Mesure et évaluation en éducation

ILSA: an automated language complexity analysis tool for French

Guillaume Loignon

Volume 44, Number spécial, 2021

Translation Issue

Reception: 12 October 2021
Final version: 16 March 2022
Acceptance: 18 May 2022

URI: https://id.erudit.org/iderudit/1095682ar
DOI: https://doi.org/10.7202/1095682ar

See table of contents

Publisher(s)

ADMEE-Canada

ISSN

0823-3993 (print)
2368-2000 (digital)

Explore this journal

Cite this article

Loignon, G. (2021). ILSA: an automated language complexity analysis tool for French. *Mesure et évaluation en éducation*, *44*(spécial), 61–88. https://doi.org/10.7202/1095682ar

Article abstract

Estimating language complexity is an important aspect of educational measurement and assessment that can be used, for instance, to control for 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 ILSA 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 ILSA for modeling the complexity of French texts.

ILSA: an automated language complexity analysis tool for French

Guillaume Loignon
University of Quebec in Montreal

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

Estimating language complexity is an important aspect of educational measurement and assessment that can be used, for instance, to control for 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 ILSA 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 ILSA for modeling the complexity of French texts.

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.*

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

ILSA, for Integrated Lexical-Syntactic Analyzer, is a natural language processing tool that extracts a set of features characterising the intrinsic complexity of text. We have created ILSA to meet certain needs in the field of educational measurement and assessment. For example, a linguistic analysis tool can help select appropriate texts according to the students' age and the educational objectives to be attained. A similar analyzer, SATO-Calibrage (Daoust et al., 1996), is currently available online, but dates back to the 1990s and has not managed to take advantage of theoretical and methodological innovations concerning the sources of text difficulty and their automated measurement. ILSA draws on more recent technical and theoretical advances, such as the *Échelle québécoise de l'orthographe lexicale* (Quebec scale of lexical spelling, or ÉQOL) database (Stanké et al., 2019) and Manulex (Lété, 2004), as well as work on the English language tool Coh-Metrix (McNamara & Graesser, 2011). This paper has two objectives: first, to present ILSA, its theoretical context and functions, and second, to conduct an initial validation test by analysing 600 texts used at the primary and secondary levels in Quebec.

Language complexity in educational measurement and assessment

In line with the Cognitive Load Theory (Clevinger, 2014), text complexity can be considered as emerging from intrinsic and extrinsic factors. The intrinsic complexity of the text is that which can be reduced to its measurable characteristics, or features. Sentence length is a classic example of a text feature (Flesch, 1948; Szmrecsányi, 2004). Extrinsic complexity depends on an array of factors that cannot be measured from the text, including reader characteristics, reading intention, situation, reader support, and the like. Similarly, Zakaluk and Samuels (1988) speak of factors inside and outside the reader›s head. In this sense, we propose the analogy of an obstacle course whose difficulty is a result of both the characteristics of the course (linguistic features) and the athlete (the person reading the

text). Modelling the complexity of the text is a significant challenge, as it requires hypothesising what would increase the cognitive load of the reader, based on measurements from the text.

Linguistic complexity analysis has multiple applications in education, including the selection, based on students' characteristics, of texts and textbooks that promote learning (Graesser et al., 2004). It is also a rarely discussed but important aspect of the test design process (Lane et al., 2015; McNamara et al., 2012; Visone, 2009). Controlling for the linguistic features of the item helps to mitigate construct-irrelevant variance due to language. Construct-irrelevant variance is the degree to which scores are influenced by processes extraneous to the purpose of a test. According to the *Standards*, the language difficulty of the item is one of the potential sources of irrelevant variance that should be controlled where possible (Joint Committee on Standards for Educational and Psychological Testing, 2014; Lane et al., 2015). The influence of language on item response has been demonstrated in several studies. For example, studies in Swedish (Persson, 2016); South African (Dempster & Reddy, 2007); and American (Martiniello, 2009) contexts have revealed the presence of language bias in standardised mathematics tests.

Linguistic aspects of assessment are not only a source of irrelevant variance. Their influence can be *relevant* when language is part of, or cannot be separated from, the skill being assessed (Avenia-Tapper & Llosa, 2015). For example, automatic language processing studies summarised by Crossley (2020) have demonstrated a statistical association between ESL writing quality scores and certain linguistic features relating to sentence complexity. This type of study supports the idea that automatic language processing can help measure linguistic complexity.

Measuring linguistic complexity

The complexity of English text has long been measured by readability formulas based on so-called "surface" features (Benjamin, 2012; Feng et al., 2010), typically average word and sentence length. The situation is similar on the French side: some readability formulas designed for English have been adapted for the French language, others created specifically for French (Mesnager, 1989). The reliance on surface features has received criticism due to its lack of consideration for the complex, subjective nature of reading (Boyer, 1992). Historical accounts of language complexity

modelling generally concur in concluding that the use of surface features is not sufficient to properly measure linguistic complexity and propose instead to move towards features theorised in psycholinguistics (Boyer, 1992; François, 2015; Kintsch & Vipond, 2014; McNamara et al., 2012; Zakaluk & Samuels, 1988). It is from this perspective that we created the linguistic analyzer presented in this study.

