédure

appliquée pour sélectionner des attributs d'intérêt. Nous rapportons des mesures d'association statistique entre les attributs (considérés individuellement) et le niveau de difficulté du texte.

Corpus utilisé

Le corpus utilisé contenait 600 textes répartis entre 11 niveaux scolaires allant de la 1re année du primaire à la 5e secondaire, selon les niveaux du système scolaire québécois. Les années scolaires fournies par le matériel ont été considérées comme des niveaux de difficulté valides pour cette étude; les textes n'ont pas été reclassés. Les critères d'inclusion dans le corpus étaient les suivants : le texte devait avoir une longueur minimale de 30 mots (pour le primaire) ou de 100 mots (pour le secondaire), ne pas être principalement composé de dialogues ou de vers, et ne pas utiliser principalement le registre familier. Ce corpus a été constitué en combinant deux banques de textes selon une procédure illustrée à la Figure 3.

Figure 3
Combinaison puis répartition des textes provenant des banques SATO et ALSI



La première banque de textes provient du développement et de l'étalonnage de l'analyseur SATO-Calibrage (Daoust et al., 1996) et contenait principalement des extraits de manuels scolaires et des examens de lecture destinés aux 11 niveaux scolaires du Québec. Après la séparation des documents contenant plus d'un texte, l'application de critères d'exclusion et l'élimination des doublons, la banque SATO contenait 656 textes. La deuxième banque de textes a été constituée dans le cadre de la présente étude et contenait principalement des extraits de manuels scolaires publiés au Québec après l'an 2000. Le niveau allait de la 6e année du primaire

à la 5e secondaire. Afin d'augmenter la taille du corpus tout en uniformisant la longueur des textes, nous avons scindé en deux les textes de la banque ALSI dont le nombre de mots était plus de deux fois supérieur à la moyenne. Après ces divisions, la banque ALSI contenait 246 textes. Les informations paratextuelles suivantes ont été retirées des deux banques : numéros de page, de paragraphe ou de ligne et autres marques ajoutées par l'éditeur, remarques et définitions ajoutées en marge, les titres et les intertitres sauf lorsque ceux-ci formaient une phrase incluant au moins un verbe conjugué. Puisque ces informations sont généralement ajoutées par l'éditeur et ne sont pas présentes pour tous les textes, elles auraient pu influencer le traitement et fausser les résultats.

Le corpus formé en combinant les banques SATO et ALSI comptait 902 textes (43820 phrases). Nous avons réservé environ le tiers de ce corpus (sélectionné aléatoirement) pour une étude ultérieure en classification du texte, portant la taille du corpus utilisé par la présente étude à 600 textes (29709 phrases). La provenance des textes et leur distribution entre les niveaux scolaires sont indiquées au Tableau 2.

Tableau 2
Provenance du corpus utilisé et distribution entre les 11 niveaux scolaires

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**TOTAL**
SATO	33	49	40	40	39	51	41	36	31	34	42	436
ALSI	0	0	0	0	0	22	22	29	25	22	44	164
TOTAL	33	49	40	40	39	73	63	65	56	56	86	600

Procédure d'extraction et de sélection d'attributs

Nous avons analysé les 600 textes avec ALSI, produisant une matrice dont chaque ligne correspond à un texte, chaque colonne est un attribut et chaque cellule est la valeur numérique de l'attribut pour le texte (voir la Figure 1 pour un exemple simplifié). Compte tenu du grand nombre d'attributs et du fait que nombre d'entre eux sont très similaires, nous avons appliqué une procédure de sélection afin d'éliminer les attributs peu pertinents pour cette étude ou ceux qui contribueraient peu d'information à l'égard de la complexité du texte. Cette procédure en trois étapes se résume comme suit :

1) Nous excluons d'emblée les attributs reflétant la longueur du texte, comme le nombre de mots, de phrases ou de paragraphes. Ces variables auraient pu introduire un biais lié à la manière dont le corpus a été formé, plusieurs textes ayant été subdivisés.

2) Suivant la chaine de traitement proposée par Taneja et al. (2014), nous calculons le gain d'information de chaque attribut, puis retirons les attributs dont le gain d'information était de zéro. Le gain d'information (GI) est une statistique indiquant, dans notre cas, dans quelle mesure l'introduction d'une variable améliore la classification des textes comparativement au niveau de la chance. Il s'agit, en termes plus techniques, de la diminution de l'entropie de Shannon conditionnelle à l'introduction de la variable (Karegowda et al., 2010; Yang & Pedersen, 2022). Retirer les attributs ayant un GI nul élimine les attributs peu susceptibles d'ajouter de l'information à l'égard du niveau de difficulté (niveau scolaire associée au texte). Cela écarte du même coup les attributs dont la variance est nulle ou très faible.
3) Nous identifions ensuite, à l'aide de la fonction *findLinearCombos* de la librairie *caret* pour R (Kuhn, 2011), les groupes d'attributs manifestant des dépendances linéaires. Ces conflits sont gérés en retirant les attributs du groupe un à un, tout en tentant de préserver les attributs ayant le GI le plus élevé. D'autres conflits sont finalement identifiés entre des combinaisons d'attributs produites à partir des mêmes mesures linguistiques ou ne différant que par l'échelle, l'attribut du groupe ayant le GI le plus élevé est conservé.

Les variables ayant passé chaque étape de sélection ont formé la sélection finale d'attributs. Nous avons en outre formé un sous-ensemble réduit de six attributs en sélectionnant le meilleur représentant (GI le plus élevé) des six catégories spécifiées dans la typologie d'ALSI.

