ATO·CI

centre de recherche en cognition et information

Le projet SATO-CALIBRAGE

François Daoust, Léo Laroche, Lise Ouellet

Claire Gélinas-Chebat, Clémence Préfontaine, Jacques Lecavalier, Jean-Charles Chebat, René Lortie, Richard Parent, Fernande Dupuis, Guy Cucumel, Georges Pelletier, Pierre Achim.

Centre de Recherche en Cognition et Information ***ATO-CI***

Université du Québec à Montréal

TABLE DES MATIÈRES

Présentation

Ce numéro des Cahiers de recherche du Centre ATO-CI est consacré à un projet qui est en cours déjà depuis plusieurs années en collaboration avec le ministère de l'Éducation du Québec. Ce projet, appelé SATO-CALIBRAGE, vise à utiliser le logiciel SATO[1] pour mesurer la difficulté de lecture d'un texte par un public cible. Le public qui nous a servi de point de référence jusqu'à ce jour est constitué des élèves de sixième année du primaire. Cependant, les instruments mis en place tiennent compte d'une échelle graduée allant du début du primaire à la fin du secondaire. Voilà pourquoi nous parlons de calibrage de textes, c'est-à-dire de classement des textes dans l'un ou l'autre niveau scolaire en fonction de divers indices de facilité ou de difficulté de lecture.

L'intérêt de ce projet dépasse cependant cet objectif immédiat. Il illustre en effet de façon élaborée le développement d'une **méthodologie expérimentale** en analyse de texte. SATO-CALIBRAGE est aussi un bon exemple de **recherche-action**. Tout au long de son déroulement, le projet a impliqué à la fois des chercheurs universitaires de plusieurs disciplines (informatique, mathématiques, linguistique, pédagogie) et des intervenants du milieu de l'éducation (professionnels du Ministère et d'organismes de soutien, conseillers pédagogiques et enseignants). Du côté de la recherche, ce projet a servi de banc d'essai et de stimulant pour les développements informatiques, linguistiques et méthodologiques.

La publication de ce Cahier ne marque pas le point final du projet. Il vise plutôt à dresser le portrait de son évolution jusqu'à ce jour et à indiquer les perspectives de son déroulement futur. Le Cahier comprend cinq parties.

Dans la première partie, traitant de la **problématique**, Lise Ouellet du ministère de l'Éducation du Québec, nous présente la problématique du calibrage des textes dans le milieu de l'éducation. Ensuite, elle dresse une chronologie détaillée du projet.

Un deuxième article situe les questions de lisibilité des textes, questions qui nous ont surtout intéressés dans le projet SATO-CALIBRAGE, par rapport à la problématique plus générale de l'intelligibité des textes. Cet article est signé par Claire Gélinas-Chébat et Clémence Préfontaine, toutes deux professeures au département de linguistique de l'UQAM, par Jacques Lecavalier du CEGEP de Valleyfield et par Jean-Charles Chébat du département des sciences administratives de l'UQAM.

Finalement, René Lortie, chargé de projet, et Richard Parent, coordonnateur du projet DELTA[2], tous deux du ministère des Communications du Québec, situeront le projet SATO-CALIBRAGE dans le contexte plus large de la lisibilité des publications gouvernementales. On trouvera aussi dans cet article le bilan d'une première tentative d'utilisation du prototype dans d'autres ministères et organismes.

La deuxième partie du Cahier est consacrée à la **méthodologie et à l'analyse**. Voici une brève description des contributions à ce chapitre.

François Daoust, chercheur responsable du projet SATO-CALIBRAGE au Centre ATO-CI, présente le cadre expérimental de l'ensemble du projet et le rôle particulier occupé par le logiciel SATO.

Lise Ouellet, du ministère de l'Éducation du Québec, nous dresse un portrait du corpus de textes sur lequel s'appuie le protocole expérimental.

Fernande Dupuis, du département de linguistique de l'UQAM et chercheur au Centre d'ATO, ainsi que François Daoust, présentent le dispositif linguistique qui est déployé à l'intérieur du prototype. Il y sera plus particulièrement question de la catégorisation grammaticale hors contexte et de la levée des ambiguïtés catégorielles sur les verbes par une stratégie de «grammaires locales».

Léo Laroche, du ministère de l'Éducation du Québec, Guy Cucumel du département des sciences comptables de l'UQAM, et François Daoust exposent divers aspects du dispositif statistique déployé dans le protocole expérimental. Divers résultats seront communiqués.

La troisième partie du Cahier porte sur le **prototype** SATO-CALIBRAGE dans sa version actuelle. Léo Laroche présente le prototype alors que deux utilisateurs, Pierre Achim et Georges Pelletier, nous livrent leurs réflexions en tant qu'utilisateurs. Aussi, Claire Gélinas-Chébat, avec Clémence Préfontaine et François Daoust, nous fournit un exemple d'utilisation de SATO sur des documents d'information de nature administrative.

Dans une quatrième partie, le Cahier nous présente des éléments de **conclusion et de prospective** sur les acquis et les développements futurs du projet.

Finalement, on trouvera en **annexe**, la liste des collaborateurs du projet et la liste des organismes qui nous ont soutenu d'une manière ou d'une autre.

Notes

[1] *SATO, Système d'analyse de texte par ordinateur*, François Daoust, Manuel de Références, Centre ATO-CI, janvier 1992.

[2] DELTA, projet sur l'informatique textuelle au Gouvernement du Québec; cf. article de René Lortie et Richard Parent, *Place de la lisibilité en gestion de l'information textuelle.*

Première partie
Problématique

Problématique liée à l'éducation

Lise Ouellet

Lise Ouellet est responsable de l'évaluation du français au primaire. Elle travaille à la Direction de la formation générale des jeunes du ministère de l'Éducation du Québec.

CONTEXTE

L'enseignement et l'évaluation de la lecture au primaire et au secondaire représentent des défis constants pour les enseignantes et les enseignants de même que pour les conseillères et les conseillers pédagogiques de français : le respect des orientations des programmes, la définition d'objectifs précis, l'intérêt des situations proposées aux élèves, le choix de textes, la difficulté des tâches à réaliser, la définition de seuils de réussite, l'interprétation des résultats, etc. Le choix de textes figure parmi les préoccupations de premier plan puisqu'il est au coeur de l'enseignement de la lecture.

Choisir un texte qui convient aux intérêts des élèves ou qui rejoint des thèmes abordés en classe peut être relativement facile; par contre, trouver un texte qui correspond au degré d'habileté à lire des élèves, cibler un texte qui présente un défi susceptible de les faire progresser, choisir rapidement un texte sans devoir faire de longues analyses, voilà tout un casse-tête pour de nombreuses personnes du milieu scolaire. Les exemples qui suivent illustrent quelques situations rencontrées couramment dans les écoles.

Plusieurs enseignantes et enseignants planifient leurs cours à l'aide du matériel didactique offert sur le marché; ils font ainsi confiance aux choix des auteurs et utilisent les textes tels quels. D'autres préfèrent présenter aux élèves des textes venant de sources diversifiées; ils se demandent alors si les 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 précis ou de stratégies de lecture à faire développer par les élèves; par exemple, on peut vouloir que les élèves s'entraînent à trouver le sens d'un mot inconnu d'après le contexte, à comprendre les liens qu'établissent certains mots de relation moins habituels ou à comprendre le sens de phrases très longues. Dans tous ces cas, au moment de la **planification**, l'enseignante ou l'enseignant 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 car c'est à partir d'eux que s'élaboreront les tâches à réaliser ou les questions à répondre. Les problèmes qu'on a énumérés pour la planification se retrouvent aussi au moment de l'évaluation : choisir le texte qui convient à la classe, quelles sont les difficultés qu'il pose,

soit pour les éviter, soit, au contraire pour vérifier si l'élève est capable de surmonter les difficultés que présente le texte. La plupart du temps, les choix de textes pour l'évaluation se fait de façon empirique, en se fiant à l'expérience des personnes qui font le choix. Par le passé, il est arrivé que les textes aient été trop faciles ou trop difficiles.

Les textes utilisés pour l'enseignement ou pour l'évaluation peuvent venir de multiples sources : matériel didactique, revues, livres, journaux, etc. Cependant, il peut arriver qu'on choisisse de rédiger des textes; il n'y a qu'à penser aux auteures et aux auteurs de matériel didactique ou aux situations d'évaluation. Au moment de la **rédaction**, la personne se verra confrontée aux mêmes questions qui ont été évoquées pour la planification et l'évaluation : on voudra un texte ni trop facile ni trop difficile ou, au contraire, un texte qui présente des défis particuliers pour que les élèves développent des stratégies spécifiques.

Que peut faire une enseignante ou un enseignant, une auteure ou un auteur devant ce problème? Bien sûr l'intuition et l'expérience viendront à la rescousse, mais est-ce suffisant? Est-on assuré d'avoir le texte qu'on cherche? Spontanément, le recours aux formules de lisibilité vient à l'esprit.

Les travaux sur la lisibilité

Plusieurs recherches ont été faites pour évaluer la lisibilité d'un texte et différentes formules ont été établies. Les indices ainsi trouvés permettent d'avoir une appréciation du degré de difficulté. Les formules les plus connues (Flesch, Henry, Gunning) ne prennent en compte qu'un certain nombre de variables et, le plus souvent, ce sont la longueur du texte, la longueur des mots et la longueur des phrases qui contribuent à calculer l'indice. Il est, de plus, très difficile d'interpréter les indices obtenus, soit que les travaux ont été faits en langue anglaise, soit qu'ils se sont surtout intéressés aux textes pour le grand public. En outre, à notre connaissance, aucun de ces travaux ne s'est préoccupé de la clientèle des élèves du primaire. Au Québec, différentes équipes de travail se sont aussi intéressées à la question de la lisibilité des textes.

Dans son livre **Apprendre à lire à 15 ans** (1989), Pierre Chamberland, conseiller pédagogique à la CECM, fournit une liste d'aspects à considérer pour évaluer la lisibilité d'un texte : le rapport sujet/lecteur; la qualité de l'information; la qualité du discours (organisation et formulation de l'information).

Dans la région de Laval-Laurentides-Lanaudière, des conseillères et des conseillers pédagogiques de français animés par Victor Guérette ont aussi fait des travaux qui allaient dans le même sens; des enseignantes et des enseignants ont même été habilités à utiliser les grilles qui permettent d'évaluer la lisibilité d'un texte.

Dans la foulée de ces travaux, la Direction de la formation générale des jeunes du ministère de l'Éducation a publié, en mars 1992, un document d'information intitulé **Grilles d'analyse pour évaluer des épreuves de lecture de primaire**; une des grilles proposées concernait la lisibilité d'un texte : elle comportait seize critères qui permettaient d'apprécier le

contenu, l'organisation, les caractéristiques linguistiques et la longueur d'un texte.

De plus, dans le programme de français de 1993, on trouve des indications concernant la complexité des textes pour chaque classe.

Jacques Lecavalier du cégep de Valleyfield et Clémence Préfontaine de l'Université du Québec à Montréal, quant à eux, mènent des travaux sur l'intelligibilité des textes. On trouvera d'ailleurs, dans ce cahier, un article de Claire Gélinas-Chébat, Clémence Préfontaine, Jacques Lecavalier et Jean-Charles Chébat sur la lisibilité et l'intelligibilité.

Il faut noter que si tous ces travaux établissent clairement les différents aspects à examiner lorsqu'on veut choisir un texte pour les élèves, aucun d'eux ne fournit de balises pour déterminer à quelle classe un texte devrait être utilisé. De plus, les aspects du texte qui facilitent la lecture ou qui, au contraire, rendent le texte plus difficile, sont pointés globalement; il faut refaire une lecture attentive du texte pour trouver les passages qui pourraient nous intéresser. En outre, les grilles proposées doivent être remplies à la main; aucune suggestion n'est faite pour que certains aspects (Ex. : la longueur du texte ou la familiarité du vocabulaire) soient évalués à l'aide de l'ordinateur.

LE DÉVELOPPEMENT DE SATO-CALIBRAGE

À la fin des années quatre-vingt, quelques personnes du ministère de l'Éducation ont pris connaissance du logiciel SATO développé par le Centre d'ATO de l'Université du Québec à Montréal. Elles ont été fascinées par les possibilités du logiciel pour l'analyse des textes. L'indice Gunning étant implanté à l'intérieur de SATO, il est apparu que le logiciel pourrait faire davantage pour étudier la question de la lisibilité des textes. Comme les personnes qui s'intéressaient au projet avaient en tête la problématique du primaire et du secondaire, les travaux ont été graduellement orientés pour résoudre la problématique décrite précédemment.

Le terme «calibrage» des textes, utilisé par certains groupes du milieu scolaire, s'est installé dans nos travaux sans qu'il soit remis en question, d'où l'appellation SATO-CALIBRAGE. Il s'agit d'un néologisme qui donne une acception nouvelle au mot calibrage. Calibrer un texte signifie donner un indice à ce texte pour le situer sur un continuum, établi de la première année du primaire à la cinquième année du secondaire. L'indice recherché veut dépasser les formules traditionnelles de lisibilité et prendre en compte les meilleures variables susceptibles de démarquer un texte d'une classe à l'autre.

Dès le départ, il est apparu que les travaux ne pourraient progresser sans une collaboration étroite entre le ministère de l'Éducation, le Centre d'ATO de l'UQAM et les milieux scolaires. D'autres personnes, des linguistes et des personnes de différents ministères, s'intéressent à nos travaux, profitent du développement mais ne participent pas directement à l'amélioration de l'outil de calibrage de textes pour le milieu scolaire.

Personnes impliquées

Une équipe de coordination s'est spontanément mise sur pied pour travailler sur ce problème à multiples facettes. Cette équipe est formée de :

- François Daoust, analyste en informatique, chercheur au Centre d'ATO et auteur du logiciel SATO;
- Léo Laroche, spécialiste en statistiques, préoccupé de l'analyse des textes; de 1986 à mars 1992, il travaillait pour la Direction générale de l'évaluation et des ressources didactiques, depuis 1992, il est à la Direction de la Recherche du ministère de l'Éducation;
- Lise Ouellet, responsable de l'évaluation du français au primaire; jusqu'en mars 1992, elle était à la Direction du développement de l'évaluation qui, à partir d'avril 1992, a été intégrée à la Direction de la formation générale des jeunes (MEQ).

La participation du milieu scolaire est essentielle au développement de SATO-CALIBRAGE. Tout d'abord, un groupe d'usagers a été mis sur pied; il s'agit de conseillères et de conseillers pédagogiques de français du primaire et du secondaire, de responsables d'élaboration d'épreuves au ministère de l'Éducation et à BIM[1]; quelques personnes du milieu collégial et du milieu universitaire font aussi partie de l'équipe. On trouvera en annexe les noms des personnes qui constituent ce groupe. Les personnes impliquées reçoivent une version de travail du logiciel, l'expérimentent et font des commentaires ou des suggestions.

On a eu aussi recours à des enseignantes et des enseignants du primaire pour établir la liste des mots connus des élèves de sixième année.

Enfin, le travail envisagé commandait la constitution d'un corpus de textes; celui-ci a pu être élaboré grâce à la participation de diverses personnes du milieu scolaire. On trouvera plus loin des explications sur le corpus.

CHRONOLOGIE DU TRAVAIL FAIT AU COURS DES DERNIÈRES ANNÉES

Année scolaire 1988-1989

Juillet 1988	Premier énoncé de la problématique de calibrage des textes pour les élèves du primaire et du secondaire.
Mars 1989	Signature d'une convention entre le ministère des Communications, le

ministère de l'Éducation et l'Université du Québec à Montréal permettant de conduire des travaux de recherche devant aboutir à la mise au point d'instruments informatiques pour l'analyse de la lisibilité de textes destinés aux élèves du primaire et du secondaire.

Publication dans le **Bulletin des systèmes experts** du texte de Léo Laroche intitulé **«L'utilisation de SATO pour calibrer des textes»**.

Avril 1989 — Début des travaux pour constituer un corpus de textes : collaboration avec BIM[1] pour utiliser des textes qui étaient déjà sur support informatique; intégration des textes utilisés pour les épreuves d'appoint du ministère de l'Éducation.

Année scolaire 1989-1990

Septembre 1989 — Inventaire des méthodes en usage pour évaluer la lisibilité d'un texte; étude des grilles d'analyse de textes utilisées dans différents milieux scolaires.

Octobre 1989 — Rencontre d'un groupe de travail à l'UQAM : objectifs poursuivis, démonstration du logiciel et échange sur la façon de faire des participantes et des participants pour juger de la lisibilité d'un texte.

Novembre 1989 — Rencontre de l'équipe de coordination : discussion sur la pertinence d'établir le lexique des mots connus des élèves; préparation d'une stratégie de consultation.

Consultation de Pierre Chamberland qui a établi des aspects à considérer pour choisir des textes; échange sur les aspects qui pourraient être mécanisés.

Février 1990 — Établissement d'un lexique à partir du corpus (environ 375 textes).

Avril 1990 — Consultation de cinq enseignantes pour juger de la familiarité du vocabulaire; ces personnes venaient de régions variées et représentaient des milieux différents : rural, semi-rural, moyen (de banlieue), défavorisé, aisé (éduqué).

Mai-juin 1990 — Entrée des données de la consultation; constitution d'un lexique de mots connus des élèves de sixième année.

Juin 1990 — Rapport d'évaluation du projet.

Année scolaire 1990-1991

Août 1990	Rencontre de l'équipe de coordination : examiner la possibilité de créer des macro-commandes pour faciliter le travail sur les textes; discussion sur la diffusion des travaux.
Septembre 1990	Publication dans la revue ICO du texte de Léo Laroche intitulé «**Calibrage des textes et lisibilité**».
Octobre 1990	Atelier au congrès de l'ADMEE (Association pour le développement de la mesure et de l'évaluation en éducation) : présentation par François Daoust et Léo Laroche.
Décembre 1990	Entente entre le ministère de l'Éducation et le Centre d'ATO pour la poursuite des travaux.
Avril 1991	Rédaction d'un mini-guide d'utilisation. Formation d'un groupe d'usagers. Première rencontre du groupe : présentation du logiciel, travail pratique et exercices sur le mini-corpus. Consultation de Michel Pagé, professeur au département de psychologie de l'Université de Montréal au sujet de la compréhension en lecture, des mots catégoriels, des connecteurs, de la richesse de l'information, etc. Atelier présenté par Léo Laroche et François Daoust au congrès «Writing Assessment» tenu à l'hôtel Sheraton à Montréal.
Mai 1991	Rencontre des usagers : mise à jour du logiciel, formation, échange sur le type de rapports produits, attentes pour l'amélioration du logiciel; demande pour l'augmentation du corpus.
Juin 1991	Travail sur l'identification des mots qui pourraient représenter des catégories sémantiques (ex. : animaux, aliments, personnes) à partir du lexique issu du corpus. (Ce travail n'a pas été concluant et aucune liste n'a été intégrée au prototype.) Le problème des mots qui représentent des catégories est apparu plus complexe que prévu et il est laissé en plan.

Année scolaire 1991-1992

Septembre 1991	Élaboration de macrocommandes pour faciliter l'utilisation de SATO-CALIBRAGE.
Octobre 1991	Rencontre des usagers : explications sur les modifications (commandes disponibles); réponse aux questions.

Novembre 1991	Présentation du logiciel et des avenues de développement au congrès de l'AQPF (Association québécoise des professeures et des professeurs de français à Québec).
Décembre 1991	Consignes pour la constitution du corpus.
	Rencontre du groupe des usagers : le point sur la constitution du corpus; commande pour enlever l'ambiguïté sur les verbes; consignes pour l'identification des fichiers transmis.
Février 1992	Rédaction du texte **SATO-CALIBRAGE : Utilisations et avenues de développement**.
	Présentation du texte et demande de subvention à la Direction des ressources didactiques du ministère de l'Éducation.
	Rencontre du groupe des usagers : utilisations possibles de SATO-CALIBRAGE; examen critique des variables disponibles; présentation des dernières analyses statistiques; pistes de développement; réponse aux questions.
Mars 1992	Bilan sur le corpus (environ 700 textes).
	Analyses statistiques.
Avril 1992	Épuration du corpus à partir des analyses statistiques.
Mai 1992	Ajout de textes pour combler les lacunes.

Année scolaire 1992-1993

Juillet 1992	Vérification du lexique afin de corriger les fautes.
	Analyses statistiques sur la variable «mots de relation».
Août 1992	Premières analyses statistiques pour développer un indice de lisibilité SATO.
Septembre 1992	Travail sur le logiciel SATO pour faciliter l'accès aux dictionnaires.
	Problématique liée à l'interface.
Octobre 1992	Amélioration de l'opération pour enlever les ambiguïtés sur les formes verbales.
Janvier 1993	Amélioration à l'interface pour le rendre plus facile d'utilisation.
Février 1993	Implantation de l'indice SATO (version provisoire).
	Rencontre des usagers : présentation de la dernière version de SATO-

	CALIBRAGE; présentation de l'indice SATO; discussion et échange sur le type d'interface à développer.
Avril 1993	Consultation de six enseignantes et enseignants venant de milieux et de régions variés pour juger de la familiarité du vocabulaire (environ 8000 mots à évaluer).
Mai 1993	Constitution du nouveau vocabulaire des mots connus.
	Analyses statistiques pour établir l'indice SATO.

Notes

[1] BIM, Banque d'instruments de mesure de la société GRICS de Gestion des ressources informatiques des commisssions scolaires.

Lisibilité - Intelligibilité de documents d'information

Claire Gélinas-Chebat, Clémence Préfontaine,
Jacques Lecavallier et Jean-Charles Chebat,

Claire Gélinas-Chébat et Clémence Préfonfaine sont professeures au département de linguistique, Université du Québec à Montréal; Jacques Lecavalier est professeur au CEGEP de Valleyfield; Jean-Charles Chebat est professeur au département des Sciences Administratives, Université du Québec à Montréal.

Novembre 1992 (revisé en juin 1993)

Introduction

Les entreprises de services (ministères, banques, compagnies d'assurance, etc.) produisent de nombreux documents dans le but d'informer le public. Or souvent, ces fascicules ne remplissent pas leur fonction puisqu'ils s'avèrent trop difficiles à comprendre pour le lecteur cible (Gélinas-Chebat, Macot, Préfontaine, Daoust, 1991).

Du point de vue linguistique, il existe un certain nombre de règles pour faciliter la lecture des textes. Ces règles touchent tout aussi bien la syntaxe (longueur et complexité des phrases, par exemple) que la sémantique (le choix des termes, par exemple). Du point de vue typographique, il existe également un certain nombre de règles; ces règles concernent par exemple, les polices de caractère (le choix, la taille du caractère), la disposition visuelle (l'utilisation de textes verticaux, la longueur des lignes de lecture).

Il est possible de mesurer le niveau de lisibilité des textes. Richaudeau, Flesch, Gunning ont proposé différentes procédures. Selon Gunning, la lisibilité repose essentiellement sur la longueur des phrases et des mots. Sato, un logiciel d'analyse de texte par ordinateur fournit, entre autres, cet indice de lisibilité.

Préfontaine et Lecavalier (1990) propose un modèle d'analyse des textes qui permet de tenir compte des différents facteurs qui contribuent à rendre la communication écrite efficace. Ces auteurs parlent d'intelligibilité des textes. Ce modèle permet une description microstructurelle (niveau des mots et des phrases), superstructurelle (niveau de l'organisation formelle des textes) et macrostructurelle (niveau sémantique et au niveau de la cohérence explicite et implicite des textes écrits).

Nous tenterons ici d'opposer les concepts et modèles de lisibilité (généralement retenu dans la littérature) à celui d'intelligibilité. C'est dans la première partie que nous tenterons de définir la lisibilité et de montrer les nombreux facteurs affectant la compréhension d'un texte écrit; ce sont des facteurs reliés au lecteur et à la lecture. Aussi, nous considérerons non seulement les aspects linguistiques, mais également des aspects graphiques ainsi que les

caractéristiques du lecteur qui le rendent plus ou moins habile à comprendre le texte.

Dans la seconde partie nous définirons le concept d'intelligibilité et décrirons le modèle.

Partie I : La lisibilité

A) Définition

La lisibilité peut être définie comme «une aptitude du texte à se faire comprendre» (Bourque, 1989). Cette définition très large implique que le lecteur sache reconnaître dans le texte les signes qui permettent sa compréhension. En anglais, comme le rapporte Morin, Sallio et Kretz (1982), on utilise «legible» pour désigner la lisibilité matérielle, typographique d'un texte et «readability» pour désigner la dimension intellectuelle et psychologique lié au processus de compréhension d'un texte lu. Timbal-Duclaux (1985) a d'ailleurs tenté de proposer les termes de «lisable - inlisable» pour distinguer l'aspect intellectuel de l'aspect matériel «lisible- illisible» (p. 14).

Comme le mentionne Richaudeau (1978), un texte efficace est un texte qui permet une lecture efficace c'est-à-dire qui permettra au lecteur d'être enrichi d'une information nouvelle. Le premier facteur d'efficacité peut-être mesuré par sa lisibilité. En linguistique, comme le mentionne Fernbach (1990),

> **la lisibilité est l'aptitude d'un texte à être lu rapidement, compris aisément et bien mémorisé.**

Depuis le début des années 1920, plusieurs formules ont été mises au point pour mesurer la lisibilité des textes. Elles utilisent différents éléments de la langue, comme la longueur des mots et des phrases, la rareté des mots, leur fréquence d'utilisation, etc. Les formules de lisibilité les plus connues sont, pour le français, celle de Georges Henry (1975) et de De Landsheere (1963, 1973). Mais les formules de lisibilité qui ont inspiré ces auteurs européens sont toutes américaines. Pensons à celles de Lively et Pressey, 1923; Washburne et Vogel, 1926; Gray et Leary, 1935; Dale et Chall, 1948; Flesch, 1948; Gunning, 1952; Taylor, 1953; Chall, 1958; Fry, 1968 et 1977.

Mais au delà des éléments considérés pour mesurer la lisibilité, il est important de comprendre que d'autres facteurs peuvent affecter la lecture; ce sont des facteurs reliés au lecteur et à la lecture.

B) Le lecteur et la lecture

Afin de saisir la complexité du processus de lecture, nous considérerons successivement les aspects physiologiques et cognitifs liés à l'acte de lecture, les difficultés du lecteur, les éléments liés à la technologie textuelle et enfin certains facteurs d'ordre linguistique.

1. Le processus mécanique de la lecture

Certains aspects doivent être considérés lorsqu'il est question des dimensions physiologiques de la lecture; il s'agit de la perception visuelle, du traitement par la mémoire de travail ainsi que par la mémoire à long terme, des éléments lus.

La perception visuelle doit être expliquée en considérant la fixation et l'empan visuel. La *fixation* est le temps de déchiffrage entre deux déplacements des yeux et l'*empan visuel* est défini par l'étendue de ce qui est vu durant chaque fixation. Lorsqu'il lit, le lecteur recherche des formes significatives, c'est-à-dire des formes connues. Un lecteur expérimenté ne fait pas une lecture lettre par lettre mais tente de retrouver des formes globales de lettres qu'il interprète. La présence de formes connues favorise son anticipation du sujet et réduit le temps de décodage, ce qui augmente ainsi la lisibilité du texte. En fait, plus le lecteur connaît de formes, plus son temps de décodage diminue.

Il est important d'éviter les ambiguïtés sémantiques et syntaxiques, c'est-à-dire les ambiguïtés qui découlent d'imprécisions au niveau du sens ou de la structure des phrases. «Le texte doit réduire le nombre des alternatives possibles par un usage rigoureux de la langue, particulièrement en donnant un sens précis aux mots et en utilisant des constructions syntaxiques prédictives» (Bourbeau, 1988, p. 16).

La mémoire à court terme ne peut traiter qu'un petit nombre d'éléments à la fois, plus précisément sept signes seulement. Il est donc essentiel que le lecteur puisse faire rapidement et simplement les liens entre les mots, sinon son attention est détournée du texte et il ne comprend plus le sens de ce qu'il lit. Comme la définit Smith (1986): «La mémoire à court terme, c'est ce à quoi nous prêtons attention dans l'instant, et ce qui se perd si notre attention est attirée ailleurs». (p. 49).

En lecture, comme le rapporte Racle (1988), les phrases élémentaires (il faut comprendre les unités sémantiques) sont considérées comme les unités de base du traitement de l'information dans la mémoire active. Lorsque les phrases élémentaires sont assemblées et intégrées, c'est-à-dire lorsque le lecteur peut établir des liens entre ce qu'il a lu et ce qu'il prévoit lire, alors ces phrases passent dans la mémoire à long terme. La mémoire active assemble, intègre, transfère à la mémoire à long terme l'information pour passer à la phrase suivante.

La mémoire à long terme, quant à elle, peut emmagasiner un très grand nombre d'informations, qui sont regroupées de façon à être retrouvées. La mémoire à long terme joue un rôle important en lecture puisqu'elle permet d'ajouter les informations saisies en les logeant au bon endroit; toutefois, cela est vrai si on comprend ce qu'on lit car: «[...] la mémoire à long terme se réorganise si efficacement et si facilement qu'on ne se rend même pas compte que l'on est en train d'apprendre» (Smith, 1986, p. 52). La mémoire à long terme fonctionne d'autant mieux que les informations nouvelles sont associées à des connaissances antérieures, comme il sera montré dans la section suivante.

Racle (1988) considère de plus qu'un lecteur ne peut lire un texte, c'est-à-dire reconnaître et comprendre, qu'en fonction de ses expériences. Les connaissances particulières d'un lecteur peuvent même l'amener, selon cet auteur, à une interprétation particulière d'une phrase ou d'un texte lu. Ehrlich et Tardieu (1985) traduisent ce phénomène en considérant que la qualité de la réception du message dépend de plusieurs types de facteurs. Il existe des facteurs liés aux motivations du sujet par rapport à l'objectif et à la finalité de la communication, ce qui déterminera ses intentions et ses attitudes à l'égard du message. Il existe évidemment des

facteurs de nature cognitive qui concernent les structures et les processus mis en jeu lors de la saisie de l'information.

Les résultats des recherches d'Erlich et Tardieu (1985), tendent à démontrer une relation étroite entre compréhension et mémoire et il semble qu'un texte narratif est jugé plus facile à comprendre qu'un texte descriptif, lui-même jugé plus facile qu'un texte théorique.

2. Les processus cognitifs de la mémoire

De nombreuses recherches tentent de mettre en évidence comment les individus acquièrent l'information et comment ils en font l'intégration à l'intérieur de leur système cognitif. L'individu traite un grand nombre d'éléments d'information. La tâche d'acquisition de l'information suppose l'individu apte à faire des choix parmi un certain nombre d'alternatives : l'individu recherche dans une accumulation d'informations qui lui sont présentées, celles qui lui seraient utiles; il cherche à atteindre un objectif et suit une route séquentielle, par étape. L'information pertinente est stockée en mémoire et s'ajoute à l'information mémorisée dans la mémoire à long terme.

Dans la tâche d'intégration de l'information, l'individu pose un jugement global sur chaque objet. Ce jugement est une évaluation de ces objets sur un certain nombre de dimensions. À chaque objet correspond un dossier, mis en mémoire, identifié (étiqueté), codé et retraçable en mémoire grâce à un système de classification.

Dans les deux sortes de tâches cognitives, acquisition et intégration, l'information à traiter se structure sur une matrice comprenant les objets (en lignes) et les attributs caractérisant ces objets (en colonnes). Dans ces deux sortes de tâches l'individu doit:

- mettre en mémoire l'information et la coder;
- réaliser un certain nombre d'opérations cognitives pour transformer ces informations en images mentales.

Depuis Piaget, les recherches sur les processus cognitifs distinguent deux stratégies cognitives de base : la combinaison systématique et l'isolement des variables. Dans le premier cas, la combinaison systématique, l'individu génère toutes les combinaisons possibles construites sur la base d'un nombre limité d'éléments (par exemple les lettres a,b,c,d,et e). Il s'agit d'une tâche qu'on ne peut réaliser qu'avec des règles de procédure systématique pour combiner les éléments de base.

Dans le cas de l'isolement des variables, l'individu fait face à une base de données comprenant plusieurs variables (par exemple, s'il s'agit de produits de consommation : les marques, les prix, les distributeurs, etc...) et plusieurs effets (par exemple, avec les produits de consommation : les performances, les coûts d'utilisation, etc...). Ici, la tâche cognitive consiste à isoler celle des variables qui constitue la cause des effets constatés (par exemple : la marque est reliée aux coûts d'utilisation).

Si ces deux stratégies mises en évidence par Inhelder et Piaget (1958) se sont révélées, en général, valides, les recherches plus récentes soulignent que pour un individu donné, le fait de posséder une habileté cognitive spécifique (par exemple l'identification de la variable causale) n'implique pas qu'il possède les autres (par exemple la capacité combinatoire).

En conséquence, la présentation de l'information doit tenir compte des contraintes propres aux individus. Certains vont chercher à identifier la variable causale qui provoque un effet recherché. Par exemple, quel numéro de téléphone appeler pour obtenir telle brochure, tel bénéfice, etc.; quelle action entreprendre pour devenir éligible à un programme ministériel, par exemple.

D'autres individus vont procéder de façon radicalement différente et vont chercher à combiner toutes les informations disponibles de plusieurs manières possibles de façon à maximiser leur utilité. Par exemple, quelles conditions doit réunir l'individu pour être éligible à un nouveau programme du ministère sans pour autant perdre son éligibilité à des programmes où il est déjà inscrit.

Il y a des degrés divers d'acquisition et d'intégration de l'information, degrés qui sont le propre de l'individu. Ainsi les capacités d'acquérir, mémoriser et traiter l'information sont interreliées. Les individus qui acquièrent peu d'informations à l'intérieur de leur mémoire à court terme sont aussi ceux qui sont le moins capables d'intégration. Ce qui a des effets cumulatifs : le ministère qui essaie de transmettre de l'information sur des programmes modifiés, a peu de chances que l'information marginale soit intégrée si la précédente information n'a pas été mémorisée, ni structurée. Inversement, ceux qui ont intégré les précédents programmes sont aussi ceux qui acquièrent le plus d'informations nouvelles.

3. Les difficultés du lecteur

Les dépliants produits par les entreprises de service demandent une compréhension univoque: il faut comprendre le sens donné par l'entreprise et aucun autre. Ceci peut expliquer certaines des difficultés rencontrées par les lecteurs, qui ne peuvent lire ces textes comme d'autres qu'ils liraient pour leur plaisir. Cette lecture fait appel à une habileté particulière. Il est alors souhaitable que les documents écrits offrent un soutien aux lecteurs, quel que soit leur niveau d'habileté : un titre clair, des numérotations, des divisions, par exemple seront des éléments facilitateurs qui aideront le lecteur à s'approprier le texte.

Les lecteurs en difficulté peuvent obtenir de l'aide lorsque des éléments facilitateurs sont ajoutés à un texte considéré comme difficile ou lorsque certains éléments qui y sont déjà présents sont mis en valeur (Tardif et Gaouette, 1986a). Ces éléments facilitateurs peuvent être des définitions claires ou l'accentuation de certaines informations en utilisant des procédés de mise en évidence, choisis selon des critères validés.

Il est également important d'indiquer clairement au lecteur l'intention de lecture qu'il devrait avoir lorsqu'il aborde un texte. Par exemple, une phrase-clé pourrait faire comprendre au lecteur la nécessité de lire le dépliant ou la brochure, car il est important que le lecteur comprenne la finalité de sa lecture (Tardif et Gaouette, 1986a). Le titre, qui a un rôle clef pour permettre l'anticipation du contenu, doit être indiqué clairement (Tardif et Gaouette, 1986a, p. 14) et un sous-titre explicite peut être ajouté, ce qui aide le lecteur à anticiper le contenu du texte.

Un autre aspect qu'il ne faut pas négliger, c'est l'activation des connaissances préalables que le lecteur a du sujet sur lequel porte le texte car il faut savoir que «la réalité affective du lecteur en regard du thème traité façonne en partie sa compréhension» (Tardif et Gaouette, 1986b, p. 5). L'utilisation, par exemple, de «vous savez déjà que...» pour rappeler une information déjà traitée, permet au lecteur de ne pas se retrouver devant la difficulté d'aborder des informations qu'il *croit* être nouvelles.

Par ailleurs, il arrive que les informations soient données explicitement dans un texte, de façon à éviter au lecteur d'avoir à faire des inférences autres que les inférences légitimes à tout acte de lecture : il est certain que le lecteur moins habile peut éprouver beaucoup de difficulté.

4. La technologie textuelle

Des recherches ont montré que la technologie textuelle (Jonassen, 1982, 1985) peut contribuer à aider le lecteur. Par technologie textuelle il faut comprendre le dynamisme des messages produits par la structuration des textes, soit le formatage, le titrage et l'apparence typographique (pour ce dernier aspect, voir Lebrun et Berthelot, 1991).

Pour soutenir la lecture, il importe de synthétiser les informations, de les assembler et de les organiser afin de «rendre interactifs la forme, l'étalage, la typographie, les descripteurs, le style, en somme l'ensemble des indicateurs et facteurs de la lisibilité [...]» (Vachon, 1988, p. 12).

L'utilisation de l'espace blanc, d'une ligne verticale à gauche, du point-repère (gros point placé à gauche d'un bloc d'informations) etc. sont autant de façons d'appliquer les principes de la technologie textuelle.

Du point de vue typographique, nous savons qu'il existe un choix impressionnant de polices de caractères. Lebrun et Berthelot (1991) souligne l'importance de l'empattement des caractères, de leur taille comme facteurs qui contribuent à la lisibilité des textes.

Selon ces auteurs, l'utilisation des majuscules, les lettres inclinées (italiques), l'espacement entre les lettres, fixe[1] d'une part et condensé[2] d'autre part, réduisent la vitesse de lecture. Selon ces auteurs toujours, il est préférable d'avoir un texte au fer à gauche et non-justifié pour un lecteur débutant.

Morin, Sallio et Kretz (1982) ont montré expérimentalement la supériorité des matrices variables sur les matrices fixes que les caractères soient de types romains ou italiques.

Macot (Gelinas-Chebat, 1991) donnait quelques règles simples du point de vue typographique pour permettre une meilleure lisibilité d'un dépliant d'information. Pour Macot les éléments graphiques doivent contribuer à la transmission d'un message et ne doivent d'aucune façon devenir plus important que le message lui-même. Les fantaisies graphiques, les couleurs ajoutées au texte, ne doivent être là que pour renforcer la compréhension du message. Il spécifiait :

«Du point de vue typographique, il est préférable :

- d'employer la même police de caractères (fonte typographique) pour l'ensemble du texte;
- de choisir un caractère normal plutôt qu'un caractère condensé;
- d'employer une police qui a des caractères carrés plutôt qu'allongés;
- d'avoir des interlignes équivalents à la hauteur du caractère typographique choisi;
- de ne pas utiliser les soulignés dans le corps du texte parce qu'ils entraînent trop

souvent des difficultés de lecture;

- pour mettre en évidence un mot ou une série de mots dans une ligne, il est préférable soit de choisir une police avec une graisse plus forte mais dans le même caractère que l'ensemble du texte;
  ou encore d'employer un jeu de couleurs, en délimitant une surface équivalente à la longueur et à la hauteur du mot ou des mots à mettre en évidence, en considérant les majuscules de ceux-ci. L'impression de cette surface se fera dans une couleur différente de la couleur choisie pour le texte. Il est bien entendu que la couleur de cette surface imprimée doit être d'une tonalité beaucoup plus légère que la couleur du bloc typographique;

- l'emploi de lettrines[3] est à déconseiller. Une lettrine mal positionnée amène une difficulté de lecture qui peut aller jusqu'à l'incompréhension de la ligne qui suit la lettrine;

- pour les index, tables des matières et tableaux divers, la plus grande ligne du texte devrait-être très rapprochée de la numérotation des pages, les autres intervalles étant calculés sur celle-ci;

- si la colonne de chiffres ne peut être près de la colonne de texte, on peut placer un léger grisé entre deux colonnes mais il faut considérer le nombre de lignes :
  si c'est un nombre impair, l'intervalle entre les colonnes 1, 3, 5, etc., devrait-être grisé;
  si c'est un nombre pair, l'intervalle entre les colonnes 2, 4, 6, etc., devrait-être grisé;

- pour les textes verticaux, la seule position acceptable est la suivante : le texte doit être perpendiculaire aux bords horizontaux de la page et être positionné pour que la lecture se fasse de bas en haut;
  cette position est de toute façon à déconseiller;

- l'emploi de fantaisies graphiques dans un dépliant doit contribuer à une plus grande compréhension du message, sinon c'est à exclure.» (Gélinas-Chebat, 1991, p. 26)

La très grande majorité des études sur le traitement de l'information porte sur les mots, peu sur les images. Or le traitement de l'image est un facilitateur puissant du traitement de l'information. Ceci est d'autant plus important que la publicité télévisée ou imprimée fait appel à l'image.

L'image a un effet cognitif non-contesté dans la littérature cognitiviste : elle permet l'amélioration sensible de la mémorisation de l'information. Les chercheurs estiment que l'image physique génère plus aisément des images mentales que les mots ne peuvent le faire. L'imagerie visuelle est en général un élément mnémonique puissant qui stimule l'apprentissage et la rétention du matériel publicitaire plus que les mots. Ceci est encore plus évident lorsqu'on veut que le public se souvienne d'interactions entre des objets (par exemple: une baleine et un cigare sont deux objets isolés; une baleine fumant un cigare constitue une interaction permettant d'illustrer un concept, par exemple que le fait de fumer fait grossir).

Nous pouvons donc retenir que le message verbal dans sa forme écrite (sans image) apparaît adéquat si le public est motivé par le sujet et s'il est apte à traiter l'information. C'est

le cas du public pour qui le sujet de l'information est important et qui peut contrôler le rythme de traitement de l'information (exemple : la brochure qu'on peut lire et relire). Mais le taux d'oubli du matériel verbal est plus élevé et il faut plus de répétitions du message pour contre-balancer cet oubli.

Le message visuel est plus adapté dans le cas où le public est moins motivé ou moins habile sur le plan cognitif. C'est le cas des publics exposés à des médias dont ils ne contrôlent pas le rythme d'exposition (exemple : la télévision). L'avantage de l'image est de nécessiter moins d'exposition répétées pour avoir le même effet à long terme. Il est tout de même important de mentionner que les représentations visuelles de l'information constituent en soi un champ de recherche très vaste qui dépasse largement la présente recherche.

5. Eléments linguistiques

Pour bien rédiger selon Fernbach (1990), spécialiste de la rédaction de textes juridiques, «il faut écrire en fonction du destinataire pour être sûr qu'il comprenne et qu'il retienne le texte (...). La bonne qualité de la rédaction s'apprécie, non seulement d'après le choix des termes adaptés, mais aussi par le choix de tournures claires (...) (Il faut éviter) les doubles négations, les formes compliquées et figées, (...) techniques de rédaction qui visent à causer des effets indirects. On déplore aussi les textes où l'auteur suppose que le lecteur dispose d'une quantité préalable d'information, lorsque ce n'est pas le cas» (Fernbach, 1990, p. 8-9)

Fernbach poursuit : «Mis à part la question typographique (...), la difficulté demeure dans le choix des mots et des structures de phrases. Parmi les caractéristiques de l'écrit qui nuisent à la lecture facile et à la bonne compréhension, citons

- les phrases trop longues;
- l'abus de substantifs (articles, pronoms démonstratifs, possessifs, etc), de mots trop longs ou de synonymes;
- le phénomène d'écran linguistique[4];
- la dépersonnalisation;
- la distanciation[5], etc.» (Fernbach, 1990, p. 9)

Erlich et Tardieu (1985), soulignent l'importance des titres, de l'importance de rendre les textes compréhensibles et de leur cohérence pour en faciliter la lecture. Mesnager (1979), établit un lien évident entre longueur des mots, rareté et difficulté de lecture : plus un mot est long, plus il est rare et plus il sera difficile à interpréter.

Timbal-Duclaux (1984) souligne plusieurs facteurs qui contribuent à rendre les textes difficiles à lire. D'abord, au niveau du style du texte, il faut éviter les textes où on retrouve un nombre élevé de mots abstraits et de génitifs qui se succèdent en cascade (de...de...de...). Il faut éviter la monotonie des structures, il faut varier les constructions pour mettre en relief les mots clefs, les mots porteurs de sens.

Cet auteur rappèle l'intérêt des phrases prédictives c'est-à-dire des phrases tournées de telle sorte que le début fait attendre une suite. Il faut laisser dominer les verbes à la voix active par opposition aux verbes à la voix passive. De plus, souvent le style passif est un style impersonnel, ce qui est à éviter.

Les textes gagnent de toute évidence à être personnalisés, à être concrétisés par l'utilisation de sujets réels.

Timbal-Duclaux dans un autre article, (1985) propose quelques règles simples pour pouvoir produire des textes «lisables» (p. 16) :

- éviter les phrases longues et complexes;
- l'abus des tournures passives;
- l'abstraction continue;
- l'abus des adjectifs, des adverbes et des noms;
- le vocabulaire inutilement technique et non-expliqué.

En fait, il suffit de produire des phrases courtes, une ponctuation fréquente, un vocabulaire simple. Pour y arriver, l'auteur suggère un test, le test de lisibilité simplifié de Rudolf Flesch.

Pour une tranche de 100 mots consécutifs, extraite au hasard dans un texte à analyser, on compte un point dans les six cas suivants :

- pour chaque majuscule ;
- pour chaque mot du texte souligné par un caractère gras ou italique;
- pour tous les nombres écrits en chiffres (pas en lettres);
- pour tous les signes de ponctuation, sauf les virgules, les traits d'union et le point quand il sert à abréger un mot;
- pour tous les symboles courants du type #,$,%,&,*, etc.;
- à chaque fin d'alinéa.

