Université de Montréal

Une approche computationnelle de la complexité linguistique par le traitement automatique du langage naturel et l'oculométrie

Par

Guillaume Loignon

Département d'administration et fondements de l'éducation

Faculté des sciences de l'éducation

Thèse présentée en vue de l'obtention du grade de Philosophiae Doctor (Ph.D.)

en sciences de l'éducation, option mesure et évaluation

Mai 2021

Université de Montréal

Unité académique : Département d'administration de l'éducation, faculté des sciences de l'éducation

---

Cette thèse intitulée

Une approche computationnelle de la difficulté du texte par le traitement automatique du langage naturel et l'oculométrie

Présenté par
Guillaume Loignon

A été évaluée par un jury composé des personnes suivantes
Sébastien Béland
Président-rapporteur

Nathalie Loye
Directrice de recherche

Phaedra Royle
Membre du jury

Luc Paquette
Examinateur externe

Résumé de la thèse

Le manque d'intégration des sciences cognitives et de la psychométrie est régulièrement déploré – et ignoré. En mesure et évaluation de la lecture, une manifestation de ce problème est l'évitement théorique concernant les sources de difficulté linguistiques et les processus cognitifs associés à la compréhension de texte. Pour faciliter le rapprochement souhaité entre sciences cognitives et psychométrie, nous proposons d'adopter une approche computationnelle. En considérant les procédures informatiques comme des représentations simplifiées et partielles de théories cognitivistes, une approche computationnelle facilite l'intégration d'éléments théoriques en psychométrie, ainsi que l'élaboration de théories en psychologie cognitive. La présente thèse étudie la contribution d'une approche computationnelle à la mesure de deux facettes de la complexité linguistique, abordées à travers des perspectives complémentaires. La complexité intrinsèque du texte est abordée du point de vue du traitement automatique du langage naturel, avec pour objectif d'identifier et de mesurer les attributs (caractéristiques mesurables) qui modélisent le mieux la difficulté du texte. L'article 1 présente ALSI (pour Analyseur Lexico-syntaxique intégré), un nouvel outil de traitement automatisé du langage naturel qui extrait une variété d'attributs linguistiques, principalement issus de la recherche en psycholinguistique et en linguistique computationnelle. Nous évaluons ensuite le potentiel des attributs pour estimer la difficulté du texte. L'article 2 emploie ALSI et des méthodes d'apprentissage statistique pour estimer la difficulté de textes scolaires québécois. Dans le second volet de la thèse, la complexité associée aux processus de lecture est abordée sous l'angle de l'oculométrie, qui permet de faire des inférences quant à la charge cognitive et aux stratégies d'allocation de l'attention visuelle en lecture. L'article 3 décrit une méthodologie d'analyse des enregistrements d'oculométrie mobile à l'aide de techniques de vision par ordinateur (une branche de l'intelligence artificielle); cette

méthodologie est ensuite testée sur des données de simulation. L'article 4 déploie la même méthodologie dans le cadre d'une expérience pilote d'oculométrie comparant les processus de lecture de novices et d'experts répondant à un test de compréhension du texte argumentatif. Dans l'ensemble, nos travaux montrent qu'il est possible d'obtenir des résultats probants en combinant des apports théoriques à une approche computationnelle mobilisant des techniques d'apprentissage statistique. Les outils créés ou perfectionnés dans le cadre de cette thèse constituent une avancée significative dans le développement des technologies numériques en mesure et évaluation de la lecture, avec des retombées à anticiper en contexte scolaire comme en recherche.

Mots-clés : complexité linguistique, compréhension de texte, traitement automatique du langage naturel, oculométrie, vision par ordinateur, mesure de l'éducation, linguistique computationnelle, apprentissage supervisé

General abstract

The lack of integration of cognitive science and psychometrics is commonly deplored - and ignored. In the assessment of reading, one manifestation of this problem is a theoretical avoidance regarding sources of text difficulty and cognitive processes underlying text comprehension. To facilitate the desired integration of cognitive science and psychometrics, we adopt a computational approach. By considering computational procedures as simplified and partial representations of cognitivist models, a computational approach facilitates the integration of theoretical elements in psychometrics, as well as the development of theories in cognitive psychology. This thesis studies the contribution of a computational perspective to the measurement of two facets of linguistic complexity, using complementary perspectives. Intrinsic text complexity is approached from the perspective of natural language processing, with the goal of identifying and measuring text features that best model text difficulty. Paper 1 introduces ISLA (Integrated Lexico-Syntactic Analyzer), a new natural language processing tool that extracts a variety of linguistic features from French text, primarily taken from research in psycholinguistics and computational linguistics. We then evaluate the features' potential to estimate text difficulty. Paper 2 uses ISLA and statistical learning methods to estimate difficulty of texts used in primary and secondary education in Quebec. In the second part of the thesis, complexity associated with reading processes is addressed using eye-tracking, which allows inferences to be made about cognitive load and visual attention allocation strategies in reading. Paper 3 describes a methodology for analyzing mobile eye-tracking recordings using computer vision techniques (a branch of artificial intelligence); this methodology is then tested on simulated data. Paper 4 deploys the same methodology in the context of an eye-tracking pilot experiment comparing reading processes in novices and experts during an argumentative text

comprehension test. Overall, our work demonstrates that it is possible to obtain convincing results by combining theoretical contributions with a computational approach using statistical learning techniques. The tools created or perfected in the context of this thesis constitute a significant advance in the development of digital technologies for the measurement and evaluation of reading, with easy-to-identify applications in both academic and research contexts.

Keywords: linguistic complexity, reading comprehension, natural language processing, eye tracking, computer vision, educational measurement, computational linguistics, supervised learning

Remerciements

« Je suis à la recherche d'un projet de doctorat en sciences de l'éducation. » Ces mots figurent tout en haut du premier message que j'ai envoyé à Nathalie Loye. La réponse : « Je suis disposée à vous rencontrer pour en discuter. » Une formule qui s'est manifestée sous de nombreuses formes dans les années suivantes. Merci, Nathalie, pour la grande générosité avec laquelle tu partages ton expertise et ton temps, pour avoir soulevé les bonnes questions, et pour m'avoir fourni un contexte dans lequel j'ai pu travailler, apprendre, et développer mes habiletés comme chercheur en éducation.

Il n'aurait pas été possible de compléter thèse sans le support de mon épouse, Émilie Courteau. Support émotionnel et logistique, mais avant tout scientifique. Dès les premiers moments, Émilie a été la co-directrice de recherche non officielle de ce projet de thèse, et lui a apporté la perspective inestimable des sciences du langage.

Je tiens ensuite à remercier Angel Arias, pour une suggestion judicieuse de modèles théoriques en compréhension de texte. Sébastien Béland, pour les conseils statistiques, et pour des suggestions de lecture qui furent très formatrices. Christophe Chénier, pour de solides réflexions sur le réalisme épistémique en psychométrie qui ont joué un rôle important dans l'élaboration du contexte théorique de cette thèse. François Daoust, merci d'avoir partagé l'imposante banque de textes de SATO-Calibrage, dont dépend largement cette thèse (articles 1 et 2), et pour votre contribution pionnière au développement des méthodologies en analyse computationnelle du texte en langue française. Sylvie Marcotte, pour une aide généreuse lors de l'élaboration de l'épreuve de compréhension de texte (article 4) et en prévision de fructueux travaux. Philippe Messier, pour une collaboration qui s'est dédoublée en initiation aux approches

multimodales et à l'écriture d'article, et pour les conseils sur la vie académique qui se sont avérés judicieux et précieux.

Je voudrais enfin à souligner la contribution des assistants de recherche Achille Vigneault et Éli Paré, qui ont abattu une immense quantité de travail de collecte et préparation de données afin que les résultats de cette thèse aient plus grande portée.

Table des matières

Index des figures

Les figures et tableaux sont numérotés par section.

Introduction générale

Article 1

Article 2

Article 3

Article 4

Discussion générale

Index des tableaux

Article 1

Article 2

Article 3

Article 4

Sigles et abréviations pour la communication de résultats	
M	Moyenne
ET	Écart-type
IC	Intervalle de confiance. Les valeurs entre crochets indiquent la borne inférieure et supérieure. Le seuil utilisé pour les intervalles de confiance était de 95%
p	valeur *p* associée à la statistique rapportée. Le sigle *p aj*. indique que la valeur *p* a été ajustée à la baisse pour tenir compte de multiples comparaisons
F	Statistique *F* sur l'égalité des variances.
r	Coefficient de corrélation de Pearson. Aussi utilisé pour exprimer la taille d'effet d'un test de Wilcoxon, laquelle s'interprète à la manière d'un coefficient de Pearson.
rs	Coefficient de corrélation de Spearman
bm	Statistique produite par le test de Brunner-Munzel.
H	Statistique *H* produite par le test de Kruskal-Wallis et son expansion le test de Scheirer-Ray-Hare
Z de Fisher	Score Z issu du transformé de Fisher appliqué aux coefficients de corrélation
d	Taille d'effet *d* de Cohen.
dl	Nombre de degrés de liberté
TELC	Taille d'effet en langage commun, équivalent à la probabilité de supériorité
JE	Justesse exacte
JA	Justesse ajustée
EAM	Erreur absolue moyenne
EQM	Erreur quadratique moyenne
sec	Secondes
ms	Millisecondes
Autres abréviations utiles	
ALSI	Analyseur Lexico-Syntaxique Intégré
MGM	Mobile Gaze Mapping
RMN	Régression multinomiale
SVM	Séparateurs à vaste marges (*support vector machines*)
TALN	Traitement automatisé du langage naturel

1. Introduction générale

Il y a près de 40 ans, Mislevy (1982) décrivait la psychométrie comme de la psychologie datant du XIXe siècle mais utilisant des méthodes statistiques contemporaines. La boutade visait le déficit théorique en psychométrie, la résistance aux apports théoriques de la psychologie. Le manque de théories explicatives est un problème toujours actuel qui n'est pas limité à la psychométrie. En raison d'un manque de référents théoriques partagés, l'ensemble des sciences comportementales sont en crise paradigmatique, ce qui nuit à la collaboration et au cumul du savoir (Kline, 2019). En psychométrie, une manifestation de ce problème est la pratique consistant à expliquer les données observées en recourant uniquement à un modèle statistique, avec peu d'égards pour la théorie scientifique du domaine (Rhemtulla et al., 2018). De même, des tâches couramment utilisées dans la recherche en psychologie ont des propriétés psychométriques questionnables (Eisenberg et al., 2019). Les invitations à mieux intégrer les théories psychologiques à la psychométrie se sont pourtant multipliées au fil des époques (Lawrence et Shea, 2008). En mesure de l'éducation plus spécifiquement, plusieurs ont soulevé que les pratiques actuelles ignorent la dimension sociocognitive de la réponse à l'item, se privant ainsi d'informations précieuses concernant, par exemple, les stratégies de réponse et la progression des apprentissages (Mislevy, 2008; Pellegrino et al., 2001). Des travaux ont également souligné le besoin de disposer d'informations riches quant aux processus cognitifs déployés par les élèves (Huff et al., 2016; Huff et Goodman, 2007), rendant l'apport de la psychologie encore plus important.

Pour sortir de la crise théorique, certains spécialistes de la mesure ont suggéré que les méthodes psychométriques soient combinées à des théories tirées de la psychologie cognitive (Borsboom, 2006b; van der Maas et al., 2011). Selon cette approche, la psychologie fournirait

une base empirique pour bon nombre des décisions clés dans l'élaboration des tests, notamment le choix des compétences à mesurer, la structure de ces compétences et les tâches pour les mesurer. Des critiques ont cependant fait valoir que les théories issues de la psychologie sont difficiles à appliquer dans le domaine de l'éducation (Leighton et Gierl, 2011). D'autres soulèvent que les théories cognitivistes sont souvent peu développées, voire inexistantes dans plusieurs domaines qui intéressent la psychométrie (Kane et Bejar, 2014). Devant ces difficultés, Mislevy (2003) et Kane (1992, 2006a, 2013) ont proposé des cadres méthodologiques dont Loye (2018) a formulé une synthèse, qui structurent la création et l'utilisation de tests dans une approche que nous qualifions de pragmatiste. L'approche pragmatiste concède qu'un modèle empirique est nécessaire pour créer et utiliser un test en respectant des exigences de validité. Ce modèle n'est cependant pas formé en puisant dans la psychologie cognitive, mais par une analyse du domaine sur lequel porte le test, et en considérant les conséquences possibles de l'utilisation du test. Le modèle ainsi créé est une description, au niveau du comportement observable, de la tâche effectuée par l'élève, d'éléments de curriculum à maitriser, etc. (Zieky, 2014). En contexte de mesure de l'éducation, cette approche présente l'avantage d'être ancrée dans l'ontologie adoptée sur le terrain, ce qui évite de devoir adapter à la réalité scolaire des modèles cognitifs issus de la recherche. Un désavantage est que les modèles empiriques créés pragmatiquement dans cette approche ne sont pas fondés dans la recherche scientifique, ce qui risque d'exacerber la crise théorique de la psychométrie (Borsboom, 2006a). Nous nous questionnons de plus quant à la capacité à émettre des inférences au sujet des processus de réponse de l'élève à partir d'un modèle formulé au niveau béhavioral. L'objectif est alors de créer un test; le modèle *ad hoc* a un rôle instrumental dans la démarche d'élaboration du test, mais n'est pas un résultat de la démarche. En phase avec la tendance actuelle représentée par les travaux de Kane (2006b, 2013),

les considérations entourant la validité du test se limitent à des questions proprement psychométriques et aux conséquences pour le public de l'utilisation prévue du test, par exemple éviter la discrimination dans l'admission aux programmes d'études; cette définition de la validité n'englobe pas l'utilisation du test comme instrument de collecte de données dans un contexte scientifique (Truijens et al., 2019). En somme, comme le suggèrent Wijsen et al. (sous presse), la psychométrie actuelle se décrit comme la construction de modèles statistiques abstraits qui peuvent être utilisés dans la recherche psychologique ou éducative mais qui n'ont pas d'interprétation particulière en termes d'attributs psychologiques spécifiques.

C'est dans ce contexte que von Davier a récemment mis de l'avant la psychométrie computationnelle, qui s'appuie sur des méthodes d'apprentissage machine et propose de faciliter la création de modèles à partir de données brutes (von Davier et al., 2019). Les sources d'information sont de natures variées : audio, vidéo faciale, questionnaires à choix multiples, temps de réponse, données de journalisation relatant l'interaction avec un système informatique (*log files*), etc. Ces données sont analysées en mettant à contribution des méthodes d'apprentissage machine afin d'extraire leurs attributs (*features*), qui sont des caractéristiques mesurables dont l'intérêt réside dans la possibilité de distinguer des entités (par exemple, des visages individuels distingués par la forme des yeux et du nez). Les attributs deviennent à leur tour les intrants de modèles psychométriques produisant un portrait multidimensionnel des habiletés de l'élève, tel qu'illustré à la Figure 1. L'approche compte sur la variété de données collectées et sur l'inclusion de séries temporelles (*time series data*) pour rendre compte des processus de réponse de l'élève.

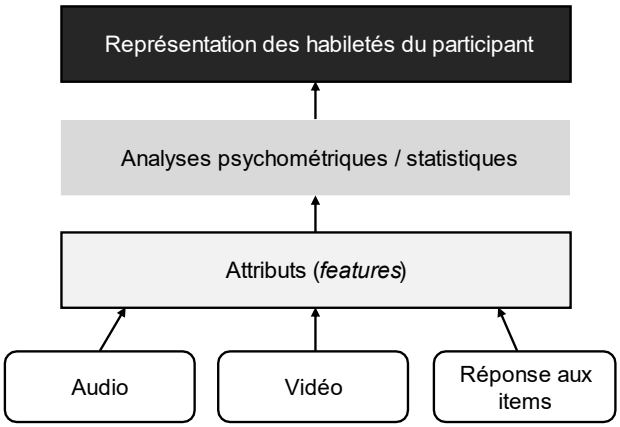


Figure 1. Schéma illustrant une approche computationnelle de la psychométrie à partir de trois sources d'information. Adapté depuis von Davier (2017).

La psychométrie computationnelle est une approche encore jeune et en construction. Il est difficile de se prononcer quant à son potentiel à solutionner – ou aggraver – la crise théorique en psychométrie. L'attrait de l'apprentissage machine pourrait même amener les tenants de cette approche, qui est fondamentalement centrée sur les données, à ignorer les théories psychologiques en faveur d'un hyper empirisme (Zumbo, 2020). Nous décelons néanmoins en psychométrie computationnelle l'opportunité d'une meilleure intégration des notions de sciences cognitives et de la psychométrie. Plusieurs auteurs, dont Johnson-Laird (1983), par exemple, ont formulé un cadre computationnel de recherche dans lequel des aspects des théories psychologiques sont mis en œuvre dans des procédures pouvant être exécutées sous forme de logiciel. En psychométrie computationnelle, les sciences cognitives pourraient être mises à contribution lors de choix méthodologiques importants, à savoir : quelles données devraient être collectées, quel dispositif expérimental employer pour les collecter, quel ensemble d'attributs extraire depuis ces données, etc. De manière plus importante, les sciences cognitives pourraient contribuer à expliquer et justifier les inférences inductives consistant à passer d'une représentation de l'information à une autre (Gärdenfors, 1990, 2004). Les publications actuelles s'inscrivant dans le cadre de psychométrie computationnelle mis de l'avant par von Davier

(Polyak et al., 2017) sont plutôt vagues quant à la justification des inférences inductives. Le système de psychométrie computationnelle décrit par Khan (2017), par exemple, traduit le fait de toucher son visage comme de l'ennui, sans référer à une théorie psychologique justifiant cette inférence. En l'absence d'un contexte de justification, les inférences inductives risquent d'être naïves, et en porte à faux avec les théories scientifiques.

1.1. Pourquoi une approche computationnelle de la difficulté du texte ?

Nous aurons l'occasion, à travers les quatre articles de cette thèse, de montrer qu'il est possible d'obtenir des résultats probants en intégrant des éléments de théories cognitives dans une approche computationnelle utilisant des techniques d'apprentissage statistique. L'objectif général de la thèse est de montrer comment une approche computationnelle peut faciliter l'intégration de la psychométrie et des sciences cognitives en mesure de la difficulté du texte. Le choix du domaine d'application se justifie d'abord par le fait que la lecture a un rôle central en éducation autant comme habileté à maitriser que comme intermédiaire par lequel l'élève développe d'autres habiletés et acquiert une grande partie de ses connaissances. La lecture est également une composante de tout test et épreuve ayant une dimension écrite, peu importe la matière. Dans l'élaboration de curriculum et de manuels scolaires, mesurer la difficulté du texte est essentiel pour sélectionner quels textes ont les caractéristiques et le niveau appropriés pour guider et évaluer les apprentissages des élèves (McNamara et Kendeou, 2011). Du côté de la recherche en sciences cognitives, la lecture occupe un rôle central comme objet d'étude, ou comme élément d'un dispositif expérimental pour étudier d'autres processus cognitifs; c'est le cas, pour prendre un exemple classique, pour la tâche de Stroop[1] (MacLeod, 2005). En

[1] Tâche classique dont les stimuli visuels prennent la forme de mots colorés indiquant eux-mêmes des couleurs, (par exemple vert ou rouge).

choisissant le domaine de la lecture, nous intervenons dans un contexte où il est plus aisé de mettre à contribution des disciplines s'inscrivant dans les sciences cognitives et qui proposent déjà des éléments théoriques pertinents, comme la psycholinguistique et la linguistique computationnelle. La lecture est, en termes simples, un terrain propice à l'intégration du domaine de la psychométrie et des sciences cognitives. Cette intégration est désirable puisqu'elle a le potentiel de fournir autant un narratif théorique à la mesure des habiletés de lecture qu'un cadre plus robuste pour l'étude de la difficulté du texte.

Le domaine de la mesure a besoin d'une meilleure spécification des processus cognitifs associés à la lecture, et des sources de difficulté des textes. Von Moere (2012) note une tendance, dans l'élaboration de tests de langue, à ignorer des notions pertinentes provenant de la psycholinguistique, comme l'influence de mots inconnus, ou à faible fréquence d'occurrence dans la langue, sur la fluidité du traitement de l'information linguistique. Les aspects linguistiques des items sont souvent négligés dans la démarche d'élaboration des tests (Lane et al., 2015; Visone, 2009). Cet évitement théorique n'est pas sans conséquence : plusieurs travaux ont démontré que les scores de tests standardisés à enjeux élevés, par exemple en mathématiques et en sciences, pouvaient être influencés par des aspects linguistiques qui auraient dû être contrôlés lors de l'élaboration du test (Dempster et Reddy, 2007; Martiniello, 2009; Persson, 2016).

De même, la recherche dans le domaine de la compréhension de texte pourrait profiter de méthodologies et outils plus rigoureux et conformes avec les théories actuelles en psychométrie. Le langage est un ensemble d'habiletés multidimensionnelles qu'on peut difficilement capturer avec les méthodologies courantes, lesquelles favorisent la mesure unidimensionnelle (Borsboom, 2006b). Par exemple, Denman et al. (2017) ont analysé la méthodologie d'études portant sur des

tests linguistiques conçus pour usage clinique, tels que le *Assessment of Literacy and Language* (Lombardino et al., 2005). Leurs résultats démontrent la récurrence de failles méthodologiques importantes, notamment l'application de méthodes statistiques requérant que tous les items mesurent un même construit, sans que l'unidimensionnalité du test n'ait d'abord été confirmée.

En somme, la lecture est un ensemble multidimensionnel d'habiletés qui joue un rôle central en éducation, et représente un défi majeur pour la mesure : mieux mesurer la lecture demande d'avoir une définition précise des entités et processus à mesurer, alors que mieux comprendre la lecture demande des méthodes rigoureuses pour mesurer ses multiples dimensions. Une approche computationnelle est particulièrement intéressante pour résoudre ce paradoxe, car elle excelle dans la modélisation de processus multidimensionnels, et permet d'implémenter dans une procédure informatique des modèles théoriques et des principes méthodologiques éprouvés, tout en contribuant à faire émerger ce qui est le plus pertinent à mesurer.

Nous définissons la complexité du texte comme l'interaction de facteurs multiples pouvant se regrouper en deux facettes : la difficulté intrinsèque au texte, et la difficulté associée aux processus de lecture – cette définition est détaillée plus loin. La thèse aborde ces deux facettes à travers des perspectives complémentaires. La difficulté intrinsèque au texte est abordée sous l'angle de l'analyse automatisée du langage naturel, avec l'objectif spécifique d'identifier et mesurer des attributs du texte permettant d'en modéliser la difficulté. La linguistique computationnelle propose déjà des méthodes et outils qui rejoignent cet objectif, tout en empruntant des éléments de théories psycholinguistiques sur la difficulté du texte. Nous constatons cependant qu'il n'existe pas d'outils de ce type qui soit adapté au contexte éducatif québécois francophone; le seul outil disponible est SATO-Calibrage (Daoust et al., 1996), qui a

maintenant 25 ans et n'a pu profiter des avancées dans le domaine. L'autre facette de la complexité linguistique, qui découle des processus de lecture, est abordée sous l'angle de l'oculométrie, avec l'objectif spécifique de créer et tester une chaine de traitement automatisé et robuste pour l'étude de la lecture par l'oculométrie mobile.

L'outil d'analyse linguistique et la chaine de traitement des données oculométriques s'inscrivent dans une approche computationnelle et ont le potentiel d'aider des équipes de recherche à faire l'arrimage souhaité entre psychométrie et sciences cognitives. Il y a une intention de validation derrière les travaux de cette thèse. Nous présentons des éléments de preuve montrant que l'outil d'analyse linguistique permet d'estimer la difficulté du texte avec une fidélité équivalente ou supérieure à l'état de l'art, et que la chaine de traitement proposée facilite et complémente l'étude oculométrique des processus de lecture. La suite de la thèse est divisée ainsi : nous introduisons d'abord les principaux éléments de cadre théorique impliqués dans cette thèse, soit la complexité linguistique, le traitement du langage naturel, l'oculométrie et la vision par ordinateur. Nous faisons ensuite un survol plus détaillé des quatre articles. Les sections suivantes sont composées des articles, entrecoupés de transitions résumant les résultats de l'article dans le contexte plus général de la thèse. Nous offrons finalement une synthèse générale des résultats puis discutons de leur intégration à la littérature récente, de même que de leurs limitations et applications possibles en recherche et en éducation.

1.2. Cadre théorique

1.2.1. Modéliser la complexité intrinsèque du texte

La difficulté d'un texte peut être définie par une combinaison de facteurs spécifiques au texte et de facteurs dépendant plutôt du lecteur et du contexte de lecture (Ravid, 2005; Zakaluk et Samuels, 1988). Nous proposons, pour illustrer cette définition, l'analogie d'un parcours à

obstacles. Bien que le parcours soit le même et que certains obstacles aient tendance à exiger un effort supplémentaire, la difficulté perçue du parcours dépend fortement des compétences des athlètes, de leur expérience, de leurs stratégies, etc. De même, la difficulté du texte émerge de l'interaction entre les attributs du texte et les connaissances et habiletés du lecteur, tel qu'illustré à la Figure 2. Les attributs sont les caractéristiques du texte telles que mesurées; deux exemples classiques étant la longueur moyenne des phrases et des mots (Solnyshkina et al., 2017).

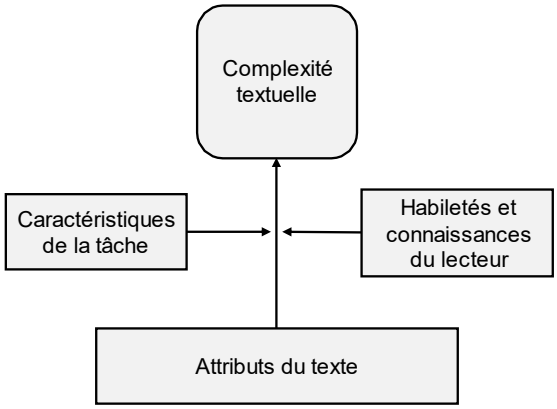


Figure 2. Schéma illustrant l'émergence de la complexité textuelle.

La complexité intrinsèque au texte peut elle-même se diviser en deux principales dimensions: la complexité lexicale et la complexité syntaxique (Ravid, 2005). La complexité lexicale est la difficulté relative au lexique du texte, qui est l'ensemble des lexèmes (mots distincts) utilisés dans un texte. La complexité syntaxique est la difficulté relative à la structure grammaticale qui relie les lexèmes. En empruntant un élément à la théorie de la charge cognitive (Clevinger, 2014; Sweller & Chandler, 1991), la complexité textuelle peut finalement se définir comme la propension des attributs lexicaux et syntaxiques du texte à induire une charge cognitive au lecteur, considérant ses habiletés et connaissances.

Suivant notre définition, la complexité intrinsèque d'un texte est estimée en fonction de ses attributs. La Figure 3 illustre les deux composantes requises pour estimer la difficulté intrinsèque au texte : un ensemble d'attributs, et un modèle de mesure ou de classification. Le

modèle de mesure est, dans ce contexte, une représentation systématique des liens entre les attributs et la difficulté du texte; il se traduit par une série d'étapes permettant de passer des valeurs des attributs à un score de difficulté. Un modèle de classification est un type particulier de modèle de mesure dont le résultat est une variable catégorielle, dans notre cas une année scolaire.

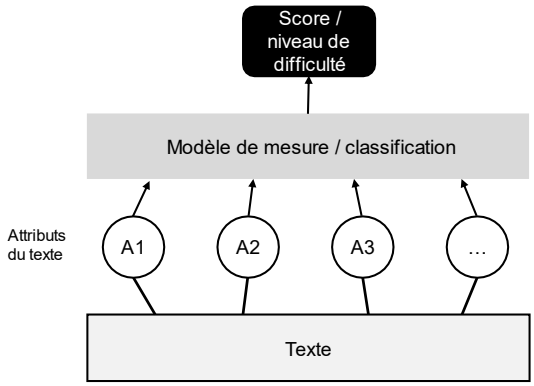


Figure 3. Schéma illustrant la mesure des attributs puis l'application d'un modèle de mesure ou de classification pour estimer la complexité linguistique du texte.

Le principe général illustré à la Figure 3 sous-tend les formules de lisibilité simples comme les approches plus sophistiquées utilisant l'apprentissage statistique. Les premiers essais systématiques pour estimer la complexité intrinsèque du texte prenaient la forme d'équations linéaires relativement simples appelées formules de lisibilité. Prenons comme exemple le *Flesch reading ease* (Flesch, 1948), qui compte parmi les formules de lisibilité les plus connues pour le texte en langue anglaise. La formule se définit comme suit, *RE* (pour *reading ease*) étant le score de difficulté, *lm* étant la longueur moyenne des mots, et *lp* étant la longueur moyenne des phrases :

$$RE = 206{,}835 - 84{,}6lm + 1{,}15lp$$

Cette formule comporte deux attributs : la longueur moyenne des mots (*lm*), qui sert d'indicateur simple de la difficulté lexicale, et la longueur moyenne des phrases (*lp*), qui approxime la complexité syntaxique. Le modèle de mesure se résume à une équation linéaire du premier degré.

Pour déterminer la valeur des paramètres de sa formule de lisibilité, Flesch (1948) a d'abord calculé la longueur moyenne des phrases et des mots d'un ensemble de textes dont le niveau de difficulté était connu et considéré comme une référence. Il a ensuite fait converger un modèle de régression linéaire dont les variables dépendantes étaient la longueur moyenne des mots et des phrases, et la variable dépendante était le niveau de difficulté du texte. Le modèle résultant décrit ainsi l'association entre les attributs et le score de difficulté pour l'ensemble des textes initialement utilisés.

Les méthodes contemporaines d'estimation de la difficulté des textes suivent le même cadre général que celui utilisé par Flesch en 1948, mais sont devenues plus sophistiquées grâce à la convergence d'innovations théoriques, technologiques et méthodologiques. Mentionnons d'abord le développement du traitement automatisé du langage naturel (TALN), dont une des applications est l'extraction automatique d'attributs textuels. Mesurer les attributs manuellement constitue un travail fastidieux, les résultats peuvent être inconsistants selon les évaluateurs (Guo et al., 2013). Plusieurs attributs textuels produits par le TALN sont raisonnablement hors de la portée d'une évaluation humaine sur un corpus de taille non triviale puisque leur mesure demanderait, par exemple, de produire l'arbre syntaxique de chaque phrase, ou de repérer chaque mot dans un lexique de référence. En employant des attributs ayant un plus haut niveau de complexité, les spécialistes souhaitent mieux estimer la difficulté du texte en tenant compte de l'aspect subjectif de la lecture (François, 2015). Ces attributs sont généralement empruntés aux théories psycholinguistiques. Par exemple, la version 3.0 du système *Coh-Metrix* extrait plus d'une centaine d'attributs linguistiques, dont plusieurs implémentent des modèles théoriques formulés dans les travaux de McNamara sur la cohésion linguistique (Graesser et al., 2011, 2014). Les technologies numériques ont également permis l'application de modèles de mesure et

de classification qui décrivent plus finement la relation entre les attributs et la difficulté des textes. Un cas notable est celui de la plateforme *Lexile*, qui est apparue à la fin des années 1980 et qui est encore largement utilisée aux États-Unis. *Lexile* utilise des attributs comparables à ceux de la lisibilité de Flesch pour estimer la difficulté des textes, mais remplace l'équation linéaire par un modèle de mesure plus complexe emprunté à la psychométrie, le modèle de Rasch (Wright et Linacre, 1994). Dans le sillage de *Lexile*, des travaux plus récents montrent une utilisation croissante de méthodes populaires d'apprentissage statistique, telles que la classification par régression multinomiale, ou par séparateurs à vaste marge (SVM).

En résumé, la modélisation de la difficulté des textes reprend aujourd'hui la méthodologie générale illustrée dans la Figure 3, en faisant appel à des attributs plus nombreux et plus complexes, ainsi qu'à des modèles de mesure ou de classification plus sophistiqués. Les deux premiers articles de la thèse font la démonstration de ce type de méthodologie. Le premier article se concentre sur l'extraction d'attributs linguistiques : nous décrivons un système de TALN créé pour les besoins de la thèse, et spécialisé dans l'analyse de textes québécois en français. Le deuxième article se concentre la classification des textes à partir des attributs discutés dans l'article 1, et décrit l'estimation du niveau de difficulté de textes répartis sur 11 années du système scolaire québécois.

1.2.2. Capturer les processus de lecture par l’oculométrie

Que la complexité du texte soit estimée par une formule simple ou un système d’apprentissage statistique, il s’agit là de la complexité découlant des attributs du texte. L’individu lisant le texte prend alors la forme idéalisée d'une masse anonyme de lecteurs ayant des compétences moyennes pour le niveau de compétence linguistique attendu (Boyer, 1992). L'aspect subjectif de la lecture n'est donc pas pris en considération autrement qu'à travers des

postulats théoriques et statistiques. Pour cette raison, le second volet de la présente thèse s'est intéressé à la compréhension du texte en tant que processus. L'oculométrie (ou suivi oculaire) est une méthode d'investigation particulièrement intéressante pour ce faire, car elle prend en compte la temporalité de la lecture et permet, au sens le plus littéral, de voir comment les participants lisent.

L'étude oculométrique de la lecture s'appuie sur des principes de physiologie humaine. Seule une petite région de la rétine, la fovéa, contient des cellules photoréceptrices suffisamment rapprochées pour capter des informations visuelles fines, dans notre cas du texte. Cette région correspond à environ deux degrés d'angle visuel dans le champ de vision d'un humain moyen (Winke, 2013), soit environ la largeur apparente du pouce lorsque le bras est tendu. Pour lire un mot, nous devons d'abord déplacer nos yeux afin que la lumière qui émane du mot frappe la fovéa. Ce mouvement rapide des yeux vers une nouvelle cible visuelle s'appelle une saccade visuelle. Le consensus actuel est que le cerveau ne traite pas les informations visuelles pendant les saccades (Rayner, 1998; Winke, 2013). Lire un mot nécessite donc d'effectuer une saccade vers le mot puis de stabiliser le regard sur le mot. La période pendant laquelle le regard reste sur un point relativement stationnaire (le point de fixation) est appelée fixation visuelle (Duchowski, 2007; Rayner, 1998). En analysant la succession des saccades et des fixations, on peut faire des inférences sur l'attention visuelle, c'est-à-dire sur les processus cognitifs qui conduisent l'individu à tourner son regard vers certaines informations plutôt que d'autres (McMains & Kastner, 2009).

La durée des fixations sur des mots, des phrases ou des passages de texte est un type de mesure fréquemment utilisé dans les études d'oculométrie de la lecture. Des études ont montré que la durée de fixation augmente à mesure que le texte devient plus difficile (Juhasz & Rayner, 2006; Kuhn & Stahl, 2003; Rayner & Well, 1996) ou que les participants ont des niveaux de

compétence de lecture plus faibles (Ashby et al., 2005). D'autres travaux ont montré que tous les mots ne sont pas fixés, et que la probabilité qu'un mot soit fixé, tout comme le temps de fixation, augmente avec la difficulté du mot (voir Rayner, 1998). Ces résultats soutiennent l'idée que la durée de fixation est influencée par la complexité du traitement de l'information, ce qui permet de déduire la difficulté perçue du texte en analysant la durée de fixation.

L'oculométrie peut également être utilisée pour étudier les processus de lecture à un niveau plus macroscopique. En mesurant le nombre de fois où un passage clé du texte a été visité, par exemple, on peut inférer si les participants ont vu les informations contenues dans ce passage. Dans la terminologie de l'oculométrie, une série de fixations sur la même région d'un stimulus visuel est désignée comme une *visite*. La visite commence lorsque le regard se pose sur la région d'intérêt, et dure tant que le regard demeure sur la région. Plusieurs études ont porté sur le nombre et la durée des visites de différentes régions d'un texte pour étudier, par exemple, les stratégies de recherche d'informations dans le texte (Şendur et Yildirim, 2015), ou les processus de réponse à des items de compréhension de texte (Bax et Chan, 2019). Une autre application fréquente consiste à comparer les schémas d'attention visuelle des novices et des experts (Gegenfurtner et al., 2011; Stofer et Che, 2014).

Un appareil d’oculométrie, ou oculomètre, prend typiquement la forme d’une caméra posée sur ou sous le moniteur qui affiche les stimuli visuels – on parle alors d’oculométrie stationnaire. Depuis quelques années, les oculomètres mobiles, qui se portent sur la tête à la manière de lunettes, gagnent en popularité. Bien que moins précis, les appareils mobiles permettent le mouvement des participants, alors que l’oculométrie stationnaire nécessite généralement que les participants demeurent immobiles. L’oculométrie mobile offre donc la possibilité de collecter les données dans un contexte plus naturel, ce qui favorise leur validité

écologique. Toutefois, cette liberté de mouvement complexifie l'analyse de données, ce qui pose un défi majeur. Les deux types d'appareils enregistrent la position du regard relativement au champ de vision. Cependant, en oculométrie mobile, le champ de vision bouge avec les mouvements du corps et de la tête; il n'est donc pas possible d'identifier directement ce que le participant regardait à un temps donné. C'est pourquoi les oculomètres mobiles sont équipés d'une caméra frontale tournée vers l'avant qui capture l'équivalent du champ de vision - un plan subjectif dans la terminologie cinématographique. L'équipe de recherche dont alors ensuite analyser, image par image, la vidéo enregistrée par cette caméra sur laquelle une cible indique l'emplacement du regard (voir Figure 4). Une heure d'enregistrement peut représenter des dizaines de milliers d'images, cette analyse manuelle est un travail fastidieux et prompt à l'erreur, ce qui impose une sérieuse limitation à l'oculométrie mobile appliquée aux processus de lecture.

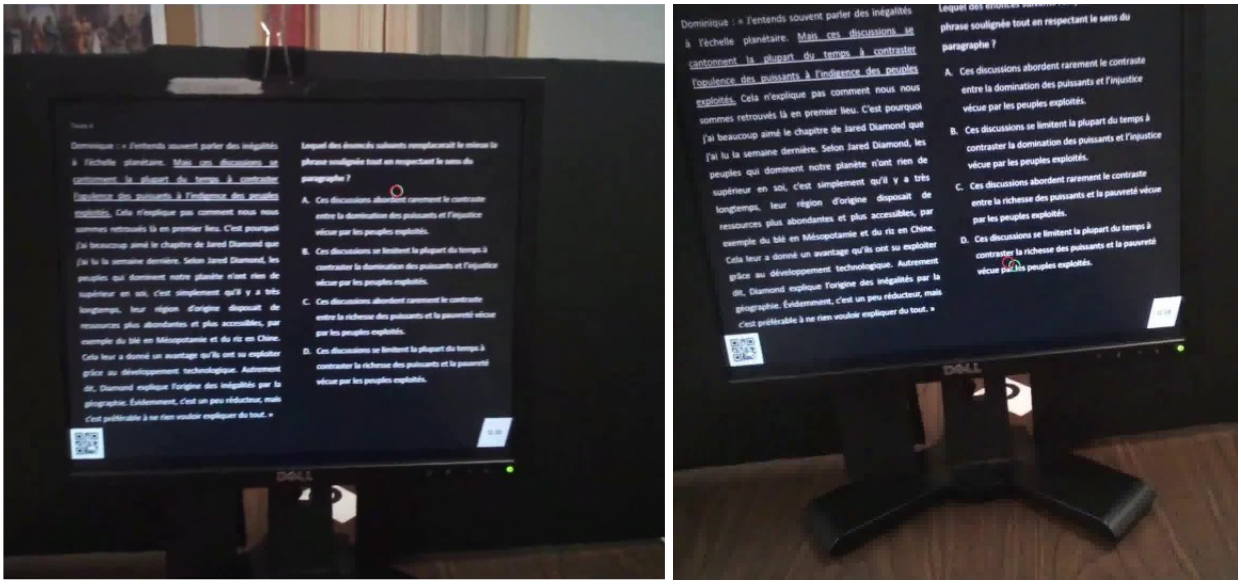


Figure 4. Arrêts sur image de la vidéo de caméra frontale d'un oculomètre mobile *Tobii Pro Glasses 2* – les cercles de couleur indiquent l'emplacement du regard. La position relative du stimulus visuel (un item de compréhension de texte) varie au fil de l'enregistrement. Les côtés droit et gauche des images ont été rognés pour mieux montrer les cercles colorés.

L'intelligence artificielle (IA) a connu plusieurs de ses succès dans l'analyse d'images, notamment le traitement automatisé de l'imagerie médicale (Robertson et al., 2018), et les enregistrements d'oculométrie (De Beugher, 2016). En combinant diverses techniques de vision par ordinateur, qui est une branche de l'IA consacrée au traitement des images, il est possible d'automatiser les opérations les plus fastidieuses et les plus sujettes aux erreurs dans le traitement des données oculométriques. Récemment, cette solution a été mise à la disposition de la communauté scientifique sous la forme de la bibliothèque *Mobile Gaze Mapping* (MGM) pour le langage Python (MacInnes, 2020). MGM combine plusieurs techniques de vision par ordinateur pour automatiser certaines étapes du traitement des données d'oculométrie mobile, facilitant ainsi leur analyse quantitative. Cette solution a pour l'instant été peu documentée et, ce qui est plus important pour cette thèse, n'a jusqu'à présent pas été testée rigoureusement sur des données issues d'une étude oculométrique de la lecture. Nous consacrons donc l'article 3 à la bibliothèque MGM, dont nous expliquons le fonctionnement et testons la robustesse à l'aide de données simulées. L'article 4 poursuit ce travail et relate une expérience pilote sur les stratégies de compréhension d'un texte argumentatif, et dont les données ont été analysées par une chaîne de traitement dont MGM était une composante. Nous résumons en plus de détails la composition de la thèse dans la section suivante.

1.3. Survol des quatre articles

Cette thèse est composée de deux grandes sections illustrant des principes de psychométrie computationnelle, conformément à l'objectif général de montrer comment une approche computationnelle peut faciliter l'intégration de la psychométrie et des sciences cognitives dans le domaine de la compréhension de texte. Chaque section est subdivisée en deux

articles, l'un décrivant un outil s'inscrivant dans une approche computationnelle, l'autre testant les performances de l'outil en situation réelle.

Le premier article décrit ALSI, un outil de TALN créé pendant le doctorat qui automatise l'analyse linguistique des textes en mesurant une variété d'attributs. ALSI est l'acronyme d'Analyseur Lexico-Syntaxique Intégré. Les attributs (*features*) extraits par ALSI sont pour la plupart issus de travaux en psycholinguistique et en linguistique computationnelle, notamment les travaux de McNamara sur la cohésion linguistique et les sources de difficulté des textes (McNamara et al., 2012; O'Reilly & McNamara, 2007). L'article passe d'abord en revue les outils actuels, leurs fonctionnalités, et leurs limites. Nous présentons ensuite une typologie simple des attributs de texte tels qu'extraits par ALSI, en établissant des ponts avec les théories psycholinguistiques sous-jacentes. Nous expliquons ensuite le fonctionnement d'ALSI et la manière dont les attributs sont produits. La suite de l'article décrit une expérimentation visant à tester la valeur des attributs produits par ALSI, et à sélectionner des attributs ayant un bon potentiel pour la classification du texte par niveau de difficulté. Nous avons utilisé ALSI pour analyser un corpus de textes scolaires en français québécois, assemblé pour cette thèse. Les résultats prennent la forme de mesures décrivant l'association statistique entre les attributs et la difficulté exprimée dans les années scolaires.

Le deuxième article aborde la question de la difficulté des textes du point de vue de l'estimation de la difficulté basée sur les attributs, et utilise le même corpus que le premier article. Après un aperçu de l'état de l'art dans la classification de la difficulté des textes, nous soulevons un défi actuel dans le domaine : rendre compte avec précision de la performance des modèles de classification de textes. Dans ce cas, la performance fait référence à la capacité du modèle à estimer le score associé au texte; diverses techniques sont utilisées à cette fin, dont

certaines peuvent donner une vision biaisée de la performance réelle du modèle. Dans l'article 2, nous présentons une procédure de sélection d'attributs textuels qui combine des éléments de théorie (présentés dans l'article 1) et des techniques d'analyse de données. Nous décrivons également les principes de mesure qui contribuent à une description plus précise de la performance des modèles de classification. Pour passer des attributs sélectionnés à un niveau de difficulté estimé, nous avons utilisé deux modèles de classification courants dans ce type d'étude, la régression logistique multinomiale et les séparateurs à marge large (aussi appelés machine à vecteurs de support, pour *support vector machines*). Les résultats indiquent que les deux modèles ont eu des performances généralement supérieures à l'état de l'art, ce qui corrobore la validité des attributs extraits par l'outil ALSI.

Le troisième article examine si l'étude oculométrique de la lecture peut être facilitée par des techniques de vision par ordinateur (*computer vision*, une sous discipline de l'intelligence artificielle). L'oculométrie mobile présente un avantage majeur en ce qu'il permet les mouvements, facilitant ainsi un contexte écologique lors de la collecte des données. Cependant, l'analyse des données d'oculométrie mobile présente des défis complexes, et les stratégies couramment utilisées pour relever ces défis ont des limites importantes, notamment pour l'application à l'étude de la lecture. L'article est composé d'un volet théorique et d'un volet expérimental. Le premier volet explique le fonctionnement de la bibliothèque Mobile Gaze Mapping (MGM) pour le langage Python (MacInnes, 2020; MacInnes et al., 2018), qui propose d'automatiser certaines étapes de l'analyse des données d'oculométrie mobile à l'aide de techniques de vision par ordinateur. Le deuxième volet teste cette solution sur des données simulées qui imitent des situations pouvant survenir lors de la collecte de données d'oculométrie mobile, par exemple lorsque les participants obstruent le capteur avec un geste de la main. Les

résultats indiquent que MGM peut effectivement stabiliser les enregistrements d'oculométrie mobile, et que des modifications aux paramètres internes pourraient en améliorer les performances.

Enfin, le quatrième article décrit une expérience pilote comparant les processus et stratégies de lecture de novices et d'experts dans une tâche de compréhension du texte argumentatif. Nous décrivons d'abord un cadre méthodologique pour l'étude de la lecture à l'aide de l'oculométrie mobile, allant de la collecte de données à la génération de résultats. Notre cadre méthodologique s'appuie sur les expérimentations présentées dans l'article 3 et intègre une version de MGM modifiée pour l'analyse robuste. Nous avons capturé, à l'aide de lunettes d'oculométrie, les mouvements oculaires d'un groupe de novices (9 étudiants de niveau postsecondaire) et d'un groupe d'experts (6 enseignants ou professionnels de l'éducation) répondant à des questions sur de courts textes argumentatifs. En utilisant des méthodes statistiques robustes, nous avons mis en évidence des différences marquées entre les processus de lecture des novices et des experts. Les résultats obtenus sont cohérents avec les travaux d'oculométrie étudiant la lecture (Rayner et al., 2006), ou comparant les patrons de mouvements oculaires de novices et d'experts (Ashby et al., 2005; Gegenfurtner et al., 2011). Cette étude confirme le potentiel de l'oculométrie mobile pour l'étude de la lecture tout en montrant l'intérêt des techniques de vision par ordinateur comme éléments de méthodologie.

1.4. Bibliographie

Ashby, J., Rayner, K. et Clifton, C. (2005). Eye movements of highly skilled and average readers: differential effects of frequency and predictability. *The Quarterly Journal of Experimental Psychology. A, Human Experimental Psychology*, *58*(6), 1065-1086. https://doi.org/10/fjj8nh

Bax, S. et Chan, S. (2019). Using eye-tracking research to inform language test validity and design. *System*, *83*. https://doi.org/10.1016/j.system.2019.01.007

Borsboom, D. (2006a). Can we bring about a velvet revolution in psychological measurement? A rejoinder to commentaries. *Psychometrika*, *71*(3), 463-467. https://doi.org/10.1007/s11336-006-1502-3

Borsboom, D. (2006b). The attack of the psychometricians. *Psychometrika*, *71*(3), 425-440. https://doi.org/10.1007/s11336-006-1447-6

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/

Daoust, F., Laroche, L. et 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

De Beugher, S. (2016). Computer vision techniques for automatic analysis of mobile eye-tracking data [thèse doctorale, Université Catholique de Louvain]. https://lirias.kuleuven.be/1668186

Dempster, E. R. et 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

Denman, D., Speyer, R., Munro, N., Pearce, W. M., Chen, Y.-W. et Cordier, R. (2017). Psychometric Properties of Language Assessments for Children Aged 4–12 Years: A Systematic Review. *Frontiers in Psychology*, *8*. https://doi.org/10/gbwbzd

Eisenberg, I. W., Bissett, P. G., Enkavi, A. Z., Li, J., MacKinnon, D. P., Marsch, L. A., & Poldrack, R. A. (2019). Uncovering the structure of self-regulation through data-driven ontology discovery. *Nature communications, 10*(1), 1-13.

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

François, T. (2015). When readability meets computational linguistics: a new paradigm in readability. *Revue francaise de linguistique appliquée*, *XX*(2), 79-97.

Gala, N., François, T., & Fairon, C. (2013). Towards a French lexicon with difficulty measures: NLP helping to bridge the gap between traditional dictionaries and specialized lexicons. Dans *Proceedings of eLex-Electronic Lexicography 2013 (p. 132-151)*. https://hal.archives-ouvertes.fr/hal-03194427/document

Gärdenfors, P. (1990). Induction, Conceptual Spaces and AI. *Philosophy of Science*, *57*(1), 78-95. https://doi.org/10/bp2q69

Gärdenfors, P. (2004). *Conceptual spaces: The geometry of thought*. MIT press.

Gegenfurtner, A., Lehtinen, E. et Säljö, R. (2011). Expertise Differences in the Comprehension of Visualizations: a Meta-Analysis of Eye-Tracking Research in Professional Domains. *Educational Psychology Review*, *23*(4), 523-552. https://doi.org/10/bpw4qb

Graesser, A. C., McNamara, D. S., Cai, Z., Conley, M., Li, H. et Pennebaker, J. (2014). Coh-Metrix Measures Text Characteristics at Multiple Levels of Language and Discourse. *The Elementary School Journal*, *115*(2), 210-229. https://doi.org/10/f6qk6f

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

Guo, L., Crossley, S. A., & McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples : A comparison study. *Assessing Writing*, 18(3), 218-238. https://doi.org/10/gcpgkq

Huff, K., & Goodman, D. P. (2007). The demand for cognitive diagnostic assessment. Dans J. Leighton, M. Gierl (Dir), *Cognitive Assessment for Education: Theory and Applications* (p. 19-60). Cambridge University Press. https://doi.org/10.1017/CBO9780511611186.002

Huff, K., Warner, Z. et Schweid, J. (2016). Large-scale standards-based assessments of educational achievement. Dans A. Rupp, Leighton, J. (Dir.), *The Wiley Handbook of Cognition and Assessment: Frameworks, Methodologies, and Applications* (p. 149-166). https://doi.org/10.1002/9781118956588.ch17

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Harvard University Press.

Juhasz, B. J. et Rayner, K. (2006). The role of age of acquisition and word frequency in reading: Evidence from eye fixation durations. *Visual Cognition*, *13*(7-8), 846-863. https://doi.org/10/dznsng

Kane, M. T. (1992). An argument-based approach to validity. *Psychological bulletin*, *112*(3), 527-535. https://doi.org/10.1037/0033-2909.112.3.527

Kane, M. T. (2006a). Content-related validity evidence in test development. Dans S. M. Downing & T. M. Haladyna (Dir.), Handbook of test development (pp. 131–153). Lawrence Erlbaum Associates Publishers.

Kane, M. T. (2006b). Validation. *Educational measurement*, 4(2), 17-64.

Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*, *50*(1), 1-73. https://doi.org/10.1111/jedm.12000

Kane, M. T. et Bejar, I. I. (2014). Cognitive frameworks for assessment, teaching, and learning: A validity perspective. *Psicología Educativa*, *20*(2), 117-123. https://doi.org/10.1016/j.pse.2014.11.006

Kline, R. B. (2019). *Becoming a Behavioral Science Researcher, Second Edition: A Guide to Producing Research That Matters* (Second edition). The Guilford Press.

Kuhn, M. R. et Stahl, S. A. (2003). Fluency: A review of developmental and remedial practices. *Journal of Educational Psychology*, *95*(1), 3. https://doi.org/10/dsck55

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

Lawrence, I. M. et Shea, E. C. (2008). *Improving Assessment: The Intersection of Psychology and Psychometrics*. Educational Testing Services.

Leighton, J. P. et Gierl, M. J. (2011). *The Learning Sciences in Educational Assessment: The Role of Cognitive Models*. Cambridge University Press.

Lombardino, L., Lieberman, J., & Brown, J. (2005). *Assessment of Literacy and Language (ALL).* The Psychological Corporation.

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

MacInnes, J. J. (2020). *Mobile Gaze Mapping* [Python]. https://github.com/jeffmacinnes/mobileGazeMapping

MacInnes, J. J., Iqbal, S., Pearson, J., & Johnson, E. N. (2018). Wearable Eye-tracking for Research: Automated dynamic gaze mapping and accuracy/precision comparisons across devices. *bioRxiv*. https://doi.org/10.1101/299925

MacLeod, C. M. (2005). The Stroop task in cognitive research. Dans A. Wenzel & D. C. Rubin (Dir.), *Cognitive Methods and their Application to Clinical Research* (p. 17–40). American Psychological Association. https://doi.org/10.1037/10870-002

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

McMains, S. A. et Kastner, S. (2009). Visual attention. Dans M. D. Binder, N. Hirokawa et U. Windhorst (Dir.), *Encyclopedia of Neuroscience* (p. 4296-4302). Springer. https://doi.org/10.1007/978-3-540-29678-2_6344

McNamara, D. S., Graesser, A. C. et 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).

McNamara, D. S. et Kendeou, P. (2011). Translating advances in reading comprehension research to educational practice. *Electronic Journal of Elementary Education, 4*(1), 33-46.

Mislevy, R. J. (1982). Foundations of a new test theory. *ETS Research Report Series*, 1982(2), i-32. https://doi.org/10.1002/j.2333-8504.1982.tb01336.x

Mislevy, R. J. (2008). How cognitive science challenges the educational measurement tradition. *Measurement: Interdisciplinary Research and Perspectives*, *6*(1-2), 124. https://doi.org/10.1080/15366360802131635

Mislevy, R. J., Almond, R. G. et Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Research Report Series*, *2003*(1). https://doi.org/10.1002/j.2333-8504.2003.tb01908.x

O'Reilly, T. et 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

Pellegrino, J. W., Chudowsky, N. et Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. National Academies Press.

