FSH UQÀM *Faculté des sciences humaines*

*D*épartement de linguistique et de didactique des langues

Centre d'analyse de texte par ordinateur de la Faculté des Sciences Humaines

- **Présentation**
- **Corpus *express***
- **Nomino**
- **Nomino, moteur de recherche pour le gouvernement du Québec**
- **SATO-Internet**
- **Projet *Visibilité***

Pour nous rejoindre...

ATO est un centre d'expertise et de consultation en analyse de textes par ordinateur rattaché à la Faculté des Sciences humaines de l'Université du Québec à Montréal.

Le Centre est devenu au Québec un véritable chef de file dans son domaine, établissant de nombreuses alliances productives avec les secteurs privé et public. Le Centre s'est particulièrement fait connaître grâce aux logiciels **SATO** et **Nomino**.

Les recherches en ATO sont aux confins de la linguistique computationnelle, des sciences cognitives et des travaux en intelligence artificielle.

Elles trouvent des applications concrètes dans de nombreux secteurs de l'industrie et des services publics, notamment au niveau du développement de systèmes d'analyse automatique de l'information.

Les recherches en ATO ont permis le développement d'outils pour l'évaluation de l'écriture, pour l'assistance au dépouillement terminologique, pour la recherche qualitative dans les textes, pour l'indexation, la classification, et l'analyse de bases de données textuelles.

FSH UQÀM *Faculté des sciences humaines*

*D*épartement de linguistique
et de didactique des langues

Centre d'analyse de texte par ordinateur de la Faculté des Sciences Humaines

ATO est un centre d'expertise et de consultation en analyse de textes par ordinateur rattaché à la Faculté des Sciences humaines de l'Université du Québec à Montréal.

Le Centre est devenu au Québec un véritable chef de file dans son domaine, établissant de nombreuses alliances productives avec les secteurs privé et public. Le Centre s'est particulièrement fait connaître grâce aux logiciels **SATO** et **Nomino**.

Les recherches en ATO sont aux confins de la linguistique computationnelle, des sciences cognitives et des travaux en intelligence artificielle.

Elles trouvent des applications concrètes dans de nombreux secteurs de l'industrie et des services publics, notamment au niveau du développement de systèmes d'analyse automatique de l'information.

Les recherches en ATO ont permis le développement d'outils pour l'évaluation de l'écriture, pour l'assistance au dépouillement terminologique, pour la recherche qualitative dans les textes, pour l'indexation, la classification, et l'analyse de bases de données textuelles.

ANALYSE DE TEXTES PAR ORDINATEUR (ATO)

ATO EBSI

[A propos du projet... | Outils et méthodes | Démonstrations SATO | Publications]
[Activités et applications | Glossaire | Forum de discussion | Carnet d'adresses]

Le projet VISIBILITÉ entend tirer profit des immenses possibilités du réseau Internet pour la diffusion d'information aux fins de rendre accessibles, pour une clientèle élargie, un ensemble de ressources en matière d'analyse de textes par ordinateur (ATO).

Cette clientèle, constituée notamment des chercheurs universitaires, des organismes publics et des entreprises, peut assurément prendre avantage des potentialités des méthodes et outils d'ATO dans la gestion et l'analyse d'une information électronique surabondante et variée.

Tout en voulant favoriser l'information et la formation en ATO, ce projet vise par la même occasion le développement d'une méthodologie intégrée, soutenue par une vision cohérente et globale, en analyse de textes par ordinateur, secteur qui, depuis ses origines, rassemble les points de vue de plusieurs disciplines autour de l'«objet textuel ».

La réalisation de ce site

est le fruit d'une collaboration entre trois partenaires:

Le Service ATO
de l'Université du Québec à Montréal
(UQAM)

L'École de bibliothéconomie
et des sciences de l'information (EBSI)
de l'Université de Montréal

La firme montréalaise DOCUMENSA

Le projet VISIBILITÉ bénéficie du support financier des organismes suivants qui l'ont subventionné dans le cadre du programme "Action concertée FCAR-CEFRIO-CRIM":

FCAR CEFRIO CRIM

[À propos du projet...]

[Démonstrations SATO]

Outils et méthodes

Au coeur de l'ATO, on trouve des logiciels, des outils, des façons de faire... à découvrir !

Présentation

Cette section présente des exemples de méthodologies d'approche de certains problèmes d'analyse de textes. Elle comporte également les informations relatives aux outils d'analyses de textes développés au Service ATO.

Méthodes générales

L'analyse de textes par ordinateur requiert des stratégies et des algorithmes en fonction du but visé. Depuis plusieurs années, des experts de différents domaines de spécialité (entre autres, communication, psychologie, bibliothécomie, sociologie, linguistique) ont élaboré ces stratégies devant leur permettre de traiter des corpus de grande taille à l'aide de Sato. Les références suivantes font état de quelques unes des méthodologies utilisées à fins d'indexation automatique de textes, d'extraction automatiques d'unités d'informations, d'analyses quantitatives et qualitives de textes, etc.

Bien entendu cette liste ne couvre pas tous les problèmes en ATO, mais elle propose par l'exemple, des solutions utilisées dans la résolution de problèmes traités à l'aide de Sato. Voir également la section publications.

- **Analyse du discours**
  - L'analyse du contenu textuel ...
  - Protocole de description ...
  - Système d'analyse de contenu ...
  - Exemple d'analyse de documents ...
- **Analyse quantitative et/ou qualitative de textes**

Les traitements statistico-linguistiques ...

Programme d'analyse de lisibilité ...

- **Approche méthodologique de la catégorisation**

  La catégorisation socio-sémantique ...

- **Construction assitée de thésaurus**

  L'analyse du contenu textuel ...

- **Extraction d'informations et acquisition des connaissances**

  Analyse de textes et acquisition des connaissances ...

- **Problème d'ordre linguistique**

  Algorithme de désambiguïsation catégorielle ...

  Le dispositif linguistique ...

- **Techniques d'indexation des textes**

  L'analyse du contenu textuel ...

  Prototype d'un système expert ...

- **Traitement des corpus**

  Étude d'un corpus ...

  Le traitement des textes du primaires ...

Outils logiciels

Ils sont disponibles au Service ATO, pour achat ou démonstration.

Logiciels généraux

- **SATO**
  - Présentation du logiciel
  - Démonstrations et téléchargement du logiciel (mode démonstration)
  - Manuel de référence SATO 4.0

- **Atelier FX** et **Nomino**
  - Présentation des logiciels
  - Télécharger les logiciels (mode démonstration)

Modules

- **Base de données lexicales (BDL)**

Il s'agit d'un dictionnaire consultable par SATO qui permet d'attribuer des catégories grammaticales aux mots d'un texte.

- Présentation de la BDL
  - Notes explicatives sur les descriptions lexicales
  - Nomenclature des valeurs symboliques
- Installation et ressources de la BDL
- Télécharger [gramr.exe]

Tutoriels

Ressources pouvant assister l'apprentissage des logiciels ou de la méthodologie d'ATO.

Remarque : il existe plusieurs fiches récapitulatives des procédures des commandes Sato pouvant vous aider à traiter vos textes. La mise à jour des fiches pour une diffusion sur l'internet n'est encore terminée. En guise d'exemple, nous donnons une fiche résumant les préparations de bases pour une 'satogération'. N'hésitez donc pas à revenir souvent dans cette section !

- Procéduriers
  - Codification de base d'un texte pour Sato
    - version HTML
    - version SGML [lire d'abord cette note sur la consultation des textes en SGML]

- Commentaires -

L'analyse du contenu textuel en vue de la construction de thésaurus et de l'indexation assistées par ordinateur; applications possibles avec SATO (système d'analyse de textes par ordinateur)

Par

Suzanne Bertrand-Gastaldy*

et

Gracia Pagola**

Résumé

L'intervention de l'ordinateur, longtemps réservée aux tâches mécaniques effectuées en aval de l'analyse des documents et de la constitution des thésaurus, se déplace en amont vers l'analyse elle-même. Des logiciels existent désormais qui assistent l'exploration des textes. On montre comment SATO (système d'analyse de textes par ordinateur), utilisé par des chercheurs de plusieurs disciplines, peut faciliter les tâches de contrôle et de structuration du vocabulaire ainsi que l'indexation. On présente ses caractéristiques importantes: possibilité d'ajouter des propriétés aux mots et aux segments textuels, génération de lexiques, analyses lexico-statistiques diverses, définition de sous-ensembles de textes et de lexiques. On examine ensuite l'aide apportée dans l'élaboration de thésaurus: extraction d'unités lexicales simples et complexes, pondération pour faciliter le choix, regroupements divers, repérage en contexte de formes équivalentes, de synonymes, de termes génériques et spécifiques, de termes associés. L'indexation assistée par ordinateur est également illustrée, avec des perspectives d'analyse "sur mesure", de même que plusieurs stratégies d'interrogation. On conclut sur la nécessité de résoudre les questions théoriques auxquelles l'analyse du contenu confronte désormais les spécialistes de l'information.

INTRODUCTION

L'intervention de l'ordinateur a été longtemps réservée aux tâches mécaniques, notamment à celles qui sont effectuées en aval de l'analyse des documents et de la constitution des thésaurus, mais avec la disponibilité des textes sur ordinateur cette intervention se déplace en amont, vers l'analyse elle-même, vers l'exploration du contenu des textes.

La construction de vocabulaires contrôlés et l'indexation, loin de disparaÎtre avec la multiplication des bases de données en plein texte, comme certains l'avaient prédit à la fin des années 1970, connaissent au contraire un regain d'intérêt. *Information Processing & Management* a consacré, en 1990, tout un numéro à l'analyse textuelle et une livraison entière de *International Classification* (no3-4, 1990) a été réservée aux logiciels de gestion de thésaurus. La lecture de publications récentes confirme que l'intérêt gagne même les

cercles non bibliothéconomiques: dans le numéro d'avril 1991 de *Byte* sur le bureau sans papier ("Paperless Office"), Locke entreprend de convaincre les gestionnaires qui ont opté pour un système de gestion électronique des documents (GED en français et DIP - "Document Image Processing" en anglais) de la nécessité de l'indexation et du contrôle de vocabulaire pour accéder au contenu des textes. Il souligne la nature complexe de la tâche "[...] indexing is not a low-level task , and it becomes more complex as larger volumes of text are involved" (p.194). Il ajoute: "In fact, the subject analysis that librarians perform to create these categories and relationships is strongly akin to what the AI literature calls *knowledge engineering* ."

Parce que le contenu des bases de données textuelles ne peut pas être exploité de façon satisfaisante pour les utilisateurs de plus en plus nombreux et diversifiés avec les méthodes traditionnelles de mots-clés ou de chaÎnes de caractères des systèmes bibliographiques, les spécialistes en sciences de l'information n'ont d'autre choix que de collaborer avec les autres spécialistes de la langue et des textes. Leurs travaux trouvent naturellement une place dans les industries de la langue, comme on peut le constater d'après le programme des congrès sur ce thème (Colloque "Les industries de la langue: Perspectives des années 1990" en novembre 1990 à Montréal). Il ne s'agit pas de leur part d'une démarche opportuniste, mais bien de la reconnaissance de la nature linguistique du matériau à traiter (Bertrand-Gastaldy, 1990a). D'ailleurs la collaboration ne s'effectue pas à sens unique. Les thésaurus ou des outils approchants deviennent nécessaires aux linguistes pour le traitement automatique des langues:

"In a rather rough statement one could say that is now the linguists, who on the one hand, need thesauri, or thesaurus-like conceptual structures, to solve their problem of meaning, i.e. the problem of language understanding, whereas the IR [information retrieval] systems designers, in turn, finally came to know that basic LE [linguistic engineering] is required in their systems to come up with more efficient, intelligent, machine-aided IR systems." (Schmitz-Esser, 1990: 130).

De plus en plus de parallèles sont établis d'une part entre les thésaurus et les méthodes d'indexation avec grilles et, d'autre part, certains modes de représentation des connaissances en intelligence artificielle, comme les réseaux sémantiques et les "frames".

Bien qu'un faible pourcentage de textes soit disponible sur support lisible par ordinateur (de 3 à 5% selon les estimations), il n'en reste pas moins que la proportion augmente rapidement et que des outils sont désormais disponibles pour mieux exploiter les textes. Nous n'avons pas l'ambition de brosser un portrait des diverses zones d'intervention des logiciels dans le traitement de la langue des textes et des questions ni des difficultés à surmonter; nous l'avons fait ailleurs (Bertrand-Gastaldy, 1990a et 1990b). Nous tenterons de montrer comment les tâches de construction de thésaurus, d'indexation et, par le fait même, de repérage peuvent bénéficier de l'assistance de l'ordinateur pour explorer le contenu de bases de données textuelles. Nous prendrons pour cela l'exemple du logiciel SATO qui n'est en aucune façon dédié aux opérations documentaires.

PRÉSENTATION D'UN OUTIL D'AIDE À L'EXPLORATION DU CONTENU DES TEXTES: SATO

Les utilisateurs du logiciel

Le logiciel SATO (système d'analyse de textes par ordinateur) a été conçu par Jean-Guy Meunier et développé par François Daoust, responsable de l'équipe ITC (Ingénierie cognitive et textuelle) au Centre d'ATO de l'Université du Québec à Montréal. Il tourne sur des ordinateurs IBM-PC et compatibles. C'est un logiciel général destiné à un public de chercheurs en sciences humaines et sociales, désireux de faire de l'analyse de contenu pour des raisons fort diverses.

Au ministère de l'Éducation, on l'utilise pour évaluer la lisibilité des textes destinés aux élèves du primaire et du secondaire, en fonction de leur connaissance du vocabulaire et de la syntaxe (Laroche, 1990).

Il a servi, avec Termino, à faire une analyse lexicologique des réponses lors d'une pré-enquête sur "Les lycéens 91" en France (*Le Monde*, juin 1991). On peut tout aussi bien y recourir pour les retranscriptions d'entrevues ou d'analyses de protocoles lorsqu'on fait verbaliser un sujet au cours de l'accomplissement d'une tâche. Les psychologues font de même pour analyser et comparer le matériel d'entretien thérapeutique.

Il a été exploité pour l'analyse de textes politiques, notamment par le sociologue Jules Duschatel qui a étudié, entre autres, le discours politique sous le régime Duplessis (Bourque et Duchastel, 1988), ainsi qu'un corpus de dossiers de la cour juvénile de Winnipeg (Duchastel, 1991).