Why create a new tool?

ILSA, for Integrated Lexical-Syntactic Analyzer (in French ALSI-*analyseur lexico-syntaxique intégré*), is an automatic natural language processing tool designed to model the complexity of French text used in primary and secondary education. Tools have already been proposed for similar purposes; we will summarise their characteristics herein. Developed in the 1990s, the Quebec text analysis platform SATO-Calibrage (Daoust et al., 1996) is still available online. SATO-Calibrage extracts relatively simple features, compared to English-language tools such as Coh-Metrix (Grasser et al. 2011), which we describe in the following sections of our paper. *DMesure* and *Amesure* are based on work in computational linguistics (François, 2009; François & Fairon, 2012; François & Miltsakaki, 2012). *Dmesure* classifies French second language texts according to the six levels of the Common European Framework of Reference. *Amesure* specialises in estimating the readability of business French documents, which makes it less relevant for primary and secondary education. *ReaderBench* was designed in a similar way to *Dmesure* to analyze text in several languages, including French (Dascalu et al., 2013) and produces many linguistic features. However, *Dmesure* and *ReaderBench* were no longer available at the time of publication of this paper, prompting the creation of a new French text analyzer to meet current needs.

The present study

The general aim of this study is to present a new tool for linguistic complexity analysis and to make a two-part argument for its validity (Loye, 2018). The first part provides an overview of the ILSA tool, summarising its general functioning. It describes the types of features extracted and the procedures used to extract them. We draw on work in psycholinguistics and computational linguistics to explain what links these features to linguistic complexity. The second part explains the use of ILSA on a

corpus of 600 texts. We identify features that have an interesting potential for estimating the difficulty of texts, expressed on the 11-grade scale of the Quebec primary and secondary school system.

The ILSA tool

General operation of ILSA

ILSA is a natural language processing tool designed specifically for extracting features that characterize the linguistic complexity of French texts. The text is first decoded and then transformed into a list of annotated words, referred to as tokens.[1]The annotations include the lemma (canonical form of the word); the part of speech or word class (noun, verb, adjective, etc.); and the hierarchical relations between words and peripheral information (verb tenses, gender, number, etc.). Other annotations are added by cross-referencing with specialised databases, which we describe later herein. The result, shown in Figure 1, is a matrix where each row represents a word and each column represents a basic feature of the word.

Operations on the word matrix then produce various linguistic features at the sentence and text level. For example, the number of words divided by the number of sentences yields a linguistic feature: the average sentence length of the text. Similarly, analysing the word matrix allows us to identify which words are conjugated verbs, while dividing their number by the number of sentences in the text yields a feature indicating the average number of conjugated verbs per sentence. In the following paragraphs, we will detail the types of features extracted by ILSA as well as their theoretical basis and extraction procedures.

Typology of features extracted by ILSA

The features extracted by ALSI are part of a simple typology that aims at grouping the features into homogeneous categories based on similar characteristics of the text, while expressing a more subtle view of its complexity.

This typology is composed of two dimensions: 1) lexical complexity, which is associated with the words in the text; and 2) syntactic complexity, which is associated with the arrangement of words in sentences and the

1. The token is the smallest linguistic unit extracted by the analyzer; for simplicity, we use the term "word" in the rest of our paper.

Figure 1
Example of automatic analysis of a text excerpt

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	Lemma	Part of speech	Frequency according to ÉQOL	No. of characters	…
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. Decoding, lemmatisation, and part of speech identification with the *UDPipe* library for R (Straka et al., 2016). Word frequency and length from the ÉQOL database (Stanké et al., 2019).

role that words play in a given sentence. This choice is motivated by the fact that text complexity is frequently defined as the intersection of a lexical and a syntactic component (Ravid, 2005), a division consistent with the *Simple View of Reading* conceptual framework (Gough & Tunmer, 1986) while also being in line with the choice of features of the ATOS (Milone, 2014) and Lexile (Smith et al., 1989) English language analysis platforms. As illustrated in Table 1, the two dimensions are subdivided into three strata: 1) surface features; 2) features whose extraction requires the use of lexical databases or an automated syntactic analysis procedure; and 3) features that qualify linguistic complexity in a more comprehensive way (e.g. cohesion measures).

Table 1
Typology of features extracted by ILSA

	Lexicon	**Syntax**
Stratum 1 *Surface*	Measures of orthographic (number of characters) or syllabic (number of syllables) length	Length of sentence, number of commas
Stratum 2 *Intermediate*	Frequency of the word or lemma in a reference lexicon; age of exposure to the word in the school curriculum	Presence of certain sentence constituents (e.g. conjugated verbs); presence and length of phrases of interest (e.g. relative subordinate), sentence hierarchy
Stratum 3 *Comprehensive*	Lexical diversity; lexical cohesion	Syntactic cohesion

The features extracted by ILSA and discussed in this paper are listed in Table 4 in the supplementary materials. Note that ILSA uses a nomenclature where the suffix indicates which aggregation function was employed: *m* is a mean; *logm* is the average of the transformed values on a logarithmic scale; *p* is a proportion; *90* is the 90th percentile; and *i* is an index.