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

Polyak, S. T., von Davier, A. A. et Peterschmidt, K. (2017). Computational psychometrics for the measurement of collaborative problem solving skills. *Frontiers in psychology*, *8*, 2029. https://doi.org/10/gcnjd5

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

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372. https://doi.org/10/b5gdv6

Rayner, K., Chace, K. H., Slattery, T. J. et Ashby, J. (2006). Eye movements as reflections of comprehension processes in reading. *Scientific Studies of Reading*, *10*(3), 241-255. https://doi.org/10.1207/s1532799xssr1003_3

Rayner, K. et Well, A. D. (1996). Effects of contextual constraint on eye movements in reading: A further examination. *Psychonomic Bulletin & Review*, *3*(4), 504-509. https://doi.org/10/fcv33j

Rhemtulla, M., van Bork, R., & Borsboom, D. (2020). Worse than measurement error: Consequences of inappropriate latent variable measurement models. *Psychological Methods*, *25*(1), 30. https://doi.org/10.1037/met0000220

Robertson, S., Azizpour, H., Smith, K. et Hartman, J. (2018). Digital image analysis in breast pathology—from image processing techniques to artificial intelligence. *Translational Research*, *194*, 19-35. https://doi.org/10/gc97vx

Şendurur, E. et Yildirim, Z. (2015). Students' web search strategies with different task types: an eye-tracking study. *International Journal of Human-Computer Interaction*, *31*(2), 101-111. https://doi.org/10/gjnswt

Sijtsma, K. (2006). Psychometrics in psychological research: Role model or partner in science? *Psychometrika*, *71*(3), 451. https://doi.org/10/b6v9dp

Solnyshkina, M., Zamaletdinov, R., Gorodetskaya, L. et Gabitov, A. (2017). Evaluating text complexity and Flesch-Kincaid Grade Level. *Journal of Social Studies Education Research*, *8*(3), 238-248. https://www.bulenttarman.com/index.php/jsser/article/view/225

Stofer, K., & Che, X. (2014). Comparing experts and novices on scaffolded data visualizations using eye-tracking. *Journal of Eye Movement Research*, 7(5). https://doi.org/10.16910/jemr.7.5.2

Sweller, J., & Chandler, P. (1991). Evidence for cognitive load theory. Cognition and instruction, 8(4), 351-362. https://doi.org/10.1207/s1532690xci0804_5

Truijens, F. L., Cornelis, S., Desmet, M., De Smet, M. M. et Meganck, R. (2019). Validity beyond measurement: Why psychometric validity is insufficient for valid psychotherapy research. *Frontiers in psychology*, *10*. https://doi.org/10/gjq5g8

van der Maas, H. L., Molenaar, D., Maris, G., Kievit, R. A. et Borsboom, D. (2011). Cognitive psychology meets psychometric theory: on the relation between process models for decision making and latent variable models for individual differences. *Psychological review*, *118*(2), 339. https://doi.org/10.1037/a0022749

Van Moere, A. (2012). A psycholinguistic approach to oral language assessment. *Language Testing*, *29*(3), 325-344. https://doi.org/10/gjrbrm

Visone, J. D. (2009). The validity of standardized testing in science. *American Secondary Education*, *38*(1), 46-61. https://www.jstor.org/stable/41406066

von Davier, A. A. (2017). Computational psychometrics in support of collaborative educational assessments. *Journal of Educational Measurement*, *54*(1), 3-11. https://doi.org/10/gcphmd

von Davier, A. A., Deonovic, B., Yudelson, M., Polyak, S. T. et Woo, A. (2019). Computational psychometrics approach to holistic learning and assessment systems. *Frontiers in Education*, *4*. https://doi.org/10/ggjp3q

Winke, P. M. (2013). Eye-Tracking technology for reading. Dans A. J. Kunnan (Dir.), *The Companion to Language Assessment* (p. 1029-1046). John Wiley & Sons. https://doi.org/10.1002/9781118411360.wbcla030

Wright, B. D. et Linacre, J. M. (1994). The Rasch model as a foundation for the Lexile Framework [article non-publié].

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

Zieky, M. J. (2014). An introduction to the use of evidence-centered design in test development. *Psicología Educativa*, *20*(2), 79-87. https://doi.org/10.1016/j.pse.2014.11.003

Zumbo, B. (2020). *On computational psychometrics as a validity framework for process data* [présentation de conférence]. The Annual Meeting of the National Council for Measurement in Education (NCME), San Francisco, États-Unis. https://www.youtube.com/watch?v=dfN26b65adw

2. Premier article :

Un nouvel outil d'analyse automatique de la complexité linguistique pour le français québécois

Guillaume Loignon

Université de Montréal

Résumé

La mesure de la complexité des textes est importante en l'éducation pour soutenir l'apprentissage, et favoriser l'évaluation dans une perspective d'équité. Les méthodes et les outils actuels d’estimation de la complexité linguistique sont souvent rudimentaires et reposent sur une conception naïve des sources de difficulté des textes. Des outils automatisés plus conformes aux théories psycholinguistiques ont été développés au cours des dernières années, mais ne sont pas adaptés au contexte québécois. Dans cet article, nous présentons un nouvel outil d'analyse de texte appelé ALSI, pour Analyseur Lexico-Syntaxique Intégré. Conçu pour le contexte scolaire québécois francophone, ALSI extrait automatiquement 48 attributs linguistiques. Après un aperçu des outils disponibles, nous présentons une typologie d'attributs caractérisant la difficulté du texte, puis nous expliquons comment ALSI extrait ces attributs. Nous faisons la démonstration d’ALSI à travers une expérience portant sur un corpus de 600 textes décrivant les attributs les plus associés statistiquement avec l’année scolaire du texte. Les résultats montrent le fort potentiel d'ALSI pour modéliser la complexité des textes en français québécois.

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

Abstract

Despite the importance of using appropriate text difficulty to promote learning, the methods and tools used to do so are often simplistic and do not capture important dimensions of text complexity. Tools more consistent with psycholinguistic theories have been developed in recent years, but they are not adapted to the Quebec context. In this article, we introduce a new text analysis tool called ILSA, for Integrated Lexico-Syntactic Analyzer. Designed for French-speaking Quebec school contexts, ILSA automatically extracts 48 text attributes. After an overview of the available tools, we present a typology of attributes characterizing text difficulty, then explain how ILSA extracts these attributes. We demonstrate ILSA through an experiment on a corpus of 600 texts that focuses on selecting sets of attributes that best characterize text difficulty. Results describe the statistical association between attributes and the text's school year, and show the strong potential of ILSA to characterize text difficulty in Quebec French.

Keywords: text features, corpus analysis, automatic natural language processing, readability, French

Premier article :

Un nouvel outil d'analyse automatique de la complexité linguistique pour le français québécois

ALSI, pour Analyseur Lexico-Syntaxique Intégré, est un outil automatisé de traitement du langage naturel qui extrait une variété d'attributs caractérisant la complexité intrinsèque du texte. Nous avons créé ALSI pour répondre à des besoins dans le domaine de la mesure et de l'évaluation en éducation. Il n'existait pas d'outil pour l'extraction des attributs du texte en français québécois qui soit à jour, compte tenu des innovations des dernières années en psycholinguistique et en linguistique computationnelle. Le seul outil disponible actuellement, SATO-Calibrage (Daoust et al., 1996), date des années 1990 et n'a pu profiter des innovations théoriques et méthodologiques concernant les sources de difficulté du texte et leur mesure automatisée. ALSI a été créé pour le contexte scolaire québécois et extrait automatiquement une cinquantaine d'attributs ancrés dans la théorie psycholinguistique. Le présent article a deux objectifs : d'abord introduire l'outil ALSI, le contexte théorique dans lequel il s'inscrit et ses fonctionnalités; ensuite, faire la démonstration de l'outil tout en déterminant quels attributs mesurés par ALSI ont le meilleur potentiel pour estimer la complexité linguistique.

2.1.1. Contexte théorique de la complexité linguistique

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 au texte. La complexité intrinsèque au texte est déterminée par ses caractéristiques mesurables, appelées *attributs*; 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. Boyer (1992) propose ainsi de concevoir la complexité du

texte entre trois pôles : le texte, le lecteur et la situation. De manière similaire, Zakaluk et Samuels (1988) parlent de facteurs « en dehors 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 et de l'athlète. Modéliser la complexité du texte représente un défi important puisqu'il faut, en s'appuyant sur des mesures faites à partir du texte, estimer la difficulté *perçue* par le lecteur, ce qui implique d'émettre des hypothèses quant à ce qui serait susceptible d'augmenter la charge cognitive.

Les comptes-rendus historiques portant sur la modélisation de la complexité linguistique s'articulent assez unanimement autour de trois thèmes communs que nous résumons comme suit: la critique du recours exclusif aux attributs « de surface », la nécessité d'introduire des attributs résultant du caractère subjectif de la lecture, et la recommandation de recourir à la psycholinguistique pour y arriver (Boyer, 1992; François, 2015; Kintsch et Vipond, 2014; McNamara, 2012; Zakaluk et Samuels, 1988). La complexité du texte a 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 furent adaptées pour la langue française, d'autres furent créées spécifiquement pour le français (Mesnager, 1989). L'usage intensif des attributs de surface a été largement critiqué, principalement car ceux-ci ne capturent pas certains éléments de complexité découlant du caractère subjectif de la lecture. Le développement de l'informatique n'a pas mené immédiatement une caractérisation plus sophistiquée de la complexité linguistique : des outils populaires, tels ATOS et Lexile, emploient toujours des attributs de surface (Milone, 2014; Smith et al., 1989).

Complexité linguistique en mesure et évaluation de l'éducation

L'analyse de la complexité linguistique a de nombreuses applications dans le domaine de l'éducation, notamment pour la sélection de textes et manuels favorisant l'apprentissage en fonction des caractéristiques des élèves (Graesser et al., 2004). La complexité linguistique est de plus un aspect peu abordé, mais important de la démarche de conception des tests (Lane et al., 2015; Visone, 2009; McNamara et al., 2012). Contrôler les attributs linguistiques de l'item permet premièrement de mitiger la variance indésirable due 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*, les attributs linguistiques des items sont une des sources potentielles de variance indésirable qu'il faut contrôler lorsque possible (Joint Committee on Standards for Educational and Psychological Testing, 2014; Lane et al., 2015). L'influence des attributs linguistiques 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 et Reddy, 2007) et américain (Martiniello, 2009) ont révélé la présence de biais linguistiques dans des tests standardisés de mathématiques.

Les aspects linguistiques de l'item ne sont pas uniquement une source de variance indésirable. Leur influence sur les processus de réponse à l'item peut être *désirable* lorsque la langue fait partie, ou ne peut être séparée, de la compétence évaluée (Avenia-Tapper et Llosa, 2015). Par exemple, des études ont analysé les attributs linguistiques du *Test of English as a Foreign Language* (TOEFL) dans le but de vérifier la validité des construits, d'inférer les stratégies des répondants, ou pour déterminer quelles caractéristiques linguistiques prédisaient le mieux le score (Kyle et al., 2016). Dans une approche d'évaluation diagnostique, quelques études ont analysé les caractéristiques linguistiques d'items dans l'optique de fournir à l'élève

une rétroaction plus détaillée quant à ses forces et défis d'apprentissage (Alavi et Ranjbaran, 2018; Buck et al., 1997; Buck et Tatsuoka, 1998; Chen et Chen, 2016).

Les outils d'estimation de la complexité linguistique

Les comptes-rendus historiques portant sur la modélisation de la complexité linguistique s'articulent assez unanimement autour de trois thèmes communs que nous résumons comme suit: la critique du recours exclusif aux attributs « de surface », la nécessité d'introduire des attributs résultant du caractère subjectif de la lecture, et la recommandation de recourir à la psycholinguistique pour y arriver (Boyer, 1992; François, 2015; Kintsch et Vipond, 2014; McNamara et al., 2012; Zakaluk et Samuels, 1988). 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 furent adaptées pour la langue française, d'autres furent créées spécifiquement pour le français (Mesnager, 1989). L'usage intensif des attributs de surface a été largement critiqué, principalement car ceux-ci ne capturent pas certains éléments de complexité découlant du caractère subjectif de la lecture, que Boyer (1992) nomme la *compréhensibilité* du texte.

S'il est reconnu depuis au moins les années 1980 que la modélisation du texte devrait inclure des attributs plus complexes et en phase avec les théories psycholinguistiques, les premiers outils de TALN analysant la complexité linguistique, principalement *Lexile*, ont plutôt consacré l'usage d'attributs simples (François, 2015). Pour l'analyse du texte en langue anglaise, *Lexile* est généralement considérée comme la première plateforme d'estimation de la complexité linguistique à être disponible au public. Conçu initialement pour estimer la complexité d'items de compréhension de textes (Smith et al., 1989), *Lexile* fonctionne en appliquant le modèle de

Rasch à des attributs calculés par segments de 125 mots (Wright et Linacre, 1994). Les attributs exacts derrière *Lexile* demeurent un secret de commerce, mais sont vraisemblablement dérivés de la longueur des phrases et de la fréquence d'occurrence des mots dans un lexique de référence.

ATOS est un système similaire à *Lexile* et qui estime le niveau de difficulté du texte en s'appuyant sur deux attributs simples (longueur de la phrase et du mot) de même que sur le niveau scolaire moyen des mots, lequel est obtenu par un lexique de référence contenant plus de 100 000 mots et indiquant l'âge auquel l'élève est réputé avoir acquis chaque lexème (Milone, 2014).

Coh-Metrix a été conçu dans le but d'aller au-delà des statistiques « de surface » en produisant des indices de complexité basés sur les données de la recherche en psycholinguistique. L'outil produit un total de 108 attributs. Plusieurs des attributs de *Coh-Metrix* concernent la cohésion du texte, qui est la présence de relations et de similitudes entre les phrases du texte (Dowell et al., 2015; McNamara et Graesser, 2011 Graesser et al., 2004).

D-Level Analyzer s'appuie sur des travaux sur le développement de la lecture chez l'enfant (Lu, 2009) et produit des attributs au niveau de la phrase, avec un accent sur les attributs syntaxiques.

Parmi les outils conçus pour le texte français, on trouve SATO-Calibrage (pour Système d'Analyse de Textes par Ordinateur) qui a été créé au Québec dans les années 1990 (Daoust et al., 1996). SATO-Calibrage est un outil qui se situe entre deux générations, soit les outils basés sur quelques attributs simples, dont le *Lexile* est le cas paradigmatique, et les outils de nouvelle génération employant des attributs plus nombreux et complexes. L'objectif initial de SATO-Calibrage était de permettre aux auteurs et aux enseignants d'obtenir des informations sur le texte afin d'adapter leur matériel, ou d'orienter leur enseignement en tirant parti des difficultés

inhérentes au texte (F. Daoust, communication personnelle, janvier 2021). Toujours disponible sur le Web, l'outil mesure 14 attributs linguistiques et estime le niveau scolaire du texte en appliquant une équation de régression linéaire.

DMesure et *AMesure* sont deux outils d'analyse s'appuyant sur les travaux de linguistique computationnelle de leur créateur Thomas François (François, 2009; François et Fairon, 2012; François et Miltsakaki, 2012). *DMesure* classifie des textes en français langue seconde selon les six niveaux du Cadre européen commun de référence. *AMesure* se spécialise plutôt dans l'estimation de la lisibilité de documents en français des affaires. Les deux outils emploient des attributs inspirés d'efforts précédents comme SATO-Calibrage, ou tirés de la littérature pertinente en linguistique computationnelle.

Enfin, *ReaderBench* a été conçu dans une approche similaire à *DMesure* pour analyser du texte en plusieurs langues, dont le français. *ReaderBench* produit un grand nombre d'attributs complexes, incluant des attributs portant sur la structure argumentative et les inférences (Dascalu et al., 2013).

2.1.2. Pourquoi créer un nouvel outil ?

Nous avons choisi de créer un nouvel outil afin de répondre à des désidératas qui n'étaient pas couverts par les plateformes d'analyse disponibles actuellement, soit : (1) pouvoir analyser le français québécois comme langue d'enseignement; (2) extraire des attributs ancrés dans des principes de psycholinguistique afin de modéliser la complexité linguistique mieux que ne l'auraient fait les attributs de surface employés seuls.

Au meilleur de nos connaissances, SATO-Calibrage est la seule plateforme répondant pleinement au premier critère puisque ses coefficients de régression ont été ajustés à un corpus de textes réparti entre les 11 niveaux du système scolaire québécois. Les attributs extraits par

SATO-Calibrage sont cependant assez simples comparativement à l'état de l'art, par exemple l'outil n'analyse pas directement la structure syntaxique. Nous nous questionnons également quant à la présence simultanée, dans le modèle prédictif, d'indices risquant d'être fortement corrélés, tels le nombre de phrases et le nombre de points, et le nombre de mots inconnus et le nombre de mots longs.

Les outils *DMesure* et *ReaderBench* produisent des attributs complexes en phase avec la linguistique computationnelle, mais ne sont pas adaptés au cursus scolaire québécois francophone. De plus, au moment de rédiger cet article, les deux plateformes n'étaient plus disponibles; nous n'avons donc pu les tester sur un corpus québécois. *AMesure* est disponible en ligne, mais son orientation vers le français des affaires la rendait moins intéressante considérant nos besoins.

Une alternative plausible aurait été de produire un cadre de recommandations pour l'analyse manuelle du texte, une méthode encore employée (voir par exemple Martiniello, 2009). La mesure par l'humain d'attributs du texte peut toutefois être fastidieuse, et les notes rapportées par différents évaluateurs peuvent varier en fonction des attributs pris en compte, de l'approche utilisée et du niveau d'expérience (Guo et al., 2013; pour un survol des problèmes d'accord interjuge dans le domaine de l'analyse textuelle voir aussi Lim, 2019). Dans l'ensemble, les outils de TALN que nous avons décrits possédaient tous certaines des fonctionnalités requises. Certains auraient pu être adaptés à nos besoins et au français québécois, mais cela aurait demandé un travail considérable, nous motivant à produire notre propre plateforme.

2.1.3. La présente étude

L'objectif général de la présente étude est d'introduire et faire la démonstration de l'outil ALSI. Nous décrivons d'abord une typologie simplifiée des attributs du texte produits par ALSI,

qui se divisent en deux grandes familles soit les attributs lexicaux (portant sur les mots) et syntaxiques (portant sur la structure de la phrase). Nous détaillons ensuite le fonctionnement d'ALSI en présentant ses lexiques de référence et les procédures utilisées pour extraire les attributs du texte. La dernière section de l'article relate une expérimentation visant à examiner le potentiel des attributs produits par ALSI pour modéliser la complexité intrinsèque de textes en français québécois. Ce faisant, nous proposons une procédure de sélection des attributs qui s'appuie à la fois sur l'analyse des données et la typologie introduite dans cet article. Considérant un cas d'utilisation où ALSI servirait à estimer le niveau de difficulté des textes, un ensemble d'attributs trop nombreux ou mal assortis pourrait causer ou amplifier les problèmes de surajustement des modèles, gonflant ainsi les performances apparentes tout en diminuant la généralisabilité des résultats. Le choix des attributs devrait donc résulter d'une analyse rigoureuse des données, guidée par une théorie substantielle (Judd et al., 2011).

2.2. Une typologie simplifiée des attributs du texte

Les attributs extraits par ALSI s'inscrivent dans une typologie simple composée de deux dimensions (lexique, syntaxe), subdivisées en trois strates définies par le niveau d'abstraction (simple, intermédiaire, globale). Le rôle de cette typologie est de regrouper les attributs en catégories cohérentes reposant sur des caractéristiques similaires du texte, tout en exprimant une vision nuancée de la complexité du texte. La typologie peut aussi contribuer à la sélection de variables de types différents pour un modèle d'estimation du niveau de difficulté du texte. La première dimension de la typologie est le lexique, qui est la liste des lexèmes (mots distincts) utilisés dans un texte. Les attributs lexicaux caractérisent donc la complexité associée aux mots du texte. La seconde dimension est la syntaxe, c'est-à-dire la structure grammaticale qui relie les lexèmes. Les attributs syntaxiques caractérisent la complexité émergeant de l'agencement des

mots en phrases, et du rôle que jouent les mots dans la phrase. Le choix des dimensions 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 et Tunmer, 1986) tout en étant sous-entendue par le choix d'attributs des plateformes d'analyse *ATOS* et *Lexile*.

Les deux dimensions se subdivisent en trois strates. La première strate est composée d'attributs obtenus par des statistiques simples qui correspondent grosso modo aux attributs de surface. La strate intermédiaire regroupe des attributs dont l'extraction demande le recours à des bases de données lexicales ou une procédure automatisée d'analyse syntaxique. La troisième strate correspond aux mesures qualifiant la complexité linguistique de manière plus globale (par ex. : mesures de cohésion). Le Tableau 1 montre des attributs ou groupes d'attributs représentant les six types, ces attributs sont décrits dans la suite de cette section.

Tableau 1

Typologie simplifiée de la complexité linguistique telle qu'employée par ALSI

	Lexique	**Syntaxe**
Strate 1 *Simple*	Mesures de longueur (orthographique ou syllabique) portant sur le lexème ou le lemme.	Longueur de la phrase, nombre de virgules.
Strate 2 *Intermédiaire*	Occurrence du lexème ou du lemme dans un lexique de référence; âge de première apparition dans le cursus scolaire.	Présence de certains constituants de la phrase (ex.: verbes conjugués); présence et longueur de syntagmes d'intérêt (ex.: subordonnée relative), niveau de hiérarchisation de la phrase.
Strate 3 *Globale*	Densité lexicale; cohésion lexicale.	Cohésion syntaxique.

2.2.1. Attributs lexicaux

Nous définissons la complexité lexicale comme la propension des mots du texte à produire une charge cognitive. Selon le modèle de Coltheart (2005), la lecture procède par deux

voies selon si le mot est connu (présent dans le lexique mental, ou vocabulaire de l'élève) ou inconnu. Les mots connus ne sont pas décodés, mais reconnus automatiquement, entrainant une charge cognitive minimale. Des travaux d'oculométrie ont en effet montré que plusieurs mots ne sont pas fixés lors de la lecture, et que la probabilité de fixer un mot est inversement proportionnelle à sa difficulté estimée (Juhasz et Rayner, 2006; Rayner et Duffy, 1986). Les mots inconnus sont décodés par un autre processus[1] : la conversion des graphèmes (parties du mot) en phonèmes (sons), ce qui entraine une charge cognitive plus élevée (Goigoux, 2003). En somme, les attributs lexicaux peuvent servir à estimer charge cognitive à partir de la probabilité du mot de figurer dans le lexique mental de l'élève, et de la difficulté de décodage dans les cas où le mot est inconnu.

La première strate de complexité lexicale regroupe les attributs portant sur les caractéristiques « de surface » du lexème, notamment le nombre de caractères ou de syllabes. Ces attributs peuvent également être mesurés à partir du lemme, qui est la forme canonique du lexème. Bien que l'usage de ce type de mesures ait été critiqué (Kim et al., 2007) nous avons jugé pertinent de les inclure dans ALSI pour des fins de recherche, mais aussi, parce qu'il demeure possible que la longueur du mot soit un intermédiaire intéressant pour estimer sa rareté et sa difficulté de décodage.

La deuxième strate de complexité lexicale regroupe les attributs produits par croisement du lexique du texte (liste des mots distincts du texte) avec un ou plusieurs lexiques de référence. Les lexiques de référence sont formés à partir de corpus de textes jugés représentatifs d'un

[1] Nous présentons ici une vue simplifiée de l'identification des mots. Comme le remarque une évaluatrice de la thèse, d'autres processus cognitifs, tel le traitement de la morphologie, influencent le processus d'identification.

domaine, par exemple un ensemble de manuels scolaires. Depuis les premières listes orthographiques telles *Le Français Fondamental* de Gougenheim (Gala et al., 2013), des lexiques de référence plus complets ont vu le jour, nommément Manulex (Lété, 2004) et ÉQOL (Stanké et al., 2019), que nous décrivons plus loin. Les lexiques peuvent être vus comme des bases de données associant des lexèmes à des statistiques estimant leur complexité, par exemple la fréquence d'occurrence du lexème, qui est son nombre d'apparitions dans le corpus ayant servi à bâtir le lexique. La fréquence d'occurrence est typiquement rapportée en nombre d'occurrences par million de mots ou en indice de fréquence standardisé (IFS), calculé selon la formule suivante (Lété, 2004) où *U* est la fréquence par million de mots:

$$IFS = 10 \times (log_{10}U + 4)$$

À partir des fréquences d'occurrence par niveau scolaire, on peut inférer d'autres attributs, par exemple l'année où le lexème apparait pour la première fois dans cursus (*age of first exposure*), attribut utilisé notamment dans *ATOS* (Milone, 2014). Conformément avec le modèle de Coltheart (2005), l'année scolaire ou l'âge à laquelle l'élève est réputé avoir vu le lexème nous offre une information complémentaire pour estimer la probabilité que le lexème soit inconnu, et indirectement sa propension à produire une charge cognitive plus élevée.

La troisième strate de complexité lexicale porte sur les attributs offrant une vision plus globale du lexique du texte. Telle qu'opérationnalisée dans ALSI, la cohésion lexicale est la tendance des phrases à reprendre de l'information de phrases adjacentes. Une plus grande cohésion entre les phrases signifie que les entités mentionnées dans une phrase ont plus de chance d'être à nouveau abordées dans la phrase suivante (Graesser et al., 2004). Ces reprises d'information, que Kintsch et Van Dijk (1978) nomment relations coréférentielles, aident à construire le sens du texte, et sont donc associées à une plus grande facilité en lecture. Une

diversité faible indique donc que les mots ont tendance à être réutilisés plusieurs fois, ce qui devrait en général diminuer la charge cognitive (Graesser et al., 2004, p. 198). Le domaine de la linguistique computationnelle propose plusieurs formules estimant la densité lexicale (voir Fergadiotis et al., 2015). Par exemple, le rapport type-jeton (*type-token ratio*) est le rapport entre la taille du lexique du texte (nombre de mots uniques) et le nombre total de mots. L'indice de Maas (1972) est une autre formule de diversité lexicale, plus robuste aux variations dans la longueur du texte. Soit *T* le nombre de mots et *U* le nombre de mots uniques d'un texte, l'indice de Maas se calculera comme suit :

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

Un rapport type-jeton ou un indice de Maas relativement plus élevé indique une plus grande diversité lexicale.

2.2.2. Attributs syntaxiques

La première strate de complexité syntaxique regroupe les attributs obtenus à partir de statistiques simples[2] portant sur la longueur des phrases, des paragraphes ou du texte, de même que sur le nombre de phrases et de paragraphes. La longueur moyenne des phrases du texte est possiblement le plus ancien estimateur de complexité syntaxique, et est encore fréquemment utilisé (Szmrecsányi, 2004). Par exemple, les variables liées à la longueur de la phrase étaient, dans le modèle de François et Fairon (2012), les meilleurs attributs syntaxiques pour estimer la complexité de textes en français. La pertinence de la longueur de la phrase pour estimer la complexité linguistique perdure, même dans un contexte où des attributs syntaxiques plus

[2] Les attributs sont classés par type et décrits au Tableau 7, situé en annexe.

sophistiqués sont disponibles. Bien qu'une phrase longue ne soit pas nécessairement complexe syntaxiquement, il demeure possible que la longueur de la phrase soit un intermédiaire (proxy) efficace pour un ensemble d'attributs syntaxiques (Szmrecsányi, 2004). Par ailleurs, lire une phrase longue demande de conserver plus de mots en mémoire de travail, ce qui augmente la charge cognitive et donc la difficulté (Graesser et al., 2004).

La deuxième strate de complexité syntaxique porte premièrement sur les attributs qui requièrent l'identification du rôle grammatical des mots. C'est le cas, par exemple, pour le nombre de verbes conjugués, ou pour la présence de verbes utilisant des conjugaisons qui représentent souvent des défis pour les élèves, tel le subjonctif présent (Daoust et al., 1996; François et Fairon, 2012). Ce niveau contient également les attributs portant sur des syntagmes d'intérêt, comme les clauses subordonnées ou les groupes nominaux complexes (Lu, 2010; Wu et al., 2020). Le caractère argumentatif du texte est évalué à partir d'une liste de connecteurs argumentatifs (*en conclusion*, *donc*, *par ailleurs*, etc.) en nous basant sur le postulat selon lequel ces connecteurs indiquent une structure logique plus complexe, et donc un texte plus difficile (Graesser et al., 2004). La profondeur de l'arbre syntaxique, comptée en nombre de nœuds (*nodes*) est également un indicateur de complexité syntaxique fréquemment employé (Sherstinova et al., 2020; Szmrecsányi, 2004). Soit une phrase représentée comme un graphe hiérarchique, la profondeur de la structure syntaxique (Blache, 2010) correspond au plus long chemin reliant un mot à la racine de la phrase.

La troisième strate porte sur les attributs syntaxiques ayant une portée plus globale. Dans cette première version d'ALSI, le seul attribut de ce type est la cohésion syntaxique, que nous définissons comme la similitude des structures grammaticale entre les phrases (Crossley et al.,

2016). Nous décrivons dans ce qui suit l'outil ALSI, en montrant comment les attributs de notre typologie sont extraits.

2.3. L'outil ALSI

ALSI est un outil de TALN spécialisé dans l'extraction d'attributs caractérisant la complexité linguistique. L'adaptation d'ALSI au français québécois et au domaine de l'éducation vient principalement de deux choix de conception : ALSI a recours à des bases de données lexicales issues du système scolaire québécois, et privilégie l'extraction d'attributs syntaxiques ayant un équivalent dans le cursus éducatif (par ex.: les groupes nominaux avec complément). Construit en langage R (Ihaka et Gentleman, 1996), ALSI est constitué de modules fonctionnant de concert pour produire une variété d'attributs linguistiques s'inscrivant dans une typologie simple.

2.3.1. Lexiques de référence

Dans sa première version, ALSI s'appuie sur trois lexiques de référence: Manulex (Lété, 2004), ÉQOL (Stanké et al., 2019) et la liste orthographique de 2013 du ministère de l'Éducation du Québec (par la suite, LOMEQ2013). Ces lexiques ont été choisis puisqu'ils contenaient des informations quant au niveau scolaire des mots et, dans le cas d'ÉQOL et de LOMEQ2013, car ils étaient adaptés au système d'éducation du Québec.

Manulex

Nous avons utilisé deux déclinaisons de Manulex : l'une contenant les lexèmes, l'autre version se basant sur les lemmes (formes canoniques des lexèmes). Le lexique Manulex contient environ 49 000 lexèmes et a été compilé à partir d'un corpus extrait de 54 manuels scolaires (niveaux scolaires CP à CM2 du système français) représentant environ deux millions de mots (Lété, 2004). Les variables incluses dans Manulex sont le nombre de lettres, la catégorie

grammaticale, et la fréquence d'occurrence à différents niveaux scolaires, calculée par million de mots (fréquence U) et en indice de fréquence standardisé (IFS).

ÉQOL

Créé récemment pour le système scolaire québécois, ÉQOL contient 16652 lexèmes tirés de manuels et ouvrages de littérature jeunesse dont le niveau allait de la 1re à la 6e année du primaire (Stanké et al., 2019). Les variables incluses dans ÉQOL sont la catégorie grammaticale, les fréquences d'occurrence pour les niveaux scolaires, la longueur du mot exprimée en caractères et en syllabes, ainsi que d'autres variables périphériques donnant des informations sur le mot et qui n'ont pas été utilisées lors de cette étude. La documentation d'ÉQOL ne spécifie pas la taille du corpus analysé pour créer ce lexique; nous l'avons estimée à 492 570 mots en nous basant sur les fréquences d'occurrence rapportées.

LOMEQ2013

Ce lexique créé par le ministère de l'Éducation du Québec est disponible sur le Web dans l'environnement *Franqus*, développé en collaboration avec l'Université de Sherbrooke. Il contient 3314 lexèmes classifiés en six niveaux scolaires allant de la 1re à la 6e année du primaire, et associés à une catégorie grammaticale, de même qu'à des informations périphériques sur le lexème (phonétique, genre et nombre). Les fréquences d'occurrence ne sont pas incluses et la majorité des formes plurielle sont absentes. Nous avons ajouté les formes plurielles des noms communs, portant la taille du lexique LOMEQ2013 à 4921 lexèmes.

Modification et fusion des lexiques de référence

Pour Manulex et EQOL, nous avons ajouté une variable représentant le niveau scolaire où le lexème apparait pour la première fois dans le corpus, exprimé en âge. Le niveau scolaire était déjà inclus dans LOMEQ2013, et a été converti en âge équivalent. Les fréquences

d'occurrence de Manulex étaient déjà standardisées (IFS); nous avons fait la conversion pour les fréquences d'ÉQOL et LOMEQ2013. Les catégories grammaticales des trois lexiques ont été recodées afin de respecter la typologie *Universal Dependency* (De Marneffe et al., 2014). Les formes lemmatisées ont été récupérées depuis la base de données Dicollecte (*Dicollecte*, s.d.) en utilisant le lexème et sa catégorie grammaticale comme clé de recherche; les lexèmes n'existant pas dans Dicollecte ont été lemmatisés avec la bibliothèque *UDPipe* pour le langage R (Straka et al., 2016). La longueur syllabique du lexème a été déterminée en utilisant en priorité le nombre de syllabes spécifiées dans ÉQOL, puis dans la base de données Lexique3 (New et al., 2005), puis par un dictionnaire de césure (Rudis, 2019). Le Tableau 2 résume les opérations sur les variables des lexiques.

Tableau 2

Disponibilité des variables d'intérêt dans les trois lexiques utilisés

Variable	Lexiques		
	Manulex	**ÉQOL**	**LOMEQ2013**
Fréquence d'occurrence	Globale et par niveau scolaire.	Globale et par niveau scolaire.	*ND*
Année d'apparition dans le cursus scolaire	ND. Inférée à partir des fréquences d'occurrence par année scolaire.		Disponible sous forme de niveaux scolaire.
Nombre de syllabes	*ND,* obtenu par croisement avec le lexique ÉQOL ou Lexique3.	Disponible	*ND,* obtenu par croisement avec le lexique ÉQOL ou Lexique3.
Nombre de caractères	Disponible.	Disponible	*ND,* obtenu en comptant les caractères.
Rôle grammatical	Disponible, recodé selon le standard *Universal Dependency*.		
Formes lemmatisées	*ND,* obtenues par croisement avec le lexique Dicollecte, ou en lemmatisant avec *UDPipe*.		

Note. ND indique une variable qui était non disponible dans le lexique.

Par suite des opérations sur les lexiques, ALSI disposait en tout de 53 654 lexèmes distincts dont 3 249 étaient communs aux trois lexiques. Le Tableau 3 résume la proportion des lexèmes partagés entre les lexiques et se lit comme suit : sur les 48 886 lexèmes de Manulex, 25,6% se retrouvent aussi dans ÉQOL, 9% se retrouvent dans LOMEQ2013, et 6,6% sont à la fois dans les deux lexiques.

Tableau 3

Proportion de lexèmes partagés entre les lexiques

Lexique	Taille (lexèmes)	Lexèmes communs (%)			
		Manulex	ÉQOL	LOMEQ2013	Trois lexiques
Manulex	48886	--	25,6	8,9	6,6
ÉQOL	16652	75,3	--	20,0	19,5
LOMEQ2013	4920	88,9	67,8	--	66,0

Afin de tester la validité du traitement et de la fusion des trois lexiques, nous avons produit une matrice de corrélation des fréquences d'occurrence par million (fréquences U), des fréquences standardisés (IFS), et des âges d'apparition. Nous répliquons ainsi la méthode employée par Stanké et al. (2019) pour tester la spécificité de leur lexique ÉQOL. Le Tableau 4 présente les résultats. La corrélation entre les fréquences par million de Manulex et ÉQOL (r = 0,93) est similaire à ce qui était rapporté par Stanké et al. (2019), soit r = 0,94. L'âge d'apparition dans le cursus était davantage corrélé avec les fréquences IFS qu'avec les fréquences U, ce qui corrobore la pertinence des unités IFS pour capturer la complexité lexicale.

Tableau 4

Matrice de corrélation de Manulex, ÉQOL et LOMEQ2013

Variable	1	2	3	4	5	6
1. Manulex fréq. U						
2. Manulex fréq. IFS	0,16					

3. ÉQOL fréq. U	0,93	0,18				
4. ÉQOL fréq IFS	0,24	0,52	0,31			
5. Âge Manulex	-0,07	-0,62	-0,06	-0,33		
6. Âge ÉQOL	-0,09	-0,42	-0,11	-0,65	0,34	
7. Âge LOMEQ2013	-0,20	-0,49	-0,18	-0,49	0,30	0,41

Note. Corrélations de Pearson calculées pour les paires complètes d'observations. Toutes les corrélations rapportées dans cette matrice avaient une valeur *p* inférieure à 0,001. Les fréquences U sont les fréquences par million de mots, IFS désigne les indices de fréquence standardisés.

2.3.2. Extraction des mesures linguistiques

ALSI est formé d'un ensemble de scripts et de fonctions en langage R. La Figure 2 illustre la structure de traitement d'information : ses intrants, ses modules et leur organisation, de même que les données produites.

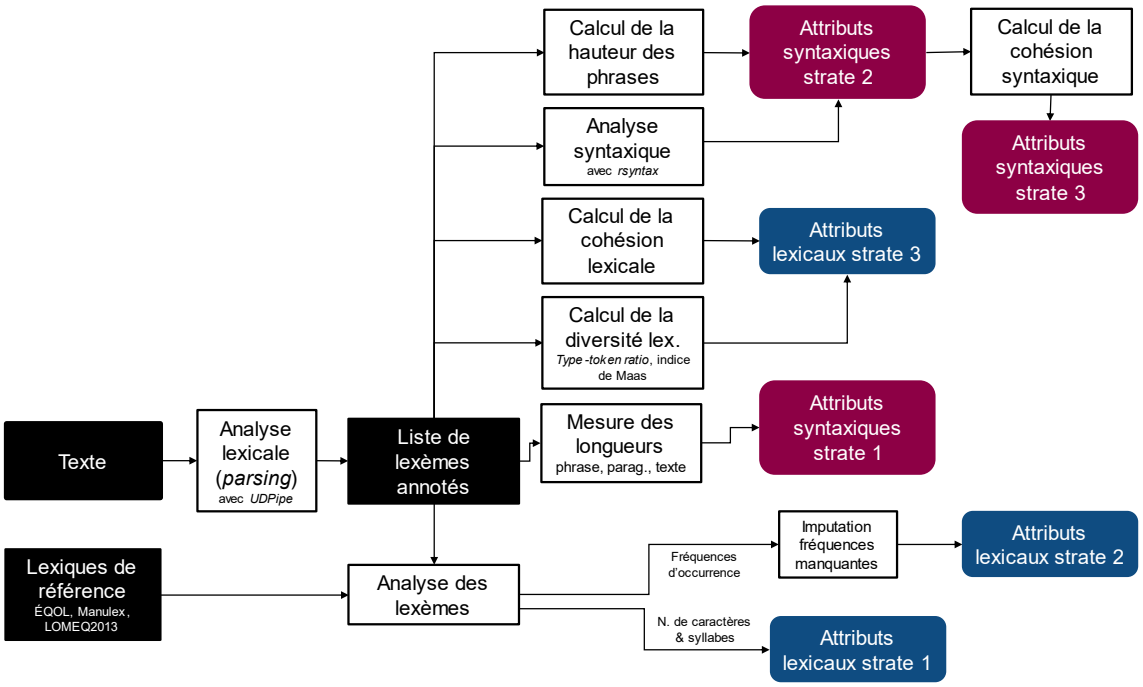


Figure 1. Architecture des modules de traitement d'ALSI. Les boites noires indiquent des structures de données utilisées en intrant, les boites blanches indiquent des processus, les boites colorées indiquent des matrices d'attributs lexicaux (bleu) ou syntaxiques (rouge). Les processus sont détaillés dans les sous-sections suivantes.

Annotation du corpus

ALSI reçoit initialement un dossier contenant des fichiers en format .txt, chaque fichier représentant un texte. Les textes sont « lus » et annotés par l'analyseur de texte *UDPipe* (Straka et al., 2016), qui convertit le texte en une base de données structurée. Chaque observation de la base de données représente un lexème (*token*) avec sa forme lemmatisée, des annotations indiquant sa catégorie grammaticale, ses caractéristiques (ex. : genre, nombre, temps du verbe), son parent dans la structure hiérarchique de la phrase, et le type de relation de dépendance qui l'unit avec son parent. Le parent désigne, dans l'arborescence de la phrase, le mot qui domine le lexème. Les rôles syntaxiques et les relations de dépendance sont tirés de la typologie universelle développée à l'Université de Stanford (De Marneffe et al., 2014).

La Figure 3 illustre quelques annotations de type *Universal Dependency* produites pour une phrase ainsi qu'une représentation graphique de sa structure hiérarchique. Dans cet exemple, *marcherai* est le parent de *je*, et la relation entre les mots est de type *nsubj*, pour sujet nominal; une liste complète des types de relation est disponible dans Marneffe et al. (2014; voir aussi Guillaume et al., 2019).

Id	Mot (*token*)	Lemme	Catégorie grammaticale	Id. parent	Relation
1	Je	il	PRON	2	nsubj
2	marcherai	marcher	VERB	--	--
3	sans	sans	ADP	4	obl
4	bruit	bruit	NOUN	2	case
5	.	.	PUNCT	2	--

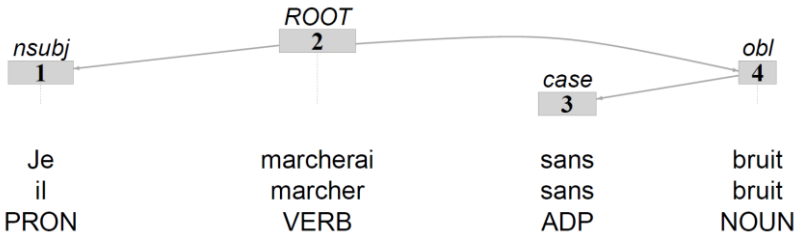


Figure 2. Exemple d'analyse d'une phrase. En haut, tableau de lexèmes annotés; l'identifiant du lexème et de son parent permet d'établir la structure hiérarchique de la phrase. En bas, représentation graphique de la structure hiérarchique de la phrase, basée sur les dépendances entre les mots. Voir de Marneffe et al. (2014) pour la liste des sigles.

L'étiquetage par UDPipe requiert une banque syntaxique (*treebank*), qui est un modèle de la langue entrainé par apprentissage machine supervisé sur un corpus de texte dont les mots étaient déjà annotés. La banque syntaxique utilisée par ALSI est *French-GSD 2.5*, elle a été entrainée sur un corpus d'environ 16 000 phrases (Guillaume et al., 2019). D'autres modèles similaires existent pour la langue française, GSD est celui ayant été entrainé sur le plus vaste corpus, et qui produisait les annotations les plus justes lors d'essais préalables avec vérification manuelle (meilleure identification de la catégorie grammaticale des lexèmes et de leur structure hiérarchique).

Appariement avec les lexiques de références

ALSI récupère dans une base de données les fréquences d'occurrence, ainsi que d'autres mesures décrites au Tableau 2. Afin de minimiser les erreurs d'appariement tout en maximisant le nombre de lexèmes appariés, ALSI tentera d'abord un couplage par lexème, catégorie grammaticale et lemme. Si aucune entrée n'est trouvée dans la base de données, ALSI fera une recherche moins spécifique par lexème et catégorie grammaticale, et enfin par lexème seulement. Au but du processus d'appariement, les lexèmes non trouvés dans le lexique de référence auront une fréquence d'occurrence inconnue. ALSI applique la méthode d'estimation de Good-Turing (Gale et Sampson, 1995) pour imputer ces données manquantes, ce qui permet d'estimer plus précisément la fréquence d'occurrence moyenne pour le texte. Le principe de la méthode Good-Turing est d'estimer la fréquence des lexèmes inconnus à partir des lexèmes dont la fréquence non ajustée était de 1 (le lexème ne figurait qu'une seule fois dans le corpus ayant servi à former le lexique de référence). Pour notre cas d'application, la formule se traduit ainsi, où N_1 est le

nombre d'entrées du lexique de référence dont la fréquence est de 1, N_0 est le nombre de lexèmes inconnus observés, et N_{ref} est la taille du lexique de référence; la première partie de l'équation donne la probabilité associée des lexèmes inconnus, celle-ci est multipliée par le nombre de mots observés (N_{obs}) pour produire un pseudocompte c :

$$c = \frac{N_1}{N_{ref} N_0} N_{obs}$$

Le pseudocompte est une estimation du nombre d'occurrences (ou fréquences « brutes ») dans le corpus ayant servi à former le lexique de référence; nous calculons ensuite son équivalent en fréquences U et SFI. La Figure 4 présente un exemple simplifié d'estimation de fréquence de Good-Turing lors de l'analyse d'un texte de huit lexèmes; la fréquence d'occurrence des deux lexèmes inconnus est estimée à 0,19 pour cet exemple.

Fréquence	N. de lexèmes
1	115
2	32
3	25
4	19
5	10
6	7
TOTAL	2430

$$c = \frac{115}{2430 \times 2} \times 8 = 0,19$$

Lexèmes du texte	Fréq. dans lexique		Fréq. dans lexique
Je	5		5
aime	3		3
aller	4		4
à	6		6
la	4		4
patinoire	?	→	*0,19*
en	2		2
hiver	?	→	*0,19*

Figure 3. Estimation de fréquences manquantes par méthode de Good-Turing. À gauche, le tableau montre que 115 lexèmes sur 2430 apparaissent une seule fois dans le lexique de référence. À droite : le lexique du texte comprend 8 lexèmes, dont 2 ayant une fréquence d'occurrence inconnue : patinoire et hiver. Ces données permettent d'estimer à environ 0,19 la fréquence d'occurrence des lexèmes inconnus, ce qui donnerait pour cet exemple une fréquence U d'environ 78.

Calcul de la profondeur syntaxique de la phrase

ALSI utilise la hauteur de l'arbre représentant les dépendances syntaxiques entre les mots comme équivalent pour la hauteur de la phrase. ALSI détermine d'abord la distance entre chaque mot et la racine de la phrase; c'est-à-dire le nombre de segments devant être parcourus pour atteindre la racine de la phrase. La profondeur syntaxique de la phrase est la plus longue distance trouvée. Par exemple, pour la phrase représentée à la Figure 5, la profondeur est de 4 puisque le plus long chemin possible remontant vers la racine de la phrase (*endormirent*) compte 4 segments.

Analyse syntaxique avec *rsyntax*

Le module d'analyse syntaxique permet d'identifier, de compter et de mesurer la longueur de différents constituants complexes de la phrase. Dans cette première version d'ALSI, nous avons ciblé les groupes verbaux et les groupes nominaux complexes. Le groupe verbal (GV) est opérationnalisé comme un groupe de mot dominé par un verbe conjugué, incluant le verbe avec auxiliaire, le groupe infinitif. Le groupe nominal complexe (GNC) est opérationnalisé dans ALSI comme un groupe de mots dominé par un nom, en incluant ses expansions. Dans la version actuelle, ALSI peut détecter les expansions suivantes : l'adjectif, le groupe participal (illustré à la Figure 4), la subordonnée relative, le groupe prépositionnel et le groupe infinitif agissant comme sujet du verbe. Ces analyses syntaxiques utilisent *rsyntax* pour R (Welbers et al., 2020). Les subordonnées complétives sont un autre type d'expansion du groupe nominal qui est plus complexe à détecter et n'a pas été incluse dans cette version d'ALSI.

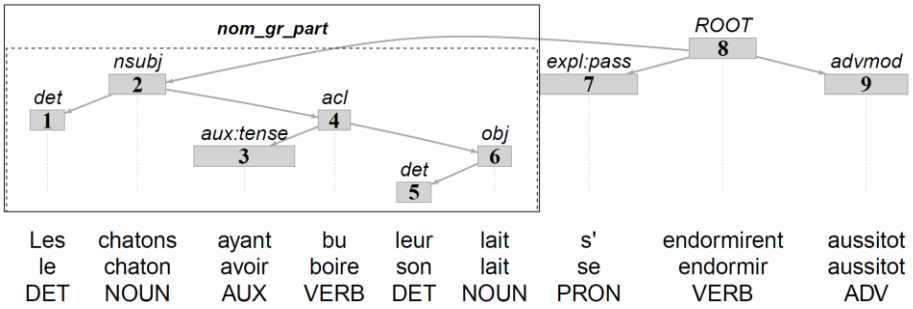


Figure 4. Représentation graphique d'une analyse avec *rsyntax.* L'encadré indique un groupe nominal avec groupe participal (indiqué comme *nom_gr_part*). Voir de Marneffe et al. (2014) pour la liste des sigles.

Mesures de cohésion

Pour estimer la cohésion lexicale, ALSI produit une première mesure en calculant l'indice de Jaccard (un coefficient de similitude) entre les lemmes uniques des paires de phrases adjacentes. Une seconde mesure est calculée selon la même méthode à partir des lexèmes étiquetés comme des noms ou noms propres.

Pour la cohésion syntaxique, ALSI produit d'abord pour chaque phrase du texte un vecteur contenant trois attributs syntaxiques: le nombre de groupes du verbe, la hauteur de l'arbre syntaxique, et le nombre de mots constituant la phrase. 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 le vecteur de la phrase et celui de la phrase suivante. Soit un espace dont les dimensions seraient les trois attributs syntaxiques, et dans lequel on représenterait chaque phrase comme un point, la distance entre deux phrases peut donc être comprise comme la distance qui sépare les deux points; une plus grande distance indique que les phrases sont moins similaires.

Sommaire des attributs extraits par ALSI

Dans sa première version, ALSI produit 48 attributs, soit 23 attributs lexicaux et 25 attributs syntaxiques. Les attributs sont résumés au Tableau 7 en annexe. Dans les noms des

variables, des suffixes indiquent 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 *d* indique une variable dichotomique dont la valeur passe à 1 dès que l'élément recherché est présent dans le texte.

2.4. Méthodologie

Nous décrivons dans ce qui suit une expérimentation dont l'objectif était de mettre à l'épreuve la capacité de l'outil ALSI à extraire des attributs qui caractérisent la complexité linguistique de textes en français québécois. L'expérimentation consistait à extraire les attributs d'un corpus de 600 textes à l'aide d'ALSI, puis à appliquer une procédure de sélection des attributs. Les résultats décrivent l'association entre les attributs sélectionnés et l'année scolaire associée au texte.

2.4.1. Corpus utilisé

Le corpus utilisé contenait 600 textes répartis entre 11 niveaux scolaires allant du primaire 1 au secondaire 5, selon les niveaux du système scolaire québécois. 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 6.

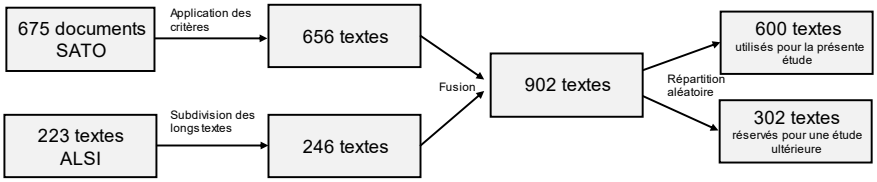


Figure 5. Combinaison puis répartition des textes provenant des banques SATO et ALSI.

La première banque de textes est un ensemble de textes utilisé pour le développement et la calibration de l'analyseur SATO-Calibrage (Daoust et al., 1996). Elle contenait initialement 675 documents au format .txt, répartis entre les 11 niveaux scolaires et tirés de manuels scolaires ou d'épreuves nationales. Après avoir scindé les documents qui contenaient plusieurs textes, supprimé les doublons et appliqué les critères d'exclusion, nous avons retenu 656 textes de la banque SATO-Calibrage. La deuxième banque de textes a initialement été créée afin d'étudier les attributs de macrogenres textuels qu'on retrouve principalement au secondaire dans le cursus du Québec (expressif et argumentatifs), nous la désignons comme banque ALSI. Les textes de la banque ALSI proviennent principalement des manuels scolaires français publiés au Québec après l'an 2000 pour les classes de 6e primaire à 5e secondaire. Elle contenait initialement 223 textes répondant à nos critères. Afin d'augmenter la taille du corpus tout en uniformisant leur longueur, les textes de la banque ALSI ont été scindés en deux lorsqu'ils comptaient un nombre de mots supérieur à deux fois la moyenne. Après ces subdivisions et l'application de nos critères d'exclusion, la banque ALSI contenait 246 textes. Les informations paratextuelles suivantes ont été retirées des textes des deux banques: numéros de page, de paragraphe ou de ligne et autres marques ajoutées par l'éditeur, remarques et définitions ajoutées en marge, de même que les titres et intertitres sauf lorsque ceux-ci formaient une phrase incluant au moins un verbe conjugué.

Le corpus formé en combinant les banques SATO et ALSI comptait 902 textes (43 820 phrases). Nous avons réservé environ le tiers de ce corpus (sélectionné aléatoirement) pour une expérimentation ultérieure en estimation du niveau de difficulté du texte (article 2 de la présente thèse) portant la taille du corpus utilisé par la présente étude à 600 textes (29 709 phrases). La classification de chaque texte était spécifiée par le manuel scolaire ou la ressource pédagogique

de provenance, aucun texte n'a été reclassifié. La provenance des textes et leur distribution entre les niveaux scolaires est indiquée au Tableau 5.

Tableau 5

Provenance du corpus utilisé et distribution entre les 11 années 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

2.4.2. Procédure d'extraction et de sélection d'attributs

Nous avons utilisé ALSI pour analyser automatiquement les 600 textes, 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. Considérant qu'ALSI, à l'instar de *Lexile* ou *ATOS*, pourrait être employé pour classifier des textes par année scolaire, nous avons ensuite appliqué une procédure de sélection aux attributs. La procédure visait à éliminer les attributs qui contribueraient peu ou mal à la classification, ou qui risqueraient de causer des problèmes de surajustement des modèles, notamment en raison d'une multicolinéarité. Le surajustement (*overfitting*) signifie dans ce contexte que les attributs feraient dévier les paramètres du modèle vers une solution bien adaptée aux textes de la phase d'entrainement, mais qui ne se généraliserait pas à de nouveaux textes.

Nous avons exclu d'emblée les attributs qui reflétaient 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. Nous avons également exclu les mesures dichotomiques portant sur la présence de verbes d'une catégorie spécifique (par ex. : des verbes conjugués au conditionnel présent) puisque ces attributs nous

semblaient fortement liés à des caractéristiques sans lien direct avec la complexité, notamment le genre textuel. Puisque certains genres textuels sont étudiés à des années spécifiques du parcours scolaire, de tels attributs auraient pu gonfler les résultats en contribuant peu à capturer la complexité linguistique du corpus utilisé.

Suivant la chaine de traitement proposée par Taneja et al. (2014), nous avons calculé le gain d'information de chaque attribut, et retiré les attributs dont le gain d'information était de zéro. Le gain d'information (GI) se définit comme la quantité d'information concernant la classe qui est ajoutée lorsque l'attribut est présent, il équivaut à l'entropie de Shannon totale du système moins l'entropie conditionnelle à l'ajout de l'attribut (Karegowda et al., 2010; Taneja et al., 2014). Cette étape écarte du même coup les variables ayant une variance nulle.