Des linguistes recourent à SATO pour découvrir des structures syntaxiques ou des structures argumentatives dont ils ne posssèdent pas encore de description. Dans sa thèse de doctorat, Anaïd Donabedian s'en est servi pour découvrir des régularités dans l'emploi de l'article en arménien ancien. Monique Lemieux y trouve une aide précieuse pour l'analyse de la grammaire du moyen français.

Le logiciel sert d'appui aux cogniticiens chargés de dépouiller les textes pour l'extraction des connaissances et la construction de systèmes experts (Paquin *et al.*, 1990), aux terminologues pour l'élaboration de terminologies (Auger, 1990) et finalement, comme nous allons le montrer, à une variété de tâches reliées à la gestion de l'information textuelle en vue d'en faciliter le repérage. D'ailleurs plusieurs des contributions de ce numéro illustrent des réalisations documentaires effectuées à l'aide de SATO.

Les caractéristiques principales

SATO présente des caractéristiques bien différentes de la plupart des logiciels documentaires commercialisés. Quelques explications préalables sont nécessaires pour apprécier les différentes opérations qu'il permet.

Sa grande originalité de SATO réside dans le fait qu'il permet d'**ajouter des propriétés aux mots ou segments textuels**.

Au point de départ, tout texte est représenté comme une suite de caractères (y compris le caractère blanc), ce qui est en soi très banal. Restitués à l'écran, ces caractères et les suites de caractères ne peuvent prendre de signification que s'ils sont interprétés par un agent cognitif humain doté de multiples connaissances:

- connaissances de l'alaphabet et des divers signes graphiques

- connaissances des morphèmes et des mots de la langue générale, connaissances de la

signification des termes particuliers au domaine

- connaissances de la syntaxe

- connaissances de sens commun

- connaissances du domaine lui-même

- connaissances de la structure des textes de tel ou tel type (lois, articles scientifiques, rapports de diagnostic médical, etc.)

- connaissances de la signification de certains types de caractères dans le contexte particulier de tel ou tel texte, de tel ou tel ensemble de textes (par exemple, les italiques peuvent indiquer qu'il s'agit d'un titre, d'un mot étranger, les caractères gras sont la marque d'un titre ou d'un sous-titre, les capitales signalent un mot qui est défini dans un glossaire, etc.)

C'est donc l'ensemble de ces connaissances qui interagissent au cours de la lecture pour que du sens soit produit à partir de ce qui n'est que traces sur le papier ou à l'écran. Des opérations cognitives très complexes ont alors lieu qui peuvent consister en généralisation, élimination de détails inutiles, catégorisation, comparaison, reconstruction, etc. Mais la mémoire humaine à court terme a des capacités très limitées, alors que les bases de données peuvent atteindre des gigabytes, d'où l'intérêt de "passer la main" à un logiciel pour des analyses sur des corpus d'envergure.

L'inconvénient c'est que le processeur informatique, contrairement au "processeur humain", ne connaÎt rien ou presque rien. Cette métaphore permettra de mieux comprendre ce qu'il faut faire pour qu'un logiciel puisse procéder lui aussi à des catégorisations grammaticales, sémantiques, textuelles, pragmatiques, à des généralisations, à des comparaisons, etc. sur des ensembles de caractères. Des caractéristiques ou propriétés doivent donc être rajoutées, surimposées soit aux mots du lexique, soit aux mots en contexte. Nous en verrons des exemples plus loin. Retenons que SATO conserve en mémoire une représentation fidèle du texte de départ (ce qui permet d'obtenir des références très exactes sur la position des mots à l'intérieur des documents, pages et lignes) et en fait, selon une image empruntée à Maurice Gingras, une photocopie sur laquelle il est possible d'annoter à volonté. Les annotations (qui sont constituées de valeurs de propriétés) peuvent d'ailleurs apparaÎtre en couleurs comme autant de traits de surligneurs.

L'ajout de propriétés peut être:

1) le résultat d'une opération automatique (les mots en capitales reçoivent la valeur Cap de la propriété Édition, la fréquence devient une propriété numérique des formes);

2) le résultat d'une opération automatique déclenchée par l'analyste (les codes propres à un logiciel de traitement de texte comme le souligné ou les caractères gras peuvent être convertis en valeurs de la propriété Typo; la projection d'un thésaurus ou d'une base de données lexicales sur le lexique résulte en l'attribution à chaque forme des valeurs déclarées);

3) le résultat d'une opération humaine effectuée au cas par cas dans le texte (segmentation et nomination des diverses subdivisions) ou dans le lexique (catégorisation sémantique du vocabulaire hors contexte). À la demande, il est possible de faire passer les propriétés textuelles dans le lexique et vice versa (on parle alors d'héritage de propriété).

L'exemple suivant illustre l'ajout de deux propriétés dans le texte (notice et zone); il s'agit d'une notice de MEDIADOQ dont la structuration en champs a été exploitée et qui a subi un pré-traitement léger, comme on peut le déceler d'après le doublement du point d'abréviation à ne pas confondre avec le point de fin de phrase pour la segmentation automatique en phrases:

***notice=1**

***zone=na** Do ... #253

***zone=no** A880333

***zone=au** Rhéaume, Luc

***zone=tm** Analyse du traitement journalistique de l'information politique au Québec: le cas du projet de restructuration scolaire de 1982

***zone=so** ix, 116, x-xxix f..: tableaux, graph..; 29 cm.. - Mémoire (M..A..) - Université Laval, 1984

***zone=re** Les formes de dépendance des journalistes à l'égard des sources d'information sont examinées par le biais d'une analyse de contenu de la couverture accordée par la presse écrite francophone au débat entourant le projet de restructuration scolaire proposé par le Gouvernement du Québec en 1982. La première partie porte sur l'information en tant qu'enjeu d'une lutte politique. La deuxième partie expose la méthodologie de la recherche. Cinq journaux quotidiens québécois ont été analysés: La Presse, Le Devoir, Le Soleil, Le Journal de Québec et Le Journal de Montréal. L'analyse porte sur 4 aspects précis de la couverture: les sources d'information, les thèmes abordés, l'approche de traitement et les tendances exprimées. La troisième partie présente les résultats de l'analyse. Il ressort que tous les intervenants ont eu un droit de parole dans les pages des journaux, mais que la couverture a été superficielle et axée sur les aspects conflictuels du dossier. La presse ne serait pas en mesure d'expliquer et de vulgariser des problèmes complexes. Par ailleurs les chroniqueurs spécialisés et les journalistes font preuve d'une plus grande autonomie à l'égard des sources d'information que les rédacteurs de nouvelles. Ceux-ci ne font que réagir aux initiatives des sources et adoptent une approche très descriptive; ils se limitent à une fonction de transmission du message des sources. bibliogr..: ff.. x-xiv, (J..C..)

***zone=dep** TRAITEMENT DE L'INFORMATION; PRESSE; INFORMATION POLITIQUE; SOURCE D'INFORMATION

***zone=idp** SOLEIL, LE; DEVOIR, LE; PRESSE, LA; JOURNAL DE Québec, LE; JOURNAL DE Montréal, LE

Comme on peut modifier à volonté les propriétés et leurs valeurs, il n'est pas pénalisant de

revenir sur le découpage des textes, au fur et à mesure que les besoins d'analyse se précisent.

Du texte, on procède à la **génération du lexique** dans lequel on peut faire apparaÎtre les propriétés avec leurs valeurs en autant de colonnes.

La fréquence - absolue ou relative - est, tout comme la propriété Édition ou la propriété Alphabet une propriété.prédéfinie. Les autres peuvent être définies au besoin et on peut leur donner n'importe quelles valeurs, symboliques ou numériques selon le cas. Voici un extrait de lexique avec affichage de quatre propriétés: l'alphabet (français), la fréquence absolue, la fréquence relative et la (ou les) zones de provenance des formes extraites:

alph	fréqabs	fréqrel	zon	(lexique)
fr	1	0.02	re	abord
fr	1	0.02	re	aborde
fr	1	0.02	re	abordés
fr	1	0.02	so	abreuver
fr	1	0.02	re	accent
fr	1	0.02	re	acceptation
fr	7	0.12	re	accepter
fr	1	0.02	(tm,re,dep,idp)	accès
fr	1	0.02	re	accord
fr	1	0.02	re	accorde
fr	3	0.5	(so,re)	accordée
fr	2	0.4	re	acheminer
fr	10	0.18	re	acteurs
fr	5	0.9	(tm,re)	activités
fr	1	0.02	re	actuel
fr	1	0.02	re	actuels
fr	1	0.02	re	adhésion
fr	1	0.02	re	administrateurs

Avec SATO on peut effectuer des **analyses lexico-statistiques**. L'exemple suivant indique successivement la moyenne, l'écart-type, la répartition, l'indice de discrimination de Salton et le chi2:

Moyenne	Écart	Répart.	Discri.	Chi2	
0.05	0.22	5.0%	0.80	0.20	abord
0.05	0.22	5.0%	1.02	0.26	accent
0.05	0.22	5.0%	0.59	0.15	acceptation
0.35	1.31	10.0%	4.33	0.55	accès
0.05	0.22	5.0%	0.81	0.21	accord
0.15	0.36	15.0%	0.56	0.19	accordée
0.50	1.24	15.0%	2.77	0.48	acteurs

0.25	0.89	10.0%	2.80	0.43	activités
0.05	0.22	5.0%	1.01	0.26	adhésion
0.05	0.22	5.0%	0.94	0.24	administrateurs
0.20	0.68	10.0%	2.96	0.51	administration
0.10	0.44	5.0%	2.22	0.40	agence
0.55	1.36	25.0%	3.08	0.52	agences
0.05	0.22	5.0%	0.59	0.15	aide
0.05	0.22	5.0%	0.80	0.20	ailleurs
0.05	0.22	5.0%	1.02	0.26	alarmiste
0.05	0.22	5.0%	1.09	0.28	alliances
0.10	0.44	5.0%	1.76	0.32	américaines

Une seule commande suffit pour obtenir la participation d'une forme dans les textes ou sous-textes d'un corpus:

Participation de information

---

nombre de lexèmes: 1

---

domaine vocabulaire/domaine domaine/texte

---

fréqtot 101 mots (1.80%) 100%

Cet affichage indique que le mot *information* apparaÎt 101 fois dans le corpus et constitue 1.80% de l'ensemble du texte.

On peut travailler séparément sur le lexique - et le corpus - de chaque langue , dans le cas de textes multilingues (maximum de quatre langues).

On peut connaÎtre les formes qui distinguent le plus un texte d'un autre, un champ d'un autre dans les notices bibliographiques, grâce à la commande Distance. Ainsi, à partir des subdivisions introduites par une propriété textuelle, on peut comparer, par exemple, le vocabulaire de deux bases de données ou celui des différentes zones des notices, celui des titres et des sous-titres, celui des introductions et celui des conclusions, celui d'un chapitre de livre par rapport à un autre (ou à l'ensemble des autres, comme dans l'exemple suivant):

fréqtot	**fréqchap**	**autrchap**	**explique**	**cumul**	**(lexique)**
0.10	2.54	0.00	2.18	2.18	préparation_des_surfaces*
0.05	1.69	0.00	1.94	4.12	cuivrage*
0.05	1.69	0.00	1.94	6.06	décapage*
0.05	1.69	0.00	1.94	8.00	oxydes*
0.05	1.69	0.00	1.94	9.95	étamage*
0.12	2.54	0.00	1.75	11.69	revêtement_de_surface*

0.07	1.69	0.00	1.29	12.99	électrodéposition*
0.02	0.85	0.00	0.97	13.96	absorbante*
...					

Un indice de lisibilité est fourni à la demande. Il est calculé selon la formule de Gunning (dont on conteste cependant la validité pour le français) et tient compte de la longueur des phrases et de la proportion de mots longs (9 caractères et plus). Mais les chercheurs du ministère de l'Éducation travaillent à d'autres indices qui sont fonction des capacités de compréhension des élèves selon leur âge:

---

201 mots de 1 car. (4%) 1372 mots de 2 car. (28%)

---

619 mots de 3 car. (13%) 360 mots de 4 car. (7 %)

---

286 mots de 5 car. (6%) 390 mots de 6 car. (8%)

430 mots de 7 car. (9%) 298 mots de 8 car. (6%)

267 mots de 9 car. (5%) 229 mots de 10 car. (5%)

203 mots de 11 car. (4%) 96 mots de12 car. (2%)

80 mots de 13 car. (2%) 43 mots de 14 car. (1%)

31 mots de 15 car. (1%) 10 mots de 16 car. (0%)

2 mots de 17 car. (0%) 1 mots de18 car. (0%)

2 mots de 19 car. (0%) 0 mots de 20 car. (0%)

3 mots de 21 à 25 car. (0%) 0 mots de 26 à 30 car. (0%)

0 mots de plus de 30 car. (0%)

---

nombre de mots ..... 4923 longueur moyenne : 5.3 car.

---

nombre de phrases ..... 318 longueur moyenne : 15.5 mots

nombre de paragraphes..... 1 longueur moyenne : 4923.0

---

pourcentage de mots de 9 lettres et plus : 20%

---

indice de lisibilité de Gunning : 14.0

Nous arrêterons là notre présentation générale. D'autres fonctionnalités seront expliquées au fur et à mesure des besoins.

Depuis trois ans environ, nous avons testé, dans des cours et des projets de recherche, les capacités de SATO à traiter les textes pleins ou les textes réduits (notices bibliographiques et

analytiques) pour différentes fonctions documentaires. Nous examinerons d'abord comment SATO peut aider à contrôler et structurer un vocabulaire de domaine.

Les opérations que nous allons présenter supposent que les textes ont été soumis au préalable en format ASCII, avec au minimum la déclaration de la propriété Alphabet et un titre.

CONSTRUCTION DE THÉSAURUS

Effectuée de façon traditionnelle, soit à partir des résultats de l'indexation manuelle des textes (méthode *a posteriori*), soit en consultant les ouvrages terminologiques ou lexicographiques ainsi que les experts du domaine (méthode *a priori*), la construction de thésaurus est une tâche trop onéreuse, insuffisamment rigoureuse et le résultat trop décalé par rapport aux besoins pour que les organisations y accordent volontiers les ressources nécessaires.