Le chiffre total obtenu en additionnant tous les points est l'indice de lisibilité. Flesch donne le barème suivant :

Score total	type de texte (traduction de Timbal-Duclaux)
0 à 20 points	soutenu, noble, guindé, académique
21 à 25	registre médian, courant
26 à 30	assez lisible, assez grand public
31 à 35	grand public
plus de 35 points	très grand public

Pourquoi ces facteurs sont-ils retenus et méritent qu'on leur attribue des points ?

Parce que chacun des éléments soulignés permettra d'apporter une contribution particulière à la lisibilité du texte. En fait, la majuscule marque soit un nom propre, soit un début de phrase. Les mots soulignés sortent de la grisaille du texte Les nombres introduisent de la variété dans le texte. Les signes de ponctuation (autre que les virgules, les traits d'union et les points) donnent des phrases mieux articulées, plus vivantes. Les symboles introduisent de la variété et les alinéas permettent de faire des intertitres.

Un second test de Flesch, décrit par Timbal-Duclaux (1985) touche l'aspect sémantique. Il permet d'assurer la bonne compréhension des textes. Il suffit de produire des textes dont le nombre de mots «concrets» est supérieur au nombre de mots «abstraits».

Comment procéder ? Pour chaque tranche de 100 mots consécutifs, accorder un point à chaque

mot concret. Les mots concrets ont la propriété de ne pas changer de signification en changeant de langue. i.e.:

- les noms de gens;
- les nombres et les mots signifiant des chiffres;
- les dates (années, saisons, jours, heures, ...);
- les mots qui désignent un sexe (jument/cheval, homme/femme);
- les mots qui désignent des personnes précises (moi, tu, il, mon, votre, etc.);
- ajoutons tous les mots qui deviennent concrets dans le contexte des 5 catégories citées. (par exemple le mot «idée», il est abstrait mais devient concret dans l'idée de Luc, la troisième idée, l'idée de la fille, etc.) .

Le pourcentage total est obtenu en additionnant tous les points. Flesch donne le barème suivant:

Score total	Niveau d'abstraction
0 à 20 % de mots concrets	hautement abstrait
20 à 30 %	plutôt abstrait
30 à 45 %	plutôt concret
45 % et plus	hautement concret

En fait, pour conclure sur l'écrit:

> «Il n'est point besoin de vocabulaire bizarre, compliqué, nombreux et chinois (...) pour fixer toutes les nuances de la pensée. Mais il faut discerner avec une extrême lucidité toutes les modifications de la valeur d'un mot suivant la place qu'il occupe. Ayons moins de noms, de verbes et d'adjectifs aux sens presque insaisissables, mais plus de phrases différentes, diversement construites, ingénieusement coupées, pleines de sonorités et de rythmes savants. Efforçons-nous d'être des stylistes excellents plutôt que des collectionneurs de termes rares [...]. La nature de cette langue [le français] est d'être claire, logique et nerveuse. Elle ne se laisse pas affaiblir, obscurcir ou corrompre.» (Extrait de la préface à Pierre et Jean, de Guy de Maupassant cité par Louis Timbal-Duclaux 1985, p. 17).

Partie II : L'intelligibilité

A) Définition

Nous empruntons la terminologie mais surtout le modèle d'analyse des textes à Préfontaine et Lecavalier (1990). Nous pouvons ainsi traiter de l'intelligibilité des textes en considérant à la fois l'aspect microstructurel, qui s'intéresse aux composantes de surface du texte (mots et phrases); l'aspect macrostructurel, qui s'intéresse aux liens entre les idées; l'aspect superstructurel, qui s'intéresse aux divisions et subdivisions du texte, ainsi qu'à tous les éléments de mise en évidence des informations.

B) Description microstructurelle

La microstructure du texte correspond aux mots et aux phrases qui sont les porteurs et les organisateurs de la signification du texte. La description microstructurelle sera traitée en considérant les critères suivants: l'indice de lisibilité, les paragraphes, les phrases, les mots, la fréquence de mots en fonction du nombre de caractères et le lexique. Cette description découle des données recueillies suite au traitement du logiciel SATO.

Il existe de nombreuses formules de lisibilité. Richaudeau (1978) compare l'efficacité de trois techniques de mesure de lisibilité, les formules de Flesh, d'Henry et la procédure dite de closure. Il élaborera d'ailleurs une formule fort complexe tenant compte de nombreuses variables dont le nombre de mots-outils indicateurs (c'est-à-dire substantifs), le nombre de répétitions, le nombre de termes d'énumération, les enchâssements, les structures monotones, etc. (Timbal-Duclaux, 1984). Cette formule difficile à appliquer ne se prête pas à des calculs informatisés.

Comme Bourbeau (1988) l'explique, les formules de Flesh, d'Henry ou encore la formule de Gunning sont des mesures de lisibilité plus ou moins complexes. Flesh prend comme facteurs le nombre moyen de mots par phrase et le nombre moyen de syllabe par 100 mots. Henry considère le nombre de mots par phrase, le pourcentage de mots différents d'une liste de mots usuels (la liste de Gougenheim) et le pourcentage de mots dits «variables du dialogue». Gunning lui prend en compte le nombre de mots par phrase et le pourcentage de mots de plus de 3 syllabes. Ces formules ont chacune des portées diverses.

Nous adoptons la mesure de la lisibilité de Gunning (1952) parce que c'est une mesure qui nous permet d'obtenir efficacement un indice de lisibilité grâce au logiciel SATO. Comme il est très long de faire tous les calculs nécessaires pour trouver l'indice de lisibilité d'un texte, nous choisissons de nous conformer aux possibilités offertes par un logiciel accessible pour nous, puisqu'il est disponible au Centre d'analyse de texte par ordinateur (ATO) de l'Université du Québec à Montréal (UQAM).

Selon Gunning, la lisibilité repose essentiellement sur la longueur des phrases et des mots. Sa formule est relativement simple, puisqu'elle ne retient que deux facteurs: le nombre moyen de mots par phrase (L) et le pourcentage de mots de plus de trois syllables (M), multiplié par un certain facteur de pondération. La formule se calcule ainsi : (L + M) x 0,4.

Ce facteur permettait, selon son auteur, d'établir une correspondance entre niveau de difficulté du texte et niveau scolaire pour des américains moyens.

Il est intéressant de savoir à quels niveaux de difficulté correspondent les indices de lisibilité ainsi trouvé; Bourbeau (1988) propose les associations suivantes en fonction des types de textes. Le tableau suivant donne ces informations, nous avons ajouté la correspondance avec les niveaux scolaires à titre informatif :

indice de lisibilité	degré de difficulté	textes-types	niveau scol. (U.S.A. 1952)
6 et -	très facile	bandes dessinées	6e et moins
9-10	moyen	Reader's Digest	9e - 10e
13 et +	difficile	revues spécialisées	13e et plus

C) Description superstructurelle

La superstructure du texte correspond à son organisation formelle, c'est-à-dire à son plan. Elle permet de rendre accessible au lecteur, dès le survol du texte, son aspect organisationnel. Il est important de guider la compréhension du lecteur, quel que soit son niveau d'habileté en lecture (ce qui prend encore plus d'importance pour un lecteur malhabile) pour l'aider à identifier les éléments d'information, regroupés en sections, qui seront abordés successivement dans le texte. Il est fondamental d'activer le processus des préconnaissances dès le début de la lecture.

La superstructure, lorsqu'elle est bien établie, facilite la réalisation des inférences pour comprendre le lien entre les idées développées dans le texte et pour déduire par lui-même la structure logique du texte; le processus cognitif mis en cause se trouve ainsi facilité, même si toutes les difficultés de compréhension ne se trouvent pas pour autant surmontées.

Une façon de clarifier la superstructure est l'utilisation de numérotation des divisions et subdivisions, ce qui facilite le repérage et la compréhension pour le lecteur. Il faut également éviter que le titre soit trop long et trop complexe lexicalement et syntaxiquement. De plus, il peut également être utile qu'une synthèse soit faite sur le sujet du texte à partir de mots-clés qui seraient repris tel quel dans le texte.

D'autres éléments de la superstructure contribuent également à clarifier ou à complexifier le texte; ce peut être le cas de symboles visuels non linguistiques, des images, des graphiques, des tableaux, etc., mais comme l'icônographie est un domaine de recherche en soi, nous ne le traitons pas ici.

D) Description macrostructurelle

La macrostructure d'un texte est la représentation sémantique, l'image que le lecteur se fait du sens d'un texte; elle se divise en trois niveaux d'analyse qui sont, dans l'ordre croissant de leur difficulté: le niveau descriptif, le niveau relationnel et le niveau structurel.

1. Le niveau descriptif

Ce niveau de compréhension demande strictement une analyse de premier niveau de la part du lecteur, c'est-à-dire une analyse qui s'appuie surtout sur l'identification des mots

et de leur sens et sur la reconnaissance des composantes du texte, sans les interrelier entre elles et sans inférer leur compréhension.

Les textes devraient être rédigés de façon à garantir que le lecteur trouve le sens :
- des mots eux-mêmes ou des expressions (expresssions figées et proverbes);
- des symboles visuels non linguistiques (tirets, points gras, etc.);

Au niveau descriptif, le texte devrait être tel que le lecteur puisse également:
- faire le lien entre les référents (pronom et nom);
- comprendre un sigle (MMS...)
- identifier la présence d'un exemple, etc.;
- identifier un titre ou un sous-titre qui contient des mots-clés qui sont repris explicitement dans le texte.

2. Le niveau relationnel (cohérence du texte explicite)

Ce niveau de compréhension peut être défini par la capacité du lecteur de faire la ou les relation(s) qui s'imposent et qui doivent obligatoirement être faites entre les idées du texte. Ce niveau demande une analyse qui permettra de dépasser la complexité du texte.

Les textes devraient être rédigés en donnant toutes les indications nécessaires au lecteur, afin qu'il puisse:

- repérer les définitions présentes dans le texte, en indiquant qu'il s'agit d'une définition;
- identifier les liens logiques explicites (cause, motif, explication, évolution) présents dans le texte;
- identifier une analogie explicite (comparaison, opposition, parallèle);
- identifier une prescription explicite (invitation, ordre, prescription, interdiction, obéissance, transgression, refus, acceptation, souhait, voeu, velléité);
- reconnaître le niveau hiérarchique d'une division par titre et sous-titre; identifier l'idée principale;
- identifier une séquence chronologique;
- reconnaître un classement (mise en ordre rigoureuse à partir de critères explicites, d'ordre temporel, logique ou analogique).

3. Le niveau structurel (cohérence du texte implicite)

Ce niveau de compréhension peut être défini par la capacité du lecteur d'établir des relations non explicites entre des propositions logiques et à se représenter une progression de la pensée incomplètement marquée dans le texte. Ce niveau demande une analyse qui permettra de comprendre les composantes très complexes du texte.

Les textes devraient être rédigés en donnant toutes les indications nécessaires au lecteur, afin qu'il puisse:

i) retrouver, restructurer l'idée, le sens, les objectifs d'un texte

a. inférer l'idée principale implicite;

b. inférer un lien logique implicite (cause, motif, explication, évolution, obligation, interdiction) nécessaire à la compréhension du texte;

c. inférer une analogie implicite (métaphore, etc.) nécessaire à la compréhension du texte;

d. inférer une prescription implicite nécessaire à la compréhension du texte (invitation, ordre, prescription, interdiction, obéissance, transgression, refus, acceptation, souhait, voeu, velléité).

ii) saisir les éléments absents du texte parce qu'ils ne sont pas exprimés par l'auteur.

a. interpréter une incohérence, une ambiguïté ou une erreur dans la structure du texte, par exemple, si les mots du titre ne sont pas repris dans le corps du texte;

b. inférer un changement de division ou de subdivision du texte comme l'absence d'un titre, d'un sous-titre, d'une transition.

Conclusion

Nous avons tenté de cerner et d'expliquer un certains nombre de variables qui semblaient pertinentes dans le processus de communication écrite. Il s'agit d'un processus complexe où il faut tenir compte des objectifs de communication mais surtout des caractéristiques de l'interlocuteur.

Nous avons traité dans la première partie de cette notion de lisibilité. Nous avons également abordé les processus liés à la lecture des textes (processus mécaniques, cognitifs) et les difficultés que ces processus entraînent pour le lecteur. Nous avons tenté surtout de rapporter les suggestions de nombreux auteurs pour rendre les textes plus faciles à lire.

À notre avis ces suggestions ne tiennent pas suffisamment compte de l'organisation formelle du texte ni de sa représentation sémantique pour le lecteur. Le modèle, proposé par Préfontaine et Lecavallier, permet une analyse beaucoup plus globale des textes mais offre surtout des éléments clefs lors de la rédaction de ces textes. Ces auteurs parlent à juste titre d'intelligibilté d'un texte. Ce modèle est exposé en deuxième partie. Enfin des suggestions d'applications sont proposées en dernière partie.

Références

Bourbeau, Nicole, (1988), *C'est pas lisable ! La lisibilité des textes didactiques*, Guide pratique, Sherbrooke, Collège de Sherbrooke, 166p.

Bourque, G. (1989), *Des mesures de lisibilité*, Communication présentée au 57e Congrès de l'ACFAS. Montréal: [inédit].

Chall, J. S. (1958), *Readibility: An appraisal of research and application*, Colombus: Ohio State University Press.

Dale, E. et Chall, J. S. (1948), *A formula for predicting readibility*, Columbus: Bureau of Educational Research, Ohio State University.

Ehrlich, Marie-France, et Tardieu, Hubert, (1985), *Lire, comprendre, mémoriser les textes sur écran vidéo*, Communication et langages, #65, p.91-106.

Fernbach, Nicole, (1990), *La lisibilité dans la rédaction juridique au Québec*, Ottawa, Le Centre de promotion de la lisibilité, Centre Canadien d'information juridique, 128p.

Flesh, R. (1948), *A new readibility yardstick*, Journal of Applied Psychology, 32, 221-233.

Fry, E. B. (1968), *A readibility formula that saves time*, Journal of Reading, 2, 513-516.

Fry, E. B. (1977), *Fry's readibility graph: Clarification, validity and extension to level 17*, Journal of Reading, 20, 242-252.

Gélinas-Chebat, C., Macot, M., Préfontaine, C., et Daoust, F. (1991), *La lisibilité de documents d'information du ministère de la Main d'oeuvre, de la Sécurité du revenu et de la Formation professionnelle*, Avis professionnel présenté au ministère de la Main d'oeuvre, de la Sécurité du revenu et de la Formation professionnelle, Gouvernement du Québec, 50 p.

Gray, W. S., Leary, B. W. (1935), *What makes a book readable*, Chicago: University of Chicago Press.

Gunning, R. (1952), *The technique of clear writing*, New York: McGraw-Hill.

Henry, Georges, (1975), *Comment mesurer la lisibilité*, Paris, Fernand Nathan, Editions Labor, 176p.

Jonassen, D. H. (1982), *The Technology of Text: Principals for Structuring, Designing and Displaying Text*, Vol. 1 Englewood Cliffs, NJ: Educational Technology Publications.

Jonassen, D. H. (1985), *The Technology of Text: Principals for Structuring, Designing and Displaying Text*, Vol. 2 Englewood Cliffs, NJ: Educational Technology Publications.

Lebrun, Nicole et Berthelot, Serge, (1991) *Lisibilité typographique et ordinateur*, ronéotypé, Séminaire du groupe LEO, Département des Sciences de l'éducation, UQAM.

Lively, B. A., Pressey, S. L. (1923), *A method for measuring the «vocabulary burden» of*

textbooks, Educational Administration and Supervision, 9, 389-398.

Mandel, Ladislas, (1988), *L'écriture typographique expression d'une identité culturelle*, Communication et langages, #68, p.86-95.

Mesnager, Jean (1989), *Lisibilité des textes pour enfants : un nouvel outil?*, Communication et langages, #79, p.18-38.

Morin, C., Sallio, P et Kretz, (1982), *Nouvelle étude de lisibilité typographique*, Communication et langages, #54, p.60-76.

Préfontaine, Cl. et Lecavalier, J. (1990), *La mesure de la lisibilité et de l'intelligibilité des textes*, Communication présentée à l'Association pour le développement de la mesure et de l'évaluation en éducation (ADMEE). Montréal, 25-27 octobre.

Racle, Gabriel, (1988), *La lisibilité : quelques aperçus neuropsychologiques*, Communication et langages, #76, p. 20-41.

Richaudeau, François, (1978) *Le texte le plus efficace que je connaisse*, Communication et langages, #37, p.6-26.

Rivaix, Yak, (1984), *L'écriture verticale*, Communication et langages, #59, p.21-34.

Smith, Fr. (1986), *Devenir lecteur*, Paris: Armand Colin - Bourrelier.

Tardif, J., Gaouette, D. 1986a), *Comment faciliter la lecture du lecteur en difficulté?* Vie pédagogique, 44, 12-15.

Tardif, J., Gaouette, D. 1986b), *Comment le lecteur en difficulté devrait-il utiliser ses connaissances antérieures?*, Vie pédagogique, 45, 4-7.

Taylor, W. L. (1953), *Cloze procedure: A new tool for measuring readibility*, Journalism Quarterly, 30, 415-433.

Timbal-Duclaux, Louis, (1984), *La transparence du texte : pour mesurer sa lisibilité*, Communication et langages, #59, p.9-20.

Timbal-Duclaux, Louis, (1985), *Textes «inlisable» et lisible*, Communication et langages, #66, p.13-31.

Vachon, M. (1988), *Effets de la structuration textuelle sur la compréhension en lecture assistée de l'ordinateur*, Université du Québec à Monréal: mémoire de maîtrise en sciences de l'éducation, inédit.

Washburne, C. W., Vogel, M. (1926), *Winnetka graded book list*, Chicago: American Library Association.

Notes

[1] En typographie, il est possible d'avoir le même espace entre les lettres, espacement fixe, quelles que soient les lettres ou au contraire d'avoir des espacements variables. Ainsi lorsque «V» et «A» se voisinent en espacement variable ces lettres seront plus rapprochées et visuellement plus faciles à lire qu'en espacement fixe. Ainsi, en espacement variable, « pour une meilleure lisibilité de VA »
devient en espacement fixe « pour une meilleure lisibilité de V A »

[2] En typographie il est possible de réduire l'espacement entre les lettres pour ajouter plus de caractères sur une même ligne.
Ainsi « pour une meilleure lisibilité »
devient en condensé « pour une meilleure lisibilité »
et en dilaté « p o u r u n e m e i l l e u r e l i s i b i l i t é ».

[3] Selon le Robert 1 (1990), on entend par lettrine, lettre ornée ou non, placée au commencement d'une section de texte.

[4] L'écran linguistique est un terme utilisé pour décrire tout mot ou groupe de mots placé entre deux mots qui devraient être juxtaposés comme par exemple entre le sujet et le verbe. Par exemple : Mon père, fort comme un taureau, a ...»

[5] On peut entendre «distanciation» dans le sens d'éloignement des référents.

Place de la lisibilité en gestion de l'information textuelle

René Lortie et Richard Parent

René Lortie et Richard Parent sont chargés de projet au ministère des Communications du Québec. Richard Parent est aussi coordonnateur du projet DELTA.

L'administration publique québécoise s'est dotée de plusieurs comités interministériels qui s'occupent du développement et de l'implantation des nouvelles technologies de l'information au bénéfice premier du fonctionnement de l'appareil d'État. C'est ainsi que, à titre d'exemple, des comités sur le génie logiciel et la géomatique cherchent à améliorer l'efficacité et la productivité de l'administration, respectivement par la mise au point de méthodes de développement de systèmes et le traitement informatique des données à référence spatiale. Dans le même esprit, un comité de ce type, qui a acquis le nom de DELTA en 1989 mais qui existe de fait depuis 1987, se donne pour mission de contribuer, par le développement et l'implantation des technologies de l'information, à "l'analyse et la gestion des textes et des connaissances" dans l'administration publique. Conçus dans cette optique, les travaux menés sous l'égide du Comité DELTA convergent vers la conception et la mise en place d'une architecture de gestion de l'information textuelle et des documents électroniques divers au sein des ministères et organismes du gouvernement québécois.

Au cours des ans, la composition du Comité DELTA a évolué, mais plusieurs personnes du noyau de départ s'y associent encore, dont Léo Laroche. Ce dernier véhicule au Comité DELTA les préoccupations de lisibilité des textes, dans le cadre général d'analyse et de gestion des textes et des connaissances évoqué plus haut. Le projet SATO-CALIBRAGE donne suite à cette préoccupation pour la lisibilité des textes. Dans la vision globale poursuivie par DELTA, SATO-CALIBRAGE constitue un module avec ses algorithmes et dictionnaires, en conformité avec les formats d'échange de données textuelles et documentaires qui doivent se situer à la base de l'architecture de gestion des documents électroniques. Soutenu par cette vision d'ensemble de DELTA, SATO-CALIBRAGE influe à son tour sur les développements qui sont menés en collaboration avec le Centre d'ATO.

Le projet SATO-CALIBRAGE dispose d'appuis pour devenir partie des outils d'aide à la rédaction, de façon analogue aux correcteurs qui viennent s'adjoindre aux logiciels de traitement de textes. La lisibilité est perçue comme utile pour l'aide à la rédaction ou à la révision de textes. Sa mesure est ressentie comme un besoin dans plusieurs secteurs: dans le monde de la documentation technique, où l'on cherche un "français de base" ou un "français simplifié"; en documentation administrative, à des fins de communication interne à l'organisation et, en particulier, quand un texte est rédigé à l'intention du public.

Initialement, la préoccupation des développeurs de SATO-CALIBRAGE avait trait spécifiquement aux textes de nature pédagogique employés aux niveaux primaire et secondaire. Par la suite, une utilisation sur des textes d'une autre nature en fut d'abord faite à la Direction des communications du ministère de la Main-d'oeuvre, de la Sécurité du revenu et de la Formation professionnelle: on y soumet les textes destinés au public, après leur rédaction par les agents d'information. Son directeur, Luc Poirier, ayant fait une présentation publique de cette utilisation, l'intérêt suscité a conduit à l'organisation d'une démarche de formation offerte aux agents d'information. Cette organisation a été rendue possible par une concertation du Centre d'ATO, de DELTA et du Forum des directeurs des communications du gouvernement du Québec.

Lors d'une rencontre-bilan sur cette activité de formation en mai 1993, la discussion a fait ressortir des éléments à améliorer dans la présentation des résultats de SATO-CALIBRAGE. Mais, ce qui est ressorti comme le problème principal, c'est une extension du calibrage des mots "connus" en fonction du grand public. En transposant la démarche suivie pour le calibrage des textes scolaires, il faudrait donc constituer un corpus tenant compte des clientèles variées de l'administration publique: un niveau général moyen (fin du secondaire?), un public adulte plus ou moins informé, des publics spécialisés, des clientèles concentrant les plus faiblement alphabétisés. Il y aurait d'ailleurs lieu de différencier la langue interne de l'administration et la langue plus générale de communication avec les citoyens et personnes morales, de façon à ajuster les dictionnaires de mots connus et les formules de mesure de la lisibilité en distinguant entre les textes pour usage interne (ex.: notes de service) et ceux à l'intention du public (ex.: collection de dépliants gratuits distribués par Communication-Québec).

Des agents d'information se sont montrés intéressés à collaborer à une telle piste de développement d'un outil mieux calibré pour les besoins de l'administration publique, aisément adaptable aux vocabulaires de domaines de façon étagée, et tenant compte des niveaux de qualification des lecteurs visés dans le grand public.

Dans l'architecture de gestion de l'information textuelle proposée par le Comité DELTA, la lisibilité se rattache aux fonctions de saisie, de rédaction et d'édition des données textuelles. Les autres grandes fonctions concernées par cette architecture de gestion de l'information textuelle sont: la transmission et la messagerie; la gestion, l'emmagasinage et le repérage; l'analyse, la catégorisation et l'annotation.

L'accessibilité aux ressources partagées entre ces diverses fonctions semble être particulièrement tributaire du recours à la normalisation des données textuelles, des formats de documents et des opérations référentielles (du catalogage classique à l'hypertexte) au moyen des normes SGML et HyTime. La mesure de la lisibilité pourra elle aussi prendre appui sur ces formats aptes à représenter plus d'information dans des modèles où le contrôle peut s'exercer le plus possible dans la sémantique de chaque domaine.

C'est du moins la vision, le rêve qui moule les objectifs à court terme, dans un contexte budgétaire âpre. Il faut que la demande prenne de la force pour convaincre aux investissements, et que la coopération bénévole soit vigoureuse.

Deuxième partie

Méthodologie et analyse

SATO-CALIBRAGE, cadre expérimental

François Daoust.

François Daoust est informaticien et chercheur au Centre d'analyse de texte par ordinateur -Cognition et information-. Il est responsable du projet SATO-CALIBRAGE au Centre ATO-CI.

Nous voulons, dans cet article, décrire le cadre expérimental dans lequel se déploie le projet SATO-CALIBRAGE[1]. Ce faisant, nous pourrons indiquer la place respective des divers dispositifs déployés qu'ils soient d'ordre linguistique ou mathématiques[2]. Enfin, nous en profiterons pour décrire le logiciel SATO[3] et sa place centrale dans dans le protocole expérimental.

Hypothèses sur le discours

Comme tout projet en analyse de texte, SATO-CALIBRAGE est fondé sur un certain nombre d'hypothèses sur la nature du discours dont les textes individuels constituent la manifestation. 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 fonction spécifique d'apprentissage.

En d'autres mots, 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 processus d'apprentissage.

À partir de cette hypothèse générale de cohérence, on va vouloir étudier le fonctionnement discursif en observant une collection 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.

En termes métaphoriques, nous pouvons imaginer qu'à un espace de pratiques sociales, correspond un ou plusieurs espaces discursifs. Il faut donc, dans un premier temps, justifier la constitution (cohérence, pertinence sociale) de l'espace discursif que l'on veut étudier. Ensuite, on doit rassembler un corpus qui soit significatif de cet espace. Cela veut dire que l'on vise à choisir des textes individuels qui se répartissent sur l'ensemble de cet espace. Cela veut dire aussi que l'on doit disposer d'une quantité suffisante de textes pour pouvoir dépister des régularités significatives.

La constitution du corpus SATO-CALIBRAGE, décrite dans l'article de Lise Ouellet, traduit bien ces préoccupations. Ainsi, comme notre projet est basé sur une hypothèse de stratification du discours en niveaux d'enseignement, on a pris un soin particulier pour que le corpus soit représentatif des divers niveaux. Cet objectif de représentativité quantitative peut aussi impliquer une réduction volontaire de la variabilité du corpus, et du discours qui le fonde. Par

exemple, nous avons dû exclure certains genres littéraires dont le fonctionnement s'écartait trop de l'ensemble. Sinon, pour tenir compte de cette dimension de façon satisfaisante, il aurait fallu augmenter la taille du corpus et introduire dans notre modèle interprétatif une variable supplémentaire pour tenir compte du genre littéraire.

Conception de modèles interprétatifs

Comme on vient de le voir, la constitution du corpus implique déjà un certain nombre d'hypothèses sur l'existence et les caractéristiques de l'espace discursif que l'on veut analyser. Le protocole expérimental auquel on veut soumettre le corpus implique aussi l'existence d'un modèle interprétatif que l'analyse permettra de corriger et de compléter.

Dans un premier temps, le modèle peut être très sommaire. En fait, il s'agit d'abord d'hypothèses sur la nature des régularités discursives qui seraient associées à une intentionnalité, explicite ou non, des textes. Dans notre cas, le modèle interprétatif était déjà, au départ, assez développé. En effet, le ministère de l'Éducation dispose de grilles permettant d'évaluer la pertinence d'un texte pour un niveau scolaire donné. Parmi les éléments de cette grille, on trouve, par exemple, la longueur du texte, la familiarité avec le vocabulaire, etc.

Notre objectif était donc, dans un premier temps, d'automatiser et de valider certains éléments de cette grille par une expérimentation rigoureuse.

Dans un deuxième temps, l'objectif était d'utiliser le protocole expérimental afin d'enrichir le modèle interprétatif. C'est ainsi que l'on peut non seulement confirmer ou infirmer des modèles interprétatifs existants mais que l'on peut aussi les développer. Par exemple, nous avons émis l'hypothèse que certains termes fonctionnels (adverbes, prépositions, conjonctions, pronoms, etc.) peuvent nous aider à démarquer les textes selon les niveaux scolaires. Plusieurs résultats de l'analyse statistique semblent confirmer cette intuition.

Cela nous conduit naturellement au deuxième temps de la recherche expérimentale, après la constitution du corpus, à savoir la construction d'un dispositif expérimental qui va permettre de confronter le modèle aux données.

Les dispositifs expérimentaux dont nous voulons parler sont des dispositifs computationnels faisant appel à l'ordinateur et au texte électronique. Le texte, dans son format électronique, est simplement une suite de codes de caractères. C'est donc par une suite de calculs sur cette chaîne que l'on doit dépister des procédés discursifs et des effets de sens.

L'objectif du dispositif expérimental est de produire des indices textuels qui mesurent la présence de ces procédés discursifs. Cependant, il est très difficile et peu naturel de concevoir des indices textuels qui se situent directement dans cet espace linéaire que constitue la séquence de caractères. Voilà pourquoi nous travaillons depuis le début des années '70 à développer un appareillage expérimental, un «laboratoire textuel» qui puisse situer le texte dans un espace qui nous soit plus familier. Il est alors plus facile de construire des indices interprétables pour la validation de nos modèles d'interprétation du texte. Ce laboratoire s'appelle SATO.

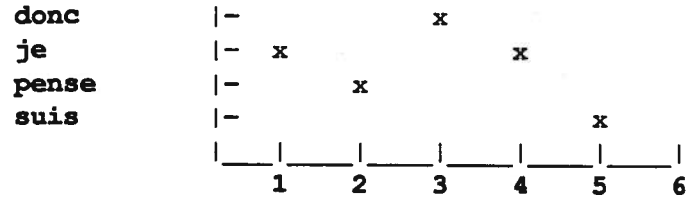
SATO, un outil pour le dépistage d'indices textuels

Si, d'un point de vue matériel, le texte se donne d'abord comme une suite de caractères, du point de vue du lecteur, le texte se présente d'emblée comme une suite de «mots qui font sens». Pourquoi ces mots font-ils sens? D'abord parce qu'ils sont perçus comme des unités, c'est-à-dire des groupes de signes délimités. Ensuite, parce que ces groupes de signes ont une signification à l'intérieur d'une langue dont on a la maîtrise, et en fonction d'un monde connu auquel ils réfèrent.

Donc, au delà de sa définition purement matérielle (cette feuille de papier imprimée), le texte met en oeuvre une dimension implicite, la dimension lexicale et une dimension explicite qui correspond à l'ordre séquentiel de la lecture.

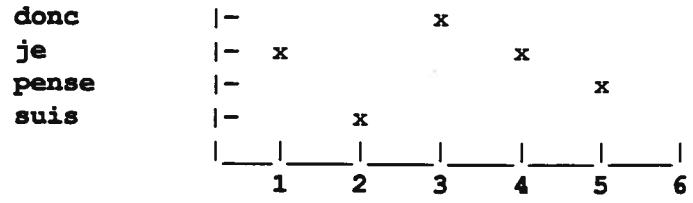
En d'autres mots, SATO situe le texte dans un espace, un plan composé de deux axes. On a d'abord un axe lexical qui dresse la liste du vocabulaire utilisé dans le texte. Ce vocabulaire (les lexèmes) a un sens dans l'univers de la langue et du discours dans lequel s'inscrit le texte. Le deuxième axe représente la linéarité du texte qui se donne en fait comme une suite d'occurrences des lexèmes. De façon abstraite, on pourrait donc voir un texte donné comme un nuage de points tracé sur ce plan.

Représentation d'un texte dans le plan lexique/occurrence



Représentation ASCII:
je pense donc je suis

On pourrait aussi avoir



Représentation ASCII:
je suis donc je pense

Cette représentation du texte dans sa double dimension, lexème et occurrence, est un choix fondamental qui va dicter le modèle informatique de SATO et le type d'opérations logiques mises en oeuvre dans les stratégies d'analyse de texte supportées par le système. C'est le

programme **SATOGEN** (pour SATO-génération) qui permet de réaliser la transformation du texte de sa représentation en termes de chaînes de caractères à sa représentation logique en termes de lexèmes qui occurrent dans leur contexte.

Destiné à soutenir des activités d'analyse, SATO offre la possibilité d'annoter le texte. Le travail d'annotation sur le texte est cette opération matérielle qui permet de marquer par un symbole le dépistage d'une unité cognitive.

Cette unité cognitive peut s'établir sur l'axe lexical. Par exemple, on peut reconnaître que tel lexème appartient à un vocabulaire familier pour les élèves d'un niveau donné. On peut constater qu'il s'agit d'un adverbe, d'un marqueur d'argumentation, etc. Ou, l'unité dépistée peut se définir sur le plan textuel (occurrences). Par exemple, le lexème «le» qui précède le mot «lexème» agit ici comme article. Ou la phrase précédente définit un exemple, etc.

Dans SATO, on utilise le terme de «**propriété**» pour désigner un système catégoriel permettant de marquer des lexèmes ou des occurrences. Par exemple, une propriété «connu» et ses valeurs «oui», «p6», etc. pourrait servir à identifier les lexèmes connus de tous (comme les nombres), et ceux connus par les élèves de sixième année. Une propriété «syntaxe» pourrait permettre d'identifier la fonction grammaticale précise de l'occurrence d'un lexème alors que la propriété «gramr» pourrait servir à définir l'ensemble des fonctions grammaticales possibles du lexème.

En recoupant les systèmes catégoriels des propriétés avec la représentation bi-dimensionnelle du texte, on obtient donc le modèle suivant:

Texte augmenté de propriétés

fréqtot	**gramr**							
1	conjonc	**donc**	|-			x		
2	pron-pers	**je**	|-	x			x	
1	verbe	**pense**	|-		x			
1	verbe	**suis**	|-					x
			|___|___|___|___|___|___|					
				1	**2**	**3**	**4**	**5**	**6**
		édition		maj	nil	cap	nil	nil	
		partie		prém	prém	conn	conc	conc	

Représentation ASCII:
*partie=prém **Je pense** *partie=conn **DONC** *partie=conc **je suis**

Dans cet exemple, nous avons défini sur l'axe lexical deux propriétés ou fonctions catégorielles.

La propriété «fréqtot» est une propriété dont les valeurs sont des entiers et qui contient le nombre

total d'occurrences du lexème dans le corpus de textes.

La propriété «gramr» est une propriété dont les valeurs possibles sont des noms de catégories grammaticales. La propriété «fréqtot» est une propriété pré-définie de SATO alors que «gramr» est une propriété ajoutée.

Sur l'axe textuel, nous avons deux propriétés.

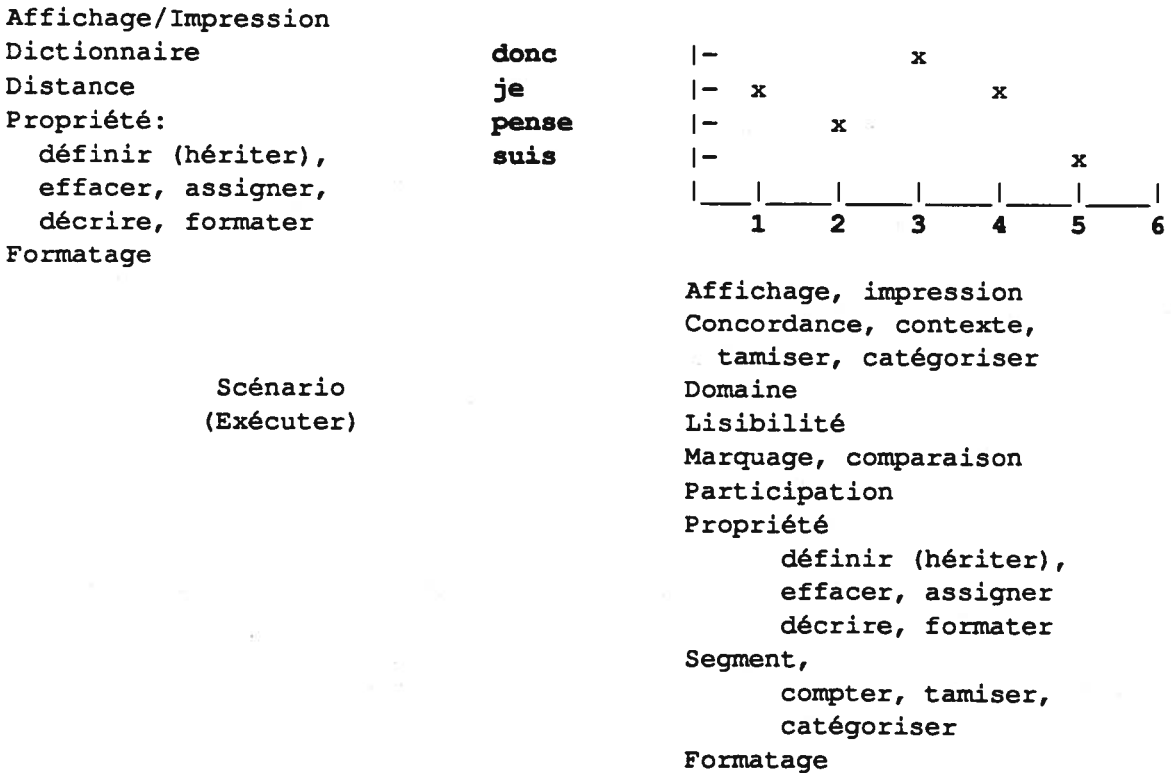
La propriété «édition» est une propriété pré-définie de SATO et dont les valeurs sont des symboles qui définissent des attributs de mise en page de l'occurrence. Par exemple, le symbole «maj» indique que la première occurrence du lexème «je» débute par un «J» majuscule.

La propriété «partie» est une propriété ajoutée qui illustre un classement des occurrences selon qu'ils appartiennent à une prémisse, à un connecteur ou à une conclusion.

Si le programme SATOGEN permet de construire la représentation du texte décrite dans la première partie, c'est le programme **SATOINT** (pour SATO-INTERROGATION) qui est l'outil du dialogue avec le texte. SATOINT est un programme interactif largement paramétrable conçu comme poste de travail du lecteur-analyste.

Rapportées sur le plan lexique-occurrences de SATO, les opérations que permet d'effectuer SATO (module SATOINT) se distribuent selon le schéma suivant:

<u>Opérations sur le plan lexique/occurrence</u>



On remarquera que certaines opérations sont disponibles tant sur le plan du lexique que sur celui des occurrences.

Il s'agit en particulier des opérations d'**affichage** et d'**impression**, qui sont aussi des opérations de **sélection** par l'utilisation du mécanisme des patrons de fouille. Les patrons de fouille permettent de désigner des lexèmes ou des occurrences par la concaténation de filtres portant sur leurs caractères ou leurs valeurs de propriété. Voici quelques exemples:

Exemples de patrons de fouille

parle	le mot «parle»;
parle$	tous les mots débutant par «parle»; «$» est un opérateur de troncature à droite;
p|ent	tous les mots débutant par «p» et se terminant par «ent»; «|» est un opérateur représentant une chaîne quelconque de caractères;
p_rle	tous les mots débutant par «p» suivi d'un caractère quelconque (opérateur «_») et se terminant par «rle» comme «parle» ou «perle»;
parl(e,ent,ure)	«parle», «parlent», «parlure»; ici, les parenthèses et la virgule permettent de définir des chaînes alternatives;
|ent*fréqtot=5,>5	tous les mots se terminant par «ent» et dont la propriété fréqtot (fréquence totale introduite par l'astérisque) est égale à 5 ou est plus grande que 5;
ab$*alphabet=(fr,en)	tous les mots débutant par «ab» et provenant des alphabets «fr» (français) ou «en» (english); on voit que le patron alternatif s'applique ici aux valeurs de la propriété alphabet;
$*alphabet~fr*fréqtot=1	tous les mots qui ne sont pas (opérateur différent «~») en français et dont la fréquence totale est 1.

Le deuxième ensemble d'opérations qui est disponible sur les deux axes de notre plan concerne la définition et l'exploitation des systèmes de **propriété**.

On peut **définir** ou **effacer** une propriété. Une des modalités intéressantes de l'opération de définition est l'**héritage**. Par exemple, si on a défini une propriété «gramr» sur le lexique, on pourrait définir une propriété «syntaxe» sur le texte qui «hérite», au moment de sa création, des valeurs de la propriété «gramr». Il s'agit de propriétés distinctes qui pourront par la suite être modifiées de façon distincte. En particulier, on pourra modifier «syntaxe» pour chacune des occurrences du lexème. Comme on le verra dans l'article sur le dispositif linguistique, on se sert

de ce dispositif pour la levée des ambiguités catégorielles sur le verbe.

Une autre opération commune aux systèmes des propriétés est l'**assignation** de valeurs. Cette opération se réalise au moyen d'une commande d'affectation ou par manipulation directe: on pointe l'objet au moyen des curseurs ou de la souris et on assigne des valeurs à l'une ou l'autre de ses propriétes. On peut aussi assigner des valeurs en faisant appel au mécanisme de sélection des patrons de fouille.

On peut aussi **décrire** une propriété. Cette opération fait appel aux techniques de la statistique descriptive et permet de dresser un portrait de l'utilisation des valeurs de la propriété.

Finalement, on peut **formater** une propriété, c'est-à-dire définir son format d'affichage: nombre de colonnes pour afficher le propriété, sa couleur, etc.

Les prescriptions de formatage sont aussi disponibles directement au niveau du lexique et des occurrences. Ainsi, on peut formater le lexique pour déterminer les propriétés que l'on veut visualiser. Il est en de même des propriétés textuelles dans le cas de l'affichage ou de l'impression du texte. En fait, SATO permet de modifier librement un très grand nombre de paramètres de visualisation.

Un type d'opérations spécifiques à l'axe lexical concerne la manipulation de **dictionnaires**. Pour SATO, un dictionnaire est un fichier externe, une base de données, qui permet d'associer des valeurs de propriété à des chaînes de caractères qui représentent normalement des formes lexicales. SATO fournit un ensemble de dispositifs pour créer, consulter et modifier des dictionnaires. On peut aussi les fouiller avec une syntaxe de patrons comme on le fait pour le lexique et le texte.

Finalement, SATO fournit des analyseurs lexicométriques. En particulier, **DISTANCE** permet de mesurer la différence de vocabulaire entre deux parties du texte. Cet analyseur permet aussi d'indiquer quels sont les lexèmes, ou les valeurs de propriété de ces lexèmes qui distinguent les deux parties du texte soumises à la mesure.

La première opération qui concerne spécifiquement l'axe des occurrences a trait au repérage de segments textuels, c'est-à-dire des portions de texte possédant diverses caractéristiques. C'est le cas en particulier de la **CONCORDANCE** qui permet de repérer des passages qui contiennent un ou plusieurs mots avec divers types de contraintes. On peut aussi se servir de la concordance pour réaliser une catégorisation en contexte automatique (cf. l'article *Le dispositif linguistique*).

Bien sûr, on pourra afficher ou imprimer les passages repérés, en soulignant les mots dépistés en position de contrainte. Cette édition des contextes est accompagnée de références de pagination aussi précises que l'on désire. On peut aussi **tamiser** les contextes, c'est-à-dire dresser le lexique d'une classe quelconque des mots qui apparaissent dans la concordance.

Outre les concordances, il existe en SATO un deuxième mécanisme de répérage de contextes. Il s'agit de la fonction de **SEGMENTATION** qui a pour fonction de partitionner le texte en segments. Par exemple, on pourrait découper le texte en documents, en paragraphes, en phrases possédant une certaine longueur, ou en segments de longueur fixe.

On peut afficher ou imprimer un ou plusieurs des segments ainsi repérés. On peut aussi

compter des classes de mots décrits par des patrons à l'intérieur de chacun de ces segments. On peut calculer divers indices de répartition et de dispersion des mots comptés. Par exemple, on pourrait segmenter le texte en phrases et compter le nombre de verbes conjugués par phrase. On pourrait segmenter le corpus en documents et compter le nombre d'occurrences de lexèmes susceptibles d'être des descripteurs de contenu. Divers indices nous indiqueront lesquels de ces lexèmes semblent discriminants. On trouvera des exemples d'utilisation de cette commande dans l'article sur la préparation des données pour l'analyse statistique.

En faisant appel aux patrons de fouille ou au repérage de segments et concordances, on peut définir un **DOMAINE** textuel. Le domaine est une restriction sur l'axe des occurrences de SATO. La commande permet aussi de dresser un sous lexique qui correpond au domaine.

Finalement, on dispose dans SATO de quelques analyseurs qui concernent directement l'axe des occurrences. Il s'agit en particulier de LISIBILITÉ, MARQUER et PARTICIPATION.

LISIBILITE fournit divers indices de difficulté/facilité de lecture du texte.

PARTICIPATION permet d'évaluer la part relative d'une classe quelconque de mots dans un ensemble de sous-textes.

MARQUER permet de marquer les différences entre deux segments textuels quasi-identiques.

Comme on peut le constater à la lecture des sections précédentes, SATO fournit très peu d'analyseurs complets. SATO fournit surtout des fonctions et opérations dont la combinaison permet à l'utilisateur de construire ses propres analyseurs. Ceux-ci vont prendre la forme de scénarios que l'on fait exécuter. Sur le plan lexèmes-occurrences de la représentation SATO, nous avons placé le scénario à la jonction des deux axes. En effet, le plus souvent, les scénarios déploient des stratégies faisant appel à la combinaison de fonctions qui agissent sur l'un ou l'autre des deux axes.