Les cas de multicolinéarité les plus évidents étaient éliminés en retirant les attributs fautifs un à un, et en tentant de préserver ceux ayant le GI le plus élevé. Pour détecter les groupes de variables ayant une multicolinéarité, nous avons utilisé la fonction *findLinearCombos* dans la bibliothèque *caret* pour le langage R (Kuhn, 2011). Nous avons ensuite détecté les combinaisons d'attributs produits à partir des mêmes mesures linguistiques, par exemple la longueur moyenne de la phrase et le logarithme moyen de la longueur de la phase. Ces combinaisons étaient éliminées en conservant seulement l'attribut ayant le GI le plus élevé. Les variables ayant passé chaque étape de sélection formaient la sélection complète d'attributs.

Nous avons également retenu un sous-ensemble de six attributs choisis depuis la sélection complète. Chaque attribut de cette sélection réduite était le meilleur représentant (GI le plus élevé) de son type, tel que spécifié dans la typologie d'ALSI.

2.4.3. Analyses statistiques

L'objectif des analyses était de produire de statistiques décrivant l'association entre les attributs et le niveau de difficulté du texte, exprimé en années scolaires. Nous avons produit une liste des variables sélectionnées, associés à des mesures de l'association statistique entre l'attribut et le niveau scolaire. Les mesures étaient le GI, le coefficient de corrélation de Spearman avec intervalles de confiance à 95% avec correction de Fieller (Bishara et Hittner, 2017), et la log-vraisemblance pour une régression logistique multinomiale. Nous avons en outre calculé la valeur médiane des attributs par classe afin de pouvoir examiner la progression des valeurs obtenues au fil des niveaux scolaires.

2.5. Résultats

La procédure de sélection a été appliquée à un groupe initial de 48 attributs produits par l'outil ALSI et considérés comme pertinents pour cette étude. Une liste complète de ces attributs se trouve au Tableau 7 en annexe, et précise la raison du rejet le cas échéant. Sur les 48 attributs considérés, un attribut (*longGNC_m*) a été retiré en raison d'un gain d'information (GI) égal à zéro, un attribut (*ageEqol*_m) a été retiré afin d'éviter des problèmes de multicolinéarité détectés par la fonction *findLinearCombos*, et 27 attributs ont été retirés pour éviter les conflits entre variables formées à partir des mêmes mesures linguistiques ou de mesures similaires. La sélection finale comptait 21 attributs (10 lexicaux, 11 syntaxiques).

Le Tableau 6 décrit l'association statistique entre l'année scolaire et les attributs de la sélection finale en présentant le gain d'information, le coefficient de Spearman, et la statistique χ^2 du test de rapport de vraisemblance. Pour les 21 attributs de la sélection finale, les coefficients de Spearman et les statistiques $\chi 2$ étaient significatifs à un seuil de $p < 0{,}001$. Le Tableau 8 en annexe donne les valeurs médianes des attributs de la sélection finale par année scolaire.

Tableau 6

Mesures de l'association statistique entre l'attribut et l'année scolaire

#	Attributs	*GI*	*rs* [IC 95%]	*LR*	Type
1	**ageManulex_m**	**0,494**	**0,70 [0,65, 0,74]**	**584,99**	**Lexique 2**
2	freqManulexSfi_m	0,489	-0,69 [-0,73, -0,65]	575,92	Lexique 2
3	freqEqolSfi_m	0,437	-0,66 [-0,70, -0,61]	511,82	Lexique 2
4	**longPh_90**	**0,425**	**0,71 [0,66, 0,75]**	**506,42**	**Syntaxe 1**
5	**longMotOrtho_m**	**0,421**	**0,62 [0,57, 0,67]**	**509,91**	**Lexique 1**
6	**cohesionSynDist_m**	**0,342**	**0,71 [0,67, 0,75]**	**437,04**	**Syntaxe 3**
7	motSeuilSyll_p	0,329	0,61 [0,56, 0,66]	473,49	Lexique 1
8	ageMels_m	0,320	0,58 [0,53, 0,64]	430,00	Lexique 2
9	**GNC_m**	**0,319**	**0,62 [0,57, 0,67]**	**405,43**	**Syntaxe 2**
10	**maas_lemma**	**0,319**	**0,54 [0,48, 0,59]**	**415,35**	**Lexique 3**
11	dansManulex_p	0,309	-0,68 [-0,72, -0,63]	407,33	Lexique 2
12	hauteurPh_m	0,306	0,61 [0,56, 0,66]	400,33	Syntaxe 2
13	verbesConju_m	0,286	0,63 [0,57, 0,67]	352,17	Syntaxe 2
14	GV_m	0,241	0,59 [0,53, 0,64]	284,54	Syntaxe 2
15	dansMels_p	0,234	-0,60 [-0,65, -0,55]	369,56	Syntaxe 2
16	virgule_m	0,221	0,64 [0,59, 0,68]	335,70	Syntaxe 1
17	partPres_m	0,174	0,49 [0,43, 0,55]	214,40	Syntaxe 2
18	partPass_m	0,153	0,52 [0,46, 0,58]	240,87	Syntaxe 2
19	verbeComplexe_m	0,152	0,44 [0,37, 0,50]	165,69	Syntaxe 2
20	cohesionLemma_m	0,117	-0,39 [-0,46, -0,32]	147,54	Lexique 3
21	phMarqueur_m	0,117	0,44 [0,37, 0,51]	132,26	Syntaxe 2

Note. Statistiques calculées à partir d'un corpus de 600 textes, pour les 21 attributs sélectionnés et 11 années scolaires. Les six attributs de la sélection réduite sont indiqués en caractères gras. *GI* indique le gain d'information, *rs* indique le coefficient de Spearman avec intervalle de confiance, *LR* est la statistique χ^2 produite par un test du rapport de vraisemblance. Toutes les corrélations de Spearman de ce tableau étaient statistiquement significatives au seuil $p < 0{,}001$. Le type renvoie à la typologie décrite dans le présent article.

Les attributs de la sélection réduite sont indiqués en caractères gras dans le Tableau 6. Il s'agit, en ordre de gain d'information, de l'âge moyen de première apparition dans le lexique Manulex (*ageManulex_m*), du 90^{e} percentile de la longueur des phrases (*longPh_90*), de la longueur orthographique moyenne (*longMotOrtho_m*), de la cohésion syntaxique de phrase à phrase exprimée comme une mesure de distance (*cohesionSynDist_m*), du nombre moyen de groupes nominaux complexes par phrase (*GNC_m*), et de l'indice de Maas calculé sur les formes lemmatisées (*maas_lemma*). La Figure 7 montre les distributions de ces six attributs, par année

scolaire, et permet de visualiser la progression des attributs de même que la présence d'observations aberrantes (*outliers*).

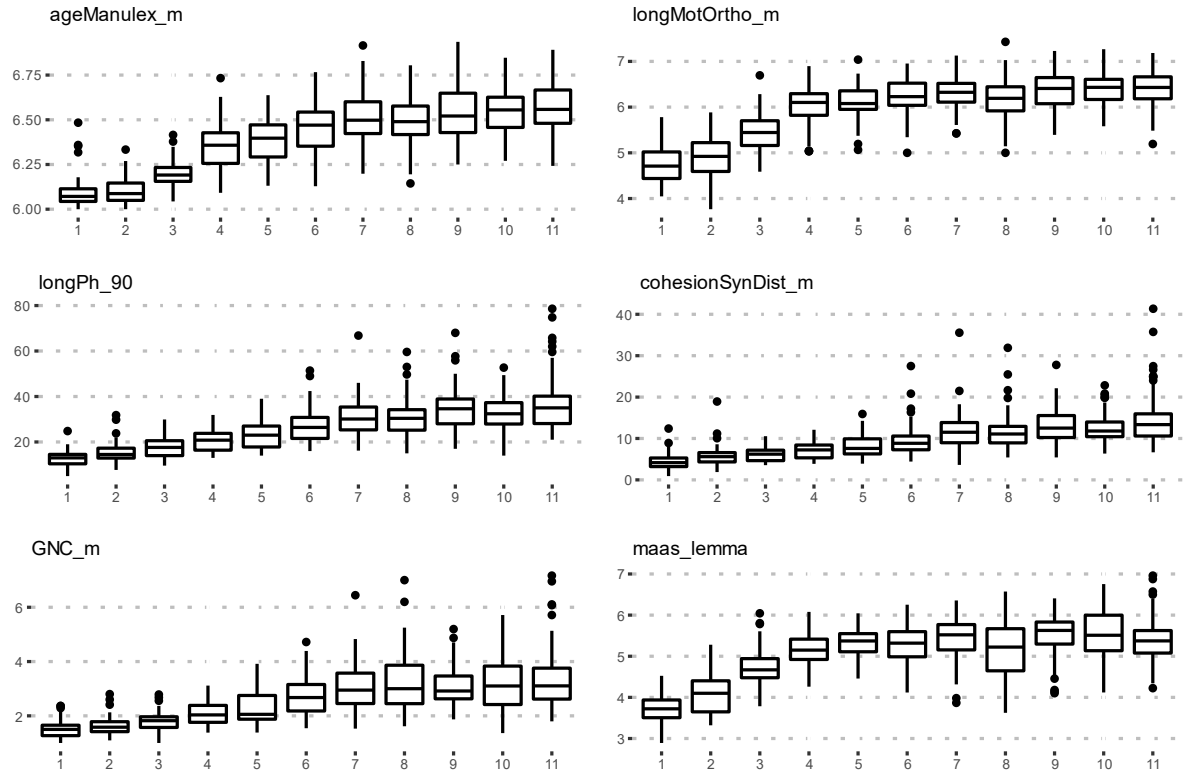


Figure 6. Diagrammes en boite des six attributs de la sélection réduite, par année scolaire. La boite indique les percentiles 25 à 75, la ligne dans la boite indique la médiane. L'axe des abscisses indique l'année scolaire dans le système primaire (1 à 6) et secondaire (7 à 11) du Québec.

2.6. Discussion

Les résultats de la procédure de sélection d'attributs montrent que la majorité des 48 attributs considérés étaient possiblement contributifs à la classification du texte par niveau scolaire; un seul attribut avait un GI nul. L'association statistique entre les 21 attributs sélectionnés et le niveau scolaire était assez élevée, le coefficient de Spearman moyen, calculé par transformé Fischer en employant les valeurs absolues, était de *rs* = 0,574, ce qui correspond à une association statistique modérée selon les barèmes rapportés par Akoglu (2018) pour la recherche en psychologie.

Les attributs affichaient des progressions différentes au fil des années scolaires. Certains attributs, comme *cohesionSynDist_m*, semblent varier de manière assez linéaire, tandis que d'autres attributs manifestent un effet de plateau, par exemple l'indice de Maas calculé sur les lemmes (*maas_lemma*) augmente jusqu'à la fin du primaire, puis se stabilise. Ces résultats sont en phase avec les observations de Daoust et al. (1996). Les plateaux observés suggèrent que certains attributs linguistiques atteignent leur complexité limite durant le parcours scolaire. Une autre explication des effets de plateau est que l'ALSI n'est peut-être pas apte à mesurer l'augmentation de la complexité au-delà d'un certain point. Ainsi, un plateau a pu être introduit 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 capturer la complexité lexicale au-delà de la 6e année. Enfin, la 11e année (5e secondaire) semblait plus prompte aux observations aberrantes (*outliers*); d'autres analyses seraient requises afin d'investiguer si ces textes se démarquant des autres avaient des caractéristiques communes (par ex. : maison d'édition, genre textuel).

Nous retenons deux résultats importants concernant la nature des attributs retenus par la procédure de sélection. Premièrement, nos résultats montrent que les attributs « de surface » peuvent effectivement contribuer à caractériser la complexité linguistique. Les attributs *longMotOrtho_m* (longueur moyenne des mots) et *longPh_90* (90e percentile de la longueur des phrases du texte) répondent au concept d'attribut « de surface », mais comptaient parmi les attributs ayant le GI le plus élevé. Ce résultat peut s'expliquer en partie par la manière dont ALSI mesure ces attributs. Pour les attributs portant sur la longueur du mot ou sur un lexique de référence, ALSI opère à partir de la liste des lexèmes uniques, sans tenir compte de la répétition de certains lexèmes. De plus, ALSI ignore lors de ce calcul certaines catégories de mots qui

pourraient fausser le calcul, tels les noms propres et les déterminants. Il ne s'agit donc pas de la longueur orthographique telle que mesurée typiquement dans les formules de lisibilité, mais de sa variante plus complexe. Selon le même principe, *longPh_90* est une alternative à la longueur moyenne de la phrase qui capturait mieux, pour le corpus utilisé, la complexité du texte – le Tableau 7 en annexe de cet article présente les résultats complets pour la sélection d'attributs.

Deuxièmement, nos résultats montrent que la cohésion linguistique, principalement la cohésion syntaxique, peut contribuer à modéliser la complexité du texte : les attributs *cohesionSynDist_m* (cohésion syntaxique moyenne) et *cohesionLemma_m* (cohésion lexicale calculée à partir des formes lemmatisées) occupaient respectivement le 6^{e} et 20^{e} rang de la sélection finale de 21 attributs. Ce résultat est important puisqu'il ajoute un soutien empirique à l'hypothèse selon laquelle la cohésion affecte la compréhension (O'Reilly et McNamara, 2007). Le rang obtenu par *cohesionLemma_m* indique que la cohésion lexicale pourrait contribuer, bien que faiblement, à estimer le niveau de difficulté du texte, ce qui est en phase avec les résultats rapportés par Todirascu et al. (2016). Ce résultat demeure cependant paradoxal puisque, alors que la reprise d'information est jugée comme facilitante (O'Reily et McNamara, 2007), des travaux d'oculométrie montrent plutôt que cette répétition alourdit le texte lorsqu'on la compare au remplacement du nom par un pronom (Kennison et Gordon, 1997).

Dans l'ensemble, nos résultats dépendent de l'échantillon de textes, et pourraient différer si l'expérience était répétée sur un corpus différent; la portée de cette étude dépend donc de la prémisse selon laquelle le corpus utilisé était représentatif des textes trouvés dans le cursus scolaire québécois. 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, les textes des années 1 à 5 sont donc en moyenne plus datés; une piste à explorer serait

donc d'ajouter des textes plus récents couvrant la période allant de la 1ère à la 5e année 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 la complexité morphologique. Enfin, d'autres analyses seraient requises pour évaluer la contribution réelle des attributs à un modèle de classification des textes par année scolaire.

2.7. Conclusion

Nous avons décrit dans cette étude un nouvel outil d'analyse linguistique, ALSI, qui produit une variété d'attributs linguistiques dans l'objectif de mesurer la complexité intrinsèque du texte en français québécois. Après avoir motivé la création d'un nouvel outil, l'article décrivait les fondements théoriques d'ALSI. Nous avons ensuite présenté les bases de données lexicales incluses dans ALSI, de même que les procédures déployées pour extraire les attributs du texte. Le second volet de l'article avait pour objectif de déterminer quels attributs étaient les plus prometteurs pour estimer l'année scolaire des textes corpus en français québécois. Une procédure de sélection théorique et empirique a retenu 21 attributs sur 48 considérés, ainsi qu'une sous-sélection constituée du meilleur attribut pour chacun des 6 types. Dans l'ensemble, les statistiques décrivant l'association statistique entre les attributs et l'année scolaire appuient la validité de l'outil ALSI en montrant qu'il extrait des attributs décrivant la complexité du texte. Les résultats montrent que les attributs à faible complexité sont toujours d'actualité, et que leur contribution peut être améliorée en raffinant leur procédure de mesure, par exemple en optant pour le 90^{e} percentile de la longueur des phrases plutôt que la longueur moyenne. Notre étude met en évidence le potentiel des attributs mesurant la cohésion linguistique, particulièrement la

cohésion syntaxique, apportant un appui empirique aux travaux ayant motivé la création de l'outil *Coh-Metrix* (McNamara et Graesser, 2011).

La présente étude a, en somme, proposé des attributs qui peuvent être mesurés avec l'outil ALSI et sont associés à la complexité linguistique des textes employés en milieu scolaire au Québec. Nous voyons plusieurs applications d'ALSI dans le domaine de l'éducation. ALSI pourrait s'intégrer à une chaine d'analyse visant à estimer automatiquement le niveau de difficulté de textes; d'autres travaux permettraient d'évaluer la performance des attributs d'ALSI lorsqu'utilisés pour classifier un corpus. L'outil pourrait ainsi contribuer à une démarche de validation d'épreuves et de tests en estimant *a priori* la difficulté linguistique des items. ALSI pourrait également aider à sélectionner ou créer du matériel didactique ayant un niveau linguistique approprié, ou qui favorise l'apprentissage de certains objets d'enseignement en français. Enfin, une prochaine version de l'outil pourrait prendre la forme d'une application Web afin de simplifier son utilisation hors du milieu de la recherche.

2.8. Annexe du premier article

Tableau 7

Liste complète des attributs avec mesures d'association statistique

#	Attributs	GI	*rs* [IC 95%]	LR	Type	Statut	Description
1	**ageManulex_m**	**0,494**	0,70 [0,65, 0,74]	584,99	**Lexique 2**	**SÉLECTION**	**Âge moyen de 1ère apparition dans le lexique Manulex.**
2	freqManulexSfi_m	0,489	-0,69 [-0,73, -0,65]	575,92	Lexique 2	SÉLECTION	Fréquence standardisée moyenne selon Manulex
3	freqEqolSfi_m	0,437	-0,66 [-0,7, -0,61]	511,82	Lexique 2	SÉLECTION	Fréquence standardisée moyenne selon ÉQOL
4	**longPh_90**	**0,425**	0,71 [0,66, 0,75]	506,42	**Syntaxe 1**	**SÉLECTION**	**90e percentile de la longueur des phrases**
5	**longMotOrtho_m**	**0,421**	0,62 [0,57, 0,67]	509,91	**Lexique 1**	**SÉLECTION**	**Longueur orthographique moyenne (nombre de caractères).**
6	dansEqol_p	0,405	-0,68 [-0,72, -0,64]	449,84	Lexique 2	DOUB (3)	Proportion de mots présents dans ÉQOL.
7	ageMoy	0,400	0,66 [0,62, 0,71]	552,08	Lexique 2	DOUB (1)	Âge moyen de 1ère apparition – combinaison des trois lexiques.
8	longMotSyll_m	0,377	0,63 [0,58, 0,68]	512,90	Lexique 1	DOUB (5)	Longueur moyenne des mots, en syllabes.
9	longPh_m	0,377	0,66 [0,61, 0,71]	460,62	Syntaxe 1	DOUB (4)	Longueur moyenne des phrases.
10	longPh_logm	0,371	0,66 [0,61, 0,71]	461,27	Syntaxe 1	DOUB (4)	Longueur moyenne des phrases, échelle logarithmique.
11	longPh30_p	0,368	0,68 [0,63, 0,72]	352,56	Syntaxe 1	DOUB (4)	Proportion de phrases comptant plus de 30 mots.
12	**cohesionSynDist_m**	0,342	0,71 [0,67, 0,75]	437,04	**Syntaxe 3**	**SÉLECTION**	**Distance euclidienne moy. entre trois attributs syntaxiques des phrases adjacentes.**
13	motSeuilSyll_p	0,329	0,61 [0,56, 0,66]	473,49	Lexique 1	SÉLECTION	Proportion de mots comptant plus de trois syllabes.
14	freqEqol_m	0,327	-0,56 [-0,62, -0,51]	412,14	Lexique 2	DOUB (3)	Fréquence moyenne selon ÉQOL.
15	motSeuilOrtho_p	0,323	0,59 [0,53, 0,64]	438,74	Lexique 1	DOUB (5)	Proportion de mots comptant plus de huit caractères.
16	ageMels_m	0,320	0,58 [0,53, 0,64]	430,00	Lexique 2	SÉLECTION	Âge moyen de 1ère apparition dans la liste orthographique du MELSQ.
17	cohesionSynDist_logm	0,320	-0,69 [-0,73, -0,64]	440,84	Syntaxe 3	DOUB (12)	Distance syntaxique des phrases adjacentes, échelle logarithmique.
18	**GNC_m**	0,319	0,62 [0,57, 0,67]	405,43	**Syntaxe 2**	**SÉLECTION**	**Nombre moyen de groupes nominaux complexes par phrase.**
19	**maas_lemma**	0,319	0,54 [0,48, 0,59]	415,35	**Lexique 3**	**SÉLECTION**	**Indice de Maas calculé sur les formes lemmatisées.**
20	dansManulex_p	0,309	-0,68 [-0,72, -0,63]	407,33	Lexique 2	SÉLECTION	Proportion de mots dans Manulex.
21	ageEqol_m	0,308	0,65 [0,6, 0,69]	501,00	Lexique 2	CO (1, 7, 16)	Âge moyen de 1ère apparition dans ÉQOL.
22	freqEqol_90	0,307	-0,54 [-0,59, -0,48]	341,84	Lexique 2	DOUB (3)	90e percentile de la fréquence d'occurrence dans ÉQOL.
23	freqManulex_m	0,307	-0,55 [-0,61, -0,49]	430,98	Lexique 2	DOUB (2)	Fréquence d'occurrence moyenne dans Manulex.
24	hauteurPh_m	0,306	0,61 [0,56, 0,66]	400,33	Syntaxe 2	SÉLECTION	Profondeur syntaxique moyenne.
25	maas_token	0,302	0,52 [0,46, 0,58]	395,17	Lexique 3	DOUB (18)	Indice de Maas calculé sur les lexèmes.
26	hauteurPh_logm	0,289	0,56 [0,51, 0,62]	348,16	Syntaxe 2	DOUB (24)	Profondeur syntaxique moyenne - échelle logarithmique.
27	verbesConju_m	0,286	0,63 [0,57, 0,67]	352,17	Syntaxe 2	SÉLECTION	Nombre moyen de verbes conjugués par phrase.

28	GNCGr_m	0,255	0,57 [0,51, 0,62]	311,88	Syntaxe 2	DOUB (17)	N. moy. de groupes nominaux avec groupe complément, participal ou prépositionnel.
29	GV_m	0,241	0,59 [0,53, 0,64]	284,54	Syntaxe 2	SÉLECTION	Nombre moyen de groupes verbaux par phrase.
30	freqManulex_90	0,236	-0,53 [-0,58, -0,46]	375,53	Lexique 2	DOUB (2)	90^{e} percentile de la fréquence d'occurrence dans Manulex.
31	dansMels_p	0,234	-0,60 [-0,65, -0,55]	369,56	Lexique 2	DOUB (16)	Proportion de mots présents dans la liste orthographique du ministère de l'éducation du Québec.
32	virgule_m	0,221	0,64 [0,59, 0,68]	335,70	Syntaxe 1	SÉLECTION	Nombre moyen de virgules par phrase.
33	GVFin_m	0,187	0,57 [0,51, 0,62]	266,75	Syntaxe 2	DOUB (29)	Nombre moyen de groupes verbaux finis par phrase.
34	partPres_m	0,174	0,49 [0,43, 0,55]	214,40	Syntaxe 2	SÉLECTION	Nombre moyen de participes présents par phrase.
35	partPres_p	0,171	0,33 [0,26, 0,4]	69,12	Syntaxe 2	DOUB (34)	Proportion de phrases ayant un participe présent.
36	partPass_m	0,153	0,52 [0,46, 0,58]	240,87	Syntaxe 2	SÉLECTION	Nombre moyen de participes passés par phrase.
37	TTR_token	0,152	-0,32 [-0,39, -0,24]	211,90	Lexique 3	SÉLECTION	Diversité lexicale (*type-token ratio*) mesurée à partir des lexèmes.
38	verbeComplexe_m	0,152	0,44 [0,37, 0,5]	165,69	Syntaxe 2	SÉLECTION	Proportion des verbes conjugués considérés comme complexes (voir Daoust et al., 1996)
39	phVerbeComplexe_p	0,135	0,40 [0,33, 0,47]	151,06	Syntaxe 2	DOUB (38)	Proportion des phrases contenant au moins un verbe complexe.
40	cohesionLemma_m	0,117	-0,39 [-0,46, -0,32]	147,54	Lexique 3	SÉLECTION	Indice de cohésion syntaxique calculé depuis les formes lemmatisées.
41	phMarqueur_m	0,117	0,44 [0,37, 0,51]	132,26	Syntaxe 2	SÉLECTION	Nombre moyen de marqueurs argumentatifs par phrase.
42	TTR_lemma	0,106	-0,26 [-0,33, -0,18]	165,71	Lexique 3	DOUB (37)	Diversité lexicale (*type-token ratio*) mesurée à partir des lemmes.
43	verbeComplexe_p	0,096	0,31 [0,23, 0,38]	89,69	Syntaxe 2	DOUB (38)	Proportion des verbes étant complexes selon la définition de Daoust et al. (1996).
44	cohesionNom_m	0,090	-0,36 [-0,43, -0,29]	109,65	Lexique 3	DOUB (40)	Indice de cohésion lexicale mesurée à partir des noms communs uniquement.
45	phMarqueur_p	0,087	0,39 [0,32, 0,46]	130,52	Syntaxe 2	DOUB (41)	Proportion de phrases contenant au moins un marqueur argumentatif.
46	GVSub_m	0,085	0,36 [0,29, 0,43]	114,28	Syntaxe 2	DOUB (29)	Nombre moyen de groupes verbaux avec subordonnée complément de phrase.
47	partPass_p	0,066	0 [-0,08, 0,08] *ns*	39,77	Syntaxe 2	DOUB (36)	Proportion des verbes conjugués étant des participes passés.
48	longGNC_m	0,000	0,18 [0,09, 0,25]	55,00	Syntaxe 2	GI = 0	Longueur moyenne des groupes nominaux complexes.

Note. Statistiques calculées sur 600 textes. Les caractères gras indiquent les attributs de la sélection réduite. Les valeurs pour la colonne Statut sont les suivantes : SÉLECTION indique que l'attribut fait partie de la sélection finale, DOUB indique que l'attribut a été retiré en raison d'un conflit potentiel avec un autre attribut indiqué entre parenthèses, CO indique que l'attribut a été retiré en raison d'une multicolinéarité avec les attributs entre parenthèses, GI = 0 indique que l'attribut a été retiré, car son gain d'information était nul. Les coefficients rho de Spearman (*rs*) et ratios de vraisemblance (*LR*) étaient tous significatifs à un seuil de $p < 0,001$, à l'exception de l'attribut *partPass_p* dont le rho de Spearman n'était pas significatif.

Tableau 8

Médiane des attributs par classe (sélection complète)

Attributs	1	2	3	4	5	6	7	8	9	10	11
ageManulex_m	6,07	6,08	6,19	6,36	6,36	6,47	6,5	6,48	6,5	6,56	6,56
freqManulexSfi_m	64,97	63,47	60,6	56,91	57,07	56,12	54,41	55,57	54,67	53,37	53,99
freqEqolSfi_m	63,63	63,02	60,22	57,39	56,13	56,88	55,35	55,62	54,91	54,11	54,82
longPh_90	12,7	14,4	17,7	21,1	22,2	25,5	30	31	34,4	30,75	34,9
longMotOrtho_m	4,66	4,91	5,47	6,07	6,07	6,21	6,32	6,19	6,39	6,45	6,43
cohesionSyn_distm	4,07	5,58	6,29	7,58	7,6	8,55	11,25	10,97	12,51	11,47	13,52
motSeuilSyll_p	0,13	0,16	0,21	0,28	0,28	0,32	0,34	0,32	0,33	0,34	0,34
ageMels_m	7,22	7,33	7,55	7,84	7,73	7,97	8,04	7,86	7,94	8,05	8,01
GNC_m	1,43	1,59	1,83	2,05	2	2,73	2,88	2,95	2,93	2,87	3,06
maas_lemma	1,92	2,02	2,16	2,26	2,3	2,3	2,35	2,29	2,37	2,36	2,32
dansManulex_p	0,98	0,98	0,98	0,96	0,96	0,95	0,95	0,94	0,94	0,93	0,92
hauteurPh_m	3,57	3,75	4,23	4,54	4,38	5,12	5,11	5,26	5,38	5,2	5,54
verbesConju_m	1	1,33	1,38	1,55	1,65	1,78	1,94	2	2,15	2	2,15
GV_m	1,67	2,05	2,12	2,37	2,43	2,57	2,7	2,78	2,99	2,83	3,07
dansMels_p	0,7	0,68	0,65	0,59	0,57	0,58	0,56	0,57	0,56	0,54	0,55
virgule_m	0,43	0,5	0,54	0,63	0,73	0,95	1,05	1,14	1,3	1,2	1,27
partPres_m	0	0	0	0,02	0,04	0,05	0,09	0,08	0,1	0,08	0,07
partPass_m	0,08	0,2	0,17	0,3	0,34	0,39	0,47	0,55	0,53	0,51	0,53
verbeComplexe_m	0,06	0,11	0,11	0,12	0,43	0,25	0,35	0,38	0,46	0,37	0,35
cohesionLemma_m	0,16	0,15	0,14	0,16	0,14	0,15	0,12	0,13	0,13	0,13	0,13
phMarqueur_m	0,11	0,17	0,23	0,44	0,45	0,56	0,71	0,66	1,18	0,76	0,85

Note. Valeurs médianes par classe des 21 attributs de la sélection complète, calculées depuis 600 textes.

2.9. Bibliographie

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

Alavi, S. M. et Ranjbaran, F. (2018). Constructing and validating a Q-matrix for cognitive diagnostic analysis of a reading comprehension test battery. Journal of English Language Teaching and Learning, *10*(21), 1-31.

Avenia-Tapper, B. et 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

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

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

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

Buck, G. et Tatsuoka, K. K. (1998). Application of the rule-space procedure to language testing: examining attributes of a free response listening test. *Language Testing*, *15*(2), 119-157. https://doi.org/10.1177/026553229801500201

Buck, G., Tatsuoka, K. et Kostin, I. (1997). The subskills of reading: Rule-space analysis of a multiple-choice test of second language reading comprehension. *Language Learning*, *47*(3), 423-466. https://doi.org/10.1111/0023-8333.00016

Chen, H. et Chen, J. (2016). Retrofitting non-cognitive-diagnostic reading assessment under the generalized DINA model framework. *Language Assessment Quarterly*, *13*(3), 218-230. https://doi.org/10.1080/15434303.2016.1210610

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/

Coltheart, M. (2005). Modeling reading: the dual-route approach. Dans M. J. Snowling & C. Hulme (Eds.), *The science of reading: A handbook* (p. 6–23). Blackwell Publishing. https://doi.org/10.1002/9780470757642.ch1

Crossley, S. A., Kyle, K. et McNamara, D. S. (2016). The development and use of cohesive devices in L2 writing and their relations to judgments of essay quality. *Journal of Second Language Writing*, *32*, 1-16. https://doi.org/10/f8rtzh

Daoust, F., Laroche, L. et 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, May). Universal Stanford dependencies: A cross-linguistic typology. Dans *LREC* (Vol. 14, pp. 4585-4592).

Dempster, E. R. et 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

Dicollecte. (s. d.). freegreek. https://github.com/freegreek/dicollecte

Feng, L., Jansche, M., Huenerfauth, M. et 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. et 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).

François, T. (2015). When readability meets computational linguistics: a new paradigm in readability. *Revue francaise de linguistique appliquée*, *XX*(2), 79-97.

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).

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).

Gala, N., François, T., & Fairon, C. (2013). Towards a French lexicon with difficulty measures: NLP helping to bridge the gap between traditional dictionaries and specialized lexicons. Dans *Proceedings of eLex-Electronic Lexicography 2013 (p. 132-151)*. https://hal.archives-ouvertes.fr/hal-03194427/document

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

Goigoux, R. (2003). Quelques points de repère pour une didactique de la compréhension. *Langage & pratiques*, *31*, 51-60.

Gough, P. B. et 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., Louwerse, M. M. et 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., Marneffe, M.-C. de et 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

Guo, L., Crossley, S. A., & McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples : A comparison study. *Assessing Writing*, 18(3), 218-238. https://doi.org/10/gcpgkq

Ihaka, R. et Gentleman, R. (1996). R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, *5*, 299-314.

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

Judd, C. M., McClelland, G. H., & Ryan, C. S. (2011). *Data analysis : A model comparison approach, second edition.* Routledge.

Juhasz, B. J. et Rayner, K. (2006). The role of age of acquisition and word frequency in reading: Evidence from eye fixation durations. *Visual Cognition*, *13*(7-8), 846-863. https://doi.org/10/dznsng

Karegowda, A. G., Manjunath, A. S. et 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.

Kennison, S. M., & Gordon, P. C. (1997). Comprehending referential expressions during reading: Evidence from eye tracking. *Discourse Processes, 24*(2-3), 229-252. https://doi.org/10.1080/01638539709545014

Kim, H., Goryachev, S., Rosemblat, G., Browne, A., Keselman, A., & Zeng-Treitler, Q. (2007). Beyond surface characteristics: a new health text-specific readability measurement. In *AMIA Annual Symposium Proceedings* (Vol. 2007, p. 418). American Medical Informatics Association.

Kintsch, W. et 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. et 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

Kyle, K., Crossley, S. A. et McNamara, D. S. (2016). Construct validity in TOEFL iBT speaking tasks: Insights from natural language processing. *Language Testing*, *33*(3), 319-340. https://doi.org/10/gg5jds

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

Lim, J. (2019). An investigation of the text features of discrepantly-scored ESL essays: A mixed methods study. *Assessing Writing,* 39, 1-13. https://doi.org/10.1016/j.asw.2018.10.003

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, *15*(4), 474-496. https://doi.org/10.1075/ijcl.15.4.02lu

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. et 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). https://doi.org/10/ghp3zg

McNamara, D. S., Graesser, A. C. et 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).

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.

New, B., Pallier, C. et Ferrand, L. (2005). *Manuel de Lexique 3.* http://lexique.org/_documentation/Manuel_Lexique.3.pdf

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

O'Reilly, T. et 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

Rayner, K. et Duffy, S. A. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition*, *14*(3), 191-201. https://doi.org/10/c8s2wf

Rudis, B. (2019). *Package hyphenatr*. https://cran.r-project.org/web/packages/hyphenatr/index.html

Sherstinova, T., Ushakova, E. et Melnik, A. (2020, septembre). Measures of Syntactic Complexity and their Change over Time (the Case of Russian). Dans *2020 27th Conference of Open Innovations Association (FRUCT)* (p. 221-229).

Smith, D. R., Stenner, A. J., Horabin, I., & Smith, M. (1989). *The Lexile Scale in Theory and Practice.* MetaMetric.

Stanké, B., Le Mené, M., Rezzonico, S., Moreau, A., Dumais, C., Robidoux, J., Dault, C. et 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, May). 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* (pp. 4290-4297).

Szmrecsányi, B. (2004). On operationalizing syntactic complexity. Dans *Le poids des mots. Proceedings of the 7th International Conference on Textual Data Statistical Analysis.* Louvain-la-Neuve (Vol. 2, p. 1032-1039).

Taneja, S., Gupta, C., Goyal, K. et Gureja, D. (2014, février). 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)* (pp. 987-997).

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

Wright, B. D. et Linacre, J. M. (1994). *The Rasch model as a foundation for the Lexile Framework.*

Wu, X., Mauranen, A. et Lei, L. (2020). Syntactic complexity in English as a lingua franca academic writing. *Journal of English for Academic Purposes*, *43*. https://doi.org/10/ggffs6

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

3. Transition entre les articles 1 et 2

Le premier article a introduit ALSI, un outil innovant qui répond à un besoin en analyse automatique du texte en français québécois. Nous avons montré qu'ALSI pouvait produire une variété d'attributs linguistiques ayant une association statistique avec le niveau de difficulté du texte exprimé en années scolaires. Cependant, nous n'avons pas testé si ces attributs pouvaient estimer l'année scolaire d'un texte. Le deuxième article s'intéresse donc à l'estimation du niveau de difficulté du texte, et évalue de manière robuste la performance de modèles de classification appliqués aux attributs produits par ALSI.

4. Deuxième article :

L’estimation robuste de la difficulté de textes en français par le traitement automatique du langage naturel

Guillaume Loignon

Université de Montréal

Nathalie Loye

Université de Montréal

Contribution des auteurs :

Loignon a conçu et effectué les analyses, et a rédigé la majorité de l’article.

Loye a fait d’importantes suggestions méthodologiques et a apporté un grand nombre de révisions au texte de l’article.

Résumé

Les avancées conjointes de la linguistique computationnelle et de l'apprentissage supervisé ont ouvert un terrain fertile pour la recherche en classification du texte par niveau de difficulté. Cependant, comme la recherche dans ce domaine fait généralement appel à des méthodologies hétérogènes, il est difficile d'évaluer de manière comparative les diverses approches afin d'identifier les meilleures pratiques. Cet article s'intéresse à la mesure robuste de la performance de modèles d'apprentissage supervisé classifiant le texte par niveau de difficulté. Le corpus utilisé comptait 902 textes en français analysés avec ALSI, un outil de traitement automatique dont nous mettons à l'épreuve la capacité à caractériser la difficulté du texte. Les modèles de classification étaient la régression multinomiale et SVM, et ont été appliqués par deux procédures : validation croisée répétée par *bootstrap*, et généralisation à de nouveaux textes. Les résultats montrent que les deux modèles ont obtenu des performances égales ou supérieures à celles des études comparables que nous résumons. Notre étude contribue ainsi au développement de méthodologies robustes de classification de textes par apprentissage supervisé, tout en soutenant la validité de l'outil ALSI pour l'extraction d'attributs linguistiques.

Mots-clés: classification du texte, lisibilité, complexité linguistique, français, apprentissage supervisé, validation croisée, méthodologies robustes

Abstract

Joint advances in computational linguistics and statistical learning have opened fertile ground for research in text classification by difficulty level. However, because research in this area uses heterogeneous methodologies, it is difficult to comparatively evaluate the performance of different approaches to identify best practices. This paper focuses on the application of robust methods for text corpus classification and performance estimation. The corpus used consisted of 902 French texts analyzed with ALSI, an automatic processing tool whose capacity to characterize linguistic complexity is put to the test. Text difficulty was estimated through two procedures (bootstrapped cross validation and generalization to new texts), using two classification models (multinomial regression and SVM) and two sets of attributes. Our study thus contributes to the development of robust methodologies for text classification by supervised learning, while supporting the validity of the ALSI tool for linguistic attribute extraction.

Keywords: text classification, readability, linguistic complexity, French, supervised learning, cross-validation, robust methodology

Thèse de doctorat

La difficulté des textes utilisés en classe est-elle appropriée ? Qu'est-ce qui rend un texte difficile ? La difficulté du texte est un enjeu important en éducation, où le texte occupe une place centrale en apprentissage et en évaluation. Historiquement, on estimait la difficulté du texte à l'aide de formules de lisibilité, typiquement des équations linéaires dont les variables sont des attributs du texte relativement faciles à mesurer (par ex. : la longueur des phrases et des mots.) À la fin des années 1980 sont apparus des outils informatiques comme *Lexile* et *ATOS,* qui extraient automatiquement des attributs textuels simples, puis estiment un score de difficulté pour le texte. Le score obtenu peut ensuite être converti en niveau scolaire en consultant une table. Bien que massivement utilisés en contexte anglo-saxon, entre autres pour situer des ouvrages dans la progression des années scolaires, ce type d'outil est critiqué en raison d'une conceptualisation réductrice de la difficulté du texte (Balyan et al., 2020; François, 2015). De nouveaux outils ont été proposés dans l'objectif de modéliser la difficulté du texte d'une manière qui serait davantage conforme avec les avancées en psycholinguistique et en linguistique computationnelle. Nous retenons à titre d'exemple la contribution importante des travaux de McNamara à l'opérationnalisation de la complexité linguistique (Graesser et al., 2014; McNamara et Graesser, 2011; McNamara et al., 2012), l'amélioration des méthodes d'extraction automatisée de structures syntaxiques (Lu, 2010; Straka et al., 2016; Welbers et al., 2020; Wu et al., 2020), et la publication de nouveaux lexiques de référence pour la langue française (Lété et al., 2004; New et al., 2005; Stanké et al., 2019). Du côté de la théorie de la mesure, l'estimation du niveau de difficulté du texte a profité du perfectionnement de modèles de classification, tels les séparateurs à vaste marge (SVM), et de l'intégration d'éléments tirés du domaine de l'intelligence artificielle (Tong et Koller, 2001). Un modèle de classification est, dans ce

contexte, un ensemble d'opérations effectué sur les attributs du texte et dont le résultat est la *classe* estimée du texte, c'est-à-dire un niveau de difficulté, souvent exprimé en années scolaires.

4.1.1. Estimer la performance d'un modèle de classification

Les études de TALN portant sur la classification du texte s'inscrivent généralement dans une approche d'apprentissage supervisé : un modèle de classification « apprend » les caractéristiques saillantes d'un ensemble de textes dont la classification est connue; ce modèle peut ensuite être appliqué à de nouveaux textes pour les classifier automatiquement (pour un survol, voir Kowsari et al., 2019). Dans ce contexte, l'entrainement du modèle désigne l'action de faire converger les paramètres du modèle vers des valeurs qui optimisent la classification, dans l'optique d'obtenir un modèle ayant une bonne capacité à classifier adéquatement des textes n'ayant pas été « vus » lors de la phase d'entrainement. La performance du modèle est évaluée en comparant les classes estimées aux classes associées aux textes (par ex. : l'année scolaire du manuel dont le texte est tiré). Différentes mesures témoignent de la qualité de la classification, comme la proportion de textes rangés dans la classe réelle ou la corrélation entre les classes estimées et réelles. Le Tableau 1 présente des indicateurs couramment utilisés pour quantifier la performance du modèle de classification. Plus un modèle d'apprentissage supervisé est complexe, mieux il performera sur ses données d'entrainement, et moins cette performance se généralisera à de nouvelles données (Friedman et al., 2001). L'entrainement du modèle implique donc une forme de compromis entre performance et généralisabilité.

Tableau 1

Indicateurs de performance courants pour la classification du texte

Indicateur	Définition
Coefficient rho de Spearman	Statistique non-paramétrique quantifiant la force de l'association entre deux variables, typiquement la classe réelle et la classe estimée; sa

	robustesse à la non-normalité de la distribution en fait un indicateur de choix pour exprimer la qualité d'une classification (Nelson et al., 2012).
Justesse	Proportion de textes rangés dans leur classe réelle (justesse exacte) ou dans la classe réelle ou adjacente (justesse adjacente).
Écart absolu moyen	Moyenne de la valeur absolue des erreurs de mesures (différence entre la classe réelle et la classe estimée). Une valeur plus faible indique une meilleure classification.
Écart quadratique moyen	Alternative à l'écart absolu moyen pour quantifier l'erreur de mesure, calculée selon la formule suivante, *e* et *r* étant la classe estimée et réelle : $EQM = \sqrt{moyenne(e - r^2)}$. Une valeur plus faible indique une meilleure classification.

L'erreur de généralisation du modèle, souvent indiquée par l'erreur quadratique moyenne entre la classe réelle et la classe estimée, est la tendance à mal classifier des observations dont la classe est inconnue du modèle. Une stratégie courante pour estimer l'erreur de généralisation est la validation croisée à *k* blocs (*k-fold cross-validation*). On partitionne d'abord le corpus aléatoirement en *k* échantillons appelés blocs. Le modèle est entrainé sur *k* - 1 blocs, puis validé sur le bloc mis de côté; on répète ensuite la procédure *k* fois en changeant le bloc de validation (voir Figure 1). Les indicateurs de performance sont ensuite calculés à partir de la série de classifications formée en combinant les *k* blocs de validation.

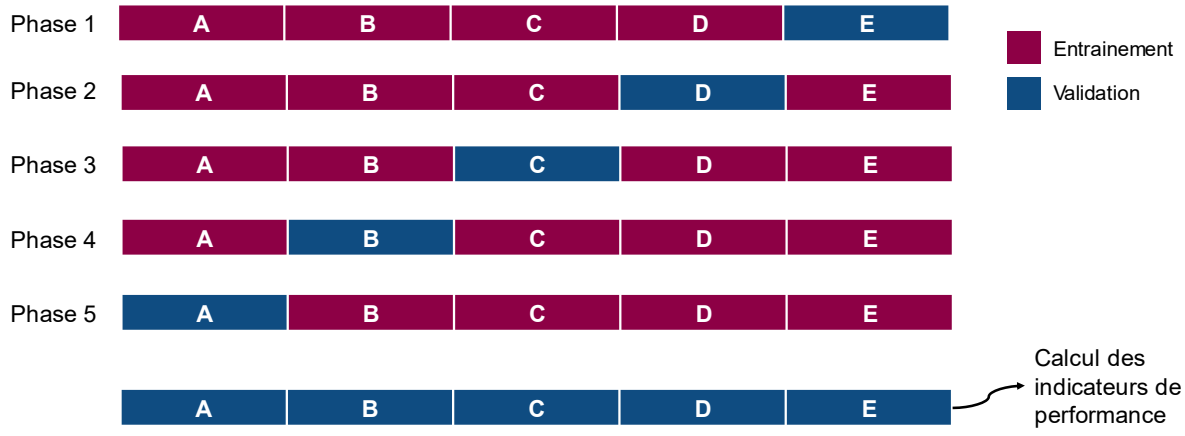


Figure 1. Schéma illustrant une procédure de validation croisée à 5 blocs.

Des travaux théoriques et études de simulation ont toutefois montré que les résultats en VC dépendent en partie du hasard de l'allocation des observations entre les *k* blocs; si une expérience de classification avec VC est répétée, la performance peut varier parfois de manière importante (Bengio et Grandvalet, 2008; Nadeau et Bengio, 2003; Varma et Simon, 2006). Pour stabiliser l'estimation, plusieurs auteurs proposent justement de répéter la procédure VC et de rapporter l'intervalle obtenu (Kim, 2009; Krstajic et al., 2014). L'estimation de la performance d'un modèle de classification peut en outre être biaisée en raison d'un nombre insuffisant d'observations, ou d'une distribution inégale des observations entre les classes (Figueroa et al., 2012; Fu et al., 2005).

4.1.2. État de l'art

Nous présentons dans ce qui suit un état de l'art de la classification par niveau de difficulté de corpus français, et montrons comment des biais d'estimation de la performance peuvent affecter ce type d'étude. Nous avons identifié six publications traitant de la classification par niveau de difficulté d'un corpus de langue française et qui rapportaient une estimation de la généralisabilité du modèle. Le Tableau 1 présente les caractéristiques saillantes de ces études. Nous employons les coefficients de corrélation comme estimateur de la performance puisque ce type de mesure est commune aux publications identifiées en plus d'être insensible au nombre de classes; l'équivalence des coefficients de Spearman (*r*) et de Pearson (*rs*) est assumée bien que le *r* tende à être légèrement plus élevé pour ce type de données (Arndt et al., 1999).

Tableau 2

État de l'art en classification du texte en langue française par niveau de difficulté

#	Référence	Corpus	Procédure	Modèle de classification	Attributs	Corrélation
1	Collins-Thompson et Callan (2004)	$N = 189$; $k = 5$	VC à 10 blocs	Rég. multinomiale	4	$r = 0,64$

2	François (2009)	N = 440; k = 9	VC à 10 blocs	Rég. multinomiale	10	r = 0,77
				Rég. ordinale	10	r = 0,78
3	François et Fairon (2012)	N = 1852; k = 6	VC à 10 blocs	SVM	46	rs = 0,73
				SVM	8	rs = 0,73
				Rég. ordinale	4	rs = 0,71
				Rég. multinomiale	8	rs = 0,70
4	François et Miltsakaki (2012)	N = 408; k = 6	VC à 10 blocs	Rég. linéaire	17	r = 0,75
				SVM	20	r = 0,68
5	François (2014)	N = 300; k = 6	*Bootstrap*	Rég. ordinale	2	r = 0,62 *
		N = 120 (entrainement sur 300); k = 6	Généralisation			r = 0,64 *
		N = 300; k = 6; déséquilibré	*Bootstrap*			r = 0,65 *
		N = 122 (entrainement sur 300); k = 6	Généralisation			r = 0,66 *
6	Dascalu et al. (2014)	N = 200; k = 3	VC à 3 blocs	SVM	54	rs = 0,63

Note. N est la taille du corpus, *k* le nombre de classes. *r* indique un coefficient de Pearson, *rs* indique un coefficient de Spearman. Les corrélations indiquent l'association statistique entre la classe réelle et la classe estimée. * indique un coefficient rapporté comme R^2 que nous avons converti en faisant la racine carrée de la valeur rapportée. Pour Dascalu et al. (2014), nous avons calculé le *rs* à partir de la matrice de confusion rapportée.

Les 6 publications retenues étaient tirées d'actes de conférence. Les résultats portaient sur des corpus de 120 à 1852 textes, répartis entre 3 à 9 classes. Les modèles de classification les plus utilisés étaient la régression multinomiale et SVM, et le nombre d'attributs utilisés variant entre 2 et 54. La VC à 10 blocs était la procédure la plus populaire; l'étude 5 a toutefois comparé les performances estimées par *bootstrap* et par généralisation (entrainement sur 300 textes puis validation sur environ 120 textes). Les 14 résultats rapportés au Tableau 2 décrivent une plage de coefficients de corrélation allant de r = 0,62 à r = 0,78. Ces performances se comparent assez favorablement aux résultats obtenus en généralisation par Nelson et al. (2012), qui ont estimé la difficulté de 683 textes en langue anglaise à l'aide de cinq outils commerciaux, dont *Lexile*, obtenant des indices de corrélation de Pearson allant de r = 0,46 à r = 0,78.

Toutefois, les corrélations du Tableau 2 devraient être comparées avec prudence (entre elles ou avec les résultats d'autres études) en raison des disparités entre les corpus et les méthodologies, et de la présence possible de biais d'estimation. Les études 2 et 3 ont retiré, après

avoir performé certaines analyses quantitatives, des textes ou types de textes jugés difficiles à classifier. Les études 1, 5 et 6 rapportent des résultats calculés sur une taille d'échantillon près ou inférieure au seuil minimal de 200 observations suggéré par Figueroa et al (2012). Concernant l'équivalence des effectifs (nombre équivalent de textes par niveau de difficulté), les études 1 et 6 ne spécifient pas comment le corpus était distribué. L'étude 3 a employé un corpus ayant des effectifs fortement déséquilibrés, ce qui peut biaiser l'estimation de la performance en faveur des classes ayant les plus forts effectifs. Enfin, les études ayant eu recours à la VC n'ont pas rapporté comment les indicateurs de performance auraient pu varier si la procédure était répétée.

4.1.3. La présente étude

L'état de l'art montre le caractère prometteur des techniques d'apprentissage supervisé pour classifier des textes de langue française par niveau de difficulté. Vu la diversité des méthodologies et la présence de certaines lacunes, comme des corpus de petite taille ou dont les classes ont été altérées de manière *post hoc*, il est cependant difficile il est difficile d'évaluer rigoureusement la performance rapportée. Nous remarquons également le manque de travaux en apprentissage supervisé portant sur des corpus en français québécois.

La présente étude a comme objectif d'évaluer, de manière robuste et dans des conditions variées, la capacité d'un système d'apprentissage supervisé à classifier un corpus en français québécois en fonction de la difficulté, exprimée en années scolaires. Afin de rejoindre cet objectif, nous avons comparé la performance en classification de deux modèles (régression multinomiale et SVM) telle qu'estimée par une procédure de VC, puis par généralisation. Nous avons testé une sélection de 21 attributs linguistiques ainsi qu'une sous-sélection comptant 6 attributs. Les résultats sont mis en perspective en les comparant à l'état de l'art.

4.2. Méthodologie

4.2.1. Corpus utilisé

La présente étude porte sur un corpus de 902 textes distribués entre 11 niveaux scolaires, en prenant comme classe réelle l'année scolaire qui était initialement associée au texte (par ex. : l'année spécifiée par le manuel scolaire dont le texte est tiré). Le corpus utilisé a été créé en combinant 656 textes récupérés depuis la banque de textes ayant servi à l'élaboration de l'outil SATO-Calibrage (Daoust et al., 1996) et 246 autres textes tirés de ressources pédagogiques ou de manuels scolaires québécois. La composition du corpus, de même que les procédures d'extraction et de sélection des attributs linguistiques sont décrites en plus de détails dans un autre article (Loignon, soumis), nous en résumons ici les grandes lignes. Sur les 902 textes, 600 ont été sélectionnés aléatoirement pour constituer le sous-corpus d'entrainement, 302 ont été réservés pour le sous-corpus de test. Le Tableau 3 illustre la distribution des textes entre les sous-corpus et années scolaires. L'association de chaque texte avec une année scolaire était déterminée par le matériel lui-même, nous n'avons pas réévalué la classification.

Tableau 3

Répartition des textes entre les sous-corpus et les classes (niveaux scolaires)

Sous-corpus	1	2	3	4	5	6	7	8	9	10	11	TOTAL
ENTRAINEMENT	33	49	40	40	39	73	63	65	56	56	86	600
TEST	22	33	23	18	21	34	27	24	22	33	45	302

Les attributs des textes ont été extraits avec ALSI (pour Analyse Syntaxique et Lexicale Intégrée) (Loignon, soumis). ALSI utilise le système d'annotation *UDPipe* pour R (Straka et al., 2016); le modèle de langue française utilisé pour l'annotation était French-GSD 2.5 (Guillaume et al., 2019). Les structures syntaxiques sont extraites avec *rsyntax* pour R (Welbers et al., 2020). Dans sa version actuelle, ALSI utilise trois bases de données lexicales : Manulex (Lété et al.,

2004), qui est un lexique adapté au niveau primaire du système éducatif français; EQOL (Stanké et al., 2019), un lexique tiré d'ouvrages scolaires pour le niveau primaire québécois, et la liste orthographique produite pour le niveau primaire par le ministère de l'Éducation du Québec en 2013. Une fois extraits par ALSI, les attributs des textes sont représentés dans une matrice où chaque ligne représente un texte et chaque colonne un attribut, les cellules contenant la valeur numérique de l'attribut pour le texte, tel qu'illustré au Tableau 4.

Tableau 4

Exemple de matrice de corpus

Identifiant du texte	Classe	Longueur moy. des mots	Proportion de mots dans ÉQOL	Longueur moy. des phrases	…
P1001	1	4,77	0,94	9,23	…
P1002	1	5,01	1,00	4,00	…
P2001	2	5,25	0,97	6,00	…
P3001	3	5,14	0,87	8,23	…
P3002	3	5,21	0,90	10,92	…
…	…	…	…	…	…

Nous avons réutilisé la sélection des attributs établie par une expérimentation précédente (Loignon, soumis), qui se résume comme suit : (1) calcul du gain d'information (GI) pour chaque attribut et élimination des attributs dont le GI est nul; (2) élimination des attributs posant des problèmes de multicolinéarité; (3) élimination des combinaisons d'attributs dérivés des mêmes mesures linguistiques. Aux étapes 2 et 3, les attributs fautifs sont éliminés un par un, en essayant de préserver les attributs ayant un GI plus élevé. Sur la cinquantaine d'attributs lexicaux et syntaxiques extraits par ALSI, nous avons retenu 21 attributs qui semblaient avoir un bon potentiel pour estimer la difficulté du texte tout en évitant la multicolinéarité et le surajustement du modèle. Nous avons en outre retenu un sous-ensemble composé de six attributs, soit trois attributs par domaine linguistique (syntaxe et lexique).

