Revue québécoise de linguistique

SATO-CALIBRAGE : présentation d'un outil d'assistance au choix et à la rédaction de textes pour l'enseignement

François Daoust, Léo Laroche et Lise Ouellet

Volume 25, numéro 1, 1996

Lisibilité et intelligibilité

URI : https://id.erudit.org/iderudit/603132ar
DOI : https://doi.org/10.7202/603132ar

Aller au sommaire du numéro

Éditeur(s)

Université du Québec à Montréal

ISSN

0710-0167 (imprimé)
1705-4591 (numérique)

Découvrir la revue

Citer cet article

Daoust, F., Laroche, L. & Ouellet, L. (1996). *SATO-CALIBRAGE* : présentation d'un outil d'assistance au choix et à la rédaction de textes pour l'enseignement. *Revue québécoise de linguistique*, *25*(1), 205–234. https://doi.org/10.7202/603132ar

Résumé de l'article

Depuis quelques années, un projet conjoint mené par l'Université du Québec à Montréal et le ministère de l'Éducation porte sur le développement d'un indice de lisibilité des textes, appelé SATO-CALIBRAGE. Des préoccupations de choix et de rédaction de textes en contexte scolaire ont été à l'origine du projet. Dans cet article, de nature descriptive, on présente l'application SATO-CALIBRAGE et on explique la méthodologie qui a été adoptée en exposant les dispositifs linguistique et mathématique qui ont été mis en place. L'indice lui-même et la façon de le calculer sont aussi présentés. En outre, quelques exemples d'utilisations en didactique, en évaluation et en rédaction sont fournis afin d'aider à comprendre les contextes d'utilisation éventuels.

Revue québécoise de linguistique, vol. 25, n° 1, 1996, © RQL (UQAM), Montréal

SATO-CALIBRAGE: PRÉSENTATION D'UN OUTIL D'ASSISTANCE AU CHOIX ET À LA RÉDACTION DE TEXTES POUR L'ENSEIGNEMENT

François Daoust
Université du Québec à Montréal
Léo Laroche
Lise Ouellet
Ministère de l'Éducation

1. Introduction

La lecture est au coeur des apprentissages fondamentaux de l'école. On souhaite non seulement que les élèves sachent lire et qu'ils aiment la lecture, mais aussi qu'ils puissent développer des stratégies de lecture efficaces afin de comprendre tout texte qu'ils doivent lire ou qu'ils ont le goût de lire. L'enseignement et l'évaluation de la lecture représentent donc des défis constants pour les enseignants[1] du primaire et du secondaire. Parmi leurs préoccupations, le choix des textes figure en tête de liste puisqu'il est le moteur de tout apprentissage de la lecture.

SATO-CALIBRAGE est un outil informatique qui veut répondre à ce besoin en offrant une assistance pour le choix et la rédaction de textes. L'une de ses caractéristiques est de donner un indice à un texte pour le situer sur un continuum, établi de la première année du primaire à la cinquième année du secondaire.

Après avoir présenté la problématique du projet, nous expliquerons la méthodologie suivie et nous ferons une description du prototype. Quelques exemples d'applications pédagogiques seront ensuite décrits.

[1] La féminisation des titres présente dans la version originale de cet article n'a pas été retenue par la rédaction.

Le présent article, de nature descriptive, présente sommairement nos travaux. Un rapport de recherche, cf. Daoust, Laroche, Ouellet & al. (1993), expose en détail l'historique du projet, ainsi que ses dimensions linguistique et statistique.

2. Problématique

Dans le monde de l'enseignement, trois responsabilités professionnelles requièrent un éclairage particulier face aux textes: la planification de l'enseignement, l'évaluation de la lecture et la rédaction de texte. Nous décrirons rapidement ces situations et nous verrons, dans la suite de l'article, comment un outil informatisé peut faciliter la tâche du choix de texte.

Plusieurs enseignants planifient leurs cours à l'aide du matériel didactique offert sur le marché. D'autres préfèrent présenter aux élèves des textes venant de sources diversifiées. Ils doivent alors se demander si ces textes sont accessibles aux élèves. Dans d'autres occasions, la planification de la lecture ne se fera pas en fonction d'une thématique mais plutôt à partir d'objectifs d'apprentissage déterminés; par exemple, trouver le sens d'un mot inconnu d'après le contexte, comprendre les liens qu'établissent certains mots de relation ou comprendre le sens de phrases très longues. Dans tous ces cas, au moment de l'**enseignement**, on a besoin de savoir si le texte est adapté à son groupe d'élèves et quels sont les défis qu'il présente.

Choisir le ou les textes qui serviront à l'**évaluation** de la lecture est une tâche lourde de conséquence. Le texte choisi est en effet le point de départ de la définition de la tâche et de l'élaboration du questionnaire. Il arrive parfois qu'une épreuve de lecture n'atteigne pas ses objectifs parce que le texte est trop facile ou trop difficile. On veut donc un texte qui soit du bon niveau. On veut aussi repérer les difficultés du texte, soit pour les atténuer, ou soit au contraire, pour vérifier si l'élève est capable de surmonter les obstacles.

Les textes utilisés pour l'enseignement ou pour l'évaluation peuvent venir de sources externes. On peut aussi choisir de rédiger des textes. Toutefois, au moment de la **rédaction**, les mêmes questions se posent: on veut un texte ni trop facile ni trop difficile ou on veut rédiger un texte qui présente des défis particuliers pour que les élèves développent des stratégies de lecture précises. Comment peut-on affronter ce problème? Bien sûr, l'intuition et l'expérience viendront à la rescousse, mais est-ce objectif et suffisant? Est-on assuré d'avoir le texte que l'on recherche? Spontanément, le recours aux formules de lisibilité vient à l'esprit.

Il existe un certain nombre d'indices pour mesurer la lisibilité d'un texte[2]. Cependant, la plupart de ces mesures ont été conçues pour des textes rédigés en langue anglaise. Voilà pourquoi nous avons voulu développer un instrument de mesure de la lisibilité qui soit adapté au milieu scolaire francophone du Québec. Pour ce faire, nous avons opté pour une démarche expérimentale basée sur une analyse comparative de larges corpus de textes. Ces corpus se veulent représentatifs du matériel fourni aux élèves québécois dans les cours de français langue maternelle au primaire et au secondaire.

Aussi, dès le début de nos travaux en 1989, il est apparu nécessaire d'établir une collaboration étroite entre le ministère de l'Éducation, le Centre d'ATO de l'UQAM et les milieux scolaires. On a donc mis sur pied un comité d'appui composé de conseillers pédagogiques de français (primaire et secondaire), de responsables d'élaboration d'épreuves (ministère de l'Éducation et Banque d'Instruments de Mesure — BIM) ainsi que de quelques personnes du milieu collégial et universitaire.

3. Méthodologie

Pour réaliser notre recherche, nous nous sommes appuyés sur certaines hypothèses touchant la nature du discours écrit. Pour valider nos hypothèses, nous avons mis en place un dispositif linguistique. Finalement, des traitements statistiques nous ont permis d'élaborer un indice rendant compte de la lisibilité des textes.

3.1 Cadre expérimental

Comme tout projet en analyse de texte, notre démarche est fondée sur un certain nombre d'hypothèses sur la nature du discours. Ainsi, par exemple, on peut considérer que les textes fournis aux élèves de première année devraient, au-delà des variations individuelles propres à chaque texte, partager des caractéristiques communes qui les destinent à leur utilisation dans un contexte d'apprentissage ou d'évaluation. En d'autres termes, nous posons en postulat l'hypothèse générale de cohérence du discours social, plus spécifiquement ici, du discours produit dans le cadre de l'institution scolaire et destiné à un public cible composé d'enfants en situation d'apprentissage. D'un point de vue sociologique donc, les textes constituant le matériel scolaire s'inscrivent dans un

[2] Pour une présentation de la problématique de la lisibilité, cf. Gélinas-Chebat & al. (1993). Pour une brève recension des indices les plus connus, cf. Laroche (1990).

cadre institutionnel déterminé avec des acteurs sociaux historiquement définis. Les indices produits au cours de notre recherche seront donc marqués par ce contexte social. Cela n'exclut pas que l'on puisse utiliser nos indices dans un cadre social différent. Il faudra cependant considérer ces différences dans l'évaluation des résultats.

En s'appuyant sur l'hypothèse générale de cohérence du discours scolaire, nous allons étudier le fonctionnement discursif en observant un ensemble de textes individuels. La question de la représentativité des données, à savoir ici les textes fournis aux élèves, est donc une des premières questions à poser dans une approche expérimentale. Cette représentativité implique des hypothèses sur l'objet à analyser, sur sa cohérence et sa variabilité. La constitution du corpus utilisé pour cette recherche traduit bien ces préoccupations. Pour chaque texte inclus dans ce corpus, un certain nombre de renseignements ont été recueillis: la classe d'enseignement où le texte est utilisé, sa provenance et, dans certains cas, le type de discours. Des données statistiques sur une quinzaine de variables linguistiques ont ensuite permis d'épurer le corpus disponible en rejetant les textes atypiques. Par exemple, un texte où il n'y a pas de points ou presque est probablement un poème. Enfin, les cas litigieux ont été soumis à un groupe d'experts qui ont confirmé ou révisé le classement des textes par rapport aux classes d'enseignement.