LA NÉCESSITÉ DU CONTR&OCIRC;LE DU VOCABULAIRE

Pourtant les résultats des évaluations du repérage en plein texte prouvent la nécessité d'une telle opération. Le taux de rappel est bas si un réseau de relations ne vient pas suggérer des moyens d'élargir les stratégies par des termes génériques ou par d'autres termes spécifiques d'une même classe. Par exemple, si l'on veut récupérer tout ce qui traite des *crucifères cultivées*, il faut savoir que l'information peut être dispersée sous chacun des termes spécifiques: *choux, choux-fleurs, brocolis, choux-de-Bruxelles, choux chinois, rutabagas, radis*. C'est ainsi que les évaluations de Blair (1986) et Blair et Maron (1985) sur le système STAIRS/TLS révèlent un taux de rappel de 20% seulement. L'étude de Balcer et Gonin (1979) avait abouti à un taux de 41,3 %, performance peu reluisante que les auteurs attribuaient à l'absence de renvois et par conséquent à des stratégies trop limitées. En plein texte, le taux de précision peut, à force de reformulations de la stratégie de recherche, atteindre un score très satisfaisant, mais c'est souvent au prix de beaucoup d'effort et de temps. D'après Garson et Love (1985), une référence pertinente dans ACS Journals Online, interrogée sur le serveur BRS, coûte en moyenne le double d'une notice trouvée dans Chemical Abstracts sur le serveur DIALOG. Tenopir (1985) a obtenu le même genre de résultats dans une étude qui comparait la recherche sur le plein texte et la recherche sur les résumés et les descripteurs, dans Harvard Business Online .

L'INFORMATISATION DES OPÉRATIONS LIÉES À LA CONSTRUCTION DES THÉSAURUS

L'automatisation a surtout touché jusqu'à présent la gestion des résultats de la collecte, du contrôle et de la structuration des termes. De nombreux logiciels existent qui assurent la validation, la réciprocité des relations, la mise à jour du contenu des thésaurus, l'édition, etc. en conformité avec les normes nationales et internationales.

Depuis longtemps déjà, des méthodes statistiques ont été testées pour extraire le "vocabulaire" de corpus textuels et en construire une représentation structurée selon des algorithmes de classification automatique. On pense surtout aux travaux de Salton (Salton, 1971) et de Sparck Jones (1971)et à tout le débat sur les "thésaurus de recherche" (Lancaster, 1977; Bertrand-Gastaldy, 1984; Richer, 1986). L'intérêt des résultats de ces

méthodes est cependant limité à certains types d'utilisation: les représentations orientent les utilisateurs qui connaissent peu le contenu de la base de données interrogée ou bien offrent des suggestions de "mots" liés d'une certaine manière à ceux qui figurent dans les stratégies de recherche, mais l'interprétation de la nature des liens reste à la charge des utilisateurs. Couplés à un système d'inférences chargés d'élargir ou de restreindre les stratégies, de tels outils risquent de diriger la recherche dans toutes sortes de direction . D'autre part les unités de représentation extraites sont hétérogènes: ce sont tantôt des "mots", tantôt des lexies, tantôt des termes, tantôt des radicaux.

L'opération de sélection et de structuration fine des termes d'un domaine, affaire de spécialistes qu'ils soient indexeurs, terminologues ou "ingénieurs cogniticiens", souffre encore de la rareté d'outils adéquats pour l'assister, ce qui a fait écrire à Ranjard (1991): "[...] force est de reconnaÎtre que les outils de gestion de thésaurus n'aident en rien à la conception des vocabulaires contrôlés. [...], ce qui contraint les documentalistes à attendre "que les outils informatiques d'analyse de contenu des textes soient à la portée de tous".

L'AIDE APPORTÉE PAR SATO DANS LES DIFFÉRENTES ÉTAPES D'ÉLABORATION D'UN THÉSAURUS

SATO peut apporter une aide non négligeable dans la conception de vocabulaires de domaines, comme nous allons tenter de l'illustrer.

Choix du moment opportun pour entreprendre la construction d'un thésaurus

Mentionnons en passant que, par ses fonctions statistiques, SATO fournit des indices utiles pour qui doit décider de l'opportunité d'entreprendre la construction de thésaurus. La fréquence moyenne des formes, ou taux de répétitivité, la proportion des mots de fréquence faible (1,2 et 3) et le calcul des formes nouvelles à chaque ajout d'une certaine quantité de textes permettent d'apprécier le plafonnement du renouvellement du vocabulaire, moment qu'il est sage d'attendre si l'on ne veut pas faire face à de nombreux problèmes de mise à jour.

Extraction du vocabulaire

Les formes simples

Comme tout autre logiciel, SATO fournit le lexique des formes simples contenues dans la base de données. Il tient compte de tous les caractères et n'élimine ni les signes de ponctuation ni les formes fonctionnelles, sauf si on le demande. Nous verrons plus loin l'intérêt des les conserver.

La catégorisation grammaticale des formes simples

SATO permet d'affecter à chaque forme du lexique la ou les catégories grammaticales qu'elle peut prendre hors contexte. L'affectation s'effectue de façon automatique, avec la procédure Dogramr à partir de la consultation de bases de données lexicales. Notons qu'on peut recourir aux sources lexicographiques ou terminologiques de son choix. Le résultat se présente ainsi:

	fréq	**gramr**

1	nomp	Canada
2	(adj,nomc)	canadienne
2	(adj,nomc)	canadiens
1	(adj,nomc)	capitalistes
1	nomc	caractère
1	v_conj	caractérise
1	(adj,nomc)	caractéristiques
1	nomc	cas
1	nomc	catégories
2	adj	catholique
3	adj	catholiques
1	nomc	causes
11	(détdém,p_dém)	ce

Une simple commande permet de connaÎtre le pourcentage de formes qui sont à la fois verbe et nom, nom et adjectif, etc.:

DÉCRIRE gramr composé pour $

nombre de lexèmes: 1422

nombre	pourcent	symbole
449	31.58%	nomc
136	9.56%	(adj,nomc)
133	9.35%	détnum
1278.	93%	nomp
94	6.61%	adj
93	6.54%	v_conj

etc.

On peut faire afficher la liste des noms, des verbes, des formes qui n'ont reçu aucune valeur, et, si on le souhaite, procéder à des ajouts et à des corrections dans le lexique ou en contexte.

Pour les formes nouvelles, celles qui ne figurent pas dans les sources consultées, il faut les catégoriser à la main. On peut le faire en bloc dans certains cas, d'après le suffixe (les mots qui se terminent par les caractères *able* ou *ables* sont en général des adjectifs, à part quelques exceptions comme *table*) ou au cas par cas, selon l'ampleur de la mise à jour nécessaire et la diversité des formes. Le lexique d'une base de données avec les valeurs de catégorie grammaticale peut être sauvegardé sous forme de dictionnaire et être réutilisé au besoin.

Les expressions complexes

L'ajout de valeurs de propriété grammaticale rend possible la recherche dans le texte de toute séquence de valeurs comme:

Nom commun + nom propre:*Île Dupas, lac Saint-Pierre, rue Ste-Catherine*

Nom commun + adjectif ou participe passé: *acide chromique, chaux hydratée*

Nom commun + de ou d'+ nom commun ou nom propre: *cours d'eau, plan d'urbanisme, village de Bernierville*

Nom commun + de ou d' + nom commun + adjectif ou participe passé: *bureau d'audiences publiques, température d'ébullition normale*

Nom commun + adjectif ou participe passé + de ou d' + nom commun ou nom propre:*agents chimiques de coagulation*

etc.

Cette méthode dite de patrons catégoriels s'assortit d'une procédure exécutable (l'équivalent d'une macro-commande dans WordPerfect) Marqterm qui "bloque" au moyen d'un trait de soulignement les différents éléments de l'expression complexe, de telle sorte qu'elle soit désormais considérée comme une simple unité lexicale. Selon le degré d'exhaustivité cherché, divers patrons peuvent être spécifiés qui correspondent aux modes les plus courants de formation des lexies complexes. Le lexique issu de la procédure ressemble à ceci:

fonction_publique

hautes_eaux_printanières_moyennes

métaux_sous_forme_de_carbonate

Office_des_ressources_humaines

personnel_engagé_à_honoraires

On constate donc l'avantage d'un logiciel qui n'élimine pas les "mots-outils" par un anti-dictionnaire, puisque les prépositions jouent un rôle important dans la formation des lexies complexes, et qui permet de formuler des stratégies de recherche incluant autant les chaÎnes de caractères (*de*, *d'*) que leurs valeurs de propriétés (préposition, par exemple.). Il s'agit d'une fouille exhaustive et systématique qui n'oblige pas à repérer et à énumérer auparavant les mots susceptibles d'entrer dans la composition de termes du domaine. Les opérations effectuées sur les propriétés constituent un raccourci pour traiter l'ensemble des unités pertinentes.

Pour les différentes variantes possibles de la procédure et sur ses limites évidentes (en particulier le bruit généré par la non-désambiguïsation en contexte des valeurs multiples), le lecteur pourra se reporter au chapitre 9 de l'ouvrage de Bertrand-Gastaldy (1992).

Ce dispositif aide donc à repérer les fameux multitermes qui sont un véritable casse-tête pour l'établissement des thésaurus. On évalue, en effet, que jusqu'à 80% des concepts d'un domaine sont exprimés par ce que les linguistes appellent des termes complexes (Boulanger,1989:361).

Les termes du domaine

Une fois la liste des expressions complexes épurées de ses scories évidentes et une fois prises les décisions délicates de la décomposition en unités plus petites, il reste à déterminer ce qui correspond à un terme du domaine, c'est-à-dire à un sous-ensemble du lexique faisant appel à la notion d'usage au sein d'une communauté scientifique donnée.

Le choix définitif des termes du domaine, qu'ils soient simples (unitermes) ou complexes (multitermes), demeure l'apanage des experts. Nous pouvons profiter de diverses fonctionnalités de SATO pour offrir à ces experts une liste exhaustive, mais ordonnée selon un ordre probable de pertinence décroissante. La pondération part de différentes propriétés des unités lexicales, fréquentielles , typographiques, textuelles, etc. Les propriétés et les valeurs qui entrent en ligne de compte ne sont pas déterminées d'avance; elles sont laissées à la discrétion des concepteurs du thésaurus, ce qui permet de les ajuster aux particularités des corpus.

Si, dans un corpus, tous les mots importants sont en gras ou en italiques, on peut décider d'accorder un certain poids à cette propriété. En outre, les titres et les sous-titres, les légendes de tableaux, les tables des matières et les index constituent de bons réservoirs de termes et l'on a vu qu'on peut caractériser explicitement ces différents éléments .On pourrait même décider d'accorder un poids supérieur aux unités lexicales qui figurent dans les deux premières et les deux dernières phrases de chaque paragraphe (cette possibilité a été rajoutée à notre demande, dans un logiciel qui peut travailler en amont de SATO, ICON par un des collaborateurs dans deux de nos projets). On peut aussi pondérer non pas à partir de la fréquence, mais de la valeur discriminante de Salton implantée depuis peu dans SATO ou bien en fonction du chi2 qui est disponible depuis longtemps avec la commande Distance.

Voici illustrés les résultats de la procédure qui a consisté à ajouter à la fréquence: + 14 aux termes complexes, + 10 aux noms communs, + 2 aux verbes et + 1 aux adjectifs et à ajouter + 5 si les formes se trouvent dans la macro-structure des textes (titres, sous-titres, etc.):

fréqtot	gramr	division	poids	(lexique)
57	terme	(titre, résumé, intro) (stitre, texte, conclu)	35	métaux_lourds
59	nom	(titre,résumé,intro)	31	effluents
		(stitre,texte,conclu)		
6	terme	(résumé,intro)	30	échange_ionique
		(texte)		
16	terme	(résumé,texte)	30	osmose_inverse
		(conclu)		
11	nom	(résumé,intro)	20	cémentation
		(stitre,texte)		

Ce qu'on peut constater donc, c'est l'éventail et la complémentarité des approches possibles pour mieux cerner le vocabulaire du domaine. Le logiciel offre des "prises de vue" multiples sur le vocabulaire. Un faisceau d'indices est ainsi disponible pour assister le choix.

Une fois la liste éditée, on peut encore recourir à SATO pour enregistrer le jugement de plusieurs experts et procéder à l'élimination des unités non pertinentes sur la base d'un simple calcul. Voici un exemple qui montre l'avis de trois spécialistes:

fréqtot	poids	expert1	expert2	expert3	(lexique)
57	35	oui	oui	oui	métaux_lourds
59	31	oui	oui	oui	effluents
6	30	oui	oui	oui	échange_ionique
16	30	oui	oui	oui	osmose_inverse
...					
1	14	oui	non	non	contrôle_simple
...					
1	14	non	non	non	demande_faible
1	14	non	non	non	demande_forte
...					

Après le choix définitif, il est possible de sauvegarder la liste des termes du domaine sous forme d'un dictionnaire qui pourra être réutilisé par la suite sur tout nouveau texte entré dans la base de données.

Contrôle du vocabulaire

Dans les bases de données en plein texte, on peut entendre par contrôle tout ce qui contribue à favoriser l'univocité du vocabulaire: le regroupement *a posteriori* des variantes flexionnelles autour d'une forme canonique , la mise en relation des différentes graphies d'un même terme, la mise en évidence de la polysémie en autant qu'elle existe dans un domaine de spécialité, au moyen de définitions et de contextes, et finalement le regroupement de synonymes ou quasi-synonymes. Le contrôle est d'autant plus nécessaire que les unités sont plus nombreuses et plus variées.