Nous désignons le résultat de cette procédure comme la sélection complète d'attributs. Afin de tester les performances des modèles employés avec un nombre réduit d'attributs, nous avons également retenu un sous-ensemble d'attributs constitué du meilleur attribut (GI le plus élevé) par type - il s'agit des six types de la typologie décrite dans Loignon (soumis). Cette typologie divise les attributs caractérisant la difficulté intrinsèque du texte en deux domaines (attributs lexicaux et syntaxiques) subdivisés en trois strates de complexité (surface, intermédiaire et global), ce qui donne six catégories d'attributs.

4.2.2. Modèles de classification

Deux modèles de classification ont été entrainés et testés lors de la présente étude : RMN est une régression logistique multinomiale, et SVM est une classification par séparateurs à vaste marge. Ces modèles ont été choisis car ils étaient les plus utilisés selon l'état de l'art résumé au Tableau 2; une description plus complète des modèles choisis et d'autres algorithmes de classification du texte se trouve dans Aggarwal et Zhai (2012).

La régression multinomiale est une généralisation de la régression logistique pour des cas où l'issue est une variable catégorielle ayant plus de deux niveaux. Elle est fréquemment utilisée dans des études portant sur la classification de textes (Kowsari et al., 2019). Durant la phase d'apprentissage, les coefficients de l'équation de régression sont déterminés à partir d'un échantillon de texte. Pour classifier un nouveau texte, la fonction calcule la *probabilité conditionnelle* des classes, ou probabilité d'appartenir à chaque classe compte tenu de la valeur des attributs (Agresti, 2012; Hosmer et al., 2013). Nous avons suivi la pratique courante voulant que l'on range le texte dans la classe ayant la probabilité associée la plus élevée (Indurkhya et Damerau, 2010). L'implémentation logicielle était *multinom*, tirée de la bibliothèque *nnet* pour R (Ripley et al., 2016), qui optimise l'ajustement de la régression multinomiale par le truchement

de réseaux de neurones artificiels. Cette implémentation possède un paramètre d'ajustement *decay*, un terme de régularisation qui pénalise l'utilisation de coefficients élevés afin de mitiger le surajustement du modèle aux données d'entrainement.

La classification par séparateurs à vaste marge (ou machines à vecteurs de support) a retenu l'attention de plusieurs chercheurs s'intéressant à la classification de textes, notamment en raison de la capacité de ce modèle à intégrer de multiples attributs du texte tout en demeurant assez robuste à la surspécification (Indurkhya et Damerau, 2010). Pendant la phase d'apprentissage, chaque texte est représenté comme un point dans un espace à *k* dimensions, où *k* est le nombre d'attributs. SVM trouve les hyperplans qui divisent le mieux les points par classe, tout en maximisant la marge entre les groupes de points (Noble, 2006). Les hyperplans déterminent ainsi la classification de nouveaux points qui seraient ajoutés au même espace. L'implémentation logicielle utilisée était la méthode *svmLinear* de la bibliothèque *caret* pour R (Kuhn, 2009). SVM utilise un paramètre d'ajustement *C* ajustant la tolérance du modèle aux erreurs de classification; une valeur de C plus élevée permet généralement un meilleur ajustement du modèle aux données d'entrainement, mais risque de diminuer la performance en généralisation à de nouvelles données (Hsu et al., 2016).

4.2.3. Procédures

Les procédures sont illustrées à la Figure 2 et détaillée dans les sections qui suivent; elles reprennent plusieurs aspects des méthodologies d'apprentissage machine supervisé de Topçuoğlu et al. (2020) et de Taneja et al. (2014). L'intérêt de cette chaine de traitement des données est qu'elle reproduit deux méthodologies couramment employées en classification du texte, soit la validation croisée et la généralisation à de nouveaux textes.

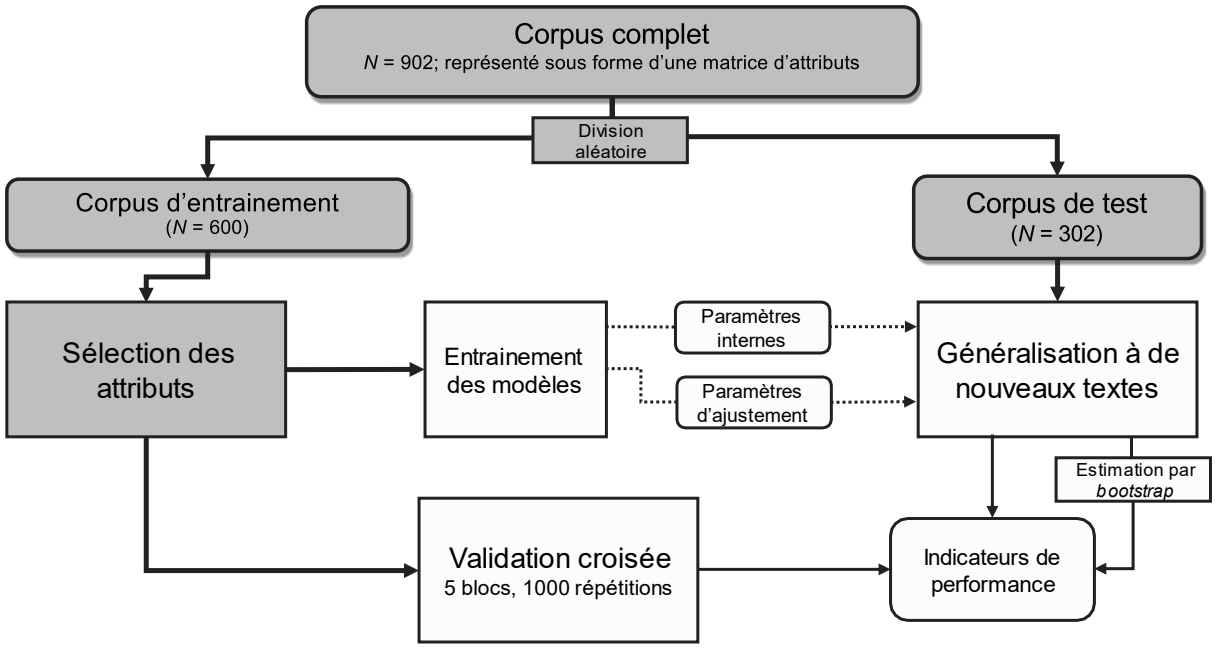


Figure 2. Chaine complète de traitement des données. Les régions en gris indiquent des ensembles de données et procédures qui sont détaillées dans une autre étude (Loignon, soumis).

Procédure d'entrainement et généralisation

Pour cette étude, l'entrainement du modèle fait référence aux opérations qui font converger les paramètres internes du modèle (par exemple, les coefficients) et ses paramètres d'ajustement (ou hyperparamètres) vers des valeurs optimales. Les valeurs optimales, pour notre cas d'application, signifient celles qui maximisent la corrélation de Spearman lorsque le modèle est appliqué à de nouveaux textes. RMN et SVM ont été ainsi été entrainés sur la sélection complète (21 variables) et réduite (6 variables) d'attributs du sous-corpus d'entrainement (600 textes). L'entrainement prenait la forme d'une validation croisée à 5 blocs, répétée pour tester une gamme de valeurs plausibles pour les paramètres d'ajustement. Les valeurs testées pour le paramètre de d'ajustement *decay* de RMN étaient {0,1, 0,01, 0,001, 0,0001, 0}, suivant une recommandation de Hsu et al. (2016). Les valeurs testées pour les paramètres d'ajustement *C* de

SVM étaient {1, 5, 10, 25}, ces valeurs étaient tirées de l'état de l'art. Les valeurs optimales ont été utilisées lors de la généralisation aux 302 textes n'ayant pas été « vus » par les modèles.

Procédure de validation croisée

Afin d'estimer de manière réaliste la performance des modèles à partir du sous-corpus d'entrainement lui-même, nous avons appliqué une procédure de validation croisée à 5 blocs, elle-même répétée 1000 fois. Pour notre étude, la VC s'implémentait donc comme suit : (1) permuter aléatoirement les données afin de s'assurer que la répartition des textes entre les blocs de VC varie; (2) appliquer une procédure de VC à 5 blocs; (3) répéter 1000 fois les étapes 1 et 2 en conservant toutes les classifications obtenues. La procédure de VC était appliquée avec le module *caret* pour R (Kuhn, 2009). Pour chaque répétition de VC, les paramètres d'ajustement (*decay* pour la RMN et *C* pour le SVM) étaient automatiquement sélectionnés à partir d'une gamme de valeurs possibles. Les valeurs testées étaient les mêmes que pour la procédure d'entrainement et de généralisation; la méthode de sélection des paramètres d'ajustement est détaillée dans la documentation du module *caret*.

4.2.4. Analyses statistiques

Les analyses visaient à évaluer la performance en fonction de la procédure (VC ou généralisation), du modèle de classification (RMN ou SVM) et de la sélection de variable employée (complète ou réduite). Les indicateurs de performance étaient ceux utilisés par François et Fairon (2012) : justesse exacte (proportion de textes rangés dans la bonne classe), justesse adjacente (proportion de textes rangés dans la bonne classe ou la classe adjacente), coefficient de Spearman, écart absolu moyen, et écart quadratique moyen.

Tous les indicateurs découlant de la procédure de VC ont été calculés d'abord pour chaque répétition, en combinant les classes estimées des 5 blocs de validation selon la procédure

recommandée par Forman et Sholz (2010); nous rapportons la médiane et l'intervalle de confiance à 95% (percentiles 2,5 et 97,5) des 1000 répétitions.

La procédure de généralisation produisait une seule série de résultats par combinaisons de modèle et sélection d'attributs. Pour les résultats obtenus par généralisation à 302 nouveaux textes, nous avons estimé les indicateurs de performance de manière robuste en faisant appel à une technique de *bootstrap* appliquée de manière *post hoc*, une stratégie suggérée par Tsamardinos et al. (2018). Ce rééchantillonnage *bootstrap* a été répété 1000 fois, le tirage était effectué par méthode de probabilité inverse (Nahorniak et al., 2015) afin de mitiger le biais de déséquilibre des classes, nous avons calculé la médiane et l'intervalle de confiance à 95% à-travers les 1000 répétitions. La méthode vise ainsi à donner un aperçu de la performance sur des corpus équilibrés qui seraient tirés de la même population. Pour qualifier plus précisément la classification obtenue par généralisation, nous avons produit des matrices de confusion détaillant la répartition des textes selon leur classe estimée et réelle, et calculé l'écart quadratique moyen par classe.

Le niveau de la chance, utilisé comme référence, a été calculé en classifiant au hasard les textes de l'échantillon; les valeurs *p* associées sont la proportion d'échantillons où l'indicateur affiche une performance inférieure à celle obtenue par une classification aléatoire des mêmes textes. Pour comparer les indicateurs par procédure ou par modèle, nous avons utilisé des tests de Wilcoxon sur la somme des rangs avec correction pour la continuité; nous rapportons la taille d'effet *r* associée, calculée avec la bibliothèque *rstatix* pour R (Kassambara, 2020). Les seuils d'interprétation du *r* étaient ceux proposés par Akoglu (2012) pour la recherche en psychologie, soit 0,1 pour faible, 0,4 pour modéré et 0,7 pour un effet de forte ampleur.

Afin de mettre les résultats en relation avec l'état de l'art, nous avons comparé les indices de corrélation obtenus par la présente étude à la gamme de valeurs rapportées par les études de l'état de l'art, résumées au Tableau 2, et avec les résultats de Nelson et al. (2012).

4.3. Résultats

Nous présentons dans cette section d'abord les résultats de l'entrainement des modèles et de la classification par validation croisée sur le sous-corpus d'entrainement ($N = 600$), puis de la généralisation des modèles au sous-corpus de test ($N = 302$).

4.3.1. Entrainement des modèles

Les modèles ont été entrainés sur 600 textes afin d'obtenir les paramètres optimaux en vue d'une généralisation à 302 nouveaux textes. Nous rapportons ici les valeurs obtenues pour les paramètres d'ajustement (hyperparamètres). La valeur optimale du paramètre d'ajustement *decay* de RMN était de 0,1 pour la sélection complète (21 attributs) et de 0,0001 pour la sélection réduite (6 attributs). La valeur optimale du paramètre d'ajustement *C* de SVM était de 1 avec les deux sélections d'attributs.

4.3.2. Procédure de VC répétée

Nous avons testé les modèles de classification sur le sous-corpus d'entrainement en appliquant une procédure de VC (validation croisée à 5 blocs sur 600 textes, répétée 1000 fois). Le Tableau 5 présente les résultats de la VC répétée, par modèle de classification (RMN ou SVM), et par ensemble d'attributs (sélection complète de 21 attributs ou sélection réduite de 6 attributs). Chaque répétition VC prenait environ 6 secondes sur un ordinateur Intel Core i5-2500 cadencé à 3,30GHz et muni de 12 Go de mémoire vive.

Tableau 5

Performance en VC

Modèle	Sélec. d'attrib.	Justesse		*rs*	EAM	EQM
		Exacte	Adjacente			
RMN	Complète	**0,34**	0,63	0,79	1,35	1,93
		[0,31, 0,36]	[0,61, 0,65]	[0,77, 0,80]	[1,29, 1,40]	[1,87, 2,00]
	Réduite	0,33	0,61	0,77	1,46	2,10
		[0,31, 0,35]	[0,57, 0,63]	[0,75, 0,78]	[1,40, 1,58]	[2,03, 2,22]
SVM	Complète	0,28	**0,64**	**0,79**	**1,28**	**1,74**
		[0,26, 0,30]	**[0,61, 0,65]**	**[0,78, 0,80]**	**[1,24, 1,33]**	**[1,68, 1,79]**
	Réduite	0,26	0,60	0,76	1,37	1,81
		[0,24, 0,28]	[0,58, 0,62]	[0,75, 0,77]	[1,33, 1,41]	[1,77, 1,85]
Chance (référence)	--	0,10	0,26	0	3,55	4,39
		[0,07, 0,12]	[0,23, 0,30]	[-0,08, 0,09]	[3,34, 3,72]	[4,19, 4,55]

Note. Médiane et l'intervalle de confiance à 95% pour 1000 répétitions d'une validation croisée à 5 blocs. Les valeurs les plus performantes sont indiquées en caractères gras. *rs* indique le coefficient de corrélation rho de Spearman. EAM est l'erreur absolue moyenne, EQM est l'erreur quadratique moyenne.

Avec sélection complète ou réduite d'attributs, les deux modèles ont mieux performé que le niveau de la chance pour l'ensemble des indicateurs rapportés dans ce tableau, $p < 0{,}001$. Les différences observées entre les résultats VC de RMN et SVM ont été examinée par des tests de Wilcoxon, et étaient statistiquement significatives selon un seuil $p < 0{,}001$; le détail des comparaisons se trouve au Tableau 8 en annexe de cet article.

Avec une sélection complète de variable, RMN parvenait ainsi à classifier environ le tiers des textes dans l'année scolaire exacte (*JE = 0,34)*, et près du deux-tiers des textes dans l'année exacte ou adjacente (*JA* = 0,63); la Figure 3 illustre la justesse exacte de RMN calculée pour les 1000 répétitions. SVM avait une justesse exacte (*JE* = 0,28) inférieure à RMN. La justesse adjacente de SVM était plus performante (*JA* = 0,64) que celle de RMN (*JA* = 0,63), la taille d'effet associée était de *r* = 0,29, ce qui correspond à un effet de faible ampleur selon les seuils sélectionnés pour cette étude.

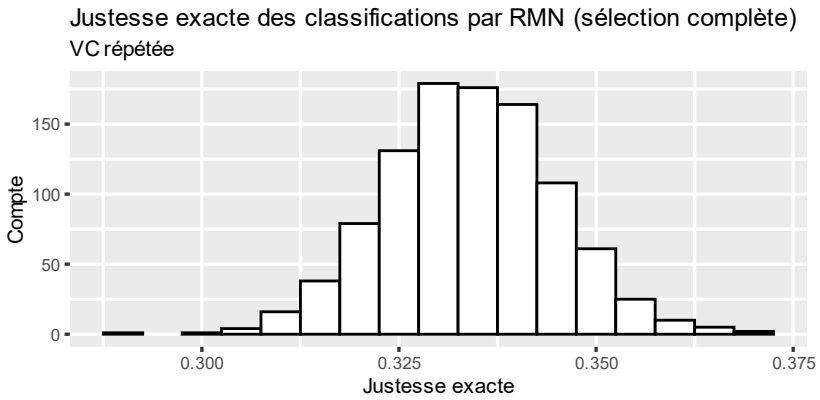


Figure 3. Histogramme de la justesse exacte obtenue par 1000 répétitions de validation croisée en employant le modèle de classification RMN avec une sélection de 21 attributs.

La corrélation de Spearman de SVM (*rs* = 0,79 [0,78, 0,8]) était supérieure à RMN (*rs* = 0,79 [0,77, 0,8]), indiquant une meilleure performance de SVM, la taille d'effet associée était de *r* = 0,27. Les mesures d'erreurs indiquaient que les classifications de SVM (*EAM* = 1,28; *EQM* = 1,74) étaient généralement plus près de leur classe réelle, comparativement à RMN (*EAM* = 1,35; *EQM* = 1,93), ces différences étaient associées des tailles d'effet de forte ampleur.

Avec une sélection réduite d'attributs, la justesse exacte était supérieure pour RMN (*JE* = 0,33) comparativement à SVM (*JE* = 0,26), *r* = 0,87. La justesse adjacente était également meilleure pour RMN (*JA* = 0,61) que pour SVM (*JA* = 0,60), *r* = 0,3. La corrélation de Spearman était également plus performante pour RMN (*rs* = 0,77) que SVM (*rs* = 0,76), *r* = 0,5. À l'égard des mesures d'erreur, SVM (*EAM* = 1,37; *EQM* = 1,81) était cependant plus performant que RMN (*EAM* = 1,46; *EQM* = 2,1), avec des tailles d'effet associées de forte ampleur.

4.3.3. Généralisation à de nouveaux textes

Afin de vérifier la capacité de généralisation des modèles à de nouveaux textes, nous avons testé les modèles sur le sous-corpus de test, lequel n'était pas impliqué dans la phase d'entrainement des modèles. Les indicateurs de performance calculés directement depuis les résultats de la procédure de généralisation, sont montrés au Tableau 7 en annexe de l'article.

Pour fins de vérification, les matrices de confusion montrant la classification par procédure de généralisation, ainsi que les écarts quadratiques moyens par classe, sont données aux Tableaux 10 à 13 en annexe.

Pour tester statistiquement les différences de performance tout en mitigeant les biais liés au déséquilibre des classes, nous avons estimé les indicateurs avec une méthode de *bootstrap* (tirage par probabilité inverse, 1000 répétitions). Les résultats sont montrés au Tableau 6.

Tableau 6

Performance en généralisation (estimation robuste par bootstrap)

Modèle	Sélec. d'attrib.	Justesse		*rs*	EAM	EQM
		Exacte	Adjacente			
RMN	Complète	0,31 [0,26, 0,36]	0,70 [0,65, 0,75]	**0,87 [0,83, 0,90]**	1,20 [1,07, 1,34]	1,71 [1,53, 1,90]
	Réduite	0,31 [0,26, 0,37]	0,62 [0,57, 0,67]	0,83 [0,78, 0,87]	1,41 [1,25, 1,56]	1,98 [1,79, 2,17]
SVM	Complète	**0,31 [0,26, 0,37]**	**0,73 [0,67, 0,78]**	**0,87 [0,83, 0,90]**	**1,09 [0,98, 1,22]**	**1,53 [1,37, 1,68]**
	Réduite	0,27 [0,23, 0,32]	0,60 [0,55, 0,66]	0,82 [0,76, 0,86]	1,34 [1,20, 1,46]	1,77 [1,63, 1,93]
Chance	--	0,09 [0,07, 0,13]	0,26 [0,21, 0,31]	0 [-0,10, 0,11]	3,62 [3,3, 3,94]	4,47 [4,13, 4,78]

Note. Les résultats portent sur la généralisation des modèles de classification à 302 nouveaux textes, et ont été estimés par méthode de bootstrap (B = 1000, n = 30 par classe). *rs* indique le coefficient de corrélation rho de Spearman, EAM est l'erreur absolue moyenne et EQM est l'erreur quadratique moyenne. Les valeurs les plus performantes sont indiquées en caractères gras.

Tous les indicateurs de performance estimés par *bootstrap* ont performé au-dessus du niveau de la chance d'une manière statistiquement significative considérant un seuil de $p <$ 0,001. Les résultats des tests de Wilcoxon répétés, utilisés pour vérifier si les différences observées entre les modèles étaient statistiquement significatives, se trouvent au Tableau 8 en annexe.

Avec la sélection complète de 21 attributs, la justesse exacte de SVM en généralisation (0,31 [0,26, 0,37]) était supérieure à celle de RMN (0,31 [0,26, 0,36]), $p = 0{,}002$, $r = 0{,}07$. La justesse adjacente de SVM (0,73) était également supérieure à celle de RMN (0,70), $p < 0{,}001$, $r = 0{,}44$. Les deux modèles étaient équivalents à l'égard de la corrélation de Spearman, $p = 0{,}106$. Les mesures d'erreurs indiquaient une meilleure performance pour SVM (*EAM* = 1,09; *EQM* = 1,53) que pour RMN (*EAM* =1,20; *EQM* = 1,71), $p < 0{,}001$, avec des tailles d'effet associées d'ampleur modérée à forte.

Avec une sélection réduite d'attributs, RMN (*JE* = 0,31; *JA* = 0,62) était plus performant que SVM (*JE* = 0,27; *JA* = 0,60) à l'égard des mesures de justesse, $p < 0{,}001$, $r = 0{,}63$. Le coefficient de Spearman de RMN (*rs* = 0,83) était supérieur à celui de SVM (rs = 0,83), $p < 0{,}001$, $r = 0{,}31$. En ce qui concerne les mesures d'erreur, SVM (*EAM* = 1,34; *EQM* = 1,77) était plus performant que RMN (*EAM* = 1,41; *EQM* = 1,98), avec des tailles d'effet associées modérées à forte.

Nous avons investigué un résultat imprévu : une légère supériorité apparente de la justesse exacte pour le modèle RMN lorsqu'employé avec une sélection de 6 attributs (*JE* = 0,31 [0,26, 0,37]) comparativement au même modèle avec 21 attributs (*JE* = 0,31 [0,26, 0,36]). Cette différence était significative selon un test Wilcoxon, $p = 0{,}001$, avec une taille d'effet ($r = 0{,}07$) inférieure au seuil de $r = 0{,}1$ choisi pour un effet de faible ampleur.

4.3.4. Synthèse et comparaison avec l'état de l'art

Afin de vérifier si les résultats obtenus par VC étaient des indicateurs fiables de la performance en généralisation, nous avons comparé la performance obtenue par les deux procédures à l'aide de tests de Wilcoxon répétés dont les résultats sont décrits au Tableau 9, en annexe de cet article. La généralisation affichait une performance supérieure pour la plupart des

indicateurs, et ce dans la plupart des conditions. Cependant, lorsque le modèle RMN était employé avec une sélection complète d'attributs, la justesse exacte était supérieure pour la VC (*JE* = 0,34) comparativement aux résultats de généralisation (*JE* = 0,31), $p < 0{,}001$, $r = 0{,}5$. De même, avec une sélection réduite d'attributs, RMN affichait une justesse exacte supérieure lorsqu'appliqué par VC répétée (*JE* = 0,33), comparativement aux performances obtenues par généralisation (*JE* = 0,31), $p < 0{,}001$, $r = 0{,}38$. Enfin, considérant le modèle SVM employant une sélection réduite d'attributs, la différence entre les justesses adjacentes obtenues par les deux procédures étaient non statistiquement significative, $p = 0{,}99$.

La Figure 4 compare les performances obtenues par la présente étude à l'état de l'art dont le Tableau 2 a fait le résumé. Pour faciliter la comparaison et contrôler le biais lié à un nombre de classes variant entre les études, nous avons employé comme point de comparaison le coefficient de corrélation converti en *Z* de Fisher. Les plages de résultat correspondent aux intervalles à 95% selon la procédure d'estimation (VC ou généralisation), le modèle de classification (RMN ou SVM) et la sélection d'attributs (complète ou réduite). Dans l'ensemble, les résultats de VC étaient inférieurs aux résultats de généralisation. En validation croisée, RMN et SVM, lorsqu'employés avec 21 attributs, produisaient une corrélation entre la classe réelle atteignant un niveau équivalent ou supérieur à l'état de l'art (voir Tableau 2). Lorsqu'employés avec 6 attributs, RMN et SVM performaient à un niveau s'approchant de la borne supérieure de l'état de l'art. En généralisation, les performances de RMN et SVM, avec les deux sélections d'attributs, étaient comparables ou supérieures aux résultats que Nelson et al. (2012) ont obtenu à l'aide d'outils commerciaux appliqués à un corpus de langue anglaise.

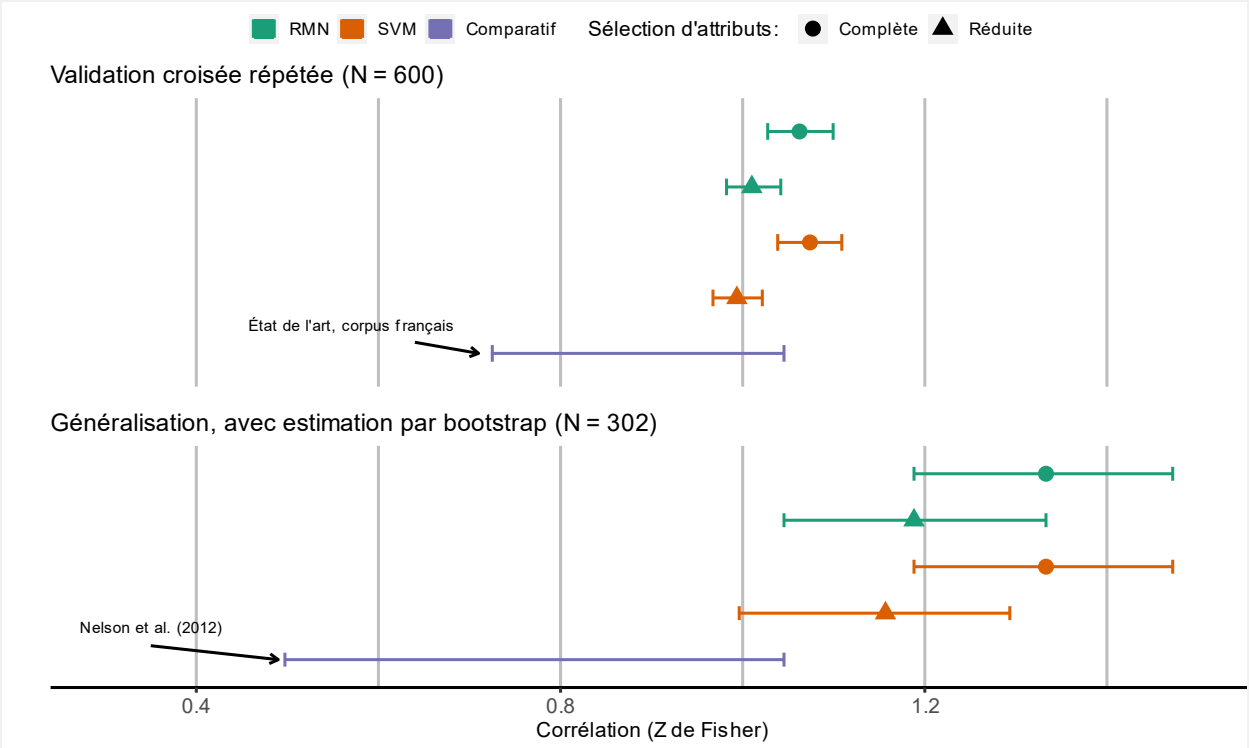


Figure 4. Sommaire des coefficients de corrélation obtenus par la présente étude selon la procédure, le modèle de classification et la sélection de variables. La valeur médiane est indiquée par la forme (cercle pour RMN et triangle pour SVM), les barres indiquent l'intervalle de confiance à 95%. Le trait mauve indique le niveau de référence; pour les résultats VC la référence est de la plage de corrélations rapportée par des études sur corpus en français (voir Tableau 2), pour les résultats de généralisation la référence est une plage de corrélations obtenues en appliquant des outils commerciaux à un corpus de langue anglaise (Nelson et al., 2012). Pour fins de comparaison, tous les coefficients de corrélation ont été normalisés en *Z* par transformé de Fisher.

4.4. Discussion

L'objectif principal de cette étude était d'évaluer rigoureusement la capacité d'un système d'apprentissage supervisé à identifier le niveau de difficulté de textes en langue française. Les analyses statistiques consistaient à décrire et comparer la performance selon la procédure, le modèle de classification, et la sélection d'attributs. Trois conclusions principales se dégagent de nos analyses.

Premièrement, la procédure VC telle que l'avons implémentée a permis une évaluation assez transparente de la plage des résultats pouvant émerger de la validation croisée. Considérant le rho de Spearman, la VC produisait un patron de résultats similaire à la généralisation, mais

avec un biais pessimiste. Ce résultat est consistant avec les travaux théoriques sur la VC (Fushiki, 2011) et peut s'expliquer par le fait que la procédure VC entrainait les modèles sur 480 textes alors que la généralisation a pu profiter d'un entrainement sur 600 textes. L'estimation par *bootstrap* des résultats de généralisation a produit des intervalles de confiance plus larges que ceux observés pour la procédure de VC répétée, une différence prévisible puisque la VC réutilisait le même ensemble de texte à chaque répétition et modifiait seulement la composition des blocs. La justesse exacte estimée par VC avait néanmoins un intervalle de confiance d'environ 5 points de pourcentage, soit ±7% à ±10% de la médiane selon le modèle et la sélection d'attributs, ce qui pourrait représenter un écart important dans certains cas d'application où plus de précision est nécessaire. D'autres travaux seraient requis afin d'étudier comment la stabilité de l'estimation VC est affectée par des facteurs tels nombre de blocs de VC, la taille du corpus et le nombre de classes.

Deuxièmement, nos résultats suggèrent que RMN et SVM ont un profil de performance distinct en classification du texte. Bien que SVM ait été généralement plus performant, RMN semblait plus robuste à la réduction du nombre d'attributs, autant en VC qu'en généralisation. Par ailleurs, RMN tendait à produire des classifications ayant une meilleure justesse, suggérant que RMN tend à faire moins d'erreurs, mais fait des erreurs plus importantes que SVM. Une implication pour la pratique est que le choix du modèle de classification devrait considérer le corpus d'entrainement, le profil de performance recherché et le coût associé aux classification aberrantes. En ce sens, une limite de cette étude est que nous n'avons pas vérifié directement si la performance était stable à-travers les classes, c'est-à-dire si les modèles classifiaient mieux les textes de certaines années scolaires.

Troisièmement, nos modèles de classification ont généralement performé à un niveau équivalent ou supérieur à l'état de l'art. Si on exclut le rôle des modèles de classification, ces performances pourraient s'expliquer par la sélection et la nature des attributs, qui rendraient mieux compte de la complexité linguistique des textes. Cette explication est appuyée par le fait que nos modèles avaient une performance équivalente ou supérieure tout en employant relativement peu de variables. Une explication connexe est que les attributs, tout en étant de nature similaire à ce qu'on retrouve dans l'état de l'art, étaient extraits de manière plus fidèle et contenaient moins de bruit statistique que dans les études rapportées. Il n'est pas exclu que les algorithmes auxquels l'outil ALSI fait appel, notamment *UDPipe*, annotent le texte avec plus de précision que leurs équivalents datant d'une dizaine d'années. La performance obtenue pourrait également s'expliquer par la nature du corpus. Bien que le corpus provînt de sources diverses couvrant plus de 30 ans, et que les classes n'aient pas été altérées, il demeure possible que les valeurs des attributs aient été faciles à diviser par année scolaire, s'inscrivant dans la progression d'un programme d'enseignement national.

Des travaux ultérieurs pourraient investiguer les raisons de la performance observée en testant d'autres analyseurs sur le corpus utilisé pour cette étude, ou en répétant l'expérience sur des textes dont l'échelle de difficulté décrit un autre type de progression que des années scolaires, telle l'échelle québécoise des niveaux de compétence en français des personnes immigrantes adultes (Desbiens et al., 2011). Concernant plus spécifiquement la comparaison avec les résultats de Nelson et al. (2012), celle-ci est limitée par le fait que ces travaux ont employé des outils commerciaux sur des textes possiblement différents de ceux employés pour développer l'outil, tandis que nos données d'entrainement et de test ont été produites par division aléatoire d'un corpus initial et avait donc une composition assez homogène. Une autre limite de

cette étude est que nous n'avons pas examiné directement si les performances obtenues étaient constantes à-travers les classes. Enfin, la portée de nos résultats dépend en grande partie de la composition du corpus, et de sa représentativité des textes utilisés dans le système scolaire primaire et secondaire québécois.

4.5. Conclusions

La présente étude portait sur l'estimation de la complexité linguistique. Ce sujet touche à une variété de domaines dont la psycholinguistique et l'apprentissage machine, et a des retombées évidentes dans le domaine de l'éducation puisqu'il pourrait contribuer, notamment, à identifier quels textes proposer aux élèves afin de mieux guider leurs apprentissages. L'objectif était d'évaluer de manière robuste l'estimation du niveau de difficulté (année scolaire) du texte par deux modèles de classification et deux sélections d'attributs. La performance a été mesurée par VC répétée en rapportant les intervalles de confiance, puis confirmée par un essai de généralisation à de nouveaux textes. Dans l'ensemble, SVM affichait une légère supériorité comme modèle de classification, bien que RMN performât mieux dans certaines conditions. En considérant les coefficients de corrélation, les deux modèles de classification tendaient à performer de manière similaire ou supérieure au niveau de l'art. Ce résultat est important car il montre qu'une méthodologie robuste ne nécessite pas de renoncer à la performance, et appuie la validité d'ALSI comme outil automatisant la mesure d'attributs qui caractérisent la complexité linguistique. En plus de valider la qualité des attributs extraits par l'outil ALSI, notre étude a montré qu'il est possible d'estimer la difficulté de textes scolaires de langue française dont les classes ont été utilisées telles quelles, c'est-à-dire telles que spécifiées par la ressource pédagogique dont le texte est issu. Nous avons validé la capacité de classification des sélections d'attributs utilisées dans une étude précédente (Loignon, soumis) et proposé une méthodologie

robuste qui pourrait servir de base à des travaux similaires de classification de corpus par niveau de difficulté. Une analyse plus approfondie nous permettrait de quantifier la contribution individuelle des attributs, et d'identifier d'autres facteurs qui pourraient expliquer les performances obtenues par cette étude.

4.6. Annexes du deuxième article

Tableau 7

Résultats de la généralisation

Modèle	Sélec. d'attrib.	Justesse		*rs*	EAM	EQM
		Exacte	Adjacente			
RMN	Complète	0,35	0,72	0,86	1,15	1,69
	Réduite	0,34	0,63	0,81	1,39	2,01
SVM	Complète	0,31	0,71	0,85	1,14	1,59
	Réduite	0,27	0,58	0,79	1,38	1,84

Note. Indicateurs de performance calculés pour 302 textes divisés entre 11 années scolaires.

Tableau 8

Comparaison de la performance des modèles (RMN ou SVM)

Procédure	Sélection	Indicateur	Test de Wilcoxon[a]			Supériorité[b]
			W	***p***	***r***	
VC	Complète	*JE*	999580,5	<0,001	0,87	RMN
		JA	334911,5	<0,001	0,29	SVM
		rs	342487,0	<0,001	0,27	SVM
		EAM	959100,0	<0,001	0,80	SVM
		EQM	999950,0	<0,001	0,87	SVM
	Réduite	*JE*	1000000,0	<0,001	0,87	RMN
		JA	671415,0	<0,001	0,30	RMN
		rs	787371,0	<0,001	0,50	RMN
		EAM	995018,0	<0,001	0,86	SVM
		EQM	1000000,0	<0,001	0,87	SVM
Généralisation	Complète	*JE*	459912,0	0,002	0,07	SVM
		JA	244426,0	<0,001	0,44	SVM
		rs	479147,0	0,106	0,04	--
		EAM	867338,5	<0,001	0,64	SVM
		EQM	929012,5	<0,001	0,74	SVM
	Réduite	*JE*	861375,0	<0,001	0,63	RMN
		JA	686524,0	<0,001	0,32	RMN
		rs	681587,0	<0,001	0,31	RMN
		EAM	743834,0	<0,001	0,42	SVM
		EQM	955850,0	<0,001	0,79	SVM

Note. *JE* : justesse exacte; *JA* : justesse adjacente; *r* : coefficient rho de Spearman; *EAM* : erreur absolue moyenne; *EQM* : erreur quadratique moyenne. [a]Test de Wilcoxon sur la somme des rangs avec correction pour la continuité; nous rapportons la taille d'effet *r*. [b]Indique quel modèle était le plus performant pour cette condition.

Tableau 9

Comparaison de la performance selon la procédure (VC ou généralisation).

Sélection	Modèle	Indicateur	Test de Wilcoxon			Supériorité
			W	***P***	***r***	
Complète	RMN	*JE*	787241	<0,001	0,50	VC
		JA	6839	<0,001	0,85	Géné.
		rs	107	<0,001	0,87	Géné.
		EAM	970995	<0,001	0,82	Géné.
		EQM	984320	<0,001	0,84	Géné.
	SVM	*JE*	123640	<0,001	0,65	Géné.
		JA	279	<0,001	0,87	Géné.
		rs	403	<0,001	0,87	Géné.
		EAM	997936	<0,001	0,86	Géné.
		EQM	992216	<0,001	0,85	Géné.
Réduite	RMN	*JE*	719508	<0,001	0,38	VC
		JA	335632	<0,001	0,28	Géné.
		rs	6216	<0,001	0,86	Géné.
		EAM	763666	<0,001	0,46	Géné.
		EQM	875869	<0,001	0,65	Géné.
	SVM	*JE*	305972	<0,001	0,34	Géné.
		JA	499844	0,99	0,00	--
		rs	21397	<0,001	0,83	Géné.
		EAM	687256	<0,001	0,32	Géné.
		EQM	659483	<0,001	0,28	Géné.

Note. Voir annotations du Tableau 8.

Tableau 10

Matrice de confusion – MNR avec sélection complète

		Classe réelle											EQM
		1	2	3	4	5	6	7	8	9	10	11	
Classe estimée	1	**2**	1	0	0	0	0	0	0	0	0	0	1,19
	2	18	**27**	12	0	0	1	0	1	0	0	0	0,72
	3	1	3	**6**	4	2	1	0	0	0	0	0	1,25
	4	1	1	2	**4**	5	2	0	0	0	1	0	1,43
	5	0	1	1	4	**2**	1	0	0	1	0	0	2,07
	6	0	0	2	5	8	**18**	5	4	2	1	1	1,86
	7	0	0	0	1	2	5	**6**	5	2	1	1	2,22
	8	0	0	0	0	0	3	6	**6**	1	6	4	2,18
	9	0	0	0	0	0	0	2	0	**4**	2	4	1,97
	10	0	0	0	0	1	1	2	3	2	**5**	8	1,78
	11	0	0	0	0	1	2	6	5	10	17	**27**	1,50

Note. Classification de 302 textes. EQM indique l'erreur quadratique moyenne, calculée par classe. Les caractères gras indiquent une classification dans la classe réelle.

Tableau 11

Matrice de confusion – MNR avec sélection réduite

		Classe réelle											EQM
		1	2	3	4	5	6	7	8	9	10	11	
Classe estimée	1	**10**	12	2	0	0	0	0	0	0	0	0	0,74
	2	12	**17**	6	0	0	0	0	1	0	0	0	0,70
	3	0	4	**10**	4	2	2	0	0	0	0	0	1,10
	4	0	0	2	**6**	3	4	0	1	0	0	0	2,07
	5	0	0	3	1	**2**	1	2	0	0	1	0	2,00
	6	0	0	0	5	7	**17**	7	4	2	3	7	2,18
	7	0	0	0	0	4	2	**1**	0	1	2	1	2,97
	8	0	0	0	1	1	2	2	**5**	2	6	2	2,66
	9	0	0	0	0	1	1	1	0	**3**	3	3	1,83
	10	0	0	0	1	1	2	1	3	2	**7**	7	1,98
	11	0	0	0	0	0	3	13	10	12	11	**25**	2,25

Note. Classification de 302 textes. EQM indique l'erreur quadratique moyenne, calculée par classe. Les caractères gras indiquent une classification dans la classe réelle.

Tableau 12

Matrice de confusion – SVM avec sélection complète

		Classe réelle											EQM
		1	2	3	4	5	6	7	8	9	10	11	
Classe estimée	1	**12**	4	0	0	0	0	0	0	0	0	0	0,67
	2	10	**25**	12	0	0	1	0	1	0	0	0	0,89
	3	0	2	**4**	4	2	0	0	0	0	0	0	1,34
	4	0	1	3	**6**	1	1	0	0	0	0	0	1,55
	5	0	0	2	1	**3**	1	0	0	0	0	0	1,73
	6	0	1	2	5	9	**16**	3	4	2	1	3	1,46
	7	0	0	0	2	5	8	**5**	5	1	1	3	1,67
	8	0	0	0	0	0	4	8	**5**	6	7	7	1,77
	9	0	0	0	0	0	3	7	6	**7**	13	13	1,24
	10	0	0	0	0	1	0	4	3	6	**11**	19	1,41
	11	0	0	0	0	0	0	0	0	0	0	**0**	2,39

Note. Classification de 302 textes. EQM indique l'erreur quadratique moyenne, calculée par classe. Les caractères gras indiquent une classification dans la classe réelle.

Tableau 13

Matrice de confusion – SVM avec sélection réduite

		Classe réelle											EQM
		1	2	3	4	5	6	7	8	9	10	11	
Classe estimée	1	**9**	8	1	0	0	0	0	0	0	0	0	0,77
	2	13	**20**	8	0	0	0	0	1	0	0	0	0,63
	3	0	5	**8**	4	1	2	0	0	0	0	0	1,56
	4	0	0	0	**0**	0	0	0	0	0	0	0	1,93
	5	0	0	2	1	**2**	0	1	0	0	1	0	2,18
	6	0	0	4	11	8	**12**	3	5	2	4	4	1,89
	7	0	0	0	2	5	7	**6**	3	4	3	4	1,80
	8	0	0	0	0	3	4	1	**5**	1	3	4	1,77
	9	0	0	0	0	1	9	13	8	**12**	14	26	1,31
	10	0	0	0	0	1	0	3	2	3	**8**	7	2,07
	11	0	0	0	0	0	0	0	0	0	0	**0**	2,63

Note. Classification de 302 textes. EQM indique l'erreur quadratique moyenne, calculée par classe. Les caractères gras indiquent une classification dans la classe réelle.

4.7. Bibliographie

Aggarwal, C. C. et Zhai, C. (2012). A Survey of Text Classification Algorithms. Dans C. C. Aggarwal et C. Zhai (éd.), *Mining Text Data* (p. 163-222). Springer US. https://doi.org/10.1007/978-1-4614-3223-4_6

Agresti, A. (2012). *Categorical Data Analysis* (3rd edition). Wiley.

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

Arndt, S., Turvey, C. et Andreasen, N. C. (1999). Correlating and predicting psychiatric symptom ratings: Spearmans r versus Kendalls tau correlation. *Journal of Psychiatric Research*, *33*(2), 97-104. https://doi.org/10/fmz4w3

Balyan, R., McCarthy, K. S. et McNamara, D. S. (2020). Applying Natural Language Processing and Hierarchical Machine Learning Approaches to Text Difficulty Classification. *International Journal of Artificial Intelligence in Education*, *30*(3), 337-370. https://doi.org/10/gg484m

Cawley, G. C. et Talbot, N. L. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research*, *11*, 2079-2107.

Collins-Thompson, K. et Callan, J. (2005). Predicting reading difficulty with statistical language models. *Journal of the American Society for Information Science and Technology*, *56*(13), 1448-1462. https://doi.org/10/c82b4x

Collins-Thompson, K. et Callan, J. P. (2004). *A language modeling approach to predicting reading difficulty* (p. 193-200).

Cristianini, N. et Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press.

Dahlke, J. A. et Wiernik, B. M. (2019). psychmeta: An R package for psychometric meta-analysis. *Applied Psychological Measurement*, *43*(5), 415-416. https://doi.org/10/gfgt9t

Dascalu, M., Stavarache, L. L., Trausan-Matu, S., Dessus, P. et Bianco, M. (2014, novembre). *Reflecting Comprehension through French Textual Complexity Factors*. 2014 IEEE 26th International Conference on Tools with Artificial Intelligence (p. 615-619). https://doi.org/10/ghjqf5

Daoust, F., Laroche, L. et 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

Desbiens, D., Laurier, M. D., & Leroux, J. (2011). *Échelle québécoise des niveaux de compétence en français des personnes immigrantes adultes.* Gouvernement du Québec.

Field, A. P. (2001). Meta-analysis of correlation coefficients: a Monte Carlo comparison of fixed-and random-effects methods. *Psychological methods*, *6*(2), 161. https://doi.org/10/bwpvg4

Figueroa, R. L., Zeng-Treitler, Q., Kandula, S. et Ngo, L. H. (2012). Predicting sample size required for classification performance. *BMC Medical Informatics and Decision Making*, *12*(1), 8. https://doi.org/10/gb345p

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

Forman, G. et Scholz, M. (2010). Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. ACM SIGKDD Explorations Newsletter, 12(1), 49–57. https://doi.org/10.1145/1882471.1882479

François, T. (2009, avril). *Combining a Statistical Language Model with Logistic Regression to Predict the Lexical and Syntactic Difficulty of Texts for FFL*. Athens, Greece (p. 19-27). https://www.aclweb.org/anthology/E09-3003

François, T. (2015). When readability meets computational linguistics: a new paradigm in readability. *Revue Francaise de Linguistique Appliquee*, *Vol. XX*(2), 79-97. https://www.cairn.info/revue-francaise-de-linguistique-appliquee-2015-2-page-79.htm

François, T. et Fairon, C. (2012, juillet). *An "AI readability" Formula for French as a Foreign Language*. CoNLL-EMNLP 2012, Jeju Island, Korea (p. 466-477). https://www.aclweb.org/anthology/D12-1043

Fu, W., Carroll, R. et Wang, S. (2005). Estimating misclassification error with small samples via bootstrap cross-validation. *Bioinformatics (Oxford, England)*, *21*, 1979-86. https://doi.org/10/dqkdhx

Fushiki, T. (2011). Estimation of prediction error by using K-fold cross-validation. *Statistics and Computing*, 21(2), 137-146. http://doi.org/10.1007/s11222-009-9153-8

Graesser, A. C., McNamara, D. S., Cai, Z., Conley, M., Li, H. et Pennebaker, J. (2014). Coh-Metrix Measures Text Characteristics at Multiple Levels of Language and Discourse. *The Elementary School Journal*, *115*(2), 210-229. https://doi.org/10/f6qk6f

Guillaume, B., Marneffe, M.-C. de et 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, H. et Jiang, X. (2014). Overcome Support Vector Machine Diagnosis Overfitting. *Cancer Informatics*, *13s1*, CIN.S13875. https://doi.org/10/gjh4t2

Heilman, M., Collins-Thompson, K. et Eskenazi, M. (2008, juin). *An Analysis of Statistical Models and Features for Reading Difficulty Prediction*. Columbus, Ohio (p. 71-79). https://www.aclweb.org/anthology/W08-0909

Hosmer, D. W., Lemeshow, S. et Sturdivant, R. X. (2013). *Applied Logistic Regression* (3 edition). Wiley.

Indurkhya, N. et Damerau, F. J. (dir.). (2010). *Handbook of Natural Language Processing*. Chapman and Hall/CRC.

Kassambara, A. (2020). rstatix : Pipe-Friendly Framework for Basic Statistical Tests (0.6.0). https://CRAN.R-project.org/package=rstatix

Kim, J.-H. (2009). Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational Statistics & Data Analysis*, *53*(11), 3735-3745. https://doi.org/10/dtqx5t

Kowsari, K., Meimandi, K. J., Heidarysafa, M., Mendu, S., Barnes, L. E. et Brown, D. E. (2019). Text Classification Algorithms: A Survey. *Information*, *10*(4), 150. https://doi.org/10/gf4cnm

Kuhn, M. (2009). The caret package. *Journal of Statistical Software*, *28*(5).

Lai, C. S., Tao, Y., Xu, F., Ng, W. W., Jia, Y., Yuan, H., Huang, C., Loi, L. L., Xu, Z., Locatelli, G. (2019). A robust correlation analysis framework for imbalanced and dichotomous data with uncertainty. *Information Sciences*, 470, 58-77. https://doi.org/10.1016/j.ins.2018.08.017

Lantz, B. (2019). *Machine Learning with R: Expert techniques for predictive modeling, 3rd Edition*. Packt Publishing.

Lété, B., Sprenger-Charolles, L. et Colé, P. (2004). MANULEX: A grade-level lexical database from French elementary school readers. *Behavior Research Methods, Instruments, & Computers*, *36*(1), 156-166. https://doi.org/10/djzxmb

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, *15*(4), 474-496. https://doi.org/10.1075/ijcl.15.4.02lu

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

Nadeau, C. et Bengio, Y. (2003). Inference for the Generalization Error. *Machine Learning*, *52*(3), 239-281. https://doi.org/10/fnhnff

Nelson, J., Perfetti, C., Liben, D. et Liben, M. (2012). *Measures of Text Difficulty: Testing their Predictive Value for Grade Levels and Student Performance* (p. 58).

New, B., Pallier, C. et Ferrand, L. (2005). Manuel de Lexique 3. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 516-524.

Noble, W. S. (2006). What is a support vector machine? *Nature Biotechnology*, *24*(12), 1565-1567. https://doi.org/10/frp7k2

O'reilly, T. et 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

Ripley, B., Venables, W. et Ripley, M. B. (2016). Package 'nnet'. *R package version*, *7*, 3-12.

Scholkopf, B., Kah-Kay Sung, Burges, C. J. C., Girosi, F., Niyogi, P., Poggio, T. et Vapnik, V. (1997). Comparing support vector machines with Gaussian kernels to radial basis function classifiers. *IEEE Transactions on Signal Processing*, *45*(11), 2758-2765. https://doi.org/10/cgsc84

Stanké, B., Le Mené, M., Rezzonico, S., Moreau, A., Dumais, C., Robidoux, J., Dault, C. et 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).

Straka, M., Hajic, J. et Strakova, J. (2016). UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. *LREC*, 8.

Tong, S. et Koller, D. (2001). Support vector machine active learning with applications to text classification. *Journal of machine learning research*, *2*(Nov), 45-66.

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

Wu, X., Mauranen, A. et Lei, L. (2020). Syntactic complexity in English as a lingua franca academic writing. *Journal of English for Academic Purposes*, *43*, 100798. https://doi.org/10/ggffs6

5. Transition entre les articles 2 et 3

Les deux premiers articles de la thèse ont introduit et testé ALSI, un outil de traitement automatique du langage naturel spécialisé dans la mesure d'attributs linguistiques influençant la complexité des textes en français. Les résultats ont montré, par des méthodes robustes, qu'un modèle de classification employant une sélection d'attributs d'ALSI fonctionne à un niveau similaire ou supérieur à l'état de l'art en matière de classification de textes par niveau de difficulté. Il s'agit cependant de la difficulté telle qu'elle peut être estimée par les attributs du texte, en faisant abstraction des habiletés et connaissances des individus lisant le texte. Dans la deuxième partie de la thèse, nous nous tournons vers la dimension subjective de la lecture et présentons des travaux d'oculométrie, un domaine qui connaît actuellement un renouveau avec l'arrivée de dispositifs mobiles et faciles à utiliser. L'article 3 présente et met rigoureusement à l'épreuve une boîte à outils en langage Python utilisant l'intelligence artificielle pour automatiser les étapes fastidieuses de l'analyse des enregistrements captés par des lunettes d'oculométrie.

6. Troisième article:

Peut-on automatiser l'analyse de données d'oculométrie mobile ? Une étude de faisabilité basée sur la vision par ordinateur

Guillaume Loignon

Université de Montréal

Résumé

La récente démocratisation de l'oculométrie mobile offre une opportunité majeure pour la recherche sur l'attention visuelle. L'oculométrie mobile présente un avantage en ce qu'elle permet le déplacement des participants, favorisant la collecte des données dans un contexte plus naturel. Cependant, l'analyse des données d'oculométrie mobile présente des défis complexes, et les stratégies couramment utilisées pour relever ces défis ont des limites importantes, en particulier pour l'application à l'étude de la lecture. Cet article examine si l'étude oculométrique de la lecture peut être facilitée par des techniques de vision par ordinateur, une branche de l'intelligence artificielle. Nous expliquons d'abord le fonctionnement de la bibliothèque *Mobile Gaze Mapping* (MGM) pour le langage Python (MacInnes, 2020; MacInnes et al., 2018a, MacInnes et al., 2018b), qui propose d'automatiser certaines étapes de l'analyse des données d'oculométrie mobile en utilisant des techniques de vision par ordinateur. Nous testons ensuite cette solution sur des données simulées qui imitent des situations pouvant survenir lors de la collecte de données en oculométrie mobile, par exemple lorsqu'un geste de la main obstrue le capteur. Les résultats montrent que, dans un contexte d'étude oculométrique de la lecture, MGM stabilise les enregistrements assez fidèlement pour un usage scientifique. Nous proposons en outre des modifications aux paramètres internes de MGM qui pourraient améliorer les performances.

Mots-clés: dispositifs portables, oculométrie, vision par ordinateur, intelligence artificielle, apprentissage non-supervisé

Abstract

The recent democratization of mobile eye tracking offers a major opportunity for research on visual attention. An advantage of mobile eye-tracking is that it allows for the movement of participants, thus facilitating an ecological context when collecting data. However, the analysis of mobile eye-tracking data presents complex challenges, and the strategies commonly used to address these challenges have significant limitations, particularly for application to the study of reading. This paper examines whether the assessment of reading processes through eye tracking can be facilitated by computer vision, a branch of artificial intelligence. We first explain how the Mobile Gaze Mapping (MGM) package for the Python language (MacInnes, 2020; MacInnes et al., 2018) works, proposing to automate some steps in the analysis of mobile eye tracking data using computer vision techniques. We then test this solution on simulated data that emulate situations that can occur during mobile eye-tracking data collection, such as when the sensor is obstructed by a hand gesture. The results show that, in the context of an oculometric study of reading, MGM stabilizes the recordings sufficiently well for scientific purposes. We propose modifications to the internal parameters of MGM that could improve performance.