3.1.1 Constitution du corpus

Les textes qui composent notre corpus proviennent du matériel didactique approuvé par le ministère de l'Éducation du Québec (MEQ). Il peut aussi s'agir de textes utilisés pour l'évaluation. Dans les deux cas, il était donc possible d'attribuer une classe d'appartenance à chaque texte. Au début, le corpus contenait des textes qui correspondaient à tous les types de textes imposés dans les programmes d'études. Par la suite, les poèmes, les chansons, les comptines, les charades, les faits divers, les lettres d'invitation, les contrats et les extraits de pièce de théâtre ont été exclus et ce, pour deux raisons: d'une part, ces types de textes étaient marginaux dans le corpus et, d'autre part, leurs formes linguistiques les démarquaient nettement de l'ensemble. Nous avons donc décidé de nous concentrer sur des textes homogènes, les plus couramment utilisés dans l'enseignement du français au Québec; nous avons cherché à constituer un corpus d'environ 50 textes par classe. Le tableau suivant présente la composition du corpus utilisé dans l'élaboration de l'indice SATO-CALIBRAGE.

Tableau 1
Description du corpus

Ordre d'enseignement / Classes	PRIMAIRE	SECONDAIRE
Première année	65	63
Deuxième année	82	43
Troisième année	63	49
Quatrième année	60	57
Cinquième année	60	67
Sixième année	70	–
TOTAL	400	279

Au total, le corpus compte 679 textes.

Le MEQ dispose de grilles permettant d'évaluer la pertinence d'un texte pour une classe d'enseignement donnée ou, à tout le moins, pour un cycle d'enseignement. Parmi les éléments de cette grille, on retrouve des caractéristiques qu'il est possible de détecter à l'aide d'un programme informatique, par exemple la longueur du texte, la longueur des phrases, etc. Notre objectif était donc, dans un premier temps, d'automatiser et de valider certains éléments de cette grille d'évaluation utilisée dans le milieu scolaire. Dans un deuxième temps, l'objectif visé était d'élaborer un protocole expérimental afin d'enrichir le modèle interprétatif.

3.1.2 Élaboration du protocole expérimental

Pour réaliser ce protocole, nous avons retenu le logiciel SATO qui nous permettait de dépister rapidement une variété d'indices textuels. Cet outil informatique représente le texte sur un plan composé de deux axes. On a d'abord un axe lexical qui dresse la liste du vocabulaire utilisé dans le texte. On a ensuite un axe syntagmatique qui restitue la linéarité du texte qui se donne en fait comme une suite d'occurrences des lexèmes.

Destiné à soutenir des activités d'analyse, le logiciel offre aussi la possibilité d'annoter et de catégoriser le texte, permettant ainsi de marquer le dépistage de processus discursifs; ces marques peuvent s'inscrire sur l'axe lexical ou sur l'axe syntagmatique. Nous utilisons le terme 'propriété' pour désigner une classe d'annotations ou de catégories permettant de marquer des lexèmes ou des occurrences.

Les dispositifs expérimentaux utilisés dans ce cadre ont pris la forme de scénarios de commandes qui aboutissent à la production de données quantitives correspondant à divers indices ou variables: nombre ou proportion de phrases utilisant telle ou telle construction, fréquence d'utilisation de tel lexème ou telle classe de lexèmes. Appliqués sur chacun des textes du corpus, ces scénarios produisent finalement une très grande quantité de données. Il peut arriver que des données puissent être interprétées directement par des spécialistes. Le plus souvent cependant, pour être évaluées correctement, les données doivent être examinées à travers une modélisation mathématique (cf. 3.3).

3.2 Le dispositif linguistique

Nous désignons par dispositif linguistique, l'ensemble des ressources linguistiques déployées à l'intérieur du projet SATO-CALIBRAGE. Ces ressources sont les suivantes: 1) bases de données lexicales; 2) procédures pour repérer des noms propres; 3) procédures pour identifier les verbes conjugués; 4) procédures de dépistage des locutions grammaticales; 5) et, s'appuyant sur les dispositifs précédents, procédures permettant le dépistage de phrases potentiellement complexes.

3.2.1 Bases de données lexicales

Les bases de données lexicales prennent la forme de dictionnaires. Ce sont des fichiers qui contiennent des informations sur des formes lexicales. En consultant ces dictionnaires, on peut annoter le lexique d'un texte en transférant sur une propriété lexicale les renseignements se trouvant dans le dictionnaire. Deux bases de données sont ici utilisées. La première[3] contient la catégorie

[3] Cette base de données, appelée couramment 'la BDL', a été développée au départ par Luc Dupuy dans le cadre du projet SACAO (Système d'analyse de contenu assistée par ordinateur), Programme Actions spontanées, FCAR 1989-1991; ce projet était dirigé par Jules Duchastel alors qu'il était directeur du Centre d'ATO de l'UQAM. La version de la BDL utilisée jusqu'à maintenant ne contient que les catégories majeures sans trait de genre, nombre ou temps et personne. La nouvelle version de la BDL contient ces traits ainsi que le lemme des formes fléchies. Il serait sans doute intéressant, pour une expérimentation future, de vérifier la prédictivité de ces nouvelles variables sur la classe d'appartenance des textes.

grammaticale hors contexte d'environ un demi million de formes lexicales. Quant à la seconde liste, elle a été développée à l'intérieur du projet; il s'agit d'un dictionnaire de mots familiers aux élèves de sixième année du primaire constitué à partir des formes lexicales rencontrées dans le corpus utilisé pour ce projet.

Le dictionnaire des mots connus par les élèves de sixième année a été constitué en faisant valider le lexique de l'ensemble du corpus par des enseignants de sixième année. Le corpus s'étant enrichi au cours des années, cette validation a dû être faite deux fois pour tenir compte des nouveaux mots. Dans chacun des cas, la validation a été effectuée par un groupe de cinq enseignants d'expérience provenant de régions différentes du Québec et oeuvrant dans des milieux sociaux différents. On a accepté comme connus les lexèmes jugés familiers par au moins quatre personnes. La consigne donnée aux enseignants demandait de considérer connu un mot qu'au moins les trois quarts des élèves connaissent à l'oral. Dans certains cas, des vérifications ont été faites auprès des élèves eux-mêmes, afin de vérifier leur connaissance de certains mots.

Notons que la validation des mots a été effectuée sur les formes fléchies des lexèmes, c'est-à-dire dans la forme où ils se présentent dans le texte. Par la suite, nous avons élaboré des dispositifs de fléchissement permettant d'ajouter les flexions régulières des mots connus. En ce qui concerne les verbes, le problème de l'extension de la couverture est plus délicat. Il n'est pas question évidemment d'accepter toutes les formes conjuguées des verbes. Finalement, nous avons convenu de conjuguer les verbes aux temps simples selon les prescriptions du programme de sixième année en écriture[4]. Nous établissons présentement les critères pour sélectionner les verbes à conjuguer à partir des formes déjà authentifiées comme connues. Par exemple, si seulement le participe passé a été reconnu, est-ce que l'on peut conclure automatiquement que le verbe est connu dans ses formes conjuguées?

3.2.2 Repérage des noms propres

Même si notre corpus est volumineux, la liste des mots soumis à l'évaluation du milieu n'est pas exhaustive. On verra plus loin que l'on peut compléter la liste par un dictionnaire personnel. Par ailleurs, les noms propres, qu'il n'est pas possible évidemment de rassembler dans une base de données lexicale,

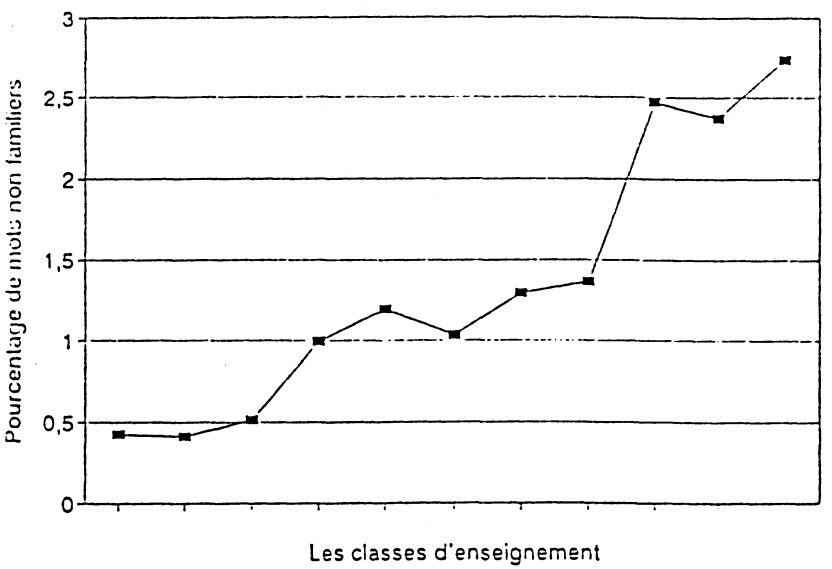
[4] Le programme du MEQ prescrit «la formation des temps simples (radical et terminaison) à l'indicatif présent, à l'imparfait, au futur simple, au conditionnel présent, à l'infinitif présent, au participe passé, au subjonctif présent (quand la forme est différente de celle de l'indicatif présent) et au passé simple (3e personne du singulier et du pluriel)».

peuvent être considérés connus en raison de leur contexte d'utilisation. Pour faciliter l'identification des noms propres, un scénario a été bâti dans le but de dresser pour chaque texte une liste de noms propres potentiels. Cette liste peut, au choix de l'utilisateur, être présentée pour fins de validation.