Annotation of the corpus

The texts to be analysed initially take the form of files in .txt, each file containing one text or excerpt. The text is decoded and annotated using the UDPipe library for R language version 0.8.9 (R Core Team, 2022; Wijffels, 2022). The annotations follow the typology of the Universal Dependency framework (De Marneffe et al., 2014). Annotation with *UDPipe* requires a French language model pre-trained by machine learning techniques. This model is what makes it possible to identify the part of speech (noun, verb, etc.) and the syntactic relations between words. The model used was *French-GSD 2.5* (Guillaume et al., 2019).

Lexical analysis

Lexical analysis produces features estimating the difficulty associated with words. In its first version, ILSA is based on three reference lexicons: Manulex, ÉQOL, and the *Liste orthographique du ministère de l'Éducation du Québec* (suggested spelling list compiled by Quebec's ministry of education). Manulex (Lété, 2004) contains about 49,000 words and was compiled from 54 textbooks (CP to CM2 school levels in the French system)

representing about two million words. ÉQOL (Stanké et al., 2019) is a lexicon created for the Quebec school system and contains 16,652 words from textbooks and children's literature for Grades 1 to 6. The MEQ spelling list is available through the *Franqus* project of the Université de Sherbrooke and contains 3,314 words classified into six grade levels from Grade 1 to 6, or 4,921 words after adding the missing plural forms for common nouns.

For frequency features, ILSA uses the standard frequency index (SFI) logarithmic scale. Lexical features in strata 1 and 2 are generated from the lexicon (list of unique words) of the text, with each lexeme counting only once.[2] If a word is missing from Manulex or ÉQOL, the missing frequency is imputed using the Good-Turing frequency estimation method (for an explanation see Gale and Sampson, 1995).

ILSA also estimates lexical diversity, which is the tendency to use a diverse vocabulary, as simpler texts are more likely to reuse the same words. Several formulas exist for this (Fergadiotis et al., 2015); ILSA calculates the type-token ratio and the Maas (1972) index. The type- token ratio estimates lexical diversity by dividing the number of unique words by the total number of words (text length). The Maas index is a similar measure, calculated according to this formula, where T is the total number of words and U the number of unique words:

Maas Index

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

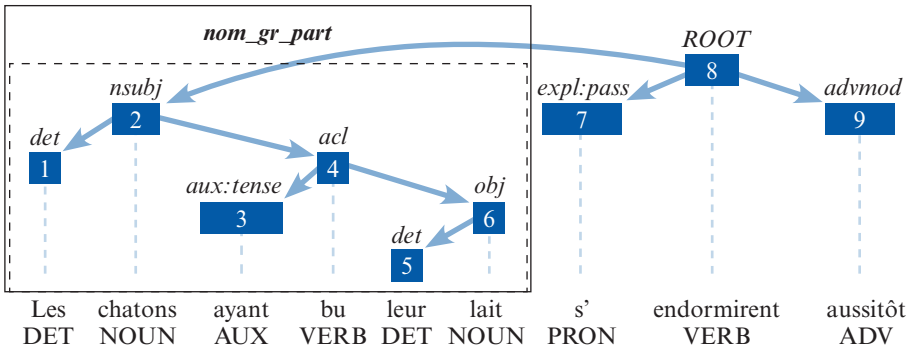
Syntactic analysis

While sentence length features are calculated directly from the annotated word list (see Figure 1), other syntactic features require additional analysis. A frequently used indicator of syntactic complexity is the height (or depth) of the sentence, counted in number of nodes (Sherstinova et al., 2020). Given a sentence represented as a hierarchical graph, its height corresponds to the most lengthy path connecting a word to the root of the sentence (Blache, 2010). ILSA uses for this calculation the tree representing

2. Words in the following categories are not considered: auxiliaries, proper nouns, numbers, determiners, and non-alphanumeric symbols.

the syntactic dependencies between words. Figure 2 shows an example of a syntax tree whose height is 4, the longest path from the word *leur* to the root of the sentence (*endormirent*).

Figure 2
Graphical representation of a sentence



Note. The box indicates a complex nominal group, in this case a noun with a participial phrase, detected using the *rsyntax* library (Welbers et al., 2020). See De Marneffe et al. (2014) for the list of acronyms. Figure produced with *rsyntax*.

Furthermore, ILSA extracts frequency or length features of syntactic constituents such as the verbal group, detected with the *rsyntax* library for R (Welbers et al., 2020).[3] In this first version of ILSA, we have targeted verbal groups and complex nominals. The verbal group (VG) is operationalised as a word group dominated by a conjugated verb. The complex nominal group (CNG) is operationalised in ILSA as a word group dominated by a noun, including its expansions. ILSA can detect the following expansions: the adjective, the participial group (see Figure 2), the relative subordinate, the prepositional group, and the infinitive group acting as the subject of the verb (e.g. *bien dormir est important*, "to sleep well is important").