Les variantes flexionnelles

Une des lacunes de SATO est qu'il ne possède pas d'analyseur syntaxique; il est donc impuissant à résoudre les problèmes d'identité (homographie) que l'on trouve par exemple dans:

lit _____> sorte de meuble (nom commun, masculin, singulier)

*lit*_____> une des formes possibles de *lire* (verbe)

avions _____> appareil de locomotion aérienne (nom commun, masculin, pluriel)

avions _____> une des formes possibles de *avoir* (verbe)

ainsi que les problèmes d'altérité, c'est-à-dire du rattachement de deux chaÎnes distinctes à un même lexème, comme dans:

lits _____|_ lit (nom commun, masculin)

lit _____|

lit _____| _ lire (verbe)

Il existe encore peu de logiciels qui procèdent à une lemmatisation automatique car, pour cela, il faut reconnaÎtre la fonction syntaxique du mot en contexte, lui attribuer une valeur grammaticale unique et finalement renvoyer la forme flexionnelle à la forme canonique. On notera cependant que, dans la phase 2 du projet Delta du ministère des Communications du Québec, il est prévu que le logiciel Termino, mis au point par l'équipe RDLC (Recherche et développement en linguistique computationnelle) du Centre d'ATO pour le dépistage de lexies complexes à l'Office de la langue française, fournisse à SATO les résultats de son analyse syntaxique.

En attendant, il est toujours possible d'effectuer ce renvoi pour toutes les formes qui, hors contexte, n'ont qu'une valeur, c'est-à-dire sont uniquement noms, uniquement adjectifs ou uniquement verbes, mais il s'agit d'une solution "boÎteuse" parce qu'incomplète.

Formes équivalentes: graphies différentes, abréviations, sigles et acronymes

Par rapport aux autres logiciels documentaires, l'aide originale qu'apporte SATO consiste dans le repérage des régularités d'ordre typographique ou linguistique susceptibles de marquer une variante graphique, une abréviation, un sigle ou un acronyme. La plupart du temps, les sigles et acronymes sont inscrits en lettres capitales et sont accompagnés, lors de leur première mention, de la forme pleine équivalente. Une parenthèse, parfois des tirets, distingue les deux. On peut se faire une idée du genre de stratégie de recherche et de réponses fournies d'après cet exemple où l'on cherche toute séquence qui comporterait une parenthèse ouvrante suivie de n'importe quelle chaÎne de caractères écrite en capitales et où l'on demande d'écrire le contexte (pour alléger la présentation, on a supprimé les références):

Concordance stricte \($*édit=cap

Écrire concordance*

...

La matrice du polymère de l'échangeur peut être un co-polymère_de_styrène et de benzène_divinylique (DVB), phénol et formaldehyde ou des polymères_naturels, tels que le charbon_sulfoné.

...

---

le thénoyltrifluoroacétone (TTA) possède une sélectivité très élevée pour le cuivre, à ph très bas;

On constate une fois de plus l'intérêt d'un logiciel qui permet de fouiller non seulement les mots, mais aussi n'importe quel caractère comme un signe de ponctuation.

Synonymie

La stratégie utilisée précédemment peut être adaptée pour repérer des synonymes. Évidemment, les réponses du système vont comporter beaucoup de bruit, mais on sera sûr de l'exhaustivité de la fouille en autant qu'on aura su répertorier les divers marqueurs de synonymie: *et, comme, ou, est désigné(e) par, sont désigné(e)s par*, etc. Il restera à user de ses connaissances du domaine - et souvent du simple bon sens - pour ne conserver que les passages pertinents. L'établissement de relations de synonymie est une opération sémantique qu'un ordinateur ne peut effectuer. On sait que la synonymie dépend beaucoup du contexte et de la possibilité de visionner ces contextes (qui, soit dit en passant, sont paramétrables dans SATO). Voici un exemple de résultats:

Le revêtement de surface est désigné par les termes d'électrodéposition, de galvanotechnique ou de galvanostégie.

Nous verrons plus loin comment enregistrer les relations entre formes Nous allons continuer, pour le moment, à illustrer l'aide que le logiciel peut apporter dans le dépistage de ces relations.

Les définitions

De la même façon, on peut repérer des définitions qui pourront donner lieu à des notes d'application dans le thésaurus. Selon les régularités des textes, on pourra chercher une séquence de deux-points et de guillemets ouvrants, ou bien des expressions comme *définir*, *entendre par*, etc.

...

"emploi excédentaire": un emploi autorisé provisoirement en surplus des effectifs réguliers

---

...

La complexation se définit comme la formation d'un composé_complexe par un agent_complexant_ou_chelatant.

...

Le revêtement_de_surface est désigné par les termes_d'_électrodéposition, de galvanotechnique ou de galvanostégie. On peut le définir comme l'action de déposer, par voie_électrique, un métal ou un alliage sur un autre métal ou des plastiques.

Structuration du vocabulaire

Pour ordonner la multiplicité des unités lexicales recueillies et fournir une image cohérente du domaine, on procède habituellement à plusieurs types de regroupements qui en facilitent la consultation et fournissent des suggestions pour élargir ou rétrécir les stratégies de recherche.

Regroupement par thèmes

Pour identifier les grandes thématiques, il est possible de s'appuyer sur un simple affichage du lexique. Dans la zone des fréquences élevées, là où l'on retrouve habituellement les formes grammaticales courtes, se détachent des formes longues, habituellement des noms qui correspondent à ce que l'on appelle les mots-thèmes. Ensuite, à l'aide de la commande Tamiser, on fait afficher les mots qui co-occurrent le plus souvent avec ces mots-thèmes et l'on obtient une première organisation. Voici les fréquences les plus élevées d'un corpus:

Écrire lexique $ tri fréqtot

alphabet	fréqtot	(lexique)
fr	902	de
fr	695	.
fr	693	,
...		
fr	148	métaux <----
fr	101	ph <----
...		
fr	81	précipitation <----
fr	76	effluents <----
fr	76	élimination <----
fr	73	a
fr	73	qui
fr	70	concentration <----
fr	69	procédé <----
...		
fr	66	cuivre <----
fr	63	lourds <----
...		
fr	59	traitement <----
fr	56	ions <----
...		
fr	47	effluent <----
fr	47	hydroxyde <----
fr	44	forme <----
fr	44	il

fr	42	métal <----

Regroupement par facettes

La subdivision du corpus en plusieurs thèmes ou domaines (domaine juridique, administratif, économique, social, etc.) n'est pas suffisante pour introduire une structure. Il faut souvent chercher les différentes facettes des domaines en question.

Les suffixes peuvent servir à détecter des formes dont les membres partagent une caractéristique commune. Par exemple, les formes qui finissent par *-ité* expriment souvent des qualités, celles en *-ation* correspondent souvent à des actions ou à des processus.

acidité	accumulation
alcalinité	anodisation
étanchéité	augmentation
humidité	automatisation
polarité	caractérisation
réactivité	cémentation
solubilité	centrifugation
stabilité	chloration
toxicité	clarification
...	...

La recherche de concordances de termes qui peuvent désigner explicitement cette caractéristique, par exemple *procédés* ou *processus* , complète utilement la fouille:

---

procédés

#2 *page=doc1/3/10 ... *page=doc1/3/15/6

---

*divis=résumé La première regroupe les procédés courants pour éliminer les métaux_lourds des effluents, tels: la précipitation_chimique_sous_forme_d'_hydroxydes, de carbonates, de sulfures, de phosphates ou de métal_élémentaire; la coprécipitation; la séparation_solide-liquide_par_filtration, flottation, sédimentation ou centrifugation; la coagulation-floculation, et le traitement_biologique.

---

processus

#10 *page=doc1/8/17/6 ... *page=doc1/8/21/2

---

*divis=intro Ces processus sont : la précipitation_chimique, la coagulation-floculation, l'échange-ionique, l'extraction_par_solvant, la cémentation, la complexation, les traitements_électrochimiques ou biologiques, l'évaporation et la séparation_par_membranes.

Regroupement par familles de mots

Pour contrer la dispersion de l'information due à la variété de la nature grammaticale des mots beaucoup plus grande dans un corpus en plein texte que dans une liste de mots clés contrôlés où ne sont conservés en général que les noms, il est commode de procéder à des regroupements autour d'un même radical. Du même coup, on contrôle partiellement certaines conséquences fâcheuses de la synonymie phrastique. Cette opération s'effectue à partir de la troncature à droite, qui peut pallier en même temps l'absence de lemmatisation, et aussi à partir de la lemmatisation à gauche si l'on veut établir des listes de mots de la même famille. Ainsi avec la stratégie |**plant$**, on récupère:

---

planter

---

plants

---

transplant

transplants

transplanter

transplantation

---

replanter

Le rattachement des mots de même famille n'est malheureusement pas aussi automatique qu'il y paraÎt, à cause des modifications du radical (*voir, vue*), de la formation savante ou populaire (*caprin, chèvre*) de l'origine latine ou grecque du radical (*aquatique, hydrique*) et de la différenciation sémantique des dérivés (*receveur, récepteur*).

Voici des passages qui mentionnent, sous différentes formes, la notion de récupération des métaux:

Concordance libre récup$ méta$

---

récupéré

...

La solution peut être réutilisée et le métal éliminé est récupéré.

récupérés

...

La quantité des métaux récupérés est trop petite pour intéresser les principales compagnies qui les utilisent.

---

récupèrent

...

Les procédés de séparation physique, tels la précipitation, la filtration, la floculation et l'évaporation, récupèrent les métaux sous forme d'oxydes, d'hydroxydes ou de sels.

---

récupérer

...

L'objectif de ce rapport est de faire une revue critique de la littérature sur les procédés en usage et des nouvelles technologies qui permettent d'éliminer, de récupérer et de recycler les métaux lourds contenus dans les effluents industriels.

---

récupération

...

La récupération de chacun des métaux n'est pas toujours possible.

Relations hiérarchiques

On sait que les relations hiérarchiques constituent l'ossature du thésaurus. Leur dépistage s'effectue au moyen d'expressions comme: *est un(e), sont, tel(le)(s) le, la, te(le)(s) que* ou bien par une stratégie qui recherche tout nom précèdant un caractère de ponctuation comme la parenthèse ouvrante ou encore les deux-points, caractère lui-même suivi d'un ou plusieurs noms (c'est ce qu'on appelle une tournure cataphorique), avec, le cas échéant, une fermeture de la parenthèse. Voici un exemple de statégie de recherche :

Concordance stricte est un

Concordance stricte est une

Concordance stricte sont :

Concordance stricte tels l'

Concordance stricte tels la

Concordance stricte tels que

Concordance ordonnée $*gramr=nomcommun \($*gramr=nomcommun \)

Concordance ordonnée $*gramr=nomcommun \:$*gramr=nomcommun

avec des extraits de résultats:

...

L'effluent est passé en premier à travers une résine_cationique pour éliminer les cations tels que le Fe, le Cu, le Zn, le Ni et le Cr3+.

...

Les produits_chimiques qui sont utilisés fréquemment pour précipiter les métaux_lourds_sous_forme_d'_hydroxydes sont la chaux, la soude, la soude_caustique et l'oxyde_de_magnésium.

...

Le procédé est d'autant plus intéressant qu'il y a présence de métaux_précieux (or, argent,platine) dans les effluents.

...

L'utilisation de Fes comme co-précipiteur des métaux_lourds (cuivre, cadmium, nickel, chrome et zinc) s'avère avantageuse comparativement aux hydroxides.

...

---

toutes les crucifères cultivées: chou, chou-fleur, brocoli, chou de Bruxelles, chou chinois, rutabaga et radis

---

De plus, un paradigme peut être constitué autour d'un nom suivi de ses différentes déterminations comme ceci:

Congé

Congé à temps plein

Congé à traitement différé

Congé de maladie

Congé de préretraite

Congé de maternité

Congé hebdomadaire

Congé partiel

Congé pour adoption

Congé pour affaires judiciaires

Congé pour événements familiaux

Congé pour responsabilités parentales

Congé sabbatique

Congé sans traitement

On peut alors obtenir des sous-ensembles regroupant les concepts par une de leurs caractéristiques communes. Le choix des caractéristiques se fait de façon empirique, d'après la liste des termes recueillis dans le corpus:

. Selon le motif:

Congé pour affaires judiciaires

Congé pour événements familiaux

Congé de préretraite

Congé de maladie

Congé de maternité

Congé pour adoption

Congé pour responsabilités parentales

. Selon la fréquence:

Congé hebdomadaire

Congé sabbatique

. Selon les conditions de rémunération:

Congé à traitement différé

Congé sans traitement

. Selon la "complétude":

Congé à temps plein

Congé partiel

Des termes spécifiques peuvent provenir d'une combinaison de caractéristiques:

Congé partiel sans traitement

Quant aux relations partitives, qui sont rangées avec les relations hiérarchiques dans les normes de thésaurus, elles sont détectées par des expressions comme *partie de*, *membre de* ... L'article de Serge Houde dans ce même numéro montre l'intérêt de l'exploitation de tournures de ce genre dans les définitions de dictionnaires électroniques.

Recherche de termes associés

Les relations associatives, parce qu'elles sont souvent établies sans grande rigueur, ont déjà été qualifiées de véritable fourre-tout. Elles regroupent des termes qui, dans les énoncés textuels entretiennent des relations lexico-syntaxiques, du type action - objet de l'action - instrument de l'action comme *Communication, Communiqué, Radio*, des relations entre déterminé et déterminant comme *Permis d'absence , Absence*, des relations entre termes de sens voisin comme *Efficacité, Efficience*, des relations entre désignation du concept et propriétés du concept comme *Sol, Humidité*. Il a été suggéré à plusieurs reprises de les remplacer par des relations de co-occurrence non étiquetées, mais établies de façon plus systématique et reflétant mieux le contenu réel de la base de données à interroger.

Dans SATO, la commande Tamiser filtre, pour une forme donnée (*chaux vive* dans l'exemple ci-dessous), les formes utilisées dans le même contexte, la phrase par exemple, et indique la fréquence de co-occurrence:

freqass	**(lexique)**
4	chaux_vive
3	chaux_hydratée
2	équipements_spéciaux
2	nécessite
2	utiliser
1	brûlures
1	cadmium
1	causer

Il est possible de préciser, pour la forme spécifiée, la nature grammaticale des mots associés: les verbes seulement, ou les noms et adjectifs, par exemple. Le texte de Cossette dans cette même livraison de la revue illustre l'utilisation de cette méthode pour établir des classes de mots propres à un domaine.

Enregistrement et gestion des relations entre mots

Les différents types de relations sont considérés comme des propriétés et les mots reliés sont autant de valeurs de propriété. Ainsi peut-on avoir:

. Une propriété Lemme

lemme	**(lexique)**
informatif	informatif
informatif	informatifs
informative	informatives

. Une propriété Équivalent

équivalent	**(lexique)**

dvb	benzène_divinylique
revêtement_de_surface	électrodéposition
revêtement_de_surface	galvanostégie
revêtement_de_surface	galvanotechnique
tta	thénolyltrifluoroacétone

. Une propriété Facette

facette	**(lexique)**
propriété	acidité
propriété	alcalinité
matériaux	cadmium
processus	cémentation
processus	coagulation-floculation
processus	complexation
produits	chaux
produits	chaux_caustique
etc.	