Keywords: wearable devices, eye tracking, computer vision, artificial intelligence, unsupervised learning

Peut-on automatiser l'analyse de données d'oculométrie mobile ? Une étude de faisabilité basée sur la vision par ordinateur

L'oculométrie est l'étude des mouvements des yeux et de la position du regard relativement à un stimulus visuel, généralement dans le but d'émettre des inférences concernant les processus cognitifs associés à l'attention visuelle (Duchowski, 2007). Les appareils enregistrant les données d'oculométrie, ou oculomètres, se divisent typiquement en deux grandes familles : stationnaires et mobiles. L'oculomètre stationnaire est placé devant les participants à la manière d'une webcam, et enregistre en continu la position de son point de regard (*gaze point*) dans le système de référence fixe du stimulus visuel. L'oculomètre mobile est porté sur la tête et muni d'une caméra frontale qui capte le champ de vision; la position du point de regard est enregistrée relativement à cette vidéo. Les deux types d'appareils produisent des données prenant la forme d'une liste de coordonnées cartésiennes indiquant la position du regard relativement à une image de référence. Nous désignons ces données comme des observations oculométriques. Chaque observation oculométrique contient un couple de coordonnées cartésiennes indiquant la position du regard, ainsi qu'un horodatage (marque indiquant le temps où l'observation a été enregistrée). En analysant les observations oculométriques en séquences ordonnées par l'horodatage, on peut ainsi reproduire le mouvement des yeux, tel un film composé d'images statiques montrées en succession rapide.

Les processus de lecture sont un domaine d'application fréquent de l'oculométrie (voir par exemple Bax, 2013; Sánchez et al., 2018; Zhan et al., 2020). Les travaux d'oculométrie s'intéressant à la lecture utilisent presque exclusivement l'oculométrie stationnaire en raison de la cadence d'échantillonnage (nombre d'observations oculométriques collectées par seconde) et une résolution plus élevée comparativement aux appareils mobiles, permettant des

enregistrements de meilleure qualité (Cognolato et al., 2018). La qualité d'un enregistrement oculométrique est typiquement mesurée en termes d'exactitude (*accuracy*) et de précision, qui équivalent respectivement à l'erreur systématique et à la dispersion statistique (Dalrymple et al., 2018; Hooge et al., 2019). Tel qu'illustré par la Figure 1, l'exactitude indique à quel point les coordonnées enregistrées par l'oculomètre correspondent à l'emplacement réel du regard, et la précision désigne la stabilité des coordonnées enregistrées, qui ne devraient pas varier si les participants fixent une cible immobile.

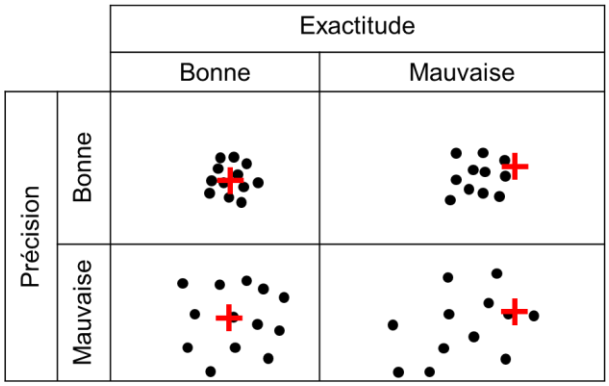


Figure 1. Deux caractéristiques définissant la qualité de données oculométriques. Les points indiquent les coordonnées du regard telles qu'enregistrées par un appareil d'oculométrie. La croix rouge indique l'emplacement réel du point scruté. Adapté depuis Dalrymple et al. (2018).

Le type d'expérience pouvant être réalisé en oculométrie stationnaire est cependant limité par l'affichage des stimuli sur un écran bidimensionnel et de taille relativement petite. Les participants doivent de plus demeurer immobiles, avec dans certains cas la tête retenue par un appuie-menton[1], dans une position qui peut être inconfortable après un certain temps (MacInnes et al., 2018). Les oculomètres stationnaires sont par ailleurs généralement plus coûteux et complexes à installer et à calibrer. C'est donc comme une réponse à un besoin de la communauté

[1] Un autre dispositif implique de mordre une barre pour stabiliser la tête, à la manière des radiographies du dentiste.

de chercheurs que les premiers oculomètres mobiles apparaissent sur le marché au début des années 2000 (Cognolato et al., 2018).

L'oculométrie mobile présente un intérêt pour l'étude de la lecture puisqu'elle favorise la validité écologique en permettant la lecture dans un contexte plus naturel. Cette mobilité accrue vient au prix de deux limitations majeures affectant l'analyse des données: 1) la position du regard est enregistrée relativement à la vidéo de la caméra subjective, et non par rapport au stimulus visuel; 2) l'appareil étant porté sur la tête, le mouvement des yeux ne peut être directement analysé sans tenir compte des autres mouvements (par ex. : rotations de la tête). Surmonter ces difficultés, que nous expliquons en plus de détails dans les sections qui suivent, faciliterait donc l'utilisation de l'oculométrie mobile.

6.1.1. Les défis d'analyse des données d'oculométrie mobile

Les données d'oculométrie prennent la forme d'une liste horodatée de coordonnées correspondant à l'emplacement du regard dans un système de référence. En oculométrie stationnaire, le système de référence est le stimulus visuel montré apparaissant sur le moniteur; l'appareil stationnaire enregistrera donc des coordonnées reflétant directement les régions fixées par les participants. Pour les oculomètres mobiles, le système de référence est la vidéo subjective (provenant de la caméra frontale). La Figure 2 illustre la même séquence de fixations et saccades dans les deux systèmes de référence.

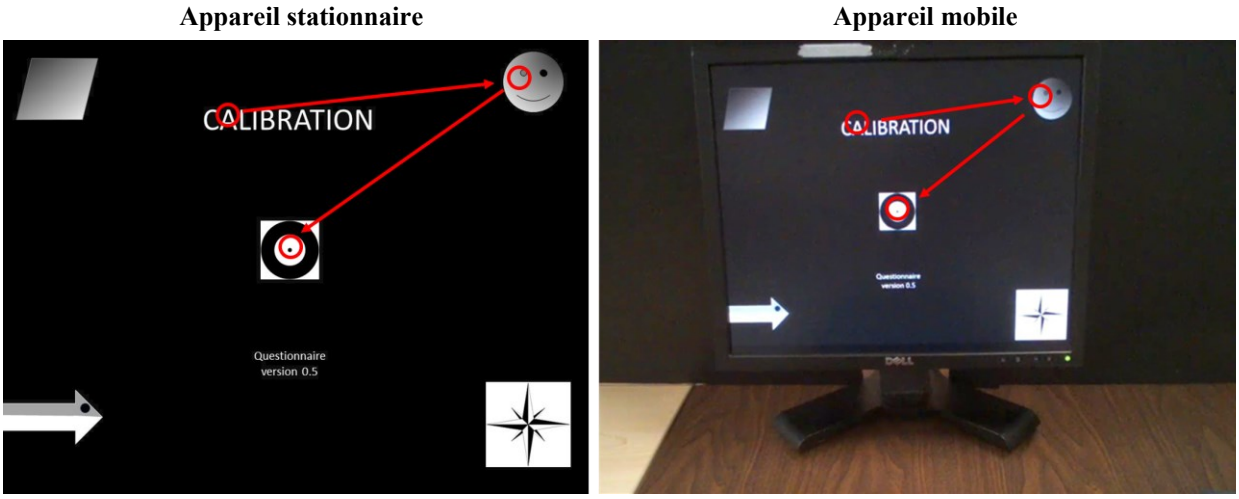


Figure 2. Illustration du problème de système de référence en oculométrie mobile. La même séquence de fixations et saccades (simulée) montrée dans deux systèmes de référence. Les cercles représentent les fixations et les flèches les saccades. À gauche : les données « objectives » d'un oculomètre stationnaire utilisent le stimulus visuel comme système de référence pour indiquer les coordonnées du regard. À droite : l'oculomètre mobile enregistre les coordonnées relativement à la vidéo subjective captée par la caméra frontale; nous supposons pour fins de démonstration qu'il n'y avait aucun mouvement de la tête durant la séquence.

Comme l'oculomètre mobile est porté sur la tête, les données oculométriques seront affectées par les mouvements des yeux, mais aussi par les mouvements de la tête et, pour certaines applications, par les mouvements du participant dans l'environnement. Par exemple, si on tourne la tête vers la gauche, le stimulus visuel se déplacera vers la droite dans le cadre de la caméra subjective. La Figure 3 illustre ce phénomène avec une séquence de trois observations oculométriques où un participant (fictif) fixait une cible visuelle tout en bougeant légèrement la tête.

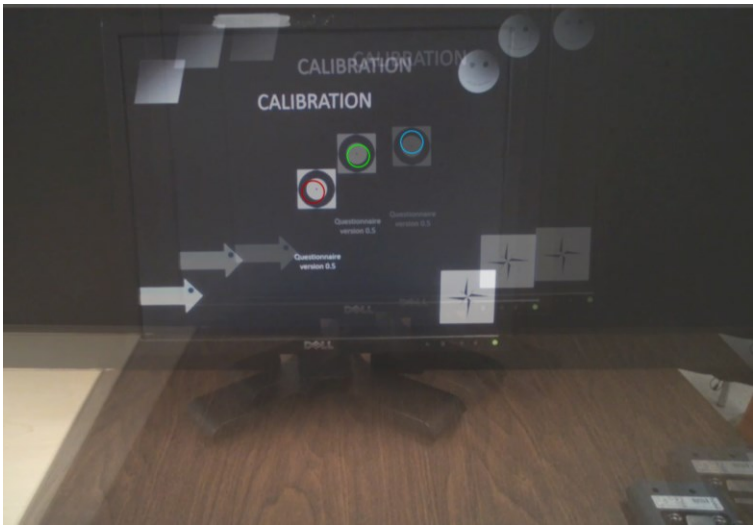


Figure 3. Simulation d'un enregistrement d'oculométrie mobile. Les cercles colorés indiquent trois fixations sur la cible de calibration au centre du stimulus visuel. Les coordonnées cartésiennes représentant la position du regard vont varier même si le participant fixait un même point.

6.1.2. Stratégies d'analyse des données d'oculométrie mobile

Nous survolons dans ce qui suit trois stratégies couramment utilisées pour analyser les données d'oculométrie mobiles, puis introduisons une solution s'appuyant sur la vision par ordinateur pour transformer les données « subjectives » en données s'apparentant plutôt à ce que produit un appareil stationnaire (pour un exposé similaire, voir Fong et al. 2016).

Stratégie 1 : Restreindre l'analyse à des mesures globales

Une manière de contourner le problème posé par des observations oculométriques contenant des coordonnées « subjectives » est de restreindre les analyses à certaines mesures globales ne tenant pas compte de la trajectoire du regard. La vidéo de la caméra frontale n'est alors pas utilisée. Ainsi, Miranda et al. (2018) ont comparé certaines mesures oculométriques de participants lisant des textes sur papier et support électronique (durée moyenne des fixations, amplitude des saccades, dilatation pupillaire et distance du point de convergence). Les auteurs n'ont pas considéré quelles régions du texte étaient scrutées par les participants, s'intéressant plutôt à la manière dont les indicateurs psychophysiques de la lecture varient selon le support.

Puisque ce type de méthodologie ne permet pas de connaitre à quel endroit le participant regarde à un moment spécifique, elle n'est pas suffisante pour une étude des stratégies déployées lors d'une tâche de compréhension de texte.

Stratégie 2 : Analyse manuelle de la vidéo de caméra frontale

Une méthode d'analyse fréquemment utilisée en oculométrie mobile consiste à visionner la vidéo captée par la caméra frontale de l'appareil sur laquelle le point de regard est affiché en superposition (Fong et al., 2016). Les chercheurs, parfois aidés par des experts, peuvent alors produire des annotations. Par exemple, Kahraman (2019) a eu recours à cette méthode afin d'identifier la présence d'une stratégie de recherche d'information lors de la réponse aux items de la version papier d'un test d'anglais langue seconde. Cette méthode peut être assistée par un logiciel ayant des fonctionnalités d'annotations vidéo, comme *NVivo* ou la bibliothèque *GazeCode* pour MATLAB. Ce type d'annotation demande cependant beaucoup de temps, et il peut être difficile de produire des mesures ayant une résolution temporelle fine, la durée d'une fixation étant typiquement entre 200 et 300 ms (Rayner, 1998).

Stratégie 3 : Analyse semi-automatisée avec un logiciel spécialisé.

Des logiciels spécialisés, tel *Tobii Pro Lab*, permettent à l'usager d'indiquer la position du stimulus visuel dans un enregistrement d'oculométrie mobile. Le logiciel transpose ensuite les coordonnées de fixation vers le plan fixe de l'image de référence, et peut ensuite produire des mesures ayant une résolution temporelle ou spatiale plus fine que ce que permettrait la stratégie 2. L'usager doit cependant délimiter l'emplacement du stimulus visuel ou de la zone d'intérêt; dans *Tobii Pro Lab* cette opération, montrée à la Figure 4, se fait au clic de souris et doit être répétée pour chaque cadre vidéo. Le logiciel peut alors produire des mesures et graphiques similaires à ce que l'équipe de recherche aurait pu obtenir par l'oculométrie stationnaire, par

exemple la durée moyenne des fixations sur une zone, ou encore une carte thermique illustrant les régions les plus longuement scrutées. Si elle permet des analyses plus fines, cette stratégie implique néanmoins un travail fastidieux et susceptible d'introduire un biais lié à l'erreur humaine (Imbert et al., 2015; Vansteenkiste et al., 2013).

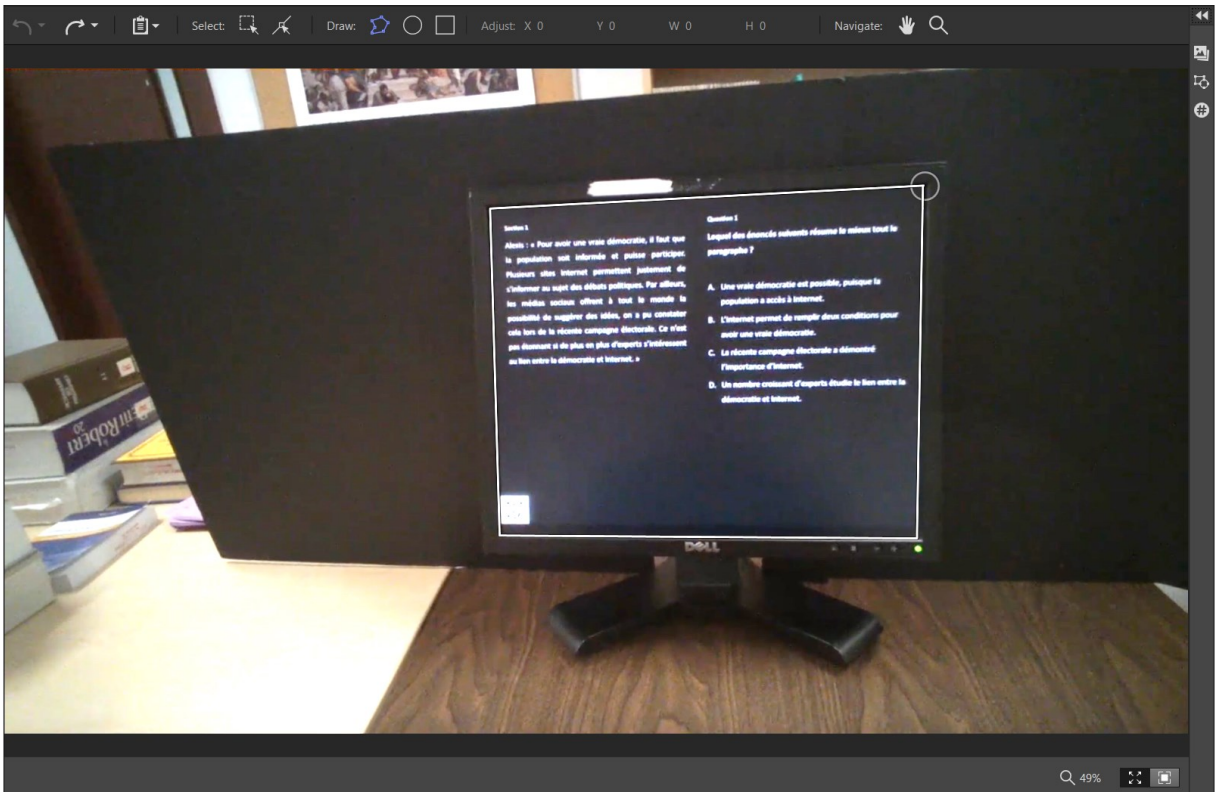


Figure 4. Capture d'écran montrant l'outil de délimitation de zone dans le logiciel Tobii Pro Lab. Le stimulus visuel a été délimité (en blanc) dans le logiciel, une opération faite à la souris et devant être répétée à chaque cadre vidéo.

Une nouvelle stratégie d'analyse basée sur la vision par ordinateur

Dans le but d'automatiser davantage l'analyse d'enregistrements d'oculométrie mobile, quelques auteurs ont utilisé une solution s'appuyant sur des techniques de vision par ordinateur (*computer vision*). La vision par ordinateur est une branche du génie informatique ancrée dans les théories sur la cognition et qui s'intéresse à la reproduction des mécanismes de la vision par des ordinateurs (Ikeuchi, 2014). Décrite sommairement, cette nouvelle approche d'analyse

consiste à détecter l'emplacement du stimulus visuel par des algorithmes de détection d'image, c'est alors l'algorithme qui délimite l'emplacement du stimulus visuel. Les coordonnées cartésiennes contenues dans les observations oculométriques sont ensuite transposées vers le plan du stimulus visuel, ce qui reproduit le type de données enregistré par un appareil stationnaire.

Cette application de la vision par ordinateur à l'oculométrie mobile est apparue récemment. Un précurseur se trouve dans les travaux de Toyama et al. (2012) qui décrivent un algorithme s'appuyant sur la reconnaissance d'objet pour détecter si le participant regardait une cible, et proposent une application de réalité augmentée pour les musées. Dans le domaine de la lecture, Kunze et al. (2013) ont appliqué des techniques de vision par ordinateur à des enregistrements d'oculométrie mobile afin de détecter les mots plus difficiles à lire dans un texte. Il peut cependant être ardu d'intégrer la vision par ordinateur à une chaine de traitement des données oculométriques vu les connaissances pointues que cette technique exige, tant en vision informatique qu'en programmation.

L'intégration de la vision par ordinateur à l'analyse d'enregistrements d'oculométrie mobile est devenue plus accessible lorsque MacInness et collègues ont publié la boite d'outils *Mobile Gaze Mapping* et en ont partagé le code source avec la communauté des chercheurs (MacInnes, 2020). Créée en langage Python, *Mobile Gaze Mapping* (ci-après MGM) déploie un ensemble de techniques de vision par ordinateur pour transformer les données d'oculométrie mobile vers un format qui s'apparente à ce qu'aurait produit l'oculométrie stationnaire. Le stimulus visuel est détecté automatiquement dans chaque cadre vidéo, ce qui élimine le travail fastidieux de délimiter la zone d'intérêt.

À l'heure actuelle, la bibliothèque MGM n'a été l'objet que de trois publications. Un premier article décrit succinctement la méthodologie de MGM et relate une expérimentation pilote où les participants observaient une œuvre d'art à des distances et angles d'observation variés (MacInnes et al., 2018b). Les auteurs ont avec succès converti les données « subjectives » de plusieurs modèles de lunettes d'oculométrie vers le plan « objectif » d'une image représentant l'œuvre. Un second article présente succinctement la bibliothèque MGM pour Python (MacInnes, 2018b). La troisième publication relate l'utilisation de MGM pour transformer des données d'oculométrie mobile portant également sur l'observation d'œuvres d'art, les auteurs ont ainsi pu produire des cartes thermiques illustrant les patrons d'attention visuelle des visiteurs (Grazioso et al., 2020).

6.1.3. La présente étude

Si l'expérimentation pilote avec MGM a démontré l'intérêt de l'outil, des tests plus rigoureux demeurent nécessaires pour démontrer que MGM est prêt pour une utilisation dans un cadre de recherche scientifique sur la lecture de textes. Nous nous interrogeons notamment quant à la robustesse de MGM face aux enregistrements réalisés dans des conditions variables découlant de la nature même de l'oculométrie mobile. De plus, bien que l'étude de la lecture soit une application fréquente de l'oculométrie, MGM n'a pas été testé pour l'analyse de données oculométriques où les stimuli sont du texte. Nous voyons donc un intérêt à tester la validité de MGM pour l'étude des processus de lecture par l'oculométrie mobile. Notre étude se divise en deux parties : un volet théorique dont l'objectif est de détailler le fonctionnement de MGM, et un volet empirique dont l'objectif est d'évaluer la qualité (exactitude et précision) des données d'oculométrie mobile transformées par MGM en utilisant différentes configurations.

Le volet théorique s'appuie sur la littérature scientifique et technique du domaine, et sur une analyse du code source de MGM. MGM se présente actuellement comme une « boite noire » qu'il serait possible d'utiliser en ignorant le fonctionnement interne. Notre objectif dans le premier volet est de permettre une évaluation *a priori* de MGM en ouvrant cette boite noire, en décrivant les opérations effectuées sur les données de même que les principes mathématiques sous-jacents.

Le volet empirique s'appuie sur une expérimentation à partir de la simulation d'un enregistrement d'oculométrie mobile, créé dans le cadre de cette étude. L'avantage des données simulées pour cette étude est de permettre de contrôler la position du regard dans les observations oculométriques, ce qui permet en retour de comparer la position réelle du regard à la position résultant du traitement par MGM. La simulation imite un participant scrutant en alternance deux points dans une image représentant un item d'une épreuve de compréhension de texte. Afin de tester la robustesse de MGM dans différentes conditions, nous avons simulé divers événements perturbateurs, c'est-à-dire des situations qui surviennent typiquement lors d'enregistrements d'oculométrie mobile et qui pourraient altérer la qualité des données collectées. Nous avons également cherché à déterminer si les paramètres de MGM choisis par ses auteurs, étaient optimaux pour une application d'étude oculométrique de la lecture. Pour fins de comparaison, nous avons testé MGM dans sa configuration proposée et dans des configurations alternatives où les paramètres internes avaient été modifiés. L'objectif du second volet est donc de tester la robustesse de MGM dans des conditions variées, et de vérifier si certaines manipulations des paramètres internes de MGM permettraient d'en améliorer la qualité de traitement. Les résultats permettent de comparer la qualité des données transformées et la robustesse des configurations testées dans le contexte de différents événements perturbateurs

(image bloquée par la main, mouvements de la tête, flous, etc.) L'article se conclut sur une discussion synthèse des deux volets.

6.2. Volet 1 : l'analyse de données d'oculométries mobiles avec MGM

MGM s'appuie sur l'*application projective*, un traitement couramment employé en vision par ordinateur afin de, par exemple, faire correspondre deux images photographiées sous des angles différents. Au sens strict, l'application projective est une transformation géométrique linéaire qui projette une figure vers un autre plan ayant une taille et une inclinaison différentes, tout en préservant la structure générale de la figure (Ikeuchi, 2014). Soit une forme s'inscrivant dans un plan et que l'on souhaite transposer vers un second plan, l'*homographie* désigne l'ensemble des transformations linéaires qui traduisent les coordonnées du premier plan vers le second; chaque point de la forme est ainsi projeté vers son équivalent dans le second plan. En vision par ordinateur, l'homographie est typiquement représentée dans une matrice *H* de dimensions 3 x 3. En multipliant par *H* un vecteur contenant les coordonnées cartésiennes d'un point (ici *P1x* et *P1y*), on obtient les coordonnées transposées dans l'autre plan, selon l'équation :

$$\begin{pmatrix} P1x & P1y & 1 \end{pmatrix} \begin{pmatrix} a & b & 0 \\ c & d & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} P2x & P2y & 1 \end{pmatrix}$$

Selon les valeurs que prennent *a*, *b*, *c*, et *d*, la multiplication des coordonnées par la matrice *H* produira donc une, ou une combinaison de trois transformations linéaires : l'homothétie, la transvection et la rotation. La composante *a* indique l'homothétie horizontale, *b* indique la transvection horizontale, *c* indique la transvection verticale et *d* l'homothétie verticale. La rotation est produite en manipulant les quatre composantes : soit une rotation de *x* degrés, *a* et *d* indiqueront le cosinus de *x*, *b* le sinus, et *c* sera égal à -sin(*x*). Les autres composantes de *H* sont constantes. La Figure 5 donne un exemple où une image de référence (à gauche) a été

projetée vers son équivalent dans une photographie; la matrice *H* est montrée dans le coin inférieur gauche.

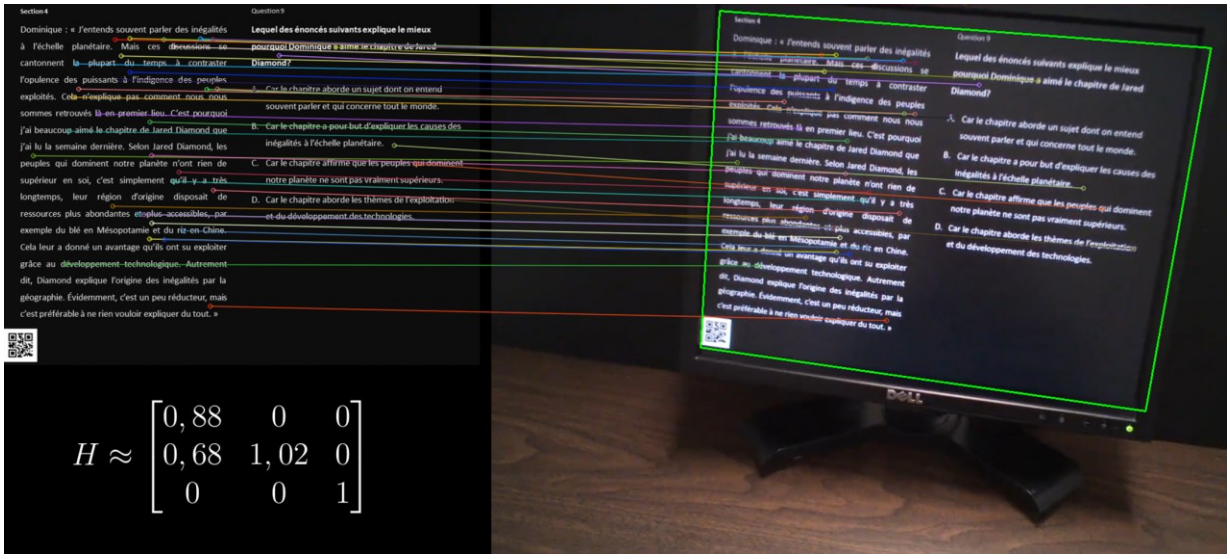


Figure 5. Démonstration de l'appariement d'image par méthode de vision par ordinateur. Les points-clés d'une image de référence (à gauche) sont appariés à leurs équivalents dans un cadre vidéo tiré d'un enregistrement oculométrique. La zone délimitée en vert indique l'emplacement estimé de l'image de référence dans le cadre vidéo. Dans le coin inférieur gauche : la matrice *H* décrivant la transformation (valeurs arrondies à deux décimales).

Les cercles colorés indiquent le centre des *points-clés*, qui sont appariés avec leur équivalent dans l'autre image. En vision par ordinateur, le point-clé est une région de l'image dense en informations identifiantes et qui est associé à une position (coordonnées) dans l'image et à un descripteur, qui est un ensemble de variables qui décrivent le point-clé et le rendent unique. Ce qui constitue un point-clé dépend de l'algorithme de détection, il peut s'agir de lignes, de coins, de contours, de taches, etc. La détection de points-clés est l'identification de ces régions et de leur emplacement (coordonnées) dans l'image (Lowe, 2004). C'est en associant les points-clés de l'image de référence aux points-clés du cadre vidéo que MGM calcule la matrice *H*, qui permet à son tour de transformer les données oculométriques.

Nous avons vu que les appareils d'oculométrie mobile enregistrent les mouvements des yeux sous la forme d'une liste horodatée de coordonnées cartésiennes relatives à la vidéo captée

par la caméra frontale, et que l'homographie est un ensemble de transformations permettant de projeter une image vers un autre plan. L'objectif de MGM est de trouver l'homographie (matrice *H*) qui permet de transformer les coordonnées relatives à la vidéo en coordonnées relatives au stimulus visuel qui était scruté par les participants. Une matrice *H* est calculée pour chaque image de la vidéo subjective. Ces matrices permettent de projeter l'image mouvante de la caméra frontale vers un plan stable comportant l'image de référence. En boucle, MGM analyse un à un les cadres de vidéo subjective et calcule *H* pour chacun d'entre eux en comparant le cadre à l'image de référence. La Figure 6 illustre le fonctionnement de MGM.

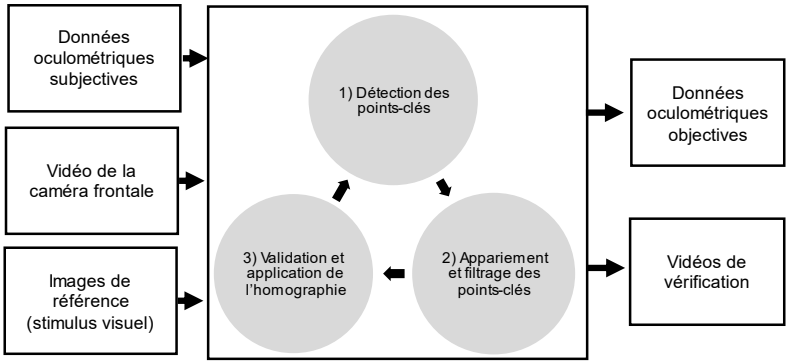


Figure 6. Schéma illustrant les intrants, la boucle de traitement, et les extrants de MGM.

6.2.1. Les intrants

MGM requiert trois sources de données : la vidéo captée par la caméra frontale, les données oculométriques en tant que telles, et une image de référence représentant le stimulus visuel. Les données oculométriques prennent la forme d'un vaste tableau dont chaque ligne est une observation oculométrique, et dont les colonnes indiquent l'horodatage (*timestamp*), le numéro du cadre vidéo correspondant, et les coordonnées de l'emplacement du regard relatives à la vidéo subjective. Un script de traitement est inclus dans la bibliothèque afin de convertir dans le format requis par MGM les données générées par les appareils d'oculométrie mobile les plus courants.

6.2.2. Les processus

MGM opère via une boucle de traitement qui détecte d'abord les points-clés et extrait leurs descripteurs, puis les associe aux points-clés correspondants dans l'image de référence préalablement analysée, formant ainsi des *pairages*. Un descripteur désigne dans ce contexte un ensemble de variables qui définissent le point-clé et le rendent distinct. Les pairages sont ensuite filtrés afin de minimiser l'incidence d'erreurs de pairages (association inadéquate, le second point-clé n'était pas l'équivalent du premier dans l'autre image). Les pairages jugés comme valides servent à calculer l'homographie permettant le mieux de projeter les données oculométriques subjectives vers le plan fixe de l'image de référence. Après validation, l'homographie est finalement appliquée aux coordonnées du point de fixation. Sous forme de pseudo-code, le processus peut se résumer ainsi :

```
Détecter et décrire les points-clés de l'image de référence.
Pour chaque cadre vidéo, faire :
    Détecter et décrire les points-clés du cadre.
    Apparier les points-clés du cadre à ceux de l'image de référence;
    Sélectionner les pairages produisant la meilleure homographie.
    Si l'homographie résultante est valide, faire :
        Appliquer l'homographie aux données oculométriques.
        Sinon, inscrire des données manquantes.
```

Nous détaillons dans les sections suivantes les trois grandes étapes de la boucle de traitement de MGM, puis en décrivons les résultats.

Étape 1 : Détection et description des points-clés

L'algorithme de détection utilisé par MGM est le Scale Invariant Feature Transform, ou SIFT (Lowe, 2004). La procédure SIFT est expliquée en plus de détails dans Krig (2016, p. 209-213), nous en présentons ici un survol. L'image est préalablement convertie en tons de gris

et représentée dans une matrice dont les dimensions sont la taille de l'image (en pixels) et dont chaque cellule contient une variable numérique indiquant l'intensité de la teinte de gris pour le pixel correspondant. SIFT identifie comme points-clés les pixels dont la teinte de gris se démarque des huit pixels adjacents. Afin d'être plus robuste aux changements d'échelle et au bruit visuel, la procédure est répétée à différents niveaux de redimensionnement (appelés octaves) et niveaux de flou (appelés *layers*, ou couches). Le type de flou appliqué est de type gaussien, qui est un filtre de traitement de l'image atténuant les fréquences spatiales supérieures à un seuil. SIFT combine l'information obtenue à-travers les octaves et leurs couches, puis affine la sélection de points-clés et l'estimation de leurs coordonnées cartésiennes.

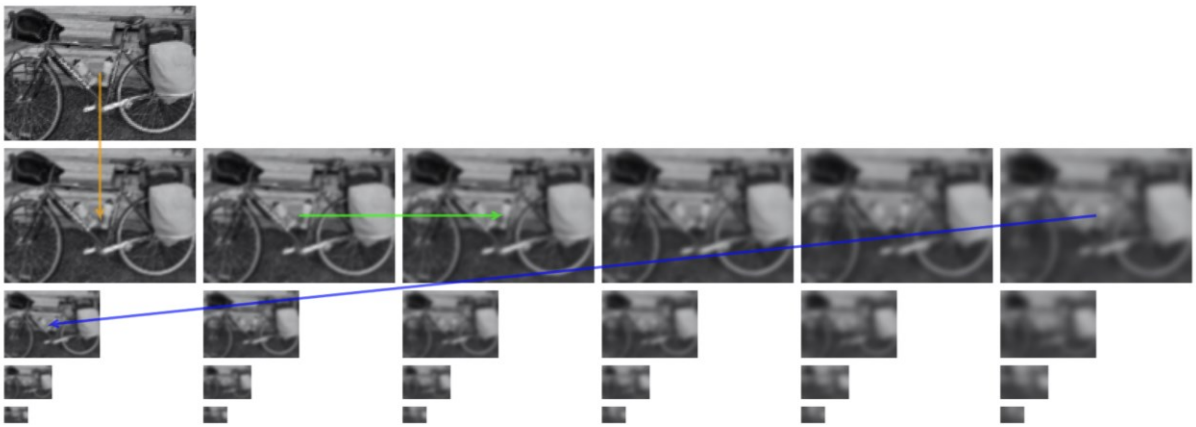


Figure 7. Schéma illustrant quatre octaves (montrés verticalement) et six couches ou niveaux de flou (montrés horizontalement) produits par l'algorithme SIFT lors de la détection des points-clés. Les flèches indiquent les transformations effectuées : un flou gaussien initial (orange), une série de flous gaussiens produisant les couches (vert) et le redimensionnement (bleu). La figure a été réalisée avec l'outil disponible au http://weitz.de/sift/

Une fois les points-clés détectés, SIFT leur calcule chacun un descripteur qui prend la forme d'un vecteur contenant 128 composantes. Tel qu'illustré par la Figure 8, la région du point-clé est divisée en 16 sous-régions de 4 x 4 pixels. Un histogramme des gradients (amplitude et orientation dominante dans la variation des tons de gris) est produit pour chaque sous-région. Les bâtonnets de l'histogramme correspondent à 8 orientations possibles pour le

gradient. Le descripteur regroupe les valeurs numériques des histogrammes, soit 8 variables par sous-région pour 16 sous-régions, totalisant 128 composantes.

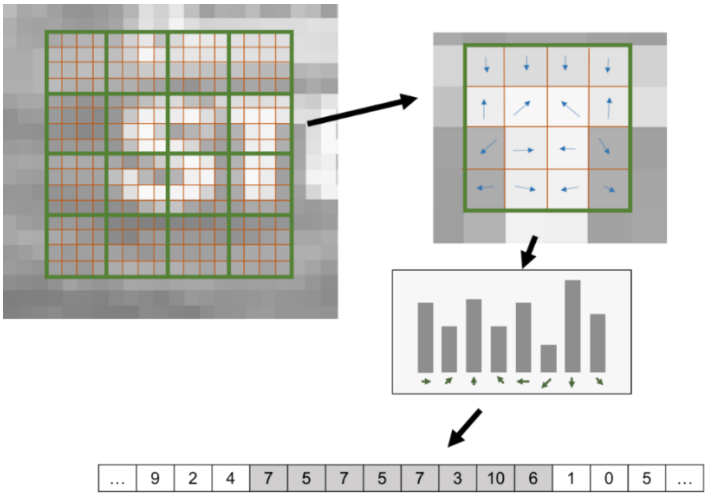


Figure 8. Schéma illustrant la description d'un point-clé par l'algorithme SIFT. Une des sous-régions est montrée en gros plan, avec les gradients de chaque pixel et leur sommation dans un histogramme. Les données de l'histogramme sont finalement ajoutées au descripteur.

Au terme de l'étape 1, on obtient un ensemble de points-clés ayant chacun un descripteur sous la forme d'un vecteur de données, et une position exprimée comme une coordonnée cartésienne.

Étape 2 : Appariement et filtrage des points-clés

L'objectif de la deuxième étape est de former les appariements de points-clés les plus plausibles entre le cadre vidéo et l'image de référence. La ressemblance entre les points-clés est estimée à partir de la mesure de distance entre leurs descripteurs; plus la mesure de distance est petite, plus les points-clés sont jugés comme étant similaires. La mesure de distance est la distance Euclidienne. Le résultat de l'étape 2 est une liste de points-clés du cadre vidéo avec leurs deux meilleurs pairages. Le deuxième meilleur pairage ne sera pas utilisé pour le calcul de l'homographie, il est conservé à cette étape en vue d'une validation par le test des ratios de Lowe (Lowe, 2004).

Le test des ratios de Lowe compare les deux meilleurs pairages possibles pour chaque point-clé, vérifiant que la qualité du meilleur pairage se distingue suffisamment de celle du deuxième meilleur pairage. L'intuition sous-jacente à ce test est que, chaque point-clé du cadre ne pouvant avoir qu'un seul équivalent dans l'image de référence, alors le deuxième meilleur pairage devrait être aberrant, ce qui devrait se refléter par une distance euclidienne plus élevée. Soit *pc1* un point du cadre dont les deux meilleurs parages sont avec *pr1* et *pr2*, qui sont deux points-clés de l'image de référence, et *d*() une fonction retournant la distance euclidienne entre les descripteurs de deux points-clés. Le test des ratios de Lowe vérifie si le ratio des distances euclidiennes découlant des deux pairages est inférieur à un seuil :

$$\frac{d(pc1, pr1)}{d(pc1, pr2)} < seuil$$

Lorsque l'équation est satisfaite, le point-clé est considéré comme valide et son meilleur pairage sera utilisé lors du calcul de l'homographie. Sinon, le point-clé est retiré de la liste et ses pairages seront ignorés lors du calcul de l'homographie. Le seuil de tolérance utilisé par MacInness et collègues était de 0,5. Les pairages conservés sont traités par l'algorithme RANSAC.

RANSAC, pour *Random Sample Consensus,* est à la base une méthode robuste de régression linéaire. Nous décrivons ici l'implémentation de RANSAC en vision par ordinateur, telle qu'utilisée dans MGM. Le principe est de trouver l'homographie produisant le plus possibles de points valides (*inliers*), qui sont les points dont les résidus sont inférieurs à un seuil. Les modèles produits sont alors des homographies. Les résidus sont les différences entre les coordonnées réelles et les coordonnées projetées, décrits par l'équation suivante, où *H* est l'homographie et *P1* et *P2* sont les coordonnées cartésiennes d'un point-clé et de son équivalent présumé dans l'autre image :

$$Résidus = H(P1) - P2$$

RANSAC calcule une homographie à partir d'une grappe de quatre pairages sélectionnés au hasard, puis teste cette homographie sur l'ensemble des pairages. Les pairages dont les résidus sont inférieurs à un seuil sont alors considérés comme valides (*inliers*) et les autres comme aberrants (*outliers*). Cela signifie, par exemple pour un seuil de deux pixels, qu'un pairage sera valide si son résidu est de deux pixels ou moins. RANSAC recommence ensuite avec une autre grappe aléatoire de pairages, et ces opérations sont répétées en boucle jusqu'à ce que qu'on ait atteint le nombre maximal d'itérations, qui est déterminé dans MGM par un paramètre. Au terme du processus, la meilleure homographie est celle ayant produit le plus de pairages valides.

Étape 3 : Validation et application de l'homographie

Avant d'appliquer l'homographie pour transformer les coordonnées, MGM effectue encore une étape de validation. Des homographies aberrantes peuvent en effet survenir, par exemple lorsque le participant ne regardait pas l'image de référence ou lorsque cette dernière est incorrectement détectée. MGM vérifie donc si l'application de l'homographie projette les coordonnées des observations oculométriques vers un point qui existe dans l'image de référence. Nous avons ajouté une vérification supplémentaire qui exclut certaines homographies improbables, notamment lorsqu'une rotation est si prononcée que l'image se retrouverait inversée. Lorsque l'homographie calculée pour un cadre vidéo est rejetée comme aberrante, celle-ci n'est pas utilisée dans la transformation des coordonnées des points de fixations qui lui sont associés; on inscrit plutôt des données manquantes.

6.2.3. Les extrants

Le résultat final est un tableau de données oculométriques objectives où chaque ligne correspond à un point de fixation, avec sa position dans le temps (exprimée en millisecondes) et

ses coordonnées relatives à l'image de référence. Afin d'aider à diagnostiquer des problèmes survenant durant le traitement, MGM produit en outre trois vidéos de vérification (voir Figure 9). La première vidéo reprend les images tirées de la caméra frontale de l'appareil et ajoute, en superposition, un cercle représentant l'emplacement du point de fixation; cette vidéo pourrait servir pour une annotation manuelle des enregistrements. La seconde vidéo présente le résultat de la transformation : un plan fixe montrant l'image de référence avec le point de fixation en superposition. La troisième vidéo montre l'alignement de l'image de référence par-dessus la vidéo de la caméra frontale et peut être utilisée pour valider le traitement de MGM.

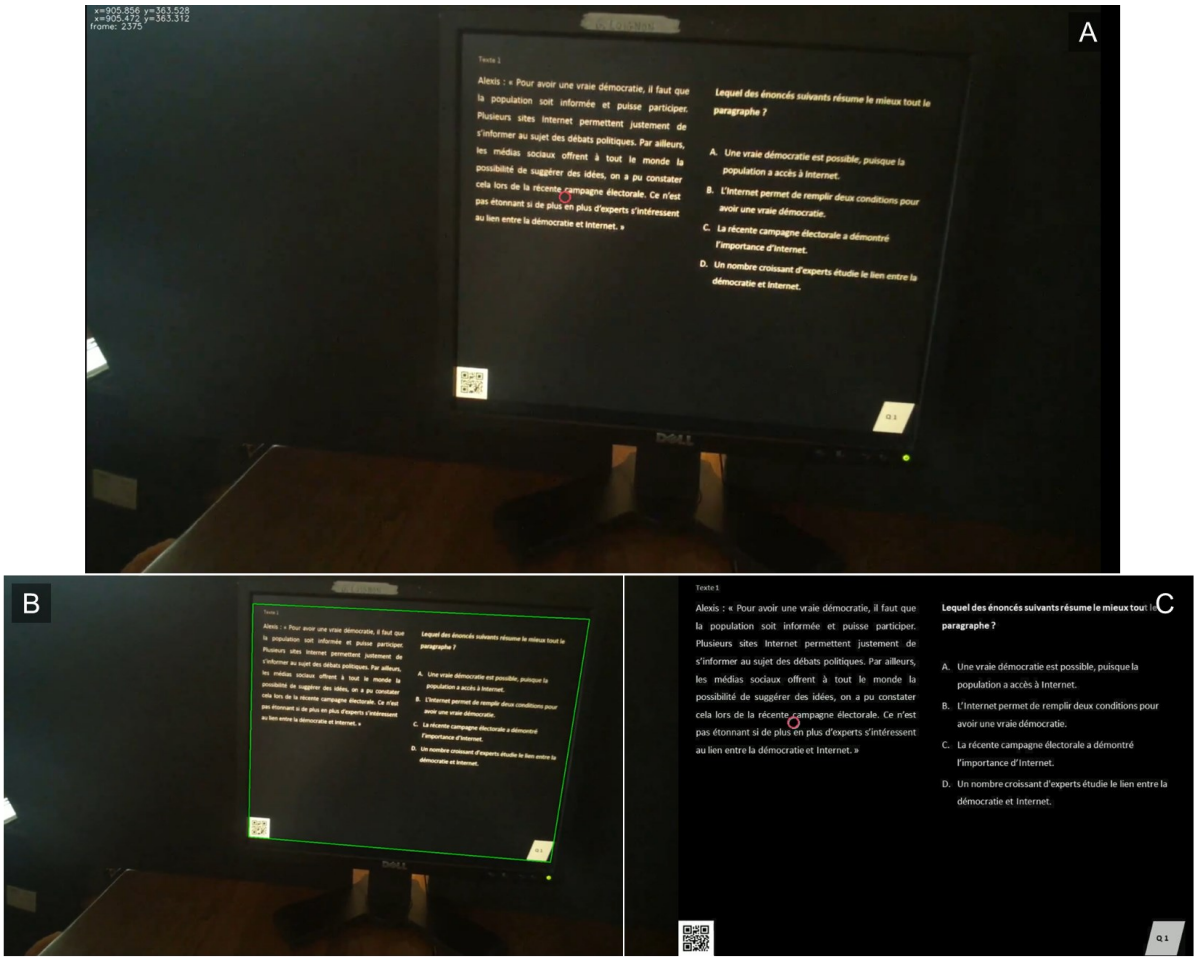


Figure 9. Arrêt sur image synchronisé de trois vidéos de vérification produites par MGM. (A) Vidéo de la caméra frontale avec point de fixation indiqué par un cercle coloré. (B) Emplacement du stimulus visuel sur la vidéo captée par la caméra frontale (C) Plan fixe du stimulus visuel avec point de fixation indiqué par un cercle coloré. Le code source de MGM a été légèrement altéré afin d'ajouter des informations de diagnostic sur la vidéo

A et un périmètre vert indiquant le stimulus visuel sur la vidéo B. La vidéo est disponible au : https://www.youtube.com/watch?v=wJGMtqKpZrg

6.2.4. Synthèse du volet 1

L'objectif du volet 1 était de détailler le fonctionnement de MGM. Dans un contexte de démocratisation de l'oculométrie mobile, MGM répond à un problème méthodologique causé par le déplacement du point de référence à partir duquel l'appareil mobile enregistre la position du regard. L'utilisation de MGM évite aux équipes de recherche le travail fastidieux d'identifier manuellement la position du stimulus visuel dans chaque image de la vidéo. Nous avons vu comment MGM combine des techniques de vision par ordinateur afin de transformer automatiquement des données « subjectives » en les projetant vers un plan fixe. À l'interne, MGM utilise l'algorithme SIFT pour détecter les points-clés des images et extraire leurs descripteurs, qui sont des vecteurs représentant les gradients de gris de la région entourant le point-clé. MGM paire les points-clés du cadre vidéo à ceux de l'image de référence en se basant sur la distance euclidienne entre leurs descripteurs. Le test des ratios de Lowe agit comme premier filtre pour éliminer des pairages aberrants. L'algorithme RANSAC a ensuite le double rôle d'éliminer davantage de pairages aberrants et de calculer l'homographie optimale en s'appuyant sur les pairages les plus plausibles. L'homographie est finalement appliquée aux coordonnées (relatives à la vidéo) du point de fixation afin d'obtenir sa position dans l'image de référence.

La qualité (précision et exactitude) des données telles que transformées par MGM dépend largement de la capacité de ses algorithmes à traiter adéquatement les images. Si l'emplacement de l'image de référence est incorrectement détecté, ou si des pairages aberrants se glissent dans le calcul de l'homographie, alors les coordonnées produites ne correspondront pas à l'emplacement qui était fixé par le participant. MGM produit des vidéos de vérification pour

identifier de potentiels problèmes; nous avons altéré la partie du code source produisant ces vidéos afin de les rendre plus informatives.

6.3. Volet 2 : expérimentation avec données simulées

Une application projective réussie devrait transposer adéquatement les coordonnées des fixations visuelles du plan de la vidéo subjective (caméra frontale de l'appareil d'oculométrie mobile) vers le plan de référence de l'image fixe. Si MGM a correctement fait son travail, alors le point de fixation enregistré relativement au cadre vidéo sera transposé avec précision vers l'image de référence (voir Figure 9). Plusieurs obstacles à la validité de la méthode MGM peuvent survenir dans les conditions normales d'un enregistrement d'oculométrie mobile. Des essais ont démontré que SIFT avait une certaine sensibilité aux changements d'échelle, aux effets de flou, et aux déformations sphériques (Karami et al., 2017; Khan et al., 2011; Wu et al., 2013; Zheng-Jian Ding et al., 2012). Or, un mouvement avant-arrière de la tête introduit un changement d'échelle, la taille apparente du stimulus visuel variant selon le mouvement. De même, un mouvement rapide peut momentanément introduire un effet de flou à la vidéo, et certains oculomètres mobiles captent les images à travers une lentille grand-angle qui introduit d'emblée une déformation sphérique. Enfin, l'algorithme SIFT peut échouer lorsque l'image contient des détails répétitifs tels que les caractères d'un texte (Uchiyama et Saito, 2009), ce qui pourrait limiter la portée de la méthode MGM pour l'étude de la lecture.

La Figure 10 illustre des situations survenant lors d'enregistrement oculométrique où la robustesse de MGM pourrait être mise à l'épreuve. Il s'agit d'arrêts sur image tirés d'enregistrements oculométriques captés avec des lunettes Tobii Pro Glasses 2. Les images A à C de la Figure 10 sont tirées d'essais de calibration. L'image A rend plus apparente la déformation sphérique introduite par la lentille grand-angle de l'appareil. La déformation est

surtout visible près des bordures verticales. Les images B et C montrent une transvection verticale et horizontale, l'image C étant en outre plus sombre que A et B. Le fait que le stimulus visuel soit montré sur un écran introduit un effet de flou lumineux (*blooming*) qui se remarque notamment près du périmètre des carrés noirs.

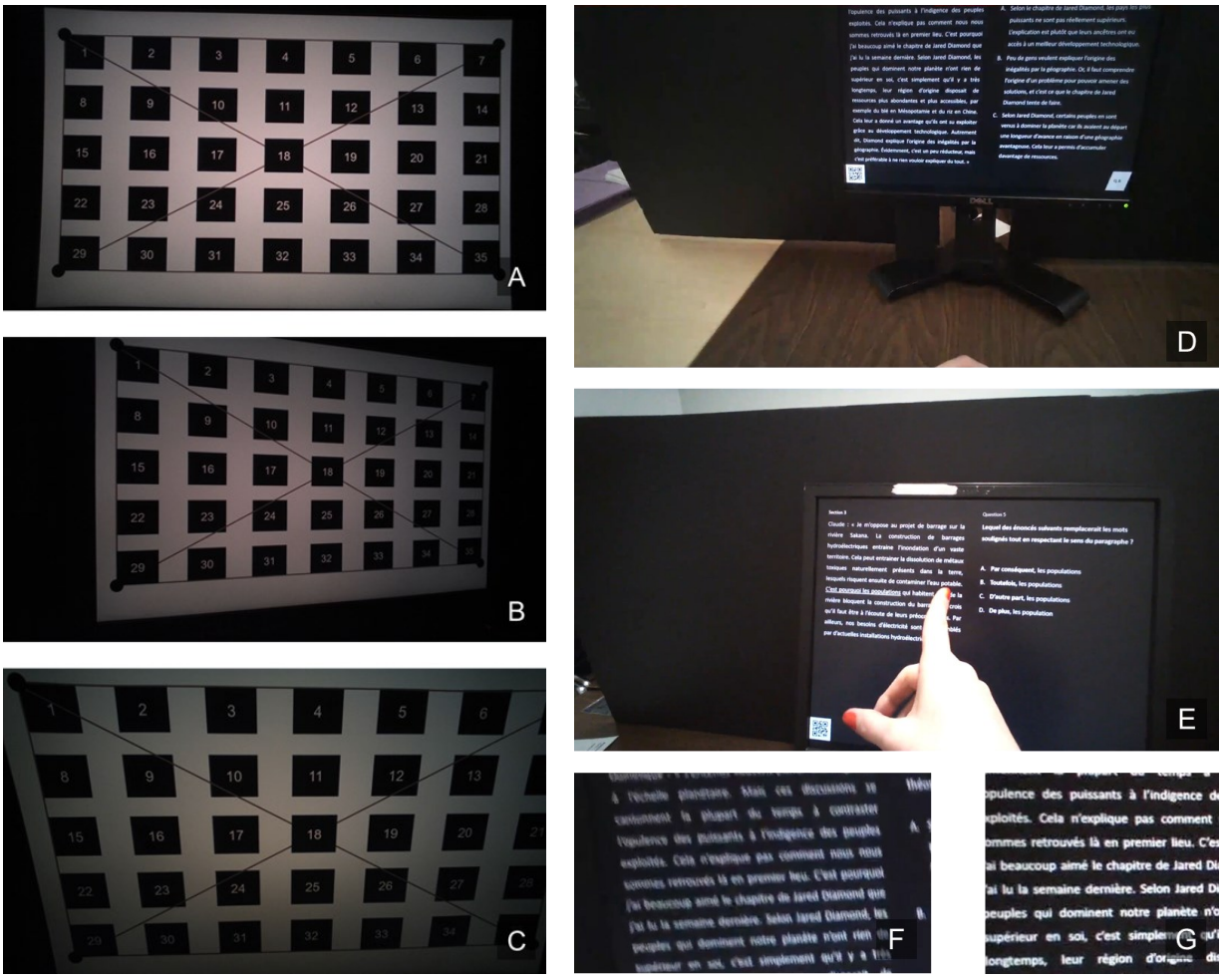


Figure 10. Images tirées de la vidéo de caméra frontale d'un enregistrement oculométrique réalisé avec les lunettes Tobii Pro 2 lors d'essais de calibration (A à C) et lors d'une expérience réelle (D à G). A : déformation sphérique. B et C : transvection. D : stimulus visuel tronqué. E : occultation partielle. F : flou induit par le mouvement (gros plan). G : flou lumineux induit par le moniteur (gros plan).

Les images D à G de la Figure 10 sont tirées des enregistrements d'une expérience portant sur la compréhension de texte relatée dans une autre publication (article 4 de la présente thèse). En D, le stimulus visuel se retrouve partiellement hors champ après un mouvement de la

tête. Dans l'image E, une participante pointe du doigt une région du stimulus visuel, occultant partiellement l'image. Ce type d'occultation est attendu lors d'enregistrements qui tentent d'imiter un contexte naturel, et serait inévitable dans un dispositif expérimental dans lequel, par exemple, les participants devaient interagir avec une interface apparaissant à l'écran. De plus, l'oculomètre est légèrement plus près du moniteur en E qu'en D, faisant varier la taille apparente du texte. Les images F et G sont des gros plans détaillant des artéfacts visuels : F montre un flou induit par le mouvement, et G montre l'effet de flou lumineux produit par l'écran, les contours des caractères devraient normalement être nets.