3. Another analysis solution has recently been proposed by the fsca library for the R language (Vandeweerd, 2021), but at the time of publishing we had not yet had the opportunity to test it.

Cohesion measures

Increased lexical cohesion between sentences means that entities referenced in one sentence have a higher probability of being referenced again in the next sentence, which can facilitate reading (Graesser et al., 2004; Kintsch & Van Dijk, 1978). ILSA produces two measures of lexical cohesion: one compares all single lemmas in adjacent sentences, while the other compares only common and proper nouns. Lexical cohesion is then estimated by calculating the cosine similarity between adjacent sentences, which are then represented as word vectors (for an explanation of the calculation, see Han et al. 2012). This technique is notably used by the Coh-Metrix tool (Grasser et al., 2004).

To estimate syntactic cohesion, ILSA creates for each sentence of the text a vector containing three syntactic features that have been previously converted into standardised scores so they are one the same scale: sentence length, syntax tree height, and number of complex nominal groups. These features were chosen due to their being, in our preliminary tests, the three syntactic features most correlated with school grade level. Syntactic cohesion is then estimated by calculating the Euclidean distance between the vectors of adjacent sentences. The distance is converted into a measure of cohesion (similarity) by using the formula $1/(d + 1)$, where d is the distance.

Methodology

Overview of the methodology

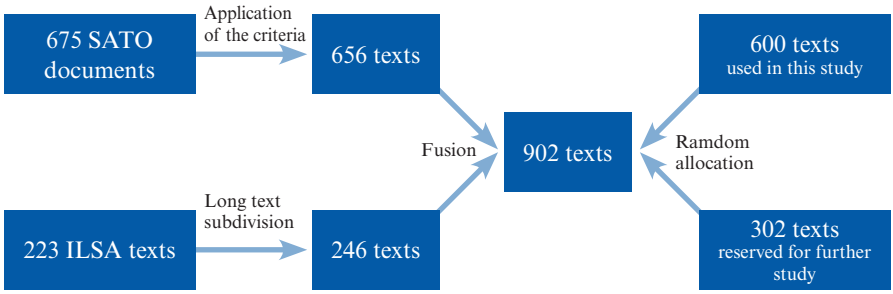
The objective of the analyses was to test the ability of the ILSA tool to extract features that characterise the linguistic complexity of French language texts. We first describe the composition of the corpus of 600 texts that we analysed using ILSA, then the feature selection procedure. We report measures of statistical association between the features (considered individually) and the grade level associated with the text.

Corpus used

The corpus used contained 600 texts distributed among 11 grades ranging from Grade 1 to 11 (secondary 5), according to the levels of the Quebec school system. The grade levels provided in the material were considered valid levels of difficulty for this study; the texts were not reclassified. The criteria for inclusion in the corpus were as follows: the text

had to be at least 30 words long (for primary school) or 100 words long (for secondary school), not mainly composed of dialogue or verse, and not mainly using the colloquial register. This corpus was constituted by combining two text sets according to a procedure illustrated in Figure 3.

Figure 3
Combining and allocating texts from the SATO and ILSA sets



The first text set originated from the development and calibration of the *SATO*-Calibrage analyzer (Daoust et al., 1996) and mainly contained excerpts from school textbooks and reading tests for Quebec's 11 grade levels. After splitting documents containing more than one text, applying exclusion criteria, and eliminating duplicates, the SATO set encompassed 656 texts. The second text set was created for this study and mainly included excerpts from textbooks published in Quebec after the year 2000. The level ranged from Grade 6 to 11 (secondary 5). To increase the size of the corpus while standardising text length, we subdivided the ILSA set texts whose number of words was more than twice the average. After these divisions, the ILSA set contained 246 texts. The following paratextual information was removed from both sets: page, paragraph, or line numbers and other marks added by the editor; remarks and definitions added in the margin; and titles and intertitles except when these formed a sentence including at least one conjugated verb. Since this information is usually added by the editor and is not present for all texts, it could have influenced the processing and biased the results.

The corpus formed by combining the SATO and ILSA sets contained 902 texts (43,820 sentences). We reserved about a third of this corpus (selected randomly) for a later study in text classification, bringing the size of the corpus used in the present study to 600 texts (29,709 sentences). The origin of the texts and their distribution across grade levels are shown in Table 2.

Table 2
Origin of the corpus used and distribution among the 11 school grade levels

	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
ILSA	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

Feature extraction and selection procedure

We analysed the 600 texts with ILSA, producing a matrix where each row is a text, each column is a feature, and each cell is the numerical value of the feature for the text (see Figure 1 for a simplified example). Given the large number of features and the fact that many features are very similar, we applied a selection procedure to eliminate features that are not very relevant for this study or that would contribute little information regarding the complexity of the text. This three-step procedure can be summarised as follows:

1) We exclude from the outset features reflecting the length of the text, such as the number of words, sentences, or paragraphs. These variables could have introduced a bias related to the way the corpus was formed, since some texts were subdivided.