Analyses statistiques

L'objectif des analyses était de décrire l'association statistique entre les attributs sélectionnés et le niveau de difficulté du texte, exprimé en niveaux scolaires (1ère année primaire à 5e secondaire) et considéré comme une variable ordinale. Les mesures d'association statistiques étaient le GI et le coefficient rho de Spearman avec intervalles de confiance à 95%. Les intervalles ont été calculés à l'aide de la méthode de Fieller, moins biaisée lorsque les données ont une distribution non normale (Bishara & Hittner, 2017). Afin de pouvoir examiner la progression des valeurs obtenues, nous avons de plus calculé la valeur médiane des attributs par niveau scolaire.

Résultats

La procédure de sélection a été appliquée à un groupe initial de 42 attributs produits par l'outil ALSI et considérés comme pertinents pour cette étude. Une liste complète des attributs considérés se trouve au Tableau 4 (matériel supplémentaire de l'article) et précise la raison du rejet, le cas échéant. Sur les 42 attributs considérés, 6 ont été retirés en raison d'un GI nul, aucun n'a été retiré en raison de dépendances linéaires, 18 attributs ont été retirés pour éviter des conflits entre attributs similaires (sur une échelle différente ou dérivés des mêmes mesures). La sélection finale comptait 20 attributs (8 lexicaux, 12 syntaxiques).

Le Tableau 3 indique l'association statistique entre le niveau scolaire et les attributs de la sélection finale en présentant le GI, le coefficient de Spearman et le type d'attribut selon la typologie décrite dans le présent article. Pour les 20 attributs sélectionnés, les coefficients de Spearman étaient significatifs à un seuil de $p < 0{,}001$ et les intervalles de confiance des coefficients de corrélation ne contenaient pas la valeur 0. La magnitude des corrélations était de faible à forte selon les barèmes d'interprétation suggérés par Akoglu (2018) pour la recherche en psychologie. Dans l'ensemble, la direction des corrélations était cohérente avec la nature des attributs mesurés, c'est-à-dire une corrélation positive lorsque la valeur numérique de l'attribut est censée augmenter avec la difficulté du texte, et vice versa.

Les attributs de la sélection réduite (GI le plus élevé de leur type) sont indiqués en caractères gras dans le Tableau 3. Il s'agit de : l'âge moyen de première apparition dans le lexique Manulex (*ageManulex_m*), la longueur orthographique moyenne (*longMotOrtho_m*), la longueur des phrases exprimée en nombre de mots (*longPh_m*), la cohésion syntaxique de phrase à phrase (*cohesionSyn_m*), la hauteur moyenne de l'arbre syntaxique de la phrase (*hauteurPh_m*), et l'indice de diversité lexicale de Maas calculé sur les lemmes (*maas_lemma_i*). La Figure 4 montre les distributions de ces six attributs par année scolaire, permettant de visualiser leur progression ainsi que la présence de valeurs aberrantes (*outliers*). Ainsi, cinq des attributs montrés à la Figure 4 avaient une progression généralement croissante; par exemple, la longueur moyenne des phrases passait d'environ 10 mots en 1re année à environ 20 mots en 7e année (première secondaire) à un peu moins de 30 mots en 11e année (5e secondaire). Dans le cas de la cohésion syntaxique, la progression était décroissante,

suggérant que la cohésion diminue lorsque les textes deviennent plus complexes. Le Tableau 5 (matériel supplémentaire de l'article) dresse une liste des valeurs médianes par attribut et par niveau scolaire.

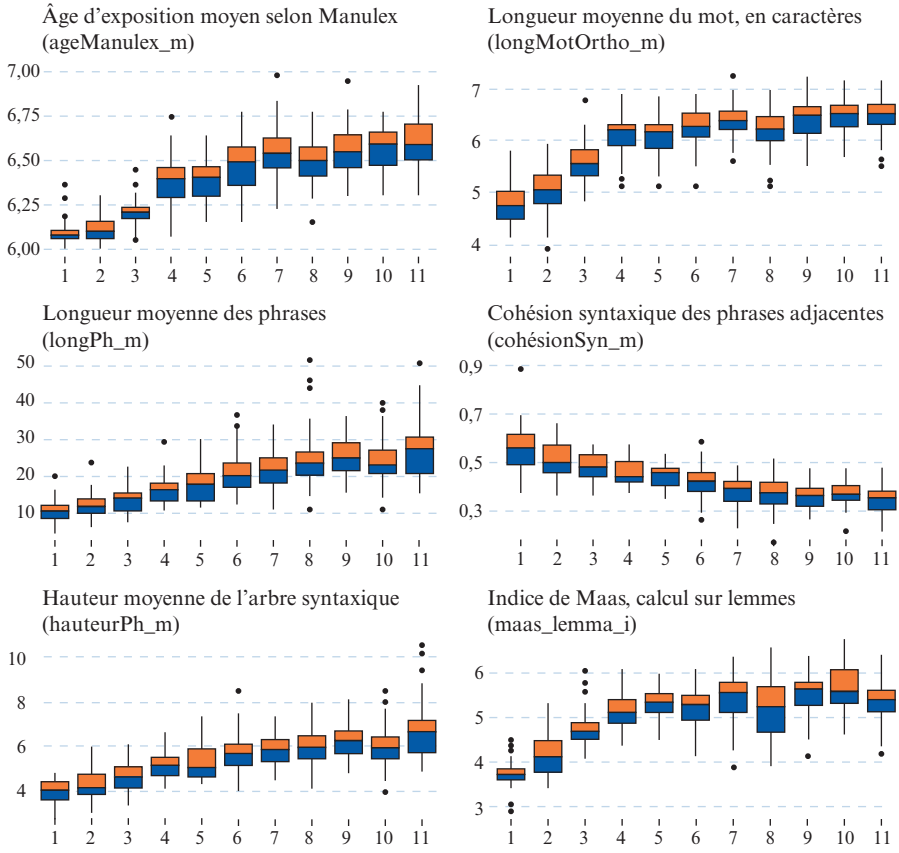
Tableau 3
Mesures de l'association statistique entre l'attribut et le niveau scolaire