Les scénarios prennent la forme de fichiers ASCII composés de séquences de commandes SATO. Ils sont, le plus souvent, composés à partir d'extraits du journal associé à une session de travail dans SATOINT. Une fois que l'on a mis au point des stratégies en mode interactif, on reprend ces stratégies et on en fait des analyseurs.

Les scénarios de commandes ont un double statut. D'un point de vue technique d'abord, ce sont des programmes permettant de reproduire des stratégies d'analyse et de les appliquer dans un cadre de production. Mais ce sont aussi des objets scientifiques qui sont la matérialisation d'un savoir descriptif ou analytique.

Voilà pourquoi le logiciel SATO constitue un outil central pour la construction de dispositifs expérimentaux appliqués à des corpus textuels. Il correspond tout à fait à l'idée que l'on se fait d'un laboratoire, à savoir un ensemble d'outils que l'on peut combiner à loisir. De plus, SATO combine à la fois les avantages d'un système interactif et d'un système à base de commandes. Le mode interactif permet de naviguer rapidement à l'intérieur du matériau textuel. En cela il est un outil d'exploration et de découverte. D'autre part, comme ce cheminement laisse une trace dans le journal sous la forme de commandes exécutables, SATO est un outil permettant de construire des dispositifs reproductibles qui pourront prendre place à l'intérieur de protocoles expérimentaux serrés.

Analyse des résultats

Les dispositifs expérimentaux qui seront élaborés dans SATO vont généralement prendre la forme de scénarios. C'est en déployant ces scénarios que l'on produit les indices sur le texte. Ces indices peuvent être interprétés directement par le chercheur dans le cadre de son modèle. Souvent, cependant, les indices, pour être interprétés correctement, vont devoir être examinés à travers une modélisation mathématique.

Par exemple, d'après notre modèle interprétatif, nous considérons qu'un texte est facile à lire s'il contient un nombre limité de mots inconnus du lecteur. La familiarité du vocabulaire va pouvoir être dépistée dans le modèle SATO en confrontant le lexique du texte (axe lexical) avec un dictionnaire de mots connus validés par les enseignants. Cette consultation du dictionnaire va permettre d'annoter le lexique du texte en affectant à la propriété «connu» la valeur appropriée. Par la suite, il sera facile de demander à SATO de nous fournir la proportion d'occurences des lexèmes connus.

Aussi, en parcourant l'axe des occurrences, on pourra utiliser un dispositif comme les concordances pour identifier, par exemple, les phrases qui contiennent plusieurs mots inconnus. On pourra aussi faire la proportion de ces phrases par rapport au nombre total de phrase dans le texte. On a ainsi un deuxième indice visant à mesurer la difficulté associée à l'emploi de mots inconnus.

Résumons-nous. Nous utilisons SATO pour construire des indices qui visent à mesurer des comportements discursifs. On trouvera plusieurs exemples de ces indices dans les divers articles sur le dispositif linguistique. Le comportement de ces indices est interprétable dans le cadre de notre modèle d'interprétation du texte. Nous appliquons ces indices à un corpus de textes représentatif. Cela veut dire que les scénarios SATO sont appliqués sur chacun des textes ou sur des ensembles de textes selon le cas. Il nous reste maintenant à juger des résultats obtenus. Les résultats bruts sont sans doute intéressants d'un point de vue qualitatif. Cependant si on veut vraiment savoir ce qu'ils nous révèlent, on a avantage à les interpréter d'abord d'un point de vue 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 qui vont nous faire découvrir dans l'espace des données des régularités appréhendées ou insoupçonnées.

Finalement, il faudra évaluer les résultats observés dans l'espace mathématique des mesures pour leur donner un sens à l'intérieur de notre modèle interprétatif.

Les modèles mathématiques que l'on peut utiliser pour l'analyse des résultats issus des indices textuels ne sont pas très différents de ceux qui sont employés dans d'autres sciences. Néammoins, comme c'est généralement le cas en sciences appliquées, on doit bien examiner les conditions d'application des modèles et faire les ajustements nécessaires.

On trouvera dans les articles suivants une présentation élaborée du dispositif mathématique qui a été utilisé dans le projet SATO-CALIBRAGE.

L'analyse des résultats issus de l'expérimentation va nous conduire à perfectionner notre

protocole expérimental. Ainsi, on va pouvoir juger de la pertinence de nos indices. La mécanique du fonctionnement discursif est souvent étonnante. Souvent, comme lecteur humain, on saisit, ou on croit saisir, le sens du texte alors que son fonctionnement interne reste implicite et souvent caché.

A cet égard, la lecture par ordinateur est impitoyable. La construction des indices textuels oblige à débusquer le fonctionnement interne du texte. Celui-ci respecte les règles générales de la langue mais ces règles ne permettent pas de rendre compte du niveau proprement discursif.

La mise au point d'un protocole d'analyse de texte constitue donc véritablement une entreprise de description du discours. Le protocole n'est donc pas simplement utilitaire. Il fait partie intégrante du processus général d'analyse et devient partie prenante du modèle interprétatif.

La méthodologie que nous venons d'illustrer permet donc à la fois d'unir et de distinguer les différents niveaux d'une science du texte.

Distinguer d'abord parce que chaque niveau possède son autonomie et ses règles propres. Un premier niveau concerne la théorie du discours et possède une dimension sociologique évidente. Le discours a une fonction sociale qui permet d'en éclairer la finalité et les règles générales de fonctionnement.

D'un autre côté, les textes individuels possèdent aussi leurs propres règles de fonctionnement. Ils mobilisent de façon particulière des procédés linguistiques, stylistiques, narratifs, argumentatifs, etc. Aussi, en analyse de texte par ordinateur, nous développons des outils pour dépister ces procédés. Il s'agit donc là d'un deuxième niveau qui possède ses règles propres et ses modèles.

Finalement, le niveau mathématique est aussi essentiel pour manipuler la complexité des indices textuels et pour nous faire découvrir les régularités et les singularités textuelles.

En somme, c'est l'ensemble de ces niveaux qui doit être mobilisé dans une approche expérimentale. On ne s'étonnera donc pas de voir la diversité de formation des rédacteurs ayant collaboré à la confection du présent Cahier de recherche.

Notes

[1] On trouvera quelques-unes des idées développées dans cet article dans un chapitre de l'ouvrage collectif *Le droit saisi par l'ordinateur*, publié sous la direction de Claude Thomasset, René Côté et Danièle Bourcier (Éditions Yvon Blais, Cowansville, Québec, 1993). Ce chapitre, intitulé *La méthode expérimentale en analyse de texte par ordinateur* (François Daoust, p. 441-448) s'intéresse plus particulièrement aux textes de nature juridique.

[2] La méthodologie utilisée dans le projet SATO-CALIBRAGE partage plusieurs points en commun avec le projet de prototype de système expert pour l'aide à l'analyse des jugements, présenté dans la deuxième parution des Cahiers du Centre ATO-CI: Suzanne Bertrand-Gastaldy, François Daoust, Jean-Guy Meunier, Gracia Pagola et Louis-Claude Paquin, présenté au congrès de l'AQDIJ en octobre 92.

[3] SATO, Système d'analyse de texte par ordinateur, Manuel de références, François Daoust, Centre d'ATO, janvier 1992.

Description du corpus de textes

Lise Ouellet

Lise Ouellet est responsable de l'évaluation du français au primaire. Elle travaille à la Direction de la formation générale des jeunes du ministère de l'Éducation du Québec.

Création du corpus

Dans un premier temps, nous avons cherché des textes qui étaient déjà sur support informatique. Ainsi, les responsables de la Banque d'instruments de mesure (BIM) nous ont fait parvenir des disquettes sur lesquelles on trouvait des épreuves de lecture de la première à la sixième année du primaire, épreuves qui avaient été élaborées durant les années précédentes. La Direction du développement de l'évaluation du ministère de l'Éducation a aussi fait parvenir des textes utilisés dans les épreuves de lecture de troisième et sixième année du primaire, et de troisième et cinquième année du secondaire. Un de nos collaborateurs avait sur disquette des textes de première à sixième année qui venaient d'un matériel didactique approuvé. Lorsque le groupe des usagers a été formé, chaque personne était invitée à enrichir le corpus par des textes qui étaient utilisés dans son milieu. L'équipe de coordination visait à constituer une banque d'au moins cinquante textes par classe.

Pour chaque texte, nous avions donc une information sur la classe où le texte était utilisé et sa provenance; dans certains cas, nous avions aussi le type de texte, cette information a été mise en commentaire.

Validation du corpus

Dans un deuxième temps, nous avons vérifié le corpus pour nous assurer que, d'une part, il n'y avait pas de texte en double et que, d'autre part, les textes étaient valables pour les travaux que nous poursuivions. Des données statistiques sur une quinzaine de variables ont servi de guide pour faire l'épuration. Ainsi, 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. Ces formes de textes, aux caractéristiques syntaxiques assez particulières, étaient marginales dans le corpus et nous avons donc décidé de nous concentrer sur des textes homogènes, les plus couramment utilisés dans l'enseignement du français au Québec.

Cette validation a permis d'éliminer un certain nombre de textes, réduisant par le fait même l'objectif de cinquante textes par classe. En examinant les types de textes présents pour chaque classe et la provenance des textes, nous avons complété le corpus en choisissant des textes qui venaient de divers manuels de lecture récemment édités.

Description du corpus actuel

Le corpus actuel comprend 679 textes répartis de la façon suivante :

Ordre d'enseignement	Nombre de textes
Primaire Secondaire	400 279
TOTAL	679

Corpus du primaire

Le programme de français du primaire stipule que les élèves doivent lire des textes variés : à caractère informatif, imaginaire, expressif, incitatif, poétique et ludique. Les textes qui composent le corpus du primaire représentent cette diversité; toutefois, les proportions varient d'un type de texte à l'autre. Il faut rappeler que notre corpus a été créé à la fois avec des textes issus du matériel didactique et d'autres utilisés dans des situations d'évaluation. Or, en évaluation, les textes imaginaires, poétiques et ludiques ont été rarement retenus ces dernières années. On trouvera dans le tableau ci-dessous la description du corpus du primaire.

Tableau I : Description du corpus du PRIMAIRE

Types de textes / Classes	Informatif	Imaginaire	Expressif	Incitatif	Poétique ludique	Nombre de textes
Première année	26%	22%	18%	21%	13%	65
Deuxième année	25%	41%	18%	10%	6%	82
Troisième année	48%	16%	13%	14%	9%	63
Quatrième année	56%	11%	10%	15%	8%	60
Cinquième année	42%	23%	12%	10%	13%	60
Sixième année	87%	6%	6%	1%	---	70
Pourcentage moyen	**47%**	**20%**	**13%**	**12%**	**8%**	**400**

Corpus du secondaire

Le programme de français du secondaire précise, pour chaque classe, les types de discours qui doivent être lus par les élèves. Les enseignantes et les enseignants respectent donc ces types de textes pour l'enseignement et pour l'évaluation. On trouvera dans le tableau ci-dessous la composition du corpus du secondaire. On se rappellera que certaines formes de textes ont été exclues parce qu'elles ne convenaient pas au type de travail amorcé pour établir un indice qui rendra compte de la difficulté relative des textes du primaire au secondaire.

Tableau 2 : Description du corpus du SECONDAIRE

Types de textes / Classes	Informatif	Littéraire	Expressif	Incitatif	Nombre de textes
Première secondaire	Texte informatif	Récit		Mode d'emploi	
	29%	59%	5%	7%	63
Deuxième secondaire	Article de revue	Récit		Règles de jeu	
	29%	58%	---	13%	43
Troisième secondaire	Article de revue	Courte légende			
	72%	28%	---	---	49
Quatrième secondaire	Article analytique	Nouvelle littéraire	Lettre d'opinion		
	49%	44%	7%	---	57
Cinquième secondaire	Article critique	Extrait de romans			
	86%	14%	---	---	67
Pourcentage moyen	53%	41%	2%	4%	279

Utilisation du corpus

Le corpus qui a été constitué dans le cadre des travaux de développement du prototype

SATO-CALIBRAGE est utilisé à des fins de recherche seulement. Il a servi, dans un premier temps, à faire une liste de mots sur lesquels des enseignantes et des enseignants de sixième année ont fourni un jugement de familiarité; ces mots formeront le lexique des mots connus des élèves. Dans un deuxième temps, le corpus a servi de base aux analyses statistiques en vue de composer un indice SATO qui fournit des indications quant à la difficulté d'un texte; cet indice permet de situer un texte sur le continuum scolaire du primaire au secondaire, mais aussi de savoir quelles sont les variables qui rendent un texte plus facile ou plus difficile.

Le dispositif linguistique

François Daoust, Fernande Dupuis.

François Daoust est informaticien et chercheur au Centre d'analyse de texte par ordinateur -- Cognition et information. Il est responsable du projet SATO-CALIBRAGE au Centre ATO-CI. Fernande Dupuis est professeure associée au département de linguistique de l'UQAM. Elle est aussi chercheure au Centre ATO-CI.

Nous désignons, par dispositif linguistique, l'ensemble des ressources linguistiques déployées à l'intérieur du prototype SATO-CALIBRAGE. Ces ressources sont de trois ordres. Il y a d'abord des bases de données lexicales. On a ensuite des procédures permettant de repérer les noms propres et d'identifier en contexte les verbes conjugués. Enfin, s'appuyant sur les dispositifs précédents, on a des procédures permettant de dresser une typologie des phrases susceptibles d'être plus difficiles à lire.

Comme indiqué dans l'article SATO-CALIBRAGE, cadre expérimental, les divers dispositifs linguistiques prennent la forme de scénarios de commandes SATO.

Les bases de données lexicales

Les bases de données lexicales prennent la forme de dictionnaires SATO. Ce sont des fichiers externes 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 informations se trouvant dans le dictionnaire.

La première base de données lexicales que nous utilisons dans SATO-CALIBRAGE contient la catégorie grammaticale hors contexte. Elle contient plus d'un demi million de lexèmes[1]. La philosophie qui a guidé la construction de cette base de données, appelée simplement **BDL**, est de fournir à la communauté des chercheurs un dictionnaire de base pour l'analyse de texte par ordinateur. La BDL profite donc à plusieurs projets, dont SATO-CALIBRAGE, en même temps qu'elle bénéficie de l'apport des divers projets pour son entretien.

La catégorisation grammaticale que fournit la BDL est orientée vers la grammaire d'usage plutôt que vers la résolution de problèmes de parsage[2]. Ce choix est justifié par le fait que nous visons la communauté des chercheurs, y compris une majorité de non-linguistes. Par ailleurs, nous entendons nous inspirer de méthodologies éprouvées, comme celles du LADL[3] en France, pour assurer dans le futur un meilleur entretien et une meilleure validation de la BDL. Cela implique, entre autres, de séparer la BDL en formes simples, et en formes «marquées» calculées à partir des règles de dérivation et de conjugaison. Ces règles sont connues. Cependant l'effort pour les colliger, en faire des procédures SATO reproductibles, modifiables et publiques, est loin d'être négligeable.

Le deuxième dictionnaire que nous utilisons a été développé à l'intérieur du projet. Il s'agit du **dictionnaire des mots connus par les élèves de sixième année**. Comme l'indique Lise

Ouellet dans l'article *Description du corpus textuel*, ce dictionnaire 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 reprise pour tenir compte de 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. Ont été acceptés comme connus les lexèmes jugés tels par au moins quatre enseignants. La consigne donnée aux enseignants demandait de considérer connu un mot que les trois quarts au moins des élèves connaissent à l'oral. Plusieurs enseigants ont consulté leurs élèves.

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 des flexions régulières manquantes aux mots connus. Ce travail reste à compléter pour les conjugaisons les plus simples des verbes connus. Une analyse plus poussée des réponses des enseignants permettrait aussi de voir si certaines formes dérivées d'une même racine posent des difficultés particulières.

Le repérage assisté des noms propres

Divers types de lexèmes ont été exclus du dictionnaire des mots connus. On considère en effet que les nombres, plusieurs formes fonctionnelles (articles, pronoms, conjonctions et prépositions usuels) et les noms propres devraient être considérés connus. Les nombres peuvent être identifiés par les patrons morphologiques de SATO. Les formes fonctionnelles sont identifiées par la BDL. Il reste les noms propres.

Pour faciliter l'identification des noms propres, nous utilisons un scénario qui permet de dresser une liste de candidats. Cette liste peut être validée hors contexte ou en contexte s'il y a lieu. Le dispositif repère les mots débutant par une majuscule et qui ne sont pas des formes fonctionnelles. Il fournit le nombre de fois où le mot apparaît, le nombre de fois où il débute par une majuscule et le nombre de fois où il débute par une majuscule et n'est pas précédé d'une ponctuation forte.

Pour illustrer comment une telle tâche peut être facilement programmée à l'intérieur de SATO, voici le scénario complet du dispositif. Les lignes débutant par un astérisque sont des lignes de commentaires.

Scénario NOMP

```
* !DESCRIPTION : Identification des noms propres (avec assistance)
* !DATE        : Janvier 1993
* !AUTEUR      : François Daoust, Centre ATO-CI, UQAM
* !NOTE        : On doit procéder à la catégorisation syntaxique du
* !              lexique avant d'appeler cette procédure (propriété gramr)


* La commande suivante permet d'associer à l'abréviation 1
* l'ensemble des mots qui satisfont aux trois critères suivants:
```

```
* tous les caractères sont admissibles: (opérateur «$»);
* la propriété «édition» indique qu'ils débutent par une majuscule
*      ou sont en lettres capitales;
* la propriété «gramr» indiquent qu'ils ne sont pas (opérateur «~»)
*      des abréviations, des adverbes, des articles, des déterminants,
*      des conjonctions, des noms propres déjà identifiés,
*      des prépositions, des pronoms ou des codes (résidus)
* Rem.  Le double astérisque en fin de ligne indique que la
* commande se poursuit sur la ligne suivante
*
Abréviation 1 $*édition=(maj,cap)*gramr~**
(abréviation,adv,art,dét$,conjonction,nomp,pronom,prép,p\_$,résidu)

* Dans les concordances qui suivent, on va utiliser un contexte numérique
* d'un mot avant et après le mot pôle (ou central) de la concordance
*
Contexte numérique de 1

* On crée la propriété «maj» qui va contenir le nombre de fois
* où apparaît un des mots désignés par l'abréviation 1
*
Propriété créer maj entière pour lexique

* Par la concordance, on repère les contextes d'apparition de
* ces mots et on augmente de 1 la propriété «maj» du lexème associé
* désigné par l'abréviation 1
*
Concordance stricte *1*maj:+1

* La propriété «libre» va contenir le nombre de fois où un mot
* désigné par l'abréviation 1 n'est pas précédé d'une ponctuation forte
*
Propriété créer libre entière pour lexique

* Par la concordance, on repère les contextes d'apparition de
* ces mots et on augmente de 1 la propriété «libre» du lexème associé
* à la condition que le mot qui précède ne soit pas (opérateur «*~»)
* une des ponctuations désignées
*
Concordance stricte (.,?,:,...,-,«,!,;)*~ *1*libre:+1

* On crée une propriété «décision» sur le lexique dont les valeurs sont:
* nil    : pour indiquer qu'il ne s'agit pas d'un nom propre
* np     : pour nom propre
* voir   : pour visualiser les contextes d'apparition du lexème
*
Propriété créer décision symbolique pour lexique np voir

* On va suggérer de citer comme nom propre les mots désignés par
* l'abréviation 1 et qui apparaissent en «position libre» dans la phrase
*
Valeur décision np pour $*libre>0
```

```
* On affiche les mots susceptibles d'être des noms propres. La propriété
* décision permet d'agir sur ces candidats. «np» va permettre leur
* transformation automatique en nom propre. «nil» laisse les mots en
* majuscules de ponctuation. «voir» permet de souligner dans le texte
* les mots dont le statut est à confirmer. Cette confirmation se fait
* par la catégorisation de la propriété édition (valeur +np).
* Pour ce faire, on utilise la manipulation directe en pointant
* le lexème que l'on veut catégoriser.

* On définit des touches pour faciliter le travail de
* catégorisation en associant une lettre à chaque décision
*
Touche n décision valeur nil
Touche p décision valeur np
Touche v décision valeur voir

* On choisit les propriétés à afficher avec le lexique
*
Format lexique maj libre décision

* On passe en affichage pas à pas et on informe l'utilisateur de la
* tâche à accomplir
*
Format affichage normal
Écrire message Liste des lexèmes susceptibles d'être des noms propres
Écrire message La touche «n» force la décision à «nil» : pas un nom propre
Écrire message La touche «p» force la décision à «np»  : nom propre
Écrire message La touche «v» force la décision à «voir»: voir les contextes

* On affiche le lexique en laissant à l'utilisateur la possibilité
* d'utiliser les touches pour modifier la valeur de la propriété «décision»
*
Écrire lexique $*maj>0 Tri alphabet

* On supprime l'affichage pour donner suite aux décisions
*
Format affichage expert

* On change la catégorie grammaticale des mots désignés
* comme noms propres et on ajoute à la propriété «édition»
* la valeur «np» afin que SATO cite la majuscule
*
Valeur gramr + nomp pour $*décision=np
Valeur édit + np pour $*décision=np*édition=(maj,cap)

* On définit des touches pour faciliter le travail de
* catégorisation en associant une lettre à chaque décision
* qui, cette fois, sera prise sur l'occurrence plutôt que
* sur le lexème
*
```

```
Touche n édition - valeur np
Touche p édition + valeur np

* On définit un contexte de phrase pour la concordance
*
Contexte délimité de (.,;,:,...,!,?) exclus à (.,;,:,...,!,?) inclus

* On repère toutes les phrases où apparaît un mot que l'on veut
* voir, s'il est en majuscules ou capitales;
* Rem. l'opérateur «*@» indique que l'on veut trier les concordances
* selon l'ordre alphabétique des mots
*
Concordance libre $*décision=voir*édition=(maj,cap)*@

* On rétablit l'affichage normal, on informe l'utilisateur de la
* tâche à accomplir et on affiche les concordances
*
Format affichage normal
Écrire message Liste des contextes à valider; les noms propres potentiels **
sont soulignés; les noms propres retenus doivent être confirmés par la **
touche «p»
Écrire message La touche «n» force la décision à «nil» : pas un nom propre
Écrire message La touche «p» force la décision à «np»  : nom propre
*
Écrire concordance *
```

Le dépistage des locutions fonctionnelles

Comme indiqué dans l'article sur le cadre expérimental, une des hypothèses que nous voulions examiner dans SATO concerne l'apport, en termes de facilité ou de difficulté de lecture, de certaines formes fonctionnelles. On entend ici, par forme fonctionnelle, les adverbes, les articles, les déterminants, les conjonctions, les prépositions, et les pronoms.

Comme plusieurs de ces formes sont des locutions, nous avons aussi bâti un dispositif pour repérer ces locutions. Pour le moment, les locutions sont bloquées pour être lexicalisées dans une deuxième version du texte. L'inconvénient de cette méthode est qu'elle modifie la structure de surface du texte (longueur des phrases et des mots). Dans la prochaine version du prototype, l'étape du blocage sera éliminée. On doit aussi noter que le problème du repérage des locutions fonctionnelles est un problème difficile à régler[4]. On s'est donc concentré sur les cas les moins ambigus. La liste comprend tout de même plus de 500 locutions comportant chacune plusieurs variantes, par exemple la locution prépositive «sous prétexte (de,d',du,des)» ou la locution conjonctive «tant et si bien (que,qu')».

La dépistage en contexte des verbes conjugués

Suite à la projection d'un dictionnaire sur le lexique d'un texte, des unités graphiques, souvent parmi les plus fréquentes, reçoivent plus d'une catégorie syntaxique. Ces catégories renvoient à la nature grammaticale de chacun des lexèmes. Lorsque l'on examine les phrases dans lesquelles ces lexèmes sont employés, on peut voir que la syntaxe, notamment, permet de préciser laquelle des catégories grammaticales est active. Par exemple, dans la phrase *la femelle construit habituellement son nid sous un tas de larges branches*, le mot *branches*, qui possède les catégories nom et verbe, ne peut pas être un verbe conjugué à cause de la présence à sa gauche de la préposition *de*. On peut donc éliminer certaines ambiguïtés catégorielles en s'appuyant sur les catégories grammaticales des mots qui précèdent ou suivent la catégorie ambiguë. Dans SATO-CALIBRAGE, en particulier, nous voulons repérer les véritables verbes conjugués puisque nous allons les utiliser pour dénombrer le nombre de propositions par phrase.

Le modèle de traitement

Notre traitement de l'ambiguïté des catégories Nom/Verbe s'apparente à ce que Silberztein (1989, pp. 137-140), appelle des «grammaires locales». Le modèle de traitement que nous allons illustrer ici pour un problème syntaxique pourrait être repris pour toute autre ambiguïté catégorielle susceptible d'être levée à partir de l'examen des contextes immédiats.

Avec SATO, il s'agit de décrire, sous la forme de patrons de fouille (concordances SATO), les contextes désambiguïsants. Du même coup, on associe aux patrons de fouille des actions de désambiguïsation catégorielle. La solution développée ici comporte deux étapes incorporées dans une seule procédure : l'élagage ou la suppression des catégories grammaticales «indésirables» («émondage» dans Habert 1990, pp.179-183) et l'ajout d'une propriété («règle») permettant de visualiser le résultat de la règle et de retracer le contexte de son application. Cette procédure de trace a été utilisée dans la phase de validation de l'algorithme. Dans le prototype final, nous procédons directement à la catégorisation.

Rappelons qu'avant de procéder à la désambiguïsation, on doit d'abord procéder au dépistage des expressions figées et à la catégorisation grammaticale.

On s'appuie sur un certain nombre de caractéristiques du nom commun et du verbe pour lever l'ambiguïté. Le nom commun, tête du syntagme nominal (SN), doit généralement être déterminé; mais le déterminant (à quelques exceptions près) ne peut apparaître sans le nom de sorte que la présence du déterminant force celle d'un nom à proximité. Dans le syntagme verbal, les positions qui précèdent le verbe sont très contraintes et l'on sait que les pronoms personnels clitiques sont strictement ordonnés par rapport au verbe. La désambiguïsation se trouve facilitée par ce fonctionnement positionnel. Ce type de caractéristiques nous servira d'appui dans la formulation des règles. Il faut ajouter cependant que ces règles assument que les unités sur lesquelles on s'appuie ne sont pas elles-mêmes ambiguës.

Les règles

Les règles utilisées sont contextuelles et opèrent de manière essentiellement locale (sur

le contexte immédiat à gauche et à droite). Elles sont ordonnées selon une priorité décroissante et chaque règle s'applique dans un contexte modifié par l'application de la règle précédente. On commence par lever les cas d'ambiguïtés fréquents ou aisés. Cela signifie que l'efficacité d'une règle particulière dépend de l'ordonnancement global. Certaines règles ont un caractère probabiliste, c'est-à-dire qu'elles s'appliquent aux constructions les plus fréquentes. Pour les applications qui ne tolèrent aucune erreur, on devra vérifier les contextes d'application de ces règles.

Il y a trois types de règles. D'abord, on a des règles dites lexicales qui s'appppliquent aux lexèmes spécifiques particulièrement fréquents. Ensuite, on a des règles de confirmation des catégories Verbe et Auxiliaire. Finalement, on trouve des règles qui retirent la catégorie V_conj (verbe conjugué) à la séquence ambiguë.

On attribue le numéro de la règle déclenchée à la propriété *règle. L'ajout de cette propriété permet de repérer la règle qui a opéré la désambiguïsation. Du point de vue informatique, l'ensemble du dispositif prend la forme d'un scénario (fichier de commandes SATO) que nous avons appelé **DESAMBIC**.

Rem. Dans les commandes qui suivent:
*«*1» tient lieu de «Concordance stricte»;*
«-» est un opérateur qui désigne un patron facultatif;*
«+» est un opérateur qui désigne un patron répétable;*
«~» est un opérateur qui désigne un patron qui doit être absent;*
«:» est un opérateur qui implique l'affectation de la valeur qui suit;
«:-» est un opérateur qui implique le retrait de la valeur qui suit;
«:+» est un opérateur qui implique l'ajout de la valeur qui suit;
«==» signifie une égalité stricte (la catégorie désignée et rien d'autre);
«~~» signifie la non-égalité stricte (NON ==)
*«**» en fin de ligne signifie que la commande se poursuit sur la ligne suivante*

Règles lexicales

11- Puis

```
*1      je*~                                                    **
        puis*syntaxe:-v_conj*411                                **
        -*~
```

12- Ni

```
*1      ni                                                      **
        $*syntaxe=v_conj*syntaxe~~v_conj*syntaxe:-v_conj*412
```

13- L'un

```
*1      l'                                                      **
        un*syntaxe:p_indéf
```

Règles de confirmation

1- Pronom personnel - Verbe

Une forme, qui peut être un nom ou un verbe, précédée d'un pronom personnel sujet qui peut être suivi facultativement d'un adverbe de négation et d'un pronom clitique objet, est un verbe conjugué.

Exemple : *Il ne **dépense** ses énergies que pour lutter contre le froid*

```
*1      (je,j',tu,il,elle,on,nous,vous,ils,elles)                    **
        (ne,n')*-                                                    **
        (me,te,nous,vous,le,la,les,m',se,s',t',l',lui,leur)*-        **
        (le,la,les,lui,leur,nous,vous,en,y)*-                        **
        $*syntaxe=v_conj*syntaxe~~v\_conj*syntaxe:v_conj*règle:+c1
```

2- Auxiliaire - Participe passé - (été,eu) - Participe passé

Une forme verbale suivie d'un adverbe facultatif et répétable, de été ou eu facultatifs, et d'une forme pouvant être un adjectif ou un participe passé, est un auxiliaire; été et eu ne sont pas des auxiliaires et la forme adj/ppassé est un participe passé.

La forme été et eu, qui suit un auxiliaire suivi d'un adverbe facultatif et répétable, est un participe passé.

Exemple : *Les plantes sauvages ont souvent été **arrachées** des terres agricoles...*

```
* Cas général
*1      $*syntaxe=aux*syntaxe:aux*règle:+c2                          **
        $*gramr=adv*-*+                                              **
        (été,eu)*syntaxe:ppassé*règle:+c2

* Cas particulier
*1      $*gramr=aux*syntaxe:aux*règle:+c2                            **
        $*gramr=adv*-*+                                              **
        (été,eu)*syntaxe:aux*règle:+c2*-                             **
        $*gramr=(adj,ppassé)*syntaxe:ppassé*règle:+c2

Valeur règle      c2 pour $*syntaxe=aux*syntaxe=v_conj
Valeur syntaxe - aux pour $*syntaxe=aux*syntaxe=v_conj
```

3- ne - Verbe

Une forme précédée de ne, et précédée de pronoms objets clitiques facultatifs et répétables, est un verbe.

Exemple : *La marmotte ne **cache** pas de réserves...*

```
*1      (n',ne)                                                    **
        $*gramr=p_pers*-*+                                         **
        $*syntaxe=v_conj*syntaxe~~v_conj*syntaxe:v_conj*règle:+c3
```

4- Catégorisation des verbes par le pronom objet

Une forme, précédée de pronoms objets (directs ou indirects), et suivie facultativement de d'autres pronoms objets, est un verbe.

Exemple : *La marmotte se terre dans sa résidence d'hiver...*

```
*1      (me,te,se,m',t',s',lui,nous,vous)                          **
        (le,la,les,lui,leur,nous,vous,en,y)*-                      **
        $*syntaxe=v_conj*syntaxe~~v\_conj*syntaxe:v_conj*4c4
```

5- Catégorisation des verbes par le pronom sujet inversé

Une forme, suivie d'un trait d'union, et suivie d'un pronom personnel, est un verbe.

Exemple : ***Amenez-en; Donne-la; Va-t-il à l'école?***

```
*1      $*syntaxe=v_conj*syntaxe~~v\_conj*syntaxe:v_conj*4c5       **
        -                                                          **
        (je,nous,tu,vous,il,elle,on,elles,ils,ce,t,le,la,les,lui,y,en)
```

Règles de désambiguïsation

1- préposition - Verbe

Une forme, qui peut être soit un nom soit un verbe, 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...***

```
*1      $*gramr==prép                                              **
        $*gramr=(art$,dét$)*-                                      **
        $*gramr==adj$*-*+                                          **
        $*syntaxe=v_conj*syntaxe=nomc*syntaxe:-v_conj*&*règle:+d1
```

(Remarque: l'opérateur *& oblige la concordance à se déployer à partir du verbe ambigu. Sinon SATO pourrait, par optimisation, choisir une autre position.

2- (au,aux,du,des,un,une) - Verbe

Une forme, qui peut être un nom ou un verbe, précédée d'un adjectif facultatif lui-même précédé de au,aux,du,des,un,une, n'est pas un verbe conjugué. On exclut la construction *l'un, l'une*.

Exemple : *Cela produit d'abord une **série** de sons sourds...*

```
*1      l'*~                                                    **
        (au,aux,du,des,un,une)                                  **
        $*gramr==adj$*-*+                                       **
        $*syntaxe=v_conj*syntaxe=nomc*syntaxe:-v_conj*règle:+d2
```

3- Pronom personnel (1,2p) - Verbe

Une forme qui se termine par es et qui peut être un nom ou un verbe n'est pas un verbe conjugué si elle n'est pas précédée par tu, nous ou vous. Ces pronoms peuvent être suivis facultativement de ne et/ou de pronoms clitiques objets.

On doit exclure les inversions interrogatives. Cette règle, juste dans la majorité des cas, est cependant erronnée pour les formes impératives des verbes comme *faire* et *dire*. Si l'on a besoin d'une fiabilité absolue, on devra donc vérifier les contextes d'application de la règle.

Exemple : *Nous ne devrions jamais circuler sur les **voies** publiques...*

```
*1      (tu,nous,vous)*~                                        **
        (n',ne)*-                                               **
        $*gramr=p_pers*-                                        **
        les*syntaxe=v_conj*syntaxe=nomc*syntaxe:-v_conj*règle:+d3   **
        -*~
```

4- Déterminant possessif - Verbe

Une forme qui peut être un nom ou un verbe précédée d'une forme qui ne peut être qu'un déterminant possessif n'est pas un verbe conjugué.

Exemple : *Les avions pourront partir à sa **recherche**...*

```
*1      $*gramr==détposs                                        **
        $*syntaxe=v_conj*syntaxe=nomc*syntaxe:-v_conj*règle:+d4     **
```

5- Déterminant démonstratif - Verbe

Une forme qui peut être un nom ou un verbe, précédée d'une forme qui ne peut être qu'un déterminant démonstratif, n'est pas un verbe conjugué.

Exemple : *Les béliers ressemblent en plusieurs points à ceux qui sont nés sous ce **signe** astrologique...*

```
*1      $*gramr==détdém                                          **
        $*syntaxe=v_conj*syntaxe=nomc*syntaxe:-v_conj*règle:+d5   **
```

6- Article - Adjectif - Verbe

Une forme qui peut être un nom ou un verbe, précédée facultativement d'une forme qui ne peut être qu'un adjectif répétable, lui-même précédé d'un article quelconque, n'est pas un verbe conjugué.

Exemple : *Le jeune athlète était choisi sur la meilleure **équipe** du tournoi...*

```
*1      $*gramr=art$                                             **
        $*gramr==adv*-                                           **
        $*gramr==adj*+                                           **
        $*syntaxe=v_conj*syntaxe=nomc*syntaxe:-v_conj*règle:+d6
```

7- Auxiliaire - Verbe

Une forme précédée d'un auxiliaire n'est pas un verbe conjugué.

Exemple : *Conrad et Denis en seront **quittes** pour un sommeil agité.*

```
*1      $*gramr==aux                                             **
        $*syntaxe=v_conj*syntaxe:-v_conj*règle:+d7                **
```

8- Verbe - Verbe conjugué

Une forme qui peut être un nom ou un verbe et qui suit ou qui précède une forme qui ne peut être qu'un verbe conjugué n'est pas un verbe conjugué.

Il s'agit ici d'une règle approximative qui génère des erreurs avec les propositions relatives du genre *l'homme qui arrive repart.* Si une application requiert un minimum d'erreur, on doit prévoir une validation manuelle des phrases où la règle a été déclenchée en présence d'un pronom relatif. Une concordance SATO pourra facilement identifier les cas à vérifier.

Exemple : *Le **groupe a produit** cinq albums qui ont marqué l'histoire...*

```
*1      $*syntaxe=v_conj*syntaxe=nomc*syntaxe:-v_conj*règle:+d8      **
        $*syntaxe==v_conj

*1      $*syntaxe==v_conj                                            **
        (pas,jamais)*-                                               **
        $*syntaxe=v_conj*syntaxe=nomc*syntaxe:-v_conj*règle:+d8
```

9- Verbe infinitif - Verbe

Une forme qui peut être un nom ou un verbe, précédée d'un article facultatif lui même précédé d'un verbe à l'infinitif, n'est pas un verbe conjugué.

Exemple : Le pilote doit avoir **recours** aux instruments de vol.

```
*1      $*gramr=v_inf                                                **
        $*gramr=art$*-                                               **
        $*syntaxe=v_conj*syntaxe=nomc*syntaxe:-v_conj*règle:+d9
```

10- Point - Déterminant - Verbe

Une forme qui peut être un nom et un verbe, précédée d'un déterminant quelconque, lui-même précédé d'un point de ponctuation, n'est pas un verbe conjugué.

Exemple : *Ces **plantes** sont infiniment faciles à cultiver.*

```
*1      .                                                            **
        $*gramr=dét$                                                 **
        $*syntaxe=v_conj*syntaxe=nomc*syntaxe:-v_conj*règle:+d10
```

11- Participe passé - Déterminant - Verbe

Une forme qui peut être un participe passé et un verbe, précédée d'un élément qui est strictment un déterminant, n'est pas un verbe conjugué.

Exemple : *À plusieurs **reprises**, il tente d'ouvrir son parachute.*
*Cela a permis d'obtenir quelques **données** sur la durée de vie de l'ours sauvage.*

```
*1      $*gramr==dét$                                                **
        $*syntaxe=v_conj*gramr=ppassé*syntaxe:-v_conj*règle:+d11
```

12- Verbe - (qui,dont)

Une forme, qui peut être un nom et un verbe, suivie de qui ou dont n'est pas un verbe conjugué.

Exemple : *Une mince couche de **glace** qui se brise.*

```
*1      $*syntaxe=v_conj*syntaxe=nomc*syntaxe:-v_conj*règle:+d12      **
        (qui,dont)
```

13- Verbe conjugué - Article/Déterminant/Adjectif - Verbe

Une forme qui peut être un nom et un verbe, précédée d'un article, d'un déterminant ou d'un adjectif quelconque, lui-même précédé d'une forme qui ne peut être qu'un verbe, n'est pas un verbe conjugué.

Exemples : *Les prairies rocheuses sont autant d'endroits où se trouve la **marmotte***
*Elles étaient longues d'une dizaine de millimètres et me mangeaient toute **vive**.*
*Le co-pilote me sembla bien **calme**.*

```
*1      $*gramr==v_conj                                                **
        $*gramr=(art,dét,adj)$                                         **
        $*syntaxe=v_conj*syntaxe=nomc*syntaxe:-v_conj*règle:+d13
```

14- Verbe conjugué/préposition - nous/vous

Une forme suivie d'un pronom personnel qui n'est pas liée au verbe par un trait d'union n'est pas un verbe conjugué.

Exemples : ***Entre** nous; **Contre** vous, etc.*

```
*1      $*syntaxe=v_conj*syntaxe=prép*syntaxe:-v_conj*4d14             **
        -*~                                                            **
        (nous,vous,lui,toi,moi,eux)
```

Appartenance catégorielle, choix opérés et justifications

Certaines formes (des catégories fonctionnelles pour la plupart) sont désambiguïsées au préalable en vertu de leur fréquence d'utilisation. C'est le cas de *par* et *son* qui ont une probabilité forte pour les catégories Prép et AdjPoss alors qu'elle est assurément faible pour la catégorie Nom. De la même façon, on peut inhiber la catégorie nominale de *est* et la catégorie verbale de *cela* et *plus*.

Une approche fondée sur les statistiques et la pondération (à partir de grand corpus) permet d'envisager le traitement de la désambiguïsation dès la phase d'étiquetage non seulement pour les catégories fonctionnelles mais pour l'ensemble des catégories. Hennequin (1992, p. 17) rapporte que dans les systèmes fonctionnant ainsi, le dictionnaire a une forme un peu particulière. Chaque mot catégoriellement ambigu figurant dans le dictionnaire possède, en plus de ses diverses catégories, un marqueur qui indique la probabilité que le mot en question ait effectivement telle ou telle catégorie (voir aussi Smith 1991, p. 87).

Intérêt de la phase de désambiguïsation avec SATO

Avec SATO, on dispose d'un outil 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 semblables 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 ou d'ajouter de nouvelles règles pour augmenter l'efficacité du système. On pourrait notamment se servir des règles d'accord.

L'intérêt de la méthode développée ici, outre sa relative simplicité, notamment pour l'ajout de nouvelles règles, tient donc au protocole de validation qu'il permet de réaliser.

Après l'application des règles expliquées précédemment, on peut passer en mode assistance pour permettre à l'utilisateur de lever les ambiguïtés qui demeurent. Si on accepte un certain pourcentage d'erreur, on peut sauter cette phase de validation manuelle.

Le dépistage de phrases complexes

Dans l'article de Léo Laroche *Analyses statistiques pour la constitution d'un indice SATO-CALIBRAGE*, on trouve la liste complète des variables (indices) dépistées par SATO-CALIBRAGE. Certaines de ces variables proviennent directement de la grille d'évaluation du ministère de l'Éducation. D'autres nous ont été suggérées par les conseillers pédagogiques qui font partie du Comité des utilisateurs de SATO-CALIBRAGE. Voici le scénario produisant plusieurs des indices de difficulté (ou de facilité) utilisés par SATO-CALIBRAGE. Le rapport de calibrage, dans sa forme minimale, n'utilise qu'une fraction des indices dépistés.

```
* Scénario pour le dépistage de variables d'indices de complexité
* François Daoust
* Centre d'Analyse de Textes par Ordinateur
* Université du Québec à Montréal, juin 1993

* On crée des propriétés de marquage dont on se servira dans la suite des
  opérations;
* la propriété «diag» permet d'identifier le type de complexité dépisté
```

```
Propriété créer rejet symbolique pour lexique oui
Propriété créer marque symbolique pour texte oui
Propriété créer diag symbolique pour texte 1-15 16-20 21-25 26-30 31-99 **
DPréConj DPro3 4V ProInv Inc2 Pro3 Conj2 ProPer2 ProProV PronpV ProÉcrProV

* L'abréviation 1 contient la liste des ponctuations fortes
Abréviation 1 (.,;,:,...,!,?)

* Phrases de 1 à 15 mots
* La commande segmenter divise le texte en phrases en ne
* conservant que celles qui correspondent au patron indiqué: «<16»;
* La commande valeur affecte le diagnostic à la ponctuation
Segmenter par délimiteur *1 terminal longueur <16
Valeur diag + 1-15 pour *1 concordance *

* Phrases de 16 à 20 mots
* La commande segmenter divise le texte en phrases en ne
* conservant que celles qui correspondent au patron indiqué: «>15<21»;
* La commande valeur affecte le diagnostic à la ponctuation
Segmenter par délimiteur *1 terminal longueur >15<21
Valeur diag + 16-20 pour *1 concordance *


* Phrases de 21 à 25 mots
* La commande segmenter divise le texte en phrases en ne
* conservant que celles qui correspondent au patron indiqué: «>20<26»;
* La commande valeur affecte le diagnostic à la ponctuation
Segmenter par délimiteur *1 terminal longueur >20<26
Valeur diag + 21-25 pour *1 concordance *

* Phrases de 26 à 30 mots
* La commande segmenter divise le texte en phrases en ne
* conservant que celles qui correspondent au patron indiqué: «>25<31»;
* La commande valeur affecte le diagnostic à la ponctuation
Segmenter par délimiteur *1 terminal longueur >25<31
Valeur diag + 26-30 pour *1 concordance *

* Phrases de plus de 30 mots
* La commande segmenter divise le texte en phrases en ne
* conservant que celles qui correspondent au patron indiqué: «>30»;
* La commande valeur affecte le diagnostic à la ponctuation
Segmenter par délimiteur *1 terminal longueur >30
Valeur diag + 31-99 pour *1 concordance *

* On redéfinit le contexte pour inclure la ponctuation qui
* précède le début de la phrase.
Contexte délimité de *1 inclus à *1 inclus

* Phrases commençant par une préposition ou une conjonction
* On cherche la ponctuation immédiatement suivie d'une
* préposition ou d'une conjonction à laquelle on affecte le diagnostic.
Concordance stricte *1 $*gramr=(prépo$,conjon$)*diag:+DPréConj
```

```
* Phrases qui débutent par un pronom à la 3ième personne
* On cherche la ponctuation immédiatement suivie du
* pronom auquel on affecte le diagnostic.
Concordance stricte *1 (il,elle,ils,elles)*diag:+DPro3

* Phrases contenant quatre propositions ou plus
* On cherche quatre verbes conjugés successifs;
* on affecte le diagnostic au dernier verbe.
Concordance ordonnée $*syntaxe=v\_conj  $*syntaxe=v\_conj **
$*syntaxe=v\_conj $*syntaxe=v\_conj*diag:+4V

* Phrases interrogatives contenant un pronom inversé
* On cherche un verbe conjugué immédiatement suivi (opérateur «*.»)
* d'un trait d'union et d'un pronom; Ensuite, on cherche la présence
* de «?» auquel on affecte le diagnostic.
Concordance ordonnée $*syntaxe=v\_conj*. -*. $*gramr=p\_$ ?*diag:+ProInv

* Phrases contenant au moins 2 mots inconnus
* On cherche deux mots inconnus;
* on affecte le diagnostic au dernier.
Concordance ordonnée $*connu=nil  $*connu=nil*diag:+Inc2

* Phrases contenant une séquence de 3 pronoms
* On cherche trois pronoms successifs;
* on affecte le diagnostic au dernier.
Concordance stricte $*gramr=p\_$ $*gramr=p\_$ $*gramr=p\_$*diag:+Pro3


* Phrases contenant deux conjonctions ou plus
* On cherche deux conjonctions successives;
* on affecte le diagnostic à la dernière.
Concordance ordonnée $*gramr=conjonction $*gramr=conjonction*diag:+Conj2

* Phrases contenant deux pronoms personnels
* de la première ou deuxième personne
* On cherche deux pronoms désignés successifs;
* on affecte le diagnostic au dernier.
Concordance ordonnée (j',je,tu,t',nous,vous) (j',je,tu,t',nous,vous)*diag:+ProPer2

* Phrases contenant le patron  pronom pronom verbe
* On exclut les formes pronominales simples que l'on va marquer par «*marque:oui»;
* Finalement, on cherche un pronom, immédiatement suivi d'un pronom
* nom-marqué et d'un verbe auquel on affecte le diagnostic.
Concordance stricte je me*marque:oui $*syntaxe=v\_conj
Concordance stricte tu te*marque:oui $*syntaxe=v\_conj
Concordance stricte (il,elle,ils,elles) se*marque:oui $*syntaxe=v\_conj
Concordance stricte nous nous*marque:oui $*syntaxe=v\_conj
Concordance stricte vous vous*marque:oui $*syntaxe=v\_conj
Valeur marque oui pour y
Concordance stricte $*gramr=p\_$ $*gramr=p\_$*marque~oui **
```

```
      $*syntaxe=v\_conj*diag:+ProProV

* Phrases contenant le patron  pronom-non-personnel verbe
* en excluant «c'» et «on» que l'on va marquer par «*rejet:oui»;
* Finalement, on cherche un pronom désigné qui n'est pas rejeté,
* immédiatement suivi d'un verbe conjugué auquel on affecte le diagnostic.
Valeur rejet oui pour (c',on)
Concordance stricte $*gramr=p\_(relatif,indéf,dém,poss)*rejet~oui **
$*syntaxe=v\_conj*diag:+PronpV
Valeur rejet nil pour (c',on)

* Phrases contenant le patron suivant: pronom écran pronom verbe; L'écran
* est n'importe quoi sauf une ponctuation; On affecte le diagnostic au verbe
Concordance stricte $*syntaxe=p\_$ **
$*syntaxe~p\_$*gram~(ponctuation,virgule) **
$*syntaxe=p\_$ $*syntaxe=v\_conj*diag:+ProÉcrProV
```

Une fois que le dispositif linguistique a effectué ses diagnostics , SATO-CALIBRAGE peut afficher les phrases marquées dans le rapport de calibrage. Il peut aussi les compter pour produire les données qui seront traitées par les analyseurs statistiques.