. Une propriété Famille

fréqtot	**famille**	**(lexique)**
1	récupérer	récupérables
52	récupérer	récupération
1	récupérer	récupère
14	récupérer	récupéré
1	récupérer	récupèrent
9	récupérer	récupérer
1	récupérer	récupérés

. Une propriété Hiérarchie

hiérarchie	**(lexique)**
(métaux, métaux_précieux)	argent
(métaux, métaux_lourds)	cadmium
(métaux, métaux_lourds)	chrome
(métaux, métaux_lourds)	cuivre
(métaux)	métaux
(métaux, métaux_lourds)	métaux_lourds
(métaux, métaux_précieux)	métaux-précieux

(métaux, métaux_lourds)	nickel
(métaux, métaux_précieux)	or
(métaux, métaux_précieux)	platine
(métaux, métaux_lourds)	zinc

La mention de deux niveaux hiérarchiques est nécessaire à cause de la difficulté de manipulation des relations dans SATO.

L'ajout d'une propriété **Associés** est possible, mais le nombre de valeurs associées à chaque forme risque d'être encombrant:

. Une propriété Associés

associés	**(lexique)**
...	
maladies	accident_d'_empoisonnement
chaux_vive	brûlures
brûlures	chaux_vive
métaux,métaux_lourds	élimination
accident_d_'empoisonnement	maladie
élimination,récupération	métaux

Bilan sur les relations dans SATO

La recherche d'information peut se faire soit sur chaque forme répertoriée dans le lexique de la base de données soit sur sa propriété, de cette façon: tous les passages de textes contenant des occurrences ayant pour valeur de la propriété générique *métaux précieux*. Avec l'exemple fourni plus haut, on récupèrerait toutes les concordances comportant *argent, or, platine, métaux-précieux*. Si l'on formulait la requête avec *métaux*, on obtiendrait tous les passages comportant un des termes du lexique affiché plus haut: *argent, cadmium,chrome*, etc.

L'ennui, c'est que SATO n'a pas été prévu, au départ, pour gérer ce genre de relations. Il n'assure pas la réciprocité, ce qui expose le concepteur du thésaurus à toutes sortes d'erreurs, d'omissions, de contradictions.

L'apport réel du logiciel réside clairement dans le dépistage des relations, ce que ne font pas en général les autres logiciels. En somme, SATO et les logiciels documentaires sont complémentaires. Le premier assiste les opérations intellectuelles d'appréhension du contenu, de sélection des termes et de leurs relations, les seconds gèrent les résultats de ce travail, contrôlent les doublons, s'assurent de la réciprocité des relations, bref prennent en charge toutes les tâches, très lourdes si elles sont effectuées à la main, de vérification et de contrôle (Rohou, 1987). C'est ce qui a incité les responsables du projet VIXIT au Conseil du Trésor du gouvernement du Québec à coupler SATO et Seconde (de la firme Destin Inc.) pour l'analyse des textes en gestion des ressources humaines à l'aide d'un vocabulaire contrôlé *a posteriori* .

Mise à jour: extraction des mots nouveaux

La mise à jour du vocabulaire est facilitée dans SATO car, au fur et à mesure que des textes sont inclus dans le corpus, le logiciel permet de connaÎtre les nouvelles formes introduites par ces textes et d'appliquer systématiquement les traitements seulement à ces formes. Il suffit, pour cela de faire deux sous-ensembles du corpus: le premier comporte tous les textes entrés avant la mise à jour, le second tous ceux qui ont été rajoutés. Lorsqu'on fait éditer le lexique, les formes qui ont une fréquence zéro dans le corpus le plus ancien correspondent aux formes nouvelles.

sous_ensemble1	sous_ensemble2	(nouvelles formes)
fréq1	fréq2	
0	1	accumulateurs
0	1	actifs
0	1	admissible
0	1	agent_chelatant
0	1	agent_réducteur
0	1	aluminium
0	1	amalgame
0	1	anodique
0	1	anodisation
0	1	bioxyde_de_sodium
0	1	bisulfite_de_sodium
0	1	boue_contaminée
0	1	cadmium_résiduel

Les autres sources pour l'élaboration de thésaurus

Les questions des utilisateurs

Nous n'avons parlé jusqu'ici que de l'exploitation des textes qui constituent la base de données. Cependant, la même méthodologie peut être appliquée aux questions des utilisateurs. Si on choisit le mode témoin au début d'une session de travail avec SATO, toutes les commandes et les réponses sont enregistrées. On peut donc trier et éditer le tout pour enrichir le thésaurus avec les termes - et éventuellement les relations - employés par les utilisateurs.

Les sources terminologiques et lexicographiques

Les banques de terminologie et les dictionnaires sur support lisible par ordinateur peuvent être exploités par SATO de la même façon que les textes. Leur régularité et leur normalisation sont d'ailleurs plus grandes. L'expérience rapportée dans ce même numéro par Serge Houde témoigne de l'intérêt de la démarche. Le fait que, dans la version du Robert-E destinée au MacIntosh, chaque mot-clé de l'article soit directement accessible, accroÎt encore la facilité d'exploitation avec SATO. Un autre essai a été tenté avec la base de terminologie

Termium sur CD-ROM.

Vers des thésaurus personnalisés

On souligne de plus en plus le caractère "privé" des structures cognitives. Un terme peut, en effet, appartenir à plusieurs catégories, selon le point de vue adopté: "[...] different persons, in different occupations may possess different world views and make different demands upon sources od knowledge as a consequence. " (Wilson, 1984: 200).

Contrairement à ce qu'affirme la norme ISO 2788 Organisation internationale de normalisation, 1986), il n'existe pas à proprement parler de catégories *a priori* , mais des catégories imposées par des agents cognitifs dans un domaine donné, les auteurs des textes et reconnues par d'autres agents cognitifs, les lecteurs. Le choix des caractéristiques sur lesquelles s'appuie la catégorisation dépend du contexte d'utilisation (Frohmann, 1983). L'idéal pour un utilisateur est de pouvoir accéder au regroupement le plus parlant en fonction de la tâche à accomplir.

La technologie de l'hypertexte permet d'ores et déjà d'envisager une superposition de plusieurs visions différentes d'un domaine, comme celle d'un groupe d'utilisateurs par rapport à un autre (Agosti *et al.* , 1989) ou bien d'un individu par rapport à un groupe (Belkin *et al*., 1991). Avec SATO, il est possible de catégoriser le vocabulaire de différentes façons et de conserver chacune des "prises de vue" sous forme de dictionnaire.

En guise de conclusion partielle

Lorsqu'il s'agit de construire un thésaurus de qualité, on ne peut pas compter sur la machine uniquement. On est confronté à ce défi qui, d'après Locke (1991: 200) accompagne tout effort de repérage en plein texte: "[...] the knowledge engineering that goes into constructing a first-class thesaurus of relevant concepts." Mais tout comme les cogniticiens chargés d'élaborer des bases de connaissance, les spécialistes de l'information peuvent désormais s'appuyer sur des outils qui assistent l'appréhension du contenu des sources dépouillées pour structurer à la fois le lexique et les concepts d'un domaine, donc qui agissent en amont de l'analyse plutôt qu'en aval comme les logiciels documentaires traditionnels.

Il ne reste pas moins beaucoup de recherches à effectuer sur les bases théoriques des thésaurus, sur leur contenu et l'organisation de leur contenu pour qu'ils puissent répondre adéquatement aux nouveaux besoins d'analyse de textes par ordinateur et de manipulation automatique pour la recherche documentaire:

"[...] it may now be possible to leave to a human specialist's intervention only the more tricky cases, - interventions like disambiguation, word selection, etc., which then would have to be done by interaction.

So, the question being posed today is this: How must a machine-operated thesaurus look like, and can it be built and maintained, if it is to meet the needs of such machine or machine-aided natural language processing - among others: IR." [information retrieval] (Schmitz-Esser, 1990: 130).

AIDE À L'INDEXATION

LE DÉCLIN DE L'INDEXATION HUMAINE

L'indexation humaine est devenue un véritable goulot d'étranglement. Beaucoup trop subjective, beaucoup trop longue, beaucoup trop coûteuse, elle est de plus en plus remplacée par l'indexation automatique rendue possible par la numérisation des données textuelles. Nous ne reprendrons pas ici la démonstration de l'impossibilité d'obtenir des résultats de qualité avec des méthodes aussi crues que l'indexation des chaÎnes de caractères autres que les formes fonctionnelles ou même de méthodes statistiques uniformes pour des types de discours aussi dissemblables que le droit, la littérature, la chimie, etc. et des genres aussi divers que l'article scientifique, la correspondance et le règlement administratif. Des processus cognitifs aussi complexes que la lecture et la condensation pour d'autres individus ne peuvent être formalisés, entre autres parce qu'ils sont encore partiellement inconnus et beaucoup trop complexes. D'autre part, les méthodes d'analyse linguistique des textes en langue naturelle sont encore très rudimentaires, malgré de belles réalisations. Nous avons montré par ailleurs que la meilleure solution résidait sans doute dans l'indexation assistée (Bertrand-Gastaldy, 1990b). De plus en plus de voix s'élèvent pour proposer une complémentarité des traitements linguistiques, statistiques et procéduraux, de type système expert (Chaumier et Dejean, 1992; Doszkocs, 1986; Meunier *et al.* , 1986, entre autres).

L'INDEXATION ASSISTÉE PAR SATO

Actuellement, avec SATO on peut surtout bénéficier d'une indexation qui recourt aux équivalents. C'est au concepteur du système d'indexation de déterminer la spécificité de son indexation en choisissant la granularité de son contrôle de vocabulaire. Il peut considérer comme équivalents seulement les variantes flexionnelles, y ajouter certaines variantes dérivationnelles (pour regrouper par exemple les noms et les verbes d'action), les synonymes, les quasi-synonymes et même les antonymes, selon le taux de rappel souhaité, puisque la précision est assurée, en principe, par les formulations en langage naturel. La profondeur de l'indexation dépend en grande partie de la segmentation préalable des textes: on peut prendre comme unité documentaire le texte dans son entier, chaque chapitre, chaque paragraphe.

SATO va souligner dans les textes (seulement dans les titres et les sous-titres si on désire une indexation superficielle) les formes répertoriées comme termes du domaine ou équivalents, ajouter la valeur de propriété à côté de la forme détectée et assigner en début de texte une forme unique, celle qui aura été déterminée comme le "descripteur" accepté. Voici un exemple de sous-titre ainsi indexé.

---

*index=revêtement_de_surface

---

*divis=stitre 1.2 Électrodéposition*index=revêtement_de_surface

---

---

*divis=texte Le revêtement_de_surface est désigné par les termes d'électrodéposition, de galvanotechnique ou de galvanostégie. On peut le définir comme [...]

Il est possible de se fixer une grille d'indexation et de faire repérer systématiquement dans les textes les descripteurs qui correspondent aux "facettes" souhaitées, par exemple dans les

conventions collectives: mesure, personnel concerné, personnel exclu, durée. Le rôle de l'indexeur se borne à réviser l'indexation en contexte, à éliminer les termes repérés qui ne correspondent pas au sujet, à rajouter des termes pour les concepts implicites ou ceux qui sont exprimés par des pronoms, à remplacer un terme général par un terme spécifique plus adéquat. Sa tâche est facilitée par le soulignement de toutes les occurrences qui ont déclenché l'indexation et par une fonction de "catégorisation" en contexte. Nous avons ainsi construit un petit modèle d'analyse de la correspondance de la division des évaluations environnementales au ministère de l'Environnement qui retient la phase du projet, s'il y a lieu, l'action (l'intervention sur l'environnement), l'objet de l'action, le lieu, les limites spatiales . L'exemple ci-dessous ilustre l'indexation d'une lettre après correction humaine (on voit toutes les annotations surimposées par les opérations d'attribution de valeurs de propriétés; l'indexat - ou ensemble de mots-clés assignés - a été disposé automatiquement à la tête du texte et il est suivi du nom des facettes correspondantes):

---

*page=let2/1/4

Sainte-Foy, le

---

*index=(réaménagement, route_148, Aylmer, chemin _Rivermead_et_ Pont_Champlain)*facet=(action, objet, lieu, limites)

Destinataire 2

Aylmer*index= Aylmer*facet=lieu (Québec)

J9H-3M2

Mxxxxx,

Le projet de réaménagement*index=réaménagement*facet=action de la route_148*index=route_148*facet=objet entre_le_chemin_Rivermead _et_le_Pont_Champlain *index=entre_le_chemin_Rivermead_et_le Pont

---

_Champlain*facet=limites à Aylmer*index= Aylmer*facet=lieu du ministère des

---

transports, est un projet assujetti à la procédure_d'_évaluation_et d'_examen_des_impacts_sur_l'_environnement*index=nil*facet=nil.

[...]

Le ministère de l' Environnement est actuellement en attente de l'étude_ d'_impact*index=nil*facet=nil. La procédure_d'_évaluation_et_

d'_examen_des_impacts_sur_l_'environnement*index=procédure_d'_

évaluation_et_d'_examen_des_impacts_sur_l'_environnement*facet=

phase se poursuivra lorsque l'étude_d'_impact *index=nil*facet=nil sera

déposée au ministère de l'Environnement.