Attributs	GI	r_s [IC 95%]	Type
ageManulex_m	**0,502**	**0,71 [0,67, 0,75]**	**Lex. 2**
freqManulexSfi_m	0,488	-0,70 [-0,74, -0,66]	Lex. 2
longMotOrtho_m	**0,441**	**0,63 [0,58, 0,68]**	**Lex. 1**
freqEqolSfi_m	0,432	-0,66 [-0,7, -0,61]	Lex. 2
longPh_m	**0,407**	**0,70 [0,66, 0,74]**	**Syn. 1**
cohesionSyn_m	**0,398**	**-0,70 [-0,74, -0,66]**	**Syn. 3**
hauteurPh_m	**0,386**	**0,67 [0,62, 0,71]**	**Syn. 2**
ageEqol_m	0,374	0,65 [0,60, 0,70]	Lex. 2
motSeuilOrtho_p	0,340	0,61 [0,56, 0,66]	Syn. 1
ageMels_m	0,332	0,59 [0,53, 0,64]	Lex. 2
maas_lemma_i	**0,322**	**0,54 [0,48, 0,60]**	**Lex. 3**
verbesConju_m	0,293	0,65 [0,60, 0,70]	Syn. 2
GNC_m	0,287	0,60 [0,54, 0,65]	Syn. 2
virgule_m	0,228	0,64 [0,59, 0,69]	Syn. 1
partPass_m	0,219	0,55 [0,49, 0,60]	Syn. 2
partPres_m	0,172	0,49 [0,42, 0,55]	Syn. 2
GV_m	0,159	0,56 [0,50, 0,61]	Syn. 2
phMarqueur_m	0,108	0,44 [0,37, 0,50]	Syn. 2
adp_p	0,106	0,30 [0,22, 0,37]	Syn. 2
simCosinNom_m	0,079	-0,31 [-0,38, -0,23]	Lex. 3

Note. Statistiques calculées à partir d'un corpus de 600 textes, pour les 20 attributs sélectionnés et les 11 niveaux scolaires. GI indique le gain d'information. Les caractères gras indiquent l'attribut ayant le GI le plus élevé par type. R_s indique le coefficient de corrélation de Spearman entre chaque attribut et le niveau scolaire du texte, avec intervalle de confiance à 95%. Toutes les corrélations de Spearman de ce tableau étaient statistiquement significatives au seuil $p < 0{,}001$. Les types d'attributs lexicaux et syntaxiques sont résumés dans le présent article.

Figure 4

Diagrammes en boîte des six attributs de la sélection réduite, par niveau scolaire



Note. Résultats portant sur 600 textes. L'axe des abscisses indique l'année scolaire au primaire (1 à 6) et au secondaire (7 à 11) du Québec. L'axe des ordonnées montre l'unité de mesure propre à l'attribut. La boîte indique les percentiles 25 à 75.

Discussion

Dans la présente étude, des analyses de corrélation ont été utilisées pour tester la capacité des attributs extraits par ALSI à estimer la difficulté de textes en français. Notre présentation des résultats s'est concentrée sur une sélection de 20 attributs (8 lexicaux, 12 syntaxiques) qui semblaient particulièrement intéressants pour estimer la difficulté des textes utilisés dans un contexte scolaire. Nous retenons trois résultats importants concernant la nature des attributs retenus par la procédure de sélection.

Premièrement, plusieurs attributs manifestaient un effet de plateau. Par exemple, l'indice de diversité lexicale de Maas calculé sur les lemmes (*maas_lemma_i*) augmente jusqu'à la fin du primaire, puis se stabilise. Ces effets de plateau ont aussi été décrits par Daoust et al. (1996) et suggèrent que certains attributs linguistiques atteignent leur complexité limite durant le parcours scolaire. Une autre explication possible est qu'ALSI n'est peut-être pas apte à mesurer la progression de certains attributs au-delà d'un certain point. Ainsi, certains des plateaux pourraient s'expliquer par le fait que les lexiques de référence ne couvrent pas le niveau secondaire (années 7 à 11). De futurs travaux pourraient tester l'inclusion dans ALSI de lexiques couvrant aussi le niveau secondaire dans le but de mieux estimer la complexité lexicale au-delà de la 6e année.

Deuxièmement, nos résultats montrent que les attributs dits « de surface » peuvent effectivement contribuer à estimer la difficulté du texte. La longueur moyenne des mots (r_s = 0,63) et le 90e percentile de la longueur des phrases (r_s = 0,69) comptaient parmi les attributs ayant la corrélation la plus forte avec le niveau de difficulté du texte. Ces résultats remettent en question les conclusions d'autres auteurs voulant que ce type d'attribut soit sans valeur. Ils concordent toutefois avec l'étude similaire de François et Fairon (2012), selon laquelle la longueur des mots et la longueur des phrases faisaient partie des attributs davantage corrélés avec le niveau de difficulté du texte (r_s = 0,48 et r_s = 0,61, respectivement). Une explication plausible est que les attributs de surface, malgré leur simplicité apparente, restent des intermédiaires efficaces pour évaluer la difficulté du texte. Cette explication va dans le sens des conclusions de Szmrecsányi (2004) à l'égard de la longueur de la phrase comme estimateur de la complexité syntaxique.