NOTES

[1] 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 1989-91) dirigé par Jules Duchastel alors qu'il était directeur du Centre d'ATO.

[2] L'équipe RDLC du Centre d'ATO a produit un analyseur morphologique (LCMF, L. Dumas avec la collaboration de P. Plante, D. Perras et A. Plante) destiné à fournir l'information catégorielle nécessaire au parseur ALSF (analyseur lexico-syntaxique du français, J.M. Marandin, S. David et P. Plante). Contrairement à l'approche que nous avons adoptée avec la BDL, l'analyseur morphologique LCMF est un module informatique fermé plutôt qu'une base de données modifiable.

[3] COURTOIS, Blandine (1990) "Un système de dictionnaires électroniques pour les mots simples du français". *Langue Française 87.*

[4] Que l'on pense par exemple à l'expression «en fait» qui souvent n'agit pas comme locution: «en fait, il en fait trop». Dans cet exemple, on a la locution adverbiale «en fait» suivie d'une forme conjuguée du verbe «faire».

Bibliographie

COURTOIS, Blandine (1990) "Un système de dictionnaires électroniques pour les mots simples du français". *Langue Française 87.*

FUJISAKI, T. F. JELINEQ, J. COCKE, E. BLACK, T. NISHINO (1989) « A Probalilistic Parsing Method for Sentence Disambiguation», présenté au International Workshop on Parsing Technologies, CMU, repris in *Current Issues in Parsing Technology*, Masaru Tomita (éds), Kluwer Academic, 1991.

GUILLET, Alain (1990) «Reconnaissance des formes verbales avec un dictionnaire minimal». *Langue Française 87.*

HABERT, Benoît (1991) *Olmes : un système d'exploration et de structuration de textes,* Thèse de doctorat, Université Paris 7, Institut Blaise Pascal.

HENNEQUIN, Marie-Pierre (1992) «Émontage et analyse syntaxique automatique», DEA en Linguistique Théorique et Formelle, Université Paris 7, UFR en Linguistique.

MILNE, Robert (1988) «Lexical Ambiguity Resolution in a Deterministic Parser» in *Lexicla Ambiguity Resolution,* Steven L. Small, Garrison W. Cottrel and Michael K. Tanenhaus (eds), Morgan Kaufman Publishers.

SILBERZTEIN, Max (1992) «Reconnaissance automatique des mots d'un texte : les premières étapes» à paraître dans les Actes du Colloque Lexique Grammaire, UQAM.

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

SMITH, Georges W. (1991) Computers and Human Language, Oxford University Press, New York.

Le dispositif mathématique

François Daoust.

François Daoust est informaticien et chercheur au Centre d'analyse de texte par ordinateur -Cognition et information-. Il est responsable du projet SATO-CALIBRAGE au Centre ATO-CI.

Nous désignons, par dispositif mathématique, l'ensemble des méthodes quantitatives utilisées à l'intérieur du projet SATO-CALIBRAGE 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 (indices) qui varient de façon significative par rapport aux textes provenant des divers niveaux scolaires. Ainsi, on peut confirmer ou infirmer nos 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 niveau scolaire d'un texte.

Dans SATO-CALIBRAGE, 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 le niveau d'enseignement auquel ils sont destinés, nous avons utilisé des tests d'hypothèses 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 (progiciel SPSS). Nous 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 (cf. l'article de Guy Cucumel) destiné à grouper les termes ayant des distributions similaires sur l'échelle scolaire. L'interprétation des groupes a permis d'éliminer certains groupes 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 (indices SATO-CALIBRAGE) permettant de classer un texte dans un niveau d'enseignement. Pour ce faire, nous avons utilisé les régressions simples et multiples (cf article de Léo Laroche), et l'analyse discriminante.

La sélection des indices

Dans la partie précédente sur le dispositif linguistique, nous avons illustré comment, à partir d'hypothèses linguistiques et pédagogiques, nous avons construit des scénarios SATO qui dépistent la réalisation de certaines constructions discursives. Ces scénarios produisent des indices numériques à savoir le nombre, absolu ou relatif, d'occurrences du phénomène recherché.

La première question que l'on se pose est la suivante: est-ce que ces indices varient de façon significative lorsqu'on analyse des textes qui proviennent de niveaux d'enseignement différents?

Revenons à notre hypothèse selon laquelle l'usage de certains termes fonctionnels distingue les textes selon leur niveau scolaire. Pour valider une telle hypothèse, il nous faut établir un indice qui va prendre la forme d'une fonction à deux paramètres. Le premier paramètre représente un terme de l'ensemble du vocabulaire dont la liste nous est donnée naturellement par l'axe lexical construit par SATO. Un des dispositifs linguistiques auquel fait appel SATO-CALIBRAGE permet d'effectuer une catégorisation grammaticale hors contexte d'un lexique (cf. article de Fernande Dupuis et François Daoust). Par suite, il est donc facile de restreindre le vocabulaire soumis à notre indice aux lexèmes ayant reçus la catégorie grammaticale recherchée.

Ensuite, nous avons besoin d'un paramètre qui signe la variation des textes selon le niveau scolaire. Ce paramètre sera construit sur l'axe textuel (chaîne des occurrences) de SATO. Il suffit en effet de construire une partition sur cette axe qui distingue les occurrences des lexèmes selon qu'elles appartiennent à un texte issu de première année, deuxième, etc.

Cette partition peut s'opérer de diverses façons. Nous allons illustrer ici la façon la plus simple. Elle consiste à nommer chacune des classes d'enseignement lors même de la constitution du corpus.

Constitution d'un corpus par niveaux scolaires

```
*page=PR1          { document primaire 1 composé ...}
*page=@texte1.     { d'un texte sur le fichier texte1}
*page=@texte2.     { et texte2}

*page=PR2          { document primaire 2 composé ... }
*page=@texte3.     { de texte3 }
*page=@texte4.     { et texte4 }
```

Finalement l'indice lui-même se définit comme la somme des occurrences d'un lexème donné dans une classe donnée pondérée par le nombre total d'occurrences dans la classe. Il s'agit donc d'une fréquence relative, par exemple, la fréquence de la locution «en_outre» dans le domaine des textes de sixième année.

L'indice étant défini, on doit l'appliquer à notre corpus en faisant varier les paramètres. Voici un exemple d'un telle procédure en termes de commandes SATO :

Commandes SATO pour trouver des termes discriminants
(procédure CANDIDAT)

```
* On définit la propriété «motfonc» qui va contenir l'ensemble
* mots fonctionnels potentiellement discriminants
Propriété CRéer motfonc symbolique pour lexique oui non

* Pour ce faire, on sélectionne d'abord les lexèmes qui
* possèdent les catégories grammaticales suivantes:
* adv: adverbes;
* art, artdéf, artgén, artind, artpart:
```

```
*      articles, articles définis, génériques, indéfinis et partitifs;
* conjonction;
* détdéf, détdém, détindéf, détposs:
*      déterminants définis, démonstratifs, indéfinis, possessifs;
* délim: délimiteurs;
* p_dém, p_indéf, p_pers, p_poss, p_relatif:
*      pronoms démonstratifs, indéfinis, personnels, possessifs,    relatifs;
* ponctuation: ponctuation forte;
* prép: préposition;
* virgule: ponctuation faible.

Valeur motfonc oui pour **
$*gramr=(adv,art,art(déf,gén,ind,part),conjonction,**
dét(déf,dém,ind,poss),délim,**
p\_(dém,indéf,pers,poss,relatif),ponctuation,prép,virgule)

* On rajoute les locutions fonctionnelles bloquées par le caractère «_»
Valeur motfonc oui pour |\_$*mot=nil

* On élimine certaines marques d'édition
Valeur motfonc - oui pour (x,[,],%,#,+,/)

* Partition du texte en niveau scolaire
Segmenter par document

* On crée la propriété «chi2» qui va retenir l'indice calculé
Propriété CRéer chi2 entière pour lexique

* On calcule l'indice pour tous les lexèmes
* qui font partie de la liste des mots fonctionnels.
* («$» est un opérateur de troncature sur les caractères)
Compter fréquences $*motfonc=oui Sauvegarde
```

Outre, le tableau des indices de fréquence des lexèmes dans chaque domaine, la procédure CANDIDAT fournit, grâce à la commande COMPTER, une représentation mathématique des résultats.

Représentation mathématique des résultats de la procédure CANDIDAT

Moyenne	Écart	Répart.	Discri.	Chi2	
1.39%	0.18	100.0%	0.00	31.11	**à**
0.01%	0.00	83.3%	0.00	1.16	**à_cause_de**
...					

Le tableau précédent fournit un certain nombre de mesures qui nous permettent d'interpréter mathématiquement les résultats obtenus par l'application de notre indice. Ainsi, dans SATO-CALIBRAGE, nous avons utilisé la statistique du Chi2 pour ne conserver, parmi l'ensemble des termes fonctionnels, que ceux dont l'indice Chi2 dépasse un certain seuil (23.20 pour une probabilité d'erreur de 1% à 11 degrés de liberté).

SATO calcule l'indice Chi2 en comparant, pour un lexème donné, son nombre d'occurences dans chacun des segments par rapport au nombre attendu sous l'hypothèse d'une indépendance de la distribution du lexème sur la partition choisie.

Sur les 1048 termes fonctionnels soumis à l'analyse, 332 ont été retenus comme étant inégalement distribués sur les 11 niveaux d'enseignements qui constituent le corpus.

Nous avons aussi soumis au test du Chi2 les autres variables du prototype, en particulier celles qui portent sur les constructions de phrases, pour ne conserver que les variables pertinentes.

Le groupement des indices par classification automatique

Comme le nombre de variables retenues par le test du Chi2 nous semblait trop important, nous avons voulu grouper certaines variables se comportant de façon similaire. En fait, c'est le groupement des termes fonctionnels qui nous intéressait d'abord. En effet, on retrouve parmi ces termes des fréquences relativement peu élevées. Nous avons voulu combiner des lexèmes afin de produire des indices plus stables.

L'article de Guy Cucumel explique les principes de la classification automatique. Sur la base de ces principes, nous avons utilisé des progiciels différents qui ont abouti à des classifications comparables.

Nous avons donc produit la matrice d'occurrences par segments des termes fonctionnels retenus en utilisant la procédure COMPTER de SATO. Par la suite nous avons soumis cette matrice au progiciel statistique SAS. Il est à noter que pour cette analyse, nous avons profité des résultats de l'analyse de régression déjà réalisée par Léo Laroche. Cette analyse nous a conduit à combiner les textes des secondaire 1 et 2, d'une part et de secondaire 3, 4 et 5 d'autre part. Deux jeux de commandes SAS sont utilisés. Le premier vise à recoder les fréquences de départ en fonction de la taille des documents (niveaux scolaires). Ensuite, on transforme les fréquences absolues ainsi pondérées en fréquences relatives à l'ensemble du corpus. Ainsi, on pourra utiliser la distance euclédienne pour constituer les différents groupes.

```
* Production de variables pondérées

libname MEQ 'répertoire des données en format SAS';
data MEQ.FONC;
infile 'fichier de données ASCII';
length id $ 24;
input id $ V1-V11;

* On pondère les fréquences absolues en égalisant la taille
  des différentes classes;
F1=V1*36.83;
F2=V2*15.97;
F3=V3*8.51;
F4=V4*6.26;
F5=V5*5.30;
F6=V6*3.85;
F7=(V7+V8)*2.15;
```

```
F8=V9+V10+V11;

* On transforme les fréquences absolues en proportion de distribution
  entre les différentes classes;
TOTAL=F1+F2+F3+F4+F5+F6+F7+F8;
F1=F1/TOTAL;
F2=F2/TOTAL;
F3=F3/TOTAL;
F4=F4/TOTAL;
F5=F5/TOTAL;
F6=F6/TOTAL;
F7=F7/TOTAL;
F8=F8/TOTAL;


* Classification automatique
libname MEQ 'répertoire des données en format SAS';

* On peut faire varier le nombre de groupes (maxc)
proc FASTCLUS
  data=MEQ.FONC maxc=10 short maxiter=10 out=MEQ.FC;
  var f1-f8; id id;

* On classe les données en fonction de leur groupe
proc sort;
  by cluster id;

* On imprime les données en fonction de leur groupe
proc print
  data=MEQ.FC;
  var id f1-f8 total;
  by cluster;

* On imprime, pour chaque groupe, la fréquence moyenne par niveau
proc means
  data=MEQ.FC;
  var f1-f8;
  by cluster;
```

La première chose que nous avons constaté en examinant les résultats de la classification automatique, c'est l'inégalité des classes obtenues. Cela nous a permis de dépister des comportements atypiques. Il faut voir en effet que nous n'avons pas utilisé de critères sémantiques dans le choix des termes fonctionnels. Ainsi, on peut trouver des adverbes de manière dont l'usage traduit davantage un choix d'auteur qu'un niveau de complexité. En examinant la distribution moyenne des lexèmes selon les niveaux scolaires, on peut constater ce caractère atypique.

Donc, la classification automatique nous a permis d'éliminer des variables. Nous avons repris ce processus d'élagage en retraitant de façon itérative les données restantes jusqu'à l'obtention de groupes significatifs par rapport à notre objectif. Après cinq ou six de ces itérations, nous avons réduit notre nombre de variable de 332 à 162. Ce travail d'élagage va se

poursuivre par un examen plus approfondi des groupes. Finalement, dans les mois qui viennent nous soumettront les variables groupées à l'analyse de régression et à l'analyse discriminante.

Liste des mots fonctionnels potentiellement discriminants

Voici la liste des termes fonctionnels soumis à la classification automatique. Les lexèmes suivis d'un astérisque font partie de la liste des 162 lexèmes conservés après un premier travail d'élagage.

	au_moyen_d'	chèrement	en_dessous_de
- *	auprès_de	chez *	en_effet *
à *	au_sommet_de	ci *	enfin *
à_côté_de	au_sommet_du	comme *	en_haut_de
à_droite	aussi *	comment *	en_réalité
afin_de	aussitôt_que	confortablement	ensemble
afin_qu'	autour_de	consciencieusement	ensuite *
à_gauche	autour_du	contre	entre *
ainsi *	autre *	d' *	envers
à_jamais	autrefois	d'_accord_avec	en_vertu_de *
à_l'_aide_d'	autres *	d'_affilée *	environ *
à_l'_exception_des	aux *	d'_ailleurs *	en_vrac
à_l'_instant *	avant *	dans *	et *
alors *	avec *	dans_le_but_d'	et_puis
alors_que *	bas	de *	face_au
à_même_de	beaucoup *	dedans	facilement *
à_mesure_que	bien *	de_devant	fébrilement
à_moins_d' *	bien_que	dehors	finalement
amoureusement	bien_sûr	delà	fort
à_nouveau	bientôt *	demain	fréquemment
à_part	brusquement	de_nouveau *	gaiement
après *	c' *	depuis_qu'	généralement
à_présent *	ça *	des *	grâce_à *
à_proximité_d'	car *	dès	grâce_aux
à_proximité_de	ce *	dessus	guère *
à_savoir	ceci	devant	habituellement
attentivement	cela *	différents	haut
au *	celui *	directement *	hé
au_bas_d'	cependant	donc	hélas
au_bas_de	cependant_que	dont *	heureusement
aucun *	certaines *	du *	hier
au-delà *	certains *	durant *	hors_du
au-dessous *	certes *	également *	ici
au-dessus *	ces *	elle *	il *
au-devant *	cette	elles *	ils *
au_haut_d'	ceux *	en *	immédiatement
aujourd''hui	chacun	en_conformité_avec	impatiemment
aujourd'_hui	chacune	encore *	inconsciemment
au_juste	chaque *	en_dessous	ironiquement

j' *
jamais *
je *
jusqu'_à *
jusqu'_au
l' *
la *
là *
la_sienne *
le *
le_long_de *
lentement
les *
les_miens
les_siens
leur *
leurs *
loin
loin_que *
long *
longtemps
lorsqu' *
lorsque *
lui *
m'
ma
maintenant *
maints *
mais *
mal *
malheureusement
me *
même *
mes *
moi *
mon *
n' *
ne *
ni *
non
nos *
notre
nous *
obligatoirement
on *
or
ou *
où *
ou_bien *
oui
outre
par *
parce_qu'
parce_que
par_contre *
par_exemple
parfaitement
parfois *
par_la_suite
parmi *
particulièrement *
partout *
pas *
patiemment
pendant *
pendant_que
peu *
pis
plein
plus *
plusieurs *
plus_qu'
plutôt *
plutôt_qu'
point *
pour *
pour_qu'
pourquoi *
près *
près_d'
près_de *
près_de_chez
près_du
principalement
proche_de
prudemment
puis *
puisque
qu' *
quand *
quant
quatre
que *
quelque *
quelquefois
quelques *
qui *
quoi
rien *
s' *
sa *
sans *
se *
selon
sensiblement
séparément *
ses *
seul *
si *
simplement
sinon *
soi
soigneusement
son *
soudain
sous
souvent *
subitement
sur *
surtout *
t' *
ta *
tandis_qu'
tant
tantôt
tard *
te *
tel *
telle
tellement_qu'
telles
tels
tes *
tiens
toi *
ton *
tous
tout *
tout_à_coup
tout_à_l'_heure
tout_de_suite
toute *
toutefois *
toutes
traditionnellement
très *
trop *
tu *
un *
une *
unes
un_peu *
un_peu_d'
vers *
verticalement
violemment
vite
voici *
voilà
vos
votre
vous *
vraiment
vu
, *
. *
: *
; *
? *
! *
... *
(
)
« *
» *
$ *
% *

La construction de fonctions prédictives

Le dernier type de méthodes mathématiques utilisées dans SATO-CALIBRAGE vise à construire des fonctions prédictives permettant de classer un texte dans un niveau scolaire à partir des seuls indices produits par le prototype. C'est ce que l'on appellera l'indice SATO-CALIBRAGE.

Comme le jugement de la difficulté d'un texte nous vient de son classement selon le niveau scolaire auquel il est destiné, il est naturel de faire appel à des modèles de corrélation linéaire entre nos indices et la variable «niveau scolaire». Voilà ce qui justifie l'emploi des méthodes de régression décrite dans l'article de Léo Laroche.

Nous pourrions aussi utiliser des méthodes d'analyse discriminante qui ont l'avantage de pouvoir utiliser des variables qui n'auraient pas cette relation linéaire avec le niveau scolaire. Cependant, l'interprétation de la fonction de classement est beaucoup moins intuitive en analyse discriminante que dans la régression.

Classification par partition et classification hiérarchique: deux méthodes complémentaires

Guy Cucumel.

Guy Cucumel est professeur au département des sciences comptables de l'UQAM. Il est aussi spécialiste des méthodes françaises d'analyse de données.

Introduction

Les analyses de classification automatiques sont des méthodes multidimensionnelles descriptives, permettant d'explorer un échantillon de données, d'épurer ces données et d'aider à la formulation d'hypothèses de recherche. Leur objectif général est de regrouper des individus ou des objets, caractérisés par un ensemble de variables, en classes, d'une part bien homogènes, d'autre part aussi différentes entre elles que possible.

Notre approche se veut essentiellement pédagogique. Deux méthodes sont présentées ici, la classification par partition et la classification ascendante hiérarchique. A l'aide d'un exemple fictif, on montre le fonctionnement de chacune des techniques, sans entrer dans le développement mathématique. On montre par la suite en quoi ces deux méthodes sont complémentaires l'une de l'autre, puis on les applique sur un cas réel d'analyse.

1. Notion d'inertie

Le concept d'inertie est introduit ici sous une forme vulgarisée, il est à la base des deux algorithmes que nous allons présenter. On définit trois types d'inertie: l'inertie totale, l'inertie intra-classe et l'inertie inter-classe.

Considérons un nuage de points dont le centre de gravité est G (figure 1a), l'inertie totale du nuage est la somme des carrés des distances de tous les points au centre de gravité G (figure 1b).

Comme on peut l'observer sur la figure 1c, les points de ce nuage se répartissent naturellement en deux classes A et B.

L'inertie intra-classe correspond à la somme, sur les deux classes A et B, des carrés des distances des points de chacune des classes au centre de gravité de celle-ci (figure 1c). Cette mesure traduit un critère d'homogénéité interne des classes. Ainsi plus les distances entre les individus des classes et le centre de gravité de celles-ci sont faibles, plus la valeur de l'inertie intra-classe est faible, et plus les classes sont homogènes.

L'inertie inter-classe (figure 1d) correspond, pour sa part, à la somme des carrés des distances des centres de gravité G_A et G_B des deux classes au centre de gravité du nuage de points, pondérées par le poids de chaque classe (i.e.: le nombre d'objets contenus dans chaque classe). Cette mesure est un critère de séparabilité des classes. Ainsi plus les centres de gravité des classes sont éloignés du centre de gravité du nuage de points, plus la valeur de l'inertie inter-classe est élevée et mieux les classes sont séparées.

Etant donné que l'inertie totale se décompose en inertie intra-classe et inter-classe (l'inertie totale correspond à la somme des inerties intra-classes et inter-classes), on obtient la propriété suivante: en augmentant l'une ou l'autre des inerties intra-classe ou inter-classe, on obtient une diminution de sa contrepartie. Donc, en cherchant les classes les plus homogènes, on trouve simultanément les classes les mieux séparées.

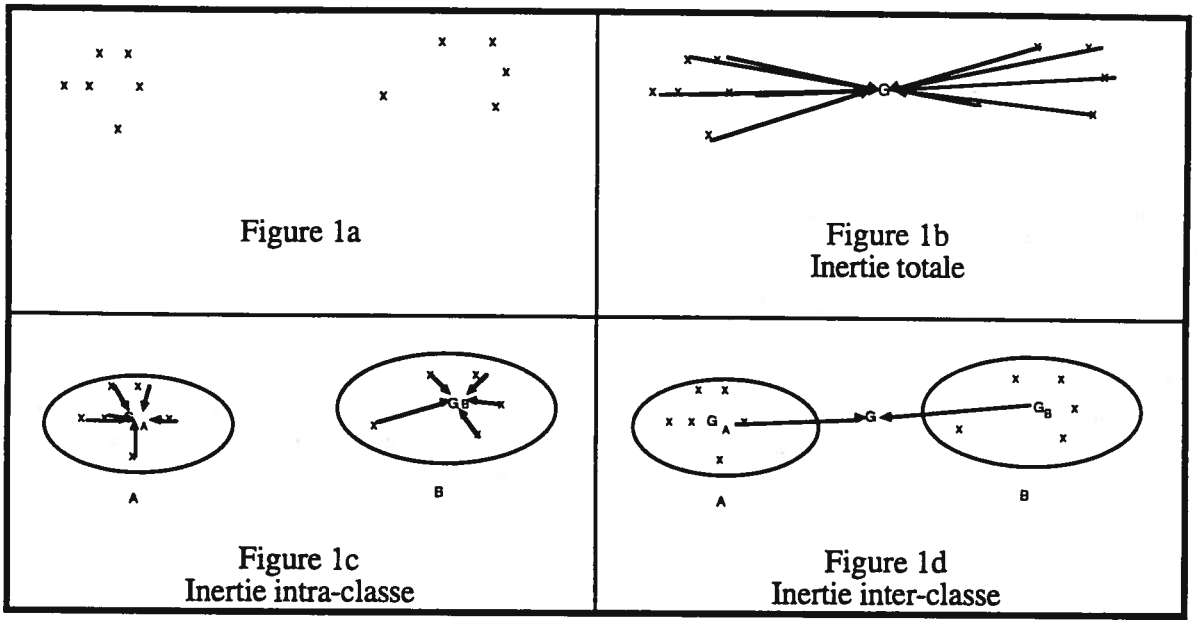


Figure 1a

Figure 1b
Inertie totale

Figure 1c
Inertie intra-classe

Figure 1d
Inertie inter-classe

2. La classification par partition

Le but de cette méthode est d'obtenir une classification d'un ensemble d'objets en classes deux à deux disjointes, le nombre de classes étant fixé a priori. Les objets les plus semblables quant aux caractéristiques (ou variables) les décrivant devant appartenir à la même classe à l'issue de la classification. Mathématiquement, la technique consiste à rechercher, pour un nombre de classes fixé, les regroupements d'objets en classes qui minimisent l'inertie intra-classe tout en maximisant l'inertie inter-classe.

Le principe de cette méthode est exposé à l'aide de l'exemple suivant:

On dispose d'un ensemble de 10 objets numérotés A à J, caractérisés par trois variables V_1, V_2 et V_3 (tableau 1).

La représentation de ces données dans un espace à trois dimensions (figure 2) montre clairement que les 10 objets s'organisent en trois classes: ABCDE, FGH et IJ. Les situations pour lesquelles une simple représentation graphique suffit à repérer des classes dans un ensemble de données sont cependant extrêmement rares. En général, d'une part le nombre de variables est trop important pour

qu'une représentation graphique soit possible à produire, et d'autre part, le nombre d'objets à classer est grand et les classes ne sont pas aussi distinctes les unes des autres que dans la présente situation.

	V_1	V_2	V_3
A	1	1	1
B	1	2	1
C	1	2	2
D	2	2	2
E	3	1	1
F	1	10	8
G	3	10	9
H	2	9	9
I	6	10	6
J	7	10	6

Tableau 1

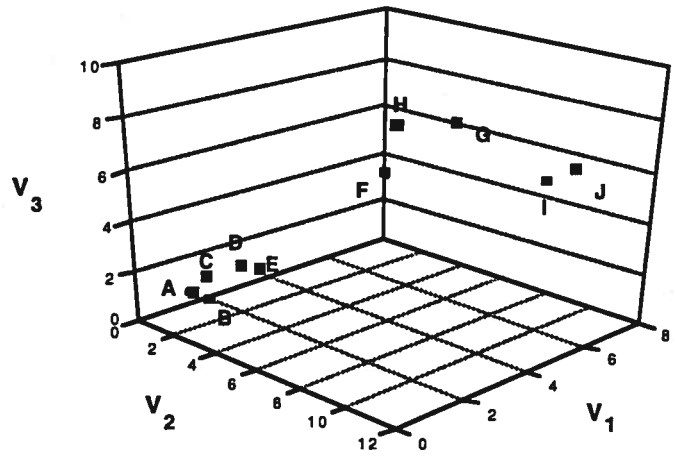


Figure 2

On doit donc recourir à un algorithme afin de déterminer par un processus automatique les classes présentes dans la population d'objets à classifier. Son fonctionnement est le suivant:

Pour trouver une partition en k classes :

i) choisir aléatoirement k points appelés noyaux.

ii) associer chaque point du nuage au noyau dont il est le plus proche. On forme ainsi une partition (ensemble de classes deux à deux disjointes) en k classes.

iii) déterminer k nouveaux noyaux en prenant les points les plus proches des centres de gravités des k classes trouvées à l'étape ii).

iv) retour à l'étape ii).

v) l'algorithme s'arrête lorsque la partition ne change plus d'une itération à une autre.

Une variante de l'algorithme consiste à l'initialiser avec une partition aléatoire et à passer immédiatement à l'étape iii).

Pour appliquer l'algorithme sur le présent exemple, on commence par dresser le tableau de distances (tableau 2) de tous les objets entre eux. La distance utilisée est la distance euclidienne classique. On ne discutera pas ici du choix de la distance.

	A	B	C	D	E	F	G	H	I	J
A	0	1.00	1.41	1.73	2.00	11.40	12.21	13.04	11.45	11.92
B	1.00	0	1.00	1.41	2.24	10.63	11.49	10.68	10.68	11.18
C	1.41	1.00	0	1.00	2.45	10.00	10.82	9.95	10.25	10.77
D	1.73	1.41	1.00	0	1.73	10.05	10.68	9.90	9.80	10.25
E	2.00	2.24	2.45	1.73	0	11.58	12.04	11.36	10.72	11.05
F	11.40	10.63	10.00	10.05	11.58	0	2.24	1.73	5.39	6.32
G	12.21	11.49	10.82	10.68	12.04	2.24	0	1.41	4.24	5.00
H	13.04	10.68	9.95	9.90	11.36	1.73	1.41	0	5.92	5.92
I	11.45	10.68	10.25	9.80	10.72	5.39	4.24	5.92	0	1.00
J	11.92	11.18	10.77	10.25	11.05	6.32	5.00	5.92	1.00	0

Tableau 2

1ère itération:

On cherche une partition en trois classes

i) Choisissons trois points au hasard: A, F et H.

ii) A l'aide du tableau de distances (tableau 2), on affecte chaque point au noyau dont il est le plus proche. On obtient ainsi une première partition en trois classes:

classe 1: A, B, C, D, E
classe 2: F,I
classe 3: H,G,J

iii) On calcule les coordonnées du centre de gravité de chaque classe (tableau 3).

	V_1	V_2	V_3
classe 1	2.67	2.67	1.40
classe 2	3.50	10.00	14.00
classe 3	4.00	9.67	8.67

Tableau 3

Les points les plus proches des trois centres de gravité sont choisis comme nouveaux noyaux, soit:

classe 1: D
classe 2: F
classe 3 G

iv) retour à l'étape ii)

2ème itération:

On définit une nouvelle partition en trois classes:

classe 1: D, A, B, C, D
classe 2: F
classe 3: G, H, I, J

iii) calcul des centres de gravité des trois nouvelles classes (tableau 4)

	V_1	V_2	V_3
classe 1	2.67	2.67	1.40
classe 2	1.00	10.00	8.00
classe 3	4.50	9.75	7.50

Tableau 4

Les points les plus proches des trois centres de gravité sont choisis comme nouveaux noyaux, soit:

classe 1: D
classe 2: F
classe 3 I

iv) retour à l'étape ii)

3ème itération:

On définit une nouvelle partition en trois classes:

classe 1: D, A, B, C, D
classe 2: F, G, H
classe 3: I, J

iii) calcul des centres de gravité des trois nouvelles classes (tableau 5)

	V_1	V_2	V_3
classe 1	2.67	2.67	1.40
classe 2	2.00	9.67	8.67
classe 3	6.50	10.00	6.00

Tableau 5

Les points les plus proches des trois centres de gravité sont choisis comme nouveaux noyaux, soit:

classe 1: D
classe 2: H
classe 3 I

iv) retour à l'étape ii)

4ème itération:

On définit une nouvelle partition en trois classes:

classe 1: D, A, B, C, D
classe 2: F, G, H
classe 3: I, J

v) la partition trouvée est la même qu'à l'étape précédente, l'algorithme a convergé. La partition "naturelle" que l'on pouvait observer à partir de la figure 2 a donc été trouvée par un processus automatique.

La classification par partition présente l'avantage de détecter des classes sans a priori, en partant d'une solution de départ aléatoire. De plus l'algorithme converge toujours très rapidement vers une solution. En effet, empiriquement, on constate que la convergence est généralement atteinte en cinq ou six itérations. Son grand inconvénient est d'obliger l'utilisateur à fixer le nombre de classes qu'il souhaite obtenir dès le départ. En d'autre termes, si on avait souhaité obtenir une solution en quatre classes ou en deux classes, l'algorithme aurait convergé vers de telles solutions, ce qui n'est pas nécessairement utile, car la structure des objets de départ s'organise en trois classes. En pratique le critère d'optimisation est la diminution de l'inertie intra-classe. L'algorithme s'arrête

lorsque celle-ci ne décroît plus de façon significative. La solution retenue n'est pas nécessairement la meilleure partition possible, mais est néanmoins une bonne partition[1].

3. La classification ascendante hiérarchique

La classification ascendante hiérarchique conduit pour sa part, à l'identification des regroupements les plus typiques issus de la population à l'étude, tout en ayant l'avantage, contrairement à la méthode précédente[2], de ne pas exiger de l'utilisateur qu'il précise lui-même, dès le départ, le nombre de classes recherchées. Pour cette raison, la classification hiérarchique se révèle particulièrement bien adaptée aux fins de recherches exploratoires.
Pour construire une hiérarchie (résultat d'une classification ascendante hiérarchique), on procède en plusieurs étapes. A l'aide du tableau de distances (tableau 2), on cherche à associer deux à deux les individus les plus proches. Par la suite, on calcule les distances entre les classes ainsi formées. La distance entre deux classes est définie par la perte d'inertie inter-classe que l'on encourt en les regroupant. On associe à leur tour ces nouvelles classes en réunissant deux à deux les plus proches. On réitère l'opération jusqu'à ce que toutes les classes soient agrégées en une seule. A chaque étape du processus, la classe nouvellement formée, appelée palier, est nécessairement composée de la réunion de deux classes existant à l'étape précédente. On construit ainsi une suite de classes emboîtées de telle sorte qu'à n'importe quelle étape de la procédure tous les individus appartiennent à une classe et à une seule.

Au terme de la procédure, on obtient une classification, représentée par un arbre hiérarchique ou dendrogramme (figure 3), tel que les classes formées successivement sont nécessairement imbriquées les unes dans les autres. La partition la plus grossière constitue le sommet de l'arbre (tous les individus étant regroupés en une seule classe) alors que les éléments terminaux en forment la partition la plus fine (chaque individu constituant à lui seul une classe de la partition).

On associe à chaque palier de l'arbre une valeur numérique, appelé indice d'agrégation, correspondant à sa hauteur sur le dendrogramme. Cette valeur est la distance des deux classes que le palier réunit. Plus on se rapproche du sommet de l'arbre, plus la distance entre les deux classes les plus proches est grande, ce qui se traduit par un indice d'agrégation de plus en plus élevé.

L'opération suivante consiste à couper l'arbre au niveau d'un saut important de l'indice d'agrégation de façon à obtenir une bonne partition. En coupant l'arbre au niveau d'un saut important de l'indice d'agrégation, on s'assure d'obtenir une partition constituée de classes bien homogènes et bien séparées, car l'importance du saut témoigne du fait que, pour constituer une nouvelle classe, il a fallu réunir des groupes qui, bien qu'étant les plus proches, se trouvaient tout de même relativement éloignés.
La coupure matérialisée sur la figure 3 par le trait pointillé donne les trois classes que l'on a obtenu précédemment par la classification par partition.

---

[1] A titre d'exemple, si un ordinateur pouvait énumérer et évaluer un millier de partitions par seconde, il lui faudrait plus de 8 jours pour déterminer la partition optimale de 20 individus en 5 classes et plus de 2444 siècles si le nombre d'individus passe à 30 (Diday et al. 1982).

[2] Qu'on parle de techniques de partitionnement direct ou de classification floue. Voir à ce sujet, Sneath et Sokal (1973); Diday et al. (1982).

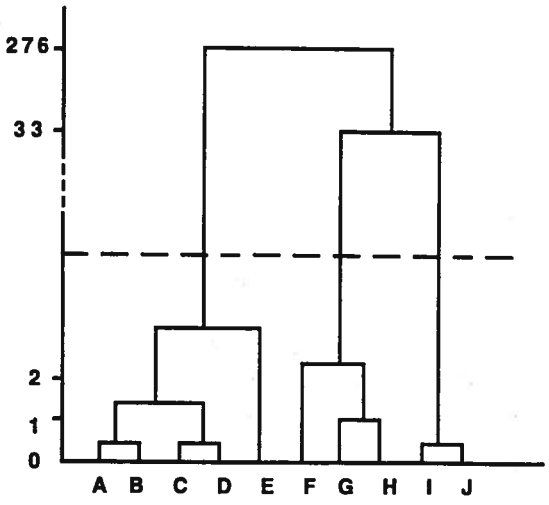


Figure 3

4. Complémentarité des deux méthodes

La classification hiérarchique présente le défaut de ne pas systématiquement fournir la meilleure partition. De plus compte tenu de la complexité de sa représentation graphique, il est peu envisageable de l'interpréter lorsque le nombre de données est important. En pratique, on se contentera de ne représenter qu'un histogramme des indices de niveaux afin de repérer le premier saut important de l'indice d'agrégation, ce qui permettra d'avoir une idée assez précise du nombre de classes présentes dans l'échantillon étudié. La partition obtenue est alors utilisée comme partition de base avec la méthode de classification par partition. Les deux algorithmes sont utilisés de façon complémentaire.

5. Exemple

On dispose d'un échantillon de 546 textes[3] appartenant aux différents niveaux du primaire et du secondaire (voir l'article de Léo Laroche) dont les profils sont définis par la liste de variables suivante:

% d'adjectifs
% d'adverbes
% d'articles définis
% d'articles partitifs
% de conjonctions
% de verbes conjugués (avec la propriété gramr)
% de verbes conjugués (avec la propriété syntaxe)
% de délimiteurs
% de déterminants indéfinis
% de déterminants démonstratifs

[3] Le corpus utilisé pour cette analyse est antérieur à celui utilisé pour les autres analyses statistiques, ce qui explique que l'échantillon n'est composé que de 546 textes.

% de déterminants numériques
% de déterminants possessifs
% de verbes infinitifs
% de noms communs
% de noms propres
% de participes passés
% de participes présents
% de ponctuations
% de prépositions
% de pronoms
% de pronoms personnels
% de pronoms relatifs
% de pronoms indéfinis
% de pronoms démonstratifs
% de pronoms possessifs
% de formes résiduelles
% de virgules
% d'interjections
% de mots inconnus (vocabulaire de 6ème année)
% de pronoms "je j' tu nous vous"
% de phrases de plus de 15 mots
% de phrases de plus de 20 mots
% de phrases de plus de 25 mots
% de phrases de plus de 30 mots
% de phrases commençant par une préposition ou une conjonction
% de phrases commençant par un pronom de la 3ème personne
% de phrases contenant 4 propositions ou plus
% de phrases interrogatives contenant un pronom renversé
% de phrases contenant au moins une subordonnée relative
% de phrases contenant au moins 2 mots inconnus (6ème année du primaire)
% de phrases contenant une séquence de 3 pronoms
% de phrases contenant 2 conjonctions ou plus
% de phrases contenant au moins un pronom de la 1ère ou 2ème personne
% de phrase contenant le patron suivant: pronoms réfléchis + verbe conjugué
% de phrases contenant le patron suivant: pronoms non personnel + verbe conjugué
% de phrases contenant le patron suivant: pronoms + écran + pronoms + conjugué

Notre objectif est de déterminer s'il existe des groupes de textes ayant des profils semblables, de déterminer quelles en sont les variables caractéristiques et de comparer ces groupes avec le niveau scolaire auquel ils correspondent d'après des experts.

Une classification ascendante hiérarchique montre qu'il existe cinq classes de textes dans cet échantillon. Les classes obtenues après coupure de la hiérarchie sont ensuite soumises à un algorithme de classification par partition. Les tableaux 6, 7 et 8 indiquent pour chacune des classes la liste des variables les plus significatives. Une variable est considérée comme étant significative d'une classe, si l'écart entre sa moyenne dans la classe et sa moyenne dans l'échantillon (moyenne générale) est grand. Une statistique de test (Morineau, 1984) appelée valeur-test est indiquée. Plus cette valeur est grande en valeur absolue plus la variable est caractéristique de la classe. Une valeur-test positive indique une moyenne pour la classe supérieure à la moyenne de l'échantillon, alors

qu'une valeur-test négative indique l'inverse. La probabilité associée à la valeur test, est la probabilité d'obtenir un écart aussi important si celui-ci est dû au hasard. Ainsi pour la classe 1/5, c'est le pourcentage de ponctuations qui est la variable la plus caractéristique, il est plus élevé dans cette classe que dans l'échantillon total. Le pourcentage de phrases contenant au moins une subordonnée relative est la deuxième variable la plus importante, mais cette fois-ci ce pourcentage est plus faible pour la classe que pour l'échantillon total.

NUM.	LIBELLE	IDEN	MOYENNES CLASSE	MOYENNES GENERALE	ECARTS TYPES CLASSE	ECARTS TYPES GENERAL	V.TEST	PROBA
	CLASSE 1 / 5					EFFECTIF = 243		
20.%	PONCTUATIONS	% PO	9.718	7.747	2.509	2.802	14.70	0.000
23.%	PRONOMS PERSONNELS	% PR	9.063	8.252	3.439	3.394	5.00	0.000
65.%	PHRASES COMMENCANT PAR UN PRONOM 3E PERS.	% PH	12.275	10.387	10.307	8.771	4.50	0.000
9.%	VERBES CONJUGUES (PROP. SYNTHAXE)	% VE	10.868	10.315	3.377	2.947	3.92	0.000
8.%	VERBES CONJUGUES (PROP. GRAMR)	% VE	13.927	13.350	3.928	3.247	3.71	0.000
32.%	PRONOMS JE TU NOUS VOUS	% PR	2.987	2.492	3.346	2.873	3.60	0.000
64.%	PHRASES COMMENCANT PAR PREP. OU CONJ.	% PH	0.719	0.369	3.545	2.517	2.91	0.002
14.%	DETERMINANTS POSSESSIFS	% DE	2.267	2.070	1.798	1.613	2.56	0.005
10.%	DELIMITEURS	% DE	0.766	0.922	1.044	1.243	-2.63	0.004
3.%	ADVERBES	% AD	5.926	6.222	2.583	2.304	-2.68	0.004
11.%	DETERMINANTS INDEFINIS	% DE	1.051	1.149	0.765	0.725	-2.80	0.003
18.%	PARTICIPES PASSES	% PA	3.739	4.019	2.252	1.875	-3.12	0.001
72.%	PHRASES CONTENANT AU MOINS 1 PRONOM PERS. 1E OU 2E PERS.	% PH	4.147	6.188	5.480	11.067	-3.86	0.000
12.%	DETERMINANTS DEMONSTRATIFS	% DE	0.634	0.751	0.562	0.542	-4.53	0.000
75.%	DE PHRASES AVEC LES PATRONS: PRON+ECRAN+PRON+CONJUGUE	% DE	4.298	6.581	5.139	9.133	-5.23	0.000
19.%	PARTICIPES PRESENTS	% PA	0.499	0.633	0.538	0.522	-5.36	0.000
7.%	CONJONCTIONS	% CO	5.479	6.010	2.005	1.918	-5.79	0.000
70.%	PHRASES CONTENANT UNE SEQUENCES DE 3 PRONOMS	% PH	1.930	4.175	2.422	7.643	-6.14	0.000
29.%	VIRGULES	% VI	4.122	4.767	1.795	2.043	-6.61	0.000
21.%	PREPOSITIONS	% PR	10.847	11.933	3.207	3.272	-6.94	0.000
66.%	PHRASES CONTENANT 4 PROPOSITIONS OU PLUS	% PH	1.241	5.066	2.453	10.287	-7.77	0.000
36.%	MOTS DE 9 LETTRES ET PLUS	% MO	7.066	8.747	3.897	4.065	-8.65	0.000
73.%	DE PHRASES AVEC LE PATRON PRONOMS REFL.+CONJUGUE	% DE	8.804	13.712	7.314	10.747	-9.55	0.000
24.%	PRONOMS RELATIFS	% PR	1.428	1.948	0.857	1.089	-9.99	0.000
63.%	DE PHRASES DE PLUS DE 30 MOTS	% DE	1.176	7.561	2.903	12.875	-10.37	0.000
74.%	DE PHRASES AVEC LES PATRONS: PRON NON-PERS.+CONJUGUE	% DE	9.634	15.350	8.005	11.358	-10.52	0.000
31.%	MOTS INCONNUS (VOCABULAIRE 6E ANNEE)	% MO	2.126	4.441	2.262	3.820	-12.67	0.000
71.%	PHRASES CONTENANT 2 CONJONCTIONS OU PLUS	% PH	12.176	21.974	8.802	15.233	-13.45	0.000
69.%	PHRASES CONTENANT AU MOINS 2 MOTS INCONNUS (6E PRIM.)	% PH	3.894	18.205	6.086	21.877	-13.68	0.000
68.%	PHRASES CONTENANT AU MOINS UNE SUBORDONNEE REL.	% PH	14.355	24.232	8.570	15.061	-13.71	0.000

Tableau 6

Compte tenu du fait que l'on a 11 niveaux scolaires, si les variables que l'on a utilisées pour la classification étaient suffisamment discriminantes on aurait pu s'attendre à trouver une classification en 11 classes, chacune des classes correspondant à un niveau particulier. Le tableau 9 montre le croisement entre la classification des textes effectuée par des experts a priori et la classification trouvée par un procédé automatique.