Cette liste de noms propres se limite aux mots qui ne font pas déjà partie de la liste des mots connus à titre de nom commun par exemple. On exclut aussi les articles, les pronoms, les adverbes, les prépositions et les conjonctions. Finalement, on dénombre l'utilisation des lexèmes de la liste restante en comptant les graphies qui débutent par une majuscule et en notant les cas où cette utilisation ne fait pas suite à une ponctuation forte telle le point. Dans ce dernier cas, il y a de bonnes chances que le mot soit un nom propre.

Les noms propres ainsi identifiés sont ajoutés à la liste des mots connus. De même, on y ajoute les nombres, ponctuations et séparateurs. La figure 1 présente la moyenne des pourcentages de mots inconnus par classe d'enseignement. Bien que la courbe indique un fléchissement au premier cycle du secondaire, on constate que le pourcentage de mots inconnus croît avec la classe d'enseignement.

Figure 1
Pourcentage moyen de mots inconnus par classe d'enseignement



3.2.3 Désambiguïsation des verbes conjugués

Il nous est aussi apparu que le décompte du nombre relatif de propositions dans le texte était susceptible de nous donner des indications sur sa complexité. Le dépistage en contexte des verbes conjugués s'est donc avéré nécessaire. Nous avons déjà eu l'occasion de présenter de façon détaillée notre stratégie de désambiguïsation des verbes, cf. Daoust & Dupuis (1996, 1994). Nous nous contenterons donc ici d'un rappel.

À l'aide de SATO, il est possible de décrire, sous forme de patrons de fouille, les contextes désambiguïsants. Du même coup, on peut associer à ces patrons des actions de désambiguïsation catégorielle. Nous adoptons donc une stratégie de 'grammaires locales', cf. Silberztein (1989). La solution développée ici comporte deux étapes incorporées dans une seule procédure. On a d'abord l'élagage ou la suppression des catégories grammaticales 'indésirables'. On a ensuite l'ajout d'une propriété permettant de visualiser le résultat de la règle et de retracer le contexte de son application. Avec le logiciel, on a construit un dispositif permettant d'évaluer la productivité des règles et de voir par quels moyens rendre la grammaire d'émondage plus efficace. On peut comparer toutes les applications réussies de telle ou telle règle ou, à l'inverse, tous les contextes où aucune règle de désambiguïsation ne s'est appliquée. Cela permet d'examiner tous les cas semblables disséminés dans un texte et de rectifier les règles déjà existantes. Cet examen peut aussi conduire à l'ajout de nouvelles règles pour augmenter l'efficacité du système. Ce dispositif, utilisé en phase de validation, a été supprimé dans le module final.

Voici un exemple de règle et sa traduction dans une commande au logiciel.

Règle: Une forme, qui peut être soit un nom soit un verbe conjugué, n'est pas un verbe conjugué si elle est précédée d'une forme qui est strictement une préposition. La préposition peut être suivie facultativement d'un article ou d'un déterminant et d'adjectifs non ambigus.

Exemple: La femelle construit habituellement son nid sous un tas de larges branches...

Commande: contexte appliquer **
$*gramr==prép*. **
$*gramr=(art$,dét$)*-*. **
$*gramr=v_conj*gramr=nomc*syntaxe:-v_conj*&*règle:+d1

Explication: La commande 'contexte appeler' effectue le repérage des contextes qui satisfont aux contraintes définies par les filtres dont la définition

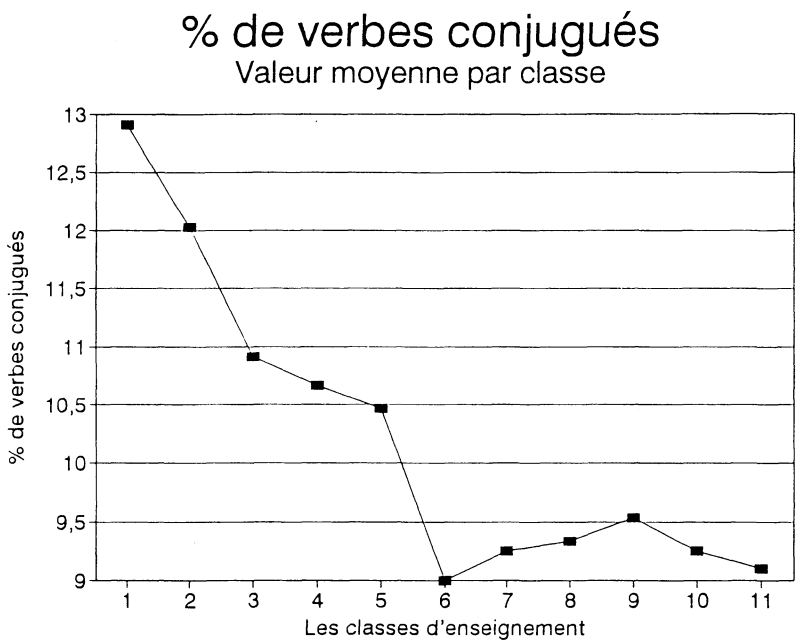
suit. Le '$' indique qu'il n'y a aucune contrainte sur les caractères du mot; 'gramr' contient la catégorie grammaticale hors-contexte; '==prép' signifie que le mot doit être une préposition non-ambiguë; '*.' est un opérateur de proximité qui indique que le filtre suivant est immédiatement adjacent. Le deuxième filtre accepte tout mot qui est un article ou un déterminant; '*-' indique que la position peut être vide. Le dernier filtre indique que le mot possède la double catégorie verbe conjugué et nom commun; la propriété 'syntaxe' est la projection de 'gramr' sur les occurrences; l'opérateur ':-' indique que l'on veut enlever la catégorie 'nomc' à la propriété 'syntaxe'; l'opérateur '*&' indique que la catégorie verbale doit être le pôle de la concordance; finalement, '*règle:+d1' indique que l'étiquette 'd1' qui identifie la règle doit être apposée sur le lexème désambiguïsé.

La couverture des règles de désambiguïsation pourrait être élargie puisque nous nous sommes limités aux règles les plus productives. On pourrait aussi faire appel à des algorithmes probabilistes, ce qui est de plus en plus la tendance comme en témoigne un numéro récent de la revue de l'ATALA, cf. Habert (1995): chaînes de Markov, N-gram etc. Cependant, comme les textes à analyser sont courts, le nombre de validations manuelles à effectuer, le cas échéant, ne constitue pas un handicap majeur.

La figure 2 à la page suivante permet de visualiser la moyenne des pourcentages par texte de verbes conjugués selon les années d'enseignement. Il apparaît donc que le nombre relatif de verbes conjugués dans un texte est un indice de facilité. Cela correspond probablement à un grand nombre de phrases courtes à construction simple du type sujet-verbe-complément.

Une première série d'analyses statistiques nous a révélé que la difficulté croissante des textes en fonction de la classe d'enseignement semblait subir un fléchissement au niveau du secondaire. En particulier, nos variables ne semblaient pas suffisantes pour caractériser la fin du secondaire. C'est alors que nous avons entrepris, cf. Daoust (1993) de vérifier si certains marqueurs pouvaient rendre compte de structures argumentaires davantage présentes dans les textes des classes les plus avancées. Nous nous sommes donc intéressés aux adverbes, aux prépositions et aux conjonctions. Comme plusieurs items de ces catégories prennent la forme de locutions, nous avons dû d'abord développer une procédure de reconnaissance de ces locutions. Faisant l'économie d'une analyse grammaticale complète des phrases, nous avons adopté encore une fois une stratégie de grammaires locales. Ce faisant, nous avons choisi d'écarter des expressions trop ambiguës comme *en fait*.

Figure 2
Pourcentage moyen de verbes conjugués par classe d'enseignement



3.2.4 Dépistage des locutions grammaticales

Les lexèmes simples et les locutions ainsi dépistées ont ensuite fait l'objet de diverses manipulations statistiques. Ces analyses nous ont conduit à définir une nouvelle variable basée sur le décompte des lexèmes suivants: *alors que, à l'instant, à présent, au-delà, au-dessous, au-dessus, au-devant, certes, dont, guère, parmi, particulièrement, séparément, toutefois, d'ailleurs, en effet, en vertu de, le long de, tel, telle, telles, tels, vous.* La présence du pronom *vous* peut apparaître étonnante à première vue. Le mot lui-même n'est pas difficile mais il annonce probalement une forme conjuguée plus difficile. On verra plus tard dans la présentation de l'indice que le pronom *tu* a aussi été retenu par les analyses statistiques. L'usage relatif du *tu* diminue avec la classe d'enseignement. À l'inverse du *vous*, il s'agit donc d'un indice de facilité. L'usage contrasté du *tu* et du *vous* est sans doute relié à des considérations psycholinguistiques, le tutoiement étant souvent relié à un univers familier plus caractéristique des premières classes du primaire. Les figures 3 et 4 confirment que le pourcentage d'utilisation de ces mots augmente avec la classe d'enseignement alors que le pourcentage d'utilisation du *tu* diminue.