Troisièmement, nos résultats suggèrent que la cohésion linguistique peut contribuer à modéliser la complexité du texte. L'attribut de cohésion syntaxique (*cohesionSyn_m*) affichait une corrélation de r_s = -0,66, corrélation d'ampleur modérée selon les barèmes suggérés par Akoglu (2018). Ce résultat est important puisqu'il ajoute un soutien empirique à l'hypothèse selon laquelle la cohésion affecte la compréhension (O'Reilly & McNamara, 2007). La cohésion lexicale (*simCosinNom_m*) a toutefois présenté une corrélation plus modeste (r_s = -0,31), rejoignant les résultats obtenus par Todirascu et al. (2016) sur un corpus de langue française.

Nous avons identifié plusieurs limites à la présente étude dont la portée est basée sur la prémisse que les textes utilisés sont représentatifs de ce que l'on trouve dans le curriculum québécois, et possiblement dans d'autres curriculums francophones. Nous avons également supposé que le niveau scolaire indiqué par le matériel peut être considéré comme une référence fiable. De manière plus spécifique, nos résultats sont limités par le fait que l'ensemble de textes plus récents (la banque ALSI) ne couvre pas les 11 années du parcours scolaire. En effet, les textes des années 1 à 5 sont globalement plus anciens, provenant de la banque employée par Daoust et al. (1996). Une piste à explorer serait donc d'ajouter des textes plus récents, couvrant la période allant de la 1re à la 5e année du primaire. Les résultats dépendent également des attributs linguistiques que la version actuelle d'ALSI peut extraire. Des travaux ultérieurs pourraient intégrer des types d'attributs portant sur d'autres aspects de la langue, notamment sur la complexité morphologique. Enfin, comme notre étude s'est limitée à des analyses considérant les attributs un à un, il faudrait procéder à des analyses multivariées afin de modéliser la difficulté du texte et d'évaluer la contribution des attributs. La validité externe de l'instrument, sa capacité à estimer le niveau scolaire de nouveaux textes, pourrait être testée en appliquant un modèle multivarié à un nouveau corpus.

Conclusion

Dans cette étude, nous avons décrit ALSI, un nouvel outil d'analyse linguistique qui génère une variété d'attributs dans le but d'évaluer la complexité d'un texte. Après avoir justifié le développement d'un nouvel outil, l'article a décrit les bases théoriques d'ALSI et a présenté les procédures d'extraction des attributs. Le second volet de l'article avait pour objectif de déterminer les attributs qui étaient les plus prometteurs pour évaluer le niveau scolaire des textes du corpus en français québécois. Nous avons pour cela appliqué ALSI à un corpus de 600 textes répartis entre les 11 années scolaires considérées comme indicateurs de la difficulté du texte. Des analyses corrélationnelles ont montré le potentiel des attributs pour évaluer la difficulté du texte, ce qui appuie la validité de l'outil ALSI. Les résultats montrent de plus que les attributs de surface sont toujours d'actualité et mettent en évidence le potentiel des attributs mesurant la cohésion linguistique, particulièrement la cohésion syntaxique. La présente étude a, en somme, proposé des attributs qui peuvent être extraits

avec l'outil ALSI et qui sont associés à la complexité linguistique de textes employés en milieu scolaire au Québec. Il s'agit d'une première étape dans la validation de l'outil, d'autres travaux étant requis afin d'en tester la validité externe.

En plus de l'évaluation de la difficulté du texte, nous voyons plusieurs applications d'ALSI dans le domaine de l'éducation. L'outil pourrait contribuer à une démarche de validation d'épreuves et de tests en évaluant *a priori* la difficulté linguistique des items. ALSI pourrait de plus aider à sélectionner ou à créer du matériel didactique ayant un niveau linguistique approprié, ou qui favorise l'apprentissage de certains objets de savoir en français. En contexte d'évaluation linguistique, ALSI pourrait être appliqué à des productions écrites d'apprenants du français langue seconde pour évaluer le développement du vocabulaire et de la syntaxe. Enfin, une prochaine version de l'outil est prévue et prendra la forme d'une application Web afin de simplifier son utilisation[4].

Réception : 12 octobre 2021
Version finale : 16 mars 2022
Acceptation : 18 mai 2022

4. Un prototype est disponible au https://gloignon.shinyapps.io/ALAIN_v3/

LISTE DE RÉFÉRENCES

Akoglu, H. (2018). User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine*, *18*(3), 91-93. https://doi.org/10/ggw2tg

Avenia-Tapper, B., & Llosa, L. (2015). Construct Relevant or Irrelevant? The Role of Linguistic Complexity in the Assessment of English Language Learners' Science Knowledge. *Educational Assessment*, *20*(2), 95-111. https://doi.org/10.1080/10627197.2015.1028622

Benjamin, R. G. (2012). Reconstructing readability : Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review*, *24*(1), 63-88. https://doi.org/10/bdjfkd

Bishara, A. J., & Hittner, J. B. (2017). Confidence intervals for correlations when data are not normal. *Behavior research methods*, *49*(1), 294-309.