En résumé, pour être intégré à une chaine de traitement de données oculométriques, MGM devrait être robuste aux différents types de mouvements survenant lors de la collecte de données, mais aussi aux occultations, déformations et artéfacts visuels introduits par le participant ou par l'équipement lui-même. Notre objectif dans le volet 2 de cette étude est donc de reproduire, par le biais de données simulées, les situations présentées dans les exemples précédents afin de tester la robustesse de MGM dans des conditions variées. L'utilisation de données simulées permet d'isoler différentes conditions afin d'identifier avec plus de précision ce qui pourrait affecter la validité des données produites par MGM.

6.3.1. Méthodologie

Afin de tester la robustesse de MGM lorsque le stimulus visuel est un texte et dans des conditions variées, nous avons analysé la qualité des transformations produites par MGM en fonction de l'événement perturbateur de la configuration utilisée. Les événements perturbateurs (mouvements, flous, etc.) sont des situations pouvant affecter la qualité des données et sont listés au Tableau 1. Les configurations de MGM sont des variantes apportées aux paramètres internes de MGM et à l'algorithme de détection de points-clés. Les configurations utilisées sont listées au

tableau 2, leur code source est disponible en annexe du présent article. Comme alternative à l'algorithme SIFT utilisé par défaut dans MGM pour détecter les points-clés, nous avons aussi testé l'algorithme AKAZE, considéré comme plus robuste (Alcantarilla et al., 2013; 2012). Des indicateurs de performance ont été produits pour chaque configuration et événement perturbateur afin de tester MGM tout en déterminant si d'autres configurations améliorent la performance dans certaines conditions.

Simulation d'un enregistrement oculométrique

Nous avons simulé des données imitant le format et certaines particularités d'un enregistrement produit par l'oculomètre mobile *Tobii Pro Glasses 2*. La démarche avait pour objectif de reproduire des événements perturbateurs tels que montrés à la Figure 10; nous nous sommes inspirés de travaux en vision par ordinateurs comparant la performance d'algorithmes de transformation homographique (Karami et al., 2017; Tareen et Saleem, 2018), et de travaux évaluant la performance d'appareils d'oculométrie dans des conditions variées (Hessels et al., 2014; Niehorster et al., 2017).

La vidéo subjective simulée a été produite avec le logiciel PowerPoint. Elle est constituée de 22 diapositives regroupées en huit événements perturbateurs décrits au Tableau 1. Les diapositives sont montrées durant une seconde chacune, formant une vidéo d'une durée totale de 22 secondes. Le choix des événements perturbateurs a été motivé par le visionnement préalable d'environ 20 heures d'enregistrements d'oculométrie mobile. Le stimulus visuel apparaissant dans la vidéo imite un item d'une épreuve de compréhension de texte[2].

[2] La vidéo et les données simulées, de même que le code et les images utilisés pour produire la simulation, sont disponibles au https://github.com/gloignon/fakegazing

Tableau 1

Sommaire des événements perturbateurs composant la simulation

#	Événement	Description	# Diapositive
1	Translation avec hors-champ	Déplacement rectiligne vers le haut et la droite de manière que le haut du stimulus visuel n'apparaisse plus dans la scène visuelle.	1 - 3
2	Rotation	Rotation dans le sens horaire (5, 10 puis 15 degrés).	4 - 6
3	Transvection	Transvection horizontale puis verticale.	7 - 9
4	Luminosité	Variation de la luminosité de l'image de référence (-20%, -40%, -60%).	10 - 12
5	Flou	Flou graduel du stimulus visuel (flou gaussien de 1, 2 et 4 pixels).	13 - 15
6	Occultation	Un pictogramme représentant une main se déplace au-dessus du stimulus visuel tout en augmentant de taille, occultant différentes régions du stimulus visuel, mais sans cacher les cibles scrutées par le participant fictif.	16 -18
7	Changement d'échelle	La taille du stimulus visuel change, imitant un participant qui avance et recule (+10%, -10%, -20%).	19 - 21
8	Stimulus absent	Le fond apparait sans le stimulus visuel.	22

La cadence de la vidéo est de 25 cadres par seconde et le format est de 1920 par 1080 pixels, ce qui reproduit les spécifications de l'appareil *Tobii Pro 2*. La distorsion sphérique introduite par la lentille de l'appareil a également été émulée; nous avons obtenu par tâtonnement les paramètres réduisant le plus possible la déformation sphérique des images tirées de l'appareil, puis avons appliqué la transformation inverse aux images de la simulation. Le résultat imite donc la déformation imposée par la lentille de l'appareil *Tobii Pro Glasses 2*.

La Figure 11 est tirée de la vidéo de simulation et montre le stimulus visuel (texte et choix de réponse) tel qu'il apparait dans la 19^{e} diapositive de la séquence. On remarque que le stimulus visuel est montré conjointement à des pictogrammes et des fragments de texte qui jouent le rôle des informations visuelles qui pourraient apparaitre en périphérie dans une expérience réelle; l'objectif était alors de voir si les algorithmes de MGM seraient bernés par la présence de ce « bruit » visuel.

Figure 11. Image tirée de la vidéo créée pour simuler un enregistrement oculométrique. La région inférieure indique, de gauche à droite, le nom de l'événement, le numéro du cadre vidéo et le numéro de la diapositive.

Les données oculométriques associées à cette vidéo imitent un participant scrutant deux points du stimulus visuel : le milieu de la lettre « a » qui débute la troisième ligne et le point suivant la lettre C dans les choix de réponse. L'enregistrement simule un regard posé alternant entre ces deux cibles (500 ms par cible), testant MGM pour des coordonnées situées en périphérie et vers le centre de l'image. En raison des animations, l'emplacement exact des cibles varie selon les diapositives et a été annoté manuellement. Dans le système de coordonnées de l'image de référence, ces points devraient toujours correspondre aux positions (37, 157) et (511, 289).

Pour imiter l'appareil *Tobii Pro Glasses 2,* les enregistrements oculométriques simulés ont été cadencés à 50 Hz et la vidéo est cadencée à 25 Hz. Chaque cadre vidéo est donc associé à deux points de données, pour un total de 1100 observations. La structure des enregistrements oculométriques simulés imite le format requis par MGM : un point de données correspond à une

ligne contenant un horodatage (en millisecondes écoulées depuis le début de l'enregistrement), le numéro du cadre vidéo correspondant, et les coordonnées cartésiennes indiquant la position du regard dans la vidéo.

Traitement et analyse des données avec MGM

Les données simulées ont été analysées avec MGM, en utilisant différentes configurations (choix d'algorithmes et de valeurs de paramètres) décrites au Tableau 2. Les variables manipulées étaient le choix de l'algorithme (SIFT ou AKAZE), les paramètres de l'algorithme et le seuil de Lowe utilisé. Chaque configuration a été testée avec un seuil de Lowe de 0,5, qui est la valeur par défaut dans MGM, et avec un seuil alternatif de 0,7 qui s'était avéré prometteur lors d'essais préalables. La nomenclature des configurations indique l'algorithme, ses valeurs de paramètre, et le seuil de Lowe utilisé, par exemple AKAZE-DEF 70 pour AKAZE avec ses valeurs par défaut, et un seuil de Lowe de 0,70. Le code Python produisant ces configurations est donné en annexe et contraste les valeurs des paramètres.

Tableau 2

Configurations de MGM utilisées pour analyser les données de simulation.

Configuration		Descriptif
SIFT-DEF	50	SIFT avec ses paramètres par défaut.
	70	
SIFT-MGM	50	SIFT avec paramètres utilisés par MGM.
	70	
SIFT-ALT	50	SIFT avec paramètres alternatifs, basés sur ceux utilisés par MGM (flou
	70	moins intense, plus grand nombre de couches par octave).
AKAZE-DEF	50	AKAZE avec ses paramètres par défaut.
	70	
AKAZE-ALT	50	AKAZE avec paramètres alternatifs (descripteurs s'adaptant à
	70	l'orientation, plus grand nombre d'octaves et de couches par octave).

Les configurations SIFT-DEF et AKAZE-DEF sont basées sur les valeurs de paramètre par défaut de leurs algorithmes respectifs. Les configurations SIFT-MGM reprennent les choix des auteurs de MGM en ce qui concerne les valeurs de paramètres de SIFT, SIFT-MGM 50 est la configuration de MGM non modifiée. Les configurations SIFT-ALT et AKAZE-ALT ont été ajoutées comme alternatives robustes par suite d'essais préliminaires. SIFT-ALT est basé sur SIFT-MGM et diminue la valeur du paramètre *sigma*, lequel contrôle l'intensité du flou lors de la détection des points-clés. AKAZE-ALT utilise un descripteur s'adaptant à l'orientation, et utilise davantage d'octaves lors de la détection de points-clés. Les valeurs de paramètre de chaque configuration sont données en annexe.

Indicateurs de performance

Les configurations ont été évaluées en calculant et comparant les mesures que nous décrivons dans ce qui suit. Une explication plus détaillée des mesures de qualité pour les données d'oculométrie se trouve dans Holmqvist et al. (2012).

Exactitude. L'exactitude a été opérationnalisée comme la moyenne des distances euclidiennes entre les coordonnées cartésiennes des cibles visuelles et leurs coordonnées calculées par application projective, selon la formule suivante :

$$distance = \sqrt{(x_{réel} - x_{calculé})^2 + (y_{réel} - y_{calculé})^2}$$

Les cordonnées *x* et *y* réels sont l'emplacement réel de la cible, *x* et *y* calculés sont l'emplacement calculé par MGM.

Précision. La précision a été définie comme la distance standardisée des écarts-types, et est calculée selon la formule suivante, où sd_x et sd_y sont l'écart-type des coordonnées sur l'axe des X et des Y :

$$précision = \sqrt{{sd_x}^2 + {sd_y}^2}$$

La précision a été calculée sur des plages de données correspondant aux diapositives.

Observations manquantes. Nous considérons comme *manquantes* les observations oculométriques dont la transformation par MGM a échoué, et qu'il serait possible d'identifier comme telles dans le cadre d'une expérience empirique. Par exemple, si MGM projette le point de fixation vers des coordonnées négatives ou n'existant pas dans le stimulus visuel, les observations seront considérées comme *manquantes* dans le tableau de données résultant de MGM.

Observations aberrantes. Nous désignons *aberrantes* les observations oculométriques pour lesquelles la différence entre l'emplacement réel du point de fixation et son emplacement calculé par MGM était supérieure à un seuil. Deux seuils d'exactitude ont été utilisés : 5 et 10 pixels, qui correspondent grosso modo à la largeur des lettres minuscules et majuscules dans la simulation. Il ne serait généralement pas possible, dans le cadre d'une expérience sur données réelles, de déterminer si une donnée est ainsi aberrante puisque la position réelle du point de fixation n'est alors pas connue, d'où l'intérêt d'en estimer la proportion dans le cadre d'une stimulation. Les observations sont considérées comme *valides* si elles ne sont ni manquantes ni aberrantes.

Temps d'exécution. Le temps requis a été calculé en altérant le script Python de MGM afin d'y inclure une fonction chronomètre. Le chronomètre démarre au début de la boucle de traitement qui analyse les cadres de la vidéo. Il s'arrête lorsque le dernier cadre a été analysé. Les durées ont été arrondies à la seconde près.

Logiciels utilisés

Les images ont été produites dans PowerPoint puis exportées en format JPEG. L'effet de distorsion sphérique a été ajouté avec le logiciel *Gimp*. La conversion des images en vidéo,

l'ajout du numéro de cadre et la simulation du flou lumineux (*blooming*) ont été réalisés avec *FFmpeg*. Les coordonnées des cibles visuelles ont été extraites en examinant les diapositives à l'aide d'un programme d'édition d'image conventionnel. MGM a été exécuté dans Python version 3.6.9, sur un ordinateur Intel Xeon cadencé à 2,3 GHz et muni de 12 Go de mémoire vive. Un seul processeur était utilisé à la fois, et l'accélération du traitement par la carte vidéo était désactivée. Les simulations basées sur SIFT ont utilisé la version 3.4.2 de la bibliothèque *OpenCV* de Python – la plus récente version permettant d'utiliser SIFT, qui est depuis devenu un logiciel sous licence. Les simulations utilisant AKAZE ont utilisé OpenCV 4.2.0. Les analyses statistiques ont été effectuées avec R, version 3.6.3.

Plan d'analyses statistiques

Les analyses statistiques visaient à comparer l'issue de la transformation d'observations oculométriques par dix configurations de MGM et portaient d'abord sur la capacité à transformer adéquatement les observations, puis sur l'exactitude et la précision des observations étant demeurées valides après leur transformation. Une observation était considérée comme valide si sa transformation homographique associée n'était pas aberrante, et si l'erreur de projection (différence entre la position projetée du point et sa position réelle) était inférieure à un seuil.

Afin de vérifier si MGM analysait adéquatement des données d'oculométrie mobile, nous avons examiné la proportion d'observations valides et les types de rejets de données selon la configuration utilisée pour l'analyse. Un test du χ^2 de Pearson a été appliqué afin de vérifier si la proportion d'observations valides (calculée à un seuil souple de 10 pixels) était différente pour au moins une des configurations. Advenant la significativité statistique du test du χ^2, nous avons comparé les configurations par des tests du χ^2 *post hoc*, en utilisant SIFT-MGM 50 comme niveau de base puisque cette configuration est celle fournie avec MGM. Nous avons ensuite

examiné par statistiques descriptives quels types d'événements perturbateurs étaient susceptibles d'entrainer des observations manquantes ou aberrantes pour les différentes configurations.

La seconde partie de nos analyses porte sur l'exactitude et la précision des observations telles que transformées par MGM. Des tests de Kruskal-Wallis ont servi à déterminer si l'exactitude d'au moins une des configurations était statistiquement différente. Advenant un test statistiquement significatif, nous avons procédé à des comparaisons multiples par tests de Brunner-Munzel (Karch, 2020) en utilisant SIFT-MGM 50 comme niveau de base et en ajustant la valeur *p* par la méthode de Holm-Bonferroni. Le choix du test de Brunner-Munzel était motivé par la distribution asymétrique des données, l'implémentation était celle de la bibliothèque *brunnermunzel* pour R (Ara, 2020). Nous avons ensuite approfondi de manière descriptive les différences observées, par configuration et par événement.

6.3.2. Résultats

Les analyses statistiques portaient sur un total de 11 000 observations oculométriques, soit 1100 par configuration (50 observations par seconde, durant 22 secondes, pour 10 configurations). L'ensemble des configurations a produit des données manquantes lorsque lors de l'événement « Stimulus absent », un résultat attendu puisque le stimulus n'était pas détecté. Nous avons donc éliminé les observations associées à cet événement, portant le total à 10 500.

Tel que résumé au Tableau 3, les configurations utilisant l'algorithme SIFT avaient un temps d'exécution généralement plus long (entre 453 et 1399 secondes) que les configurations utilisant AKAZE (entre 414 et 427 secondes). La proportion de données valides variait entre 85,8% et 99,2% selon un seuil relativement souple de 10 pixels, et entre 85,4% et 99,2% en adoptant un seuil strict de 5 pixels. En excluant les données non valides, chaque configuration a produit une erreur de projection (exactitude) en moyenne inférieure à deux pixels, et une

dispersion (précision) moyenne inférieure à un pixel. Les résultats du Tableau 3 sont détaillés dans les sections suivantes.

Tableau 3

Sommaire des résultats pour les configurations testées

Configuration		Temps d'exécution (sec)	Observations valides (%)		Exactitude (pixels)		Précision (pixels)	
			Seuil 10 px	Seuil 5px	*M*	*ET*	*M*	*ET*
SIFT-DEF	50	453	90,5	88,1	1,06	2,47	0,14	0,29
	70	456	90,7	90,7	**0,72**	**0,75**	0,15	0,21
SIFT-MGM	50	*823*	*90,5*	*88,1*	*1,11*	*2,52*	*0,06*	*0,17*
	70	828	90,5	90,5	**0,72**	**0,69**	**0,05**	**0,14**
SIFT-ALT	50	1400	95,2	95,2	0,86	0,80	**0,05**	**0,14**
	70	1399	95,2	95,2	0,84	0,80	0,27	0,16
AKAZE-DEF	50	**414**	85,8	85,4	1,29	1,68	0,27	0,53
	70	419	88,8	88,8	0,96	1,14	0,28	0,44
AKAZE-ALT	50	428	97,0	97,0	1,07	1,26	0,24	0,48
	70	427	**99,2**	**99,2**	0,97	0,91	0,27	0,48

Note. Les italiques indiquent le niveau de base (configuration SIFT-MGM 50). Les caractères gras indiquent les valeurs les plus performantes. Les mesures d'exactitude et de précision ont été calculées en excluant les observations rejetées.

Validité des observations

Un test du χ^2 de Pearson a été appliqué afin d'examiner la relation statistique entre la configuration utilisée et la proportion d'observations valides (considérant un seuil de 10 pixels). Cette relation était significative, χ^2 (9) = 230,49, $p < 0{,}0001$, indiquant qu'au moins une des configurations se démarquait quant à la proportion d'observations valides. Nous avons donc procédé à des tests du χ^2 répétés afin de comparer la configuration de base, SIFT-MGM 50, aux autres configurations testées. Les configurations SIFT-ALT 50 (95,2% de données valides selon un seuil de 10 pixels), SIFT-ALT 70 (95,2%), AKAZE-ALT 50 (97%) et AKAZE-ALT 70 (99,2%) ont produit plus de données valides que la configuration de référence, SIFT-MGM 50

(90,5%). AKAZE-DEF 50 a produit moins de données valides (85,8%) que la configuration de référence. Les autres configurations ont produit une proportion de données valides statistiquement équivalente à SIFT-MGM 50. Les valeurs *p* de ces tests *post hoc*, ajustées par correction de Holm-Bonferroni, sont indiquées à la Figure 12, qui détaille l'issue des transformations non valides en fonction de la configuration utilisée. Les rejets indiqués comme « échec transformation » indiquent des observations oculométriques pour lesquelles MGM projetait le point vers des coordonnées n'existant pas dans le stimulus visuel; les rejets de type « pairages insuffisants » indiquent des données manquantes résultant d'un nombre insuffisant de paires valides de points-clés pour procéder à la transformation homographique; les autres rejets étaient causés par une erreur de projection (distance entre l'emplacement réel du point et son emplacement calculé par MGM) supérieure à 5 pixels (seuil souple) ou de 10 pixels (seuil laxe). Par exemple, pour SIFT-DEF 50, environ 10% des observations ont été rejetées en raison d'un échec lors de transformation par MGM, et 2% devraient être rejetées selon un seuil d'exactitude de 10 px, et les autres observations étaient valides.

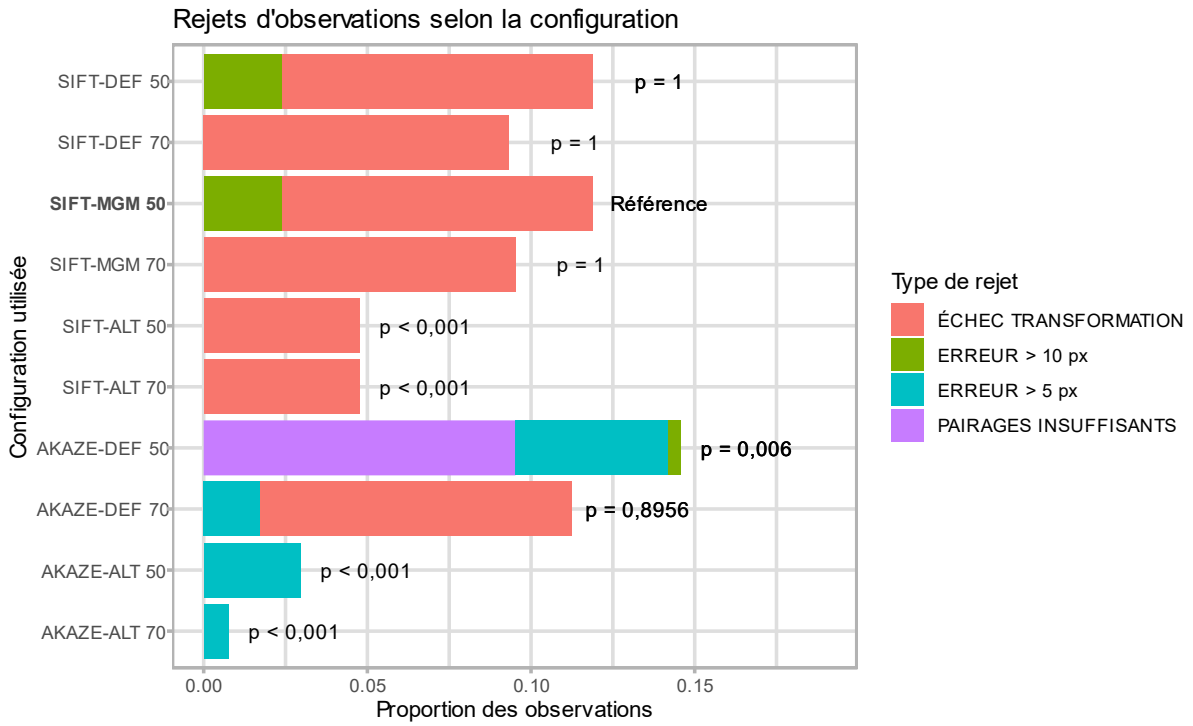


Figure 12. Proportion d'observations rejetées, selon la configuration utilisée lors des analyses. Les valeurs *p* sont issues de tests du χ^2 de Pearson et indiquent si la proportion de rejets pour cette configuration diffère du niveau de référence (SIFT-MGM 50); un ajustement de Holm-Bonferroni a été appliqué aux valeurs *p* en raison des comparaisons multiples. *ÉCHEC TRANSFORMATION* indique des observations rejetées car MGM a projeté le point vers une coordonnée inexistante, *ERREUR > 10 px* indique que le point a été projeté à plus de 10 pixels de son emplacement réel, *ERREUR > 5 px* indique que le point a été projeté à plus de 5 pixels de son emplacement réel, *PAIRAGES INSUFFISANTS* indique que la transformation ne pouvait être calculée en raison d'un manque de pairages de points-clés.

Nous avons détaillé les analyses précédentes en analysant la provenance des données non valides par configuration et par événement. Les données résultant de MGM étaient considérées comme *manquantes* lorsque le stimulus visuel n'était pas détecté (pairages insuffisants), ou lorsque la transformation était non valide d'une manière pouvant être détectée en examinant la matrice d'homographie associée. Les données étaient considérées comme *aberrantes* lorsque la différence entre la position réelle du point de fixation et sa position calculée par MGM excédait un seuil. Le Tableau 4 résume les configurations et événements ayant produit des données manquantes.

Tableau 4

Sommaire des observations manquantes

Configuration		Événement perturbateur	Cause du rejet
SIFT-DEF	50	Flou (modéré et élevé)	Transformation non valide
SIFT-DEF	70	Flou (modéré et élevé)	Transformation non valide
SIFT-ALT	50	Flou (élevé)	Transformation non valide
SIFT-MGM	50	Flou (modéré et élevé)	Transformation non valide
SIFT-MGM	70	Flou (modéré et élevé)	Transformation non valide
SIFT-ALT	70	Flou (élevé)	Transformation non valide
AKAZE-DEF	50	Rotation (10° et 15°)	Pairages insuffisants
AKAZE-DEF	70	Rotation (10° et 15°)	Transformation non valide

Toutes les configurations basées sur SIFT avaient des données manquantes, de même que AKAZE-DEF 50 et AKAZE-DEF 70. Pour les configurations SIFT, des données manquantes sont survenues uniquement lors de l'événement « Flou ». SIFT-DEF et SIFT-MGM ne parvenaient pas à analyser les images lorsque le stimulus visuel était soumis à un flou modéré ou élevé. SIFT-ALT était robuste au flou modéré, mais échouait avec un flou d'intensité élevée. AKAZE-DEF était sensible aux effets de rotation de 10 et 15 degrés. La Figure 13 illustre les invariances et sensibilités des deux algorithmes et montre que SIFT-MGM 50 détectait correctement le stimulus visuel après une rotation de 15 degrés, alors qu'AKAZE-DEF 70 échouait à partir de 10 degrés. De même, SIFT-MGM 50 ne parvenait pas à détecter le stimulus visuel avec un flou modéré, tandis que AKAZE-DEF 70 réussissait avec un flou d'intensité élevée.

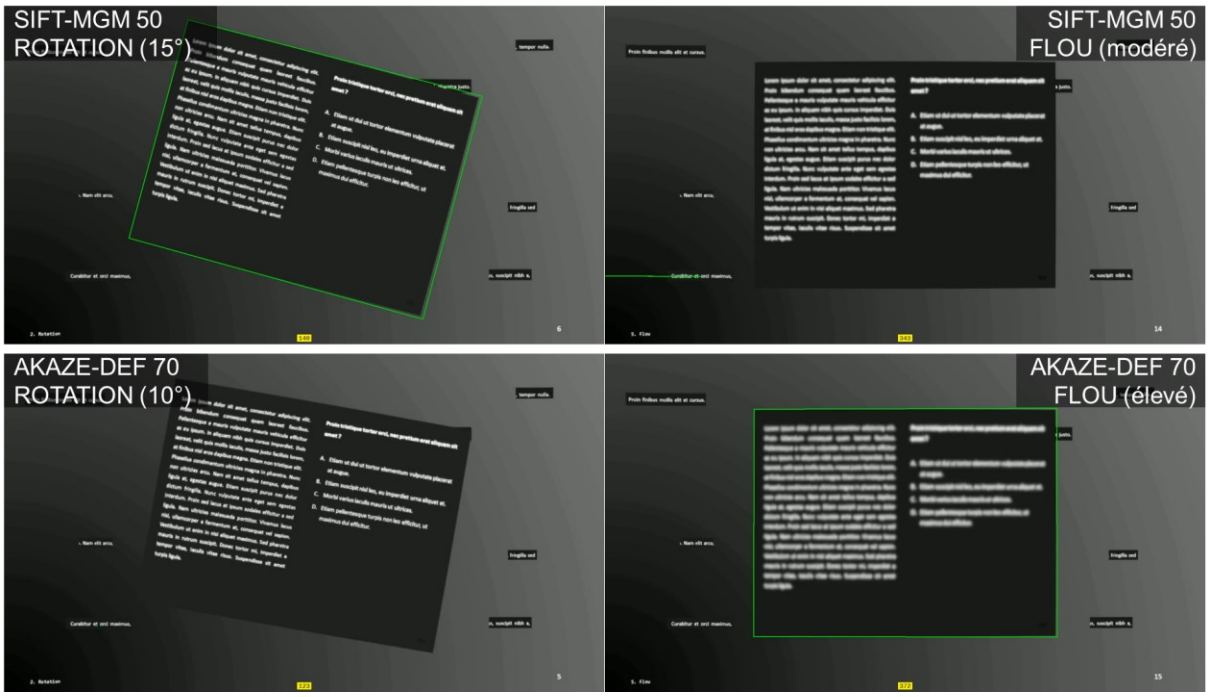


Figure 13. Arrêts sur image tirées de vidéos de débogage générées par MGM et montrant l'emplacement du stimulus visuel par un encadré vert. L'algorithme SIFT avec les paramètres de MGM est robuste aux rotations, mais sensible aux effets de flou. À l'inverse, la configuration utilisant les paramètres par défaut d'AKAZE avec un seuil de Lowe de 0,70 est sensible aux rotations, mais robuste aux effets de flou.

Le Tableau 5 montre que certaines configurations ont produit des observations aberrantes (erreurs de projection supérieure à un seuil). Considérant un seuil de 10 pixels, l'événement « Occultation » a produit 25 observations aberrantes pour SIFT-DEF 50 et SIFT-MGM 50, ce qui correspond à une demi-seconde d'enregistrement. L'événement « Transvection » a produit 4 observations aberrantes pour la configuration AKAZE-DEF 50. Considérant un seuil plus strict de 5 pixels, en plus des données aberrantes déjà résumées, « Occultation » a entrainé des observations aberrantes pour toutes les configurations employant l'algorithme AKAZE. L'événement « Transvection » a provoqué des observations aberrantes pour les configurations AKAZE-DEF 50 et AKAZE-ALT 50. L'événement « Changement d'échelle » a entrainé des observations aberrantes pour AKAZE-DEF 50.

Tableau 5

Sommaire des observations aberrantes

Configuration		Événement	N. de rejets	
			Seuil 5 px	Seuil 10 px
SIFT-DEF	50	Occultation	25	25
SIFT-MGM	50	Occultation	25	25
AKAZE-DEF	50	Transvection	4	21
		Occultation	0	24
		Échelle	0	4
AKAZE-DEF	70	Occultation	0	18
AKAZE-ALT	50	Occultation	0	6
		Transvection	0	25
AKAZE-ALT	70	Occultation	0	8

Qualité des analyses

Le but de la seconde série d'analyses statistiques était de comparer les configurations MGM à l'égard de l'exactitude et de la précision des observations une fois transformées; ces analyses excluent les données non valides. Les valeurs moyennes par configuration ont été présentées au Tableau 3. L'application d'un test de Kruskal-Wallis indique que l'exactitude différait selon la configuration utilisée, $H(9) = 36{,}464$, $p < 0{,}0001$. Les comparaisons *post hoc* par tests de Brunner-Munzel indiquent que seule la configuration AKAZE-DEF 70 produisait des observations moins exactes que la référence (SIFT-MGM 50), $p = 0{,}0116$. Les autres configurations avaient une exactitude statistiquement équivalente à la référence.

La précision différait pour au moins une des configurations, tel qu'indiqué par un test de Kruskal-Wallis, $H(9) = 34{,}185$, $p < 0{,}0001$. Les comparaisons multiples par tests de Brunner-Munzel montrent que AKAZE-DEF 70, AKAZE-ALT 50 et AKAZE-ALT 70 produisaient des données ayant une précision inférieure au niveau de référence. Les autres configurations avaient une précision statistiquement équivalente à la référence.

Le Tableau 6 fait le sommaire des valeurs moyennes d'exactitude et de précision en fonction de la configuration et de l'événement, en considérant uniquement les données valides. Les configurations SIFT-DEF 50 et SIFT-MGM 50 avaient une mesure d'exactitude supérieure à trois pixels lors de l'événement « Occultation ». La configuration AKAZE-DEF 50 avait une mesure d'exactitude supérieure à deux pixels lors des événements « Occultation » et « Rotation ». Les autres mesures d'exactitude étaient inférieures à deux pixels. La mesure de précision était inférieure à un pixel pour l'ensemble des configurations et événements perturbateurs testés. Ces valeurs demeurant sous notre seuil strict de 5 pixels, nous n'avons pas produit de statistiques inférentielles pour tester la significativité statistique des écarts observés.

Tableau 6

Exactitude et précision moyenne par événement

Configuration		Translation		Rotation		Transvection		Luminosité		Flou		Occultation		Échelle	
		E	P	E	P	E	P	E	P	E	P	E	P	E	P
SIFT-DEF	50	0,73	0,08	1,07	0,00	0,76	0,13	0,01	0,05	0,00	0,00	3,11	0,38	1,04	0,26
	70	0,51	0,05	1,28	0,11	0,30	0,16	0,17	0,23	0,43	0,16	1,03	0,18	1,06	0,19
SIFT-MGM	50	0,69	0,28	1,24	0,00	0,79	0,06	0,04	0,07	0,00	0,00	3,17	0,00	1,11	0,00
	70	0,50	0,00	1,30	0,00	0,44	0,21	0,09	0,08	0,50	0,00	0,94	0,00	1,11	0,00
SIFT-ALT	50	0,70	0,06	1,12	0,08	0,68	0,08	0,58	0,13	0,56	0,00	1,17	0,00	1,14	0,00
	70	0,71	0,00	1,03	0,12	0,52	0,16	0,53	0,16	0,50	0,00	1,17	0,00	1,28	0,00
AKAZE-DEF	50	0,65	0,18	2,29	0,00	2,01	0,40	0,65	0,16	0,09	0,14	2,12	0,60	1,88	0,26
	70	0,82	0,43	1,24	0,25	0,80	0,26	0,69	0,07	0,42	0,31	1,67	0,56	1,29	0,07
AKAZE-ALT	50	0,73	0,29	1,19	0,23	1,92	0,22	0,33	0,21	0,46	0,18	1,14	0,58	1,70	0,00
	70	0,84	0,22	1,20	0,28	0,95	0,42	0,67	0,09	0,43	0,37	1,25	0,54	1,44	0,00

Note. Exactitude (E) et précision (P) moyenne, en pixels, calculées sur les données valides uniquement. L'exactitude indique la distance moyenne entre l'emplacement estimé par MGM et l'emplacement réel; la précision indique la stabilité des mesures.

6.3.3. Discussion du volet 2

L'objectif de la présente étude était d'évaluer la capacité de MGM à transformer des données d'oculométrie mobile dans des conditions variées imitant le contexte d'une étude portant sur la lecture, tout en testant d'autres configurations de MGM qui pourraient améliorer la

performance. Nous avons produit des statistiques comparant la validité et la qualité des transformations effectuées par 10 configurations de MGM.

Synthèse des résultats

Concernant la viabilité de MGM, nos résultats indiquent que cette méthode transformait de manière valide entre 88% et 99% des observations environ, selon la configuration utilisée. En excluant les données non valides, l'erreur de projection (distance entre l'emplacement réel du regard et son emplacement calculé par MGM) demeurait sous un seuil de tolérance 5 pixels, soit environ la largeur d'un caractère dans le stimulus visuel, et moins de 1% de la diagonale du stimulus visuel (1200 pixels). La faible mesure de précision obtenue (généralement sous 1 pixel) est cohérente considérant que la dispersion des points de fixation résulte habituellement de micro-tremblements de l'œil (Holmqvist et al., 2012), lesquels n'ont pas été émulés dans notre enregistrement factice. Ce résultat demeure intéressant puisqu'il indique que les coordonnées produites par MGM étaient assez stables malgré le caractère probabiliste de l'algorithme RANSAC.

Concernant les choix d'algorithme et de paramètres de MGM, les résultats de nos analyses indiquent qu'altérer la configuration de base a permis d'améliorer la proportion de données valides, la qualité des transformations, ainsi que le temps d'exécution. Deux conclusions émergent à ce sujet. Premièrement, les valeurs de paramètre proposés par les auteurs de MGM amélioraient peu la validité et la qualité des données transformées, tout en augmentant le temps d'exécution. Adopter un seuil de Lowe moins strict (0,70 versus le 0,50 de MGM) produisait généralement des résultats similaires ou supérieurs et n'altérait pas le temps d'exécution. Deuxièmement, la configuration AKAZE-ALT 70 semble être la plus performante pour nos données de simulation. Dans l'ensemble, les configurations basées sur AKAZE traitaient les

données deux fois plus rapidement que la configuration de base SIFT-MGM 50, et trois fois plus rapidement que nos variantes SIFT-ALT. AKAZE-ALT 70 produisait la plus grande proportion d'observations valides, entrainant le rejet de moins de 1% des observations considérant un seuil de rejet strict de 5 pixels. Par comparaison, la configuration de base de MGM (SIFT-MGM 50) entrainait le rejet d'environ 12% des données considérant le même seuil. La configuration SIFT-MGM 50 était plus performante qu'AKAZE-ALT 70 pour les mesures d'exactitude et de précision, mais l'ampleur moyenne de cette différence était négligeable (inférieure à un pixel) et instable à-travers les types d'événements perturbateurs.

Notre étude a également contribué à la recherche sur la robustesse des algorithmes de vision par ordinateur SIFT et AKAZE. Les statistiques sur la provenance des rejets de données ont révélé que l'algorithme SIFT était sensible aux effets de flou, un résultat en phase avec l'étude de Ding et al. (2012) sur l'application de SIFT aux images floues. La performance des configurations SIFT-ALT indique que la sensibilité au flou peut être mitigée en augmentant le nombre de couches par octave, mais au prix d'un temps d'exécution plus long. Nous avons également relevé une sensibilité de SIFT à l'occultation, mais celle-ci disparaissait lorsqu'un seuil de Lowe moins strict (0,70, versus le 0,50 de la configuration de base de MGM) était employé. Quant à l'algorithme AKAZE, dans sa configuration par défaut, celui-ci était sensible aux effets de rotation dans sa configuration de base. AKAZE devenait robuste aux rotations en adoptant une configuration s'adaptant à l'orientation, tel que démontré par la performance d'AKAZE-ALT 70. Ce résultat va dans le sens des résultats d'essais comparatifs ayant démontré une bonne tolérance d'AKAZE aux rotations lorsque des descripteurs s'adaptant à l'orientation étaient employés (Tareen et Saleem, 2018). La configuration AKAZE-ALT 70 présentait une légère sensibilité à l'occultation du stimulus visuel qui n'était pas présente lorsque MGM utilisait

la configuration AKAZE-DEF 50. Cette différence pourrait s'expliquer par la difficulté à former des pairages valides lorsque l'image est partiellement dissimulée, combinée au fait qu'AKAZE-ALT 70 détecte les points-clés de manière moins stricte.

Limites et avenues de recherche

La présente étude constitue une première tentative d'évaluation de l'annotation automatisée de données d'oculométrie mobile. Bien que nos résultats appuient la validité de cette méthode, nous considérons que d'autres travaux pourraient préciser nos conclusions. Nos données simulées ne reproduisaient pas toutes les caractéristiques d'un enregistrement oculométrique, notamment les saccades et autres mouvements oculaires. Bien que nous ayons tenté de reproduire des événements perturbateurs authentiques (flous, déplacements, etc.) il est possible que notre simulation n'ait pas capturé de manière assez précise les caractéristiques pouvant nuire à la qualité d'un enregistrement réel. Pour éclairer cette question, il serait pertinent de répéter l'expérience en comparant la qualité des analyses manuelles et automatiques sur un enregistrement véritable. Par ailleurs, nous n'avons testé que deux algorithmes et quelques configurations de MGM. Il pourrait être avantageux, pour identifier les meilleures configurations selon les conditions, de tester systématiquement des configurations représentant les fourchettes plausibles de chaque paramètre.

Implications

Malgré ses limitations, la présente étude suggère plusieurs implications méthodologiques et théoriques. Par exemple, nous avons montré que MGM, ou une méthode similaire, pouvait être intégrée à une chaine de traitement des données d'oculométrie mobile, avec un effet minimal sur le nombre d'observations et sur la qualité des données. Considérant le type de stimulus visuel et d'événements perturbateurs utilisés durant l'expérimentation, il nous semble réaliste d'intégrer la

méthode MGM à l'étude des processus de lecture par l'oculométrie. Notre étude a de plus contribué au domaine de la vision par ordinateur en testant la robustesse des algorithmes SIFT et AKAZE sous différentes configurations et dans des contextes variés.

6.4. Sommaire et conclusions générales

Nous avons contribué à la démocratisation de l'oculométrie en testant si les limitations dues au caractère subjectif des données d'oculométrie mobile pouvaient être contournées en recourant à une méthode automatisée dérivée de l'intelligence artificielle. Dans le premier volet de l'article, nous avons présenté une documentation auparavant inexistante pour la boite à outils MGM. En illustrant le fonctionnement de MGM, notre intention était de faire une validation *a priori* de cette méthode en montrant que ses bases théoriques et méthodologiques sont solides. Les notions d'oculométrie et de vision par ordinateur présentées dans cette section pourraient s'appliquer à d'autres contextes de recherche, et servir de base théorique pour la conception de solutions analogues à MGM. Dans le second volet de l'article, nous avons testé MGM dans un contexte contrôlé reproduisant des situations survenant lors d'enregistrements d'oculométrie mobile, comme des mouvements de la tête et des flous. Combinés aux quelques études ayant testé MGM ou des méthodes similaires en situation réelle, nos essais comparatifs appuient l'idée que l'oculométrie mobile est une avenue viable pour l'étude des processus reliés à l'attention visuelle. Nous avons formulé quelques recommandations pour améliorer la qualité des données analysées par cette méthode, et anticipons un intérêt pour l'étude des stratégies de lecture, qui est une application fréquente de l'oculométrie stationnaire et pourrait profiter de l'aspect moins contraignant des appareils mobiles.

6.5. Annexe du troisième article

6.5.1. Code source pour reparamétriser MGM

SIFT DEF (valeurs par défaut de l'algorithme SIFT)

```
# définit la méthode de détection de features
featureDetect = cv2.xfeatures2d.SIFT_create(
  sigma = 1.6)

# définit la méthode d'appariement de features
index_params = dict(algorithm = FLANN_INDEX_KDTREE, trees = 4)
search_params = dict(checks=32)
matcher = cv2.FlannBasedMatcher(index_params, search_params)
```

SIFT MGM (configuration spécifiée par les auteurs de MGM)

```
featureDetect = cv2.xfeatures2d.SIFT_create(
  sigma = 1.6)  # def: 1.6
index_params = dict(algorithm = FLANN_INDEX_KDTREE, trees = 5)  # def: trees = 4
search_params = dict(checks=10)  # def: checks = 32
matcher = cv2.FlannBasedMatcher(index_params, search_params)
```

SIFT ALT (valeurs suggérées améliorant la robustesse aux flous)

```
featureDetect = cv2.xfeatures2d.SIFT_create(
  sigma = 1.4)  # diffère de la valeur défaut de 1.6, également utilisée par MGM
index_params = dict(algorithm = FLANN_INDEX_KDTREE, trees = 5)  # def: trees = 4
search_params = dict(checks = 10)  # def: checks = 32
matcher = cv2.FlannBasedMatcher(index_params, search_params)
```

AKAZE DEF (valeurs par défaut pour l'algorithme AKAZE)

```
featureDetect = cv2.AKAZE_create(
        descriptor_type = 4  # descripteur vertical,
        nOctaves = 4,  # 4 octaves
```

```
        nOctaveLayers = 4  # 4 couches de flou
    )
matcher = cv2.BFMatcher(normType = cv2.NORM_HAMMING)
```

AKAZE ALT (configuration la plus performante selon nos essais)

```
featureDetect = cv2.AKAZE_create(
    descriptor_type = 5,  # descripteur s’adaptant à l’orientation, déf = 4
    nOctaves = 6,  # valeur par défaut est 4
    nOctaveLayers = 4  # valeur par défaut est 4
)
matcher = cv2.BFMatcher(normType = cv2.NORM_HAMMING)
```

6.6. Bibliographie

Alcantarilla, P. F., Bartoli, A. et Davison, A. J. (2012). KAZE features. Dans A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato et C. Schmid (dir.), *Computer Vision – ECCV 2012* (vol. 7577, p. 214-227). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-33783-3_16

Alcantarilla, P., Nuevo, J. et Bartoli, A. (2013). *Fast explicit diffusion for accelerated features in nonlinear scale spaces*. In Proceedings of the British Machine Vision Conference 2013 (p. 13.1-13.11). https://doi.org/10.5244/C.27.13

Ara, T. (2020). *brunnermunzel: (Permuted) Brunner-Munzel Test* (version 1.4.1). https://CRAN.R-project.org/package=brunnermunzel

Bax, S. (2013). The cognitive processing of candidates during reading tests: Evidence from eye-tracking. *Language Testing*, *30*(4), 441-465. https://doi.org/10.1177/0265532212473244

Cognolato, M., Atzori, M. et Müller, H. (2018). Head-mounted eye gaze tracking devices: An overview of modern devices and recent advances. *Journal of Rehabilitation and Assistive Technologies Engineering*, *5*. https://doi.org/10.1177/2055668318773991

Dalrymple, K. A., Manner, M. D., Harmelink, K. A., Teska, E. P. et Elison, J. T. (2018). An examination of recording accuracy and precision from eye tracking data from toddlerhood to adulthood. *Frontiers in Psychology*, *9*. https://doi.org/10/gdnzp3

Duchowski, A. (2007). *Eye Tracking Methodology: Theory and Practice* (2 edition). Springer.

Fong, A., Hoffman, D. et Ratwani, R. M. (2016). Making sense of mobile eye-tracking data in the real-world: A human-in-the-loop analysis approach. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *60*(1), 1569-1573. https://doi.org/10/ggqr4r

Grazioso, M., Esposito, R., Maayan Fanar, E., Kuflik, T. et Cutugno, F. (2020). Using eye tracking data to understand visitors' behaviour. Dans *Proceedings of the AVI2CH Workshop on Advanced Visual Interfaces and Interactions in Cultural Heritage* (vol. 2687). http://ceur-ws.org/Vol-2687/paper6.pdf

Hessels, R. S., Cornelissen, T. H., Kemner, C., & Hooge, I. T. (2015). Qualitative tests of remote eyetracker recovery and performance during head rotation. *Behavior Research Methods*, 47(3), 848-859. https://10.3758/s13428-017-0863-0

Holmqvist, K., Nyström, M. et Mulvey, F. (2012). Eye tracker data quality: What it is and how to measure it. *Eye Tracking Research and Applications Symposium (ETRA)*. https://doi.org/10/f25r34

Hooge, I. T. C., Holleman, G. A., Haukes, N. C. et Hessels, R. S. (2019). Gaze tracking accuracy in humans: One eye is sometimes better than two. *Behavior Research Methods*, *51*(6), 2712-2721. https://doi.org/10/gf3gc5

Ikeuchi, K. (2014). *Computer Vision: A Reference Guide*. Springer US.

Imbert, J.-P., Hurter, C., Peysakhovich, V., Blättler, C., Dehais, F. et Camachon, C. (2015). Design requirements to integrate eye trackers in simulation environments: Aeronautical use case. Dans R. Neves-Silva, L. C. Jain et R. J. Howlett (dir.), *Intelligent Decision Technologies* (vol. 39, p. 231-241). Springer International Publishing. https://doi.org/10.1007/978-3-319-19857-6_21

Kahraman, H. (2019). Reading as a single construct: A process-oriented study. *Novitas-ROYAL (Research on Youth and Language)*, *13*(2), 206-220. https://eric.ed.gov/?id=EJ1231980

Karami, E., Prasad, S., & Shehata, M. (2017). *Image Matching Using SIFT, SURF, BRIEF and ORB: Performance Comparison for Distorted Images*. arXiv preprint arXiv:1710.02726.

Karch, J. D. (2021). Psychologists should use Brunner-Munzel's instead of Mann-Whitney's U test as the default nonparametric procedure. *Advances in Methods and Practices in Psychological Science*, 4(2). https://doi.org/10.1177/2515245921999602

Khan, N. Y., McCane, B. et Wyvill, G. (2011, décembre). SIFT and SURF performance evaluation against various image deformations on benchmark dataset. Dans *2011 International Conference on Digital Image Computing: Techniques and Applications* (p. 501-506). https://doi.org/10/gjsfnf

Krig, S. (2016). *Computer Vision Metrics: Textbook Edition*. Springer.

Kunze, K., Kawaichi, H., Yoshimura, K. et Kise, K. (2013). Towards inferring language expertise using eye tracking. Dans *CHI '13 Extended Abstracts on Human Factors in Computing Systems, Paris, France (p. 217).* https://doi.org/10.1145/2468356.2468396

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, *60*(2), 91-110. https://doi.org/10/bqrmsp

MacInnes, J. J. (2020). *Mobile Gaze Mapping* [Python]. https://github.com/jeffmacinnes/mobileGazeMapping

MacInnes, J. J., Iqbal, S., Pearson, J. et Johnson, E. N. (2018a). *Wearable Eye-tracking for Research: Automated dynamic gaze mapping and accuracy/precision comparisons across devices*. bioRxiv. https://doi.org/10.1101/299925

Macinnes, J. J., Iqbal, S., Pearson, J. et Johnson, E. (2018b). Mobile Gaze Mapping: A Python package for mapping mobile gaze data to a fixed target stimulus. *Journal of Open Source Software*, *3*(31), 984. https://doi.org/10/ggqr6q

Miranda, A. M., Nunes-Pereira, E. J., Baskaran, K. et Macedo, A. F. (2018). Eye movements, convergence distance and pupil-size when reading from smartphone, computer, print and

tablet. *Scandinavian Journal of Optometry and Visual Science*, *11*(1), 1-5. https://doi.org/10/gg2bw3

Niehorster, D. C., Cornelissen, T. H., Holmqvist, K., Hooge, I. T., & Hessels, R. S. (2018). What to expect from your remote eye-tracker when participants are unrestrained. *Behavior Research Methods*, 50(1), 213-227. https://10.3758/s13428-017-0863-0

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372. https://doi.org/10/b5gdv6

Sánchez, D. O., Ramírez, A. A. V., González-Becerra, V. H., Abundis-Gutiérrez, A., Río, J. M. del, Capilla, L. A. Z., Huerta, J. R. A. et López, M. A. (2018). Reading Comprehension and Eye-Tracking in College Students: Comparison between Low- and Middle-Skilled Readers. *Psychology*, *9*(15), 720-726. https://doi.org/10/ggbv74

Tareen, S. A. K. et Saleem, Z. (2018, mars). A comparative analysis of SIFT, SURF, KAZE, AKAZE, ORB, and BRISK. Dans *2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)* (p. 1-10). https://doi.org/10.1109/ICOMET.2018.8346440

Toyama, T., Kieninger, T., Shafait, F. et Dengel, A. (2012). Gaze guided object recognition using a head-mounted eye tracker. *Dans ETRA '12: Proceedings of the Symposium on Eye Tracking Research and Applications* (p. 91-98). https://doi.org/10/ghszph

Uchiyama, H. et Saito, H. (2009). Augmenting text document by on-line learning of local arrangement of keypoints. Dans *2009 8th IEEE International Symposium on Mixed and Augmented Reality* (p. 95-98). https://doi.org/10/cfxxwb

Vansteenkiste, P., Cardon, G. et Lenoir, M. (2013). Dealing with head-mounted eye-tracking data: Comparison of a frame-by-frame and a fixation-based analysis. Dans *Proceedings*

of the 2013 Conference on Eye Tracking South Africa. https://doi.org/10.1145/2509315.2509325

Wu, J., Cui, Z., Sheng, V. S., Zhao, P., Su, D. et Gong, S. (2013). A Comparative Study of SIFT and its Variants. *Measurement Science Review*, *13*(3), 122-131. https://doi.org/10.2478/msr-2013-0021

Zhan, Z., Wu, J., Mei, H., Wu, Q. et Fong, P. S. W. (2020). Individual difference on reading ability tested by eye-tracking: from perspective of gender. *Interactive Technology and Smart Education*, *17*(3), 267-283. https://doi.org/10/ghbttb

Zheng-Jian Ding, Yang Zhang, A-Qing Yang, et Dai-Li. (2012). Image matching of Gaussian blurred image based on SIFT algorithm. Dans *2012 International Conference on Wavelet Active Media Technology and Information Processing (ICWAMTIP)* (p. 121-124). https://doi.org/10/gh43b2

7. Transition entre les articles 3 et 4

L'article 3 a détaillé le fonctionnement de la boite à outils *Mobile Gaze Mapping* (MGM) et a testé ses capacités en simulant le contexte expérimental d'une étude de la lecture par l'oculométrie mobile. Nos essais comparatifs ont montré que, malgré quelques limitations que nous avons pu mitiger en altérant le code source de MGM, cette technique offrait de bonnes performances sur des données simulées. Nous n'avons cependant pas testé MGM pour des données collectées en situation réelle. C'est l'objectif de l'article 4, qui décrit une étude pilote d'oculométrie mobile portant la compréhension de textes argumentatifs, dont la chaîne de traitement des données comprenait MGM.

8. Quatrième article :

L’étude des processus de lecture par l’oculométrie mobile : une étude pilote

Guillaume Loignon

Université de Montréal

Nathalie Loye

Université de Montréal

Contribution des auteurs :

Loignon a conçu les items, le dispositif expérimental, et le plan d’analyses, il a dirigé la collecte de données, effectué les analyses et rédigé la majorité de l’article.

Loye a fait d’importantes suggestions méthodologiques, a guidé certains aspects de la conception des items, et a apporté un grand nombre de révisions au texte de l’article.

Résumé

Les appareils d'oculométrie de type mobile facilitent l'étude de l'attention visuelle dans un contexte naturel, un avantage par rapport aux appareils stationnaires généralement utilisés dans les études portant sur les processus de lecture. Les données d'oculométrie mobile sont cependant influencées par les mouvements du participant, ce qui en complexifie l'analyse et limite le type de mesures pouvant être produites. Récemment, une solution à ce défi méthodologique a été proposée sous la forme de la bibliothèque *Mobile Gaze Mapping* pour Python (Macinnes et al., 2018), qui s'appuie sur des techniques de vision par ordinateur (une branche de l'intelligence artificielle). Dans le cadre d'une étude pilote d'oculométrie mobile, nous avons cherché à savoir si une chaine de traitement de données intégrant *Mobile Gaze Mapping* rendait possible une analyse des processus de compréhension de texte par l'oculométrie mobile. Les participants étaient des élèves de niveau postsecondaire et un groupe d'experts constitué de professionnels de l'éducation, tous devaient répondre à des questions à choix multiple portant sur de courts textes argumentatifs. Les résultats indiquent que la méthode proposée permet d'émettre des inférences vérifiables concernant la fluidité en lecture et les stratégies de compréhension de texte. Cette méthode a donc un fort potentiel pour de futures études portant sur la lecture, mais aussi plus largement pour des travaux s'intéressant à l'attention visuelle et dans lesquelles les participants doivent demeurer libres de leurs mouvements.

Mots-clés: oculométrie, lecture, compréhension de texte, vision par ordinateur, psycholinguistique

Abstract

Mobile eye tracking devices facilitate the study of visual attention in a natural context, which is an advantage over the stationary devices generally used in studies of reading processes. However, data recorded with a mobile eye tracking device is influenced by the participant's movements, which complexifies the analysis and limits the type of measurements that can be produced. Recently, the *Mobile Gaze Mapping* (Macinnes et al., 2018) package for Python became available, providing an automated analysis solution based on computer vision techniques (a branch of artificial intelligence). Through a mobile eye tracking pilot study), we investigated whether a data processing pipeline integrating Mobile Gaze Mapping would make it possible to analyze text comprehension processes through mobile eye tracking. Participants were a group of post-secondary students, and a group of education professionals, all of whom were asked to answer multiple-choice questions about short argumentative texts. Results indicate that the proposed methodology allows researchers to make testable inferences about reading fluency and reading comprehension strategies. We discuss the potential and challenges of the proposed methodology for future studies on reading, but also more broadly for work involving visual attention and in which the participant must remain free to move.