Figure 3
Pourcentage moyen de certaines formes fonctionnelles

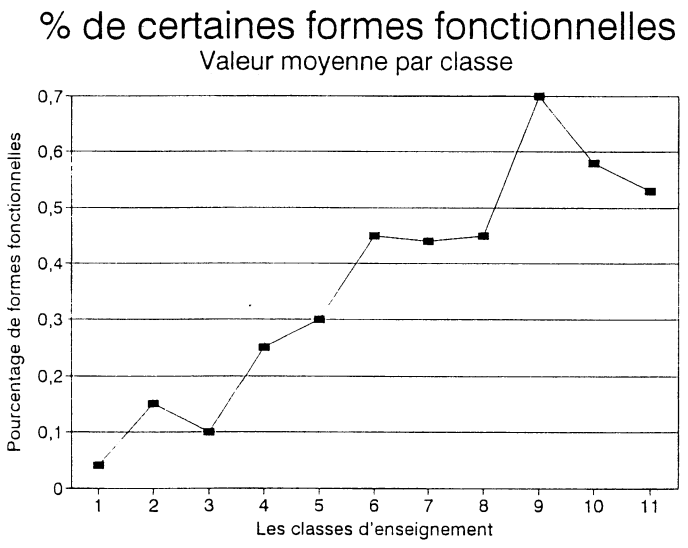
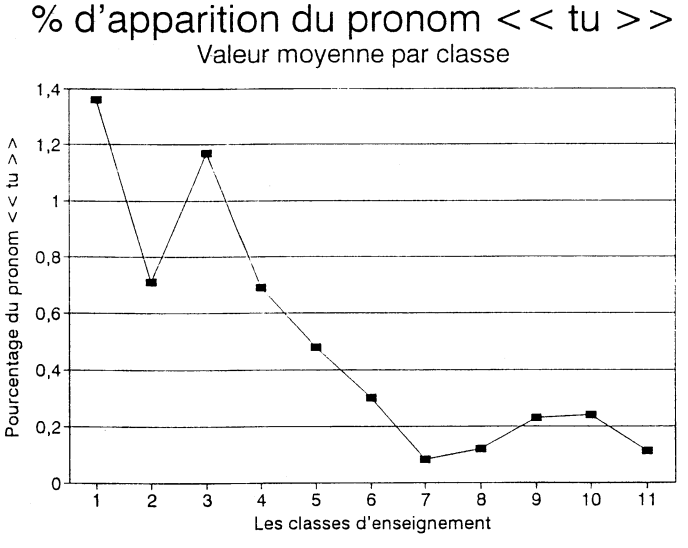


Figure 4
Pourcentage moyen du pronom *tu* par classe d'enseignement

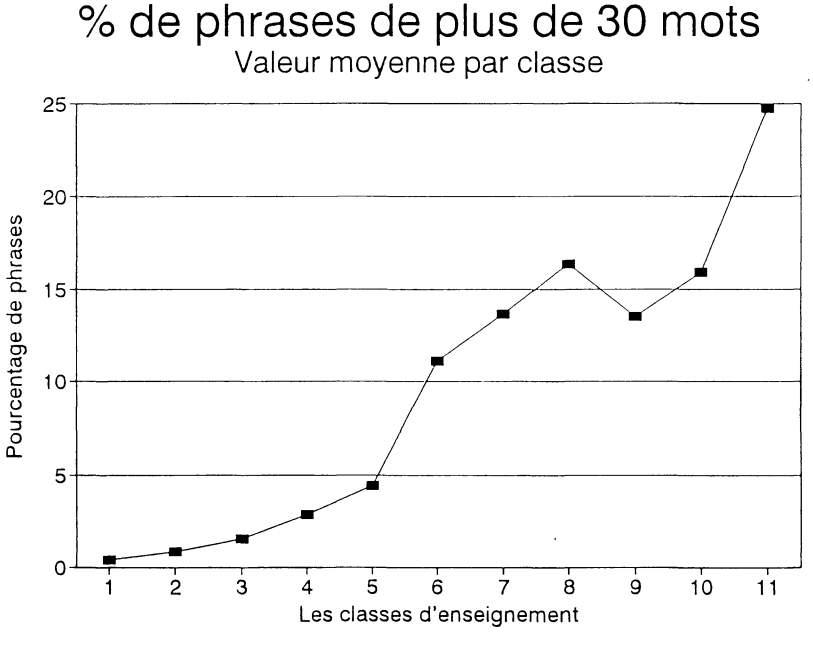


3.2.5 Repérage des phrases complexes

Le dernier élément de notre dispositif linguistique consiste à repérer certaines constructions de phrase susceptibles de contenir des éléments de complexité. Nous nous sommes contentés ici de repérer des phrases qui contenaient des éléments déjà dépistés par les opérations précédentes. De multiples suggestions nous ont été faites par le comité d'appui. Beaucoup n'ont pas résisté à l'analyse statistique qui, par nature, ne retient pas les situations trop peu fréquentes. Dans le rapport qualitatif produit par le logiciel, nous avons quand même utilisé des diagnostics que l'analyse statistique n'a pas retenu. De façon intuitive à tout le moins, et de l'avis des membres du comité d'appui, ces diagnostics peuvent suggérer des reformulations. C'est le cas par exemple des phrases qui possèdent plus de trois verbes conjugués ou celles qui possèdent des *qui, que, dont* etc.

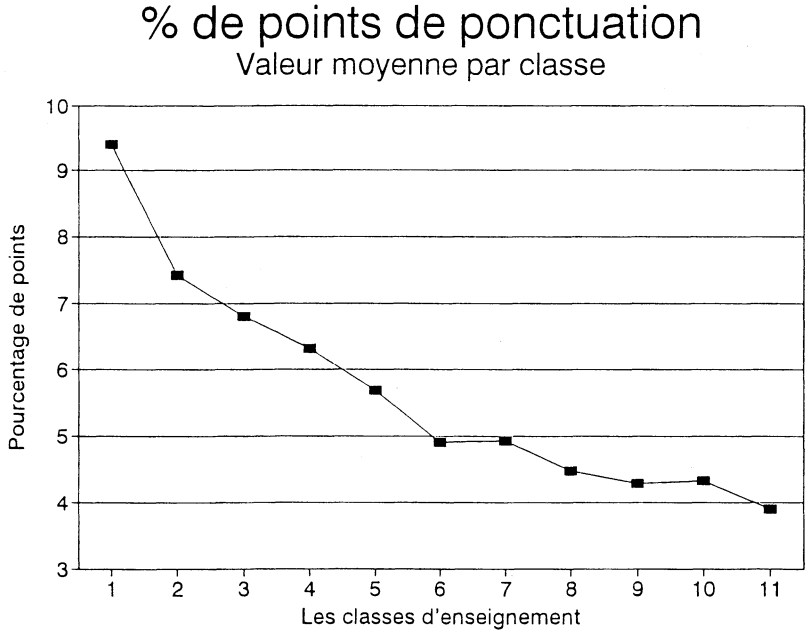
Comme on le verra en 3.3.2, la première variable en importance de l'indice SATO-CALIBRAGE est un indice simple de complexité de phrase, à savoir le pourcentage de phrases de plus de 30 mots (ponctuation incluse). Plusieurs seuils de longueur ont été utilisés mais c'est ce seuil de 30 mots que l'analyse statistique a fait ressortir (figure 5).

Figure 5
Pourcentage moyen de phrases de plus de 30 mots par classe d'enseignement



Le deuxième variable la plus importante de l'indice vient renforcer ce critère concernant la longueur des phrases. Il s'agit du pourcentage de points. Plus il y a de points dans le texte et plus il est facile. En première année, on a, en moyenne, presqu'un point à tous les 10 mots (ponctuations incluses!). En cinquième année du secondaire, on a un point à tous les 25 mots en moyenne (figure 6).

Figure 6
Pourcentage moyen de points par classe d'enseignement



Parmi les autres variables retenues par l'analyse statistique, plusieurs rejoignent l'intuition. Le nombre de phrases croît avec la classe d'enseignement mais avec une retombée étonnante en cinquième année du secondaire. Le point d'exclamation est un élément facilitant caractéristique des trois premières années du primaire. Les pronoms relatifs sont presqu'absents en première année. Leur nombre croît ensuite régulièrement mais connaît une baisse significative au premier cycle du secondaire. Ce même phénomène de fléchissement se produit pour d'autres variables, dont le pourcentage de mots de 9 lettres et plus qui, autrement, croît avec la classe d'enseignement.

Deux variables ont été retenues alors que l'on ne s'y attendait pas. Il s'agit des pourcentages d'utilisation des lexèmes *en* et *l'* respectivement. Même si l'importance statistique de ces deux variables n'est pas très grande (cf.tableau 2), elles apparaissent toutes deux comme des éléments de complexité. Il faudrait donc vérifier les contextes pour voir si leur utilisation plus fréquente dans les classes avancées correspond à des constructions de phrases déterminées. Le graphique de *en* semble indiquer une certaine stabilité de son usage à partir de la cinquième année. C'est donc surtout la faible utilisation du *en* dans les premières du primaire qui est digne d'attention (figure 7). Le graphique du *l'* nous laisse plus perplexe. Il faudra sans doute y revenir (figure 8).