Blache, P. (2010, juillet). *Un modèle de caractérisation de la complexité syntaxique* [présentation de conférence]. TALN 2010, Montréal, Canada. https://hal.archives-ouvertes.fr/hal-00576890

Boyer, J.-Y. (1992). La lisibilité. *Revue française de pédagogie*, *99*, 5-14. https://doi.org/10/ddnvf8

Clevinger, A. (2014). *Test performance : the influence of cognitive load on reading comprehension* [Thèse doctorale, Georgia State University]. https://scholarworks.gsu.edu/psych_theses/123/

Crossley, S. A. (2020). Linguistic features in writing quality and development : An overview. *Journal of Writing Research, 11*(3), 415-443. https://doi.org/10.17239/jowr-2020.11.03.01

Daoust, F., Laroche, L., & Ouellet, L. (1996). SATO-CALIBRAGE : Présentation d'un outil d'assistance au choix et à la rédaction de textes pour l'enseignement. *Revue québécoise de linguistique*, *25*(1), 205-234. https://doi.org/10/ghhd3p

Dascalu, M., Dessus, P., Trausan-Matu, Ş., Bianco, M., & Nardy, A. (2013). ReaderBench, an environment for analyzing text complexity and reading strategies. Dans H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (dir.), *Artificial Intelligence in Education* (p. 379-388). Springer. https://doi.org/10/ghjqdq

De Marneffe, M. C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., & Manning, C. D. (2014). Universal Stanford dependencies : A cross-linguistic typology. Dans *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* (p. 4585-4592). European Language Resources Association (ELRA).

Dempster, E. R., & Reddy, V. (2007). Item readability and science achievement in TIMSS 2003 in South Africa. *Science Education*, *91*(6), 906-925. https://doi.org/10/cd687q

Feng, L., Jansche, M., Huenerfauth, M., & Elhadad, N. (2010). A comparison of features for automatic readability assessment. *Dans COLING '10 : Proceedings of the 23rd International Conference on Computational Linguistics* (p. 276-284). http://www.aclweb.org/anthology/C10-2032

Fergadiotis, G., Wright, H. H., & Green, S. B. (2015). Psychometric Evaluation of Lexical Diversity Indices : Assessing Length Effects. *Journal of Speech, Language, and Hearing Research : JSLHR*, *58*(3), 840-852. https://doi.org/10/gh62rx

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*(3), 221. https://doi.org/10/bzrfs6

François, T. (2009). Combining a statistical language model with logistic regression to predict the lexical and syntactic difficulty of texts for FFL. Dans *Proceedings of the Student Research Workshop at EACL 2009* (p. 19-27). Association for Computational Linguistics.

François, T. (2015). When readability meets computational linguistics : A new paradigm in readability. *Revue française de linguistique appliquée, 20*(2), 79-97. https://doi.org/10/gh5tmg

François, T., & Fairon, C. (2012). An "AI readability" formula for French as a foreign language. Dans Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (p. 466-477). Association for Computational Linguistics.

François, T., & Miltsakaki, E. (2012). Do NLP and machine learning improve traditional readability formulas ? Dans *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations* (p. 49-57). Association for Computational Linguistics.

Gale, W. A., & Sampson, G. (1995). Good-Turing frequency estimation without tears. *Journal of Quantitative Linguistics*, *2*(3), 217-237. https://doi.org/10/bnnzxz

Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remedial and Special Education*, *7*(1), 6-10. https://doi.org/10.1177/074193258600700104

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix : Providing Multilevel Analyses of Text Characteristics. *Educational Researcher*, *40*(5), 223-234. https://doi.org/10/cwtd84

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix : Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers*, *36*(2), 193-202. https://doi.org/10/ft568w

Guillaume, B., De Marneffe, M.-C., & Perrier, G. (2019). Conversion et améliorations de corpus du français annotés en Universal Dependencies. *Traitement automatique des langues*, *60*(2), 71-95. https://hal.inria.fr/hal-02267418

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining : Concepts and Techniques (3e éd.)*. Morgan Kaufmann. https://doi.org/10.1016/B978-0-12-381479-1.00002-2

Joint Committee on Standards for Educational and Psychological Testing. (2014). *Standards for Educational and Psychological Testing*. American Educational Research Association.

Karegowda, A. G., Manjunath, A. S., & Jayaram, M. A. (2010). Comparative study of attribute selection using gain ratio and correlation based feature selection. *International Journal of Information Technology and Knowledge Management*, *2*(2), 271-277.

Kintsch, W., & Van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, *85*(5). https://doi.org/10.1037/0033-295X.85.5.363

Kintsch, W., & Vipond, D. (2014). Reading comprehension and readability in educational practice and psychological theory. Dans L.-G. Nilsson, T. Archer (dir.), *Perspectives on learning and memory* (p. 329-365). Psychology Press.

Kuhn, M. (2011). *Data Sets and Miscellaneous Functions in the caret Package*. http://ftp.uni-bayreuth.de/math/statlib/R/CRAN/doc/vignettes/caret/caretMisc.pdf

Lane, S., Raymond, M. R., & Haladyna, T. M. (dir.). (2015). *Handbook of Test Development* (2e éd.). Routledge.

Lété, B. (2004). MANULEX : une base de données du lexique écrit adressé aux élèves. Dans É. Callaque, J. David (dir.) *Didactique du lexique* (p. 241-257). De Boeck.

Loye, N. (2018). Et si la validation n'était pas juste une suite de procédures techniques… *Mesure et évaluation en Éducation, 41*(1), 97-123. https://doi.org/10.7202/1055898ar

Maas, H. D. (1972). Über den zusammenhang zwischen wortschatzumfang und länge eines textes. *Zeitschrift für Literaturwissenschaft und Linguistik*, *2*(8), 73.