Comme l'indique le tableau 9 la classe 1 regroupe majoritairement des textes du début du primaire, résultat que l'on pouvait attendre puisque les deux variables les plus caractéristiques sont le pourcentage de ponctuations et le pourcentage de pronoms personnels. La classe 2 regroupe plutôt des textes de la fin du primaire, la classe 3 des textes du début du secondaire et la classe 4 des textes de la fin du secondaire. Quant à la classe 5, qui ne regroupe que 5 textes elle est difficile à qualifier. Elle regroupe certainement des textes qui sont difficilement classifiables et dont le profil ne correspond à aucun niveau particulier.

NUM.	VARIABLES CARACTERISTIQUES LIBELLE	IDEN	MOYENNES CLASSE	MOYENNES GENERALE	ECARTS TYPES CLASSE	ECARTS TYPES GENERAL	V.TEST	PROBA
	CLASSE 2 / 5					EFFECTIF = 123		
24.%	PRONOMS RELATIFS	% PR	3.040	1.948	1.052	1.089	12.62	0.000
68.%	PHRASES CONTENANT AU MOINS UNE SUBORDONNEE REL.	% PH	35.631	24.232	12.543	15.061	9.53	0.000
73.%	DE PHRASES AVEC LE PATRON PRONOMS REFL.+CONJUGUE	% DE	21.659	13.712	9.044	10.747	9.31	0.000
7.%	CONJONCTIONS	% CO	7.311	6.010	1.819	1.918	8.54	0.000
26.%	PRONOMS DEMONSTRATIFS	% PR	1.536	1.059	0.919	0.792	7.58	0.000
70.%	PHRASES CONTENANT UNE SEQUENCES DE 3 PRONOMS	% PH	7.695	4.175	5.440	7.643	5.80	0.000
71.%	PHRASES CONTENANT 2 CONJONCTIONS OU PLUS	% PH	28.740	21.974	12.560	15.233	5.59	0.000
72.%	PHRASES CONTENANT AU MOINS 1 PRONOM PERS. 1E OU 2E PERS.	% PH	10.630	6.188	10.168	11.067	5.05	0.000
23.%	PRONOMS PERSONNELS	% PR	9.571	8.252	3.382	3.394	4.89	0.000
67.%	PHRASES INTER. CONTENANT UN PRONOM RENVERSE	% PH	4.726	2.741	8.580	5.397	4.63	0.000
3.%	ADVERBES	% AD	7.035	6.222	2.220	2.304	4.44	0.000
32.%	PRONOMS JE TU NOUS VOUS	% PR	3.463	2.492	2.707	2.873	4.25	0.000
12.%	DETERMINANTS DEMONSTRATIFS	% DE	0.931	0.751	0.559	0.542	4.17	0.000
74.%	DE PHRASES AVEC LES PATRONS: PRON NON-PERS.+CONJUGUE	% DE	18.843	15.350	9.114	11.358	3.87	0.000
15.%	VERBES INFINITIFS	% VE	3.605	3.131	1.357	1.844	3.24	0.001
11.%	DETERMINANTS INDEFINIS	% DE	1.334	1.149	0.857	0.725	3.22	0.001
9.%	VERBES CONJUGUES (PROP. SYNTHAXE)	% VE	11.054	10.315	2.795	2.947	3.16	0.001
21.%	PREPOSITIONS	% PR	11.200	11.933	2.821	3.272	-2.82	0.002
20.%	PONCTUATIONS	% PO	7.090	7.747	1.909	2.802	-2.95	0.002
13.%	DETERMINANTS NUMERIQUES	% DE	0.846	1.191	0.907	1.302	-3.34	0.000
4.%	ARTICLES DEFINIS	% AR	7.054	7.756	2.586	2.637	-3.35	0.000
22.%	PRONOMS	% PR	4.885	5.605	1.941	2.142	-4.23	0.000
16.%	NOMS COMMUNS	% NO	23.963	25.696	3.913	4.647	-4.70	0.000
31.%	MOTS INCONNUS (VOCABULAIRE 6E ANNEE)	% MO	2.622	4.441	1.861	3.820	-5.99	0.000
69.%	PHRASES CONTENANT AU MOINS 2 MOTS INCONNUS (6E PRIM.)	% PH	7.407	18.205	7.655	21.877	-6.21	0.000
	CLASSE 3 / 5					EFFECTIF = 129		
31.%	MOTS INCONNUS (VOCABULAIRE 6E ANNEE)	% MO	8.537	4.441	2.327	3.820	13.92	0.000
69.%	PHRASES CONTENANT AU MOINS 2 MOTS INCONNUS (6E PRIM.)	% PH	38.073	18.205	10.744	21.877	11.79	0.000
36.%	MOTS DE 9 LETTRES ET PLUS	% MO	10.891	8.747	3.300	4.065	6.85	0.000
29.%	VIRGULES	% VI	5.828	4.767	1.878	2.043	6.74	0.000
21.%	PREPOSITIONS	% PR	13.443	11.933	2.496	3.272	5.99	0.000
19.%	PARTICIPES PRESENTS	% PA	0.823	0.633	0.476	0.522	4.74	0.000
22.%	PRONOMS	% PR	6.262	5.605	1.672	2.142	3.98	0.000
10.%	DELIMITEURS	% DE	1.241	0.922	1.353	1.243	3.33	0.000
13.%	DETERMINANTS NUMERIQUES	% DE	1.471	1.191	1.239	1.302	2.80	0.003
70.%	PHRASES CONTENANT UNE SEQUENCES DE 3 PRONOMS	% PH	2.721	4.175	2.387	7.643	-2.47	0.007
67.%	PHRASES INTER. CONTENANT UN PRONOM RENVERSE	% PH	1.477	2.741	3.917	5.397	-3.04	0.001
8.%	VERBES CONJUGUES (PROP. GRAMR)	% VE	12.583	13.350	2.035	3.247	-3.07	0.001
15.%	VERBES INFINITIFS	% VE	2.688	3.131	1.043	1.844	-3.12	0.001
72.%	PHRASES CONTENANT AU MOINS 1 PRONOM PERS. 1E OU 2E PERS.	% PH	3.231	6.188	4.934	11.067	-3.47	0.000
26.%	PRONOMS DEMONSTRATIFS	% PR	0.838	1.059	0.524	0.792	-3.62	0.000
9.%	VERBES CONJUGUES (PROP. SYNTHAXE)	% VE	9.298	10.315	1.767	2.947	-4.48	0.000
32.%	PRONOMS JE TU NOUS VOUS	% PR	1.151	2.492	1.377	2.873	-6.06	0.000
20.%	PONCTUATIONS	% PO	6.208	7.747	1.106	2.802	-7.13	0.000
23.%	PRONOMS PERSONNELS	% PR	6.371	8.252	1.998	3.394	-7.19	0.000

Tableau 7

NUM.	VARIABLES CARACTERISTIQUES LIBELLE	IDEN	MOYENNES CLASSE	MOYENNES GENERALE	ECARTS TYPES CLASSE	ECARTS TYPES GENERAL	V.TEST	PROBA
	CLASSE 4 / 5					EFFECTIF = 46		
69.%	PHRASES CONTENANT AU MOINS 2 MOTS INCONNUS (6E PRIM.)	% PH	59.469	18.205	15.615	21.877	13.36	0.000
63.%	DE PHRASES DE PLUS DE 30 MOTS	% DE	28.968	7.561	12.415	12.875	11.77	0.000
31.%	MOTS INCONNUS (VOCABULAIRE 6E ANNEE)	% MO	9.576	4.441	2.950	3.820	9.52	0.000
71.%	PHRASES CONTENANT 2 CONJONCTIONS OU PLUS	% PH	40.652	21.974	12.038	15.233	8.68	0.000
68.%	PHRASES CONTENANT AU MOINS UNE SUBORDONNEE REL.	% PH	41.621	24.232	12.380	15.061	8.18	0.000
36.%	MOTS DE 9 LETTRES ET PLUS	% MO	12.717	8.747	3.774	4.065	6.92	0.000
74.%	DE PHRASES AVEC LES PATRONS: PRON NON-PERS.+CONJUGUE	% DE	25.678	15.350	9.375	11.358	6.44	0.000
21.%	PREPOSITIONS	% PR	14.865	11.933	2.321	3.272	6.35	0.000
66.%	PHRASES CONTENANT 4 PROPOSITIONS OU PLUS	% PH	12.836	5.066	9.781	10.287	5.35	0.000
19.%	PARTICIPES PRESENTS	% PA	0.937	0.633	0.530	0.522	4.12	0.000
22.%	PRONOMS	% PR	6.728	5.605	1.632	2.142	3.71	0.000
75.%	DE PHRASES AVEC LES PATRONS: PRON+ECRAN+PRON+CONJUGUE	% DE	10.696	6.581	7.445	9.133	3.19	0.001
4.%	ARTICLES DEFINIS	% AR	8.930	7.756	1.744	2.637	3.15	0.001
2.%	ADJECTIFS	% AD	7.970	7.070	1.741	2.169	2.94	0.002
29.%	VIRGULES	% VI	5.604	4.767	1.798	2.043	2.90	0.002
16.%	NOMS COMMUNS	% NO	27.341	25.696	3.328	4.647	2.51	0.006
14.%	DETERMINANTS POSSESSIFS	% DE	1.409	2.070	0.865	1.613	-2.90	0.002
65.%	PHRASES COMMENCANT PAR UN PRONOM 3E PERS.	% PH	5.920	10.387	5.798	8.771	-3.61	0.000
32.%	PRONOMS JE TU NOUS VOUS	% PR	0.802	2.492	1.123	2.873	-4.16	0.000
8.%	VERBES CONJUGUES (PROP. GRAMR)	% VE	11.393	13.350	2.101	3.247	-4.27	0.000
9.%	VERBES CONJUGUES (PROP. SYNTHAXE)	% VE	8.304	10.315	1.662	2.947	-4.83	0.000
23.%	PRONOMS PERSONNELS	% PR	5.343	8.252	2.129	3.394	-6.07	0.000
20.%	PONCTUATIONS	% PO	4.161	7.747	0.645	2.802	-9.06	0.000
	CLASSE 5 / 5					EFFECTIF = 5		
66.%	PHRASES CONTENANT 4 PROPOSITIONS OU PLUS	% PH	87.143	5.066	19.378	10.287	17.91	0.000
75.%	DE PHRASES AVEC LES PATRONS: PRON+ECRAN+PRON+CONJUGUE	% DE	78.571	6.581	26.342	9.133	17.69	0.000
72.%	PHRASES CONTENANT AU MOINS 1 PRONOM PERS. 1E OU 2E PERS.	% PH	88.571	6.188	22.857	11.067	16.71	0.000
70.%	PHRASES CONTENANT UNE SEQUENCES DE 3 PRONOMS	% PH	58.571	4.175	37.904	7.643	15.97	0.000
63.%	DE PHRASES DE PLUS DE 30 MOTS	% DE	90.000	7.561	20.000	12.875	14.37	0.000
74.%	DE PHRASES AVEC LES PATRONS: PRON NON-PERS.+CONJUGUE	% DE	71.429	15.350	23.474	11.358	11.08	0.000
71.%	PHRASES CONTENANT 2 CONJONCTIONS OU PLUS	% PH	90.000	21.974	20.000	15.233	10.02	0.000
73.%	DE PHRASES AVEC LE PATRON PRONOMS REFL.+CONJUGUE	% DE	61.429	13.712	37.143	10.747	9.96	0.000
68.%	PHRASES CONTENANT AU MOINS UNE SUBORDONNEE REL.	% PH	77.143	24.232	22.768	15.061	7.88	0.000
69.%	PHRASES CONTENANT AU MOINS 2 MOTS INCONNUS (6E PRIM.)	% PH	87.143	18.205	19.378	21.877	7.07	0.000
21.%	PREPOSITIONS	% PR	16.820	11.933	5.565	3.272	3.35	0.000
16.%	NOMS COMMUNS	% NO	31.800	25.696	3.259	4.647	2.95	0.002
14.%	DETERMINANTS POSSESSIFS	% DE	4.180	2.070	1.568	1.613	2.94	0.002
31.%	MOTS INCONNUS (VOCABULAIRE 6E ANNEE)	% MO	8.800	4.441	1.934	3.820	2.56	0.005
65.%	PHRASES COMMENCANT PAR UN PRONOM 3E PERS.	% PH	0.000	10.387	0.000	8.771	-2.66	0.004
20.%	PONCTUATIONS	% PO	0.880	7.747	0.798	2.802	-5.50	0.000

Tableau 8

Effectif* % colonne % ligne	classe 1	classe 2	classe 3	classe 4	classe 5	total
primaire 1	51 / 24.5 20.99 92.73	2 / 12.4 1.63 3.64	2 / 13.0 1.55 3.64	0 / 4.6 0.00 0.00	0 / 0.5 0.00 0.00	55 10.07 100
primaire 2	48 / 26.7 19.75 80.00	12 / 13.5 9.76 20.00	0 / 14.2 0.00 0.00	0 / 5.1 0.00 0.00	0 / 0.5 0.00 0.00	60 10.99 100.00
primaire 3	49 / 31.2 20.16 70.00	20 / 15.8 16.26 28.57	1 / 16.5 0.78 1.43	0 / 5.9 0.00 0.00	0 / 0.6 0.00 0.00	70 12.82 100.00
primaire 4	24 / 22.3 9.88 48.00	16 / 11.3 13.01 32.00	10 / 11.8 7.75 20.00	0 / 4.2 0.00 0.00	0 / 0.5 0.00 0.00	50 9.16 100.00
primaire 5	13 / 20.5 5.35 28.26	21 / 10.4 17.07 45.65	12 / 10.9 9.30 26.09	0 / 3.9 0.00 0.00	0 / 0.4 0.00 0.00	46 8.42 100.00
primaire 6	25 / 34.3 10.29 32.47	23 / 17.3 18.70 29.87	22 / 18.2 17.05 28.57	7 / 6.5 15.22 9.09	0 / 0.7 0.00 0.00	77 14.10 100.00
secondaire 1	12 / 28.0 4.94 19.05	3 / 14.2 2.44 4.76	32 / 14.9 24.81 50.79	16 / 5.3 34.78 25.40	0 / 0.6 0.00 0.00	63 11.54 100.00
secondaire 2	9 / 19.1 3.70 20.93	5 / 9.7 4.07 11.63	22 / 10.2 17.05 51.16	5 / 3.6 10.87 11.63	2 / 0.4 40.00 4.65	43 7.88 100.00
secondaire 3	1 / 10.2 0.41 4.35	8 / 5.2 6.50 34.78	13 / 5.4 10.08 56.52	1 / 1.9 2.17 4.35	0 / 0.2 0.00 0.00	23 4.21 100.00
secondaire 4	6 / 13.4 2.47 20.00	5 / 6.8 4.07 16.67	12 / 7.1 9.30 40.00	7 / 2.5 15.22 23.33	0 / 0.3 0.00 0.00	30 5.49 100.00
secondaire 5	5 / 12.9 2.06 17.24	8 / 6.5 6.50 27.59	3 / 6.9 2.33 10.34	10 / 2.4 21.74 34.48	5 / 0.3 60.00 10.34	29 5.31 100.00
total	243 100.00 44.51	123 100.00 22.53	129 100.00 23.63	46 100.00 8.42	5 100.00 0.92	546 100.00 100.00

Tableau 9

* Le premier effectif indiqué est l'effectif réel, le suivant est l'effectif sous hypothèse d'indépendance des deux classifications.

Conclusion

Comme on peut le constater à partir de l'exemple précédent, les méthodes de classification sont particulièrement bien adaptées aux analyses exploratoires. Dans la présente situation, on peut constater que certaines des variables dont on dispose ont un pouvoir de discrimination mais ne sont pas suffisantes pour discriminer l'ensemble des niveaux scolaires. Dans une telle situation il conviendrait d'éliminer les variables qui n'ont manifestement aucun pouvoir discriminant et d'en construire d'autres pour raffiner la qualité de la partition obtenue par classification automatique. Par ailleurs il faudrait également s'interroger sur la pertinence des cinq textes de la classe 5 qui ont peut-être des profils très particuliers et dans ce cas aucune raison de figurer dans un échantillon constitué dans le but de déterminer les valeurs des variables caractéristiques des niveaux scolaires.

Une série de classifications effectuées sur des jeux de variables différents conduirait à une liste de variables potentiellement adaptées à la discrimination des texte. Cette liste devrait être soumise à d'autres analyses statistiques telles que l'analyse de discrimination ou l'analyse de régression multiple et pourrait contribuer à construire un modèle permettant de classer de nouveaux textes directement à partir de leurs profils.

Bibliographie

DIDAY, E., LEMAIRE, J., POUGET, J., TESTU, F. Eléments d'analyse de données, Paris: Dunod, 1982.

DIDAY E., et al. Optimisation en classification automatique. Tomes 1 et 2, Rocquencourt: Inria, 1980.

LEBART, L., SALEM A. Analyse statistique de données textuelles. Paris: Dunod, 1988.

MORINEAU, A. Note sur la caractérisation statistique d'une classe et les valeurs-tests. Bulletin technique du centre statistique et informatique appliquées, volume 2, n° 1-2: 20-28, 1984.

SOKAL R.R., SNEATH P.H.A. Principles of Numerical Taxonomy. Freeman and co: San Francisco, London, 1963.

Analyses statistiques pour la constitution d'un indice SATO-CALIBRAGE

Léo Laroche.

Léo Laroche travaille au mìnistère de l'Éducation du Québec. Jusqu'en mars 1992, il appartenait à la Direction générale de l'évaluation et des ressources didactiques. Maintenant, il est à la Direction de la recherche.

Introduction

Dans son article *Description du corpus de textes*, Lise Ouellet décrit le corpus constitué dans le cadre de cette étude sur la lisibilité de documents utilisés pour supporter l'enseignement et pour réaliser l'évaluation des apprentissages. SATO nous a permis de disposer, pour chaque texte, de statistiques sur un ensemble de variables morpho-syntaxiques et stylistiques; ces «portraits quantifiés» de chacun des textes provenant des différentes classes d'enseignement du primaire et du secondaire ont servi à détecter des tendances, à tracer des profils. Plus de cent vingt variables ont ainsi été quantifiées.

Le logiciel SYSTAT[1] a permis de réaliser un ensemble de compilations dans le but d'identifier la classe d'enseignement. Les valeurs admissibles correspondent ici au nombre d'années de scolarité, c'est-à-dire de 1 à 11; la classe a ainsi été utilisée comme variable dépendante dans les différentes analyses réalisées.

Les travaux statistiques ont été accomplis en deux temps. Il s'est de plus avéré pertinent de réaliser ces calculs pour l'ensemble du corpus ainsi que pour chacun des sous-groupes de textes rattachés au primaire et au secondaire. Il y a eu tout d'abord des compilations faisant intervenir la classe d'enseignement et chacune des variables disponibles. Ces premiers résultats ont permis d'identifier les variables prises une à une les plus fortement reliées à la classe d'enseignement. Une deuxième ronde d'analyses a permis d'examiner simultanément les liens entre la classe d'enseignement et l'ensemble des variables retenues au moment de la première étape de compilation. Il a été possible de fabriquer des indices de calibrage rendant compte de la complexité des textes en référence aux classes d'enseignement.

Une première étape : des analyses univariées

Rappelons notre objectif. Il s'agit, sur la base des «portraits quantifiés» de chacun des textes, d'établir un lien entre les variables (indices de difficulté-facilité) et la classe d'enseignement auquel est destiné le texte. Pour ce faire, deux techniques d'analyse statistique ont été utilisées : la régression simple et la corrélation de Pearson. Ces deux techniques permettent en

[1] La version 5.03 de SYSTAT fonctionnant sous DOS a été utilisée pour cette recherche.

effet de mesurer le lien qui s'établit entre deux séries de valeurs. Ici, nous avons retenu d'une part la classe d'enseignement (entre 1 et 11 correspondant au rattachement des textes aux différentes classes du primaire -- 1 à 6 -- et à celles du secondaire -- 7 à 11 --) et d'autre part les valeurs obtenues à chacune des variables disponibles. Les coefficients fournis par ces techniques vont dans le même sens. Nous fournissons au tableau 1 les coefficients de corrélation de Pearson calculés pour chacune des cent vingt-deux variables en fonction des textes aux sous-ensembles suivants:

- les textes du primaire,
- les textes du secondaire,
- l'ensemble des textes.

Par ailleurs, les variables descriptives sont regroupées par rapport aux catégories suivantes:

1- Certains lexèmes, c'est-à-dire l'importance de la présence au sein de chaque texte de certains lexèmes retenus à la suite d'analyses statistiques préalables (nombre de variables : 88).

2- Certaines catégories grammaticales : il s'agit d'une mesure de l'importance de la présence dans chaque texte de mots rattachés à certaines catégories grammaticales. Sauf pour les verbes conjugués, le rattachement a été établis hors contexte en référence à la Base de Données Lexicales disponible au Centre d'ATO (nombre de variables : 11).

3- Longueurs moyennes : nous avons retenu les longueurs moyennes des phrases et des paragraphes comme susceptibles d'être reliées à la mesure de la complexité d'un texte (nombre de variables : 2).

4- Phrases possédant certaines caractéristiques : nous retrouvons ici des proportions de phrases d'un texte qui possèdent l'une ou l'autre caractéristique susceptible de servir d'indicateurs à la complexité du texte (nombre de variables : 16).

5- Autres caractéristiques : nous regroupons ici quelques statistiques qu'il est difficile de rattacher aux autres catégories présentées ci-dessus (nombre de variables : 5).

Tableau 1

Coefficients de corrélation de Pearson[2] établis entre la classe d'enseignement et un ensemble de variables

Variables	Primaire	Secondaire	Ensemble
A- Certains lexèmes			
Présence du lexème « à »	0.016	-0.005	0.033
Présence du lexème « à nouveau »	0.081	-0.182	0.001
Présence du lexème « ça »	-0.084	0.111	-0.012
Présence du lexème « « »	-0.075	0.053	0.056
Présence du lexème « » »	-0.078	0.054	0.052
Présence du lexème « ! »	-0.296	-0.021	-0.234
Présence du lexème « « »	0.289	-0.018	0.146
Présence du lexème « % »	0.050	0.054	0.134
Présence du lexème « . »	-0.593	-0.232	-0.634
Présence du lexème « ... »	0.025	0.068	0.110
Présence du lexème « ; »	0.140	-0.035	0.109
Présence du lexème « = »	0.063	0.101	0.038
Présence du lexème « au »	0.030	0.053	0.065
Présence du lexème « aujourd'hui »	0.090	0.055	0.057
Présence du lexème « auquel »	0.105	-0.025	0.028
Présence du lexème « avec »	-0.103	0.045	-0.077
Présence du lexème « c'est-à-dire »	0.124	0.063	0.132
Présence du lexème « c'est pourquoi »	0.143	-0.044	0.078
Présence du lexème « c'est que »	-0.029	0.007	0.017
Présence du lexème « car »	0.276	-0.126	0.101
Présence du lexème « ce »	0.073	0.085	-0.012
Présence du lexème « celui-ci »	0.047	-0.150	-0.011
Présence du lexème « certaines »	0.250	-0.024	0.080
Présence du lexème « certains »	0.253	0.080	0.104
Présence du lexème « ceux »	0.123	0.258	0.143
Présence du lexème « comme »	-0.038	0.054	0.014
Présence du lexème « d' »	0.304	0.075	0.315
Présence du lexème « d'ailleurs »	0.232	-0.011	0.070
Présence du lexème « d'après »	0.075	-0.086	0.043
Présence du lexème « de »	0.356	0.042	0.332
Présence du lexème « de ce que »	-0.028	0.077	0.003
Présence du lexème « du »	0.108	-0.013	0.145
Présence du lexème « elle »	-0.028	0.052	-0.074
Présence du lexème « elles »	0.153	-0.105	0.008
Présence du lexème « en »	0.266	0.008	0.217
Présence du lexème « en haut »	-0.048	-0.110	-0.042
Présence du lexème « en outre »	0.105	0.051	0.102

[2] Dans certains cas, il a été impossible d'établir un tel coefficient entre la classe d'enseignement et certaines variables lorsque la variation des valeurs est nulle.

Tableau 1 (suite)

Coefficients de corrélation de Pearson établis entre la classe d'enseignement et un ensemble de variables

Variables	Pri-maire	Secon-daire	Ensemble
Présence du lexème « enfin »	0.008	-0.194	0.000
Présence du lexème « faute d' »	0.045	-0.003	0.021
Présence du lexème « grâce à »	0.178	-0.227	0.028
Présence du lexème « jusqu'au milieu »	-0.042		-0.045
Présence du lexème « jusqu'à »	0.105	-0.144	0.014
Présence du lexème « l' »	0.269	-0.100	0.269
Présence du lexème « leur »	0.124	0.013	0.043
Présence du lexème « leurs »	0.255	-0.057	0.147
Présence du lexème « loin que »	0.074	-0.005	0.060
Présence du lexème « lui »	-0.068	-0.091	-0.026
Présence du lexème « mais »	-0.127	-0.069	0.048
Présence du lexème « me »	-0.181	-0.044	-0.187
Présence du lexème « mes »	-0.160	-0.054	-0.186
Présence du lexème « miens »	-0.082		-0.072
Présence du lexème « mon »	-0.193	0.007	-0.119
Présence du lexème « ni »	0.021	0.119	0.144
Présence du lexème « nos »	0.161	-0.052	0.076
Présence du lexème « notre »	0.053	0.040	0.039
Présence du lexème « nulle part »	0.078	0.014	0.053
Présence du lexème « néanmoins »	0.157	0.020	0.061
Présence du lexème « ou bien »	-0.057	0.130	-0.049
Présence du lexème « par exemple »	0.259	0.006	0.136
Présence du lexème « par la suite »	0.137	0.022	0.063
Présence du lexème « parmi »	0.285	0.192	0.129
Présence du lexème « quand »	-0.076	0.039	-0.128
Présence du lexème « quelques »	0.081	-0.188	0.010
Présence du lexème « qui »	0.234	0.235	0.253
Présence du lexème « quoi »	0.109	0.126	0.099
Présence du lexème « s »	0.149	-0.145	0.108
Présence du lexème « sauf que »	-0.003	0.039	
Présence du lexème « son »	0.169	0.037	0.086
Présence du lexème « sous »	-0.011	-0.013	0.001
Présence du lexème « toi »	-0.177	0.099	-0.182
Présence du lexème « ton »	-0.037	0.081	-0.176
Présence du lexème « tous »	-0.012	0.039	-0.021
Présence du lexème « tout »	-0.077	0.092	-0.020
Présence du lexème « tout de suite »	-0.035	0.021	0.008
Présence du lexème « tout à coup »	-0.105	-0.129	-0.111
Présence du lexème « toute »	-0.075	0.160	-0.018
Présence du lexème « toutefois »	0.232	-0.151	0.134
Présence du lexème « tu »	-0.206	0.072	-0.295

Tableau 1 (suite)

Coefficients de corrélation de Pearson établis entre la classe d'enseignement et un ensemble de variables

Variables	Pri-maire	Secon-daire	Ensemble
Présence du lexème « vers »	-0.004	-0.248	-0.010
Présence du lexème « vos »	-0.049	0.038	0.104
Présence du lexème « vous »	0.121	0.078	0.200
Présence du lexème « vu »	-0.071	-0.079	-0.063
Présence du lexème « au-dessus » ou « au-dessus de »	0.124	-0.325	0.068
Présence du lexème « depuis » ou « depuis que »	0.177	0.011	0.113
Présence du lexème « parce qu'' » ou « parce que »	-0.016	0.165	0.007
Présence du lexème « pour » ou « pour qu'' »	0.151	0.063	0.039
Présence du lexème « près » ou « près de »	-0.053	-0.285	-0.140
Présence du lexème « voilà » ou « voilà pourquoi »	-0.063	0.045	-0.046
B- Certaines catégories grammaticales			
Présence d'adjectifs	0.092	0.139	0.096
Présence d'articles définis	0.175	-0.143	0.094
Présence d'articles généraux	-0.099		-0.223
Présence de conjonctions	0.176	0.102	0.157
Présence de verbes conjugués	-0.405	-0.075	-0.362
Présence de déterminants numéraux	0.229	-0.170	0.165
Présence de noms communs	0.049	0.037	-0.033
Présence de noms propres	-0.044	-0.049	-0.010
Présence de pronoms	0.173	-0.172	0.165
Présence de pronoms personnels	-0.355	-0.018	-0.405
Présence de pronoms relatifs	0.271	0.301	0.320
C- Longueurs moyennes			
Longueur moyenne des phrases	0.689	0.137	0.220
Longueur moyenne des paragraphes	0.083	0.011	0.266
D- Phrases possédant certaines caractéristiques			
Présence de phrases avec le patron « je me »	0.097	0.010	0.150
Présence de phrases avec le patron « tu te »	0.011	0.029	0.081
Présence de phrases commençant par un pronom réfléchi à la 3e pers.	0.336	0.013	0.332
Présence de phrases avec le patron « nous nous»	0.135	-0.041	0.116
Présence de phrases avec le patron « vous vous »	0.112	0.083	0.137
Présence de phrases de 15 mots ou moins	-0.688	-0.142	-0.600
Présence de phrases de 16 à 20 mots	0.500	0.097	0.640
Présence de phrases de 21 à 25 mots	0.525	0.218	0.518
Présence de phrases de 26 à 30 mots	0.481	-0.143	0.168

Tableau 1 (suite)

Coefficients de corrélation de Pearson établis entre la classe d'enseignement et un ensemble de variables

	Variables	Pri-maire	Secon-daire	Ensemble
	Présence de phrases de plus de 30 mots	0.625	0.024	0.442
	Présence de phrases commençant par une préposition ou une conjonction	0.521	-0.005	0.468
	Présence de phrases commençant par un pronom de la 3e personne	0.510	0.231	0.539
	Présence de phrases contenant au moins quatre propositions	-0.226	0.037	-0.216
	Présence de phrases interrogatives contenant un pronom inversé	0.012	-0.082	-0.163
	Présence de phrases contenant au moins deux mots inconnus	0.349	0.139	0.372
	Présence de phrases contenant au moins deux pronoms relatifs	-0.078	0.229	-0.061
E-	**Autres caractéristiques**			
	Présence de mots inconnus	0.162	0.007	0.372
	Nombre total de mots	0.691	0.130	0.559
	Pourcentage de mots longs	0.660	0.142	0.547
	Indice de lisibilité Gunning	0.762	0.151	0.305
	Nombre de phrases	0.548	0.108	0.421
	NOMBRE DE TEXTES	400	279	679

Un premier examen du tableau 1 indique que les variables ne se comportent pas de la même manière lorsqu'on calcule des coefficients de corrélation de Pearson pour les textes rattachés à l'enseignement primaire ou ceux utilisés pour l'enseignement secondaire. Ce constat nous permet d'émettre certaines hypothèses quant à la constitution du corpus disponible ou bien aux caractéristiques propres à des textes utilisés au primaire ou au secondaire :

A- Cette étude sur la lisibilité des textes s'est d'abord concentrée au primaire. Le corpus contient d'ailleurs plus de textes rattachés à chacune des classes du primaire. Ces textes ont été choisi avec un soin particulier et ont fait l'objet d'examens attentifs à l'aide d'indices statistiques produits à intervalles réguliers. Pour leur part les textes du secondaire ont été inclus au corpus plus récemment; moins d'analyses préalables ont pu être réalisés sur ces derniers.

B- Cependant, nous croyons plutôt que certaines caractéristiques décrivant la structure d'un texte ne se présentent pas sur un continuum linéaire lorsqu'on établit des statistiques en fonction de la classe d'enseignement. Le tableau 2 illustre cette hypothèse à l'aide de deux exemples : le premier -- la longueur moyenne des

phrases -- indique une représentation peu linéaire du phénomène lorsque l'on considère l'ensemble des 679 textes[3]; le deuxième -- la proportion de phrases contenant 15 mots ou moins -- semble indiquer une répartition plus continue de la caractéristique par rapport aux classes d'enseignement[4]. Les coefficients de corrélation rapportés dans ce tableau synthétisent les tendances pour chacune des caractéristiques retenues.

Ces deux exemples illustrent bien une particularité des textes rattachés au secondaire : il semble y avoir une plus grande homogénéité au regard de plusieurs variables utilisées dans nos analyses statistiques pour ce sous-ensemble de textes. Il y aura lieu d'augmenter le nombre de textes du secondaire et de réaliser des analyses préliminaires avec uniquement ces textes afin d'identifier le jeu de variables caractérisant les classes du secondaire.

---

[3] Il est à noter que la valeur du coefficient de Pearson (0,689) indique cependant une forte relation au primaire entre la classe et la longueur moyenne des phrases.

[4] La tendance est forte pour les textes du primaire analysés séparément (-0,688) ainsi que pour l'ensemble du corpus (-0,600). Cette tendance est cependant négligeable (-0,143) lorsque l'on considère uniquement les textes du secondaire.

Tableau 2

Degré de linéarité d'un ensemble de statistiques
Une illustration

Les classes	Longueur moyenne des phrases (nombre de mots)	Phrases contenant 15 mots ou moins (%).
1	7,6	90,6
2	8,6	85,6
3	9,8	78,8
4	11,6	69,4
5	12,4	63,2
6	15,3	48,3
7	14,8	53,8
8	16,7	56,9
9	14,6	54,7
10	15,0	53,9
11[5]	31,6	45,6
r(primaire)	0,689	-0,688
r(secondaire)	0,137	-0,143
r(ensemble)	0,220	-0,600

Comme on peut le constater en examinant le tableau 1, certains coefficients de corrélation ont une valeur positive, d'autres une valeur négative. Ceci indique le sens du lien entre l'ensemble des statistiques choisies et la classe d'enseignement: lorsque le coefficient obtient une valeur positive, il y a une relation entre le niveau de scolarité et la valeur elle-même de la statistique; il s'agirait donc, lorsque le coefficient est significatif, d'une variable caractérisant les textes des classes supérieures. C'est la situation observée pour la première variable apparaissant au tableau 2 : les phrases sont généralement plus longues à la fin du secondaire. D'un autre côté, un coefficient de corrélation négatif indique une variable caractérisant les textes destinées aux plus jeunes élèves. Il en est ainsi pour la variable rendant compte de la proportion de phrases courtes (15 mots ou moins); la valeur du coefficient de corrélation s'établit dans ce cas à -0,600 pour l'ensemble des textes du corpus.

[5] Le nombre de textes ainsi que la présence de certains textes de genres littéraires très différents des autres textes du corpus peuvent expliquer les valeurs extrêmes de certaines statistiques des textes de la dernière année du secondaire

Le coefficient de corrélation de Pearson permet ici de mesurer l'importance des liens qui s'établissent entre la classe d'enseignement et les différentes variables disponibles. Nous avons conclu une relation entre une variable et son rattachement à une classe d'enseignement lorsque la valeur du coefficient était égale ou supérieure à 0,20. Le tableau 3 ne contient que les quarante-cinq variables répondant à ce critère pour au moins un sous-groupe de textes (ceux rattachées au primaire ou ceux rattachés au secondaire) ou bien pour l'ensemble du corpus.

Afin d'avoir une meilleure connaissance du comportement de chacune des quarante-cinq variables les plus fortement reliées de façon individuelle à la classe d'enseignement, le tableau 3 contient leurs valeurs moyennes pour chaque classe d'enseignement.

Tableau 3

Valeurs moyennes[6] par rapport aux classes d'enseignement
Les variables les plus fortement reliées à la classe

Variables	1	2	3	4	5	6	7	8	9	10	11
Lexèmes											
Présence de «!»	1,3	1,3	0,8	0,3	0,4	0,2	0,3	0,5	0,5	0,6	0,3
Présence de « « »	0,0	0,0	0,2	0,2	0,3	0,3	0,2	0,4	0,4	0,3	0,2
Présence de «.»	9,6	7,7	7,0	6,5	5,8	5,1	5,2	4,8	4,5	4,5	4,0
Présence de «car»	0,0	0,0	0,0	0,1	0,1	0,1	0,1	0,0	0,0	0,0	0,0
Présence de «certaines»	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Présence de «certains»	0,0	0,0	0,0	0,0	0,0	0,1	0,0	0,0	0,0	0,0	0,0
Présence de «comme»	0,5	0,6	0,8	0,9	1,0	1,1	1,2	1,1	1,0	1,1	1,4
Présence de «d'»	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Présence de «d'ailleurs»	2,4	2,3	2,9	3,3	3,3	3,8	3,8	3,4	3,6	3,2	4,1
Présence de «en»	0,6	0,7	0,9	0,8	1,9	1,1	1,0	0,9	1,0	1,0	1,0
Présence de «grâce à»	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Présence de «l'»	0,9	1,0	1,2	1,4	1,4	1,7	1,9	1,8	1,5	1,6	1,7
Présence de «leurs»	0,0	0,0	0,0	0,1	0,1	0,1	0,2	0,1	0,1	0,1	0,1
Présence de «par exemple»	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Présence de «parmi»	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Présence de «qui»	0,2	0,6	0,6	0,7	0.7	0,7	0,6	0,6	0,9	0,7	0,9
Présence de «toutefois»	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0

[6] Les valeurs ont été arrondies à une décimale. La valeur 0 indique un événement très rare lorsqu'on établit des statistiques par classe d'enseignement.

Variables	1	2	3	4	5	6	7	8	9	10	11
Présence de «tu»	1,3	0,6	1,0	0,6	0,4	0,3	0,1	0,1	0,2	0,2	0,1
Présence de «vers»	0,0	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,0
Présence de «vous»	0,0	0,1	0,0	0,1	0,0	0,2	0,1	0,2	0,4	0,4	0,2
Présence de «au-dessus» ou «au-dessous de»	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Présence de «près» ou «près de»	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,0	0,0	0,0	0,0
Catégories grammaticales											
Présence d'articles généraux	0,0	0,1	1,8	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Présence de verbes conjugués	13,2	13,2	8,1	11,2	10,9	9,3	9,5	10,5	10,0	9,8	9,2
Présence de déterminants numéraux	0,6	0,5	1,2	1,0	0,8	1,5	1,5	1,9	1,1	1,2	1,0
Présence de pronoms personnels	9,2	10,6	10,0	8,3	8,2	6,8	6,6	7,3	6,9	7,0	6,5
Présence de pronoms relatifs	0,8	1,8	2,0	2,0	2,1	1,9	1,8	1,9	2,3	2,4	2,5
Longueurs moyennes											
Longueur moyenne des phrases	7,6	8,6	9,8	11,6	12,4	15,3	14,8	16,7	14,6	15,0	31,6
Longueur moyenne des paragraphes	37,0	42,2	28,5	33,2	39.7	45,4	64,2	47,4	50,2	62,5	60,0
Phrases possédant certaines caractéristiques											
Phrases commençant par pron. réfléchi à la 3e personne	0,0	0,3	0,2	0,7	1,2	0,9	1,3	2,5	1,8	2,6	1,4
Phrases de 15 mots ou moins	90,6	85,6	78,8	69,4	63,2	48,3	53,8	56,9	54,7	53,9	45,6
Phrases de 16 à 20 mots	7,5	10,3	12,6	17,3	18,7	20,1	16,1	14,4	15,5	14,7	12,9
Phrases de 21 à 25 mots	1,2	2,4	4,7	7,6	9,5	14,0	10,6	8,7	10,7	9,8	10,7
Phrases de 26 à 30 mots	0,3	1,1	2,6	3,5	4,6	7,8	8,2	5,3	6,8	7,3	7,4
Phrases de plus de 30 mots	0,4	0,5	1,3	2,2	4,0	9,8	11,2	14,8	12,3	14,3	23,4
Phrases commençant par un préposition ou une conjonction	2,1	1,9	1,0	0,0	0,0	0,0	0,0	0,0	0,0	0,1	0,0

Variables	1	2	3	4	5	6	7	8	9	10	11
Phrases commençant par un pronom de la 3e personne	9,5	13,2	13,6	9,8	12,6	11,3	9,7	8,8	8,6	7,5	8,4
Phrases contenant au moins 4 propositions	0,8	1,9	0,6	2,9	4,2	5,1	6,1	10,1	7,1	8,0	12,5
Phrases contenant au moins 2 mots inconnus	1,2	4,7	2,5	10,1	17,4	23,7	35,4	34,8	34,0	38,3	36,7
Phrases contenant au moins 2 pronoms relatifs	5,7	10,6	21,8	18,4	20,4	26,7	26,8	24,6	26,5	30,0	12,9
Autres caractéristiques											
Présence de mots inconnus	1,3	4,1	1,8	3,9	5,3	4,9	8,1	7,4	7,8	8,6	7,7
Nombre total de mots	105	193	483	766	848	995	1079	1248	1471	1671	1333
Proportion de mots longs	7,6	5,3	7,0	9,9	9,6	10,8	10,8	9,1	10,6	11,1	11,8
Indice Gunning	3,8	5,6	6,7	8,6	8,8	10,5	10,2	10,3	10,0	10,4	17,3
Nombre de phrases	14,6	23,4	50,2	67,5	70,7	73,5	76,8	98,2	108,3	139,5	94,7
Nombre de textes	65	82	63	60	60	70	63	43	49	57	67

Une deuxième étape : des analyses multivariées

L'étape précédente a permis d'étudier l'importance des liens qui existent entre chaque variable sélectionnée et la complexité des textes telle que mesurée par son rattachement à la classe d'enseignement. Les prochaines analyses proposent d'inclure simultanément l'ensemble des variables énumérées au tableau 3 afin de déterminer les sous-ensembles qui peuvent le mieux expliquer le rattachement d'un texte à la classe d'enseignement.

Dans un premier temps, une analyse factorielle a été réalisée avec les variables apparaissant au tableau 3 pour chacun des regroupements de textes, c'est-à-dire ceux rattachés au primaire, ceux rattachés au secondaire et enfin pour l'ensemble du corpus. Pour chacun des ensembles de textes, nous fournissons les séries suivantes de données :

- les valeurs obtenues par les variables à deux facteurs;
- les valeurs obtenues à ces deux facteurs après l'exécution d'une rotation «varimax»;
- la localisation par rapport à un système d'axes des valeurs obtenues à ces deux facteurs[7].

[7] Pour chaque ensemble de textes, nous fournissons cette illustration graphique pour les valeurs des facteurs avant et après la rotation "varimax".