Key words : eye tracking, reading, text comprehension, computer vision, psycholinguistics

Quatrième article :

L'étude des processus de lecture par l'oculométrie mobile : une étude pilote

L'oculométrie est l'enregistrement de la position et du mouvement du regard dans le but d'étudier l'attention visuelle, définie comme l'action de sélectionner la région d'un stimulus visuel vers laquelle on oriente notre regard (Duchowski, 2007). Le mouvement des yeux est capté par un appareil nommé oculomètre. Il en existe deux grands types : les appareils stationnaires qui sont fixés au moniteur affichant les stimuli visuels, à la manière d'une caméra Web, et les appareils mobiles qui sont portés sur la tête à la manière de lunettes. Une application courante de l'oculométrie est l'étude des processus de lecture (Gernsbacher & Kaschak, 2006). Le postulat sous-jacent à l'étude oculométrique de la lecture est que l'emplacement du regard dans le texte correspond au mot en cours de traitement. Une caractéristique psychophysiologique appuie cette hypothèse : les informations visuelles à grain fin, comme le texte, ne peuvent être décodées que si elles ont été captées par la région fovéale de la rétine, qui contient la plus grande densité de cellules photoréceptrices (Rayner, 1998). Les oculomètres employés pour des expérimentations portant sur la lecture sont généralement de type stationnaire. Ce choix d'équipement est motivé d'abord par le fait que les oculomètres stationnaires ont une résolution spatiale et temporelle (cadence d'enregistrement) plus élevée que leurs équivalents mobiles (Cognolato et al., 2018). Un autre avantage de l'oculomètre stationnaire est que l'enregistrement produit est stable; la position du regard est toujours enregistrée relativement à un point fixe. Ainsi, il est relativement simple de déterminer quelle région du stimulus visuel était observée à un moment spécifique. Toutefois, la stabilité de l'oculométrie stationnaire exige généralement que les participants demeurent immobiles, la tête parfois soutenue par un appui-menton, ce qui peut devenir inconfortable. La collecte de données s'éloigne alors d'une situation naturelle, ce

qui soulève des questions quant à la généralisabilité des résultats au-dehors du contexte expérimental.

8.1.1. L'oculométrie, mobile et stationnaire

Nous nous sommes intéressés dans cet article à l'étude des processus de lecture par l'oculométrie mobile dans le but de favoriser une collecte de données dans un contexte plus écologique. Les oculomètres mobiles sont munis d'une caméra frontale qui enregistre une vidéo imitant le champ de vision des participants, et c'est dans le plan de référence de cette vidéo que la position du regard du participant est enregistrée. L'analyse de données d'oculométrie mobile pose donc un défi méthodologique important : puisque l'appareil est porté sur la tête, il peut être complexe de distinguer les mouvements des yeux des mouvements induits par les déplacements du participant. Plusieurs études d'oculométrie mobile limitent donc leurs analyses à une annotation plus ou moins qualitative des vidéos tirées de la caméra frontale de l'appareil, dans lesquelles on aura préalablement superposé une cible représentant le point fixé par le participant. Bien qu'il existe des logiciels spécialisés pour stabiliser l'enregistrement, ceux-ci exigent que l'emplacement du stimulus visuel soit identifié manuellement dans chaque image de la vidéo, un processus fastidieux et prédisposés à l'erreur subjective.

Pour automatiser la stabilisation des enregistrements d'oculométrie mobile, MacInness et collègues ont publié la boite à outils *Mobile Gaze Mapping* (ci-après, MGM) pour le langage Python. MGM propose de transformer les données captées par différents modèles d'appareils d'oculométrie stationnaires afin qu'elles prennent la forme des données normalement produites par un appareil stationnaire (Macinnes, Iqbal, Pearson et Johnson, 2018; MacInnes et al., 2018). Pour ce faire, MGM fait appel à plusieurs techniques du domaine de la vision par ordinateur. La transformation consiste à détecter, dans la vidéo produite par l'oculomètre mobile, une image de

référence correspondant au stimulus visuel utilisé lors de l'expérience. MGM peut ensuite traduire les coordonnées du point de fixation vers le plan fixe du stimulus visuel, imitant le format d'enregistrement des oculomètres stationnaires. Tout en ayant le potentiel de mitiger certains inconvénients de l'oculométrie mobile, au moment d'écrire ces lignes, MGM était l'objet de très peu de publications et n'avait pas été utilisé dans le cadre d'une étude publiée portant sur les processus de lecture. Lors d'une étude précédente, nous avons montré que MGM pouvait transformer, par application de techniques de vision par ordinateur, des données simulant une étude portant sur la compréhension de texte (article 3 de la présente thèse). L'efficacité de MGM dans le cadre d'une étude d'oculométrie portant sur la lecture demeure cependant à valider.

8.1.2. Oculométrie et processus de lecture

Le défi de l'oculométrie mobile est, en somme, d'identifier ce que regarde la personne qui porte les lunettes d'oculométrie, par exemple quelle région du texte est fixée à un moment spécifique de l'enregistrement. L'attention visuelle est ponctuée de *fixations* et de *saccades*. Une fixation est une période d'une durée d'environ 200 à 300 ms où les yeux sont relativement immobiles, le sujet fixant un point dans la scène visuelle; une saccade est un mouvement bref et rapide des yeux permettant le repositionnement du regard sur une nouvelle cible visuelle (Duchowski, 2007; Rayner, 1998). L'analyse des données oculométriques cible typiquement les fixations. Le traitement visuel étant suspendu pendant les saccades, celles-ci ne sont généralement pas considérées en soi comme des indicateurs des processus d'attention visuelle (Thiele et al., 2002). Les données d'un enregistrement oculométrique prennent ainsi la forme d'une liste chronologique de fixations indiquant les coordonnées cartésiennes du point de fixation.

En accord avec les modèles cognitifs de la lecture tels que celui de Coltheart (2005), des travaux de psycholinguistique utilisant l'oculométrie ont montré que le temps de fixation sur un mot est un indicateur de son traitement cognitif (Juhasz et Rayner, 2006). Les mots sont plus longuement fixés lorsque leur fréquence d'occurrence est plus faible, ou lorsqu'ils sont difficiles à inférer par le contexte de la phrase (Rayner et Well, 1996). De même, la durée de fixation augmente, par exemple, lors de passages de textes conceptuellement plus complexes (Rayner, 1998), ou lorsque la complexité syntaxique est accrue (Demberg et Keller, 2008). Plusieurs études ont utilisé des statistiques portant sur la durée des fixations comme des indicateurs pour mesurer la charge cognitive (Irwin, 2004; Wang et al., 2014). La durée des fixations peut ainsi être utilisée comme un indicateur de la fluidité en lecture, définie comme la capacité à reconnaitre les mots de manière automatisée (Kuhn et Stahl, 2003). La complexité du texte demeure relative au niveau d'habileté du lecteur; les lecteurs ayant un niveau d'habileté plus faible auront en moyenne des fixations plus longues (Ashby et al., 2005). Bien que peu étudiée, la variabilité de la durée des fixations peut également servir d'indicateur de fluidité, une plus grande variabilité indiquant un flot de lecture moins constant. Par exemple, Smith et al. (2018) ont montré que, lors de tâches de lecture, l'écart-type intra-individuel de la durée de fixation était plus élevé chez un groupe d'individus aphasiques que pour un groupe contrôle neurotypique; les groupes étaient cependant équivalents lors de tâches de repérage visuel simples.

D'autres mesures d'intérêt peuvent être dérivées des fixations, comme les visites, qui sont des plages de fixations pendant lesquelles le regard du participant reste dans une région définie du stimulus visuel, ou *zone d'intérêt*. L'analyse des plages de fixation consécutives sur une même région permet d'étudier l'allocation des ressources attentionnelles à un niveau plus macroscopique. Les ressources attentionnelles et temporelles étant limitées, il est attendu qu'un

lecteur plus habile répartira davantage son attention visuelle entre les zones dont le contenu est pertinent pour effectuer la tâche (Gegenfurtner et al., 2011). Le nombre et la durée des visites sur une phrase contenant une information clé peuvent donc être utilisés comme indicateurs de l'habileté en compréhension de texte (Bax, 2013; Bax et Chan, 2019). De même, la séquence dans laquelle les zones d'intérêt sont observées et la division de l'attention visuelle entre les zones peuvent indiquer différentes stratégies de compréhension qui varient également selon le niveau d'habileté (Solheim et Uppstad, 2011; Stofer et Che, 2014). En ce sens, l'étude oculométrique de la compréhension de textes s'inscrit dans la lignée des travaux en psycholinguistique et en orthophonie montrant que les stratégies de compréhension de textes varient en fonction de la compétence générale en lecture (Castles et al., 2018; Ecalle et al., 2008) et en fonction de l'expertise dans la thématique du texte (Shanahan et al., 2011).

8.1.3. Objectifs de la présente étude

La présente étude a donc cherché à déterminer si l'intégration de MGM dans une chaîne de traitement des données faciliterait l'application de l'oculométrie mobile au domaine de la compréhension de texte. À cette fin, nous avons mené une étude pilote en utilisant les lunettes *Tobii Pro Glasses 2*. Dans cette étude, un groupe d'élèves et un groupe d'experts ont été invités à répondre à des questions à choix multiples sur de courts textes argumentatifs. Les enregistrements oculométriques ont été transformés en utilisant MGM. Nous cherchons ainsi à répliquer les résultats typiquement observés par des études en oculométrie stationnaire, à savoir que les experts ont des durées de fixation plus courtes et portent une attention visuelle plus importante aux régions pertinentes du stimulus visuel (voir la métaétude de Gegenfurtner et al., 2011). Nous avons donc retenu les deux hypothèses suivantes pour notre étude pilote, celles-ci sont opérationnalisées plus loin dans la section *Plan d'analyse* :

Hypothèse 1 : fluidité de la lecture. La fluidité de la lecture, estimée en analysant la durée des fixations, devrait être moindre chez les élèves. Conformément aux principes de psycholinguistique dont nous avons fait le survol, une lecture moins fluide devrait se manifester par une durée de fixation à la fois plus longue et plus variable au niveau intra-individuel.

Hypothèse 2 : stratégies de compréhension de texte. Les deux groupes devraient différer quant à la répartition de l'attention accordée aux différentes parties de l'item. Cette hypothèse sera vérifiée en comparant la répartition des visites (nombre de visites accordées à chaque zones) et le temps total accordé aux différentes zones (*dwell time*).

8.2. Méthodologie

8.2.1. Participants

Les participants étaient 10 élèves[1] d'un cégep de la grande région de Montréal (âge moyen 21,2 ans; écart-type : 5,39) et 8 collaborateurs experts qui ont accepté de répondre à une série d'items de compréhension de texte à choix multiple avec enregistrement audio et oculométrique. Les élèves ont été recrutés par affiches et par des visites dans les cours de philosophie et de psychologie[2]. Le groupe de collaborateurs experts a été recruté via notre réseau de contacts et était constitué de cinq enseignants de philosophie au cégep, un enseignant de philosophie au secondaire, un spécialiste de la mesure ayant aussi une formation en philosophie, et une orthophoniste. Les données oculométriques d'un élève parmi les 10 et de deux experts sur huit n'ont pu être utilisées puisque la calibration des lunettes n'était pas adéquate en raison d'une

[1] Les participants élèves recevaient une carte-cadeau d'une valeur de 15 dollars. Le projet a été approuvé par le comité éthique du cégep où la collecte de données a eu lieu, de même que par Comité d'éthique de la recherche en éducation et en psychologie de l'Université de Montréal (CEREP-18-002-D).

[2] Notre étude s'inscrit dans un projet plus vaste visant à créer une épreuve de dépistage pour le domaine de la philosophie au collégial, qui repose largement sur l'étude de textes argumentatifs.

dioptrie ne pouvant être corrigée par les lentilles de la trousse accompagnant l'appareil *Tobii Pro Glasses 2*.

8.2.2. Items utilisés

Les analyses portent sur huit items tirés d'une épreuve de compréhension de texte ciblant les compétences d'argumentation. Les items prenaient la forme de 4 vignettes dans lesquels un individu fictif exprime un raisonnement dans un paragraphe de texte sur lequel portaient entre 1 et 4 questions à choix multiples. Le dispositif expérimental est montré à la Figure 1, le format de présentation des items à la Figure 2. Pour cinq des huit items, une phrase du paragraphe était soulignée; les participants devaient alors, par exemple, choisir l'énoncé qui pourrait le mieux remplacer le passage souligné. L'épreuve contenait initialement 5 vignettes et 15 items modifiés par suite de consultations avec un enseignant de philosophie, une orthopédagogue et une didacticienne. Sur ces 15 items, 7 ont été exclus des analyses car ils n'ont pas été vus par tous les participants, les items utilisés sont montrés en annexe de cet article.

8.2.3. Instrumentation

Notre étude employait des lunettes d'oculométrie *Tobii Pro Glasses 2* et leur trousse de verres correcteurs. La cadence d'enregistrement des données oculométriques était de 50 Hz. L'enregistrement vidéo de la caméra frontale avait une résolution de 1920 par 1080, un angle de vision de 82° par 52°, et était cadencé à 25 Hz. L'appareil est également muni d'un microphone qui enregistrait une piste audio. Les enregistrements ont été pilotés avec le logiciel *Tobii Pro Glasses Controler.*

Le moniteur utilisé pour afficher les items était un écran d'ordinateur conventionnel dont la surface d'affichage mesurait environ 38 cm de largeur par 31cm de hauteur, et une résolution de 1466 par 768 pixels. Il était placé à environ 60 cm du participant et la hauteur et l'angle

étaient ajustés selon la taille du participant afin de permettre une posture naturelle et de minimiser les mouvements. Un panneau de carton mousse (*foamcore*) noir de 125 cm de largeur par 50 cm de hauteur était placé derrière l'écran afin de limiter le champ visuel et diminuer la variation de l'intensité lumineuse pouvant affecter la caméra frontale. Les items du questionnaire étaient présentés à l'aide du logiciel PowerPoint à partir d'un ordinateur portable relié au moniteur par un câble VGA. Le dispositif expérimental est montré à la Figure 1.

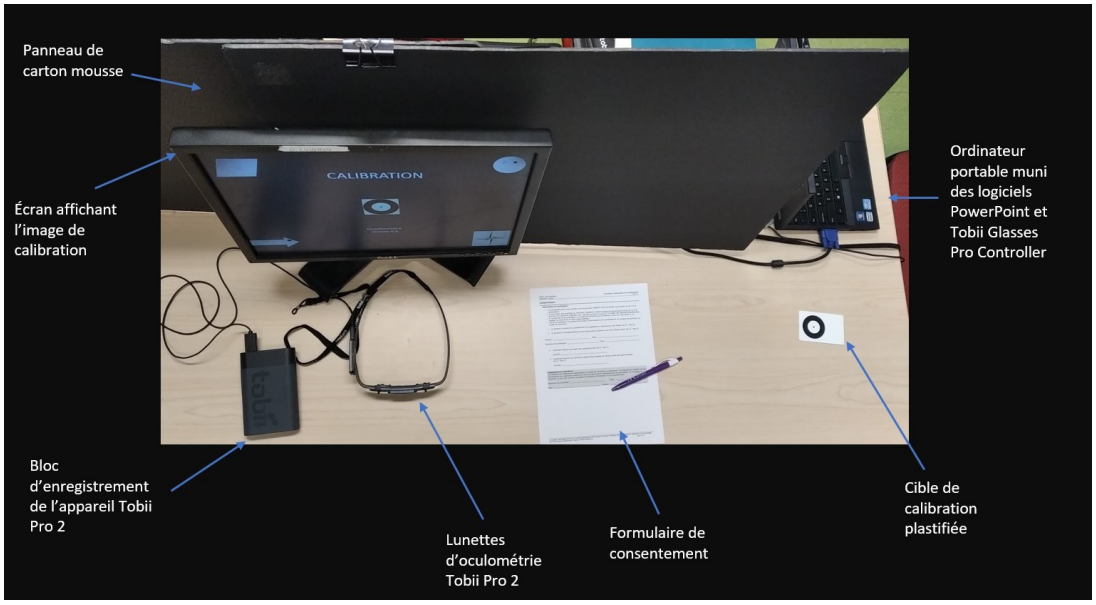


Figure 1. Configuration de la station de collecte de données.

8.2.4. Procédures

Cartographie des zones d'intérêt

Les coordonnées des zones d'intérêt dans les images représentant les items ont été cartographiées manuellement, puis notées dans un tableau. Les zones étaient composées d'une ou plusieurs régions rectangulaires comme le montre la Figure 2, il s'agissait du paragraphe de texte de la vignette, du passage souligné lorsqu'il y en avait un, de l'énoncé de la question, et des choix de réponse.

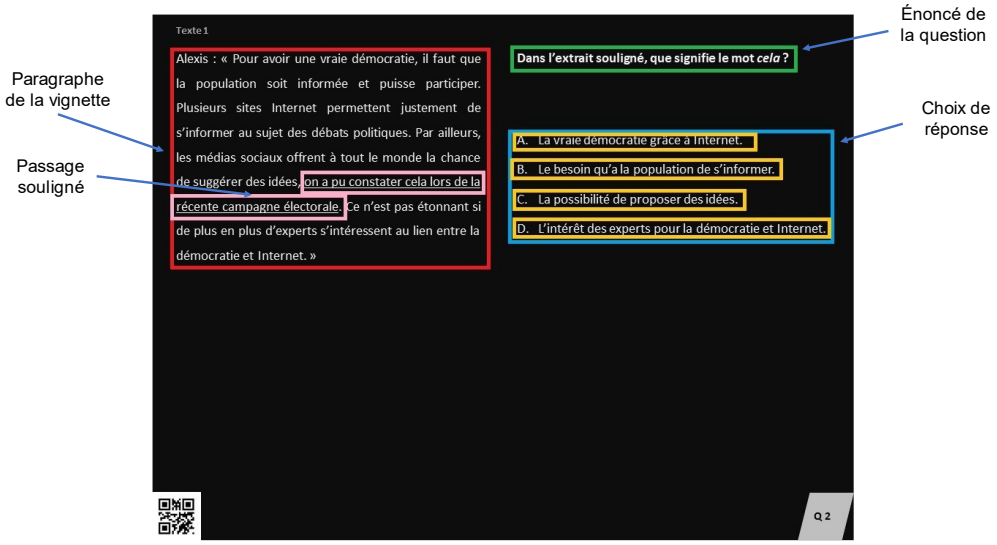


Figure 2. Cartographie des zones d'intérêts dans un stimulus visuel correspondant à l'item d'un test de compréhension de texte. Les rectangles de couleur ne s'affichent pas lors de l'expérimentation et illustrent ici l'emplacement des zones. Les choix de réponse peuvent être analysés individuellement ou comme une seule zone.

Collecte de données

Le participant s'asseyait devant le moniteur affichant les items, à une distance d'environ 60 cm. Après avoir lu et rempli le document d'information et de consentement, le participant enfilait les lunettes d'oculométrie et s'assurait de leur confort. Des lentilles correctrices étaient installées au besoin, puis les lunettes étaient calibrées à partir d'une cible apparaissant à l'écran. Le participant devait alors fixer quatre objets apparaissant dans les coins de l'écran afin de vérifier la calibration. Les consignes données au participant étaient de s'abstenir de lire à voix haute, de répondre en nommant la lettre de leur choix, puis de justifier leur réponse. Aucune instruction spécifique n'était donnée quant aux stratégies de lecture. Les participants experts recevaient en outre la consigne de commenter la qualité de l'item après avoir répondu et justifié leur réponse. La collecte prenait, incluant l'ajustement et la calibration des lunettes, environ une heure par participant.

Pour chaque item du questionnaire, l'enregistrement était structuré par la séquence illustrée à la Figure 3: l'item est affiché, le participant répond, puis justifie sa réponse à voix haute et, le cas échéant, commente l'item. Lorsque le participant termine sa justification, l'expérimentateur annonce la prochaine question, qui apparait alors à l'écran. Les justificatifs des réponses et les commentaires des experts collectés dans le cadre de cette collecte de données n'ont pas été utilisés pour le présent article, et seront analysés ultérieurement.

Chronométrage

Les enregistrements ont été visionnés afin de noter les réponses et temps de réponse, telle qu'illustré à la Figure 3. Le chronométrage a été fait directement à partir de la vidéo en utilisant une précision à la seconde près. Nous avons considéré comme temps de réponse la durée entre l'affichage de l'item à l'écran et le moment où le participant exprime sa réponse. Le chronométrage des enregistrements a été utilisé pour segmenter les vidéos, chaque segment correspondant à un item pour un participant.

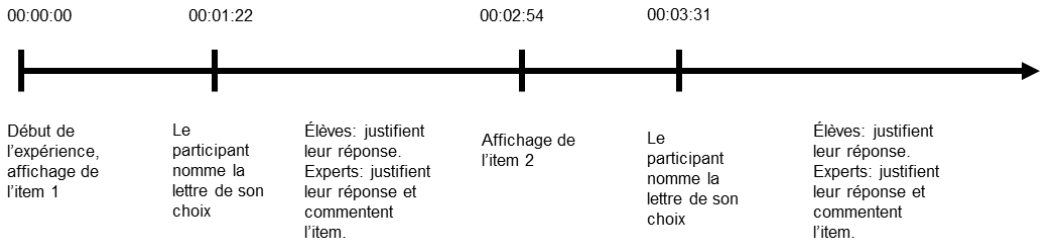


Figure 3. Ligne du temps illustrant le protocole de collecte de données, avec exemple de chronométrage.

Transformation avec MGM

Les segments ont été traités à l'aide de MGM de manière à produire des données relatives au système de coordonnées de l'image de référence. Le code source de MGM a été légèrement modifié de manière à rendre les analyses plus robustes et rapides. Les modifications sont détaillées dans une étude précédente (article 3 de la présente thèse) ayant également testé sur des

donnée simulées plusieurs configurations de MGM, nous avons employé la configuration ayant le mieux performé (AKAZE-ALT 70, décrite plus loin). La transformation avec les techniques de vision informatique a entrainé le rejet de 24 956 observations oculométriques sur un maximum théorique de 310 800 (voir Tableau 1). Ce type de rejet survient lorsque les algorithmes n'arrivent pas à détecter adéquatement le stimulus visuel (la représentation graphique de l'item). Les données oculométriques associées au cadre vidéo qui ne peuvent alors pas être transformées sont rejetées. Une inspection des vidéos de débogage produits par MGM indique que ce type de rejet se produisait le plus souvent lorsque le participant ne regardait pas le moniteur affichant l'item, ou bougeait la tête rapidement, ce qui se produisait fréquemment lors de la phase de justification de la réponse.

Le sommaire des rejets de données est présenté au Tableau 1. Environ 3% des données étaient déjà classifiées comme non valides par l'appareil d'oculométrie mobile, ce qui se produit lorsque la position du regard du participant ne peut être déterminée. Les données rejetées lors de la transformation, ce qui regroupe les rejets survenant lorsque le participant ne regarde pas le stimulus et les rejets dus à une erreur de MGM, représentaient 8% du maximum théorique. En somme, le processus complet allant de la collecte à la production des statistiques a entrainé la perte d'environ 11% des données, une proportion de rejet comparable à celle rapportée par Hareide et Ostnes (2017) pour une expérimentation en oculométrie mobile.

Tableau 1

Sommaire des observations oculométriques rejetées et analysées

	Nombre d'observations	**Proportion**
Maximum théorique d'observations (6126 sec à 50 Hz)	310 800	100%
Identifiées comme non valide par l'appareil	8 884	2,9%
Rejetées lors de la transformation	24 956	8,0%

Observations oculométriques analysées	276 960	89,1%

Note. Les proportions indiquées sont relatives au maximum théorique d'observations oculométriques, calculé en multipliant la durée totale des enregistrements (6216 sec) par la cadence d'échantillonnage (50 Hz). Les rejets lors de la transformation incluent autant les pertes de données survenant lorsque le participant ne regarde pas le stimulus que les rejets en raison d'un échec de MGM à détecter le stimulus pourtant présent; ces deux sources de rejet n'ont pas été séparées pour la présente étude.

Classifications des mouvements oculaires

La classification des données oculométriques transformées par MGM en fixations, saccades, clignements d'yeux et autres événements a été effectuée avec la bibliothèque *saccade* pour le langage R (von der Malsburg, 2015) qui implémente la méthode d'Engbert et Kliegl (2003). Cette méthode utilise la vélocité des mouvements oculaires comme critère de classification, le seuil étant calculé en multipliant l'écart absolu de la médiane par une constante lambda. Nous avons choisi un lambda de 8 en nous appuyant sur les résultats d'essais systématiques pour des données d'oculométrie mobile (Backhaus et al., 2020). Le lissage des coordonnées (*smooth coordinates*) par moyenne mobile était activé, tel que recommandé par la documentation de la bibliothèque *saccades* pour des données provenant d'un appareil ayant une cadence d'enregistrement faible (moins de 200 Hz). Le seuil inférieur de durée des fixations était de 100 ms, en deçà duquel les fixations étaient reclassifiées comme des saccades oculaires – environ 22% des fixations ont ainsi été rejetées. Un seuil de 100 ms est typique pour l'étude oculométrique des processus de lecture (van der Lans et al., 2011; Wass et al., 2012). La procédure a détecté 15615 fixations.

Production des jeux de données

Une fois les fixations classifiées, leurs coordonnées cartésiennes (qui correspondent alors à l'emplacement du point de regard sur dans le plan fixe du stimulus visuel) ont été transformées en zones d'intérêt par croisement avec le tableau contenant la cartographie des zones d'intérêt. Par exemple, si la fixation est située aux coordonnées cartésiennes (100, 150) du stimulus visuel,

et que ce point est situé à l'intérieur de la zone Z, alors nous considérons qu'il s'agit d'une fixation sur la zone Z. Les 2172 fixations dont les coordonnées ne se situaient pas à l'intérieur d'une zone d'intérêt ont été retirées. Le résultat prend la forme d'une liste chronologique des fixations avec leur durée, la zone d'intérêt fixée, le tout par participant et par item. Les données sur les visites de zones d'intérêt ont été produites en effectuant un recodage par plage des données de fixation. La Figure 4 illustre la forme que prenaient les deux tableaux de données produits.

Fixations

#	Participant	Item	Début (ms)	Fin (ms)	Durée (ms)	Zone d'intérêt
1	X101	Q2	0	210	210	Libellé de la question
2	X101	Q2	290	550	260	Passage de texte
3	X101	Q2	660	940	280	Passage de texte
4	X101	Q2	950	1260	310	Passage de texte
5	X101	Q2	1270	1380	110	Choix A

Visites

#	Participant	Item	Début (ms)	Fin (ms)	Durée (ms)	Zone d'intérêt
1	X101	Q2	0	210	210	Libellé de la question
2	X101	Q2	290	1260	970	Passage de texte
3	X101	Q2	1270	1380	110	Choix A

Figure 4. Les données des fixations sont recodées par plage afin de produire les données des visites.

8.2.5. Analyses statistiques

Nous avons comparé les élèves et les experts à partir des données comportementales et oculométriques. Les variables comportementales étaient, pour cette étude, les réponses des participants aux items et les temps de réponses mesurés à partir des enregistrements audiovidéo. Les différences intergroupes dans la proportion d'items réussis et le temps moyen de réponse à l'item ont été vérifiées par des tests de Brunner-Munzel (Brunner et Munzel, 2000; Fagerland et Sandvik, 2009; Karch, 2020). Ce test a été choisi en raison de la petite taille d'échantillon et de la distribution non-normale. La taille d'effet rapportée est la taille d'effet en langage commun

(*TELC*) ou probabilité de supériorité. Une table de conversion des tailles d'effet est disponible en annexe du présent article.

Les variables oculométriques étaient celles contenues dans les deux jeux de données (fixations et visites) décrits précédemment, c'est-à-dire la durée et le nombre de fixations, de même que la durée et le nombre de visites sur les zones d'intérêt. Les zones d'intérêt utilisées étaient celles illustrées à la Figure 2. Toutes les analyses statistiques ont été programmées en langage R (Ihaka et Gentleman, 1996). Nous présentons dans ce qui suit les analyses effectuées afin de vérifier les hypothèses de recherche.

Hypothèse 1 : fluidité en lecture

Pour vérifier si la fluidité en lecture différait entre les deux groupes, nous avons d'abord comparé la durée moyenne des fixations et l'écart-type de la durée de fixation. Les durées de fixation des deux groupes ont été comparées par un test de Brunner-Munzel, choisi en raison de la distribution non-normale et asymétrique des données. Nous avons utilisé un test de Levene pour comparer l'écart-type de la durée de fixation centrée par soustraction de la moyenne du participant. Le centrage des données visait à minimiser les biais découlant de la variabilité intra-individuelle (Bolger et Laurenceau, 2013), qui peut être élevée pour la durée des fixations (Rayner, 1998).

Pour améliorer la robustesse des analyses face à un nombre inégal d'observations par participant et à une variabilité intra-individuelle élevée, nous avons également calculé une moyenne et un écart-type par groupe en agglomérant d'abord les données par participant. Cette procédure a été répétée par procédure de *bootstrap* (2000 itérations, avec remise). Les résultats des itérations *bootstrap* ont été agrégés de la manière suivante : les moyennes et écarts-types des durées de fixation sont d'abord calculées par participant, puis combinées par groupe; les valeurs

rapportées sont les médianes des résultats obtenus pour l'ensemble des itérations, les bornes des intervalles de confiance à 95% sont les percentiles 2,5 et 97,5. Nous avons estimé la taille d'effet conformément à la méthode de Cohen (1988), qui divise la différence de moyenne par l'écart-type combiné des deux groupes; un *d* de Cohen a ainsi été calculé pour chaque itération bootstrap, nous rapportons la médiane et l'intervalle de confiance à 95%.

Hypothèse 2 : stratégies de compréhension de texte

Pour tester l'hypothèse selon laquelle les élèves et experts déploient des stratégies différentes de compréhension de texte, nous avons comparé l'attention visuelle que les groupes accordaient aux différentes zones d'intérêt. L'intérêt pour les zones d'intérêt a été opérationnalisé par deux variables : la proportion des visites accordées à chaque zone, et le temps total de visite passé sur chaque zone (*dwell time*), également transformé en proportions. Les proportions ont été calculées par participant, item et zone d'intérêt, de manière à minimiser les biais liés au nombre inégal d'observations. Nous avons utilisé le test de Scheirer–Ray–Hare (Scheirer et al., 1976) afin de vérifier qu'au moins une des zones d'intérêt différait au sein du groupe (intragroupe), et entre les groupes (intergroupe); ce test constitue une alternative multifactorielle et non-paramétrique à l'ANOVA. L'implémentation du test était celle de la bibliothèque *rcompanion* de R (Mangiafico, 2020). Afin de voir quelles zones d'intérêt étaient différentes, nous avons procédé à des comparaisons multiples par tests de Brunner-Munzel bilatéraux, choisis en raison de la distribution non-normale et asymétrique des données. Nous avons appliqué la méthode des comparaisons multiples avec le meilleur (*multiple comparisons with best)* proposée par Hsu (1996) et contrasté chaque zone avec le passage de texte, qui était la zone ayant reçu le plus de visites et de temps de visite. Pour les comparaisons intergroupes, nous avons contrasté chaque zone d'intérêt avec son équivalent dans l'autre groupe. Afin de minimiser

le taux de fausse découverte, nous avons appliqué la correction de Benjamini-Hochberg à l'ensemble des comparaisons (Benjamini et Hochberg, 1995).

8.3. Résultats

8.3.1. Résultats comportementaux

L'analyse des données comportementales a porté sur les réponses et temps de réponse des 10 élèves et 8 experts ayant produit en tout 112 segments d'enregistrement (durée moyenne : 55,5 sec; écart-type : 33,5 sec). Un segment débute au moment où l'item est présenté à l'écran et se termine lorsque le participant énonce une réponse. Le Tableau 2 fait le sommaire des temps de réponse et de la proportion de bonnes réponses pour chaque item. La proportion d'items réussis par les élèves (M = 0,65, ET = 0,29) et par les experts (M = 0,70, ET = 0,34), était statistiquement équivalente, p = 0,5851; la taille d'effet en langage commun (*TELC*) associée à ce résultat était de 0,59 [0,25, 0,92], cette statistique correspond à la probabilité de supériorité d'un élève pris au hasard. Le temps moyen de réponse à l'item des élèves (M = 56,1 sec, ET = 22,9 sec) et des experts (M = 54,5 sec, ET = 21,4 sec) était statistiquement équivalent, p = 0,8495, *TELC* = 0,47 [0,12, 0,82]. Les différences entre les deux groupes ne sont donc pas visibles à partir des données comportementales; ces résultats sont détaillés par item au Tableau 3.

Tableau 2

Descriptif des items avec statistiques comportementales (n = 18)

Question	Temps de réponse (sec)		Prop. de bonnes réponses	Présence d'un passage souligné
	Moyenne	Écart-type		
Q1	73,9	43,6	0,14	Non
Q2	38,7	15,4	0,93	Oui
Q4	40,9	17,7	0,64	Oui
Q8	98,9	40,6	0,79	Non
Q9	47,1	24,6	0,64	Non
Q11	46,6	31,1	0,50	Oui
Q12	48,1	19,0	0,71	Oui

Q13	49,7	19,0	1	Oui

Tableau 3

Données comportementales – comparaison élève-expert (n = 18)

Mesure	Élèves		Experts		Test de Brunner-Munzel	
	M	*ET*	*M*	*ET*	Statistiques	Taille d'effet [IC à 95%]
Temps de réponse à l'item (sec)	54,48	21,41	56,06	22,88	*bm* = -0,1940 *dl* = 11,628 *p* = 0,8495	*TELC* = 0,47 [0,12, 0,82]
Proportion d'items réussis	0,65	0,29	0,70	0,34	*bm* = 0,5599 *dl* = 12,994 *p* = 0,5851	*TELC* = 0,59 [0,25, 0,92]

Note. Pour le test de Brunner-Munzel, nous rapportons la statistique produite par le test (*bm*), le nombre de degrés de liberté (*dl*) et la valeur p. La taille d'effet rapportée (*TELC*) est une probabilité de supériorité, ou taille d'effet en langage commun, avec son intervalle de confiance à 95%; la valeur de référence pour ce calcul était le groupe d'élèves.

8.3.2. Résultats oculométriques – durée des fixations

Les résultats oculométriques portent sur les enregistrements de 9 élèves et 6 experts ayant produit respectivement 8639 et 4800 fixations. La Figure 5 montre que la distribution des durées de fixation décrivait une courbe asymétrique avec une borne inférieure à 100 ms, qui était le seuil minimal choisi pour la durée des fixations. L'asymétrie de la distribution, mesurée selon la méthode 1 de Joanes et Gill (1998), était de 2,67 pour le groupe d'élèves, et 3,09 pour le groupe d'experts.

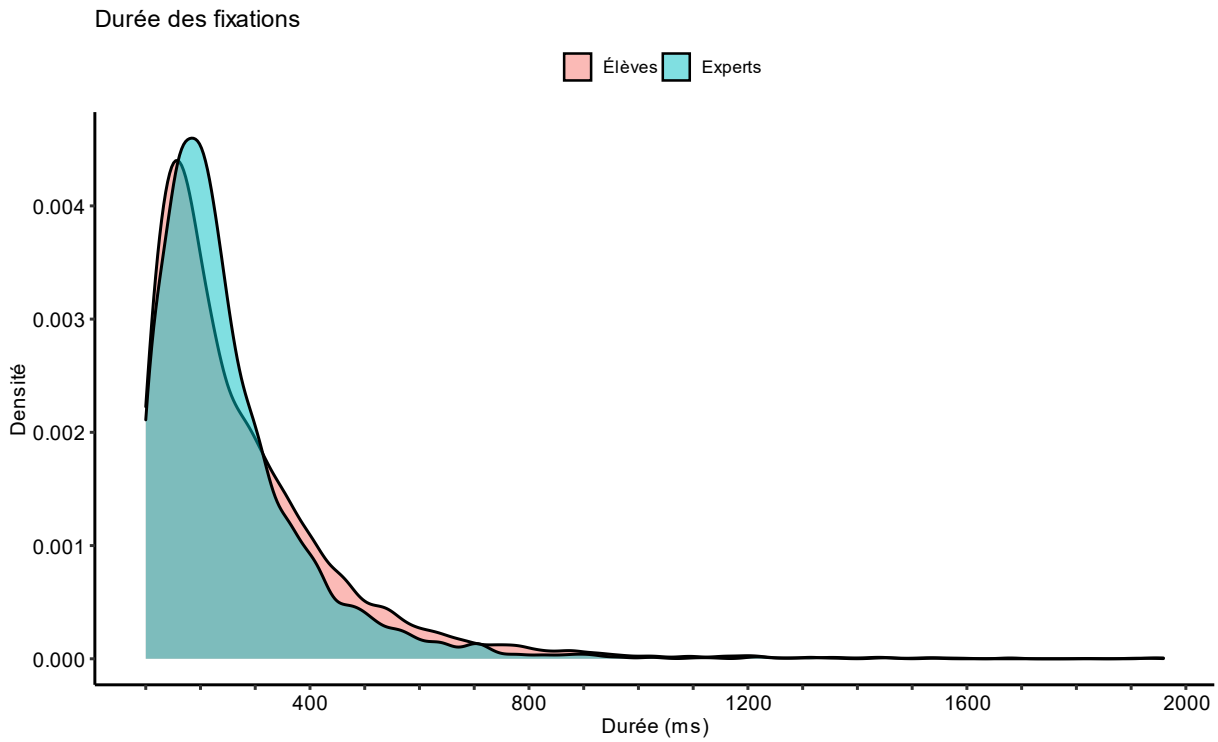


Figure 5. Distribution de la durée de fixation pour 9 élèves (8639 fixations) et 6 experts (4800 fixations), en millisecondes.

Pour l'ensemble des fixations sur des zones d'intérêt, les fixations des élèves (M = 276 ms) étaient statistiquement plus longues que celles des experts (M = 254 ms), bm = 3,6173, p = 0,0003. La taille d'effet de langage commun pour cette analyse ($TELC$ = 0,52 [0,51, 0,53]) indique qu'une fixation d'élève prise au hasard avait environ 52% de chances d'être plus longue qu'une fixation d'expert. L'écart-type de la durée de fixation centrée par participants était aussi plus élevé pour les élèves (ET = 170 ms) que pour les experts (ET = 145 ms) selon un test de Levene, $F(1)$ = 59,982, p < 0,0001.

Les statistiques descriptives agglomérées par participant puis par groupe indiquaient une durée de fixation en apparence plus longue et variable chez les élèves (M = 281 ms, ET = 176 ms) que chez les experts (M = 242 ms, ET = 144 ms). Pour vérifier cette différence, nous avons recalculé ces statistiques 2000 fois par rééchantillonnage de type bootstrap (2000 répétitions, avec remise). Tel qu'illustré à la Figure 6, les résultats de la procédure de *bootstrap* confirment

que les élèves (*M* = 278 [267, 288] ms, *ET* = 176 [169, 184] ms), et ont généralement fait des fixations de durée plus longue et variable que les experts (*M* = 245 [238, 252] ms, *ET* = 144 [136, 152] ms). La taille d'effet générale calculée depuis les résultats de la procédure de *bootstrap* était de *d* = 0,21 [0,13, 0,28], ce qui correspond à un effet de petite ampleur selon les barèmes proposés par Cohen (1988).

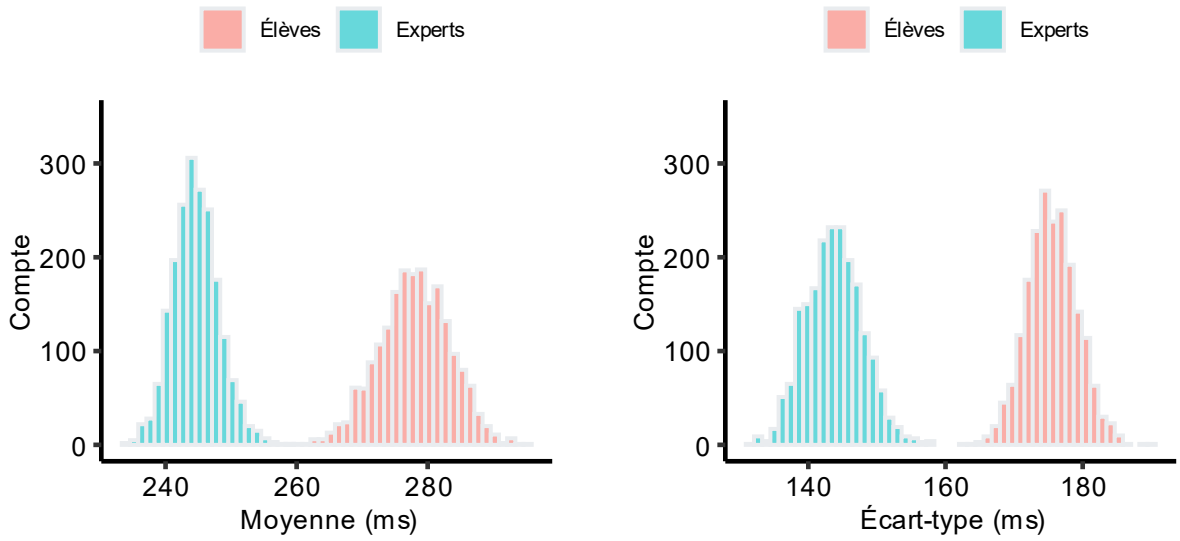


Figure 6. Résultats de l'estimation de la moyenne et de l'écart-type de la durée des fixations, pour 2000 itérations *bootstrap*.

En somme, les statistiques calculées directement sur l'ensemble des données, on observait une différence intergroupe de 24 ms pour la moyenne et de 26 ms pour l'écart-type. Les estimations par technique de *bootstrap* indiquaient une différence de 33 ms (*IC* = [21, 46]) pour la moyenne, et 32 ms (*IC* = [21, 42]) pour l'écart-type. Conformément avec l'hypothèse 1, ces résultats montrent que les élèves avaient des fixations en moyenne plus longues et d'une durée plus variable que les experts.

8.3.3. Résultats oculométriques – allocation des visites

Les résultats sur les visites de zone d'intérêt portaient sur 3196 visites (2018 pour les 9 élèves, 1178 pour les 6 experts). Nous avons produit des diagrammes à barres superposées,

montrés à la Figure 7, afin de visualiser l'allocation des visites et du temps de visite sur les zones d'intérêt par les élèves et les experts.

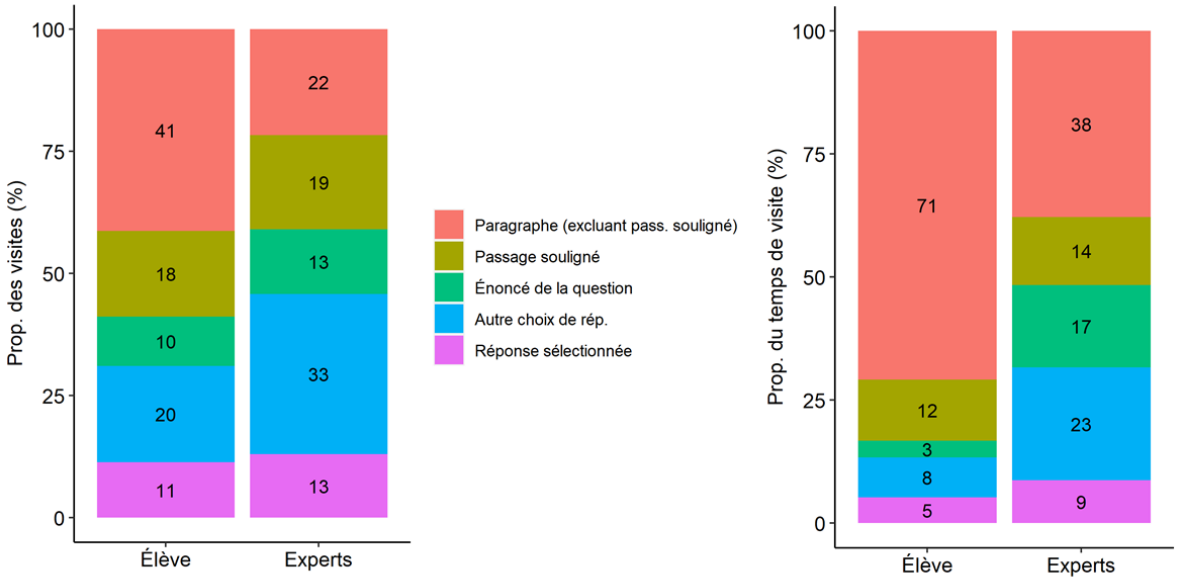


Figure 7. Allocation des visites (à gauche) et du temps de visite (à droite) sur les zones d'intérêt, par groupe. Les chiffres indiquent la proportion moyenne par zone. Par exemple, les élèves ont consacré 3% de leur temps de visite sur l'énoncé de la question, contre 17% pour les experts.

Nous avons appliqué un test de Scheirer-Ray-Hare afin de vérifier si la proportion de visites variait en fonction de la zone d'intérêt et du groupe (élève ou expert). Le test a montré un effet principal statistiquement significatif de la zone d'intérêt, $H(4) = 170{,}062$, $p < 0{,}0001$. L'effet principal du groupe n'était pas statistiquement significatif, $H(1) = 0{,}025$, $p = 0{,}8744$. L'interaction entre la zone d'intérêt et le groupe était statistiquement significative, $H(4) = 47{,}991$, $p < 0{,}0001$, indiquant que l'influence de la zone d'intérêt sur l'allocation des visites était différente selon le groupe.

Le Tableau 4 présente le résultat des comparaisons multiples par tests de Brunner-Munzel sur les proportions de visites effectuées, et permet de vérifier la significativité statistique des différences observées à la Figure 7. Les comparaisons intragroupes montrent que les élèves ont alloué, proportionnellement, davantage de visites au paragraphe (en moyenne 41% des visites)

qu'à chacune des autres zones. Chez les experts, le paragraphe (22%) était en seconde position dans la proportion des visites, derrière les choix de réponse non-sélectionnés (33%), $p < 0{,}0001$, *TELC* = 0,73. Les comparaisons intergroupes montrent que les élèves ont alloué plus de visites au paragraphe de texte comparativement aux experts, $p < 0{,}0001$, *TELC* = 0,06. Les élèves ont consacré une plus petite proportion de leurs visites aux choix de réponses autres que celui sélectionné (20%) comparativement aux experts (33%), $p < 0{,}0001$, *TELC* = 0,73. La proportion des visites sur la portion soulignée du paragraphe, sur l'énoncé et sur le choix de réponse sélectionné étaient statistiquement équivalentes entre les groupes. Les comparaisons statistiquement significatives à un seuil $p < 0{,}05$ le sont demeurées après application de la correction de Benjamini-Hochberg.

Tableau 4

Comparaisons multiples – proportion des visites sur les zones d'intérêt

Type de comparaison	Zone	Tests de Brunner-Munzel			
		bm [a]	***p***	***p aj.*** [b]	***TELC* [*IC* à 95%]** [c]
Intragroupe élève	PARAG	--	--	--	--
	ÉNONCÉ	-32,6878	< 0,0001	< 0,0001	0,03 [0,00, 0,06]
	SOULIGNÉ	-20,1999	< 0,0001	< 0,0001	0,05 [0,01, 0,10]
	CH_AUTRE	-13,0013	< 0,0001	< 0,0001	0,10 [0,04, 0,16]
	CH_SÉLEC	-28,7321	< 0,0001	< 0,0001	0,03 [0,00, 0,06]
Intragroupe expert	PARAG	--	--	--	--
	ÉNONCÉ	-4,5270	< 0,0001	< 0,0001	0,25 [0,14, 0,36]
	SOULIGNÉ	-0,5240	0,6022	0,7117	0,46 [0,31, 0,61]
	CH_AUTRE	4,1062	0,0001	0,0001	0,73 [0,62, 0,84]
	CH_SÉLEC	-4,4762	< 0,0001	< 0,0001	0,25 [0,14, 0,36]
Intergroupe	PARAG	-18,5203	< 0,0001	< 0,0001	0,06 [0,01, 0,11]
	SOULIGNÉ	0,2914	0,7719	0,7719	0,52 [0,38, 0,67]
	ENONCÉ	1,4466	0,1516	0,1970	0,58 [0,47, 0,70]

	CH_AUTRE	4,7759	< 0,0001	< 0,0001	0,73 [0,64, 0,83]
	CH_SELEC	0,4438	0,6586	0,7135	0,53 [0,39, 0,67]

Note. Comparaisons multiples portant sur 3196 visites. Pour les comparaisons intragroupes, la valeur de référence est la proportion de visites sur le paragraphe. Pour les comparaisons intergroupes, nous avons utilisé les valeurs des élèves comme niveau de référence. [a] Statistique résultant du test de Brunner-Munzel bilatéral. [b] valeur *p* après application de la correction de Benjamini-Hochberg à l'ensemble des comparaisons. [c] Taille d'effet en langage commun indiquant la probabilité de supériorité du groupe de référence; intervalle de confiance à 95%.

La méthodologie pour comparer l'allocation du temps de visite sur les zones d'intérêt était la même que pour le nombre de visites. Un test de Scheirer-Ray-Hare indique un effet principal significatif de la zone d'intérêt, $H(4) = 209,3$, $p < 0,0001$, et du groupe, $H(1) = 16,909$, $p < 0,0001$. L'interaction entre la zone d'intérêt et le statut du participant était significative $H(4) = 52,781$, $p < 0,0001$.

Le Tableau 5 résume le résultat des comparaisons multiples par test de Brunner-Munzel pour la proportion du temps de visite accordée aux différentes zones d'intérêt. Pour les deux groupes, le paragraphe a été plus longuement visité que chacune des autres zones. Les tailles d'effet en langage commun, présentées au Tableau 5, indiquent que le contraste entre le paragraphe et les autres zones était plus marqué chez les élèves que chez les experts. Les comparaisons intergroupes montrent que les élèves ont accordé une plus grande proportion de leur temps de visite au paragraphe (71%), comparativement aux experts (38%), $p < 0,0001$, $TELC = 0,07$. Les élèves ont alloué proportionnellement moins de temps à l'énoncé de la question (3%) comparativement aux experts (17%), $p < 0,0001$, $TELC = 0,88$. La proportion du temps de visite sur les choix de réponse non-sélectionnés était plus faible chez les élèves (8%) que chez les experts (23%), $p < 0,0001$, $TELC = 0,85$. De même, les élèves ont consacré une plus faible proportion de leur temps de visite au choix de réponse sélectionné (5%) comparativement aux experts (9%), $p = 0,0236$, $TELC = 0,65$. Les groupes étaient équivalents concernant la

proportion du temps alloué au passage souligné. Les comparaisons statistiquement significatives à un seul $p < 0,05$ le sont demeurées après application de la correction de Benjamini-Hochberg.

Tableau 5

Comparaisons multiples – proportion du temps de visite sur les zones d'intérêt

Type de comparaison	Zone	Tests de Brunner-Munzel			
		bm	***p***	***p aj.***	***TELC* [*IC* à 95%]**
Intragroupe élève	PARAG	--	--	--	--
	ÉNONCÉ	-63,4069	< 0,0001	< 0,0001	0,01 [0,00, 0,02]
	SOULIGNÉ	-32,5424	< 0,0001	< 0,0001	0,02 [0,00, 0,05]
	CH_AUTRE	-38,9615	< 0,0001	< 0,0001	0,02 [0,00, 0,04]
	CH_SÉLEC	-49,2079	< 0,0001	< 0,0001	0,01 [0,00, 0,03]
Intragroupe expert	PARAG	--	--	--	--
	ÉNONCÉ	-4,8613	< 0,0001	< 0,0001	0,23 [0,12, 0,34]
	SOULIGNÉ	-5,6081	< 0,0001	< 0,0001	0,19 [0,08, 0,30]
	CH_AUTRE	-3,0622	0,0036	0,0043	0,30 [0,17, 0,43]
	CH_SÉLEC	-8,1787	< 0,0001	< 0,0001	0,14 [0,05, 0,23]
Intergroupe	PARAG	-13,0819	<0,0001	< 0,0001	0,11 [0,05, 0,17]
	SOULIGNÉ	0,4851	0,6302	0,6302	0,54 [0,38, 0,69]
	ENONCÉ	11,7487	< 0,0001	< 0,0001	0,88 [0,82, 0,95]
	CH_AUTRE	8,9815	< 0,0001	< 0,0001	0,86 [0,78, 0,94]
	CH_SELEC	2,3453	0,0218	0,0236	0,65 [0,52, 0,78]

Note. Comparaisons multiples portant sur 3196 visites. Pour les comparaisons intragroupes, la valeur de référence est la proportion de visites sur le paragraphe. Pour les comparaisons intergroupes, nous avons utilisé les valeurs des élèves comme niveau de référence. [a] Statistique résultant du test de Brunner-Munzel bilatéral. [b] valeur *p* après application de la correction de Benjamini-Hochberg à l'ensemble des comparaisons. [c] Taille d'effet en langage commun indiquant la probabilité de supériorité du groupe de référence; intervalle de confiance à 95%.

En somme, les analyses comparant l'allocation de l'attention visuelle (visites et temps de visite) révèlent plusieurs différences entre les élèves et experts, ce qui appuie l'hypothèse 2. La plupart des différences significatives à un seuil de $p < 0,05$ avaient une taille d'effet qui, lorsque traduite en équivalent en *d* de Cohen, dépassait le seuil de 0,8 considéré comme un effet de

grande ampleur (Cohen, 1988). Les élèves ont consacré une plus grande part de leur attention visuelle au paragraphe, tandis que les experts ont davantage réparti leur attention sur l'ensemble des zones. Les experts ont ainsi consacré davantage de visites et de temps de visite aux choix de réponse autres que celui sélectionné, et ont alloué une plus grande part de leur temps de visite à l'énoncé de la question.

8.4. Discussion

La présente étude s'intéressait à la faisabilité d'une étude de la compréhension de texte par l'oculométrie mobile en intégrant *MGM* pour Python à la chaine de traitement des données. Bien qu'aucune différence n'ait été mise en évidence dans les données comportementales, notre expérimentation pilote a permis d'observer des différences marquées entre les patrons de mouvements oculaires d'élèves et d'experts lors de tâches de compréhension de texte.

8.4.1. Interprétation des résultats oculométriques

La durée des fixations était légèrement plus longue chez les élèves. Ce résultat tend à indiquer que les élèves avaient une charge cognitive plus élevée se traduisant par une fluidité moindre, comme le prédisait l'hypothèse 1. Notre comparaison intergroupe de la durée des fixations par *boostrap* avait une taille d'effet associée de $d = 0{,}21$. Par comparaison, la métaétude de Gegenfurtner et al. (2011) a rapporté une taille d'effet ajustée moyenne de $d = 0{,}18$ pour 44 études comparant la durée de fixation de novices et d'experts lorsque le stimulus visuel contenait à la fois des régions pertinentes et redondantes[3]; cette mise en perspective devrait cependant être considérée avec prudence, compte tenu de la diversité des approches utilisées pour définir les niveaux d'expertise. La différence d'âge entre les étudiants et les experts pourrait également

[3] Converti depuis la valeur fournie ($r_c = -0{,}09$), laquelle utilisait les experts comme groupe de référence alors que nous avons utilisé les élèves, d'où l'inversion du signe.

expliquer la différence observée ; toutefois, cette explication est mise en doute par des travaux montrant que la durée de fixation a une distribution équivalente chez les jeunes adultes et les adultes plus âgés (Rayner, 1998).