Martiniello, M. (2009). Linguistic Complexity, Schematic Representations, and Differential Item Functioning for English Language Learners in Math Tests. *Educational Assessment*, *14*(3-4), 160-179. https://doi.org/10/fcj83v

McNamara, D., & Graesser, A. (2011). Coh-Metrix : An Automated Tool for Theoretical and Applied Natural Language Processing. Dans P. M. McCarthy (dir.), *Applied natural language processing and content analysis : Identification, investigation, and resolution*, (p. 188-205). IGI Global. https://doi.org/10/ghp3zg

McNamara, D. S., Graesser, A. C., & Louwerse, M. M. (2012). Sources of text difficulty : Across genres and grades. Dans J. Sabatini (dir.), *Measuring up : Advances in how we assess reading ability* (p. 89-116). R&L Education.

Mesnager, J. (1989). Lisibilité des textes pour enfants : Un nouvel outil ? *Communication & Langages, 79*(1), 18-38. https://doi.org/10/bb9gfg

Milone, M. (2014). *Development of the ATOS readability formula*. Renaissance Learning Inc.

O'Reilly, T., & McNamara, D. S. (2007). Reversing the reverse cohesion effect : good texts can be better for strategic, high-knowledge readers. *Discourse Processes*, *43*(2), 121-152. https://doi.org/10.1080/01638530709336895

Persson, T. (2016). The language of science and readability : correlations between linguistic features in TIMSS science items and the performance of different groups of Swedish 8th grade students. *Nordic Journal of Literacy Research*, *2*(1). https://doi.org/10.17585/njlr.v2.186

Ravid, D. (2005). Emergence of linguistic complexity in later language development : evidence from expository text construction. Dans D. D. Ravid et H. B.-Z. Shyldkrot (dir.), *Perspectives on Language and Language Development : Essays in Honor of Ruth A. Berman* (p. 337-355). Springer US. https://doi.org/10.1007/1-4020-7911-7_25

R Core Team (2022). *R : A language and environment for statistical computing.* R Foundation for Statistical Computing. https://www.R-project.org/.

Sherstinova, T., Ushakova, E., & Melnik, A. (2020). Measures of Syntactic Complexity and their Change over Time (the Case of Russian). *27th Conference of Open Innovations Association (FRUCT)* (p. 221-229). https://doi.org/10.23919/FRUCT49677.2020.9211027

Smith, D. R., Stenner, A. J., Horabin, I., & Smith, M. (1989). *The Lexile scale in theory and practice. Final report*. MetaMetrics.

Stanké, B., Le Mené, M., Rezzonico, S., Moreau, A., Dumais, C., Robidoux, J., Dault, C., & Royle, P. (2019). ÉQOL : Une nouvelle base de données québécoise du lexique scolaire du primaire comportant une échelle d'acquisition de l'orthographe lexicale. *Corpus*, 19. https://doi.org/10.4000/corpus.3818

Straka, M., Hajic, J., & Straková, J. (2016). UDPipe : trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, pos tagging and parsing. Dans *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (p. 4290-4297). European Language Resources Association.

Szmrecsányi, B. (2004). On operationalizing syntactic complexity. Dans G. Purnelle, C. Fairon & A. Dister (dir.). *Le poids des mots. Proceedings of the 7th International Conference on Textual Data Statistical Analysis.* (Vol. 2, p. 1032-1039). Leuven University Press.

Taneja, S., Gupta, C., Goyal, K., & Gureja, D. (2014). An enhanced k-nearest neighbor algorithm using information gain and clustering. Dans *2014 Fourth International Conference on Advanced Computing Communication Technologies* (p. 325-329). https :// doi.org/10/ghndnz

Todirascu, A., François, T., Bernhard, D., Gala, N., & Ligozat, A. L. (2016). Are cohesive features relevant for text readability evaluation ? Dans *26th International Conference on Computational Linguistics (COLING 2016)* (p. 987-997). https ://aclanthology. org/C16-1

Vandeweerd, N. (2021). fsca : French syntactic complexity analyzer. *International Journal of Learner Corpus Research, 7*(2), 259-274. https ://doi.org/10.1075/ijlcr.20018.van

Visone, J. D. (2009). The Validity of Standardized Testing in Science. *American Secondary Education*, *38*(1), 46-61. https ://www.jstor.org/stable/41406066

Welbers, K., van Atteveldt, W., & Kleinnijenhuis, J. (2020). Extracting semantic relations using syntax : an R package for querying and reshaping dependency trees. *Computational Communication Research*, *3*(2), 1-16.

Wijffels, J. (2022). udpipe : Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the UDPipe NLP Toolkit. R package version 0.8.9. https :// CRAN.R-project.org/package=udpipe

Yang, Y., & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. Dans *Proceedings of the 14th International Conference on Machine Learning* (p. 412-420). Morgan Kaufmann Publishers.