Veuillez agréer, Mxxxxx, l'expression de mes sentiments les meilleurs

Auteur Y

Lors de l'interrogation, selon le taux de rappel et de précision souhaités, on pourra interroger soit sur n'importe quel mot du texte, soit sur les descripteurs seulement (si on veut se restreindre à ce qui est vraiment thématique), soit sur l'ensemble des mots du texte et des descripteurs. On peut aussi demander la recherche sur tous les équivalents du descripteur. Les stratégies d'indexation ou "patrons de fouille" se présenteront ainsi:

1) **concordance libre étude_d'_impact**

[récupère toutes les occurrences, telles quelles]

2) **concordance libre étude$_d'_impact$**

[récupère toutes les occurrences, au singulier et au pluriel]

3) **concordance libre $*équivalent=étude_d'_impact**

[récupère toutes les occurrences équivalentes au descripteur, qu'elles représentent ou non le sujet du document (il n'y a pas eu d'indexation)]

4) **concordance libre $*index=étude_d'_impact**

[récupère toutes les formes équivalentes au descripteur assigné au texte ou au segment de texte, parce qu'il représentait le sujet du document selon l'indexeur qui a révisé les suggestions d'indexation de SATO]

Dans le cas de l'exemple ci-dessus, la lettre sera signalée par la stratégie no 3 (ou les stratégies 1 et 2):

$*équivalent=étude_d'_impact

mais ne le sera pas si l'on précise qu'on ne cherche que les lettres dont le sujet principal porte sur cette phase du projet (stratégie no 4).

Dans un logiciel documentaire comme Seconde, un menu présentera ces options et le logiciel ira chercher dans les champs indexation et/ou texte selon les spécifications de l'utilisateur. Il pourra aussi, si on le demande, aller fouiller dans le thésaurus et afficher les descripteurs environnants. Comme le logiciel a une interface à menu, la plupart des utilisateurs le trouveront plus facile à utiliser; par contre, les experts seront privés de plusieurs fonctionnalités disponibles dans SATO. Dans l'application VIXIT, on recourt donc

au logiciel SATO pour procéder à l'indexation, puis on transfère les résultats dans Seconde pour son exploitation.

Indexation personnalisée

L'indexation est ainsi accélérée et rendue plus uniforme. C'est une façon de procéder qui convient bien dans les contextes les plus familiers aux bibliothécaires: ceux où il faut desservir une clientèle assez nombreuse, qui interroge fréquemment et dont on peut bien cerner les besoins. Mais les textes informatisés dans les bureaux peuvent servir à toutes sortes de fins. Certains individus ont des préoccupations atypiques par rapport à leurs collègues. C'est ainsi que des étudiants de l'EBSI (École de bibliohtéconomie et des sciences de l'information) ont dû proposer, pour une personne chargée de planification et d'études prospectives, un système qui repérait toutes les phrases susceptibles d'exprimer dans tout type de texte (produit à l'intérieur de l'organisation comme à l'extérieur) une notion de futur, de prévision. La solution a consisté à élaborer un dictionnaire pour les différentes formes susceptibles d'exprimer le temps et à mettre au point une série de stratégies dépistant la présence simultanée d'un de ces mots avec les terminaisons du futur ou du conditionnel (Chouinard, 1990; Domecq, 1989). La stratégie visant ce dépistage consiste donc en un mélange d'indexation au repérage et de catégorisation préalable du vocabulaire. Voici un exemple de phrases repérées:

...

Il peut encore en être ainsi demain, dans la mesure où seront domestiquées les énergies solaire, géothermique, nucléaire ...

...

Dans ces conditions, presque tous les systèmes seraient acceptables, si l'on y consacrait des ressources suffisantes.

...

Mais ce que nous indique l'analyse éco-énergétique, et qui échappait à tout autre indicateur, c'est que ce processus ne pourra pas se poursuivre indéfiniment.

On s'approche un tant soit peu de la situation idéale où l'indexation fournirait autant de données, accessibles d'autant de façons que le requerraient les différents problèmes à résoudre au sein de l'organisation. Habituellement, les gestionnaires de systèmes d'information en ont une image monolithique et cultivent une approche centralisatrice. Un bon système de gestion des ressources d'information devrait normaliser les pratiques d'analyse sans nier les différences d'approche nécessaires à l'efficacité: "You must allow for departmental and even individual requirements while still providing the standard basis for indexing." (O'Shea, 1989:18).

Il faudra encore beaucoup d'études sur la façon de lire un texte, d'utiliser l'information, de poser des questions pour découvrir la diversité des approches. En rendant le texte accessible aussi facilement, on fait tomber le carcan dans lequel les systèmes bibliographiques enfermaient les lecteurs pour des raisons évidentes de limites technologiques et économiques. Avec des "boÎtes à outils" du type de SATO, les utilisateurs atypiques peuvent mettre au point des analyseurs "sur mesure" qui se superposent aux analyseurs prêts-à-porter" pour la majorité (Bertrand-Gastaldy, 1990b).

Constitution d'index

Un index peut être constitué qui comporte les références très précises (document, page, ligne et position dans la ligne) et peut être suivi d'un formatage avec un logiciel de traitement de texte:

Concordance libre information_politique

information_politique
1 *page=mediad1/1/9/3
2 *page=mediad1/1/38/4
3 *page=mediad1/6/15/3
4 *page=mediad11/6/38/8
5 *page=mediad2/8/34/2

Autres applications en analyse

D'autres fonctionnalités de SATO peuvent être mises à profit: dans un service d'indexation, il peut être utile de disposer d'un diagnostic de lisibilité des textes pour répartir le travail en fonction de l'expérience ou des habiletés de chacun. Il n'est pas non plus inintéressant de vérifier la lisibilité des résumés rédigés par le service.

Comme certains documents administratifs connaissent de nombreuses versions, le marquage par un souligné des ajouts dans une version par rapport à une autre aide à évaluer rapidement la nécessité de réindexer la nouvelle version.

Nous utilisons actuellement SATO et SPSS pour analyser les résultats des analyses effectuées par des indexeurs afin de détecter les corrélations entre une rubrique de classification et les mots du texte, plus précisément, les mots dans certaines subdivisions de texte. De la sorte, il sera possible de concevoir une série de règles qui déclencheront l'attribution automatique de telle ou telle rubrique sur la base de la co-présence de certaines des unités lexicales discriminantes. En outre, le système expert pourra exploiter des relations thésaurales autres que les relations d'équivalence pour l'indexation. La présence simultanée d'un certain nombre de termes spécifiques pourra conduire à l'assignation du générique. La co-présence de tel ou tel terme dans un contexte déterminé pourra contribuer à désambiguïser un polysème.

LE REPÉRAGE

Nous avons montré comment on peut recourir à SATO pour mener à bien les principales activités documentaires: l'élaboration de thésaurus, l'indexation et, au travers de ces deux

fonctions, nous avons illustré certaines des possibilités d'exploration des corpus qui constituent autant de stratégies de recherche. Le lecteur se sera sans doute rendu compte de la multiplicité des combinaisons possibles. Nous allons récapituler les caractéristiques les plus saillantes.

On peut visualiser les mots dans le texte ou dans le lexique trié par ordre alphabétique, de fréquence, de longueur et d'après les propriétés assignées.

Plusieurs mots peuvent être inclus dans la stratégie de recherche avec des conditions sur leur position respective: dans n'importe quel ordre (concordance libre), dans l'ordre de la déclaration (concordance ordonnée), dans une position d'adjacence (concordance stricte).

La fouille sur les chaÎnes de caractères bénéficie de la possibilité de troncature à droite et à gauche, comme nous l'avons vu plus haut, ainsi que du masque:

La commande:

Écrire lexique pruden-e

peut récupérer:

---

prudence

---

prudente

Si l'on cherche toutes les formes de 3 lettres, on demandera:

Écrire lexique $---

ce qui donnera:

---

fréq

---

7 aux

---

1 bon

1 cas

7 ces

143 des

---

2 100

Avec la commande suivante où:

(introduit une suite de patrons alternatifs,

) termine cette suite

, sépare chacun des patrons alternatifs

Écrire lexique (0,1,3,4,5,6,7,8,9)$

on obtiendra:

1980

1984

1988

223

...

\ sert à inclure dans la stratégie un caractère spécial (comme un signe de ponctuation) ou une majuscule. La commande:

Écrire lexique \M$

extrait:

Manitoba

Matane

Mirabel

...

Lorsqu'on cherche une valeur de propriété symbolique (qui porte un nom), on peut utiliser = (pour égal) et ~ (pour différent). Si l'on veut toutes les formes qui sont des noms, on écrit:

$*gramr=nomc

et on aboutit à ce genre de résultat:

---

fréq gramr

---

1 nomc administrateur

---

2 nomc administrateurs

2 nomc agence

---

1 nomc application

Les opérateurs disponibles pour les propriétés numériques sont:

= pour: égal

< pour: plus petit que

> pour: plus grand que

, pour l'opérateur logique ou

$*fréquence>5

---

fréq

---

7 accès

---

10 acteurs

11 agences

28 analyse

...

Nous avons dit plus haut que l'on peut associer une fouille sur les caractères et une fouille sur les propriétés. Voici un exemple où l'on cherche tous les mots qui ont reçu la valeur infinitif et qui se terminent par ir, er, oir, re:

|(ir,er,oir,re) *gramr=infinitif

---

fréq gramr

---

1 infinitif avoir

---

1 infinitif conclure

1 infinitif contenir

---

2 infinitif penser

Cet autre exemple combine une troncature à gauche sur une chaÎne de caractères, une valeur de propriété grammaticale et une fréquence:

Écrire lexique |ation*gramr=nomc*fréqtot>5

---

fréq gramr

---

13 nomc communication

101 nomc information

10 nomc présentation

6 nomc relation

6 nomc situation

Dans le texte, on peut délimiter le contexte dans lequel s'effectuera la fouille: contexte numérique en fonction d'un nombre fixé par l'utilisateur de mots avant et après le mot cherché; contexte délimité en fonction de la présence dans le texte de caractères délimiteurs (comme les signes de ponctuation forte ou faible), ce qui permet par exemple de chercher dans des phrases, des portions de phrases ou des paragraphes; contexte homogène obtenu en fonction des valeurs d'une propriété textuelle donnée: les textes, les chapitres, les notices, les entrées de dictionnaire, etc. On peut afficher ou imprimer également toute portion de texte qui correspond à une valeur de propriété textuelle, par exemple tous les titres:

Écrire texte $*zone=tm

***notice=1*zone=tm** Analyse du traitement journalistique de l'information politique au Québec: le cas du projet de restructuration scolaire de 1982

***notice=2*zone=tm** La position canadienne face au nouvel ordre mondial de l'information et de la communication

***notice=3*zone=tm** L'information administrative et les moyens de puissance

On peut ensuite extraire les formes lexicales qui proviennent des titres:

fréq zone

7 tm accès

5 tm activités

3 tm administrative

11 tm agences

28 tm analyse

On n'en finirait pas de donner des exemples de combinaisons diverses. Il faut laisser aller son imagination, ce qui n'est pas aussi aisé lorsqu'on est conditionné par les contraintes des logiciels documentaires traditionnels!

La souplesse d'exploration de SATO peut d'ailleurs se transformer en difficulté d'utilisation, car il faut maÎtriser une syntaxe plus riche (tant au niveau des unités que des opérateurs pour les manipuler) que celle qu'offrent d'habitude les logiciels documentaires. C'est pourquoi SATO est bien adapté pour les concepteurs, les chercheurs et fouineurs de tout acabit! Pour les autres, ceux qui ont d'autres occupations, il est possible de dissimuler la plupart des démarches sous des macro-commandes qui rendent le logiciel transparent, ou bien de déverser les résultats d'analyse dans un logiciel documentaire qui offre, par contre, beaucoup moins de diversité d'approches.

AU-DELÀ DES PROCÉDURES D'ANALYSE DU CONTENU: LES QUESTIONS THÉORIQUES

Au-delà des procédures pour faciliter l'analyse du contenu, le dépouillement automatique de textes intégraux soulève des questions théoriques importantes que nous mentionnons brièvement (nous les avons développées davantage dans Bertrand-Gastaldy et Pagola, 1992).

En ce qui concerne le contenu d'un thésaurus, on peut s'interroger, par exemple sur le type d'unités lexicales à conserver, sur l'inclusion des verbes, des adjectifs et des adverbes à côté des noms pour une meilleure caractérisation d'un domaine. On peut aussi se demander comment interrelier les termes propres au domaine, le vocabulaire de la langue commune et les mots sélectionnés de façon privilégiée pour accompagner les termes du domaine (ce que les terminologues appellent les co-occurrents). La question de la finesse des étiquettes de relations en fonction des conditions d'utilisation, notamment dans les systèmes à base d'inférence, est également débattue. Les liens entre thésaurus, terminologies et bases de connaissances font également l'objet de recherches.

Pour ce qui est des sources à exploiter en vue de colliger les termes du thésaurus, on a mentionné les textes de la base de données à interoger, les dictionnaires et banques de terminologie, les questions des utilisateurs. On se doute qu'on ne peut manipuler au hasard les sources de données. Il faut se poser des questions sur leur validité à alimenter un thésaurus et, plus fondamentalement sur la fonction des thésaurus, donc sur le genre de connaissances à y inclure. Les relations qu'ils mettent en évidence sont-elles des relations communément acceptées dans la vie quotidienne, dans le domaine de spécialité, dans la base de données particulière à laquelle on veut faciliter l'accès? Ces relations concernent-elles les concepts qui seraient exprimés par les termes ou les relations entres les différentes unités lexicales, ou encore les deux? Rien de tout cela n'est clarifié dans la littérature sur les thésaurus.

Les phénomènes textuels qu'il faudrait maÎtriser pour arriver à une indexation automatique capable d'extraire avec précision le sujet du document ne sont pas abordés depuis très longtemps dans la littérature en sciences de l'information. Mais on sait mieux désormais comment procèdent les indexeurs humains. Avec toutes leurs connaissances de la langue, du domaine, des conditions de production des textes et du contexte d'utilisation, ils réussissent en principe à reconnaÎtre les équivalences sémantiques entre un terme et sa définition, les périphrases et les paraphrases, à distinguer les nuances fines entre énoncés presque semblables, à détecter les concepts implicites, à suppléer aux ellipses, à rattacher correctement les pronoms à leurs référents, à laisser de côté les informations superflues ou redondantes et à intègrer progressivement les détails présentés dans les micro-propositions pour dégager le sens général, les macro-propositions. Ils filtrent l'information la plus

pertinente pour les utilisateurs dont ils connaissent bien les besoins. Le problème se pose donc de savoir jusqu'où et comment on peut le mieux les assister dans ces tâches complexes, tout en sachant qu'on ne peut guère les remplacer entièrement, sauf dans des cas où les sous-titres fournissent une sorte d'auto-indexation ou encore lorsqu'une indexation de piètre qualité n'est pas trop pénalisante. La solution à privilégier doit s'appuyer sur une analyse coût-bénéfices et dépend de la disponibilité de logiciels évolués.