3.3 Le dispositif mathématique

L'analyse mathématique a deux objectifs. Il s'agit d'abord d'évaluer l'ampleur et la pertinence de la variation des indices. Il s'agit ensuite de voir comment les divers indices partiels peuvent, en se combinant, produire des indices complexes susceptibles de révéler des régularités appréhendées ou insoupçonnées.

Nous désignons, par dispositif mathématique, l'ensemble des méthodes quantitatives utilisées pour interpréter les indices fournis par SATO. Ces méthodes mathématiques sont utilisées à deux fins. D'abord, on s'en sert pour déterminer les variables qui varient de façon significative en fonction de la classe d'enseignement. Ainsi, on peut confirmer ou infirmer des hypothèses concernant divers fonctionnements discursifs. Ensuite, on s'en sert pour combiner les indices primitifs significatifs afin de construire des fonctions aptes à prédire le rattachement à une classe d'enseignement.

Dans notre projet, nous avons fait appel à quatre types de modèles mathématiques. D'abord, puisque nous visons à trouver des indices permettant de distinguer les textes selon leur rattachement à des classes d'enseignement, nous avons recours à des tests d'hypothèses (test du Chi-2 en particulier) pour réaliser une première sélection des indices.

En ce qui concerne la constitution des indices basés sur les termes fonctionnels, nous avons voulu réduire le nombre de variables. Pour ce faire, nous avons utilisé deux techniques. Dans la première, nous avons soumis les termes fonctionnels retenus à l'analyse discriminante et avons conservé les termes gardés par l'analyse. Dans la deuxième technique, nous avons d'abord soumis l'ensemble des termes retenus à un algorithme de classement destiné à grouper ceux qui ont des distributions similaires par rapport aux classes d'enseignement, cf. Cucumel (1993), pour une présentation de méthodes de classification.

Figure 7
Pourcentage moyen du mot *en* par classe d'enseignement

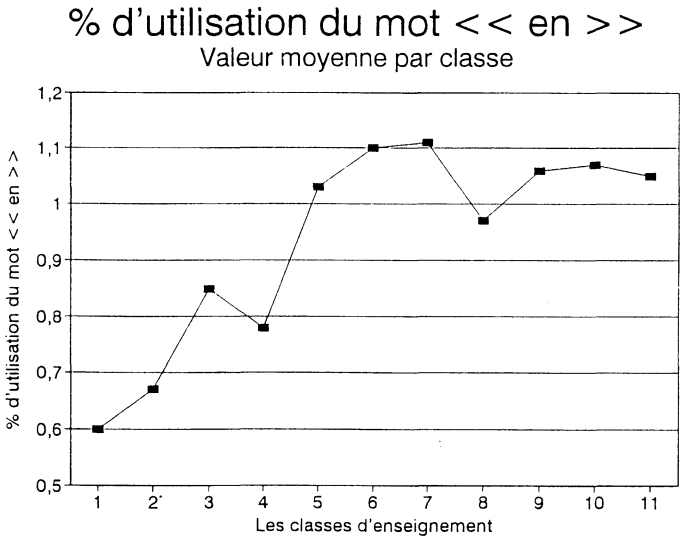
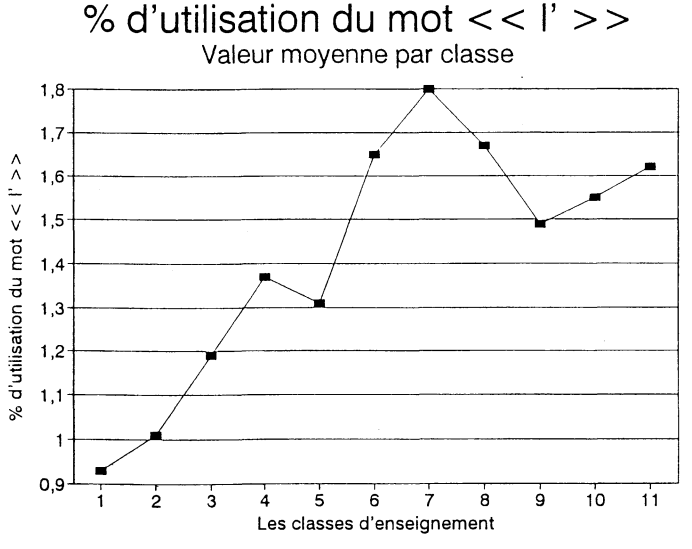


Figure 8
Pourcentage moyen du mot *l'* par classe d'enseignement



L'interprétation des groupes ainsi constitués a permis d'éliminer ceux dont le comportement semblait atypique. Elle a aussi permis de garder les autres groupes sous la forme d'indices composites.

Finalement, nous avons élaboré des fonctions prédictives permettant de classer un texte par rapport à une classe d'enseignement. Pour cela, nous avons utilisé les régressions simples et multiples et l'analyse discriminante.

Au terme de ces différentes analyses, il a été possible de définir un indice de difficulté, cf. Laroche (1993), basé sur la réalité du système scolaire québécois. Nous l'avons appelé l'indice SATO-CALIBRAGE.

3.3.1 Élaboration d'un indice

Pour être en mesure de proposer un indice susceptible d'indiquer le niveau de complexité d'un texte, les travaux statistiques sur le corpus se sont déroulés en deux temps. Au préalable, un ensemble de renseignements quantifiés a été produit à l'aide du logiciel SATO. Une première série d'analyses a consisté à produire des compilations univariées: statistiques descriptives, régressions simples, corrélations de Pearson. Dans un deuxième temps, des analyses multivariées ont permis de déterminer les sous-ensembles de variables qui peuvent le mieux expliquer le rattachement d'un texte à une classe d'enseignement.

Rappelons notre objectif. Il s'agit, sur la base des 'portraits quantifiés' de chacun des textes du corpus disponible, d'établir un lien entre les variables (indices de difficulté-facilité) et la classe d'enseignement auquel est destiné le texte. Pour ce faire, la régression simple et la corrélation de Pearson ont permis de mesurer l'importance du lien qui s'établit entre deux séries de valeurs. La première série de valeurs représente la classe d'enseignement. Il s'agit d'un nombre entre 1 et 11 correspondant aux six années du primaire et aux cinq années du secondaire. La deuxième série de valeurs représente les résultats calculés à partir du texte pour chacune des variables disponibles.

Les résultats obtenus à la suite des analyses univariées ont permis de retenir les variables les plus fortement liées à la classe d'enseignement. Nous avons pu constater une assez grande linéarité des résultats obtenus au regard de la classe d'enseignement, principalement pour les textes utilisés dans les classes du primaire. Plus de 120 variables ont été utilisées pour réaliser les travaux de cette première étape de nos analyses statistiques. Après ces compilations, 45 variables furent identifiées comme étant assez fortement reliées à la classe d'enseignement.

La seconde étape de l'analyse consiste à examiner le comportement de l'ensemble des variables retenues auparavant afin de déterminer les sous-ensembles qui peuvent le mieux expliquer le rattachement d'un texte à la classe d'enseignement. On sait que plusieurs variables mesurent des aspects semblables de la complexité d'un texte; il s'agit de diminuer cette redondance.

Des analyses factorielles ont permis de regrouper les variables et de les situer par rapport au rattachement à la classe d'enseignement. Un 'facteur' principal a ainsi pu être identifié. Il a été possible de situer certaines variables sur un axe décrivant leur plus ou moins grande 'complexité' mesurée par leur proximité à la classe d'enseignement. Cette phase de l'analyse a permis de retenir une trentaine de variables susceptibles de mieux décrire la lisibilité d'un texte.

Comme nous l'indiquions plus haut, parmi les variables disponibles pour ces analyses, plusieurs sont fortement corrélées entre elles, indiquant qu'il y a redondance. Des analyses réalisées à l'aide de la régression multiple tentent justement de diminuer ce phénomène en déterminant le jeu de caractéristiques le plus relié à la complexité des textes mesurée par leur rattachement à la classe d'enseignement. Les résultats obtenus par ces analyses ont rendu possible la fabrication d'un indice de calibrage.

3.3.2 Description de l'indice de calibrage

L'indice SATO-CALIBRAGE existe sous deux versions. La première, et la plus performante, tient compte de la longueur des textes. Plus un texte est long et plus il est susceptible d'appartenir à une classe avancée. Un deuxième indice a aussi été constitué sur la base d'une exclusion volontaire des variables tenant compte de la longueur des textes. Cet indice est moins performant mais est plus susceptible d'être utilisé sur des textes d'une autre nature, tels des romans ou autres textes longs.

Le tableau 2 donne la liste des variables constituant l'indice sensible à la longueur du texte. Cette liste a été produite par un algorithme de régression multiple qui permet d'éliminer les variables redondantes en ne gardant que celles qui sont les plus explicatives de la variance. Les 14 variables conservées permettent d'expliquer 74.2% de la variance. La variance mesure la dispersion de la variable 'classe d'enseignement' autour de sa moyenne. Lorsque cette dispersion est calculée autour de la droite de régression (l'indice SATO-CALIBRAGE), elle diminue des trois quarts. C'est donc dire que l'on a pu 'expliquer' ou prédire la classe d'enseignement avec une efficacité de près de 75% en utilisant les mesures produites par l'analyse des textes.