Tableau 4

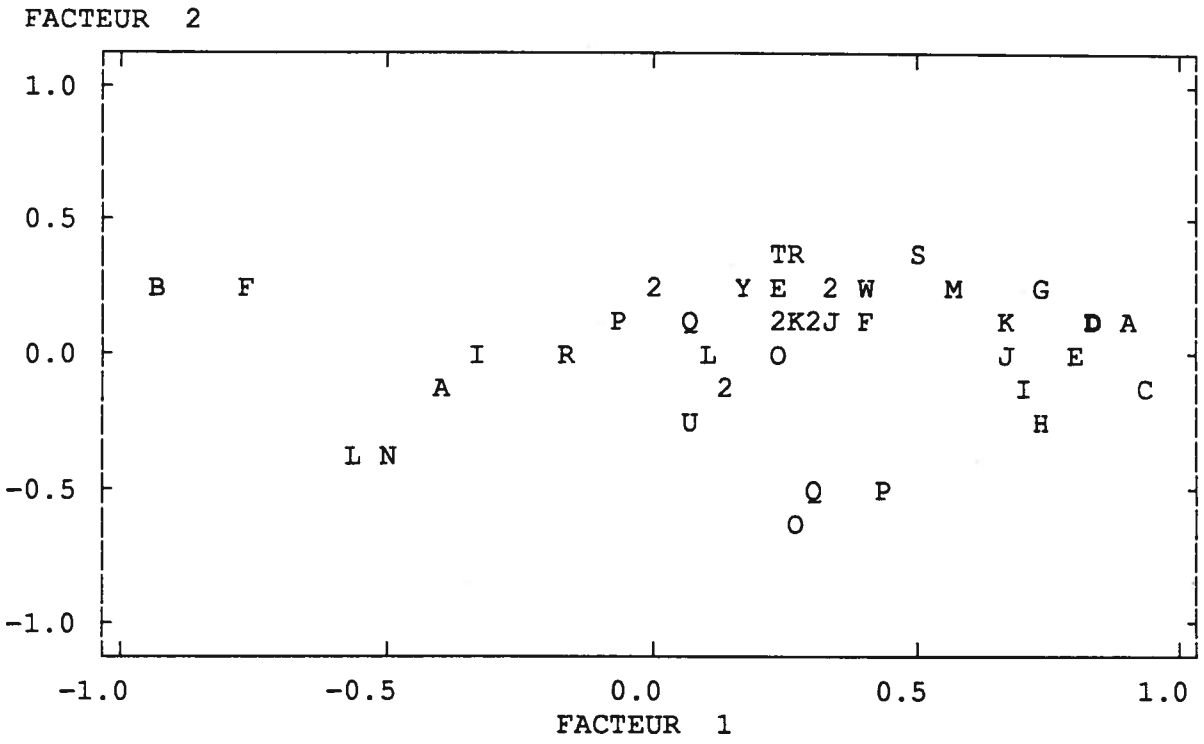
Valeurs obtenues à une analyse factorielle réalisée avec les textes du primaire

Les variables	Premier facteur	Second facteur
A-Indice Gunning	0.935	0.028
B-Phrases de 15 mots ou moins	-0.923	0.141
C-Longueur moyenne des phrases	0.921	-0.130
D-La classe	**0.838**	**0.010**
E-Phrases de 21 à 25 mots	0.786	-0.045
F-Présence du « . »	-0.754	0.237
G-Pourcentage de mots longs	0.736	0.180
H-Phrases avec deux pron. relatifs	0.723	-0.287
I-Phrases de plus de 30 mots	0.690	-0.221
J-Phrases de 16 à 20 mots	0.679	-0.099
K-Nombre total de mots	0.651	0.056
L-Présence de verbes conjugués	-0.583	-0.484
M-Phrases avec 2 mots inconnus	0.571	0.130
N-Présence de pronoms personnels	-0.503	-0.383
O-Présence de pronoms relatifs	0.272	-0.693
P-Phrases avec 4 propositions	0.435	-0.581
Q-Présence de « qui »	0.296	-0.578
R-Présence de déterminants numéraux	0.274	0.373
S-Présence de « de »	0.488	0.317
T-Présence des guillemets	0.250	0.266
U-Longueur moyenne des paragraphes	0.081	-0.263
V-Présence de « près/près de »	-0.010	0.187
W-Présence de « d' »	0.405	0.175
X-Présence d'articles généraux	-0.005	0.174
Y-Présence de mots inconnus	0.172	0.170
Z-Présence de « vous »	0.148	-0.155
a-Présence de « ! »	-0.391	-0.154
b-Présence de « ceux »	0.118	-0.139
c-Présence de « en »	0.333	0.133
d-Présence de « l' »	0.332	0.131
e-Présence de « parmi »	0.221	0.130
f-Nombre de phrases	0.401	0.125
g-Présence de « leurs »	0.225	0.120
h-Présence de « car »	0.309	0.119
i-Phrases commençant par pron. rel.	-0.336	-0.096
j-Présence de « certains »	0.326	0.091
k-Présence de « certaines »	0.257	0.086
l-Présence de « au-dessus/de »	0.084	-0.068
m-Présence de « grâce à »	0.225	0.066
n-Présence de « toutefois »	0.308	0.043
o-Présence de « par exemple »	0.245	-0.029
p-Phrases commençant par un pron. 3e p	-0.060	0.019
q-Présence de « aujourd'hui »	0.059	0.014

r-Présence de « tu »	-0.183	-0.003
Pourcentage de variance expliquée	22.813	5.742

Figure 1

Valeurs obtenues à une analyse factorielle réalisée avec les textes du primaire
Représentation graphique
Facteurs établis avant la rotation «varimax»



NOTE : La lettre en caractère gras représente la classe d'enseignement

Tableau 5

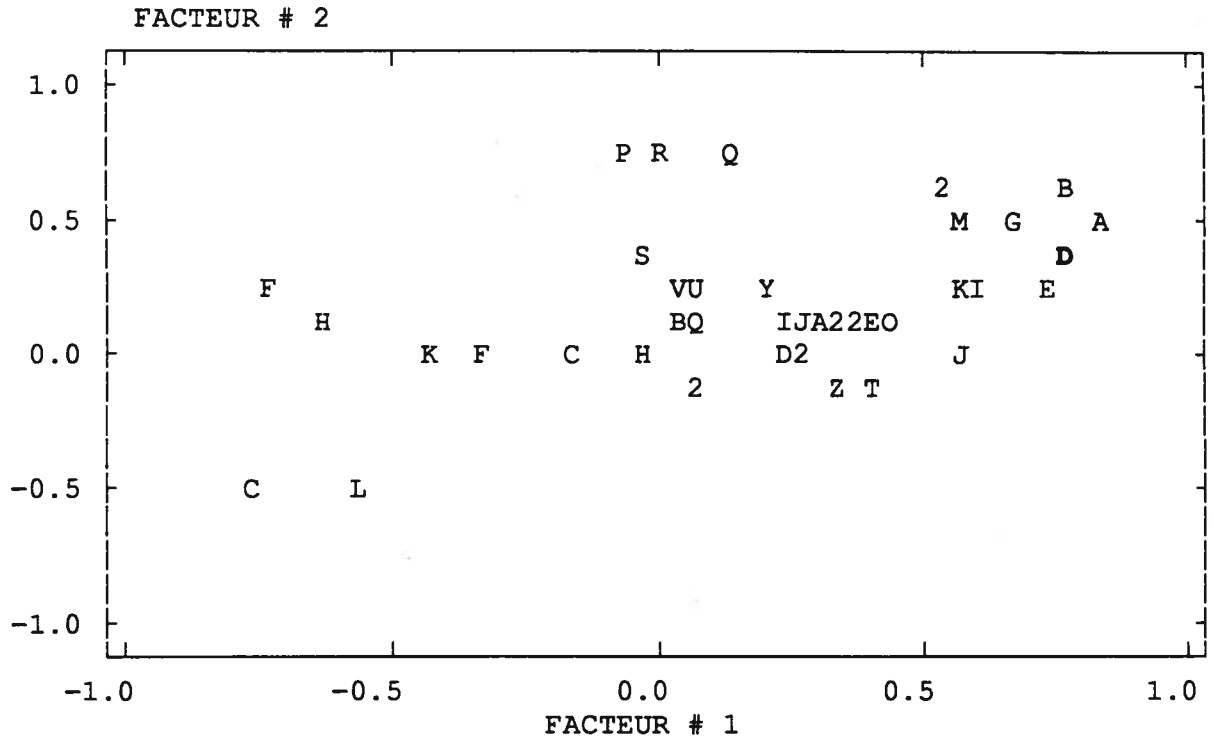
Valeurs obtenues à une analyse factorielle réalisée avec les textes du primaire
Facteurs établis après la rotation «varimax»

Les variables	Premier facteur	Second facteur
A-Indice Gunning	0.848	0.395
B-Longueur moyenne des phrases	0.764	0.530
C-Phrases de 15 mots ou moins	-0.761	-0.540
D-La classe	**0.754**	**0.368**
E-Pourcentage de mots longs	0.739	0.170
F-Présence de conjugués	-0.738	0.171
G-Phrases de 21-25 mots	0.682	0.393
H-Présence de pron. personnels	-0.622	0.116
I-Nombre de mots	0.607	0.242
J-Présence de «de»	0.578	-0.064
K-Phrases avec 2 mots inconnus	0.569	0.141
L-Présence du «.»	-0.567	-0.550
M-Phrases de 16-20 mots	0.562	0.394
N-Phrases avec 2 pron. rel.	0.517	0.581
O-Phrases de plus de 30 mots	0.517	0.508
P-Présences de pron. relatifs	-0.069	0.741
Q-Phrases avec 4 propositions	0.127	0.715
R-Présence de «qui»	0.005	0.649
S-Longueur des paragraphes	-0.046	0.271
T-Présence de déterminants num.	0.412	-0.210
U-Présence de «vous»	0.062	0.205
V-Présence de «ceux»	0.043	0.177
W-Présence de «près/de»	0.076	-0.172
X-Présence d'art. généraux	0.074	-0.158
Y-Présence de «par exemple»	0.206	0.136
Z-Présence de guillemets	0.343	-0.125
a-Présence de «toutefois»	0.295	0.100
b-Présence de «au-dessus/de»	0.044	0.098
c-Présence de «tu»	-0.164	-0.080
d-Présence de mots inconnus	0.230	-0.075
e-Nombre de phrases	0.414	0.069
f-Phrases qui commencent par conj/prép	-0.344	-0.065
g-Présence de «certains»	0.332	0.065
h-Phrases qui commencent par pron 3e p.	-0.046	-0.044
i-Présence de «grâce à»	0.230	0.042
j-Présence de «certaines»	0.268	0.039
k-Présence de «!»	-0.418	-0.038
l-Présence de «l'»	0.355	0.033
m-Présence de «car»	0.329	0.032
n-Présence de «en»	0.357	0.031
o-Présence de «d'»	0.441	0.026
p-Présence de «parmi»	0.256	-0.017

q-Présence de «aujourd'hui»	0.059	0.014
r-Présence de «leurs»	0.255	-0.006
Pourcentage de variance expliquée	19.367	9.18

Figure 2

Valeurs obtenues à une analyse factorielle réalisée avec les textes du primaire
Représentation graphique
Facteurs établis après la rotation «varimax»



NOTE : La lettre en caractère gras représente la classe d'enseignement

Tableau 6

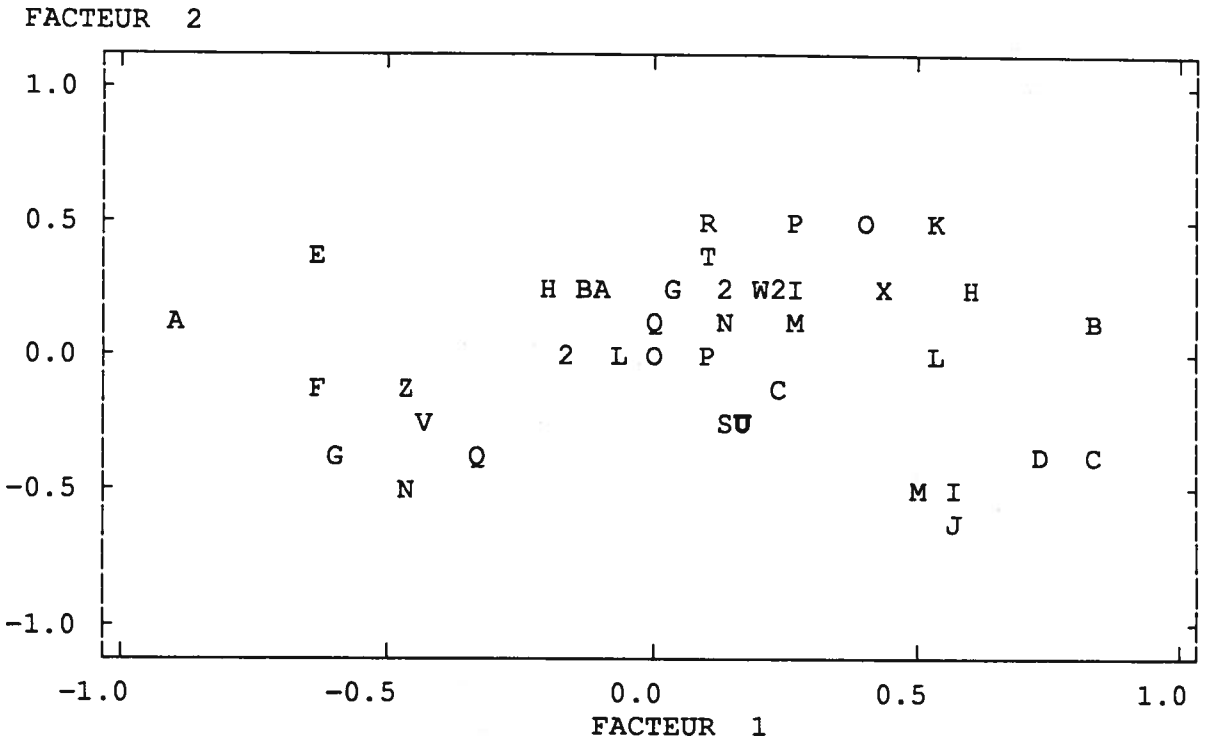
Valeurs obtenues à une analyse factorielle réalisée avec les textes du secondaire
Facteurs établis avant la rotation «varimax»

Les variables	Premier facteur	Second facteur
A-Phrases de 15 mots ou moins	-0.898	0.001
B-Phrases avec 2 mots inconnus	0.839	0.010
C-Phrases de plus de 30 mots	0.819	-0.405
D-Phrases avec 2 pron. rel.	0.732	-0.430
E-Présence du «.»	-0.648	0.302
F-Nombre de phrases	-0.636	-0.235
G-Présence de conjugués	-0.600	-0.398
H-Présence des guillemets	0.595	0.225
I-Indice Gunning	0.561	-0.573
J-Phrases avec 4 prop.	0.557	-0.704
K-Pourcentage de mots longs	0.547	0.476
L-Présence de «d'»	0.545	-0.067
M-Longueur des phrases	0.503	-0.622
N-Présence des pron. pers.	-0.455	-0.579
O-Phrases de 21-25 mots	0.413	0.428
P-Présence de «l'»	0.278	0.390
Q-Présence de «tu»	-0.333	-0.389
R-Phrases avec 16-20 mots	0.086	0.384
S-Présence de pron. rel.	0.126	-0.342
T-Présence de déterm. numéraux	0.110	0.334
U-La classe	**0.178**	**-0.316**
V-Présence de «!»	-0.428	-0.284
W-Présence de «toutefois»	0.202	0.230
X-Présence de mots inconnus	0.420	0.227
Y-Présence de «certaines»	0.219	0.200
Z-Nombre de mots	-0.459	-0.198
a-Présence de «près/de»	-0.094	0.185
b-Présence de «au-dessus/de»	-0.130	0.169
c-Présence de «qui»	0.222	-0.169
d-Présence de «par exemple»	0.137	0.162
e-Présence de «grâce à»	0.128	0.156
f-Présence de «certains»	0.248	0.153
g-Présence de «en»	0.046	0.144
h-Phrases qui commence par pron 3e p.	-0.212	0.142
i-Longueur des paragraphes	0.260	0.136
j-Présence de «vous»	-0.183	-0.096
k-Phrases qui commencent par prép./conj	-0.155	-0.076
l-Présences des guillemets	-0.077	-0.056
m-Présence de «ceux»	0.263	0.055
n-Présence de «parmi»	0.129	0.051
o-Présence de «aujourd'hui»	0.016	-0.023

p-Présence de «leurs»	0.098	-0.017
q-Présence de «car»	0.006	0.014
Pourcentage de variance expliquée	17.361	9.202

Figure 3

Valeurs obtenues à une analyse factorielle réalisée avec les textes du secondaire
Représentation graphique
Facteurs établis avant la rotation «varimax»



NOTE : La lettre en caractère gras représente la classe d'enseignement

Tableau 7

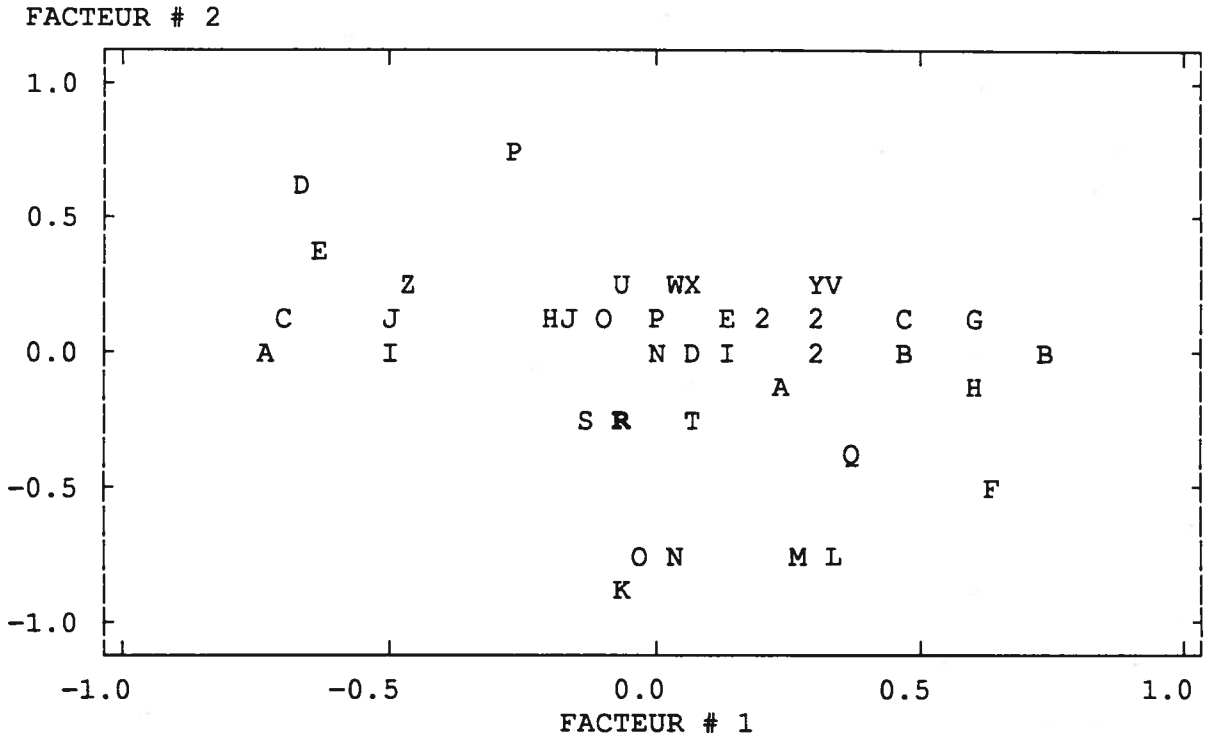
Valeurs obtenues à une analyse factorielle réalisée avec les textes du secondaire
Facteurs établis après la rotation «varimax»

Les variables	Premier facteur	Second facteur
A-Présence de pronoms personnels	-0.725	-0.123
B-Pourcentage de mots longs	0.725	-0.016
C-Présence de conjugués	-0.711	0.109
D-Phrases de 15 mots ou moins	-0.664	0.605
E-Nombre de phrases	-0.629	0.254
F-Phrases avec 2 mots inconnus	0.628	-0.556
G-Phrases de 21-25 mots	0.593	0.039
H-Présence des guillemets	0.592	-0.233
I-Présence de «tu»	-0.508	-0.064
J-Présence de «!»	-0.508	0.077
K-Phrases avec 4 propositions	-0.061	-0.895
L-Phrases de plus de 30 mots	0.334	-0.851
M-Phrases avec 2 pron. relatifs	0.253	-0.810
N-Indice Gunning	0.030	-0.801
O-Longueur des phrases	-0.046	-0.798
P-Présence du «.»	-0.277	0.659
Q-Présence de «d'»	0.358	-0.416
R-La classe	**-0.081**	**-0.354**
S-Présence de pron. relatifs	-0.136	-0.338
T-Présence de «qui»	0.050	-0.274
U-Phrases qui commencent par pron 3e p.	-0.061	0.247
V-Phrases de 16-20 mots	0.322	0.226
W-Présence de «au-dessus/de»	0.018	0.213
X-Présence de «près/de»	0.054	0.200
Y-Présence de déterminants num.	0.306	0.173
Z-Nombre de mots	-0.473	0.162
a-Présence de «ceux»	0.232	-0.136
b-Présence de mots inconnus	0.464	-0.114
c-Présence de «l'»	0.468	0.102
d-Présence de «leurs»	0.061	-0.079
e-Présence de «en»	0.131	0.075
f-Longueur des paragraphes	0.284	-0.074
g-Présence de «certains»	0.286	-0.053
h-Présence de «vous»	-0.200	0.051
i-Présence de «parmi»	0.130	-0.049
j-Phrases qui commencent par prep/conj.	-0.165	0.048
k-Présence de «toutefois»	0.304	0.035
l-Présence de «grâce à»	0.199	0.030
m-Présence de «par exemple»	0.210	0.028
n-Présence de «aujourd'hui»	-0.004	-0.028
o-Présence des guillemets	-0.095	0.011
p-Présence de «car»	0.014	0.006

q-Présence de «certaines»	0.297	0.001
Pourcentage de variance expliquée	13.673	12.889

Figure 4

Valeurs obtenues à une analyse factorielle réalisée avec les textes du secondaire
Représentation graphique
Facteurs établis après la rotation «varimax»



NOTE : La lettre en caractère gras représente la classe d'enseignement

Tableau 8

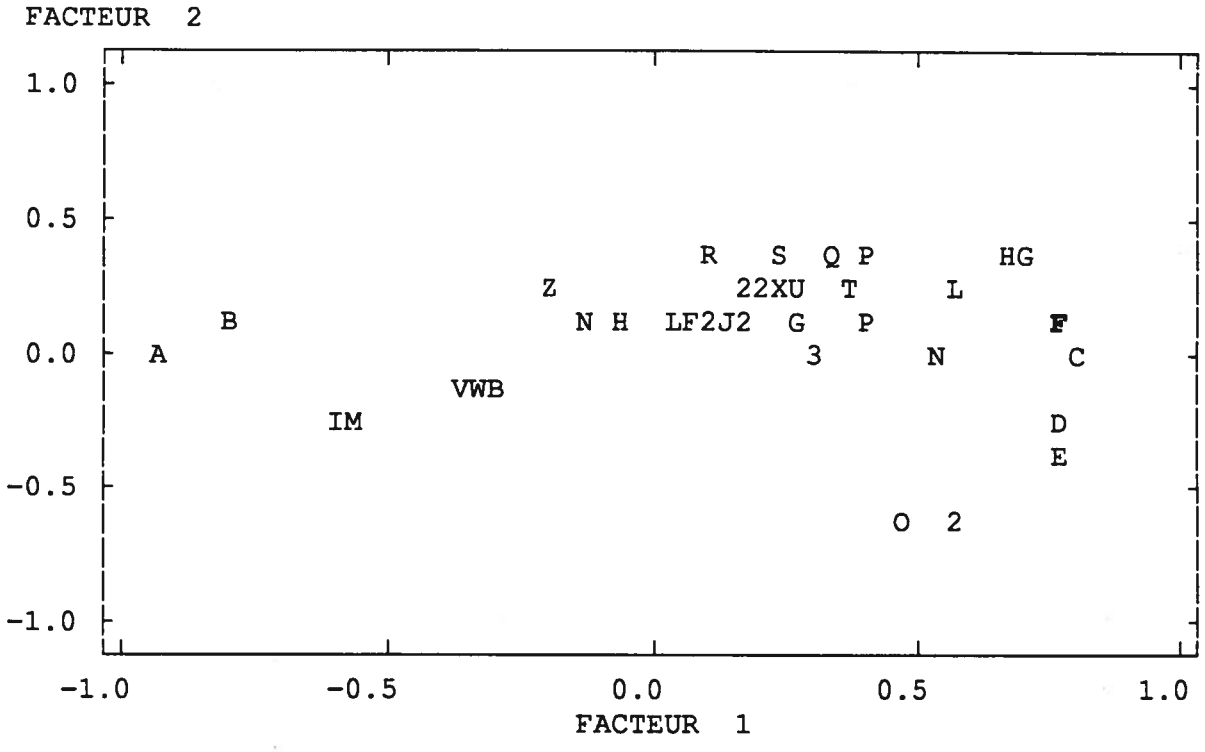
Valeurs obtenues à une analyse factorielle réalisée avec l'ensemble des textes
Facteurs établis avant la rotation «varimax»

Les variables	Premier facteur	Second facteur
A-Phrases de 15 mots ou moins	-0.923	-0.016
B-Présence du «.»	-0.801	0.029
C-Phrases avec 2 mots inconnus	0.795	-0.107
D-Phrases avec 2 pron. relatifs	0.782	-0.273
E-Phrases de plus de 30 mots	0.775	-0.458
F-La classe	**0.755**	**0.101**
G-Pourcentage de mots longs	0.701	0.340
H-Phrases de 21-25 mots	0.664	0.318
I-Présences de conjugués	-0.593	-0.315
J-Phrases avec 4 propositions	0.565	-0.696
K-Indice Gunning	0.562	-0.651
L-Présence des guillemets	0.560	0.140
M-Présence de pron. personnels	-0.557	-0.365
N-Présence de «d'»	0.520	-0.030
O-Longueur des phrases	0.456	-0.727
P-Phrases de 16-20 mots	0.404	0.352
Q-Nombre de mots	0.342	0.275
R-Nombre de phrases	0.106	0.274
S-Présence de det. numéraux	0.237	0.255
T-Présence de «l'»	0.380	0.235
U-Présence de «en»	0.268	0.207
V-Présence de «!»	-0.378	-0.183
W-Phrases qui commencent par prép/conj.	-0.317	-0.176
X-Présence de «certains»	0.232	0.168
Y-Présence de «certaines»	0.186	0.159
Z-Phrases qui commencent par pron 3e p.	-0.203	0.148
a-Présence de «car»	0.172	0.147
b-Présence de «tu»	-0.294	-0.146
c-Présence de «grâce à»	0.155	0.137
d-Présence de «par exemple»	0.192	0.134
e-Présence de pron. relatifs	0.309	-0.119
f-Présence de «au-dessus/de»	0.066	0.117
g-Présence de «toutefois»	0.271	0.113
h-Présence de «près/de»	-0.083	0.107
i-Présence des guillemets	0.106	0.102
j-Présence de «parmi»	0.141	0.100
k-Présence de «vous»	0.087	0.065
l-Présence de «aujourd'hui»	0.050	0.064
m-Présence de «leurs»	0.173	0.054
n-Présence d'articles généraux	-0.126	0.045
o-Présence de «qui»	0.311	-0.043
p-Présence de mots inconnus	0.390	0.037

q-Longueur des paragraphes	0.307	-0.037
r-Présence de «ceux»	0.179	0.027
Pourcentage de variance expliquée	19.735	6.727

Figure 5

Valeurs obtenues à une analyse factorielle réalisée avec l'ensemble des textes
Représentation graphique
Facteurs établis avant la rotation «varimax»



NOTE : La lettre en caractère gras représente la classe d'enseignement

Tableau 9

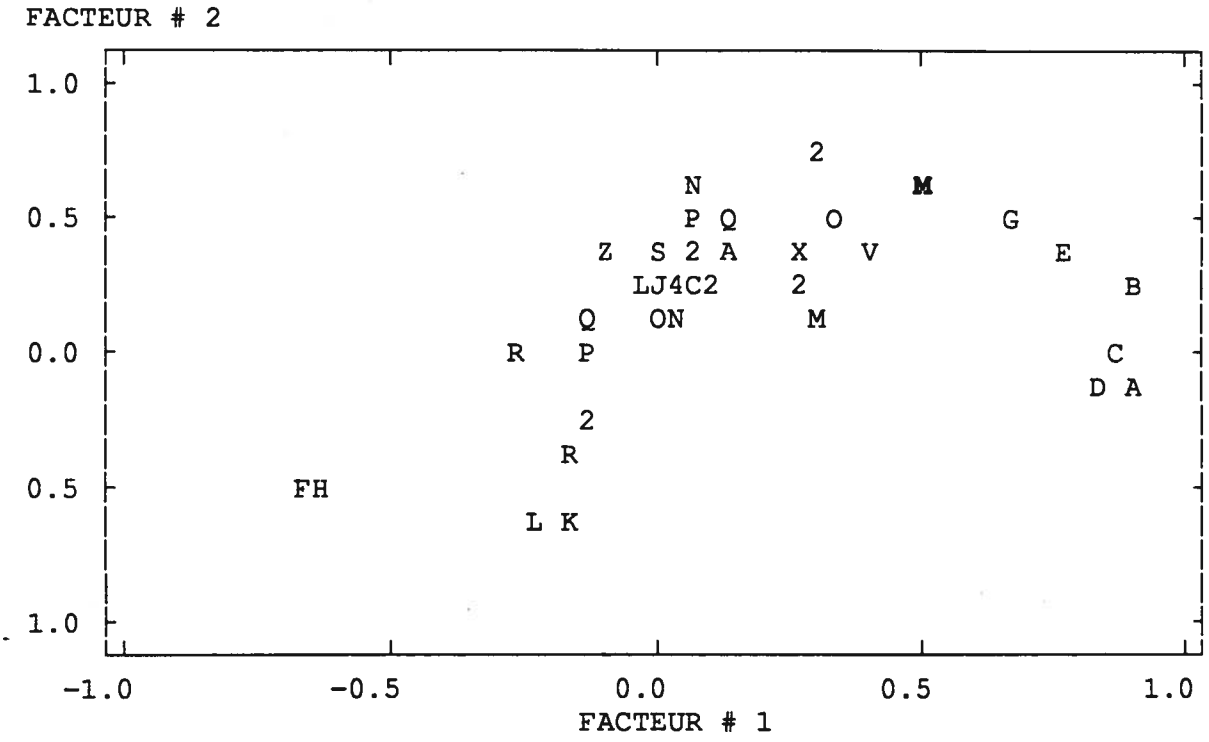
Valeurs obtenues à une analyse factorielle réalisée avec l'ensemble des textes
Utilisation de la rotation «varimax»

Les variables	Premier facteur	Second facteur
A-Phrases qui contiennent 4 prop.	0.884	-0.146
B-Phrases avec plus de 30 mots	0.883	0.172
C-Indice Gunning	0.852	-0.114
D-Longueur des phrases	0.823	-0.241
E-Phrases avec 2 pron. relatifs	0.767	0.315
F-Phrases de 15 mots ou moins	-0.680	-0.624
G-Phrases avec 2 mots inconnus	0.666	0.448
H-Présence du «.»	-0.619	-0.510
I-Pourcentage de mots longs	0.299	0.720
J-Phrases de 21-25 mots	0.286	0.678
K-Présence de pron. personnels	-0.175	-0.642
L-Présence de conjugués	-0.235	-0.629
M-La classe	**0.498**	**0.576**
N-Phrases de 16-20 mots	0.069	0.531
O-Présence des guillemets	0.326	0.477
P-Nombre de mots	0.073	0.433
Q-Présence de «l'»	0.128	0.428
R-Présence de «!»	-0.161	-0.388
S-Présence de déterminants num.	0.008	0.348
T-Phrases qui commencent par prép/conj	-0.121	-0.342
U-Présence de «en»	0.063	0.332
V-Présence de «d'»	0.409	0.323
W-Présence de «tu»	-0.123	-0.304
X-Présence de mots inconnus	0.267	0.287
Y-Présence de «certains»	0.062	0.280
Z-Nombre de phrases	-0.102	0.276
a-Présence de «toutefois»	0.128	0.264
b-Présence de «certaines»	0.034	0.242
c-Présence de «par exemple»	0.054	0.228
d-Présence de «car»	0.031	0.224
e-Présence de «grâce à»	0.025	0.206
f-Longueur des paragraphes	0.254	0.177
g-Présence de «qui»	0.261	0.174
h-Présence de «parmi»	0.039	0.169
i-Présence de «leurs»	0.093	0.155
j-Présence des guillemets	0.012	0.146
k-Présence de «ceux»	0.116	0.138
l-Présence de «au-dessus/de»	-0.029	0.131
m-Présence de pron. relatifs	0.310	0.116
n-Présence de «vous»	0.022	0.107
o-Présence de «aujourd'hui»	-0.005	0.081
p-Présence d'art. généraux	-0.125	-0.050

q-Présence de «près/de»	-0.133	0.025
r-Phrases qui commencent par pron. 3e p	-0.250	-0.024
Pourcentage de variance expliquée	14.008	12.455

Figure 6

Valeurs obtenues à une analyse factorielle réalisée avec l'ensemble des textes
Représentation graphique
Facteurs établis après la rotation «varimax»



NOTE : La lettre en caractère gras représente la classe d'enseignement

Les tableaux 10 et 11 présentent un sommaire des valeurs jugées significatives au facteur le plus relié à la classe d'enseignement; un seuil de 0,300 a été retenu comme critère de sélection. Nous présentons ces statistiques avant et après l'exécution de la rotation «varimax». Nous avons inclus la classe parmi les variables au moment de l'analyse factorielle. Il est alors possible de situer les valeurs obtenues aux différentes variables par rapport à la classe d'enseignement.

L'examen des valeurs obtenues avant la rotation «varimax» indique que le facteur numéro 1 caractérise généralement la difficulté d'un texte telle que mesurée par son rattachement à la classe d'enseignement. Le corpus du secondaire se comporte différemment. Après l'exécution d'une rotation «varimax», le premier facteur ne caractérise que les textes du primaire.

Tableau 10

Valeurs jugées significatives[8] obtenues au facteur caractérisant la classe d'enseignement

Facteurs établis avant la rotation «varimax»

Variables	Primaire (Facteur #1)	Secondaire (Facteur #2)	Ensemble (Facteur #1)
Présence de «!»	-0,391		-0,378
Présence des guillemets			0,560
Présence du point	-0,754	0,302	-0,801
Présence de «car»	0,309		
Présence de «certains»	0,326		
Présence de «d'»	0,405		0,520
Présence de «de»	0,488		0,560
Présence de «en»	0,333		
Présence de «l'»	0,332	0.390	0,380
Présence de «qui»			0,311
Présence de «toutefois»	0,308		
Présence de «tu»		-0,389	
Présence des conjugués	-0,583	-0,398	-0,593
Présence des dét. numéraux		0,334	
Présence des pron. pers.	-0,503	-0,579	-0,557
Présence des pron. relatifs		-0,342	0,309
Présence des mots inconnus			0,390
Nombre de mots	0,651		0,342
Longueur des phrases	0,921	-0,622	0,456
Longueur des paragraphes			0,307
Présence des mots longs	0,736	0,476	0,701
Indice Gunning	0,935	-0,573	0,562
Nombre de phrases	0,401		
Phrases de 15 mots ou moins	-0,923		-0,923

[8] Un seuil de 0,300 a été retenu comme critère de sélection.

Variables	Primaire (Facteur #1)	Secondaire (Facteur #2)	Ensemble (Facteur #1)
Phrases de 16-20 mots	0,679	0,384	0,404
Phrases de 21-25 mots	0,786	0,428	0,664
Phrases de plus de 30 mots	0,690	-0,405	0,775
Phrases commençant par pron. 3e pers.	-0,336		-0,317
Phrases avec 4 propositions	0,435	-0,704	0,565
Phrases avec 2 mots inconnus	0,571		0,795
Phrases avec 2 pron relatifs	0,723	-0,430	0,782
La classe d'enseignement	**0,838**	**-0,316**	**0,755**

Tableau 11

Valeurs jugées significatives[9] obtenues au facteur caractérisant la classe d'enseignement

Facteurs établis après la rotation «varimax»

Variables	Primaire (Facteur #1)	Secondaire (Facteur #2)	Ensemble (Facteur #2)
Présence de «!»	-0,418		-0,388
Présence des guillemets	0,343		
Présence du point	-0,567	0,659	-0,510
Présence de «car«	0,329		
Présence de «certains»	0,332		
Présence de «d'»	0,441	-0,416	0,323
Présence de «de»	0,578		0,477
Présence de «en»	0,357		0,332
Présence de «l'»	0,355		0,428
Présence de «tu»			-0,304
Présence de conjugués	-0,738		-0,629
Présence de deter. numéraux	0,412		0,348
Présence de pron. personnels	-0,622		-0,642
Présence de pron. relatifs	-0,338		
Nombre de mots	0,607		0,433
Longueur des phrases	0,764	-0,798	
Pourcentage de mots longs	0,739		0,720
Indice Gunning	0,848	-0,801	
Nombre de phrases	0,414		
Phrases de 15 mots ou moins	-0,761	0,605	-0,624
Phrases de 16-20 mots			0,532
Phrases de 21-25 mots	0,682		0,678

[9] Un seuil de 0,300 a été retenu comme critère de sélection.

Variables	Primaire (Facteur #1)	Secondaire (Facteur #2)	Ensemble (Facteur #2)
Phrases de plus de 30 mots	0,517	-0,851	0,678
Phrases commençant par pron. 3e personne			-0,342
Phrases contenant 4 prop.		-0,895	
Phrases avec 2 mots inconnus	0,569	-0,556	0,448
Phrases avec 2 pron. relatifs	0,517	-0,810	0,315
La classe d'enseignement	**0,754**	**-0,353**	**0,576**

Calcul d'indices de calibrage

C'est ainsi que 32 variables ont été jugées plus fortement reliées à la classe d'enseignement[10]. Le lien observé s'établit avec l'ensemble du corpus ou avec l'un des sous-corpus tel que déterminé par le rattachement à l'ordre d'enseignement.

Il est à noter que les pourcentages de variance expliquée par les deux facteurs retenues pour faire cette analyse sont faibles pour chacun des sous-ensembles de textes considérés. Ces résultats nous indiqueraient que le rattachement d'un texte à une classe d'enseignement s'effectue à l'aide de plusieurs critères dont certains ne sont pas mesurés par les variables utilisées dans cette analyse. Par ailleurs, il faut prendre en considération le fait que plusieurs variables sont fortement corrélées entre elles, indiquant qu'il y a redondance lorsque l'on choisit un nombre élevé de variables pour réaliser une analyse multivariée.

La technique de régression multiple permet de diminuer cet effet de redondance. C'est donc avec ce sous-ensemble de trente-deux variables que nous avons réalisé des analyses de régression multiple avec une méthode de sélection des variables (il s'agit de l'option «Stepwise»). Un des avantages de telles analyses est qu'elles permettent de connaître la variable la plus associée à la classe d'enseignement et de voir si l'ajout d'autres variables contribue à expliquer une plus grande partie de cette variance. Il sera ensuite possible, si le taux de variance expliqué est élevé, de bâtir une équation qui rendra compte de l'ensemble des variables ainsi identifiées; l'utilisation d'une telle équation peut permettre le calcul pour chaque texte d'un indice rendant compte de sa difficulté telle que mesurée par son rattachement à la classe d'enseignement.

Nous avons réalisé trois séries d'analyses de régression multiple pour tenir compte de l'ensemble des textes du corpus ainsi que de chacun des sous-groupes formés par les textes du primaire et ceux du secondaire. Cette stratégie a été retenue afin d'estimer le comportement des variables par rapport à la classe à l'intérieur de chaque ordre d'enseignement.

Les variables reliées à la longueur du texte (nombre de mots et de phrases) n'ont pas été retranchées de cette analyse[11]. De plus, nous avons retenu l'indice Gunning même s'il s'agit là d'une statistique générée à partir de variables déjà incluses par ailleurs (longueur moyenne des phrases et proportion de mots longs). Il sera toujours possible d'exclure ces trois variables dans une autre phase d'analyse. D'un autre côté, nous avons ajouté la variable rendant compte de la proportion de mots non familiers même si cette caractéristique n'a pas été identifiée au moment de l'analyse factorielle. Des travaux antérieurs nous laissent croire qu'il s'agit là d'une caractéristique importante à considérer pour juger de la difficulté d'un texte. Il est à noter que la variable «proportion de phrases contenant au moins deux mots non familiers» a été identifiée comme étant fortement reliée à la classe pour les textes du primaire, ceux du secondaire et pour l'ensemble du corpus; il y a peut-être là redondance lorsque l'on inclut simultanément ces deux variables dans un modèle d'analyse.

Le tableau 12 présente la liste des variables retenues par la technique de régression multiple «stepwise» pour chacun des sous-ensembles de textes. Encore une fois, on constate un

---

10 Nous avons retenu l'ensemble des variables contenues dans les tableaux 10 et 11, c'est-à-dire celles retenues avant et après l'exécution de la rotation "varimax".

11 La longueur d'un texte s'avère être une variable fortement liée à la classe d'enseignement lorsqu'il s'agit de documents utilisés dans des situations d'apprentissage ou d'évaluation pédagogique.

comportement différent au secondaire : le pourcentage de variance expliquant la classe n'est que de 26,1 p. cent avec les onze variables retenues. Ce taux s'établit à 75,9 p. cent pour les textes du primaire et à 72,0 p. cent lorsque l'on utilise l'ensemble du corpus.

Tableau 12

Les variables retenues avec l'analyse de régression multiple
Utilisation de la technique «stepwise»

Les variables	Primaire	Secondaire	Ensemble
Présence de «!»			X
Présence des guillemets	X		X
Présence du «.»			X
Présence de «car»		X	
Présence de «d'»	X		
Présence de «en»	X		X
Présence de «l'»		X	X
Présence de «qui»	X		
Présence de «toutefois»		X	
Présence de «tu»	X		X
Présence de dét.numéraux		X	
Présence de pron. personnels		X	X
Présence de pron. relatifs	X	X	X
Présence de mots inconnus	X		
Nombre de mots	X		
Longueur des phrases		X	
Longueur des paragraphes			X
Pourcentage de mots longs	X	X	X
Indice Gunning	X		
Nombre de phrases	X	X	X
Phrases de 16-20 mots			X
Phrases de 21-25 mots	X		
Phrases de plus de 30 mots		X	
Phrases commençant par pron. 3e pers.	X		
Phrases avec 4 prop.	X	X	
Phrases avec 2 mots inconnus	X		X
Phrases avec 2 pron. rel.	X		
NOMBRE DE VARIABLES RETENUES	16	11	13

VARIANCE EXPLIQUÉE	75,9%	26,1%	72,0%

Les tableaux 13, 14 et 15 fournissent pour chaque ensemble de textes le résultat obtenu après chaque étape de l'analyse de régression multiple réalisée avec la technique «stepwise». Dans chaque cas, plusieurs étapes ont été nécessaires pour établir le choix du jeu de variables le plus prédicteur de la classe d'enseignement. Parfois, certaines variables ont été retranchées de l'analyse après avoir été préalablement choisies comme fortement reliées à la classe.

Tableau 13
L'analyse de régression multiple pour les textes du primaire
Utilisation de la technique «stepwise»

Étapes d'analyse	Variables	Pourcentage de variance
1	Présence du point	35,2
2	Nombre de mots	58,0
3	Longueur moyenne des phrases	65,8
4	Indice Gunning	69,3
5	Retranchée:Longueur des phrases	69,2
6	Phrases avec au moins 2 mots inconnus	70,7
7	Nombre de phrases	72,1
8	Retranchée : Nombre de mots	72,1
9	Phrases avec 21 à 25 mots	72,8
10	Présence de pronoms relatifs	73,5
11	Retranchée : Présence du point	73,4
12	Présence des guillemets	74,1
13	Longueur des paragraphes	74,5
14	Présence de «d'»	74,8
15	Nombre de mots	75,0
16	Pourcentage de mots longs	75,6
17	Présence de mots inconnus	75,9
18	Présence de «tu»	76,1
19	Phrases contenant au moins 4 prop.	76,3
20	Présence de «qui»	76,5
21	Présence de «en»	76,6
22	Phrases commençant par un pron de 3e p.	76,8
23	Retranchée: Longueur des paragraphes	76,7
24	Phrases contenant au moins 2 mots inconnus	76,8
Variance corrigée		75,9%

Tableau 14

L'analyse de régression multiple pour les textes du secondaire
Utilisation de la technique «stepwise»

Étapes d'analyse	Variables	Pourcentage de variance
1	Présence de pronoms relatifs	9,0
2	Phrases de plus de 30 mots	12,4
3	Nombre de phrases	16,3
4	Présence de «toutefois»	18,5
5	Pourcentage de mots longs	21,1
6	Présence de «car»	22,6
7	Présence de «l'»	24,0
8	Présence de déterminants numéraux	25,2
9	Phrases contenant 4 propositions	26,6
10	Longueur moyenne des phrases	28,0
11	Présence de pronoms personnels	29,1
Variance corrigée		26,1%

Tableau 15

L'analyse de régression multiple pour l'ensemble des textes du corpus Utilisation de la technique «stepwise»

Étapes d'analyse	Variables	Pourcentage de variance
1	Présence du point	40,2
2	Nombre de mots	54,2
3	Pourcentage de mots longs	59,0
4	Phrases de plus de 30 mots	63,4
5	Phrases avec au moins 2 mots inconnus	67,0
6	Nombre de phrases	68,1
7	Retranchée : Nombre de mots	68,0
8	Présence de pronoms relatifs	68,9
9	Présence des guillemets	69,6
10	Présence de «tu»	70,2
11	Présence de «!»	70,7
12	Phrases de 16 à 20 mots	71,6
13	Retranchée : Phrases de plus de 30 mots	71,5
14	Présence de pronoms personnels	71,9
15	Longueur des paragraphes	72,1
16	Présence de «en»	72,4
17	Présence de «l'»	72,5
Variance corrigée		72,0%

Les tableaux 16, 17 et 18 présentent les équations fournies suite à l'analyse de régression multiple[12]. Les valeurs calculées seront par la suite utilisées pour compiler des indices de calibrage des textes du corpus.

[12] Ces tableaux reproduisent les rapports fournis par Systat à la suite de la régression multiple "Stepwise".