2) Following the processing chain proposed by Taneja et al. (2014), we calculate the information gain of each feature, then remove the features whose information gain was zero. Information gain (IG) is a statistic indicating, in our case, to what extent the introduction of a variable improves the classification of texts compared to the chance level. In more technical terms, it is the decrease in Shannon's entropy conditional on the introduction of the variable (Karegowda et al., 2010; Yang & Pedersen, 2022). Removing features with zero IG eliminates features that are unlikely to add information regarding the level of difficulty (grade level associated with the text). At the same time, it eliminates features with zero or very low variance.

3) We then identify, using the *findLinearCombos* function of the caret library in R (Kuhn, 2011), groups of features that exhibit linear dependencies. These conflicts are managed by removing the features from the group one by one, while trying to preserve the features with the highest IG. Other conflicts are finally identified between combinations of features produced from the same linguistic measures or differing only in scale, and the feature of the group with the highest IG is retained.

The variables that passed each selection stage comprised the final selection of features. We further formed a reduced subset of six features by selecting the best representative (highest IG) of the six categories specified in the ILSA typology.

Statistical analysis

The objective of the analyses was to describe the statistical association between the selected characteristics and the level of difficulty of the text, expressed in grade levels (Grades 1 to 11, which is known in Quebec as Secondary 5) and considered as an ordinal variable. The statistical measures of association were the IG and Spearman's rho coefficient with 95% confidence intervals. Intervals were calculated using Fieller's method, which is less biased when the data have a non-normal distribution (Bishara & Hittner, 2017). To examine the progression of the features, we also calculated the median value of the features by grade level.

Results

The selection procedure was applied to an initial group of 42 features produced by the ILSA tool and considered relevant for this study. A complete list of the features considered, along with the reason for rejection if applicable, can be found in Table 4 in the supplementary material to this paper. Of the 42 features considered, 6 were removed due to zero IG; none were removed due to linear dependencies; and 18 features were removed to avoid conflicts between similar features (on a different scale or derived from the same measures). The final selection consisted of 20 features (8 lexical, 12 syntactic).

Table 3 shows the statistical association between grade level and the final feature selection by presenting the IG, Spearman coefficient, and feature type according to the typology described in this paper. For the 20 selected

features, the Spearman coefficients were significant at the $p < 0.001$ level and the confidence intervals of the correlation coefficients did not include the value 0. The magnitude of the correlations ranged from weak to strong according to the interpretation scales suggested by Akoglu (2018) for psychological research. Overall, the direction of the correlations was consistent with the nature of the features measured, i.e. a positive correlation when the numerical value of the feature is expected to increase with the difficulty of the text, and vice versa.

Table 3

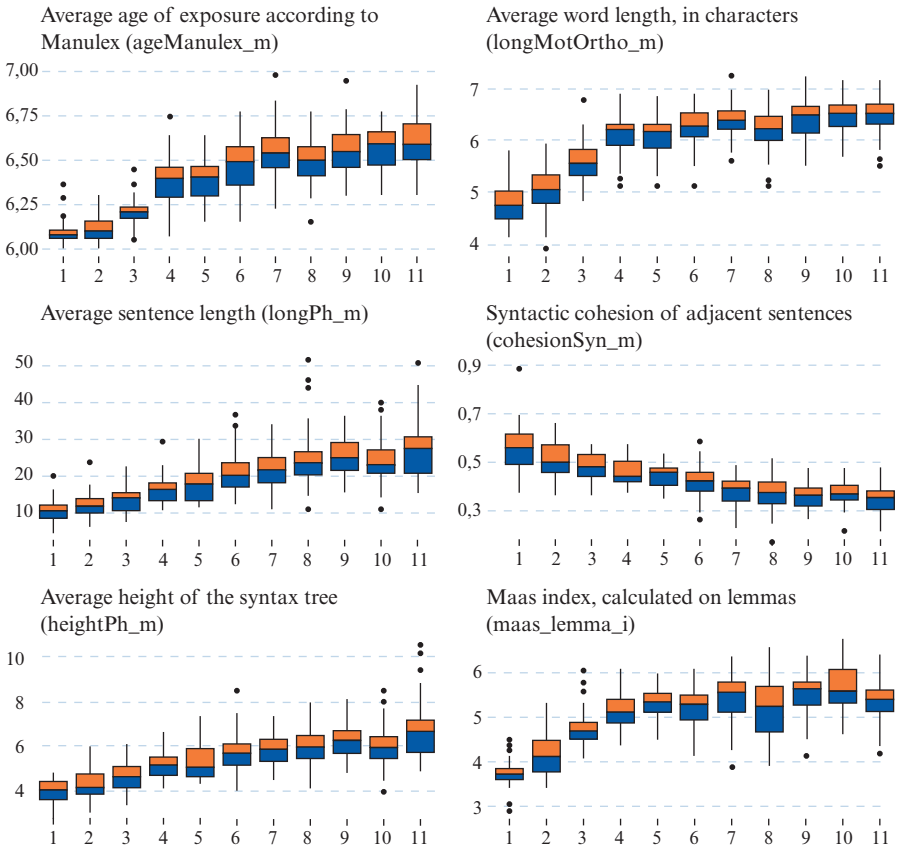
Measures of statistical association between feature and grade level

Features	IG	r_s [95% CI]	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. Statistics calculated for a corpus of 600 texts, for the 20 selected features and 11 grade levels. IG indicates information gain. Boldface indicates the feature with the highest IG per type. R_s indicates the Spearman correlation coefficient between each feature and the grade level of the text, with 95% confidence interval. All Spearman correlations in this table were statistically significant at the $p < 0.001$ threshold. The types of lexical and syntactic features are summarised in the current paper.