Zakaluk, B. L., & Samuels, S. J. (1988). *Readability : Its Past, Present, and Future*. International Reading Association. https ://eric.ed.gov/ ?id=ED292058

Tableau 4
Liste complète des attributs avec mesures d'association statistique, issue de la sélection et description

#	Attribut	*GI*[a]	r^s [IC 95 %][b]	Statut[c]	Type[d]	Description
1	ageManulex_m	0,502	0,71 [0,67, 0,75] ***	SÉLEC	Lex. 2	Âge d'exposition moyen selon Manulex
2	freqManulexSfi_m	0,488	-0,7 [-0,74, -0,66] ***	SÉLEC	Lex. 2	Fréquence standardisée moyenne selon Manulex
3	longMotOrtho_m	0,441	0,63 [0,58, 0,68] ***	SÉLEC	Lex. 1	Longueur orthographique (nombre de caractères) moyenne
4	longMotSyll_m	0,437	0,63 [0,58, 0,68] ***	DOUB(3)	Lex. 1	Longueur syllabique moyenne
5	freqEqolSfi_m	0,432	-0,66 [-0,7, -0,61] ***	SÉLEC	Lex. 2	Fréquence standardisée moyenne selon ÉQOL
6	longPh_m	0,407	0,7 [0,66, 0,74] ***	SÉLEC	Syn. 1	Longueur moyenne des phrases
7	cohesionSyn_m	0,398	-0,7 [-0,74, -0,66] ***	SÉLEC	Syn. 3	Cohésion syntaxique moyenne entre phrases adjacentes
8	hauteurPh_m	0,386	0,67 [0,62, 0,71] ***	SÉLEC	Syn. 2	Hauteur moyenne de l'arbre syntaxique des phrases
9	ageEqol_m	0,374	0,65 [0,6, 0,7] ***	SÉLEC	Lex. 2	Âge d'exposition moyen selon ÉQOL
10	longPh_90	0,362	0,69 [0,64, 0,73] ***	DOUB(6)	Syn. 1	90e percentile de la longueur des phrases
11	motSeuilOrtho_p	0,34	0,61 [0,56, 0,66] ***	SÉLEC	Lex. 1	Prop. de mots comptant plus de huit caractères
12	dansManulex_p	0,339	-0,68 [-0,72, -0,63] ***	DOUB(1)	Lex. 2	Prop. de mots présents dans Manulex
13	motSeuilSyll_p	0,339	0,62 [0,57, 0,67] ***	DOUB(11)	Lex. 1	Prop. de mots comptant plus de trois syllabes

#	Attribut	*GI*[a]	*r*[c] [IC 95 %][b]	Statut[c]	Type[d]	Description
14	ageMels_m	0,332	0,59 [0,53, 0,64] ***	SÉLEC	Lex. 2	Âge d'exposition moyen selon la liste orthographique du ministère de l'Éducation du Québec
15	dansEqol_p	0,332	-0,67 [-0,71, -0,62] ***	DOUB(5)	Lex. 2	Prop. de mots présents dans la liste ÉQOL
16	maas_lemma_i	0,322	0,54 [0,48, 0,6] ***	SÉLEC	Lex. 3	Diversité lexicale selon l'indice de Maas calculé sur les lemmes
17	freqManulex_m	0,307	-0,56 [-0,61, -0,5] ***	DOUB(2)	Lex. 2	Fréquence moyenne selon Manulex
18	freqEqol_m	0,303	-0,56 [-0,61, -0,5] ***	DOUB(5)	Lex. 2	Fréquence moyenne selon ÉQOL
19	maas_token_i	0,301	0,52 [0,46, 0,58] ***	DOUB(16)	Lex. 3	Diversité lexicale selon l'indice de Maas calculé sur les mots (*tokens*)
20	verbesConju_m	0,293	0,65 [0,6, 0,7] ***	SÉLEC	Syn. 2	Nombre moyen de verbes conjugués
21	GNC_m	0,287	0,6 [0,54, 0,65] ***	SÉLEC	Syn. 2	Nombre moyen de groupes nominaux complexes par phrase
22	longPh30_p	0,282	0,67 [0,62, 0,71] ***	DOUB(6)	Syn. 1	Prop. de phrases comptant plus de 30 mots (voir Daoust et al., 1996)
23	dansMels_p	0,238	-0,59 [-0,64, -0,53] ***	DOUB(14)	Lex. 2	Prop. de mots présents dans la Liste orthographique du ministère de l'éducation du Québec

#	Attribut	*GI*[a]	r^s [IC 95 %][b]	Statut[c]	Type[d]	Description
24	GNCGr_m	0,235	0,55 [0,49, 0,6] ***	DOUB(21)	Syn. 2	Nombre moyen par phrase de groupes nominaux complexes contenant au moins un groupe complément, participial ou prépositionnel
25	virgule_m	0,228	0,64 [0,59, 0,69] ***	SÉLEC	Syn. 1	Nombre moyen de virgules par phrase.
26	partPass_m	0,219	0,55 [0,49, 0,6] ***	SÉLEC	Syn. 2	Nombre moyen de participes passés par phrase
27	partPres_m	0,172	0,49 [0,42, 0,55] ***	SÉLEC	Syn. 2	Nombre moyen de participes présents par phrase
28	GV_m	0,159	0,56 [0,5, 0,61] ***	SÉLEC	Syn. 2	Nombre moyen de groupes verbaux par phrase
29	GVFin_m	0,152	0,54 [0,48, 0,6] ***	DOUB(28)	Syn. 2	Nombre moyen de groupes verbaux finis (excluant les verbes infinitifs) par phrase
30	TTR_token_i	0,152	-0,32 [-0,39, -0,24] ***	DOUB(19)	Lex. 3	Diversité lexicale, ratio type-jeton calculé sur les *tokens*
31	verbeComplexe_m	0,149	0,46 [0,39, 0,52] ***	DOUB(20)	Syn. 2	Prop. des verbes conjugués considérés comme complexes (voir Daoust et al., 1996)
32	phMarqueur_m	0,108	0,44 [0,37, 0,5] ***	SÉLEC	Syn. 2	Nombre moyen de connecteurs argumentatifs et organisateurs textuels par phrase
33	TTR_lemma_i	0,107	-0,26 [-0,34, -0,18] ***	DOUB(19)	Lex. 3	Diversité lexicale, ratio type-jeton calculé sur les lemmes
34	adp_p	0,106	0,3 [0,22, 0,37] ***	SÉLEC	Syn. 2	Prop. de tous les mots du texte étant des prépositions