Le tableau 2 présente l'analyse de la régression multiple hiérarchique sur les variables prédictives de la classe d'enseignement. L'analyse de la régression multiple a été incluse dans cette étude sur la lisibilité afin de vérifier s'il est possible d'obtenir une équation de prédiction qui permettrait de décrire la relation linéaire entre des variables indépendantes et le rattachement des textes à la classe d'enseignement. La valeur de la régression multiple à la seizième étape du calcul est de 0,742. Cela signifie que près de 75% de la variance des valeurs indiquant la classe de rattachement des textes est expliquée par une combinaison linéaire des quatorze variables faisant partie de l'équation de régression.

Tableau 2
Analyse de régression sur les variables prédictives de la classe d'enseignement

Variable	Variance expliquée	Description de la variable
v1	30,4	% de phrases de plus de 30 mots
v2	44,4	% de points (.)
v3	57,0	% de mots inconnus (non familiers)
	67,7	Nombre total de mots
v4	69,1	% de formes fonctionnelles difficiles + vous
v5	70,0	% de tu
v6	70,8	Nombre de phrases
–	70,7	RETRAIT de la variable 'Nombre total de mots'
v7	72,0	% de points d'exclamation (!)
v8	72,6	% de pronoms relatifs
v9	73,2	% de mots de 9 lettres et plus
v10	73,5	% de 'en'
v11	73,8	% de 'l''
v12	74,0	% de verbes conjugués
v13	74,1	% d'adjectifs
v14	74,2	% de phrases contenant plusieurs mots non familiers

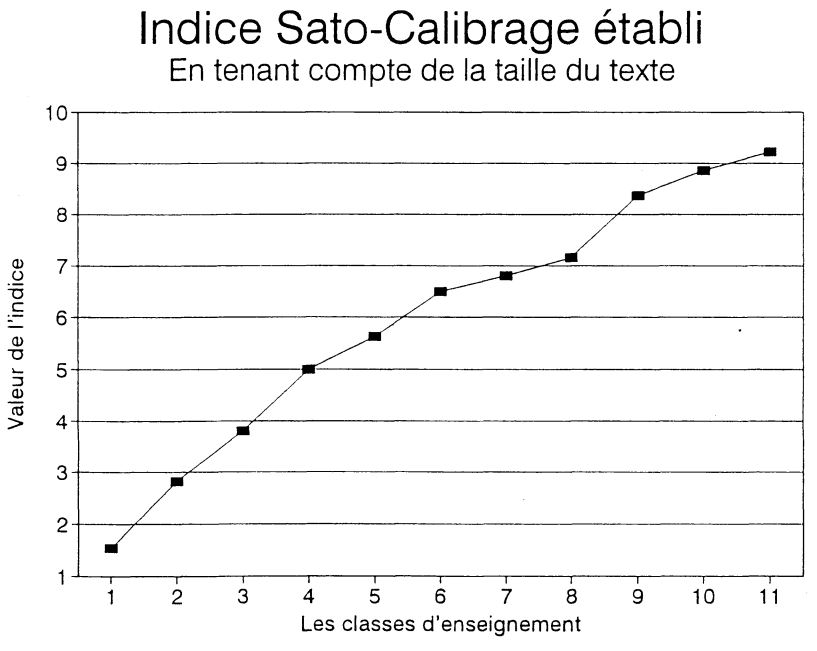
L'équation de régression construite à partir de ces variables est la suivante:

$$3.613 + 0.054\mathbf{v1} - 0.245\mathbf{v2} + 0.781\mathbf{v3} + 0.562\mathbf{v4} - 0.196\mathbf{v5} + 0.014\mathbf{v6} - 0.228\mathbf{v7} + 0.340\mathbf{v8} + 0.037\mathbf{v9} + 0.306\mathbf{v10} + 0.224\mathbf{v11} - 0.081\mathbf{v12} + 0.048\mathbf{v13} - 0.018\mathbf{v14}$$

Règle générale, un facteur positif dans l'équation indique que la variable est un indice de difficulté alors qu'un facteur négatif est la marque d'un indice de facilité. L'application de cette équation sur un texte donné nous fournit un estimé de la classe d'enseignement auquel il devrait être destiné, compte tenu de l'analyse des caractéristiques de notre corpus de 679 textes.

La figure 9 représente la courbe de la moyenne par classe de l'indice SATO-CALIBRAGE en fonction du niveau d'enseignement.

Figure 9
Indice SATO-CALIBRAGE tenant compte de la taille du texte



On peut constater que l'indice est très corrélé avec le niveau d'enseignement. On voit aussi qu'il est généralement linéaire malgré un affaissement progressif de la courbe après la quatrième année. Le rythme d'augmentation de la difficulté des textes a donc tendance à être moins forte au fur et à mesure de l'avancement scolaire. On trouvera dans le Guide de l'usager, cf. Daoust, Laroche & Ouellet (1996), des graphiques sur la répartition par classe d'enseignement de chacune des variables constituant l'indice.

4. Description du prototype SATO-CALIBRAGE

Les divers traitements réalisés par l'application sont les suivants: 1) la 'génération' du texte par le module SATOGEN; 2) son analyse par le module SATOINT suivi de la production du rapport sommaire de calibrage; 3) et finalement, la production d'un rapport qualitatif.

Le prototype utilise la version courante du logiciel SATO. Ainsi, nous pouvons profiter des améliorations au produit au fur et à mesure de son développement. Un module d'interface a aussi été développé pour faciliter la configuration de l'outil et lancer les traitements.

Ayant choisi le ou les textes à calibrer, les analyses suivantes sont déclenchées.

1- **Génération**. Il y a tout d'abord 'génération' du texte afin d'en produire une version en format interne à SATO. Cette opération consiste à lire le texte linéaire en codes ASCII afin d'en produire une représentation en deux dimensions: lexique et occurrences.

2- **Analyse et rapport sommaire**. Le module SATOINT réalise les traitements suivants.

   a) Lecture des commandes de configuration.

   b) Catégorisation grammaticale.
      La catégorisation grammaticale s'opère par la consultation d'un dictionnaire SATO et par une analyse morphologique des formes particulières telles les nombres.

   c) Repérage des mots familiers.
      - Création de la propriété 'connu':
        - oui mots outils les plus simples, noms propres, nombres, etc.;
        - p6 primaire sixième année (liste validée);
        - p6a primaire sixième année (liste personnelle);
      - Consultation du dictionnaire des mots connus (6e année).
      - Consultation du dictionnaire spécialisé (disciplinaire) des mots connus.
      - Attribution de la valeur 'oui' aux noms propres, délimiteurs, ponctuation, nombres, articles...
      - Consultation du dictionnaire personnel des mots connus.

d) Dépistage des noms propres correspondant à des mots inconnus.
Le dépistage des noms propres a pour objectif principal de marquer ces mots comme connus. En effet, le contexte permet généralement au locuteur de comprendre qu'il s'agit d'un nom de personne, de lieu, etc. Dans cette application, il n'est pas nécessaire de valider les noms propres qui correspondent déjà à une forme réputée connue. Le dépistage des noms propres est réalisé en comptant le nombre de fois où un mot (excluant les articles, pronoms, etc.) est en majuscule, et le nombre de fois où le mot en majuscule n'est pas précédé d'une ponctuation forte. Une décision est alors suggérée.

e) Validation manuelle des candidats noms propres.
Si cette option a été choisie lors de la configuration du prototype, la liste des candidats noms propres sera présentée pour validation. On peut alors confirmer ou infirmer la décision de SATO et, si nécessaire, voir la forme en contexte.

f) Catégorisation en contexte des verbes conjugués.
SATO-CALIBRAGE fait le décompte du nombre de propositions dans le texte. On identifie la proposition par le verbe conjugué. Comme plusieurs formes lexicales possèdent une catégorie grammaticale ambiguë, par exemple des verbes qui sont homographes avec des noms, des adjectifs, etc., on doit identifier les 'vrais' verbes. Les étapes de l'algorithme de dépistage sont les suivantes:

- identification des locutions grammaticales figées;
- application de patrons de concordance avec catégorisation en contexte.

g) Validation manuelle des verbes encore ambigus.
Même si les règles de grammaires locales permettent de lever la majorité des ambiguïtés sur le verbe, il reste en général un certain nombre d'ambiguïtés. Si l'option de validation a été choisie lors de la configuration du prototype, les verbes encore ambigus sont présentés pour une catégorisation manuelle.

h) Identification des indices de complexité.
À partir du marquage déjà réalisé, il est possible de dépister diverses configurations susceptibles de représenter des difficultés.

Les contextes dépistés sont alors marqués par une annotation inscrite dans la propriété contextuelle DIAGNOSTIC. Voici la liste des diagnostics dépistés sur les phrases.

4 verbes: la phrase possède au moins quatre verbes conjugés;
31 mots et Plus: la phrase contient plus de 30 mots;
Prorel-Con: la phrase possède un mot qui, au dictionnaire, peut être un pronom relatif (*qui, que, dont,* etc.). La phrase est affichée même si, en contexte, le mot agit comme conjonction.
2 mots inconnus: la phrase contient au moins deux mots inconnus.

i) Production des données quantitatives utilisées pour le calcul de l'indice SATO-CALIBRAGE (rapport sommaire).