The features of the reduced selection (highest IG of their type) are shown in bold in Table 3. These include the average age of first appearance in the Manulex lexicon (*ageManulex_m*); the average orthographic length (*longMotOrtho_m*); the sentence length expressed in number of words (*longPh_m*); the sentence-to-sentence syntactic cohesion (*cohesionSyn_m*); the average height of the syntactic tree of the sentences (*heightPh_m*); and the Maas lexical diversity index computed on lemmas (*maas_lemma_i*). Figure 4 shows the distributions of these six features by

Figure 4
Box plots of the six features of the reduced selection, by grade level



Note. Results based on 600 texts. The x-axis shows the school year at primary (1-6) and secondary (7-11) levels in Quebec. The y-axis shows the unit of measurement for the feature. The box shows the 25 to 75 percentiles.

school year, revealing their progression as well as the presence of outliers. Five of the features shown in Figure 4 had a generally increasing progression; for example, the average sentence length went from about 10 words in Grade 1 to about 20 words in Grade 7 (secondary 1) to just under 30 words in Grade 11. In the case of syntactic cohesion, the progression was downward, suggesting that cohesion decreases as texts become more complex. Table 5 (supplementary material to the current paper) lists the resulting median values by feature and grade.

Discussion

In the present study, correlation analyses were used to test the potential of features extracted by ILSA to estimate the difficulty of French texts. Our presentation of the results focused on a selection of 20 features (8 lexical, 12 syntactic) that seemed particularly useful for estimating the difficulty of texts used in a school context. Three notable results were found regarding the nature of the features retained by the selection process.

First, a number of features showed a plateau effect. For example, the Maas lexical diversity index calculated on lemmas (*maas_lemma_i*) increases until the end of primary school and then stabilises. These plateau effects have also been described by Daoust et al (1996) and suggest that some linguistic features reach their limit of complexity at a certain point in the educational pathway. Another possible explanation is that ILSA may not be able to measure the progression of some features beyond a certain grade level. For example, some of the plateaus could be explained by the fact that the reference lexicons do not cover the secondary level (Grades 7-11). Future work could test the inclusion in ILSA of lexicons that also cover the secondary level to better estimate lexical complexity beyond Grade 6.

Second, our results show that the so-called "surface" features can indeed contribute to estimating text difficulty. Average word length (r_s = 0,63) and 90th percentile sentence length (r_s = 0,69) were among the features that most closely correlated with text difficulty. These results call into question the conclusions of authors who assert that this type of feature is worthless. However, they are consistent with a similar study by François and Fairon (2012), which found that word length and sentence length were among the features that correlated most closely with level of

difficulty (r_s = 0,48 and r_s = 0,61, respectively). A plausible explanation is that surface features, despite their apparent simplicity, remain effective intermediaries for assessing text difficulty. This explanation is in line with Szmrecsányi's (2004) findings regarding sentence length as an estimator of syntactic complexity.

Third, our results suggest that linguistic cohesion can help model text complexity. The feature measuring syntactic cohesion (*cohesionSyn_m*) showed a correlation of r_s = -0,66, a correlation of moderate magnitude according to the scales suggested by Akoglu (2018). This result is important as it adds empirical support to the hypothesis that cohesion affects comprehension (O'Reilly & McNamara, 2007). However, lexical cohesion (*simCosinNom_m*) showed a more modest correlation (r_s = -0,31), in line with the results obtained by Todirascu et al. (2016) on a French language corpus.

We have identified several limitations to the present study, the scope of which is based on the premise that the texts used are representative of what is found in the Quebec curriculum, and possibly in other francophone curricula. We also assumed that the grade level indicated by the material can be considered a reliable reference. More specifically, our results are limited by the fact that the most recent text sets (the ILSA bank) does not cover all 11 grades. Indeed, the texts from Grade 1 to 5 are generally older, as they were sourced from the set compiled by Daoust et al. (1996). One avenue to explore would be to add more recent texts, covering the period from Grade 1 to 5. Our results also depend on the linguistic features that the current version of ILSA can extract. Further work could incorporate feature types measuring other aspects of language, such as morphological complexity. Because our study was limited to analyses in which features were considered individually, multivariate analyses would be required to model text difficulty and assess the contribution of features to the model. The external validity of the instrument, its ability to estimate the grade level of new texts, could be tested by applying a multivariate model to a new corpus.

Conclusion

In this paper, we described ILSA, a new linguistic analysis tool that generates a variety of features for the purpose of assessing text complexity. After justifying the development of a new tool, the paper described the theoretical basis of ILSA and presented the feature extraction procedures.