La durée des fixations était également plus variable chez les élèves, indiqué par un écart-type intra-individuel plus élevé. Ce résultat va dans le sens de ce qui a été observé par Smith et al. (2018), c'est-à-dire qu'une habileté plus faible en lecture entraine une plus grande variabilité dans la durée des fixations. Bien que cette piste demanderait davantage d'analyses, la plus grande variabilité de la durée de fixations chez les élèves, de même que la distribution des données par groupe (voir Figure 5) suggère que les élèves avaient une plus grande proportion de fixations longues, ce qui pourrait indiquer que davantage de mots ou de passages du texte étaient décodés par un processus plus laborieux.

Dans la seconde partie des analyses, nous avons comparé l'allocation des visites et du temps de visite, utilisés comme mesures de l'attention visuelle consacrée aux différentes zones de l'item. Les résultats montrent qu'il existe des différences entre les groupes, confirmant l'hypothèse 2. Les experts ont davantage réparti leurs visites et leur temps de visite sur les différentes zones d'intérêt, alors que les élèves se sont davantage concentrés sur le paragraphe de texte. De plus, les experts ont été plus attentifs à l'énoncé de la question, et ont davantage considéré les choix de réponse autres que celui sur lequel s'est arrêté leur choix. Les experts ont également été plus attentifs au passage souligné du paragraphe de la vignette lorsqu'on prend comme référence la proportion de visites ou de temps de visite allouée au reste du paragraphe. Ces résultats suggèrent une allocation plus efficace des visites et du temps de visite par les experts, ce qui était également une conclusion de la métaétude de Gegenfurtner et al. (2011),

mais demeure pour notre étude un résultat paradoxal puisque cette allocation ne semble pas avoir permis performance supérieure selon les données béhaviorales.

8.4.2. Implications pour la recherche

Notre étude démontre la viabilité d'une méthodologie innovante et qui pave la voie pour de futurs travaux portant sur les processus de lecture. Les avantages de l'oculométrie mobile pourraient ainsi s'étendre à l'étude de la lecture dans un contexte écologique, et auprès de populations qu'il serait difficile d'étudier par l'oculométrie stationnaire. De manière secondaire, nous avons fait la démonstration de méthodes statistiques pour l'analyse robuste de données oculométriques, lesquelles posent un défi en raison de leur distribution asymétrique et du nombre inégal d'observations.

Dans le domaine de la compréhension de texte, notre étude contribue à la littérature assez rare sur l'allocation de l'attention visuelle lors de tâches de lecture. Le présent article montre la plausibilité d'un programme de recherche plus complet dont les résultats pourraient se transférer vers le domaine de l'enseignement, par exemple en étudiant le patron d'allocation de l'attention visuelle chez les élèves ayant des difficultés spécifiques en lecture.

8.4.3. Limites et recommandations

Bien que l'analyse des données oculométriques ait permis de différencier les deux groupes en fonction de leur expertise, les données comportementales n'ont montré aucune différence significative entre les groupes; la proportion d'items réussis, ainsi que le temps de réponse moyen, étaient équivalents entre les groupes. Ainsi, les stratégies plus matures déployées par les experts n'ont pas conduit à de meilleures réponses aux items. Ce résultat peut indiquer des problèmes avec les items, qui étaient encore en développement au moment de l'étude et qui ont été conçus pour dépister les difficultés des étudiants qui commencent une carrière

postsecondaire, et non pour évaluer le niveau de compétence d'experts. Un autre facteur pouvant expliquer ce résultat est la nature différente de la tâche entre élèves et experts; les experts devaient en plus formuler des recommandations pour l'amélioration des items, ce qui peut avoir altéré leur performance. Le fait que les élèves et experts n'avaient pas exactement la même tâche constitue une limitation potentielle de notre étude pilote. Des études rapportées dans la synthèse faite par Rayner (1998) indiquent que les patrons de mouvements oculaires diffèrent lorsqu'on confie une tâche au participant (par exemple, jouer à un jeu vidéo versus regarder un enregistrement du même jeu). Wang et al. (2014) rapportent également une durée de fixation plus élevée lorsque la tâche de lecture est plus complexe. Dans notre cas, bien que les élèves n'étaient pas passifs, la tâche confiée aux experts (répondre à l'item et le commenter) était cognitivement plus complexe. Nous nous questionnons donc quant à la part de la différence intergroupe qui pourrait alternativement s'expliquer par la complexité accrue de la tâche des experts. Des travaux futurs pourraient uniformiser la tâche des deux groupes en leur demandant uniquement de répondre aux items sans les commenter, ou à l'inverse en demandant autant aux élèves qu'aux experts de commenter les items. La notion d'expertise a elle-même de multiples définitions pouvant influencer le contraste novice-expert, des futurs travaux devraient donc opérationnaliser plus clairement ce qui est entendu par « expert » et « novice » afin de sélectionner une tâche permettant de contraster les groupes à partir de la réponse à l'item.

Une autre limite de l'étude pilote est la petite taille de l'échantillon. Néanmoins, nous avons pu observer des différences assez importantes entre les groupes, malgré l'utilisation de tests statistiques non paramétriques, qui sont plus susceptibles de ne pas rejeter l'hypothèse nulle. Nous avons également utilisé des méthodes robustes pour confirmer les effets observés, comme le *bootstrap* et le test de Brunner-Munzel avec intervalles de confiance.

Enfin, bien que la chaîne de traitement proposée dans cet article étende les capacités de l'oculométrie mobile à la lecture, ce type d'appareil a généralement une précision et un taux d'enregistrement plus faibles que les appareils fixes. La possibilité que les lunettes se déplacent sur le visage (*splippage*) pendant l'enregistrement peut également affecter la précision des enregistrements. La méthodologie proposée n'élimine pas ces limites, et nous ne la recommandons pas pour l'étude d'aspects de la lecture nécessitant une résolution temporelle ou spatiale qui dépasse les capacités de l'oculomètre mobile. Ainsi, nos résultats portant sur des zones d'intérêt de petite taille (par ex. : les choix de réponses) sont à prendre avec précaution. Nos résultats ne permettent pas d'identifier précisément les formes de complexité linguistique qui affectent la fluidité au niveau du mot; un dispositif avec une résolution spatiale plus fine serait nécessaire pour le faire. Une étude future pourrait quantifier plus précisément la fiabilité de la méthode proposée lorsqu'elle est utilisée avec différents dispositifs mobiles, tout en offrant des conseils méthodologiques pour atténuer la production de valeurs aberrantes et les risques de décalibration pendant l'enregistrement.

8.4.4. Conclusion

Notre expérience pilote démontre qu'une chaîne de traitement des données incorporant MGM facilite l'étude de la compréhension de textes dans une approche d'oculométrie mobile. Le dispositif expérimental et la chaîne de traitement des données décrits dans cet article ont produit des résultats similaires à ceux de travaux influents en oculométrie de la lecture (Ashby et al., 2005; Rayner, 1998; Rayner et Well, 1996), et rejoignent des études comparant des groupes ayant différents niveaux d'expertise (Gegenfurtner et al., 2011). Compte tenu de ses limites, la méthodologie proposée ne remplace pas directement l'oculométrie stationnaire, mais pourrait être

une alternative intéressante pour les études portant sur la lecture et qui priorisent la mobilité et le confort des participants plutôt qu'une résolution spatio-temporelle fine.

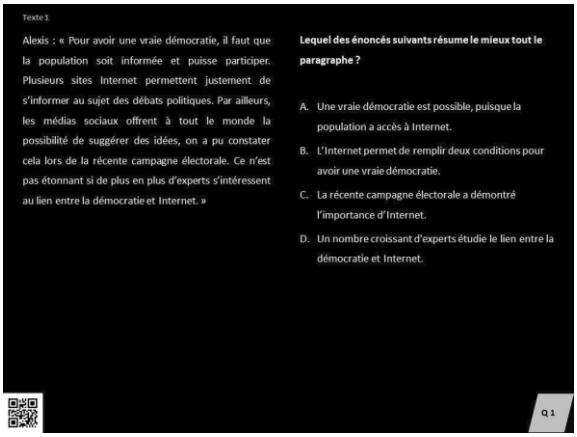
8.5. Annexe du quatrième article

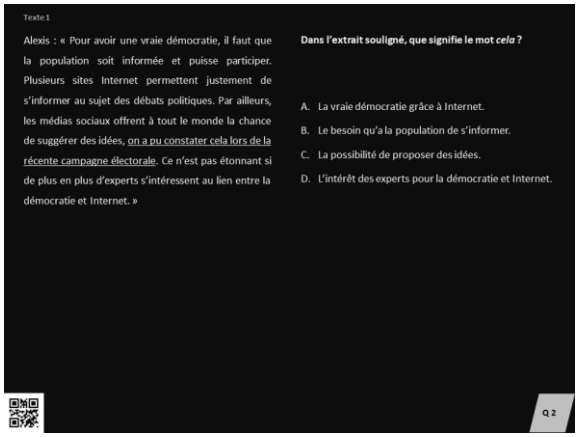
Table de conversion des tailles d'effet

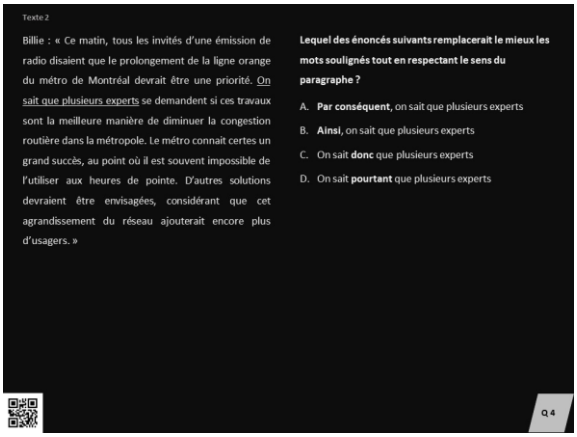
d	*TELC*	*r de Pearson*
-3	0,02	-0,83
-2,75	0,03	-0,81
-2,5	0,04	-0,78
-2,25	0,06	-0,75
-2	0,08	-0,71
-1,75	0,11	-0,66
-1,5	0,14	-0,60
-1,25	0,19	-0,53
-1	0,24	-0,45
-0,75	0,30	-0,35
-0,5	0,36	-0,24
-0,25	0,43	-0,12
0	0,50	0,00
0,25	0,57	0,12
0,5	0,64	0,24
0,75	0,70	0,35
1	0,76	0,45
1,25	0,81	0,53
1,5	0,86	0,60
1,75	0,89	0,66
2	0,92	0,71
2,25	0,94	0,75
2,5	0,96	0,78
2,75	0,97	0,81
3	0,98	0,83

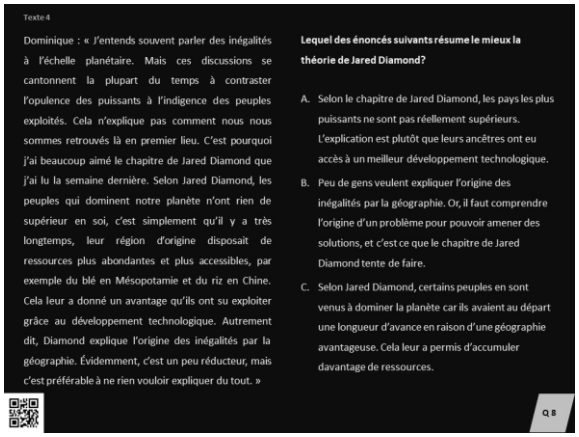
Note. TELC et *r* calculées avec la bibliothèque *effectsize* pour R.

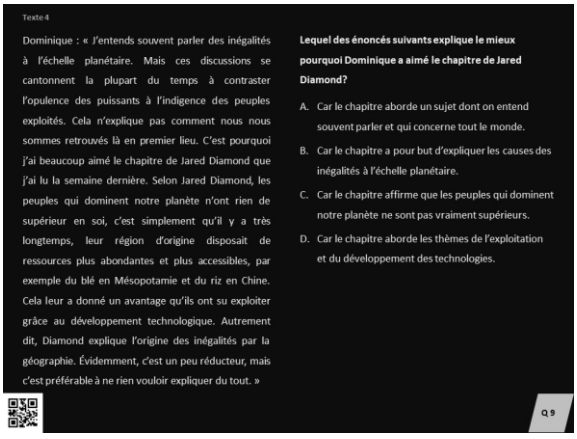
Items utilisés

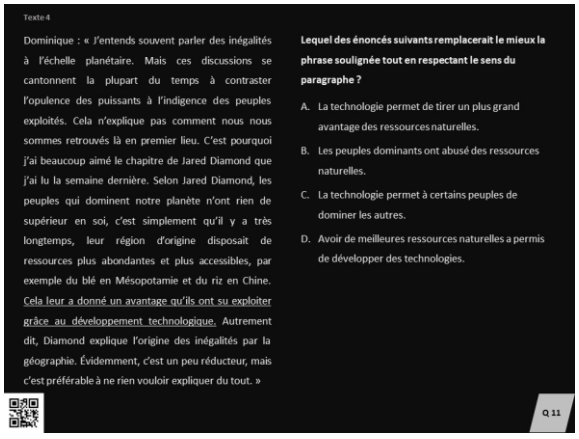


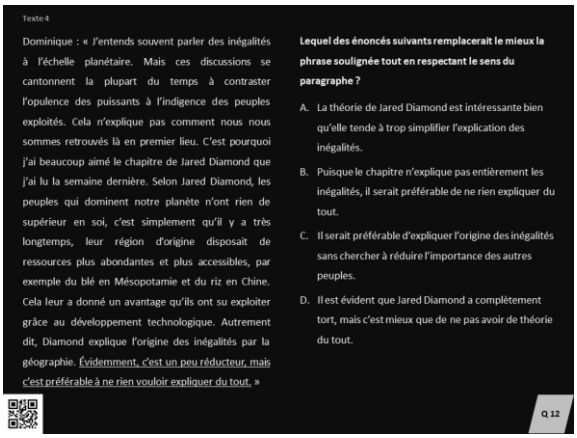


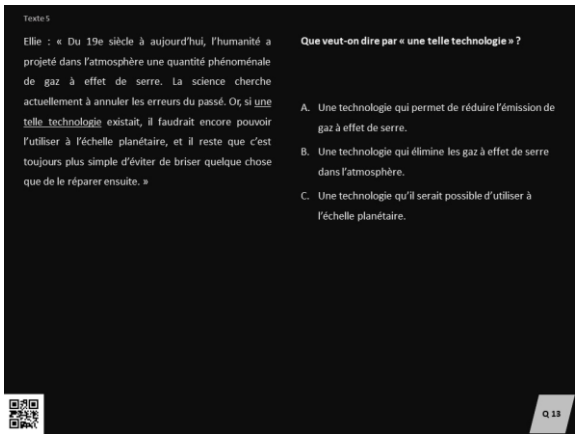









Texte 4
Dominique : « J'entends souvent parler des inégalités à l'échelle planétaire. Mais ces discussions se cantonnent la plupart du temps à contraster l'opulence des puissants à l'indigence des peuples exploités. Cela n'explique pas comment nous nous sommes retrouvés là en premier lieu. C'est pourquoi j'ai beaucoup aimé le chapitre de Jared Diamond que j'ai lu la semaine dernière. Selon Jared Diamond, les peuples qui dominent notre planète n'ont rien de supérieur en soi, c'est simplement qu'il y a très longtemps, leur région d'origine disposait de ressources plus abondantes et plus accessibles, par exemple du blé en Mésopotamie et du riz en Chine. Cela leur a donné un avantage qu'ils ont su exploiter grâce au développement technologique. Autrement dit, Diamond explique l'origine des inégalités par la géographie. Évidemment, c'est un peu réducteur, mais c'est préférable à ne rien vouloir expliquer du tout. »
Lequel des énoncés suivants remplacerait le mieux la phrase soulignée tout en respectant le sens du paragraphe ?
A. La théorie de Jared Diamond est intéressante bien qu'elle tende à trop simplifier l'explication des inégalités.
B. Puisque le chapitre n'explique pas entièrement les inégalités, il serait préférable de ne rien expliquer du tout.
C. Il serait préférable d'expliquer l'origine des inégalités sans chercher à réduire l'importance des autres peuples.
D. Il est évident que Jared Diamond a complètement tort, mais c'est mieux que de ne pas avoir de théorie du tout.
Q 12


Texte 5
Ellie : « Du 19e siècle à aujourd'hui, l'humanité a projeté dans l'atmosphère une quantité phénoménale de gaz à effet de serre. La science cherche actuellement à annuler les erreurs du passé. Or, si une telle technologie existait, il faudrait encore pouvoir l'utiliser à l'échelle planétaire, et il reste que c'est toujours plus simple d'éviter de briser quelque chose que de le réparer ensuite. »
Que veut-on dire par « une telle technologie » ?
A. Une technologie qui permet de réduire l'émission de gaz à effet de serre.
B. Une technologie qui élimine les gaz à effet de serre dans l'atmosphère.
C. Une technologie qu'il serait possible d'utiliser à l'échelle planétaire.
Q 13

8.6. Bibliographie

Ashby, J., Rayner, K. et Clifton, C. (2005). Eye movements of highly skilled and average readers: differential effects of frequency and predictability. *The Quarterly Journal of Experimental Psychology. A, Human Experimental Psychology*, *58*(6), 1065-1086. https://doi.org/10/fjj8nh

Bax, S. (2013). The cognitive processing of candidates during reading tests: Evidence from eye-tracking. *Language Testing*, *30*(4), 441-465. https://doi.org/10.1177/0265532212473244

Bax, S. et Chan, S. (2019). Using eye-tracking research to inform language test validity and design. *System*, *83*. https://doi.org/10.1016/j.system.2019.01.007

Benjamini, Y. et Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289-300. https://doi.org/10/gfpkdx

Bolger, N. et Laurenceau, J.-P. (2013). *Intensive longitudinal methods: An introduction to diary and experience sampling research*. Guilford Press.

Brunner, E. et Munzel, U. (2000). The nonparametric Behrens-Fisher problem: asymptotic theory and a small-sample approximation. *Biometrical Journal*, *42*(1), 17-25. https://doi.org/10/c485sn

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd éd). L. Erlbaum Associates.

Coltheart, M. (2005). Modeling reading: the dual-route approach. Dans M. J. Snowling & C. Hulme (Eds.), *The science of reading: A handbook* (p. 6–23). Blackwell Publishing. https://doi.org/10.1002/9780470757642.ch1

Demberg, V. et Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, *109*(2), 193-210. https://doi.org/10/bp6zw5

Duchowski, A. (2007). *Eye Tracking Methodology: Theory and Practice* (2e édition). Springer.

Engbert, R. et Kliegl, R. (2003). Microsaccades uncover the orientation of covert attention. *Vision Research*, *43*(9), 1035-1045. https://doi.org/10/fvqhbn

Fagerland, M. W. et Sandvik, L. (2009). The Wilcoxon-Mann-Whitney test under scrutiny. *Statistics in Medicine*, *28*(10), 1487-1497. https://doi.org/10/cpcc4b

Gernsbacher, M. A. et Kaschak, M. P. (2006). Psycholinguistics. Dans *Encyclopedia of Cognitive Science*. American Cancer Society. https://doi.org/10.1002/0470018860.s00601

Hareide, O. S., & Ostnes, R. (2017). Maritime usability study by analysing eye tracking data. *The Journal of Navigation*, 70(5), 927-943. https://10.1017/S0373463317000182

Hsu, J. (1996). *Multiple Comparisons: Theory and Methods*. CRC Press.

Ihaka, R. et Gentleman, R. (1996). R: a language for data analysis and graphics. *Journal of computational and graphical statistics*, *5*, 299-314.

Irwin, D. E. (2004). Fixation location and fixation duration as indices of cognitive processing. *The interface of language, vision, and action: Eye movements and the visual world*, *217*, 105-133.

Joanes, D. N. et Gill, C. A. (1998). Comparing measures of sample skewness and kurtosis. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *47*(1), 183-189. https://doi.org/10/b5rgv6

Juhasz, B. J. et Rayner, K. (2006). The role of age of acquisition and word frequency in reading: Evidence from eye fixation durations. *Visual Cognition*, *13*(7-8), 846-863. https://doi.org/10/dznsng

Karch, J. D. (2021). Psychologists should use Brunner-Munzel's instead of Mann-Whitney's U test as the default nonparametric procedure. *Advances in Methods and Practices in Psychological Science*, 4(2). https://doi.org/10.1177/2515245921999602

Kuhn, M. R. et Stahl, S. A. (2003). Fluency: A review of developmental and remedial practices. *Journal of Educational Psychology*, *95*(1), 3. https://doi.org/10/dsck55

Macinnes, J. J., Iqbal, S., Pearson, J. et Johnson, E. (2018). Mobile Gaze Mapping: A Python package for mapping mobile gaze data to a fixed target stimulus. *Journal of Open Source Software*, *3*(31), 984. https://doi.org/10/ggqr6q

Mangiafico, S. (2020). *rcompanion: Functions to Support Extension Education Program Evaluation* (version 2.3.25). https://CRAN.R-project.org/package=rcompanion

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372. https://doi.org/10/b5gdv6

Rayner, K. et Well, A. D. (1996). Effects of contextual constraint on eye movements in reading: A further examination. *Psychonomic Bulletin & Review*, *3*(4), 504-509. https://doi.org/10/fcv33j

Scheirer, C. J., Ray, W. S. et Hare, N. (1976). The Analysis of Ranked Data Derived from Completely Randomized Factorial Designs. *Biometrics*, *32*(2), 429-434. https://doi.org/10.2307/2529511

Smith, K. G., Schmidt, J., Wang, B., Henderson, J. M. et Fridriksson, J. (2018). Task-Related Differences in Eye Movements in Individuals With Aphasia. *Frontiers in Psychology*, *9*. https://doi.org/10/gftxtr

Solheim, O. J. et Uppstad, P. H. (2011). Eye-tracking as a tool in process-oriented reading test validation. *International Electronic Journal of Elementary Education; Kutahya*, *4*(1), 153-168. http://search.proquest.com/docview/912207173/abstract/6ECD5F81480A4DBDPQ/3

Stofer, K., & Che, X. (2014). Comparing experts and novices on scaffolded data visualizations using eye-tracking. *Journal of Eye Movement Research*, 7(5). https://doi.org/10.16910/jemr.7.5.2

Thiele, A., Henning, P., Kubischik, M. et Hoffmann, K.-P. (2002). Neural Mechanisms of Saccadic Suppression. *Science*, *295*(5564), 2460-2462. https://doi.org/10/bg6f8w

van der Lans, R., Wedel, M. et Pieters, R. (2011). Defining eye-fixation sequences across individuals and tasks: the Binocular-Individual Threshold (BIT) algorithm. *Behavior Research Methods*, *43*(1), 239-257. https://doi.org/10/fn8vt5

von der Malsburg, T. (2015). saccades: Detection of fixations in eye-tracking data. *R package version 0.1-1. URL http://CRAN. R-project. org/package= saccades*.

Wang, Q., Yang, S., Liu, M., Cao, Z. et Ma, Q. (2014). An eye-tracking study of website complexity from cognitive load perspective. *Decision Support Systems*, *62*, 1-10. https://doi.org/10/f556cj

Wass, S. V., Smith, T. J. et Johnson, M. H. (2012). Parsing eye-tracking data of variable quality to provide accurate fixation duration estimates in infants and adults. *Behavior Research Methods*, *45*(1), 229-250. https://doi.org/10/f4rm5f

9. Discussion générale

L'objectif général de la thèse était de montrer comment une approche computationnelle peut faciliter l'intégration de la psychométrie et des sciences cognitives dans le domaine de la compréhension de texte. Cet objectif général se divisait en deux objectifs spécifiques. Le premier objectif spécifique (articles 1 et 2) était de créer et de mettre à l'épreuve un outil d'analyse automatique du français québécois pouvant extraire des attributs associés, dans la littérature scientifique, à la difficulté du texte. Le second objectif spécifique (articles 3 et 4) était de créer une chaine de traitement des données d'oculométrie mobile s'inscrivant dans une approche de psychométrie computationnelle.

9.1. Rappel des quatre articles

L'article 1 présentait ALSI, un nouvel outil d'analyse linguistique pour le français québécois. ALSI fait appel à trois bases de données lexicales dont nous avons comparé les caractéristiques et le contenu. L'outil applique des techniques de linguistique computationnelle pour extraire une panoplie d'attributs que nous avons regroupés en une typologie simple. L'expérimentation consistait à analyser un corpus à l'aide d'ALSI, puis à appliquer une procédure de sélection afin d'identifier quels attributs étaient les plus prometteurs pour estimer la difficulté du texte. Les résultats prenaient la forme d'un tableau décrivant l'association statistique entre les attributs et le niveau de difficulté exprimé en années scolaires. Deux sélections d'attributs ont été retenues, soit une sélection « complète » de 21 variables, et une sélection réduite de 6 variables.

L'article 2 visait à tester la valeur des attributs produits par ALSI pour l'estimation du niveau de difficulté de textes, tout en faisant la démonstration d'une méthodologie rigoureuse pour estimer la performance de modèles de classification. Notre démarche s'inscrit parmi des

travaux similaires d'estimation de la difficulté du texte par l'apprentissage machine supervisé, et dont nous avons présenté un état de l'art. Afin de tester l'estimation de la difficulté du texte dans différentes conditions, nous avons fait varier la procédure (validation croisée avec *bootstrap* ou généralisation à de nouveaux textes), le modèle de classification (régression multinomiale ou SVM) et la sélection d'attributs (complète ou réduite). Les résultats soutiennent l'idée que les attributs d'ALSI permettent d'estimer la difficulté de textes avec une fiabilité équivalente ou supérieure à l'état de l'art.

L'article 3 avait pour but de tester la fidélité des données d'oculométrie mobile traitées automatiquement en recourant à des techniques dérivées de l'intelligence artificielle. L'article présentait d'abord un enjeu méthodologique découlant de l'utilisation de l'oculométrie mobile, essentiellement la difficulté à identifier ce que le participant regarde sans recourir à une analyse manuelle fastidieuse. Nous avons décrit une solution basée sur des techniques de vision par ordinateur (*computer vision*, une branche de l'intelligence artificielle) implémentée dans la boite à outils *Mobile Gaze Mapping* (MGM) pour Python. Nous avons testé MGM sur des données simulées reproduisant certaines particularités techniques des lunettes *Tobii Pro 2*, et imitant des conditions variées de collecte de données, par exemple un événement durant lequel la main couvre partiellement le stimulus visuel. Les résultats confirment que MGM peut automatiser certaines étapes de l'analyse d'oculométrie mobile, et nous proposons des modifications aux paramètres des algorithmes de MGM qui en amélioraient la performance.

L'article 4 visait à tester, sur des données réelles d'oculométrie de la lecture, une chaine de traitement de données intégrant MGM. Les participants à l'expérimentation pilote étaient des élèves de niveau postsecondaire (groupe de novices) et des enseignants ou professionnels de l'éducation (groupe d'experts) et devaient répondre à des questions portant sur de courts textes

argumentatifs. Les données provenaient de trois sources captées par les lunettes d'oculométrie : le mouvement des yeux, la vidéo de la caméra frontale, et l'audio. Le plan d'analyses statistiques puisait dans la recherche en oculométrie, et impliquait des techniques statistiques robustes justifiées par la présence de bruit statistique dans les données. Les résultats ont permis d'identifier des différences statistiquement significatives entre les patrons d'attention visuelle des deux groupes. Cette expérimentation montre qu'il est possible d'investiguer certains aspects de la compréhension de texte par l'oculométrie mobile grâce à une chaine de traitement de données combinant vision par ordinateur et statistiques robustes.

9.2. Synthèse des résultats

La création d'un programme informatique contribuant à tester des hypothèses scientifiques est déjà un résultat en soi (Johnson-Laird, 1983). En ce sens, les principaux résultats de cette thèse sont les outils et méthodes que nous avons développés et mis à l'épreuve. Les analyses statistiques effectuées dans les quatre articles visaient, dans l'ensemble, à appuyer la validité d'outils et techniques qui facilitent l'intégration de la psychométrie et des sciences cognitives. En somme, nous avons rejoint les objectifs de la thèse en montrant qu'il est possible d'étudier deux facettes de la lecture, soit la complexité intrinsèque du texte et les stratégies de compréhension de texte, à travers un cadre méthodologique s'appuyant sur des éléments de théorie et utilisant des méthodes computationnelles.

L'outil ALSI permet mesure des attributs caractérisant la complexité intrinsèque du texte, un construit multidimensionnel et complexe à mesurer. ALSI contribue à l'intégration théorique et méthodologique en tant qu'outil de mesure s'appuyant sur des éléments de théories empruntés à la psycholinguistique et à la linguistique computationnelle. Dans une approche de psychométrie computationnelle, ALSI peut être employé pour traduire du texte en attributs

d'intérêt, lesquels peuvent à leur tour être traduits en un niveau de difficulté estimé. Les articles 1 et 2 ont identifié plusieurs attributs et regroupements qui auraient un bon potentiel pour estimer le niveau de difficulté de textes en français québécois. La démarche de création d'ALSI constitue aussi un résultat puisqu'elle capture la manière dont nous avons opérationnalisé et mesuré ces attributs linguistiques. Nous proposons de plus une procédure de sélection d'attributs motivée par la théorie et par l'analyse de données. La bonne performance des modèles de classification avec sélection d'attributs complète ou réduite tend à indiquer que la procédure de sélection d'attributs (article 1) a fourni des ensembles de variables se généralisant bien à de nouveaux textes.

Nos travaux d'oculométrie ont testé la qualité des transformations opérées par MGM sur des données simulées (article 3) et réelles (article 4). Un résultat important des articles 3 et 4 est la documentation générée pour MGM, son fonctionnement interne, et son intégration à une chaine de traitement des données conforme à des principes statistiques robustes. Un autre résultat important est l'ensemble des modifications qui ont permis d'améliorer la performance de MGM : notre configuration proposée réduisant environ de moitié le temps d'exécution, et traitait avec succès 99,2% des données simulées, contre 88,1% pour la configuration par défaut de MGM. La méthodologie décrite à l'article 4 s'inscrit déjà dans une approche de psychométrie computationnelle telle que définie par von Davier (Polyak et al., 2017; von Davier et al., 2019). Conformément à cette approche, nous avons combiné des sources d'informations ayant une modalité différente (audio, vidéo, mouvement des yeux) pour former un portrait complexe des stratégies de compréhension de texte en recourant à des techniques empruntées au domaine de l'intelligence artificielle. Pour rejoindre l'objectif général de cette thèse, cette démarche a elle-même été guidée par des éléments théoriques provenant de disciplines variées, dont la vision par

ordinateur, la physiologie humaine et la psycholinguistique. Nous avons de plus appliqué des méthodes statistiques robustes de manière à vérifier et représenter clairement les différences observées et leur ampleur.

9.3. Intégration des travaux de la thèse à la recherche actuelle

L'outil et les méthodes créés ou perfectionnés dans le cadre de la thèse sont la meilleure manière de montrer que l'informatique facilite le rapprochement entre psychométrie et sciences cognitives, et constituent donc le principal résultat de nos travaux. Cela dit, plusieurs résultats au sens conventionnel du terme – des résultats empiriques – ont émergé de notre démarche, et entrent en relation avec la recherche actuelle.

9.3.1. Contributions à l'évaluation d'outils et d'éléments de méthodologie.

Nos travaux ont contribué à l'évaluation de certains outils et éléments méthodologiques. Les corrélations obtenues pour les attributs produits à l'aide de *rsyntax* (Welbers et al., 2020) montrent la capacité de cette bibliothèque R à extraire des constituants complexes de la phrase. Par exemple, l'attribut GNC_m (notre moyenne de groupe nominaux complexes par phrase) était extrait à l'aide de *rsyntax* et avait une association statistique intéressante (*rs* = 0,621**)** avec l'année scolaire. Les résultats portant sur la corrélation des attributs avec le niveau de difficulté indiquent de plus que le lexique Manulex (Lété et al., 2004) peut contribuer à caractériser la difficulté de textes québécois bien qu'il ait été formé à partir de textes français. Nous avons en outre confirmé la valeur du lexique ÉQOL pour estimer la difficulté lexicale, et répliqué avec succès les analyses de Stanké et al. (2019) montrant la corrélation d'ÉQOL avec d'autres lexiques.

Les essais de classification du texte par année scolaire (article 2) s'inscrivent à la suite d'une série d'études dont nous avons fait la synthèse, et réutilisent certaines techniques courantes

tout en apportant un regard critique sur la méthodologie en vigueur. Notre méthodologie se distinguait des pratiques courantes par l'utilisation d'une procédure VC (validation croisée) répétée combinée à une généralisation, alors que les études comparables ont appliqué la validation croisée sans répétition et le plus souvent sans généralisation. Les résultats de la procédure VC ont montré comment la performance obtenue pouvait varier selon la composition des blocs, appuyant bien-fondé de mises en garde concernant la validation croisée qui, dans plusieurs scénarios, n'est pas suffisante pour produire une estimation réaliste des performances des modèles de classification lorsqu'appliqués à de nouvelles données (Fu et al., 2005; Kim, 2009).

Du côté de l'oculométrie, nos essais sur des données simulées et réelles constituent, à notre connaissance, les premiers essais rigoureux pour valider *Mobile Gaze Mapping* (MacInnes, 2020; Macinnes et al., 2018) comme outil d'analyse automatisée des données d'oculométrie mobile. Nos travaux pourraient avoir un impact significatif dans le milieu de la recherche en stimulant le développement et l'utilisation de méthodes de vision par ordinateur en oculométrie mobile. Les résultats de l'article 3 contribuent également à la recherche comparant la performance d'algorithmes de vision par ordinateur dans différentes conditions (Andersson et Reyna Marquez, 2016; Kalms et al., 2017; Tareen et Saleem, 2018). Les travaux dans ce domaine comparent généralement les algorithmes en utilisant leur configuration par défaut – l'originalité de notre approche était de tester comparativement la performance sous différentes valeurs de paramètres, et selon différents scénarios pouvant survenir lors de la collecte de données.

9.3.2. Contributions à des terrains actifs de recherche théorique

Certains résultats observés dans le cadre de cette thèse contribuent à la réflexion théorique sur la compréhension de texte, notamment le lien entre la difficulté du texte et la cohésion linguistique (McNamara et Kintsch, 1996). Nous avons observé des corrélations positives entre les attributs portant sur la cohésion linguistique et la difficulté du texte exprimée en années scolaires (article 1), appuyant les résultats d'études précédentes (Dascalu et al., 2014; Duran et al., 2007; François et Fairon, 2012). Bien que notre étude d'expérimentation oculométrique (article 4) ne confirme pas directement les conclusions d'O'Reilly et McNamara (2007) sur l'effet combiné de la cohésion linguistique et de l'expertise, nos résultats appuient l'idée selon laquelle les experts emploient des stratégies qui intègrent davantage l'information lors d'une tâche de lecture demandant de faire des inférences.

Les résultats des articles 1 et 2 contribuent à la recherche sur la modélisation de la difficulté du texte. Nous remarquons, par exemple, que plusieurs attributs considérés comme « de surface », ou de complexité 1 dans notre typologie, avaient une assez bonne association statistique avec le niveau de difficulté (voir Tableau 7 de l'Article 1). C'est le cas, par exemple, pour la longueur moyenne des phrases (rs = 0,66) et la longueur moyenne des mots, en nombre de caractères (rs = 0,62). Nos résultats remettent en perspective le discours critique à l'endroit des attributs textuel et montrent que ce type d'attribut ne devrait pas être systématiquement exclu des discussions entourant les facteurs associés à la difficulté intrinsèque du texte. Nos travaux s'intègrent également à la littérature théorique portant sur la classification du texte par niveau de difficulté. Les résultats rapportés à l'article 2 accordaient l'avantage au modèle de classification SVM tout en montrant que RMN était supérieur dans certaines conditions et selon certains indicateurs (régression multinomiale). Cette ambiguïté va dans le sens des comparaisons

effectuées par Verplancke et al. (2008) et qui montraient que RMN et SVM avaient une performance similaire en généralisation et en validation croisée. Bien que SVM soit déjà un modèle populaire en classification du texte (Deutsch et al., 2020; Pawar et Gawande, 2012), d'autres travaux seraient requis pour évaluer les avantages de différents modèles de classification en fonction des caractéristiques du corpus et des indicateurs utilisés pour évaluer la performance.

9.4. Application et portée des travaux

Nous voyons plusieurs applications au domaine de l'éducation de l'outil ALSI développé dans le cadre de la thèse. L'estimation de la difficulté intrinsèque du texte par ALSI pourrait faire partie d'une démarche d'édition ou de sélection de matériel didactique. ALSI peut aider à contrôler la difficulté linguistique d'une évaluation – nous avons d'ailleurs ajusté la difficulté des items de compréhension de texte (article 4) à l'aide d'une version préliminaire d'ALSI. Pour l'enseignement du français, ALSI permet de sélectionner des textes favorisant l'apprentissage de certains objets d'enseignement, par exemple le participe présent ou la phrase subordonnée.

D'autres applications des outils et méthodologies développés dans le cadre de la thèse sont envisageables dans le domaine de la recherche. Les attributs extraits par ALSI peuvent servir pour des fins de recherche sur les sources de difficulté du langage, par exemple les effets de cohésion linguistique ou de certaines formes syntaxiques. ALSI pourrait être appliqué à d'autres problèmes de classification du texte, comme la classification par genre textuel (narratif, expressif, argumentatif, etc.) Une application que nous n'avons pas eu l'occasion de tester, mais qui a un fort potentiel pour la recherche est d'utiliser ALSI dans le cadre d'une étude dans le domaine de l'écriture; l'outil pourrait analyser des productions écrites afin d'examiner les liens entre les attributs linguistiques et la qualité générale de la production évaluée par le scripteur (Taguchi et al., 2013). La méthodologie d'oculométrie développée à l'article 3 et 4 a un fort

potentiel pour l'étude des processus cognitifs liés à la lecture, elle ouvre un terrain de recherche auprès de jeunes élèves dont il peut être difficile d'étudier les processus de lecture par l'oculométrie stationnaire, laquelle demande une quasi-immobilité. Une méthodologie similaire pourrait être appliquée à d'autres contextes de recherche où l'oculométrie mobile est fréquemment employée, comme la pédagogie médicale (Koester et al., 2017; Kok et Jarodzka, 2017).

À terme, l'outil ALSI et la méthodologie d'oculométrie mobile pourraient être intégrés dans un cadre de psychométrie computationnelle plus complet, tel qu'illustré à la Figure 1. Ce cadre méthodologique envisagé reprend la méthodologie d'étude de la compréhension de texte décrit à l'article 4, et lui ajoute une source d'information, soit la difficulté estimée du texte sur lequel portent les items. Il serait ainsi possible d'étudier l'attention visuelle en fonction de la difficulté intrinsèque du texte (estimée par ALSI), de la difficulté de l'item (estimée par un modèle psychométrique), et de la stratégie de réponse (inférée par analyse qualitative de l'enregistrement audio de la verbalisation du participant).

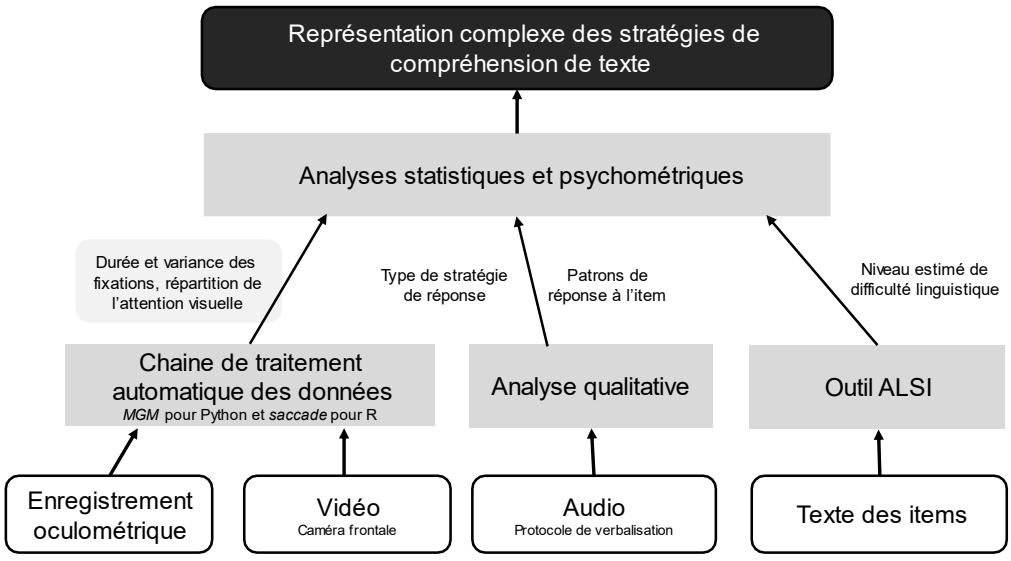


Figure 1. Cadre méthodologique de psychométrie computationnelle envisagé, intégrant l'outil ALSI et l'oculométrie mobile.

Le programme de recherche que nous proposons est à la fois la suite logique et la synthèse des quatre articles, et reprend la structure générale de la psychométrie computationnelle telle que définie par von Davier (2019). Nous jugeons cependant que, telle quelle, l'approche préconisée par von Davier ne place pas suffisamment d'emphase sur la justification scientifique des inférences inductives, qui sont les transitions vers une représentation plus complexe de l'information (Gärdenfors, 1990). Tout au long de la thèse, nous avons accordé beaucoup de place aux procédés par lesquels les données sont transformées, en décrivant autant les mécanismes que leurs appuis théoriques. Par exemple, dans la description d'ALSI, nous avons justifié les inférences reliant les attributs à la difficulté du texte en recourant à des éléments théoriques centrés autour d'un modèle cognitif de la lecture (Coltheart, 2005). En oculométrie, le même modèle appuie les inférences entre la durée des fixations visuelles et l'expertise : les novices ont tendance à faire des fixations plus longues, car moins de mots sont connus et reconnus automatiquement, donc davantage de mots sont décodés par un processus plus laborieux (Juhasz et Rayner, 2006). Nos résultats gagnent ainsi en validité scientifique, et peuvent en retour servir à améliorer nos méthodologies tout en entrant en dialogue avec la recherche sur la complexité linguistique. Une psychométrie computationnelle soucieuse d'expliciter et justifier ses inférences inductives peut donc faciliter l'intégration d'éléments théoriques en psychométrie, tout en contribuant à l'élaboration de théories en sciences cognitives.

9.5. Limites et avenues de recherche

Bien que cette thèse ait présenté plusieurs justificatifs à l'appui de sa méthodologie et de ses résultats, il convient d'en examiner les limites. Dans ce qui suit, nous présentons les limites

de la thèse qui nous ont semblé les plus saillantes, ainsi que les pistes de recherche qui découlent de celles-ci.

Les résultats portant sur la valeur d'attributs linguistiques pour estimer la difficulté du texte (articles 1 et 2) doivent être considérés avec précaution puisque nous n'avons pas directement évalué la contribution des attributs aux modèles de classification. Plusieurs méthodes et indicateurs ont été proposés pour mesurer l'importance des attributs et pourraient être exploités dans de futures publications (Altmann et al., 2010; Vidovic et al., 2016); mesurer l'importance de différents regroupements d'attributs (par ex. : lexicaux versus syntaxiques) nous semble particulièrement pertinent pour la recherche sur les facteurs de difficulté intrinsèque du texte. De plus, nos travaux ont écarté le genre textuel (par ex. : descriptif ou narratif) afin centrer nos recherches sur la difficulté linguistique attribuable aux attributs linguistiques mesurés par ALSI. Le genre textuel pourrait cependant être réintroduit dans de futurs travaux s'intéressant, par exemple, aux sources de difficulté linguistique de certains genres textuels.

Le fait qu'ALSI ait été testé sur un corpus issu du milieu de l'éducation est une limitation au sens où ces textes pourraient avoir des caractéristiques spécifiques suivant le cursus, et qui les rendrait plus faciles à regrouper par année scolaire. Nous nous interrogeons quant à la validité de l'outil ALSI lorsqu'appliqué à des textes provenant de sources autres que l'enseignement primaire ou secondaire québécois. D'autres travaux pourraient s'intéresser à la capacité d'ALSI à extraire des attributs capturant la difficulté de textes tirés d'autres contextes, notamment la formation professionnelle, ou l'information communiquée au public par différents organismes gouvernementaux. Pour tester son utilisation en contexte d'éducation aux adultes, nous voyons l'intérêt d'inclure dans ALSI des lexiques comprenant de la terminologie propre à un milieu professionnel et qui pourrait constituer une source de difficulté.

Dans les deux articles portant sur la modélisation de la complexité intrinsèque du texte, nous avons assumé une absence d'interaction entre les attributs, qui influenceraient la difficulté du texte de manière isolée. La procédure de sélection (article 1) et les modèles de classifications utilisés (article 2) présumaient que la manière dont les attributs influencent le niveau de difficulté n'est pas elle-même influencée par des effets d'interaction entre les attributs. Le choix de ne pas considérer les interactions constitue une limitation considérant notre définition initiale de la difficulté linguistique comme une interaction de facteurs (article 1). Dans le contexte de nos articles, ce choix méthodologique était motivé par le désir de simplifier les analyses et d'éviter le surapprentissage des modèles (*overfitting*) qui peut survenir lorsque trop de termes d'interaction sont ajoutés (Tu, 1996). Des travaux futurs pourraient tenter de mieux modéliser la difficulté du texte en incluant des termes d'interaction motivés par la recherche en psycholinguistique (Hagoort, 2003).

Notre expérimentation en oculométrie constituait, à notre connaissance, la première application de MGM au domaine de la lecture. Il est difficile, à ce stade, de se prononcer sur la généralisation de la méthode que nous proposons à de nouveaux contextes expérimentaux. Les lunettes d'oculométrie utilisées imposent d'emblée une limitation en raison d'une précision et d'une cadence d'échantillonnage relativement faible, ce qui diminue en retour la qualité de la détection des saccades et fixations (Leube et al., 2017). D'autres travaux pourraient répéter l'expérimentation en utilisant un appareil ayant une meilleure résolution spatiale et temporelle. Combiner la méthode proposée à de la pupillométrie serait également une avenue à explorer pour estimer la charge cognitive. Une autre limitation importante de notre expérimentation oculométrique est que nous n'avons pas contre-vérifié systématiquement la transformation des données par MGM. Une autre étude pourrait compléter nos tests sur des données simulées et

comparer la qualité des données réelles produites par la méthode MGM et la méthode d'analyse manuelle.

Enfin, une limitation de l'ensemble de la thèse est le peu de place accordée à la perspective des élèves sur la complexité du texte et les processus de lecture. Nous avons considéré comme mesure de la difficulté du texte l'année scolaire qui lui était associée. Une approche complémentaire pourrait être d'estimer la difficulté du texte par une tâche de compréhension (Dascalu et al., 2014). Une autre approche serait d'entrainer des modèles à classifier des textes selon la difficulté telle qu'évaluée par des élèves, ce qui à notre connaissance serait une première dans le domaine. Du côté de l'expérimentation oculométrique, nous nous sommes intéressés à la difficulté subjective du texte telle qu'exprimée par l'attention visuelle, mais n'avons pas utilisé lors des analyses les enregistrements audios où les participants justifient leur réponse. Ces enregistrements pourraient être analysés qualitativement dans le cadre d'une étude portant sur les stratégies de réponse aux items de compréhension de texte.

9.6. Conclusion

Malgré leurs limites, les travaux de la présente thèse ont montré qu'une approche computationnelle permettait de mesurer des facettes de la difficulté du texte selon des méthodologies guidées par la théorie. Les articles 1 et 2 ont contribué aux méthodologies d'évaluation de la difficulté des textes en introduisant l'outil de traitement automatique du langage naturel ALSI, puis en montrant par l'application de méthodes robustes que les attributs produits par ALSI permettent une modélisation de la difficulté des textes qui satisfait ou dépasse l'état de l'art. Les articles 3 et 4 ont contribué aux méthodologies d'oculométrie mobile pour la comparaison des processus de lecture de novices et d'experts en testant MGM sur des données simulées, puis en intégrant MGM à une chaine de traitement des données.

Nos travaux sont importants dans un contexte où peu de ressources de ce type sont disponibles pour étudier la complexité linguistique du français québécois de niveau scolaire. Ils constituent un pas en avant dans le développement et la validation de méthodologies explorant la bande limitrophe entre l'étude des fonctions cognitives de la lecture et l'évaluation de la compréhension de texte. Dans son ensemble, la thèse amende le cadre de psychométrie computationnelle de von Davier en prenant au sérieux le rôle des théories scientifiques dans la justification des inférences inductives reliant les niveaux d'information (par ex. : les patrons de mouvement oculaire et les processus de lecture). Nous avons contribué au développement et à la démocratisation des technologies numériques pour l'étude de la lecture, dans une approche qui stimule et facilite l'intégration théorique, plutôt que la substitution de la théorie par le logiciel. L'outil ALSI et le cadre méthodologique d'oculométrie ont de multiples avenues d'application en éducation et en recherche, allant de l'ajustement du niveau linguistique d'un test à la construction d'un cadre de psychométrie computationnelle pour l'étude de la lecture.

9.7. Bibliographie

Altmann, A., Toloşi, L., Sander, O. et Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. *Bioinformatics*, *26*(10), 1340-1347. https://doi.org/10/cm7h6d

Andersson, O. et Reyna Marquez, S. (2016). *A comparison of object detection algorithms using unmanipulated testing images : Comparing SIFT, KAZE, AKAZE and ORB* [thèse de doctorat]. KTH Royal Institute of Technology in Stockholm. http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-186503

Coltheart, M. (2005). Modeling reading: the dual-route approach. Dans M. J. Snowling & C. Hulme (Eds.), *The science of reading: A handbook* (p. 6–23). Blackwell Publishing. https://doi.org/10.1002/9780470757642.ch1

Dascalu, M., Stavarache, L. L., Trausan-Matu, S., Dessus, P. et Bianco, M. (2014). Reflecting comprehension through french textual complexity factors. Dans *2014 IEEE 26th International Conference on Tools with Artificial Intelligence* (p. 615-619). https://doi.org/10/ghjqf5

Deutsch, T., Jasbi, M. et Shieber, S. (2020). Linguistic features for readability assessment. Dans *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications* (p. 1-17). https://doi.org/10/gjnk6q

Duran, N. D., Bellissens, C., Taylor, R. S. et McNamara, D. S. (2007). Quantifying text difficulty with automated indices of cohesion and semantics. Dans *Proceedings of the Annual Meeting of the Cognitive Science Society* (vol. 29).

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).

Gärdenfors, P. (1990). Induction, Conceptual Spaces and AI. *Philosophy of Science*, *57*(1), 78-95. https://doi.org/10/bp2q69

Hagoort, P. (2003). Interplay between syntax and semantics during sentence comprehension: ERP effects of combining syntactic and semantic violations. *Journal of cognitive neuroscience*, *15*(6), 883-899. https://doi.org/10/bkcrkk

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Harvard University Press.

Juhasz, B. J. et Rayner, K. (2006). The role of age of acquisition and word frequency in reading: Evidence from eye fixation durations. *Visual Cognition*, *13*(7-8), 846-863. https://doi.org/10/dznsng

Kalms, L., Mohamed, K. et Göhringer, D. (2017). *Accelerated Embedded AKAZE Feature Detection Algorithm on FPGA* [presentation de conference]. The 8th International Symposium, Bochum, Germany (p. 1-6). https://doi.org/10/ggmtqm

Kim, J.-H. (2009). Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational Statistics & Data Analysis*, *53*(11), 3735-3745. https://doi.org/10/dtqx5t

Koester, T., Brøsted, J. E., Jakobsen, J. J., Malmros, H. P. et Andreasen, N. K. (2017). The use of eye-tracking in usability testing of medical devices. *Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care*, *6*(1), 192-199. https://doi.org/10/gg2bwz

Kok, E. M. et Jarodzka, H. (2017). Before your very eyes: the value and limitations of eye tracking in medical education. *Medical Education*, *51*(1), 114-122. https://doi.org/10/f9g8sg

Lété, B., Sprenger-Charolles, L. et Colé, P. (2004). MANULEX: A grade-level lexical database from French elementary school readers. *Behavior Research Methods, Instruments, & Computers*, *36*(1), 156-166. https://doi.org/10/djzxmb

Leube, A., Rifai, K. et Rifai, K. (2017). Sampling rate influences saccade detection in mobile eye tracking of a reading task. *Journal of Eye Movement Research*, *10*(3). https://doi.org/10.16910/jemr.10.3.3

MacInnes, J. J. (2020). *Mobile Gaze Mapping* [Python]. https://github.com/jeffmacinnes/mobileGazeMapping

Macinnes, J. J., Iqbal, S., Pearson, J. et Johnson, E. (2018). Mobile Gaze Mapping: A Python package for mapping mobile gaze data to a fixed target stimulus. *Journal of Open Source Software*, *3*(31), 984. https://doi.org/10/ggqr6q

McNamara, D. S. et Kintsch, W. (1996). Learning from texts: Effects of prior knowledge and text coherence. *Discourse processes*, *22*(3), 247-288. http://www.tandfonline.com/doi/abs/10.1080/01638539609544975

O'Reilly, T. et 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

Pawar, P. et Gawande, S. (2012). A comparative study on different types of approaches to text categorization. *International Journal of Machine Learning and Computing*, 423-426. https://doi.org/10/gjr9rh

Polyak, S. T., von Davier, A. A. et Peterschmidt, K. (2017). Computational psychometrics for the measurement of collaborative problem solving skills. *Frontiers in psychology*, *8*, 2029. https://doi.org/10/gcnjd5

Stanké, B., Le Mené, M., Rezzonico, S., Moreau, A., Dumais, C., Robidoux, J., Dault, C. et 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).

Taguchi, N., Crawford, W. et Wetzel, D. Z. (2013). What linguistic features are indicative of writing quality? A case of argumentative essays in a college composition program. *Tesol Quarterly*, *47*(2), 420-430. https://doi.org/10.1002/tesq.91

Tareen, S. A. K. et Saleem, Z. (2018, mars). A comparative analysis of SIFT, SURF, KAZE, AKAZE, ORB, and BRISK. Dans *2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)* (p. 1-10). https://doi.org/10.1109/ICOMET.2018.8346440

Tu, J. V. (1996). Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of Clinical Epidemiology*, *49*(11), 1225-1231. https://doi.org/10/fg794g

Verplancke, T., Van Looy, S., Benoit, D., Vansteelandt, S., Depuydt, P., De Turck, F. et Decruyenaere, J. (2008). Support vector machine versus logistic regression modeling for prediction of hospital mortality in critically ill patients with haematological malignancies. *BMC Medical Informatics and Decision Making*, *8*(1), 56. https://doi.org/10/b739z7

Vidovic, M. M.-C., Görnitz, N., Müller, K.-R. et Kloft, M. (2016). *Feature Importance Measure for Non-linear Learning Algorithms.* arXiv. http://arxiv.org/abs/1611.07567

von Davier, A. A., Deonovic, B., Yudelson, M., Polyak, S. T. et Woo, A. (2019). Computational psychometrics approach to holistic learning and assessment systems. *Frontiers in Education*, *4*. https://doi.org/10/ggjp3q

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