CONCLUSION

SATO a servi à illustrer le parti que l'on peut tirer d'outils d'aide à l'analyse de contenu pour mener à bien certaines tâches reliées à l'analyse et au repérage.

Le logiciel n'est pas sans défaut. Le vocabulaire des commandes peut rebuter le spécialiste de l'information documentaire. Une grosse amélioration de l'interface serait nécessaire. Il faudrait lui ajouter d'autres modules, en particulier un analyseur morpho-syntaxique. Des systèmes comme ALETH de la firme ERLI, SPIRIT, FASIT ou INDEX-D procèdent à des analyses plus fines des lexies complexes.

Par contre, il offre une panoplie étonnante d'outils destinés à des approches multiples des textes.

De plus, il peut être couplé à des logiciels aussi bien en amont (éditeurs comme PE, manipulateurs de chaÎnes de caractères comme ICON) qu'en aval (logiciel statistique comme SPSS, logiciel documentaire comme Seconde, coquille de système expert, etc.), ce qui va dans le sens d'une conception modulaire des instruments de traitement adaptée à la diversité des objectifs poursuivis. L'un des bénéfices que l'on peut tirer de son utilisation est d'ordre pédagogique. Nous prenons conscience de la complexité de l'objet textuel et du nombre de connaissances déclaratives et procédurales qu'il faut ajouter pour reproduire un tant soit peu certaines étapes de l'analyse humaine.

Avec de tels outils à notre disposition, nous nous rendons bien compte que, à l'avenir ce n'est plus tant la mise en oeuvre de certains traitements qui posera problème mais plutôt les fondements théoriques de nos façons de fonctionner et de nos outils traditionnels. De beaux défis sont à l'horizon!

BIBLIOGRAPHIE

Agosti, M.; Gradenigo, G.; Archi, A.; Inghirami, B.; Nannuci, R.; Colotti, R.; Mattiello, P.; Di Giorgi, R.M.; Ragona, M. New prospects in information retrieval techniques: a hypertext prototype in environmental law. In: *Online Information 89; Proceedings of the 13th International Online Information Meeting*, London, 12-14 December 1989: 483-494.

Association française de normalisation. *Règles d'établissement des thésaurus monolingues*. Z 47-100. Paris: AFNOR; décembre 1981.

Auger, Pierre. Terminographie et lexicographie assistées par ordinateur; état de la situation et prospectives. *Actes du Colloque Les industries de la langue: Perspectives des années*

1990, Montréal, 21-24 novembre 1990. [Québec]: Gouvernement du Québec: 659-680.

Balcer, Madeleine; Gonin, Jean-Paul. Réactions de l'usager face à l'utilisation du système de repérage en mode dialogué, BADADUQ. *Documentaliste*; 16(2); mars-avril 1979: 55-61.

Belkin, N.J.; Marchetti, P.G.; Albrecht, M.; Fusco, L.; Skogvold, S.; Stokke, H,; Troina, G. User interfaces for information systems. *Journal of Information Science* ; 17; 1991: 327-344.

Bertrand-Gastaldy, Suzanne, 1992. *Le contrôle du vocabulaire et l'indexation assistés par ordinateur; une approche méthodologique pour l'utilisation de SATO*. Avec la collaboration de Gracia Pagola. [Montréal]: Université de Montréal. École de bibliothéconomie et des sciences de l'information; janvier 1992. pagination variée [612 p.] [en pré-édition; édition définitive: été 1992].

Bertrand-Gastaldy, Suzanne, 1984. Les thésaurus de recherche; des outils pour l'interrogation en vocabulaire libre. *Argus*; 13(2); 1984: 51-58.

Bertrand-Gastaldy, S., 1990a "L'évolution de la gestion de l'information documentaire sous l'impulsion des nouvelles technologies." *Terminogramme*; Bulletin d'information terminologique et linguistique, 55, mars 1990: 25-31.

Bertrand-Gastaldy, S., 1990b. "L'indexation assistée par ordinateur: un moyen de satisfaire les besoins collectifs et individuels des utilisateurs de bases de données textuelles dans les organisations." *ICO Québec; Intelligence artificielle et sciences cognitives au Québec*; 2(4); septembre 1990: 71-91.

Bertrand-Gastaldy, S.;Pagola, G. 1992. "L'élaboration et la gestion d'un vocabulaire de domaine dans le contexte des bases de données textuelles: remises en question et méthodologies." Colloque Repérage de l'information textuelle organisé conjointement par l'Hydro-Québec et le ministère des Communications du Québec, Montréal, le 18 septembre 1991. [Montréal]: Hydro-Québec; mars 1992: 51-71.

Blair, David, C., 1986. Full text retrieval: evaluation and implications. *International Classification*; 13(1); 1986: 18-23.

Blair, David, C.; Maron, M.E., 1985. An evaluation of retrieval effectiveness for a ful-text document retrieval system. Communications of the ACM; 28(3); march 1985: 289-299

Boulanger, Jean-Claude,1989. Le statut du syntagme dans les dictionnaires généraux monolingues. *Meta*; 34(3); septembre 1989: 360-369.

Bourque, G. ; Duchastel, J.,1988. *"Restons traditionnels et progressifs". Pour une nouvelle analyse du discours politique; le cas du régime Duplessis au Québec*. Montréal: Boréal; 1988. 399 p.

Chaumier, Jacques; Déjean, Martine, 1992. L'indexation assistée par ordinateur: principes et méthodes. Documentaliste; sciences de l'information; 29(1); 1992: 3-6.

Chouinard, Daniel, 1990. *La notion d'avenir en français; une exploration au moyen du logiciel SATO*. [Montréal]: Université de Montréal. École de bibliothéconomie et des sciences de l'information; avril 1990. 33 p

Domecq, Marie-Cécile, 1989. *SATO; exemple d'une application: détermination de la valeur*

"futur" dans des textes administratifs. [Montréal]: Université de Montréal. École de bibliothéconomie et des sciences de l'information; décembre 1989. 75 p.

Doszkocs, Tamas E., 1986. Natural language processing in information retrieval. Journal of the American Society for Information Science.; 37(4); 1986: 191-196.

Duchastel, Jules, 1991. Étude d'un corpus de dossiers de la cour juridique de Winnipeg à l'aide du système d'analyse de textes par ordinateur (SATO). *Actes du Colloque Jornadas Internacionales de Anàlisis de Datos Textuales, Universitat Politècnica de Catalunya, Barcelona (Espagne)*, 1991.

Frohmann, Bernhard P., 1983. An investigation of the semantic bases of some theoretical principles of classification proposed by Austin and the CRG. *Cataloging and Classification Quarterly* ; Fall 1983; 4(1): 11-27.

Garson, L.R.; Love, R.A., 1985. Full text searching of the ACS journals online: use and abouse. In: *Online 1985 Conference Proceedings*. New York, November 4-6, 1985, Weston, Conn.: Online Inc.; 1985: 116-119.

Lancaster, F.W., 1977 Vocabulary control in information retrieval systems. *Advances in Librarianship*; 7; 1977: 1-40.

Laroche, Léo, 1990. Calibrage des textes et lisibilité. *ICO Québec; Intelligence artificielle et sciences cognitives au Québe)*; 2(4); septembre 1990: 114-118.

Locke, Christopher, 1991. The dark side of DIP. *Byte*; 16(4); April 1991: 193-206.

Meunier, J.-G., Bertrand-Gastaldy, S.; Lebel, H., 1987. "A call for enhanced representation of content as a means of improving on-line full-text retrieval." *International Classification*, 14(1), 1987: 2-10.

O'Shea, Michael, 1989. Simply defined: nonsense filing everyone can live with. *Office Equipment & Methods*; 1989 November: 16-18.

Organisation internationale de normalisation, 1986. *Principes directeurs pour l'établissement et le développement de thésaurus monolingues*. ISO 2788 (F). [Genève]; ISO; 1986. 32 p.

Paquin, Louis-Claude; Dupuy, Luc; Rochon, Yves, 1990. Analyse de texte et acquisition des connaissances: aspects méthodologiques. *ICO Québec; Intelligence artificielle et sciences cognitives au Québec*; 2(3); septembre 1990: 95-113,

Ranjard, Sophie, 1991. L'indexation manuelle: une valeur ajoutée. *Archimag*. Hors série; novembre 1991.

Richer, Pierre, 1986. La création automatique d'un thésaurus de recherche. *Argus*; 15(1); 1986: 13-19.

Rohou, Cécile, 1987. La gestion automatisée des thésaurus; étude comparative de logiciels. *Documentaliste*; 24(3); mai-juin 1987: 103-108.

Salton, Gerard, 1971. *The SMART Retrieval System* . Englewood Cliffs, NJ: Prentice-Hall; 1971.

Schmitz-Esser, Winfried, 1990. Thesauri facing new challenges. *International*

Classification; 17(3-4); 1990: 129-132.

Sparck-Jones, Karen, 1971. *Automatic Keyword Classification for Information Retrieval*. London: Archon Books; 1971.

Tenopir, Carol, 1985. Searching Harvard Business Review online; lessons in searching a full text database. *Online*; 9(2); 1985: 71-78.

Wilson, T.D., 1984. The cognitive approach to information seeking behaviour and information use. *Social Science Information Studies*; 4; 1984: 197-204.

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 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 par ordre alphabétique, avec leur fréquence d'utilisation et un indice de familiarité des mots.

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).

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.

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.

Tableau 1 : **Indices de lisibilité** des fascicules d'information

Doc. Long. moy. Long. moy. % de mots Indice de

des mots des phrases de plus Gunning

(N de car) (N de mots) de 9 lettres

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.

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.

III Journées internationales d'analyse statistique des données textuelles
Rome, 11-13 décembre 1995

La catégorisation socio-sémantique

Victor Armony et Jules Duchastel

GRADiP, Dép. de sociologie, Université du Québec à Montréal
C.P. 8888, succ. centre-ville, H3C 3P8 Montréal (Québec), Canada

Notice bibliographique

DUCHASTEL, Jules et Victor ARMONY (1995). La catégorisation socio-sémantique. In Actes des Troisièmes journées internationales d'analyse statistique de données textuelles. Rome: CISU, 1995: 193-200.

Summary : This paper describes some aspects of a socio-semantic categorization which has been applied to a large political discourse database. The authors discuss the idea of coding textual data before or during the process of analysis, referring to both the tradition of American content and qualitative analysis and French linguistic approaches to discourse. An empirical, paradigmatic, localized and sociologically-oriented categorization is proposed, and the example of the word « service(s) » in Canadian trade unions' discourse is presented.

Key words : Textual Data Analysis, Categorization, Political Discourse, Computer-Aided Analysis

---

Plan de l'article

1. Introduction

Cette communication rend compte de certains aspects d'une expérience de *catégorisation socio- sémantique* réalisée sur des discours politiques québécois et canadiens contemporains. Il s'agit d'un ensemble d'allocutions, communiqués et mémoires émanant d'institutions gouvernementales, syndicales, patronales et religieuses depuis le début des années quatre-vingt. Cette base de données textuelles de grande taille (environ un million de mots) a été compilée dans le cadre d'une recherche qui porte sur le discours politique néo-libéral et qui vise à examiner, à l'aide de l'ordinateur, l'articulation entre les nouvelles formes de représentation de la société et les transformations observables sur le plan de la régulation des rapports sociaux [1].

Nous nous attarderons d'abord sur quelques considérations autour de l'analyse sociologique du discours et le traitement des données textuelles : pourquoi et comment superposer aux mots d'un corpus un système de catégories fondé sur leur signification en contexte d'occurrence? Puis, nous exposerons brièvement les principes et les procédures de catégorisation socio-sémantique mis à l'oeuvre dans le cadre de nos travaux. Nous présenterons enfin un exemple concret d'analyse lexicale : le cas du terme « service(s) » dans le discours syndical. Cet exemple permet d'illustrer quelques-uns des avantages d'une catégorisation *paradigmatique*, *localisée* et *orientée par un découpage sociologique des référents du discours*.

2. L'analyse sociologique du discours et le traitement des données textuelles

Les données textuelles que le sociologue traite à l'aide de l'ordinateur constituent la représentation informatisée d'un ensemble de matériaux langagiers produits par des individus ou des institutions, lesquels matériaux servent de voie d'accès à un discours jugé significatif du point de vue théorique (Duchastel, 1995). Le dépouillement assisté par l'ordinateur présente l'avantage d'assurer – jusqu'à un certain point – la validité et la reproductibilité de plusieurs étapes de la recherche dans un domaine qui est extrêmement sensible aux effets de subjectivité (Duchastel & Armony, 1993). La standardisation des procédures et la réduction du volume de l'information sont en ce sens les deux axes centraux d'une démarche systématique et la catégorisation constitue à cet égard un outil particulièrement précieux. Elle permet d'établir un lien entre les données « brutes » et le cadre d'interprétation, sous forme d'interface à géométrie variable entre chacune des unités du discours et les principes d'organisation de la connaissance. La catégorisation a une valeur heuristique et expérimentale car elle facilite autant l'application de protocoles d'exploration ouverte que la réalisation de fouilles permettant le test d'hypothèses.

De manière générale, nous définissons la catégorisation des données textuelles comme l'ensemble des procédures visant à superposer aux unités d'enregistrement une ou plusieurs grilles de codage à valeur descriptive et analytique. La catégorisation sert à caractériser les

éléments du corpus en leur attribuant de l'information de type extra ou péri-textuel (renseignements sur le locuteur, les circonstances de l'énonciation, etc.) et/ou en les classifiant selon des principes d'homogénéité (fonctionnelle, sémiotique, topique, etc.). Chaque unité du corpus reçoit alors des « étiquettes » qui la spécifient vis-à-vis d'un certain nombre de règles taxinomiques établies par l'analyste. Dans le cas particulier de la catégorisation socio-sémantique, telle que nous la concevons, on vise à classer – de manière exhaustive et exclusive – les mots à valence référentielle (noms et adjectifs) en fonction d'un système de catégories thématiques.