The second part of the paper aimed at determining the features that were the most promising for assessing the academic level of texts in the Quebec French corpus. To this end, we applied ILSA to a corpus of 600 texts distributed across 11 levels of text difficulty. Correlational analyses showed the potential of the features to assess text difficulty, which supports the validity of the ILSA tool. The results further show that surface features are still relevant and highlight the potential of features measuring linguistic cohesion, particularly syntactic cohesion. The present study has, in sum, proposed features that can be extracted with the ILSA tool and which are associated with the linguistic complexity of texts used in schools in Quebec. This is a first step in validating ILSA, with further work needed to test its external validity.

In addition to the assessment of text difficulty, we see several applications of ILSA in the field of education. It could also assist in selecting or creating instructional materials at an appropriate language level, or with targeted content in French. In the context of language assessment, ILSA could be applied to written productions of learners of French as a second language to assess the development of vocabulary and syntax. Finally, a next generation version of the tool is in the works and will take the form of a web application to simplify its use.[4]

Reception: 12 October 2021
Final version: 16 March 2022
Acceptance: 18 May 2022

4. A prototype is available at https://gloignon.shinyapps.io/ALAIN_v3/

LIST OF REFERENCES

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* [conference presentation]. 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* [Doctoral thesis, 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. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (eds.), *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. In *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. *In 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. In *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. In 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? In *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 (3rd ed.)*. 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. In L.-G. Nilsson, T. Archer (ed.), *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. (ed.). (2015). *Handbook of Test Development* (2nd ed.). Routledge.

Lété, B. (2004). MANULEX : une base de données du lexique écrit adressé aux élèves. In É. Callaque, J. David (ed.) *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. In P. M. McCarthy (ed.), *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. In J. Sabatini (ed.), *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. In D. D. Ravid and H. B.-Z. Shyldkrot (ed.), *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. In *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. In G. Purnelle, C. Fairon & A. Dister (ed.). *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. In *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 ? In *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. In *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

Table 4
Complete list of features with measures of statistical association, derived from the selection and description

#	Feature	*IG*[a]	*r*[c] [95% CI][b]	Status[c]	Type[d]	Description
1	ageManulex_m	0,502	0,71 [0,67, 0,75] ***	SELEC	Lex. 2	Average age of exposure according to Manulex
2	freqManulexSfi_m	0,488	-0,7 [-0,74, -0,66] ***	SELEC	Lex. 2	Average standardised frequency according to Manulex
3	longMotOrtho_m	0,441	0,63 [0,58, 0,68] ***	SELEC	Lex. 1	Average word length, in characters
4	longMotSyll_m	0,437	0,63 [0,58, 0,68] ***	DOUB(3)	Lex. 1	Average syllable length
5	freqEqolSfi_m	0,432	-0,66 [-0,7, -0,61] ***	SELEC	Lex. 2	Average standardised frequency according to ÉQOL
6	longPh_m	0,407	0,7 [0,66, 0,74] ***	SELEC	Syn. 1	Average sentence length
7	cohesionSyn_m	0,398	-0,7 [-0,74, -0,66] ***	SELEC	Syn. 3	Average syntactic cohesion of adjacent sentences
8	heightPh_m	0,386	0,67 [0,62, 0,71] ***	SELEC	Syn. 2	Average height of the sentence syntax tree
9	ageEqol_m	0,374	0,65 [0,6, 0,7] ***	SELEC	Lex. 2	Average age of exposure according to ÉQOL
10	longPh_90	0,362	0,69 [0,64, 0,73] ***	DOUB(6)	Syn. 1	90th percentile of sentence length
11	motSeuilOrtho_p	0,34	0,61 [0,56, 0,66] ***	SELEC	Lex. 1	Prop. of words with more than eight characters
12	inManulex_p	0,339	-0,68 [-0,72, -0,63] ***	DOUB(1)	Lex. 2	Prop. of words in Manulex
13	wordMourningSyll_p	0,339	0,62 [0,57, 0,67] ***	DOUB(11)	Lex. 1	Prop. of words with more than three syllables

#	Feature	*IG*[a]	r^s [95% CI][b]	Status[c]	Type[d]	Description
14	ageMels_m	0,332	0,59 [0,53, 0,64] ***	SELEC	Lex. 2	Average age of exposure according to the Ministère de l'éducation du Québec spelling list
15	inEqol_p	0,332	-0,67 [-0,71, -0,62] ***	DOUB(5)	Lex. 2	Prop. of words in the ÉQOL list
16	maas_lemma_i	0,322	0,54 [0,48, 0,6] ***	SELEC	Lex. 3	Lexical diversity according to the Maas index calculated on lemmas
17	freqManulex_m	0,307	-0,56 [-0,61, -0,5] ***	DOUB(2)	Lex. 2	Average frequency according to Manulex
18	freqEqol_m	0,303	-0,56 [-0,61, -0,5] ***	DOUB(5)	Lex. 2	Average frequency according to ÉQOL
19	maas_token_i	0,301	0,52 [0,46, 0,58] ***	DOUB(16)	Lex. 3	Lexical diversity according to the Maas index calculated on words (tokens)
20	verbsConju_m	0,293	0,65 [0,6, 0,7] ***	SELEC	Syn. 2	Average number of conjugated verbs
21	GNC_m	0,287	0,6 [0,54, 0,65] ***	SELEC	Syn. 2	Average number of complex noun phrases per sentence
22	longPh30_p	0,282	0,67 [0,62, 0,71] ***	DOUB(6)	Syn. 1	Prop. of sentences with more than 30 words (see Daoust et al., 1996)
23	inMels_p	0,238	-0,59 [-0,64, -0,53] ***	DOUB(14)	Lex. 2	Prop. of words in the Ministère de l'éducation du Québec spelling list
24	GNCGr_m	0,235	0,55 [0,49, 0,6] ***	DOUB(21)	Syn. 2	Average number per sentence of complex noun phrases containing at least one object, participial, or prepositional phrase
25	comma_m	0,228	0,64 [0,59, 0,69] ***	SELEC	Syn. 1	Average number of commas per sentence