Le calcul de l'indice SATO-CALIBRAGE se fait par un programme externe à partir des données numériques produites par SATO.

3- **Rapport qualitatif**. Généralement, on fait suivre la production de l'indice d'un rapport qualitatif de calibrage. Ce rapport contient les éléments suivants:

- le lexique des noms, des adjectifs et des verbes apparaissant plus d'une fois;
- la longueur des mots, des phrases et des paragraphes et l'indice de Gunning[5];
- la répartition des lexèmes par rapport aux listes de mots connus;
- la liste des mots identifiés comme inconnus; l'utilisateur peut alors inscrire des lexèmes dans son dictionnaire personnel des mots connus;
- la liste des mots longs (9 lettres ou plus);
- la liste des phrases susceptibles de contenir des éléments de complexité (cf. étape h).

[5] L'indice GUNNING se calcule par la formule suivante: (longueur moyenne des phrases + % de mots de 9 lettres et plus) x 0,4.

Il est aussi possible de compléter l'analyse du texte en faisant appel à des scénarios spécialisés. En voici quelques exemples:

Lex_Tot	Lexique total trié par ordre alphabétique;
Lex_Det	Lexique des déterminants;
Lex_Lien	Lexique des mots de liaison;
Lex_Pp	Lexique des pronoms personnels;
Phr_Adj	Phrases contenant des adjectifs qualificatifs (hors contexte);
Phr_Ffd	Phrases contenant des formes fonctionnelles difficiles;
Phr_2Pp	Phrases contenant au moins deux pronoms personnels;
Phr_Prel	Phrases contenant des pronoms relatifs (hors contexte);
Phr_Conj	Phrases contenant des verbes conjugués;
Phr_Inc	Phrases contenant des mots inconnus;
Phr_9Let	Phrases contenant des mots de 9 lettres et plus;
Phr_Lon	Phrases dépassant une longueur fournie par l'utilisateur;
Phr_En	Phrases contenant *en*;
Phr_L	Phrases contenant *l'*;
Phr_Toi	Phrases contenant *toi*;
Phr_Tu	Phrases contenant *tu*;
Phr_Vous	Phrases contenant *vous*;
Phr_Excl	Phrases contenant *!*;
Phr_Mot	Phrases contenant un mot fourni par l'utilisateur;
Phr_Son	Phrases contenant une syllabe fournie par l'utilisateur;
Syn_Decr	Statistique descriptive de la propriété Syntaxe.

L'application SATO-CALIBRAGE étant 'ouverte', il est possible de la compléter et de la personnaliser à volonté. Ainsi, un utilisateur pourrait s'initier au logiciel SATO et ajouter une variété d'analyseurs pouvant fournir des avis spécialisés. Par exemple, on pourrait vouloir développer des analyses spécifiques pour identifier des difficultés qui correspondent à des populations dont la langue maternelle n'est pas le français.

5. Quelques illustrations

Afin de mieux comprendre l'utilisation qui peut être faite de SATO-CALIBRAGE, nous l'illustrerons à l'aide de quelques exemples.

5.1 Une application didactique

Une enseignante de sixième année a observé que plusieurs élèves de sa classe ont de la difficulté à comprendre les phrases longues et qu'ils ne savent

pas comment aborder ce problème. Elle a en main un texte qu'elle juge intéressant **Le travail des enfants** et qui, de surcroît, aborde un thème traité dans son cours de sciences humaines. Elle le soumet à l'analyse à l'aide de SATO-CALIBRAGE. L'outil lui donne un indice de 4,5. De prime abord, l'enseignante le juge facile pour ses élèves. En consultant les différents renseignements sur le texte, elle constate qu'il y a peu de mots inconnus mais que le texte renferme sept phrases longues, ce qui est l'objet de sa préoccupation. Elle examine ces phrases et constate qu'elle pourrait travailler les liens qu'établissent les conjonctions *et* et *ou*.

Après avoir abordé la compréhension globale du texte et proposé quelques tâches aux élèves, l'enseignante retient les phrases 1, 2, 3 et 6 (voir encadré) pour faire un travail systématique d'analyse et de compréhension; elle pourra aussi demander aux élèves de retrouver la phrase de base, notion abordée en écriture récemment. Par la suite, elle demande aux élèves de trouver dans d'autres textes des phrases longues qui comportent plusieurs *et* et *ou* et de vérifier s'ils les comprennent facilement. En outre, l'enseignante garde en réserve la phrase 7 pour une autre séance où elle abordera les compléments de phrase antéposés et les appositions, ou encore les différentes relations amenées par la préposition *pour*.

Tableau 3
Phrases longues trouvées par SATO-CALIBRAGE

1. Vers douze ans, une fille peut prendre la responsabilité de la maison et remplacer sa mère quand celle-ci est malade ou quand elle doit aller travailler aux champs.
2. Ils transportent le bois du hangar à la maison, allument et alimentent le feu dans les foyers et le poêle et transportent l'eau à l'étable et à la maison.
3. Le chef de famille doit gagner un salaire pour acheter la nourriture, les tissus ou les vêtements, le charbon ou le bois de chauffage et pour payer un loyer.
4. Le boeuf coûtait entre 5 et 10 cents la livre, le beurre, de 15 à 30 cents la livre, les oeufs, de 13 à 20 cents la douzaine.
5. Entre 1870 et 1930 environ, de nombreux enfants ont dû travailler dans des conditions très difficiles et malsaines, avant même d'avoir atteint l'âge de douze ans.
6. Entre sept et douze ans, on l'initiait donc aux travaux, aux techniques et aux responsabilités de la vie adulte pour qu'à douze ou quatorze ans il puisse fonder et entretenir une famille.
7. Enfin, avec l'arrivée des industries, surtout dans les grosses usines des villes, les patrons ont exploité les travailleurs, particulièrement les femmes et les enfants, pour augmenter leurs profits.

L'utilisation de SATO-CALIBRAGE peut donc servir directement aux enseignants pour les aider à préparer leur matériel didactique et leurs exercices.

5.2 Une application en évaluation

Un conseiller pédagogique doit préparer une épreuve sommative pour des élèves de troisième année. Il souhaite trouver un texte à caractère imaginaire qui ne soit pas trop difficile pour les élèves. Il rassemble cinq textes: trois contes et deux extraits de roman. Il soumet ces textes à SATO-CALIBRAGE et obtient les indices 2.0, 3.3, 5.2, 3.7 et 3.1. Dans un premier temps, il élimine le premier et le troisième texte, les jugeant trop facile pour l'un et trop difficile pour l'autre. En examinant de plus près les données qui concernent les trois autres textes, il constate que le quatrième texte comporte beaucoup de mots inconnus et que, pour cette raison, il causera sûrement des difficultés à plusieurs élèves. Quant aux deux autres textes, leur niveau de difficulté semble convenir pour élaborer une épreuve; ils comportent quelques phrases longues et quelques mots inconnus, ce qui semble raisonnable pour des élèves de troisième année à la fin de l'année. L'analyse du déroulement de l'histoire et l'intérêt des personnages le font opter pour le deuxième texte.

L'utilisation de SATO-CALIBRAGE peut donc s'avérer très utile pour donner des avis objectifs sur les textes utilisés dans les épreuves.

5.3 Une application en rédaction

Un auteur doit rédiger un texte à l'intention des élèves de sixième année. Il a soumis une première version de son texte à SATO-CALIBRAGE. La phrase suivante est ressortie comme étant potentiellement difficile parce qu'elle est très longue et qu'elle comporte plusieurs mots inconnus (*cohabitent, armadas, hordes, évoluant*).

> *«Déjà que les automobilistes cohabitent difficilement avec les armadas de bicyclettes dans notre centre-ville, il fallait que s'ajoutent des hordes de patineurs évoluant de façon aussi ordonnée qu'un nuage de poussière poussé par le vent.»*

L'auteur pourra donc envisager de faire deux phrases plus courtes, d'utiliser des synonymes plus familiers ou des périphrases qui feront comprendre les mots inconnus des élèves. Il pourra aussi établir différemment le lien entre cette phrase et la précédente.

L'utilisation de SATO-CALIBRAGE est donc susceptible de combler un vide important dans le domaine de l'édition scolaire. Les éditeurs, en effet, disposent d'outils et affectent du personnel pour la correction linguistique. Cependant les auteurs et les éditeurs sont beaucoup plus démunis quand il s'agit d'évaluer un texte du point de vue de la lisibilité par le public cible

auquel il est destiné. Il est aussi important qu'un logiciel ne se contente pas d'un indice chiffré. Il faut en effet que l'auteur puisse relire son texte en disposant d'indications pour en faciliter la réécriture.

6. Conclusion

Par rapport à notre objectif initial qui se limitait à fournir quelques éléments quantitatifs pour aider à évaluer le niveau de difficulté d'un texte, le prototype est allé au-delà de nos espérances. L'indice SATO-CALIBRAGE s'avère en effet très performant. Pourtant, il est vite apparu que la production d'un indice n'était pas le plus important. Ce que permet l'application et ses extensions possibles, c'est de repérer des textes possédant des caractéristiques spécifiques pour élaborer des instruments d'enseignement ou d'évaluation.