Tableau 16

Équation de régression multiple
Les textes du primaire

DEP VAR: CLASSE N: 400 MULTIPLE R: 0.876 SQUARED MULTIPLE R: 0.768
ADJUSTED SQUARED MULTIPLE R: .759 STANDARD ERROR OF ESTIMATE: 0.850

VARIABLE	COEFFICIENT	STD ERROR	STD COEF	TOLERANCE	T	P(2 TAIL)
CONSTANT	-0.838	0.278	0.000	.	-3.015	0.003
V(7)	0.437	0.126	0.092	0.849	3.461	0.001
V(27)	0.149	0.069	0.059	0.795	2.151	0.032
V(35)	0.117	0.074	0.041	0.902	1.581	0.115
V(64)	-0.267	0.113	-0.086	0.454	-2.357	0.019
V(78)	-0.061	0.035	-0.047	0.831	-1.741	0.082
V(93)	0.262	0.060	0.169	0.401	4.356	0.000
V(94)	-0.019	0.010	-0.075	0.360	-1.832	0.068
V(95)	-0.002	0.000	-0.467	0.039	-3.771	0.000
V(98)	-0.117	0.038	-0.258	0.086	-3.074	0.002
V(99)	0.485	0.079	0.753	0.040	6.125	0.000
V100	0.032	0.005	0.762	0.048	6.756	0.000
C3	0.028	0.010	0.116	0.342	2.762	0.006
C7	0.008	0.005	0.045	0.871	1.718	0.087
C8	0.027	0.012	0.067	0.655	2.197	0.029
C10	0.029	0.006	0.260	0.236	5.143	0.000
C11	-0.009	0.006	-0.067	0.374	-1.653	0.099

ANALYSIS OF VARIANCE

SOURCE	SUM-OF-SQUARES	DF	MEAN-SQUARE	F-RATIO	P
REGRESSION	916.309	16	57.269	79.333	0.000
RESIDUAL	276.481	383	0.722		

Nom des variables

V(7) les guillemets
V(27) d'
V(35) en
V(64) qui
V(78) tu
V(93) pron. rel.
V(94) mots inconnus
V(95) nombre de mots
V(98) mots longs
v(99) indice Gunning
V100 nombre de phrases

C3 phrases 21-25 mots
C7 phrases débutant par 3e p.
C8 phrases avec 4 prop.
C10 phrases avec 2 mots inconnus
C11 phrases avec 2 pron. rel.

Tableau 17
Équation de régression multiple
Les textes du secondaire

DEP VAR: CLASSE N: 279 MULTIPLE R: 0.539 SQUARED MULTIPLE R: 0.291
ADJUSTED SQUARED MULTIPLE R: .261 STANDARD ERROR OF ESTIMATE: 1.282

VARIABLE	COEFFICIENT	STD ERROR	STD COEF	TOLERANCE	T	P(2 TAIL)
CONSTANT	8.309	0.589	0.000	.	14.110	0.000
V(20)	-1.801	0.815	-0.117	0.953	-2.210	0.028
V(43)	-0.325	0.107	-0.176	0.795	-3.041	0.003
V(77)	-5.632	1.841	-0.164	0.927	-3.059	0.002
V(88)	-0.194	0.064	-0.172	0.824	-3.034	0.003
V(92)	-0.089	0.046	-0.140	0.521	-1.961	0.051
V(93)	0.552	0.097	0.315	0.864	5.687	0.000
V(96)	0.008	0.003	0.194	0.416	2.432	0.016
V(98)	0.045	0.026	0.114	0.587	1.690	0.092
V100	0.003	0.001	0.199	0.732	3.294	0.001
C5	0.035	0.009	0.385	0.248	3.723	0.000
C8	-0.039	0.013	-0.335	0.203	-2.924	0.004

ANALYSIS OF VARIANCE

SOURCE	SUM-OF-SQUARES	DF	MEAN-SQUARE	F-RATIO	P
REGRESSION	179.649	11	16.332	9.942	0.000
RESIDUAL	438.616	267	1.643		

Nom des variables

V(20)	car
V(43)	l'
V(77)	toutefois
V(88)	dét. numéraux
V(92)	pron. pers.
V(93)	pron. rel.
V(96)	longueur des phrases
V(98)	mots longs
V100	nombre de phrases
C5	phrases de plus de 30 mots
C8	phrases avec 4 prop.

Tableau 18

Équation de régression multiple
L'ensemble des textes

DEP VAR: CLASSE N: 679 MULTIPLE R: 0.852 SQUARED MULTIPLE R: 0.725
ADJUSTED SQUARED MULTIPLE R: .720 STANDARD ERROR OF ESTIMATE: 1.704

VARIABLE	COEFFICIENT	STD ERROR	STD COEF	TOLERANCE	T	P(2 TAIL)
CONSTANT	5.381	0.571	0.000	.	9.426	0.000
V(6)	-0.329	0.065	-0.119	0.757	-5.108	0.000
V(7)	0.450	0.110	0.086	0.936	4.108	0.000
V(9)	-0.367	0.042	-0.279	0.402	-8.701	0.000
V(35)	0.286	0.113	0.054	0.892	2.532	0.012
V(43)	0.135	0.081	0.038	0.809	1.668	0.096
V(78)	-0.189	0.065	-0.063	0.876	-2.910	0.004
V(92)	-0.060	0.026	-0.061	0.609	-2.335	0.020
V(93)	0.351	0.072	0.113	0.755	4.848	0.000
V(97)	0.005	0.002	0.054	0.873	2.486	0.013
V(98)	0.063	0.022	0.081	0.536	2.926	0.004
V100	0.015	0.001	0.351	0.948	16.815	0.000
C2	-0.054	0.009	-0.142	0.723	-5.955	0.000
C10	0.050	0.004	0.350	0.479	11.912	0.000

ANALYSIS OF VARIANCE

SOURCE	SUM-OF-SQUARES	DF	MEAN-SQUARE	F-RATIO	P
REGRESSION	5097.952	13	392.150	135.125	0.000
RESIDUAL	1929.918	665	2.902		

Nom des variables

V(6) !
V(7) les guillemets
V(9) le point
V(35) en
V(43) l'
V(78) tu
V(92) pron. pers.
V(93) pron. rel.
V(97) longueur des paragraphes
V(98) mots longs
V100 nombre de phrases
C2 phrases de 16-20 mots
C10 phrases avec 2 mots inconnus

En utilisant les trois équations de régression calculées par l'analyse «stepwise» présentée plus haut, trois indices de calibrage ont été compilés. Nous désignons ainsi ces trois indices :

- l'indice SATO-CALIBRAGE/PRIMAIRE;
- l'indice SATO-CALIBRAGE/SECONDAIRE;
- l'indice SATO-CALIBRAGE/ENSEMBLE.

Le tableau 19 présente les valeurs moyennes de chacun de ces indices pour chaque sous-ensemble de textes compris dans le corpus.

Tableau 19
Valeurs moyennes des indices de calibrage par classe

Les classes d'enseignement	Nombre de textes	Indice Primaire	Indice Secondaire	Indice Ensemble
1	65	1,486	7,710	1,467
2	82	2,233	8,203	2,906
3	63	2,959	8,390	3,737
4	60	3,817	8,593	4,839
5	60	4,295	8,731	5,545
6	70	5,016	8,648	6,426
7	63	5,035	8,552	7,300
8	43	5,792	8,569	7,620
9	49	5,284	9,189	8,016
10	57	6,119	9,322	8,763
11	67	8,788	9,661	8,703
Primaire	400	3,260	8,368	4,105
Secondaire	279	6,318	9,090	8,111
Ensemble	679	4,517	8,665	5,751

L'examen des statistiques contenues dans le tableau 19 confirme des constats faits plus haut : les textes du secondaire sont encore une fois peu sensibles aux indices calculés à partir des analyses précédentes. Par ailleurs, l'indice du primaire et celui établi avec l'ensemble du corpus permettent de mieux rattacher les textes du primaire à leur classe d'origine.

L'analyse des coefficients de corrélation établis entre la classe d'enseignement et chaque indice de calibrage confirme ces conclusions. Le tableau 20 contient ces coefficients de corrélation.

Tableau 20
Coefficients de corrélation de Pearson

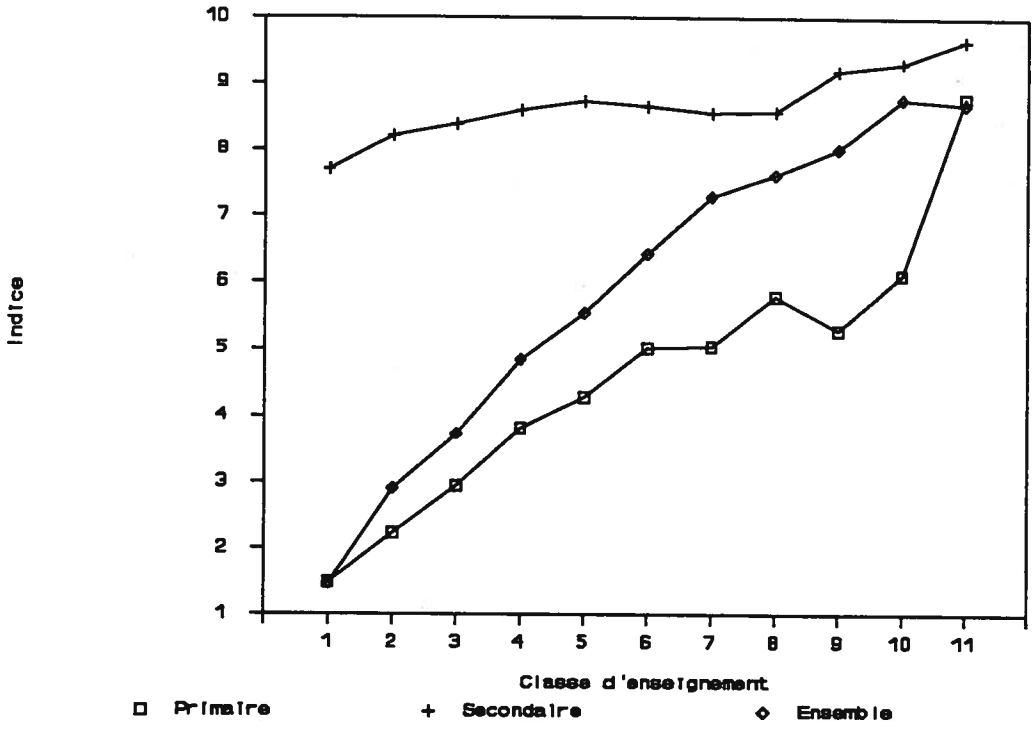
entre la classe d'enseignement et les indices de calibrage

Indices de calibrage	Corpus du primaire	Corpus du secondaire	Ensemble des textes
Primaire	0,874	0,165	0,378
Secondaire	0,404	0,539	0,566
Ensemble	0,817	0,339	0,852

Comme l'indique ce tableau, l'indice du primaire est efficace uniquement avec les textes rattachés aux classes du primaire. Par ailleurs, l'indice calculé avec l'ensemble des textes donne de bons résultats avec le corpus du primaire ainsi qu'avec l'ensemble des textes. D'un autre côté, les textes utilisés au secondaire sont plus difficiles à situer par rapport à l'une ou l'autre classe d'enseignement; l'indice «secondaire» est plus efficace (corrélation = 0,539) : le lien est cependant plutôt faible. Le tableau 19 semble indiquer qu'au secondaire, il y aurait deux groupes de textes : ceux rattachés à la 7^{ième} et à la 8^{ième} année d'une part, ceux appartenant aux trois autres classes d'autre part. La figure 7 illustre bien la distribution des indices de calibrage au regard de la classe.

Figure 7

Valeurs moyennes des indices de calibrage par classe



Conclusion

Les analyses statistiques que nous avons réalisées dans le cadre de cette étude portent sur un ensemble de documents utilisés dans des contextes d'apprentissage ou d'évaluation pédagogique. Ces analyses nous ont permis de faire les constats suivants :

A- Certains aspects propres à un texte peuvent très bien le caractériser au regard de sa lisibilité telle que mesurée par son rattachement à une classe d'enseignement. Les variables identifiées à la suite de nos analyses statistiques sont sensiblement semblables à celles retenues par des recherches antérieures proposant des indices de lisibilité (Gunning, Flesch, Dale-Chall). Ces variables concernent principalement la longueur des phrases, le choix des mots ainsi que la fréquence d'utilisation de mots appartenant à certaines catégories grammaticales.

B- Plusieurs variables caractérisant les textes sont interreliées. Par exemple, la fréquence d'utilisation des signes de ponctuation forte aura un lien avec la longueur moyenne des phrases. Les analyses réalisées dans cette étude ont permis de quantifier l'intensité de ces liens. Par ailleurs, des compilations statistiques faisant intervenir simultanément plusieurs variables ont permis de déterminer le jeu de caractéristiques le plus relié à la complexité des textes mesurée par leur rattachement à la classe d'enseignement. Les résultats obtenus à ces analyses ont rendu possible la fabrication d'indices de calibrage.

C- Les premiers examens des statistiques sur les différentes variables ont indiqué un comportement différent des textes rattachés au primaire par rapport à ceux rattachés au secondaire. Il faut rappeler que cette recherche s'est d'abord centrée sur des documents utilisés au primaire. Une grande variété de textes a été sélectionnée pour faire partie d'un corpus se voulant le plus significatif possible de documents servant à l'apprentissage de la lecture au primaire. Par ailleurs, les élèves du primaire sont en processus d'apprentissage de la lecture et de l'écriture de la langue. Les progrès réalisés entre le début et la fin du primaire sont ainsi plus facilement mesurables; les textes utilisés caractérisent plus nettement leur rattachement à l'une ou l'autre classe.

La situation est probablement différente au secondaire. Les apprentissages de base ont alors été réalisés. La difficulté des textes est probablement davantage reliée aux genres littéraires utilisés plutôt qu'à la complexité morpho-syntaxique. Après avoir complété le corpus du secondaire, il y aura lieu d'entreprendre des analyses qui permettront d'identifier les caractéristiques spécifiques à des lecteurs qui ont complété les apprentissages de base en lecture. Il faudra fabriquer des listes de mots susceptibles de faciliter ou de rendre plus complexe un texte destiné à ce public-cible.

Troisième partie

Le prototype SATO-CALIBRAGE

Présentation du prototype SATO-CALIBRAGE

Léo Laroche, ministère de l'Éducation

Léo Laroche travaille au ministère de l'Éducation du Québec. Jusqu'en mars 1992, il appartenait à la Direction générale de l'évaluation et des ressources didactiques. Maintenant, il est à la Direction de la Recherche.

Depuis plusieurs années, le ministère de l'Éducation du Québec est impliqué dans un projet d'analyse automatisée de textes pour en évaluer la lisibilité. L'origine de cet intérêt pour l'étude de textes vient du fait qu'il a été possible d'établir un lien entre le taux d'échec des élèves à certaines épreuves uniques (principalement en français, langue maternelle) et la complexité des textes choisis pour ces examens. Il est en effet reconnu par plusieurs spécialistes de la mesure et de l'évaluation que la complexité de la langue utilisée dans un instrument peut biaiser le processus d'évaluation des apprentissages dans une discipline. Le degré de lisibilité des textes est aussi un aspect majeur pour l'apprentissage et l'évaluation de la lecture à l'enseignement primaire.

Le logiciel SATO (Système d'Analyse de Textes par Ordinateur) a été retenu comme outil informatique pouvant faciliter l'analyse de la lisibilité des textes utilisés pour des fins d'apprentissage ou d'évaluation pédagogique. Les personnes impliquées dans cette étude sur la lisibilité dans un contexte d'apprentissage ou d'évaluation pédagogique ont identifié un ensemble de stratégies d'analyse qui a été désigné sous l'appellation «SATO-CALIBRAGE». Cette procédure d'utilisation du logiciel SATO a permis d'automatiser plusieurs des opérations qui doivent être réalisées pour analyser un texte à l'aide de l'ordinateur dans le but d'en connaître le degré de lisibilité. Nous présentons ci-dessous la séquence des différentes commandes qui ont été prévues pour produire les renseignements jugés utiles pour se prononcer sur le niveau de lisibilité des textes.

Description du prototype

Ce prototype utilise la version courante du logiciel SATO; cette décision permet de recourir, lorsque pertinent, aux améliorations apportées à ce produit. Outre les modules SATOGEN et SATOINT, il y a utilisation de fichiers de commandes DOS ainsi que de fichiers de commandes SATO (il s'agit des fichiers avec l'extension .CSA). Un fichier exécutable sous DOS (ce fichier porte l'extension BAT) joue le rôle de «gérant» de la procédure. Dans un premier temps, ce fichier offre à l'usager la possibilité d'exécuter l'ensemble de la procédure ou une partie uniquement. Voici les options actuellement prévues dans le prototype SATO-CALIBRAGE:

1- À partir d'un texte en codes ASCII, trois possibilités de traitement sont offertes à l'usager:

a. auto — il s'agit de l'exécution complète de la procédure avec affichage des résultats en mode continu, c'est-à-dire que l'usager devra consulter

les données produites à partir du rapport de calibrage;

b. complet — il s'agit des mêmes compilations que celles produites en a); cependant, l'usager, en plus de pouvoir consulter les résultats du traitement dans un rapport produit, pourra commander la vitesse d'affichage à l'écran;

c. prépare — cette option commande uniquement la réalisation des opérations de préparation du texte pour un traitement ultérieur dans le but de disposer des données reliées au calibrage; cette option fournit cependant un sommaire d'information sur les variables les plus reliées à la lisibilité du texte.

2- Si le texte à analyser a déjà été «préparé» à l'aide de l'une ou l'autre option présentée ci-dessus, il est possible de recourir aux options suivantes :

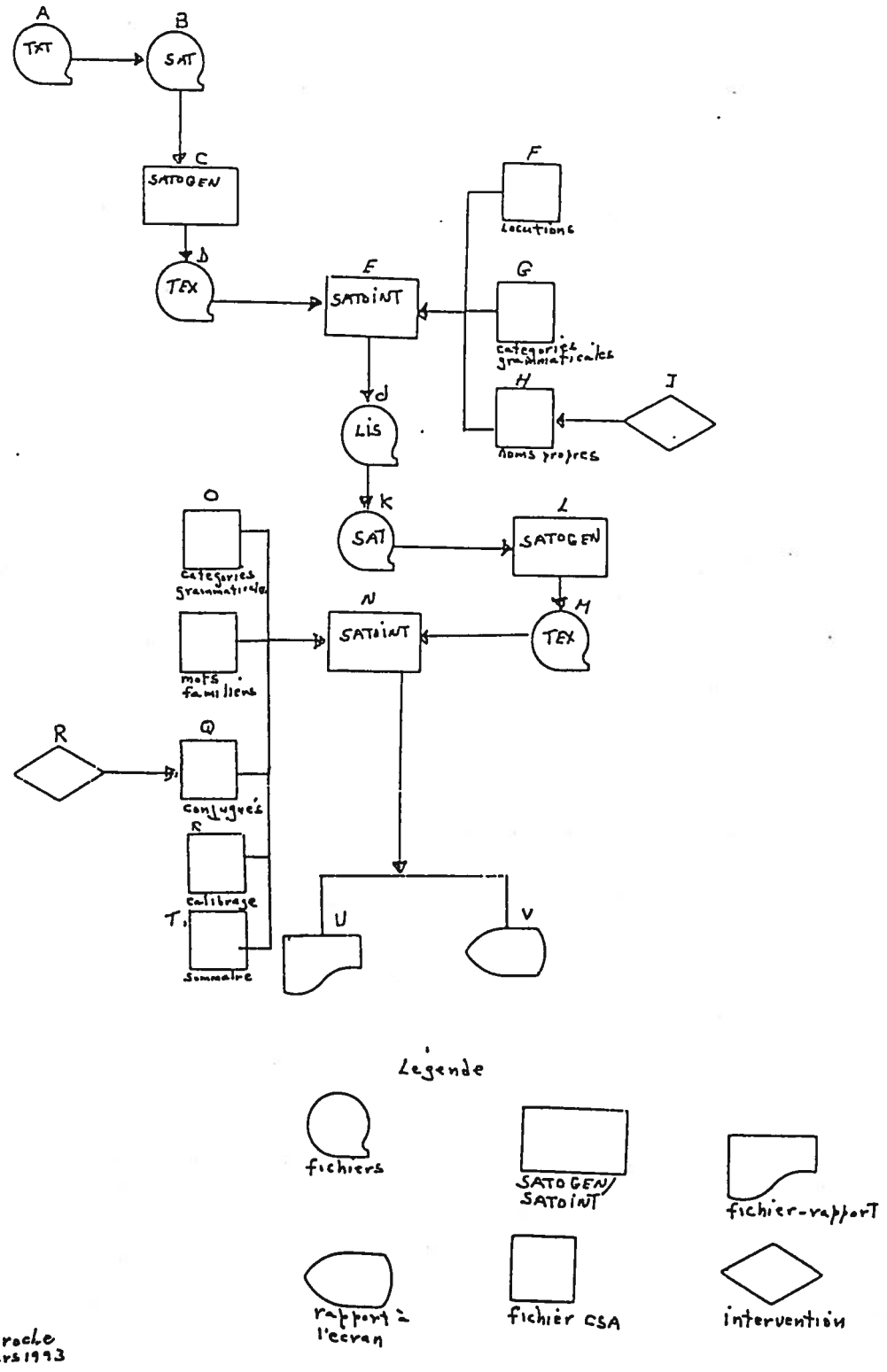
d. reprise — le recours à cette commande produit le rapport complet de calibrage;

e. sato — cette option permet de poursuivre l'analyse à l'aide du module SATOINT; l'usager peut ainsi recourir aux différentes commandes du logiciel SATO dans le but de compléter son analyse textuelle;

f. sommaire — il s'agit ici de la production du sommaire de calibrage; en plus de l'indice SATO-CALIBRAGE, ce rapport comprend des statistiques sur les variables les plus reliées à la lisibilité d'un texte.

Voici une présentation des différentes opérations réalisées lorsqu'un usager a recours à la procédure «complète» de SATO-CALIBRAGE; les autres options actuellement prévues exécutent l'une ou l'autre opération comprise dans le prototype. La figure apparaissant à la page suivante illustre le déroulement du traitement d'un texte soumis à SATO-CALIBRAGE. Comme on peut le constater, il y a un recours successif aux modules SATOGEN et SATOINT; de plus, un ensemble de fichiers exécutables directement par le logiciel SATO permet la réalisation des différents traitements nécessaires à la production d'indices reliés à la lisibilité d'un texte.

SATO-CALIBRAGE: Déroulement du traitement d'un texte



À partir du diagramme du système, voici donc un sommaire du déroulement des opérations accomplies par ce prototype; nous indiquons entre parenthèses les références aux opérations apparaissant sur le schéma.

1- Dans un premier temps, le fichier soumis à l'analyse est copié dans un fichier possédant les caractéristiques préalables à l'utilisation de SATOGEN (A et B).

2- Il y a une première «génération» du texte soumis afin d'en posséder une version «SATO» (C et D).

3- Le module SATOINT réalise les traitements suivants à l'aide d'un ensemble de fichiers exécutables (il s'agit des fichiers portant l'extension CSA) :

a) le blocage des locutions figées;
b) la catégorisation grammaticale;
c) le dépistage des noms propres (1).

Les opérations concernées par ces traitements portent les lettres E à K.

4- Il y a ensuite une nouvelle «génération» du texte comprenant les caractéristiques identifiées lors de l'étape précédente (L et M).

5- Le module SATOINT est de nouveau utilisé pour accomplir les tâches suivantes :

a) la catégorisation grammaticale;
b) le repérage des mots familiers;
c) la catégorisation en contexte des verbes conjugués (2);
d) la fabrication du rapport de calibrage;
e) le calcul de statistiques sommaires.

Il y a enfin, production du rapport de calibrage. L'ensemble de ces opérations est identifié sur le schéma par les lettres N à V.

Comme le schéma le laisse voir, il est possible de faire réaliser un nombre limité d'opérations comme c'est le cas pour l'option «prépare». Par ailleurs, l'usager pourra faire l'économie des opérations déjà réalisées lorsqu'il souhaitera reprendre une partie uniquement de la procédure.

Le schéma indique bien que les différentes opérations transforment le texte analysé pour le rendre compatible à une interrogation à l'aide du module SATOINT dans le but d'obtenir des statistiques sur les variables les plus reliées à la lisibilité d'un texte. Pour atteindre cet objectif, le prototype prévoit plusieurs travaux de préparation à cette analyse finale.

SATO, au moment du recours au module SATOGEN, réalise ses opérations à partir d'un fichier portant l'extension SAT et contenant au moins des indications sur l'alphabet utilisé et le titre attribué au document. D'autres informations peuvent aussi être fournies au logiciel : la

longueur des pages, l'indication de majuscules de noms propres, pour citer quelques possibilités. Pour l'application SATO-CALIBRAGE, nous avons préféré ne pas modifier le texte original à analyser, mais plutôt le copier, à l'aide d'une commande DOS, dans un fichier temporaire contenant cet ensemble d'indications indispensables au fonctionnement harmonieux de SATO. Ce recours à un fichier contenant déjà l'ensemble des instructions nécessaires à SATOGEN permet de plus de diminuer le nombre d'opérations à réaliser lorsque le même type d'analyses est exécuté avec des textes différents comme c'est le cas pour l'analyse de la lisibilité. Il restera toutefois à indiquer, pour chacun des textes soumis au prototype, les segments que l'on souhaite voir exclus des analyses en les plaçant en commentaires (3).

Il est, par ailleurs, possible d'indiquer au programme certaines caractéristiques rattachées aux mots à l'aide de codes ou caractères réservés à cette fin. On peut ici mentionner l'identification des noms propres qui peuvent être considérés en contexte comme des mots familiers au public-cible. Il est possible de réaliser cette identification directement dans le texte de départ en codes ASCII; nous avons plutôt préféré intégrer cette tâche aux opérations régulières du prototype. Nous pouvons ici fournir d'autres exemples de préparation des textes afin d'être en mesure de calculer les indices souhaités :

- le blocage de mots composés,
- la projection des propriétés grammaticales,
- l'identification des mots familiers pour le public-cible.

Le recours judicieux aux deux modules de SATO jumelé à l'interrogation de listes de mots (il s'agit des «dictionnaires» SATO) permet de réaliser de façon presqu'automatique ces travaux préparatoires. Lorsque les algorithmes contenus dans le prototype rencontrent des situations ambiguës, l'usager est invité à fournir l'information appropriée; nous constatons que ce type d'intervention directe de l'usager est plutôt rare dans la majorité des cas.

Développements à prévoir

Le prototype SATO-CALIBRAGE actuellement utilisé peut difficilement être modifié par un non spécialiste. Même si les fichiers de commandes inclus sont documentés à l'aide de commentaires nombreux, il peut s'avérer hasardeux d'en modifier une partie sans risquer de provoquer un mauvais fonctionnement de l'ensemble de la procédure. C'est ainsi qu'un certain nombre de fichiers de commandes ont été prévus pour obtenir des analyses sur l'un ou l'autre aspect relié à la lisibilité d'un document. On peut souhaiter avoir d'autres renseignements sur le texte analysé afin de répondre à des besoins différents de ceux prévus par la procédure choisie. Il est toujours possible de recourir au logiciel SATO lui-même pour trouver réponse à ce genre d'interrogations; il faut toutefois apprendre à utiliser cet outil polyvalent d'analyse textuelle.

Nous croyons que le prototype devra comprendre une interface qui permettrait à un usager de préciser certains paramètres pour tenir compte des objectifs poursuivis par son analyse textuelle. Une telle interface pourrait prendre plusieurs formes; idéalement, cette application donnerait lieu au développement d'un système expert.

Voici l'énumération de différents aspects qui devraient être personnalisés dans le but de tenir compte des besoins des utilisateurs.

Quelques paramètres initiaux devraient être fixés avant l'exécution de l'une ou l'autre compilation; il s'agit entre autre de :

- déterminer le public-cible visé par le texte;
- préciser la localisation des fichiers de textes à analyser;
- fixer le nombre de lignes par page de rapport;
- choisir le type d'affichage à l'écran des résultats de l'interrogation;
- indiquer les propriétés à afficher au regard du lexique.

L'usager pourrait souhaiter intervenir en cours de réalisation des travaux ou bien laisser le contrôle des opérations au programme informatique. Ce choix tiendra compte des fins poursuivis par l'usager. Il s'agit ici de la gestion des ambiguïtés qui demeurent après la mise en oeuvre d'un ensemble d'automates.

Comme nous l'indiquions plus haut, il est possible de réaliser plusieurs types de travaux à l'aide du prototype SATO-CALIBRAGE. L'usager doit pouvoir préciser sa requête pour tenir compte des objectifs poursuivis. S'agit-il d'une première analyse à partir d'un texte produit par un traitement de textes ou bien souhaite-t-on avoir des informations sur un texte qui a déjà fait l'objet d'une analyse antérieure à l'aide du prototype?

Enfin, il faut laisser à l'usager la possibilité de définir les éléments que comprendra le rapport produit à la suite de l'interrogation. La pratique nous démontre que les usagers, en plus de souhaiter avoir des indicateurs rendant compte de la complexité d'un texte, veulent disposer d'un ensemble d'informations sur la structure même du texte. S'il est facile de s'entendre sur des indicateurs reliés à la lisibilité, il est beaucoup plus difficile de cerner les aspects d'un texte rendant compte du contexte d'utilisation de la procédure.

Donc, les prochaines versions du prototype offriront une variété d'options tout en demeurant simples à utiliser.

Notes

(1) Il y a intervention de l'usager pour lever les ambiguïtés laissées par les algorithmes de traitement utilisés; cette opération peut être réalisée hors contexte ou en contexte.

(2) Comme pour les noms propres, l'usager devra résoudre les ambiguïtés sur les verbes conjugués; cette opération se réalise en contexte.

(3) En SATO, on réalise cette opération en encadrant le texte à mettre en commentaires par «*{» et «}».

Texte soumis à SATO-CALIBRAGE: texte et résultats

Voici un exemple de texte soumis au prototype SATO-CALIBRAGE. Il s'agit d'une version du prototype datant d'octobre 93.

Le texte soumis à SATO-CALIBRAGE a été utilisé au cours d'une épreuve pour des élèves de la quatrième année du primaire. On trouvera d'abord une impression du texte intégral. Ensuite, on a le rapport sommaire avec le calcul de l'indice SATO-CALIBRAGE. L'indice est obtenu en additionnant les valeurs pondérées de chacun des variables considérées. Une pondération positive indique qu'il s'agit d'un indice de difficulté, par exemple, le pourcentage de mots inconnus. Inversement, une pondération négative indique un indice de facilité, par exemple, le pourcentage de pronoms personnels.

La troisième partie du texte contient le rapport qualitatif dans sa version minimale. Cela signifie que ce rapport ne contient que certains indices, ceux que nos utilisateurs ont jugés les plus utiles. Le rapport contient d'abord un lexique du texte trié par fréquence décroissante. Le lexique exclut les termes fonctionnels: articles, pronoms, etc. Ensuite, on trouve l'indice de Gunning avec ses diverses composantes. Puis, on a la répartition des lexèmes, en nombre et en pourcentage, selon les diverses catégories de la propriété connu: *p6* (primaire 6ième année), *oui* (termes fonctionnels courants, noms propres et nombres) et *nil* (présumés inconnus).

On retrouve ensuite le lexique des mots de 9 lettres ou plus. Enfin, on retrouve une liste de phrases comprenant des éléments potentiels de difficulté. Chaque phrase est précédée de sa référence dans le texte calculée sur la position du premier mot: nom du document (appelé ici tmp), la page dans le document, le numéro de la ligne dans la page et le numéro du mot dans la ligne. La phrase est précédée d'un jugement:

4Verbes :	- la phrase contient au moins quatre verbes conjugués;
31-99 :	- la phrase contient plus de trente mots;
PronomRelatif:	- la phrase contient au moins un pronom relatif; le diagnostic est porté sur la base de la catégorie hors contexte. En contexte, le lexème peut avoir une autre fonction grammaticale, par exemple une conjonction.
2MotsInconnus:	- la phrase contient au moins deux mots inconnus.

Finalement, on retrouve la phrase elle-même. Les mots soulignés ont servi de déclencheur à l'un ou l'autre des jugements portés sur la phrase.

Texte soumis

La vie à la campagne

Au début de la colonie, en 1608, jusqu'à 1920 environ, la majorité des Québécois vivent en milieu rural. Leur vie est organisée autour des travaux saisonniers et quotidiens sur la ferme. Les familles sont

nombreuses, la vie est austère; chaque famille produit presque tous les biens nécessaires à sa subsistance.

Les parents ont besoin des enfants puisqu'ils ne peuvent assumer seuls toutes les tâches de la maison et de la ferme. Ils ne peuvent pas non plus embaucher d'hommes pour les aider puisque les revenus de la ferme sont trop maigres. Jusqu'en 1800, peu d'enfants fréquentent l'école : certains y vont quelques mois par année, quand le travail est moins exigeant; d'autres y vont quelques années seulement. Entre 1800 et 1900, de plus en plus d'enfants fréquentent l'école tout en continuant de participer aux tâches domestiques et agricoles.

Les femmes et les filles travaillent surtout à l'entretien de la famille et de la maison. Dès qu'une fille atteint l'âge de sept ans, elle s'occupe de ses frères et soeurs plus jeunes : elle les nourrit, les lave, les habille et les amène jouer dehors. Les filles nettoient la maison, servent à table, lavent la vaisselle, lavent et repassent les vêtements.

À mesure qu'elles vieillissent, elles aident leur mère à préparer les repas, à tisser et à coudre les vêtements. 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.

Les garçons s'initient avec leur père aux travaux de la ferme. Ils l'aident à faire le train, matin et soir. Ils vont chercher et ramènent les vaches aux champs, nettoient l'étable et nourrissent les animaux. 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. Au moment des semailles et des récoltes, ils apportent l'eau, les repas et les outils aux adultes occupés aux champs.

Certaines tâches peuvent être accomplies tant par les garçons que par les filles. Tout au long de l'été, la mère et les enfants vont cueillir les petits fruits et les préparent pour faire des confitures. Au moment des récoltes, ils dirigent les boeufs et les chevaux pendant que les grands ramassent et chargent le foin, les céréales ou les légumes. On confie aussi aux enfants la tâche de vider les pots de chambre dans les toilettes extérieures ou sur le tas de fumier.

L'enfant de cultivateur avait de nombreuses et lourdes responsabilités. De plus, s'il était l'aîné, qu'il soit garçon ou fille, il avait l'autorité sur ses frères et soeurs. Il se préparait donc très jeune à sa vie d'adulte et jamais il n'était rémunéré pour ses travaux.

La vie en ville

À partir de 1850, des industries s'installent dans les principales villes québécoises, surtout à Montréal et à Québec. On y trouve des industries

alimentaires, des manufactures de chaussures, de vêtements et de cigarettes, des filatures de coton. Alors qu'en 1871, seulement 23 p. 100 de la population québécoise habite la ville, en 1921, cette proportion est passée à 56 p. 100.

Avec la migration des familles rurales vers les villes, le mode et le rythme de vie changent considérablement. La famille ne produit plus elle-même les biens nécessaires à sa subsistance. 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. Par exemple, vers 1889, un loyer pouvait coûter entre 6,00 $ et 12,00 $ par mois; 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. Et l'ouvrier, qui la plupart du temps n'avait pas un métier spécialisé, gagnait entre 4,80 $ et 6,00 $ par semaine. Il devait travailler durement : dix heures et plus par jour, six jours par semaine.

Dans de telles conditions, le chef de famille ne peut à lui seul subvenir aux besoins des siens. Les femmes et les enfants, parfois dès l'âge de huit ou neuf ans, doivent contribuer au revenu familial. Les enfants sont moins bien payés que les adultes, 1,50 $ ou 2,00 $ tout au plus, pour une semaine aussi longue que celle de leurs parents. S'ils font des erreurs à cause de la fatigue ou de la distraction, ils doivent payer une amende; il peut arriver qu'un enfant perde son salaire parce qu'il doit payer trop d'amendes. L'enfant indiscipliné est sévèrement puni; il est mis «au trou» dans un cachot aménagé dans l'usine.

Les tâches des enfants varient beaucoup d'une usine à l'autre. Certains accomplissent les mêmes tâches que les adultes, d'autres font les commissions à l'extérieur et à l'intérieur de l'usine. transportent les outils et la marchandise. Dans les filatures de coton, on profite de leur petite taille : ils doivent se glisser entre les métiers à tisser pour nouer les fils cassés. À cause de leurs petites mains, certaines usines les emploient à manipuler des machines délicates, parfois dangereuses.

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. Les lois qui devaient protéger les enfants n'étaient pas toujours respectées.

L'enfance, un phénomème récent

On s'imagine mal aujourd''hui que les enfants soient entrés aussi tôt qu'à huit ou neuf ans sur le marché du travail. Quelques raisons expliquent ce phénomène.

Autrefois, on pensait qu'un enfant qui avait atteint l'âge de raison était capable de comprendre comme un adulte et de commencer à agir comme lui. 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. Il en était ainsi chez les colons français et dans les tribus amérindiennes : l'enfant participait aux travaux de la ferme et aux activités de la tribu pour s'initier aux exigences de la vie adulte.

De plus, autant les familles rurales que les familles urbaines avaient besoin du travail des enfants pour assurer leur subsistance. La famille rurale ne payait pas les enfants mais avait besoin d'eux pour produire et fabriquer les biens essentiels à la survie. En ville, la vie était très difficile, surtout dans le milieu ouvrier, et les enfants devaient contribuer au revenu familial.

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. Les enfants constituaient une main-d'oeuvre bon marché dont ils pouvaient se servir pour augmenter et accélérer la production.

Le travail des enfants a donc été bien vu et même encouragé jusqu'à ce qu'on prenne conscience de ses conséquences désastreuses. Ce n'est qu'au 20e siècle que cette situation a vraiment changé.

Les enfants qui travaillaient de longues heures dans les conditions malsaines des usines s'estropiaient souvent et voyaient leur santé atteinte. Une inspectrice constate que les enfants grandissent moins vite et n'atteignent parfois leur maturité physique qu'à vingt ou vingt-cinq ans.

De plus, le besoin économique des familles est si grand qu'on ne peut envoyer les enfants à l'école. La population est donc peu instruite. Les gouvernements commencent alors à s'inquiéter du taux élevé d'analphabétisme et des conditions de travail dans les usines.

La situation des enfants a donc beaucoup changé au cours de ce siècle. Le respect de l'enfant et de ses droits est un phénomène très récent dans l'histoire de l'humanité.

Rapport sommaire (indice SATO-CALIBRAGE)

Indice SATO-CALIBRAGE Fichier campagne.sta

Variable	Valeur	Pondération
Nbr de phrases	76.00	0.012
% de phrases de plus de 15 mots	63.16	-0.025
% de phrases de plus de 20 mots	50.00	0.020

```
% de phrases de plus de 30 mots ..................   9.21   0.089
% de phrases contenant au moins un pronom relatif   19.74   0.020
% de phrases contenant au moins 2 mots inconnus ..   0.00   0.051
% de ponctuation .................................   5.16  -0.209
% de pronoms personnels ..........................  14.46  -0.054
% de mots inconnus ...............................   0.61   0.543


Valeur de l'indice : 4.46
```

Rapport qualitatif (version minimale)

```
Rapport de calibrage du texte campagne

==========================================

Lexique des mots (non fonctionnels) apparaissant plus d'une fois

fréq gramr                          connu
21   nomc                           p6      enfants
9    nomc                           p6      vie
8    nomc                           p6      ans
7    nomc                           p6      famille
6    nomc                           p6      enfant
6    nomc                           p6      maison
5    nomc                           p6      familles
5    (adj,v_conj,nomc)              p6      ferme
5    (v_conj,nomc)                  p6      tâches
5    nomc                           p6      travail
5    nomc                           p6      travaux
4    (adj,nomc)                     p6      adulte
4    nomc                           p6      âge
4    nomc                           p6      besoin
4    nomc                           p6      conditions
4    v_conj                         p6      peut
4    (v_conj,nomc)                  p6      usines
4    nomc                           p6      vêtements
4    v_conj                         p6      vont
3    (adj,nomc)                     p6      adultes
3    (v_conj,ppassé)                p6      atteint
3    (adj,nomc)                     p6      biens
3    nomc                           p6      cents
3    nomc                           p6      champs
3    (v_conj,nomc)                  p6      doit
3    v_conj                         p6      doivent
3    nomc                           p6      école
```

3	nomc	p6	femmes
3	nomc	p6	fille
3	nomc	p6	filles
3	nomc	p6	industries
3	(adj,nomc)	p6	mère
3	v_inf	p6	payer
3	v_conj	p6	peuvent
3	nomc	p6	semaine
3	nomc	nil	subsistance
3	v_conj	p6	transportent
3	v_inf	p6	travailler
3	(v_conj,nomc)	p6	usine
3	nomc	p6	ville
3	nomc	p6	villes
2	v_conj	p6	aident
2	v_inf	p6	augmenter
2	(v_conj,nomc)	p6	bois
2	ppassé	p6	changé
2	nomc	p6	chef
2	v_inf	p6	contribuer
2	(adj,nomc)	p6	coton
2	v_conj	p6	devaient
2	nomc	p6	eau
2	(adj,v_conj,nomc)	p6	étable
2	(nomc,ppassé)	p6	été
2	v_inf	p6	faire
2	adj	p6	familial
2	nomc	p6	filatures
2	v_conj	p6	font
2	v_conj	p6	fréquentent
2	nomc	p6	frères
2	nomc	p6	garçons
2	nomc	p6	heures
2	v_conj	p6	lavent
2	(v_conj,nomc)	p6	livre
2	(v_inf,nomc)	nil	loyer
2	adj	p6	malsaines
2	(nomc,ppassé)	p6	marché
2	nomc	p6	milieu
2	nomc	p6	mois
2	nomc	p6	moment
2	(adj,nomc)	p6	nécessaires
2	v_conj	p6	nettoient
2	adj	p6	nombreuses
2	nomc	p6	outils
2	(adj,nomc)	p6	ouvrier
2	(adj,nomc)	p6	parents
2	nomc	p6	phénomène

2	nomc	p6	population
2	(v_conj,nomc,ppassé)	p6	produit
2	adj	général	récent
2	(v_conj,nomc)	p6	récoltes
2	nomc	p6	repas
2	nomc	p6	responsabilités
2	(nomc,ppassé)	p6	revenu
2	adj	p6	rurales
2	nomc	p6	salaire
2	nomc	p6	siècle
2	nomc	p6	situation
2	nomc	p6	soeurs
2	v_inf	p6	tisser

Longueur des mots, des phrases et des paragraphes

53 mots de 1 car. (4%)	319 mots de 2 car. (24%)
229 mots de 3 car. (17%)	141 mots de 4 car. (11%)
115 mots de 5 car. (9%)	129 mots de 6 car. (10%)
131 mots de 7 car. (10%)	57 mots de 8 car. (4%)
56 mots de 9 car. (4%)	44 mots de 10 car. (3%)
25 mots de 11 car. (2%)	13 mots de 12 car. (1%)
5 mots de 13 car. (0%)	2 mots de 14 car. (0%)
3 mots de 15 car. (0%)	2 mots de 16 car. (0%)
0 mot de 17 car. (0%)	0 mot de 18 car. (0%)
0 mot de 19 car. (0%)	0 mot de 20 car. (0%)
0 mot de 21 à 25 car. (0%)	0 mot de 26 à 30 car. (0%)
0 mot de plus de 30 car. (0%)	

nombre de mots..... 1324	longueur moyenne: 4.7 car.
nombre de phrases..... 76	longueur moyenne: 17.4 mots
nombre de paragraphes. 22	longueur moyenne: 60.2 mots

pourcentage de mots de 9 lettres et plus: 11%

indice de lisibilité de Gunning: 11.5

La répartition des lexèmes par rapport aux listes de mots connus.

nombre de lexèmes: 526

nombre	pourcent	symbole
466	88.59%	p6
53	10.08%	oui
7	1.33%	nil

Liste des mots identifiés comme inconnus. En vous positionnant sur un mot et en tapant une des touches suivantes, vous pouvez recatégoriser le mot:

Touche «3» pour un mot connu de 3ième année

Touche «6» pour un mot connu de 6ième année

Touche «g» pour un mot connu du «grand public»

Touche «i» pour un mot connu d'un public informé

Touche «-» pour corriger un jugement précédent

fréq	
1	assumer
1	au_cours_de
1	austère
1	humanité
1	phénomème
1	rémunéré
3	subsistance

Liste des mots longs (9 lettres ou plus).

fréq	
2	à_cause_de
1	accélérer
1	accomplies
1	accomplissent
1	activités
1	agricoles
1	alimentaires
1	alimentent
1	alors_qu'
1	amérindiennes
1	à_mesure_qu'
1	analphabétisme
1	apportent
1	atteignent
1	au_cours_de
2	augmenter
1	aujourd''hui
1	autrefois
2	certaines
1	chauffage
1	chaussures

1 cigarettes
1 commencent
1 commencer
1 commissions
1 comprendre
4 conditions
1 confitures
1 conscience
1 conséquences
1 considérablement
1 constituaient
1 continuant
2 contribuer
1 cultivateur
1 dangereuses
1 délicates
1 désastreuses
1 difficile
1 difficiles
1 distraction
1 domestiques
1 économique
1 embaucher
1 emploient
1 encouragé
1 entretenir
1 entretien
1 essentiels
1 estropiaient
1 exigences
1 expliquent
1 extérieur
1 extérieures
1 fabriquer
2 filatures
2 fréquentent
1 gouvernements
1 grandissent
1 indiscipliné
3 industries
1 inquiéter
1 inspectrice
1 installent
1 instruite
1 intérieur
1 jusqu'_à_ce_qu'
1 jusqu'_en
2 malsaines
1 manipuler
1 manufactures

1	marchandise
1	migration
2	nécessaires
2	nettoient
2	nombreuses
1	nourrissent
1	nourriture
1	organisée
1	parce_qu'
1	par_exemple
1	participait
1	participer
1	particulièrement
1	pendant_que
1	phénomème
2	phénomène
2	population
1	pouvaient
1	préparait
1	préparent
1	principales
1	production
1	proportion
1	Québécois
1	québécoise
1	québécoises
1	quotidiens
1	ramassent
1	remplacer
1	repassent
1	respectées
1	responsabilité
2	responsabilités
1	saisonniers
1	semailles
2	seulement
1	sévèrement
2	situation
1	spécialisé
3	subsistance
1	techniques
1	toilettes
3	transportent
1	travaillaient
1	travaillent
3	travailler
1	travailleurs
1	vaisselle

4 vêtements
1 vieillissent

Liste des phrases susceptibles de contenir des éléments de complexité.