#	Feature	*IG*[a]	*r*[c] [95% CI][b]	Status[c]	Type[d]	Description
26	partPass_m	0,219	0,55 [0,49, 0,6] ***	SELEC	Syn. 2	Average number of past participles per sentence
27	partPres_m	0,172	0,49 [0,42, 0,55] ***	SELEC	Syn. 2	Average number of participles per sentence
28	GV_m	0,159	0,56 [0,5, 0,61] ***	SELEC	Syn. 2	Average number of verb groups per sentence
29	GVFin_m	0,152	0,54 [0,48, 0,6] ***	DOUB(28)	Syn. 2	Average number of finite verb groups (excluding infinitive verbs) per sentence
30	TTR_token_i	0,152	-0,32 [-0,39, -0,24] ***	DOUB(19)	Lex. 3	Lexical diversity, type-token ratio calculated on tokens
31	verbComplex_m	0,149	0,46 [0,39, 0,52] ***	DOUB(20)	Syn. 2	Prop. of conjugated verbs considered as complex (see Daoust et al., 1996)
32	phMarker_m	0,108	0,44 [0,37, 0,5] ***	SELEC	Syn. 2	Average number of argumentative connectors and textual organizers per sentence
33	TTR_lemma_i	0,107	-0,26 [-0,34, -0,18] ***	DOUB(19)	Lex. 3	Lexical diversity, type-token ratio calculated on lemmas
34	adp_p	0,106	0,3 [0,22, 0,37] ***	SELEC	Syn. 2	Prop. of all words in the text being prepositions
35	simCosinNom_m	0,079	-0,31 [-0,38, -0,23] ***	SELEC	Lex. 3	Lexical cohesion measured by cosine similarity of single common nouns in adjacent sentences (see Graesser et al., 2004)
36	sconj_p	0,058	0,081 [0, 0,16] *	DOUB(32)	Syn. 2	Prop. of all words in the text being coordinating conjunctions
37	adj_p	0	0,23 [0,15, 0,31] ***	GI = 0	Syn. 2	Prop. of all words in the text being adjectives

#	Feature	*IG*[a]	r^s [95% CI][b]	Status[c]	Type[d]	Description
38	adv_p	0	0,091 [0,01, 0,17] *	GI = 0	Syn. 2	Prop. of all words in the text being adverbs
39	longGNC_m	0	0,16 [0,08, 0,24] ***	GI = 0	Syn. 2	Average length of complex noun phrases
40	noun_p	0	0,13 [0,05, 0,21] **	GI = 0	Syn. 2	Prop. of all words in the text being common nouns
41	propn_p	0	0,13 [0,05, 0,21] **	GI = 0	Syn. 2	Prop. of all words in the text being proper nouns
42	simCosinLemma_m	0	-0,13 [-0,21, -0,05] **	GI = 0	Lex. 3	Lexical cohesion measured by cosine similarity of single lemmas of adjacent sentences (see Graesser et al., 2004)

Note. Statistics calculated on a corpus of 600 texts distributed among the 11 grades of the Quebec school system. [a]Information gain (IG) between the feature and the school level associated with the text. [b]Spearman correlation between the feature and the grade level associated with the text. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. cResult of the selection procedure. SÉLEC: feature in the final selection; DOUB: feature removed because it is a near duplicate with the selected feature whose number is indicated in brackets; IG = 0: feature removed, because its IG was zero. [d]Feature type according to the classification proposed in this paper.

Table 5
Median of features by school year (selection of 20 features)

Feature	Grade level										
	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
heightPh_m	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
phMarker_m	0,07	0,08	0,12	0,16	0,16	0,22	0,22	0,21	0,32	0,28	0,29
simCosinName_m	0,13	0,14	0,12	0,13	0,08	0,11	0,04	0,06	0,07	0,05	0,05
verbsConju_m	1,18	1,55	1,57	1,78	1,87	2,03	2,21	2,4	2,61	2,37	2,71
comma_m	0,54	0,73	0,78	0,84	1	1,18	1,47	1,6	1,83	1,7	1,85

Note. Median values calculated from 600 texts; levels correspond to the 11 grades, 6 primary (1-6) and 5 secondary (7-11), of the Quebec school system.