#	Attribut	*GI*[a]	*r*[c] [IC 95 %][b]	Statut[c]	Type[d]	Description
35	simCosinNom_m	0,079	-0,31 [-0,38, -0,23] ***	SÉLEC	Lex. 3	Cohésion lexicale mesurée par la similarité cosinus des noms communs uniques des phrases adjacentes (voir Graesser et al., 2004)
36	sconj_p	0,058	0,081 [0, 0,16] *	DOUB(32)	Syn. 2	Prop. de tous les mots du texte étant des conjonctions de coordination
37	adj_p	0	0,23 [0,15, 0,31] ***	GI = 0	Syn. 2	Prop. de tous les mots du texte étant des adjectifs
38	adv_p	0	0,091 [0,01, 0,17] *	GI = 0	Syn. 2	Prop. de tous les mots du texte étant des adverbes
39	longGNC_m	0	0,16 [0,08, 0,24] ***	GI = 0	Syn. 2	Longueur moyenne des groupes nominaux complexes
40	noun_p	0	0,13 [0,05, 0,21] **	GI = 0	Syn. 2	Prop. de tous les mots du texte étant des noms communs
41	propn_p	0	0,13 [0,05, 0,21] **	GI = 0	Syn. 2	Prop. de tous les mots du texte étant des noms propres
42	simCosinLemma_m	0	-0,13 [-0,21, -0,05] **	GI = 0	Lex. 3	Cohésion lexicale mesurée par la similarité cosinus des lemmes uniques des phrases adjacentes (voir Graesser et al., 2004)

Note. Statistiques calculées sur un corpus de 600 textes répartis entre les 11 années du système scolaire québécois. [a]Gain d'information (GI) entre l'attribut et le niveau scolaire associé au texte. [b]Corrélation de Spearman entre l'attribut et le niveau scolaire associé au texte. *** $p < 0{,}001$, ** $p < 0{,}01$, * $p < 0{,}05$. [c]Issue de la procédure de sélection. SÉLEC : attribut dans la sélection finale ; DOUB : attribut retiré car il s'agit d'un quasi-doublon avec l'attribut sélectionné dont le numéro est indiqué entre parenthèses ; GI = 0 : attribut retiré, car son GI était nul. [d]Type d'attribut selon la classification proposée dans le présent article.

Tableau 5
Médiane des attributs par année scolaire (sélection de 20 attributs)

Attribut	Niveau scolaire										
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
adp_p	0,08	0,1	0,11	0,13	0,12	0,14	0,13	0,12	0,13	0,12	0,14
ageEqol_m	6,71	6,78	6,92	7,09	7,16	7,21	7,27	7,22	7,26	7,29	7,29
ageManulex_m	6,08	6,09	6,2	6,38	6,39	6,47	6,53	6,49	6,53	6,58	6,57
ageMels_m	7,31	7,4	7,58	7,86	7,74	8,01	8,07	7,89	7,98	8,1	8,05
cohesionSyn_m	0,55	0,5	0,48	0,43	0,45	0,42	0,39	0,37	0,36	0,36	0,35
freqEqolSfi_m	63,49	62,61	59,71	57,28	55,58	56,69	54,95	55,23	54,44	53,74	54,31
freqManulexSfi_m	64,21	62,67	60,26	56,63	56,42	55,68	54,26	55,36	54,28	53	53,48
GNC_m	1,43	1,59	1,8	2,05	2	2,73	2,78	2,84	2,79	2,81	2,92
GV_m	1,67	2	2,11	2,35	2,41	2,52	2,56	2,72	2,9	2,74	3,01
hauteurPh_logm	1,34	1,38	1,48	1,59	1,55	1,69	1,71	1,75	1,78	1,74	1,82
longMotOrtho_m	4,71	4,99	5,51	6,16	6,13	6,22	6,35	6,18	6,46	6,48	6,47
longPh_m	10,02	11,36	13,48	15,9	17,43	19,66	21,16	23,06	24,5	22,89	26,81
maas_lemma_i	3,69	4,09	4,66	5,1	5,31	5,28	5,53	5,25	5,63	5,58	5,4
motSeuilOrtho_p	0,05	0,07	0,11	0,17	0,17	0,19	0,2	0,19	0,21	0,21	0,22
partPass_m	0,08	0,19	0,24	0,37	0,38	0,45	0,55	0,69	0,63	0,71	0,66
partPres_m	0	0	0	0,03	0,06	0,08	0,11	0,1	0,14	0,1	0,1
phMarqueur_m	0,07	0,08	0,12	0,16	0,16	0,22	0,22	0,21	0,32	0,28	0,29
simCosinNom_m	0,13	0,14	0,12	0,13	0,08	0,11	0,04	0,06	0,07	0,05	0,05
verbesConju_m	1,18	1,55	1,57	1,78	1,87	2,03	2,21	2,4	2,61	2,37	2,71
virgule_m	0,54	0,73	0,78	0,84	1	1,18	1,47	1,6	1,83	1,7	1,85

Note. Valeurs médianes calculées à partir de 600 textes; les niveaux correspondent aux 11 années, 6 au primaire (1 à 6) et 5 au secondaire (7 à 11), du système scolaire du Québec.