Légende

4Verbes: la phrase possède au moins 4 verbes conjugués;

31-99: la phrase contient plus de 30 mots;

PronomRelatif: la phrase possède un mot qui, au dictionnaire, peut être un pronom relatif: qui, que dont, etc. La phrase est affichée même si le mot, en contexte, est une conjonction.

2MotsInconnus: la phrase contient au moins deux mots inconnus.

amène

1 *page=tmp/1/17/11

*jugement=4Verbes **elle les nourrit, les lave, les habille et les amène jouer dehors.**

lavent

2 *page=tmp/1/18/9

*jugement=4Verbes **Les filles nettoient la maison, servent à table, lavent la vaisselle, lavent et repassent les vêtements.**

.

3 *page=tmp/1/21/11

*jugement=31-99 **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.**

nourrissent

4 *page=tmp/1/25/12

*jugement=4Verbes **Ils vont chercher et ramènent les vaches aux champs, nettoient l'étable et nourrissent les animaux.**

transportent

5 *page=tmp/1/26/14

*jugement=(4Verbes,31-99) **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.**

que

6 *page=tmp/1/31

*jugement=PronomRelatif **Certaines tâches peuvent être accomplies tant par les garçons que par les filles.**

qu'

7 *page=tmp/1/39

*jugement=PronomRelatif **De_plus, s'il était l'aîné, qu'il soit garçon ou fille, il avait l'autorité sur ses frères et soeurs.**

.

8 *page=tmp/2/6/11

*jugement=31-99 **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.**

9 *page=tmp/2/9/17

*jugement=31-99 **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.**

qui

10 *page=tmp/2/11/14

*jugement=PronomRelatif **Et l'ouvrier, qui la plupart du temps n'avait pas un métier spécialisé, gagnait entre 4,80 $ et 6,00 $ par semaine.**

que

11 *page=tmp/2/17/12

*jugement=PronomRelatif **Les enfants sont moins bien payés que les adultes, 1,50 $ ou 2,00 $ tout au plus, pour une semaine aussi longue que celle de leurs parents.**

qu'

12 *page=tmp/2/21

*jugement=PronomRelatif **il peut arriver qu'un enfant perde son salaire parce_qu'il doit payer trop d'amendes.**

que

13 *page=tmp/2/24/14

*jugement=PronomRelatif **Certains accomplissent les mêmes tâches que les adultes, d'autres font les commissions à l'extérieur et à l'intérieur de l'usine.**

.

14 *page=tmp/2/31

*jugement=31-99 **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.**

qui

15 *page=tmp/2/33/5

*jugement=PronomRelatif **Les lois qui devaient protéger les enfants n'étaient pas toujours respectées.**

que
16 *page=tmp/2/35

*jugement=PronomRelatif **L'enfance, un phénomène récent**
On s'imagine mal aujourd''hui que les enfants soient entrés aussi tôt qu'à huit ou neuf ans sur le marché du travail.

qu'
17 *page=tmp/2/39

*jugement=PronomRelatif **Autrefois, on pensait qu'un enfant qui avait atteint l'âge de raison était capable de comprendre comme un adulte et de commencer à agir comme lui.**

.
18 *page=tmp/2/41
*jugement=31-99 **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.**

que
19 *page=tmp/3

*jugement=PronomRelatif **De plus, autant les familles rurales que les familles urbaines avaient besoin du travail des enfants pour assurer leur subsistance.**

.
20 *page=tmp/3/7

*jugement=31-99 **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.**

dont
21 *page=tmp/3/9/11
*jugement=PronomRelatif **Les enfants constituaient une main-d'oeuvre bon marché dont ils pouvaient se servir pour augmenter et accélérer la production.**

qu'
22 *page=tmp/3/13/10
*jugement=PronomRelatif **Ce n'est qu'au 20e siècle que cette situation a vraiment changé.**

qui
23 *page=tmp/3/15

*jugement=PronomRelatif **Les enfants qui travaillaient de longues heures dans les conditions malsaines des usines s'estropiaient souvent et**

voyaient leur santé atteinte.

que

24 *page=tmp/3/17/3

*jugement=PronomRelatif **Une inspectrice constate que les enfants grandissent moins vite et n'atteignent parfois leur maturité physique qu'à vingt ou vingt-cinq ans.**

qu'

25 *page=tmp/3/20

*jugement=PronomRelatif **De plus, le besoin économique des familles est si grand qu'on ne peut envoyer les enfants à l'école.**

---------- Terminé ----------

Exemple d'analyse de documents d'information

Claire Gélinas-Chébat, Clémence Préfontaine et François Daoust.

Claire Gélinas-Chébat et Clémence Préfontaine sont professeures au département de linguistique de l'UQAM. François Daoust est chercheur au Centre ATO-CI de l'UQAM.

Introduction

Le gouvernement du Québec, comme toute entreprise de services, diffuse des tonnes de documents écrits dans le but d'informer le public. Or, ces fascicules d'information ne semblent pas toujours remplir leur mission.

Nos services ont été requis (Gélinas-Chebat, et al. 1990) pour évaluer le niveau d'intelligibilité de fascicules d'information produits par l'un des nombreux ministère du Gouvernement du Québec. Ce ministère diffuse régulièrement des documents d'information auprès de ses bénéficiaires. Or lorsqu'il leur envoyait l'un de ces documents, la réaction typique des bénéficiaires étaient de téléphoner aux bureaux régionaux pour obtenir plus d'information. Ce ministère se retrouvait donc devant un double problème de productivité et d'efficacité: les documents envoyés ne remplissaient pas leur fonction d'information et les préposés du ministère, au lendemain d'un envoi, ne pouvaient effectuer d'autres tâches que celles de donner une information qui était en principe contenu dans le fascicule.

Nous décrirons le contexte dans lequel nous utilisons SATO pour aider à évaluer des documents, ainsi que les résultats que nous avons obtenus suite à la reformulation de l'un de leur fascicule.

Problématique

Les différents ministères des gouvernements doivent produire des documents d'information respectant des contraintes bureaucratiques et légales. S'ils sont conformes à la loi, ils ne sont pas pour autant toujours faciles à comprendre Nous tentons de l'illustrer à partir de la reproduction du document suivant:

ERRATUM
Calcul mécanographique des retenues à la source de l'impôt sur le revenu, des contributions au RRQ et de la contribution de l'employeur au RAMQ (voir TPD-107)

Nous désirons vous informer que la formule mathématique pour le calcul des retenues à la source de l'impôt du Québec sur le revenu comporte une variable erronée à la ligne T^1 du paragraphe **d)** de la page 22 qui devrait plutôt se lire comme suit:
T^1=TI - K^1 - 0,2E

De plus, la formule mathématique mentionnée au paragraphe **f)** de la page 23 devrait également être modifiée de la façon suivante:
A=[T(I + B) - K^1 - 0,2E]S^2 - M + L
Enfin, il y aurait lieu de corriger la variable I de la même page comme suit:
I=Revenu imposable annuel estimatif
=S^1(G-C-U-F-H^1-N)-Q-J ou S^1(G-C-U-F-N)-H^2-Q-J

Nous nous excusons de ce contretemps et nous vous remercions de votre collaboration.

Pour procéder à l'analyse de textes écrits, nous utilisons le modèle proposé par Préfontaine et Lecavalier (1990). Ce modèle tient compte de trois niveaux différents d'analyses de façon à tenir compte non seulement des aspects lexicaux d'un texte mais également de son aspect formel, son organisation, et de sa représentation sémantique, sa cohérence explicite et implicite.

Nous ne décrirons ici que les résultats des analyses optenues à partir du logiciel SATO. En effet, au moment de cette recherche, SATO a été un outil précieux surtout pour l'analyse microstructurelle des documents d'information.

Méthodologie

1. Matériel

Nous avons choisi 12 fascicules produits par le ministère qui nous consultait. Cette série de fascicules d'information est intitulée "Saviez-vous que...". Neuf de ces fascicules[1] se présentent sur une seule page recto-verso dont les dimensions sont de 3.5 po. par 8.5 po. environ. La ligne de lecture est de 3 po. environ.

Le document intitulé "Apte" est un fascicule de 39 pages, recto-verso dont les dimensions sont de 7.5 po. par 3.5 po. environ. La ligne de lecture varie: elle est de 3.5 po. environ pour la première page de lecture, qui est la page "Avis", elle est de 5.5 po.environ pour les pages de la table des matières et de l'index, et de 1.5 po. environ sur trois colonnes pour les pages de texte.

Le fascicule "Apport" est un document de 13 pages recto-verso dont les dimensions sont de 4 po. par 8.25 environ. La ligne de lecture est de 3.5 po environ pour l'ensemble du texte.

2. Procédures

Nous avons soumis tous les textes des formulaires au logiciel SATO.

Pour chacun des textes, nous avons obtenus le lexique complet, c'est-à-dire la liste complète des mots utilisés[2] par ordre alphabétique, avec leur fréquence d'utilisation et un indice de familiarité des mots[3].

SATO nous a également fourni différentes listes des mots utilisés, en fonction des critères suivants: lexique des mots apparaissant plus d'une fois, lexique des déterminants (sauf les articles), lexique des pronoms non-personnels (sauf les articles), lexique des pronoms personnels (sauf les articles), lexique des mots-liens (sauf les articles).

---

[1] Ce sont les fascicules F-188, F-189, F-190, F-191, F-192, F-193, F-352, F-353, F-354.

[2] "Mots", dans le sens de SATO comprenant aussi les signes de ponctuation et les chiffres.

[3] On trouvera dans Laroche (1990) le détail de la méthode utilisée pour constituer cette liste de référence.

À la requête "lisibilité" nous avons obtenu de SATO le décompte précis du nombre de mots à 1, 2, ..., n, caractères et le pourcentage correspondant, des mots de chacun des textes. De plus, il y a également le nombre total de mots pour chaque texte et la longueur moyenne des mots en fonction du nombre de caractères.

SATO nous a donné aussi le nombre de phrases en moyenne et la longueur moyenne de ces phrases en terme de nombre de mots, le nombre de paragraphe et leur longueur moyenne en terme de mots. Enfin pour chaque texte, le pourcentage de mots de 9 lettres et plus, et l'indice de lisibilité de Gunning[4].

Nous avons également obtenu différentes analyses des mots (c'est-à-dire les lexèmes et leur ventilation, en valeur absolue et relative) et des phrases. Ces analyses sont les suivantes: le rattachement des lexèmes aux catégories grammaticales, la répartition des lexèmes par rapport aux listes de mots connus, la liste des mots identifiés comme inconnus, la liste des mots longs, les phrases contenant plus de 15 mots, les phrases commençant par une préposition, une conjonction ou un adverbe, les phrases débutant par un pronom à la 3e personne, la liste des phrases contenant quatre propositions ou plus, la liste des phrases contenant un patron particulier (pronom, pronom, verbe; pronom non-personnel, verbe; pronom-écran, pronom, verbe), la liste des phrases contenant au moins une proposition subordonnée relative, les phrases contenant au moins deux mots inconnus, et enfin les phrases conprenant une séquence de 3 pronoms.

Résultats

Nous présentons ici, sous forme de tableau (tableau 1 suivant), les résultats de la requête "lisibilité", pour chacun des 12 fascicules étudiés.

[4] Dans SATO la formule utilisée est la suivante : l'indice de lisibilité = (P + M) * 0,4 où P = la longueur moyenne des phrases, et M le pourcentage de mots longs, les mots de plus de 9 lettres.

Tableau 1 : Indices de lisibilité des fascicules d'information

Doc. #	Long. moy. des mots (N de car)	Long. moy. des phrases (N de mots)	% de mots de plus de 9 lettres	Indice de Gunning[5]
F-188	4.7	15.2	10 %	10.0
F-189	4.7	16	10 %	10.5
F-190	5.1	15.5	18 %	13.5
F-191	4.9	16.6	14 %	12.2
F-192	4.9	12.8	12 %	9.8
F-193	4.8	14.1	12 %	10.4
F-352	5.1	11.5	14 %	10.3
F-353	5.4	13.9	19 %	13.2
F-354	4.9	18	11 %	11.7
Apte	4.8	18.4	14 %	12.8
S. Fin.	4.8	17.3	13 %	12.1
S. Rev.	5.1	13.5	15 %	11.6
Min	4.7	11.5	10 %	9.8
Max	5.4	18.4	19 %	13.5

Les résultats des analyses de SATO nous ont permis de donner des indices précis du niveau de difficulté des fascicules d'information. De plus, nous pouvions pointer plus précisément les responsables sémantiques ou syntaxiques de ces difficultés.

[5]Bourbeau (1988) propose les associations suivantes en fonction des types de textes.

indice de lisibilité	**degré de difficulté**	**textes-types**	**niveau scol.** (U.S.A. 1952)
6 et -	très facile	bandes dessinées	6e et moins
9-10	moyen	Reader's Digest	9e - 10e
13 et +	difficile	revues spécialisées	13e et plus

Nous avons ajouté la correspondance avec les niveaux scolaires à titre informatif.

Analyse des résultats

1. Exemple d'analyse de premier niveau

Pour un document particulier (le texte F-189), voici le genre d'indications fournies pour des analyses de premier niveau:

a. L'indice de lisibilité

SATO a calculé l'indice de lisibilité du document F-189, qui est de 10,5 (le pourcentage de mots de 9 lettres et plus est de 10%). Il s'agit donc d'un texte de difficulté moyenne selon cette mesure (Bourbeau, 1988, p. 26).

b. Les paragraphes

Pour les besoins de l'analyse faite par SATO, ce texte a été considéré comme un seul paragraphe, ce qui nous paraît discutable. Toutefois, comme cette mesure ne contribue pas au calcul de la lisibilité, il ne semble pas pertinent de considérer le nombre de paragraphes dans le texte.

c. Les phrases

La longueur moyenne des phrases est de 16,0 mots. Ce sont des phrases relativement longues, selon Bourbeau (1988) qui considère que «pour le lecteur moyen, le nombre de mots à ne pas dépasser est de 15» (p. 41). Elle ajoute que «le critère de la longueur des phrases est relié aux limites de la mémoire à court terme» (p. 41). Il faut compléter ces remarques en précisant que le nombre de propositions joue un rôle dans la compréhension des phrases: «Il n'en reste pas moins, qu'en moyenne, un texte comprenant de nombreuses subordonnées et dont les phrases seront longues, est vraisemblablement plus difficile à comprendre qu'un texte syntaxiquement plus dépouillé» (Henry, 1975, p. 67). Toutefois, une description complémentaire des phrases contenues dans le document F-189 s'impose pour que nous puissions saisir mieux l'impact de la longueur des phrases dans ce document.

d. Les mots

La longueur moyenne des mots est de 4,7 caractères, ce qui nous apparaît acceptable et porteur d'aucune difficulté particulière.

e. La fréquence de mots en fonction du nombre de caractères

Nous remarquons d'abord que 25 mots ont entre 10 et 17 caractères, dont 3 mots de 14 caractères; par une description subséquente, nous devrions mettre en évidence la difficulté sémantique de ces mots. Nous remarquons également qu'il y a 100 mots de 6 à 9 caractères sur un total de 321, ce qui signifie qu'environ le tiers des mots sont de difficulté moyenne.

f. Le lexique

Ce lexique est constitué de la liste par ordre alphabétique de tous les mots contenus dans le texte en respectant les formes morphologiques. Il faut savoir que *d'* est considéré comme un mot, de même que les éléments de ponctuation, les nombres, les suites de nombres (numéro de téléphone) ainsi que tous les symboles.

Le mot *assurance-chômage* compte 17 caractères, les mots *renseignements* (qui apparaît

2 fois) et *1-800-361-4740* (qui apparaît 1 fois) comptent 14 caractères. Il s'agit là de termes familiers, qui ne présentent pas de difficulté de compréhension.

Les mots de 11 à 14 caractères devront faire l'objet d'une description plus approfondie, afin d'évaluer leur niveau réel de difficulté. Pour une telle analyse, il peut être interressant de voir si un tel mot fait partie ou non d'un lexique courant. Pour des linguistes il existe de nombreuses listes de mots avec généralement leur fréquence d'usage. Mais ces listes de vocabulaire comprennent nécessairement des listes finies de mots et la constitution des listes représente des aires sémantiques liées aux méthodes expérimentales utilisées pour les constituer. D'un point de vue linguistique, il est important de définir ces variables afin de saisir la portée réelle de ces listes. En éducation, Fortier (1979) fournit une liste intéressante à consulter. L'application SATO-CALIBRAGE a, quant à elle, sa propre banque de données lexicales ou encore sa liste de mots connues.

Pour le cas qui nous intéresse, par exemple, le mot *subsistance* qui compte 11 caractères n'est pas présent dans le Vocabulaire fondamental de Gougenheim (*In* Henri, 1975) ni dans le Vocabulaire fondamental du québécois parlé de Beauchemin et Martel (*In* Bourbeau, 1988) ; la fréquence d'usage de ce mot est très limitée et devrait constituer une difficulté supplémentaire de compréhension. Toutefois, le mot *programmes* qui compte 10 caractères apparaît au singulier dans les deux listes.

D'autres observations peuvent être faites, notamment par la description des marqueurs de négation, des marqueurs de relation entre les propositions, des mots d'interrogation (pronoms, adverbes). Bref, il serait utile de faire sortir le lexique en fonction des catégories grammaticales, ce que SATO peut réaliser assez facilement.

L'intérêt d'une telle démarche est de mettre en évidence la complexité relative à certaines catégories grammaticales; par exemple, *ni* n'est pas en soi un mot complexe, mais lorsqu'il apparaît deux fois dans la même phrase, il en augmente nécessairement la difficulté de compréhension "*Ces frais ne s'appliquent ni au revenu de travail à votre compte ni à ceux relatifs à l'exécution d'une charge*".

La même remarque peut s'appliquer au nombre de verbes lorsqu'il est comparé au nombre de phrases ; ainsi, un nombre marqué de verbes par rappport au nombre de phrases indique la complexité des phrases. Il serait également possible de mettre en évidence d'autres éléments relatifs aux catégories grammaticales (référents des pronoms).

2. Exemple d'analyse de second niveau.

Nous présentons d'abord l'analyse du lexique faite par SATO pour le fascicule "Soutien Financier ...":

```
158 mots de 1 car. (6 %)          708 mots de 2 car. (26 %)
338 mots de 3 car. (12 %)         289 mots de 4 car. (11 %)
198 mots de 5 car. (7 %)          220 mots de 6 car. (8 %)
227 mots de 7 car. (8 %)          207 mots de 8 car. (8 %)
135 mots de 9 car. (5 %)          85 mots de 10 car. (3 %)
63 mots de 11 car. (2 %)          27 mots de 12 car. (1 %)
20 mots de 13 car. (1 %)          3 mots de 14 car. (1 %)
6 mots de 15 car. (0 %)           0 mot de 16 car. (0 %)
3 mots de 17 car. (0 %)           0 mot de 18 car. (0 %)
3 mots de 19 car. (0 %)           0 mot de 20 car. (0 %)
1 mot de 21 à 25 car. (0 %)       0 mot de 26 à 30 car. (0 %)
0 mot de plus de 30 car. (0 %)
nombre de mots .........2714      longueur moyenne : 4.8 car.
nombre de phrases.........157     longueur moyenne : 17.3 mots
nombre de paragraphes......2      longueur moyenne : 1357.0 mots

pourcentage de mots de 9 lettres et plus : 13 %
indice de Gunning : 12.1
```

On peut aussi utiliser SATO pour effectuer une catégorisation grammaticale hors contexte. Ainsi on découvre que 218 mots (28.46%) font partie de la catégorie "nom commun" alors que 80 (10.44%) sont des verbes conjugués. Il y a une soixantaine de catégories différentes pour rendre compte des nombreuses possibilités grammaticales des mots.

Nous avons ensuite soumis le lexique de ce document au lexique de SATO-CALIBRAGE (le lexique qui a été fait en collaboration avec le Ministère de l'Education). Nous constatons que 180 mots différents, sont considérés peu familiers par SATO-CALIBRAGE. Il est possible de faire apparaître à l'écran le texte où les mots considérés difficiles sont mis en évidence par un jeu différent de couleur. En imprimé, les mots peuvent être soulignés. Par exemple :

" - le remboursement d'impôts fonciers et le remboursement de la taxe de vente fédérale ; ..."

Enfin, selon SATO, ce fascicule présente 82 segments ou phrases de plus de 15 mots. À titre d'exemple,

" des montants supplémentaires sont accordés aux familles pour chacun de leurs enfants à charge de 18 ans et plus qui fréquente une école secondaire et pour chaque enfant qui réside avec ses parents et qui fréquente une école post-secondaire à temps plein ; "

Discussion

Les résultats des analyses faites par SATO nous ont aidés dans l'évaluation de l'intelligibilité de ces fascicules d'information. Ces documents s'avéraient généralement difficiles entre autres soit en fonction des éléments sémantiques ou encore des structures syntaxiques adoptées. Une reformulation des fascicules pour les rendre plus accessibles au public cible a donc été tentée. Cette reformulation intégre bien entendu les éléments propres aux niveaux de la microstructure, macrostructure et superstructure du modèle de Préfontaine et Lecavalier (1990).

Voici les résultats de l'analyse à la requête "lisibilité" pour les deux textes, avant la reformulation (version 1), et après la reformulation (version 2).

Longueur des mots,
des phrases et des paragraphes
(Sécurité du revenu, version 1)

22 mots de 1 car. (7 %)	71 mots de 2 car. (23 %)
42 mots de 3 car. (14 %)	48 mots de 4 car. (15 %)
19 mots de 5 car. (6 %)	19 mots de 6 car. (6 %)
21 mots de 7 car. (7 %)	21 mots de 8 car. (7 %)
13 mots de 9 car. (4 %)	11 mots de 10 car. (4 %)
3 mots de 11 car. (1 %)	4 mots de 12 car. (1 %)
3 mots de 13 car. (1 %)	7 mots de 14 car. (2 %)
1 mots de 15 car. (0 %)	0 mot de 16 car. (0 %)
3 mots de 17 car. (0 %)	0 mot de 18 car. (0 %)
0 mots de 19 car. (0 %)	0 mot de 20 car. (0 %)
0 mot de 21 à 25 car. (0 %)	0 mot de 26 à 30 car. (0 %)
0 mot de plus de 30 car. (0 %)	
nombre de mots 311	longueur moyenne : 5.1 car.
nombre de phrases.....23	longueur moyenne : 13.5 mots
nombre de paragraphes.....9	longueur moyenne : 34.6 mots

pourcentage de mots de 9 lettres et plus : 15 %

indice de Gunning : 11.6

Longueur des mots,
des phrases et des paragraphes
(Sécurité du revenu, version 2)

```
25 mots de 1 car. (9%)          59 mots de 2 car. (22%)
34 mots de 3 car. (13%)         38 mots de 4 car. (14%)
16 mots de 5 car. (6%)          14 mots de 6 car. (5%)
25 mots de 7 car. (9%)          20 mots de 8 car. (7%)
10 mots de 9 car. (4%)          9 mots de 10 car. (3%)
3 mots de 11 car. (1%)          1 mot de 12 car. (0%)
3 mots de 13 car. (1%)          5 mots de 14 car. (2%)
0 mot de 15 car. (0%)           0 mot de 16 car. (0%)
0 mot de 17 car. (0%)           1 mot de 18 car. (0%)
4 mots de 19 car. (1%)          0 mot de 20 car. (0%)
0 mot de 21 à 25 car. (0%)      0 mot de 26 à 30 car. (0%)
0 mot de plus de 30 car. (0%)

nombre de mots..... 267         longueur moyenne: 5.0 car.
nombre de phrases..... 29       longueur moyenne: 9.2 mots
nombre de paragraphes. 9        longueur moyenne: 29.7 mots

pourcentage de mots de 9 lettres et plus: 13%

indice de lisibilité de Gunning: 9.1
```

Conclusion

Nous avons décrit le contexte dans lequel nous avons utilisé SATO dans une tâche d'évaluation de fascicules d'information, de même que les résultats obtenus.

SATO est un outil formidable nous permettant d'obtenir des indices sûrs quant à la lisibilité de ces textes. De plus, SATO permet de faire ressortir les mots, les structures syntaxiques des textes qui contribuent à rendre ces textes difficiles.

Nous avons comparé le lexique de ces fascicules à la banque de données lexicales facilement accessible sur SATO. Cette mesure était un indicateur du niveau de difficulté des mots relativement satisfaisant en fonction de nos objectifs. Cependant, comme toutes les entreprises, ce ministère a un vocabulaire qui lui est propre et il faudrait constituer une nouvelle banque de données lexicales, propre à ce ministère. Certains mots considérés difficiles dans un certain milieu ne le sont plus dans un autre puisque très familiers. Une recherche empirique s'impose cependant parce que trop souvent le scripteur prend pour acquis que tel ou tel mot est tout à fait familier à son lecteur, ce qui n'est pas le cas. C'est donc auprès des bénéficiaires des services qu'il faudrait se référer.

Références

Bourbeau, Nicole, (1988), *C'est pas lisable ! La lisibilité des textes didactiques*, Guide pratique, Sherbrooke, Collège de Sherbrooke, 166p.

Gélinas-Chebat, C., Macot, M., Préfontaine, C., et Daoust, F. (1991). *La lisibilité de documents d'information du Ministère de la Main d'oeuvre, de la Sécurité du revenu et de la Formation professionnelle,* Avis professionnel présenté au Ministère de la Main d'oeuvre, de la Sécurité du revenu et de la Formation professionnelle, Gouvernement du Québec, 50 p.

Gunning, R. (1952). *The technique of clear writing*. New York: McGraw-Hill.

Henry, Georges, (1975), *Comment mesurer la lisibilité*, Paris, Fernand Nathan, Editions Labor, 176p.

Laroche, Léo (1990) *Calibrage des textes et lisibilité*, ICO Québec, Revue de liaison de la recherche en informatique cognitive des organisations, 2 (3), p.114 à 118.

Préfontaine, Cl. et Lecavalier, J. (1990). *La mesure de la lisibilité et de l'intelligibilité des textes*. Communication présentée à l'Association pour le développement de la mesure et de l'évaluation en éducation (ADMEE). Montréal, 25-27 octobre.

SATO au quotidien...

Georges Pelletier

Georges Pelletier est conseiller pédagogique de français à la commission scolaire de Victoriaville.

Choisir un texte pour l'élaboration d'une épreuve de lecture à soumettre à des élèves du primaire ou du secondaire est une opération courante dans l'accomplissement de ma tâche. Une des difficultés qui surgissent alors est, évidemment, l'adaptation du niveau linguistique du texte à la clientèle-cible.

Soutenir des enseignantes et des enseignants dans la planification de leur enseignement est aussi un rôle que je joue. Comment, alors, relever rapidement dans un texte à faire lire par des élèves les phrases qui contiennent des mots de relation, les phrases à multiples propositions, les phrases qui comptent plusieurs mots de substitution que sont les pronoms...?

Rédiger un texte destiné à une sensibilisation, à une formation ou à une simple information du personnel enseignant s'ajoute à la liste de mes productions. Je cherche alors à rédiger un texte facilement compréhensible, surtout si le sujet n'est pas familier aux lectrices ou aux lecteurs visés.

Dans tous ces cas, heureusement, SATO est là! En lui confiant l'analyse d'un texte, il peut rapidement en jauger la difficulté lexicale, calculer la longueur moyenne des phrases, extraire et imprimer les phrases excédant une longueur désignée, celles contenant des mots de relation, comptant un nombre précis de verbes conjugués, de pronoms, etc.

SATO est un outil de travail dont le degré d'utilité croît avec l'usage... Il est plus qu'un instrument pour établir des statistiques sur un texte écrit, il devient rapidement un outil indispensable, pour peu qu'on apprenne à l'apprivoiser.

Le calibrage de texte assisté par ordinateur, avantages et limites

Pierre Achim

Pierre Achim est conseiller pédagogique à la Commission scolaire du Long-Sault. Il est membre du groupe des usagers de SATO-CALIBRAGE.

SATO, un outil pour bien choisir ses textes en évaluation

Bien choisir un texte pour élaborer une épreuve de compréhension de la lecture est une des premières opérations susceptibles d'en garantir la validité. Sélectionner un texte trop facile revient à surévaluer l'élève. Choisir un texte trop difficile revient à le sous-évaluer. Le degré de difficulté des questions ne peut compenser la difficulté inadéquate, pourrait-on dire, du texte sans fausser la mesure exacte de l'état d'habilité à lire d'un élève.

Le calibrage d'un texte, dans l'orientation que nous venons de lui donner, c'est donc bien plus que la connaissance de son indice de lisibilité pour le simplifier, s'il y a lieu. Le calibrage sert à identifier les difficultés potentielles du texte afin de s'assurer que l'élève puisse réussir l'activité proposée. Dans certains cas, il est vrai, le calibrage peut conduire à modifier le choix des textes pour en trouver de plus simples ou de plus complexes, selon le cas. Connaissant mieux la nature du texte utilisé au cours d'une épreuve, on peut mieux interpréter les résultats pour identifier les capacités et les difficultés de l'élève.

Le calibrage, une activité essentielle pour assurer la validité d'un instrument de mesure

D'après Morissette, «la plus importante caractéristique de tout instrument de mesure, c'est la validité. On peut définir la validité comme étant la perfection avec laquelle un instrument réussit à mesurer la ou les variables qu'il propose de mesurer»[1].

La perfection est la réunion de toutes les qualités portées à leur degré le plus haut. Pour un instrument de mesure, ces qualités sont la pertinence, la congruence, la fidélité.

Sur la pertinence, Morissette indique: «Quand on se propose de démontrer logiquement la pertinence d'un test, il faut établir de façon convaincante que le test exige des sujets qui le passent, des comportments identiques à ceux spécifiés dans les objectifs du cours»[2].

Concernant la congruence, Morissette ajoute: «Il est nécessaire que l'objectif le plus spécifique soit un des éléments de l'objectif le plus général; ou que l'objectif le plus général inclut le plus spécifique»[3].

Finalement, d'après Turcotte: «Si l'évaluation, en tant qu'acte, conduit d'abord à faire la lumière sur l'état des apprentissages de l'élève par rapport à un contenu de programmes d'études, il appert que ce contenu doit être bien couvert pour éviter toute surévaluation ou sous-évaluation. Le contenu sera bien couvert par l'évaluation en autant que l'instrument de mesure alors utilisé

le mesurera entièrement. Ainsi, le degré de couverture du contenu visé par cet instrument donnera lieu à ce que l'on peut appeler la validité de contenu»[4].

La lecture, on le sait, implique la mise en branle de plusieurs objectifs simultanément. Aussi, ce qui est nécessaire pour la validité de la mesure d'un seul objectif n'en est pas moins vrai pour la mesure de plusieurs. D'où l'importance de prendre certaines précautions dans le choix du texte.

En considérant le problème du choix du texte sous cet angle, nous pourrions emprunter au rédacteur d'épreuves en mathématiques l'idée des facettes[5], les différents aspects du calibrage étant autant de facettes. Pour s'assurer d'un évaluation de qualité, nous devons avoir des textes qui proposent des difficultés de même niveau que celles que l'on retrouve dans les textes utilisés en situation d'apprentissage. D'ailleurs, les difficultés ne sont pas nécessairement dans le texte en soi mais dans l'absence de compétence ou de maîtrise du comportement adéquat (stratégie) par le lecteur. C'est pourquoi le lecteur ne peut résoudre l'embûche d'une structure de surface ou d'une structure profonde.

Le domaine étant l'ensemble des occasions où s'exerce l'habilité, le niveau de difficulté des textes doit donc être pris en compte dans l'élaboration de la situation tout autant que les éléments de programme à mesurer (connaissances déclaratives et procédurales). Le texte est l'occasion de la mise en exercice des connaissances situationnelles. Un texte ne peut être considéré comme équivalent à un autre simplement parce qu'il appartient à un même type de discours. Les connaissances situationnelles ou conditionnelles sont déterminantes sur le comment lire ces textes et sur la mise en branle des autres connaissances nécessaires à la lecture du texte (procédurales et déclaratives).

Utiliser l'information pour intervenir auprès de l'élève

La variation du degré de difficulté des textes peut expliquer les fluctuations des résultats à une épreuve. L'explication de ces fluctuations a été trop souvent laissée au hasard ou aux formules lapidaires: la forme des élèves, la température, l'humeur, la «cuvée». La malchance, pourquoi pas? Il fallait donc utiliser un autre instrument pour rectifier les résulats.

Le calibrage favorise une plus grande connaissance de l'instrument et, du fait même, une meilleure analyse des résultats. La connaissance des difficultés potentielles du texte permet d'établir une corrélation avec les erreurs et les lacunes des élèves.

Aussi, une évaluation permettant une meilleure analyse de l'habileté de l'élève permettra une meilleure orientation de l'action et l'identification précise d'interventions. Le calibrage nous permettant de mieux analyser les textes, nous permettra du même coup de mieux anticiper les stratégies nécessaires à la compréhension. Une meilleure connaissance des textes utilisés en classe permettra d'aider l'élève au bon endroit, au bon moment, avec les bonnes interventions pour favoriser l'acquisition des bonnes stratégies.

Il ne s'agit pas de réduire la difficulté, mais de la cerner pour mieux intervenir sans multiplier le nombre d'épreuves. Car, cette multiplication du nombre d'épreuves pour «tomber» sur une juste évaluation coûte très cher en temps et en énergie. Nous ne pouvons plus nous le

permettre.

Le choix des textes est tout aussi important en apprentissage qu'en évaluation. Un texte trop difficile peut décourager le lecteur et un texte trop facile n'améliore en rien ses capacités. Le texte choisi doit proposer un défi raisonnable que l'élève pourra relever avec l'aide de l'enseignant.

Les avantages et les limites d'un système informatisé

D'abord, le logiciel peut s'occuper des tâches fastidieuses: compter les mots, la longueur des phrases et des paragraphes, etc. Cependant, les résultats du calibrage sont toujours à valider par les experts. On peut, par exemple, établir la présomption qu'un mot est inconnu en le comparant à une liste de mots que nous savons connus. Il faut cependant vérifier cette présomption. Car, l'expérience des élèves dans un domaine ou un milieu donné peut permettre à ces élèves de trouver le sens de ce mot.

Aussi, l'observation des difficultés des élèves lors de la lecture de certains textes peut nous permettre d'identifier des structures syntaxiques posant problème à un moment ou l'autre du développement de l'élève. Il s'agit alors d'ajouter au logiciel des patrons de fouille qui vont nous permettre d'identifier ces structures.

Pour ce qui est des indices globaux, celui de GUNNING ou l'indice SATO-CALIBRAGE, il faut les interpréter comme des indications, à savoir qu'au niveau des structures de surface, le texte 1 peut être plus facile ou plus difficile que le texte 2. Par contre, la connaissance du sujet par le lecteur peut aplanir la difficulté présumée.

Comme le dit Tardif, «les connaissances générales en lecture qu'ont développées les bons lecteurs sont des connaissances utiles et importantes, mais (qu') elles ne peuvent leur permettre de comprendre un texte pour lequel ils n'ont pas de connaissances spécifiques»[6]. Donc, même si l'analyse des structures de surface peut nous permettre de dire qu'un texte est facile, le lecteur pourrait malgré tout le trouver difficile. Le logiciel ne peut donc se substituer au jugement du maître qui connaît ses élèves. Il peut cependant être un élément important pour qu'il se fasse un juste point de vue.

Notes

[1] MORISSETTE, Dominique, *Les examens de rendement scolaire*, PUL, 1979, p.237.

[2] idem, p.252.

[3] idem, p.20.

[4] TURCOTTE, Claude, *Un modèle d'évaluation de la compétence scolaire*, Tome 1, p. 68.

[5] SCALLION, G., La construction d'un test diagnostique selon des facettes.

[6] Tardif, J., *Pour un enseignement stratégique*, p. 56.

CINQUIÈME PARTIE

Conclusion

SATO-CALIBRAGE, perspectives de développement

François Daoust, Léo Laroche, Lise Ouellet.

Le projet SATO-CALIBRAGE est en marche depuis quatre ans. Disposant d'un petit budget, et bénéficiant de beaucoup d'implication de la part de ses principaux collaborateurs, le prototype réalisé nous fournit déjà des résultats qui dépassent nos attentes initiales.

De plus, le corpus rassemblé pour le projet, et le protocole expérimental que nous avons mis en place, nous fournissent la base pour de nombreuses autres études. Nous souhaitons donc attirer d'autres chercheurs qui pourraient profiter de cet environnement.

Dans l'immédiat, notre objectif est d'élargir l'accès au prototype. Cela implique entre autres de le rendre informatiquement de plus en plus facile à manipuler, et flexible. L'utilisation de l'atelier cognitif et textuel ACTE[1] devrait nous faciliter la tâche. Nous aimerions aussi pouvoir en élargir l'utilisation en variant le public cible. Nous pensons en particulier à un public adulte général, et à des textes de nature informative tels les prospectus gouvernementaux.

Il reste que l'évaluation de l'écriture est un problème très vaste et qu'il ne faudrait pas s'illusionner sur la possibilité, du moins dans l'immédiat, de réaliser des logiciels «clés en main». L'objectif est bien davantage de fournir des aides adaptées pour le calibrage de textes de divers types et destinés à divers publics. Aussi, nous continuons à privilégier l'approche «recherche action», c'est-à-dire une recherche qui s'inscrit dans la démarche des praticiens qui ont à manipuler les textes pour leur public cible, qui ont à en évaluer la réception auprès de leurs lecteurs. En termes pratiques, cela veut dire que nous continuerons à privilégier la constitution de comités d'utilisateurs. Cela veut dire que la diffusion du prototype devrait continuer à se faire à travers ces comités d'utilisateurs et par le biais de formules d'abonnement. Il faudra en effet pouvoir travailler à distance si on veut rejoindre des utilisateurs qui ne résident pas près des centres urbains de Montréal ou de Québec.

Au niveau de la recherche, les possibilités offertes par SATO-CALIBRAGE sont considérables. On n'ignore pas cependant que les difficultés risquent de croître au fur et à mesure que l'on s'éloignera de l'analyse microstructurelle. Les éléments d'analyse compris dans le prototype actuel sont encore embryonnaires, même si, déjà, ils fournissent des indications très utiles. Ce qui est intéressant, cependant, c'est que la méthode d'analyse mise en place peut permettre de juger expérimentalement de la pertinence des nouvelles dimensions d'analyse qui pourraient être ajoutées.

Nous invitons donc les lecteurs de ce Cahier à nous faire part de leurs commentaires et de leur désir, si tel est le cas, de participer à nos travaux.

Notes

[1] ACTE, Atelier cognitif et textuel, qui ajoute à SATO une coquille de générations de systèmes à base de connaissance. Voir *Acte, Manuel de références version 1*, 1993, Centre ATO-CI.

ANNEXES

Personnes impliquées

Comité de coordination

François DAOUST, analyste en informatique et chercheur au Centre d'ATO de l'UQAM.

Léo LAROCHE, statisticien, Direction de la recherche, ministère de l'Éducation.

Lise OUELLET, responsable de l'évaluation du français au primaire, Direction de la formation générale des jeunes, ministère de l'Éducation.

Groupe des usagers

ACHIM, Pierre	Conseiller pédagogique	C.S. du Long-Sault
COULOMBE, Johanne	Spécialiste en éducation	GRICS - Banque d'instruments de mesure
FORTIN, Marcel	Professeur de français	Collège de Sherbrooke
GÉLINAS-CHEBAT, Claire	Professeure	UQAM - dépt. de linguistique
GIROUARD, Lisette	Professeure de français	Cégep de Maisonneuve
GIROUX, Ginette	Conseillère pédagogique	C.S. Sainte-Croix
MORISSET, Diane	Conseillère pédagogique	C.S. Baldwin-Cartier
PELLETIER, Georges	Conseiller pédagogique	C.S. de Victoriaville
SARRAZIN, François	Conseiller pédagogique	C.S. Baldwin-Cartier

Personnel du Centre d'ATO

Fernande DUPUIS;
Sonia LAFOND;
Louis-Claude PAQUIN.

Enseignantes et enseignants consultés

En 1990,

BRIAND, Monique	C.S. Barraute-Senneterre	École N.D. du Sacré-Coeur
DUFRESNE, Mireille	C.S. Chomedey-de-Laval	École St-Norbert
GROLEAU, Rose-Alma	C.S. de Victoriaville	École Le Manège
HUBERDEAU, Hélène	CECM	École Baril
MARTEL, Lise	C.S. des Découvreurs	École St-Michel

En 1993,

BOURBEAU, Lise	C.S. St-Eustache	École Horizon-soleil
BUTEAU, Agathe	C.S. des Découvreurs	École Les Sources
GOSSELIN, Gilbert	C.S. des Découvreurs	École Les Sources
JOSEPH, Carole	C.S. de Victoriaville	École St-David

LAVIGNE, Yvon	C.E.C. de Verdun	Ecole Notre-Dame-de-la-Paix
MEILLEUR, Marie-Jeanne	C.S. Barraute-Senneterre	École polyvalente La Concorde

CONSULTANTS

Guy CUCUMEL, professeur au département des sciences comptables, Université du Québec à Montréal;
Pierre CHAMBERLAND, conseiller pédagogique, Commission des écoles catholiques de Montréal;
Claire GÉLINAS-CHEBAT, professeure au département de linguistique, Université du Québec à Montréal;
Michel PAGÉ, professeur au département de psychologie, Université de Montréal;
Clémence PRÉFONTAINE, professeure au département de didactique de l'Université du Québec à Montréal.

TRAVAIL TECHNIQUE

Georgette BÉLANGER, conseillère pédagogique, validation du corpus du primaire;
Maurice GÉLINAS, conseiller pédagogique, validation du corpus de secondaire;
Céline GOULET, secrétaire, entrée de textes;
Sylvie JUNEAU, animatrice, entrée de textes et de données;
Jean-François LAROCHE, entrée de données;
Marguerite TALBOT, vérification du corpus du primaire.

COLLABORATION

Robert BIBEAU, Direction des ressources didactiques;
Jean-Charles Chebat, professeur au département des Sciences Administratives, UQAM;
Nicole HUNEAULT, Direction des ressources didactiques;
Jacques Lecavalier, professeur au CEGEP de Valleyfield;
René Lortie, ministère des Communications;
Richard PARENT, ministère des Communications.

Contribution des différents partenaires

On trouvera ci-dessous la liste des différents collaborateurs qui ont apporté un soutien au développement du logiciel SATO-CALIBRAGE.

MINISTÈRE DES COMMUNICATIONS

Subvention au Centre D'ATO (1989-1990)

MINISTÈRE DE L'ÉDUCATION :

Direction générale de l'évaluation et des ressources didactiques
Subvention en 1990-1991
Ressources humaines (1986-1992) : membre du comité de coordination

Direction de la sanction des études
Subvention en 1991-1992
Ressources humaines : membre du comité de coordination

Direction du développement de l'évaluation
Ressources humaines (1988-1992) : membre du comité de coordination
Textes pour le corpus

Direction de la recherche
Ressources humaines (1992-1993) : membre du comité de coordination, analyses statistiques

Direction de la formation générale des jeunes
Ressources humaines (1992-1993) : membre du comité de coordination; coordination du jugement de familiarité sur le lexique; libération des enseignantes et des enseignants

Direction des ressources didactiques
Subvention en 1992-1993

UNIVERSITÉ DU QUÉBEC À MONTRÉAL

Centre d'ATO
Ressources humaines : développement de SATO-CALIBRAGE; membre du comité de coordination; personnel de recherche

Outils d'analyse textuelle
Locaux pour les réunions

SOCIÉTÉ GRICS (Gestion des Réseaux informatiques des commissions scolaires)

Banque d'instruments de mesure
Membres du groupe des usagers
Locaux pour les rencontres
Textes pour le corpus

COMMISSIONS SCOLAIRES

Commission scolaire du Long-Sault
Membre du groupe des usagers
Textes pour le corpus

Commission scolaire Sainte-Croix
Membre du groupe des usagers
Administration des subventions

Commission scolaire Baldwin-Cartier
Membres du groupe des usagers
Textes pour le corpus

Commission scolaire de Victoriaville
Membre du groupe des usagers
Textes pour le corpus
Libération de deux enseignantes pour juger de la familiarité du vocabulaire (1990 et 1993)

Commission scolaire Barraute-Senneterre
Libération de deux enseignantes pour juger de la familiarité du vocabulaire (1990 et 1993)

Commission scolaire Chomedey-de-Laval
Libération d'une enseignante pour juger de la familiarité du vocabulaire (1990)

Commission scolaire Saint-Eustache
Libération d'une enseignante pour juger de la familiarité du vocabulaire (1993)

Commission des écoles catholiques de Montréal
Libération d'une enseignante pour juger de la familiarité du vocabulaire (1990)

Commission des écoles catholiques de Verdun
Libération d'un enseignant pour juger de la familiarité du vocabulaire (1993)

Commission scolaire des Découvreurs
Libération de deux enseignantes et d'un enseignant pour juger de la familiarité du vocabulaire (1990 et 1993)

Maquette de la brochure: J. Ayoub
Composition et mise en pages: Fernand Daoust et Bernard Gadoua
Éditeur: Les Publications du Centre ATO.CI
Services de reprographie et d'imprimerie de l'UQAM, Canada
1994