Le logiciel s'avère très utile pour porter un premier jugement sur le degré de difficulté d'un texte. Il permet aussi d'obtenir des données utiles pour planifier un apprentissage, ou pour élaborer un instrument d'évaluation. Il donne aussi des indications pour modifier un texte en fonction de ces finalités. Cependant, il ne peut se substituer à l'enseignant pour juger de la pertinence du texte en fonction des intérêts des élèves, de leurs connaissances sur le sujet et de leur degré d'habileté à lire. SATO-CALIBRAGE a des limites comme tout outil mais il peut apporter un soutien précieux aux enseignants préoccupés de l'apprentissage de la lecture chez leurs élèves.

L'appui d'un comité d'utilisateurs tout au long du développement du prototype nous a été d'un précieux secours. Les membres de ce comité d'appui ont contribué à la constitution du corpus. Ils ont aussi utilisé l'application dans leur travail quotidien. Leurs suggestions nous permettent de penser que l'application est maintenant mûre pour être utilisée à plus grande échelle.

En outre, avec la venue de banques de textes sur CD-ROM, l'utilisation d'un outil de calibrage peut nous donner un avis supplémentaire sur le degré relatif de difficulté des textes qui composent la banque.

Le développement de l'indice SATO-CALIBRAGE avait comme préoccupation l'analyse des textes soumis aux élèves du primaire et du secondaire. Toutefois, la méthodologie développée pourrait être utilisée pour valider un indice spécifique pour les textes qui s'adressent à un vaste public. Déjà des personnes oeuvrant dans les directions des communications de différents ministères s'intéressent à nos travaux, cf. Lortie & Parent (1993). Dans une perspective d'élargissement des publics, on pourrait aussi penser aux textes destinés aux personnes pour qui le français est une langue seconde, aux textes utilisés dans les différentes disciplines au cégep ou à l'université, etc.

Enfin, le corpus utilisé et le lexique des mots familiers constituent des acquis importants pour la poursuite de la recherche. Ils servent de points de référence et de comparaison pour la construction de corpus complémentaires destinés à divers publics. Le corpus de texte pourrait sans doute servir à tester d'autres variables linguistiques dont on pourrait croire qu'ils contribuent à la complexité relative des textes. Il est clair en effet que nous sommes loin d'avoir épuisé le sujet. Nous espérons surtout que la méthodologie que nous avons utilisée puisse contribuer à stimuler les efforts de recherche dans le domaine.

Références

CUCUMEL, G. (1993) «Classification par partition et classification hiérarchique: deux méthodes complémentaires», *Cahier de recherche*, n° 3, Centre ATO-CI, UQAM, p. 83-96.

DAOUST, F. (1993) «Le dispositif mathématique», *Cahier de recherche*, n° 3, Centre ATO-CI, UQAM, p. 75-96.

DAOUST, F. & F. DUPUIS (1994) «Le dépistage en contexte des verbes conjugués à l'aide du logiciel SATO», *Revue ICO Québec*, vol 6, n° 1-2, p. 106-113.

DAOUST, F. & F. DUPUIS (1996) «Un protocole pour la mise au point d'algorithmes de désambiguïsation catégorielle», in L. Émirkanian & L.H. Bouchard (réd.), *Traitement automatique du français, Les Cahiers scientifiques de l'ACFAS*, n° 86, p. 153-173.

DAOUST, F., L. LAROCHE, L. OUELLET & al. (1993) «Le projet SATO-CALIBRAGE», *Cahier de recherche*, n° 3, Centre ATO-CI, UQAM.

Daoust, F, L. LAROCHE & L. OUELLET (1996) *SATO-CALIBRAGE, Guide de l'usager, version 1.0*, Centre ATO, UQAM.

GÉLINAS-CHEBAT, C., C. PRÉFONTAINE, J. LECAVALIER & J.C. CHEBAT (1993) «Lisibilité - Intelligibilité de documents d'information», *Cahier de recherche*, n° 3, Centre ATO-CI, UQAM, p. 19-35.

HABERT, B. (1995) «Traitements probabilistes et corpus», *T.A.L.*, revue semestrielle de l'ATALA, vol. 36, n° 1-2.

LAROCHE, L. (1990) «Calibrage des textes et lisibilité»,*Revue ICO Québec*, vol. 2, n° 3, p. 114-117.

LAROCHE, L. (1993) «Analyses statistiques pour la constitution d'un indice SATO-CALIBRAGE», *Cahier de recherche*, n° 3, Centre ATO-CI, UQAM, p. 97-139.

LORTIE, R. & R. PARENT (1993) «Place de la lisibilité en gestion de l'information textuelle», *Cahier de recherche* n° 3, Centre ATO-CI, UQAM, p. 37-38.

SILBERZTEIN, M. (1989) *Dictionnaire électronique et reconnaissance lexicale*, thèse de doctorat en informatique, LADL, Université Paris 7.

POST-SCRIPTUM

Pour illustrer notre propos, nous avons soumis le présent article à SATO-CALIBRAGE. L'indice SATO-CALIBRAGE sensible à la longueur du texte nous donne une valeur de 16.05, soit 5 ans au delà de notre corpus de référence qui ne va pas au delà de la onzième année. Par ailleurs, l'indice SATO-CALIBRAGE insensible à la longueur du texte nous donne une valeur de 10.41, ce qui ferait de cet article, propos mis-à-part (!), un texte accessible linguistiquement aux élèves de la fin du secondaire. Cette différence très nette entre les deux indices indique que la longueur du texte dépasse de beaucoup la longueur moyenne des textes que l'on donne à lire en classe aux élèves de la fin du secondaire.

Examinons plus en détails le rapport sommaire de l'indice sensible à la longueur. Rappelons que le rapport sommaire nous donne, outre l'indice, la valeur de chaque variable composant l'indice. Aussi, pour chacune des variables, on trouve en référence la classe d'enseignement dont la moyenne se rapproche le plus de la valeur observée sur notre texte. D'abord, on constate que le nombre de phrases longues n'est pas élevé. En fait, il se compare au corpus des textes de sixième année. On constate aussi que le pourcentage de points est comparable au pourcentage moyen de la huitième année, ce qui indique qu'en moyenne la longueur des phrases est comparable à celle des textes de huitième année.

Comme l'indique le rapport de l'indice insensible à la longueur, c'est d'abord le nombre élevé de mots inconnus qui fait grimper la valeur de l'indice sur notre article. La moyenne de référence pour la onzième année est de 2.73 %. Hors, nous avons 5.32 %! Il en est de même du «Nombre de mots» longs dont la moyenne pour la onzième année est de 10.1 % alors que nous avons 17 % pour cet article.

La difficulté principale du texte provient donc du vocabulaire employé qui comprend beaucoup de termes techniques qui ne font pas partie du vocabulaire courant. C'est là un des traits significatifs d'un article scientifique. Aussi, l'indice sensible à la longueur est ici abusivement élevé en raison même de la longueur d'un article de revue qui dépasse nettement celle des textes que l'on fait lire en classe. On notera que la valeur comparative de la variable «nombre de phrases» nous place au niveau de la dixième année. Cela s'explique par le fait que les textes qui comportent le plus grand nombre de phrases se retrouvent en dizième année plutôt qu'en onzième.

Est-il nécessaire de préciser que nous nous sommes servis de l'indice SATO-CALIBRAGE pour corriger notre article?

Fichier SATORQL

Indice SATO-CALIBRAGE (sensible à la longueur du texte --août 94)

Variable	Valeur	Poids	Classe
% de phrases de plus de 30 mots	12.06	0.054	6
% de .	4.53	-0.245	8
% de mots non familiers (inconnus)	5.32	0.781	11
% de formes fonctionnelles difficiles + vous	0.66	0.562	9
% de tu	0.07	-0.196	7
Nombre de phrases	481.00	0.014	10
% de !	0.02	-0.228	6
% de pronoms relatifs	1.65	0.340	2
% de mots de 9 lettres ou plus	17.00	0.037	11
% de en	1.04	0.306	11
% de l'	1.87	0.224	7
% de verbes conjugués	4.49	-0.081	6
% de d'adjectifs	11.24	0.048	6
% de phrases contenant plusieurs mots non familiers	24.95	-0.018	11

Valeur de l'indice : 16.05

Indice SATO-CALIBRAGE (insensible à la longueur du texte --août 94)

Variable	Valeur	Poids	Classe
% de .	4.53	-0.293	8
% de mots non familiers (inconnus)	5.32	0.953	11
% de formes fonctionnelles difficiles + vous	0.66	0.804	9
% de !	0.02	-0.225	6
% de tu	0.07	-0.240	7
% de pronoms relatifs	1.65	0.460	2
% de phrases de plus de 30 mots	12.06	0.030	6
% de mots de 9 lettres ou plus	17.00	0.066	11
% de en	1.04	0.301	11
% de l'	1.87	0.226	7
% de ,	3.59	0.107	1
% de toi	0.01	-0.702	7
% de phrases contenant plusieurs mots non familiers	24.95	-0.031	11
Nombre moyen de mots par phrase	16.30	0.007	6

Valeur de l'indice : 10.41