La construction de la grille de catégories suit une logique « constructiviste », c'est-à-dire qu'il s'agit d'une démarche empirique et itérative à visée interprétative, dont l'application se fait au moyen d'une lecture contextualisée : chaque occurrence est codée eu égard à sa signification dans la phrase. Cette perspective privilégie donc l'aspect *paradigmatique* mais *localisé* des unités du discours : le mot « droit », par exemple, ne sera catégorisé comme « domaine juridique » que si le sens de l'énoncé le justifie, car outre l'idée de « ce qui est conforme à une règle », il peut aussi signifier « redevance » (domaine économique) ou encore être utilisé dans une locution adverbiale comme « à bon droit ». Le diagramme suivant illustre cette logique : un même mot peut appartenir à deux catégories différentes (cas I et III), deux mots différents peuvent appartenir à une même catégorie (cas II) [2].

Diagramme

Notre grille comporte plus d'une centaine de catégories différentes, regroupées selon des critères de découpage sociologique du « monde » : acteurs et institutions, sphères d'activité, espaces sociaux, notions axiologiques, etc. Ainsi, les mots catégorisés peuvent fonctionner comme des *indicateurs socio-sémantiques* : ils renvoient, en fonction de leur sens (paradigmatique) et de leur usage (syntagmatique), à divers référents de la réalité sociale. Cette perspective s'inspire en partie de la tradition de l'analyse de contenu mais se rapproche également d'autres manières d'aborder la question du langage. Nous essayerons de la situer par rapport aux principaux courants d'analyse de textes.

Dans le contexte français, l'analyse lexicométrique ou statistique textuelle, qui vise à « traiter les mots comme des nombres » (Baudelot, 1994 : v), ainsi que, de manière plus générale, les diverses approches que l'on regroupe sous la dénomination *analyse du discours* – concernées surtout par les « problématiques de l'énonciation et de la pragmatique » (Maingueneau, 1987 : 14) –, partagent un intérêt pour *la forme de ce qui est dit ou écrit*, c'est-à-dire la facture du texte, la disposition et la distribution des unités de signification. Comme le pose Pierre Achard (1986 : 44), s'il y a quelque chose de commun dans le courant discursiviste « c'est, positivement, la prise au sérieux de la composante linguistique […] Négativement, c'est le rejet des notions de 'contenu' et du modèle de la communication ». Il n'est donc pas surprenant de constater que l'idée de *codage* ne suscite que très peu d'enthousiasme parmi les chercheurs français, alors qu'elle est centrale autant dans la tradition de l'analyse de contenu américaine que dans les écoles « qualitativistes » du monde anglo-saxon.

On sait que la catégorisation est une dimension-clé de l'analyse de contenu. Comme l'a dit Bernard Berelson (1952) : « *content analysis rises or falls by its content categories* ». C'est au moyen des catégories de contenu que l'information véhiculée par un message est réduite et uniformisée dans le but d'en produire, selon la célèbre formule, une « description objective, systématique et quantitative » [3]. Or, l'analyse de contenu est fortement associée

à l'utilisation de « dictionnaires généraux » (par exemple, le Laswell Value Dictionary, le Harvard Psychosocial Dictionary). La stratégie des dictionnaires généraux se caractérise par l'utilisation d'un nombre limité de catégories (environ 60 à 150), la discrimination des homographes à partir de normes de désambiguisation et le traitement des locutions, ainsi que par le fait que la plupart des mots du texte sont codés, que chaque catégorie comporte un nom (*tag*) et une définition de ses règles d'application et que les mots ambigus peuvent être exclus de la catégorisation (la catégorisation multiple étant déconseillée dans ce type d'approche) (Weber, 1984). On privilégie donc dans cette stratégie les schèmes de codage *a priori* plutôt que *a posteriori* (Wood, 1980).

L'analyse de contenu se veut une analyse quantitative du langage, ou plus précisément, *une quantification des données qualitatives* (Roberts & Popping, 1993). Or, depuis déjà une trentaine d'années, un courant se développe au sein de la sociologie et de l'anthropologie qui se penche lui aussi sur les données « non-numériques » mais avec une approche justement « qualitative » (fondée en grande partie sur la *grounded theory*). Cette analyse proprement qualitative vise fondamentalement à décrire et à comprendre la culture et le comportement des individus et de leurs groupes du point de vue de ceux qui sont l'objet d'étude (Bryman, 1988). Les matériaux exploités sont souvent des entrevues ou des notes de terrain ; le chercheur tente de capturer la complexité des phénomènes sociaux en faisant émerger du texte lui-même les concepts qui structureront sa théorie (Strauss, 1987). Naturellement, des ressources informatiques sont souvent mises à contribution pour gérer les masses de données que ce genre d'approche génère. Les logiciels les plus répandus dans ce domaine sont ceux de codage-repérage (*code-and-retrieve programs*). Ils permettent de diviser le texte en séquences de mots et de leur attacher des codes pour pouvoir par la suite afficher toutes les parties qui ont reçu le même code ou combinaison de codes ; certains logiciels de ce type facilitent aussi la formulation de relations entre les catégories de façon à développer des classifications conceptuelles de grande complexité (Weitzman & Miles, 1995).

Bref, l'analyse qualitative, opposée radicalement à l'analyse de contenu en ce qui concerne la quantification/réduction de l'information, partage avec celle-ci une visée classificatoire des unités de signification *à l'entrée ou durant* le traitement. En revanche, le design même des logiciels les plus employés en France révèle le souci de conserver la forme originale du texte : si l'on prend comme échantillon ceux mentionnés par Lebart & Salem (1994), il est clair que la classification des unités sémantiques (mots ou énoncés) est plutôt vue comme le *résultat* des procédures analytiques à caractère statistique. Nous avons cependant constaté que, dans le cadre d'une étude discursive à portée sociologique, il devient utile, voire nécessaire de procéder à un classement préalable des éléments du texte en fonction d'une représentation « sociologique » de la réalité. Par contre, à la différence des analyses du contenu conventionnelles qui produisent un codage hors contexte et a priori par projection de dictionnaires généraux, nous préférons nous donner comme unité d'enregistrement l'occurrence lexicale dans le discours. La catégorisation que nous proposons se rapproche enfin des méthodes qualitatives au plan du travail par « couches » – lectures successives, non linéaires du matériel et formulation d'un système flexible de codes à plusieurs niveaux d'abstraction –, mais encore une fois nous nous distançons dès lors que nous choisissons une démarche axée sur la sémantique lexicale plutôt qu'une catégorisation thématique de segments textuels.

Notre grille de catégorisation est avant tout une classification empirique (mais conceptuellement fondée) des différents référents du discours politique. Son application aux

items lexicaux n'a pourtant pas l'effet de faire disparaÎtre le mot sous la catégorie. Le système informatique utilisé – *SATO : Système d'analyse de textes par ordinateur* [4] –, permet d'apposer plusieurs catégories appartenant à des systèmes différents, tout en autorisant l'accès au mot lui-même, indépendamment des catégories qui lui sont attachées. Nous pouvons alors observer des régularités – quantitatives ou non – de comportement entre catégories et familles de catégories et d'ordonner des fouilles qui conduisent, dans un cheminement heuristique, à l'identification de certains phénomènes. Cependant, comme les équivalents ne sont pas nécessairement des synonymes et peuvent simplement comporter des traits communs, les régularités observées sur la base de cette catégorisation doivent être validées. Comme nous le verrons, la réversibilité de notre système permet de revoir en permanence le contenu de ces catégories et de valider aussi les résultats obtenus à partir de celles-ci.

3. Principes et procédures de la catégorisation socio-sémantique

Nous avons défini la catégorisation socio-sémantique comme un ensemble de procédures visant à appliquer aux unités lexicales une grille de codage à valeur descriptive et analytique d'un point de vue sociologique. La catégorisation du corpus est jugée névralgique dans l'approche que nous adoptons, car l'objectif est de faire ressortir, au sein de grands ensembles textuels, des régularités et des ruptures dans les divers axes et niveaux de structuration du discours politique (références à des valeurs, désignations des collectifs sociaux, thématisation d'enjeux, etc.). Dans le cadre de cette recherche, nous effectuons une catégorisation « en contexte » : chaque occurrence est soumise à une décision. Le codeur doit établir d'abord la pertinence de retenir le terme (a-t-il une signification « forte » et « précise », par rapport à notre grille?) et, le cas échéant, lui affecter une « étiquette » informatique.

Une catégorisation morpho-syntaxique préalable, inspirée de la grammaire de base du français, vise à déterminer si le mot est un nom, un verbe, un adjectif, une préposition, etc. Cette catégorisation est nécessaire pour déterminer les candidats à la catégorisation socio-sémantique car nous n'avons retenu à cette fin que les noms et les adjectifs. Les formes fonctionnelles ont été exclues en raison de leur faible potentiel sémantique et les verbes ignorés parce qu'ils appartiennent à une sémantique particulière qui nous éloigne de notre visée interprétative.

La catégorisation est effectuée sur l'ensemble du corpus par une équipe de codeurs sous la supervision constante d'un coordonnateur. Même si un certain nombre de mots sont catégorisés par projection de dictionnaires, la plupart des occurrences fait l'objet d'un traitement individuel avec visionnement du contexte. Les codeurs sont appelés à choisir parmi les différentes appartenances socio-sémantiques possibles d'un mot, celle qui est la plus proche de la signification en contexte de ce mot. Cela présuppose une connaissance des implications théoriques du système de catégories, mais demande avant tout de rester le plus collé sur la réalité empirique du mot en contexte, indépendamment de toute inférence analytique.

L'application de la grille se fait selon quatre principes fondamentaux : (a) la catégorisation

est exhaustive : tous les noms et adjectifs du corpus font l'objet d'une décision de catégorisation ; (b) les catégories sont exclusives : une occurrence ne peut recevoir qu'une seule catégorie, celle qui correspond à sa signification « prédominante » ; (c) la catégorisation est centrée sur la fonction référentielle des mots : deux termes qui ont le même référent reçoivent la même catégorie, indépendamment de leur « connotation » particulière ; (d) la catégorisation tient compte du contexte d'emploi des mots : deux occurrences d'une même forme lexicale peuvent avoir deux référents différents et reçoivent alors deux catégories différentes.

Nous envisageons la catégorisation comme un processus itératif : au fur et à mesure qu'il se développe, une dynamique d'aller-retour fait en sorte qu'il soit possible de : (1) détecter des régularités dans les décisions qui n'étaient pas prévues (ou « conscientes ») ; (2) détecter des inconsistances dans l'application de la grille. On peut donc dire qu'il s'agit d'un double processus d'apprentissage (sur la base de l'accumulation de décisions correctes) et de correction d'erreurs (sur la base de l'identification des décisions incorrectes). Deux documents d'appui à la catégorisation ont été créés à cet égard. Le premier regroupe, pour chaque catégorie de la grille, l'ensemble de termes du corpus qui l'ont reçue. On parle alors de « l'éventail lexical » des catégories : cette information sert à compléter la définition de chaque catégorie et permet de vérifier sa consistance interne. Le second document est l'envers du premier : il est l'index alphabétique de toutes les formes avec mention des catégories qui leur ont été affectées dans les diverses sections du corpus. Il est alors possible d'observer les différents « usages » d'un même terme. Ces documents sont mis à jour régulièrement (chaque fois que de nouveaux textes sont catégorisés) et servent à expliciter et à formaliser les critères de catégorisation ainsi qu'à effectuer un contrôle périodique de sa fiabilité (stabilité, reproductibilité et précision).

4. L'analyse des données lexicales catégorisées

Nous présenterons maintenant un exemple tiré d'une étude effectuée sur le discours de plusieurs centrales syndicales entre 1980 et 1992. Le corpus a été constitué à partir d'un échantillonnage des allocutions présidentielles aux congrès annuels ou bisannuels. Il regroupe 35 unités discursives émanant de 5 centrales syndicales différentes, pour un total de quelque 250,000 mots. Aux fins de cette communication, nous nous concentrerons sur le cas du mot « service(s) », un terme présent de manière régulière autant sur l'axe diachronique (différentes périodes) que synchronique (différents locuteurs) [5].

Comme la plupart des mots très récurrents (fréquents et répartis dans le corpus), le mot « service(s) » n'a pas une signification précise, ni constante. Il s'agit en effet d'un vocable non seulement *polysémique*, mais aussi *polyvalent* en ce qu'il désigne plusieurs champs différents de la vie sociale. Polysémique, car il peut équivaloir, selon le dictionnaire, à « fonction », « bienfait », « organisme », etc. Polyvalent parce que, tout en désignant de manière générale une « obligation et action de servir », ce mot renvoie à diverses modalités d'interaction entre des acteurs sociaux.

En fait, nous avons observé empiriquement dans le discours trois « aires » principales d'usage du mot « service(s) ». Il y a premièrement la référence globale à l'univers de

« l'utilité commune », c'est à dire de la prise en charge par l'État des questions sociales (les services sociaux) et des entreprises d'intérêt général (les services publics). Puis, on trouve la référence à une sphère particulière de l'activité économique, celle du tertiaire (le secteur des services). Enfin, le mot « service(s) » est employé pour désigner les avantages dont bénéficient ceux qui appartiennent à une association (les services fournis aux membres). Nous avons alors catégorisé toutes les occurrences (sauf quelques cas résiduels, comme dans l'expression « rendre service à quelqu'un ») en fonction de trois codes : SOCIAL, ECONOMIQUE et INSTITUTIONNEL, évoquant ainsi les domaines respectivement concernés.

Il est essentiel de comprendre que nous ne prétendons nullement que ces différents usages correspondent à des « acceptions » du vocable en question (au sens d'une sémantique lexicale). Ils correspondent plutôt à des aires discursives que nous identifions à partir de notre approche. Ce découpage vise donc à mieux circonscrire les « domaines de la réalité sociale » posés par le discours.

Voici des exemples de phrases où les catégories ont été appliquées.

Domaine social	Il faut continuer d'exiger la socialisation de l'ensemble des coûts et la gestion collective publique des **services** de santé et des **services** sociaux. (Centrale de l'Enseignement du Québec, 1988)
Domaine économique	Des emplois bien rémunérés du secteur primaire et des industries de la fabrication ont été remplacés par des emplois moins lucratifs dans le commerce et le secteur des **services**. (Centrale des Syndicats Démocratiques, 1986)
Domaine institutionnel	Nous serons en mesure de mettre nos ressources en commun, ce qui nous rendra plus efficaces et nous permettra d'améliorer encore le **service** que nous donnons à nos membres. (Fédération des Travailleurs du Québec, 1989)

Nous avons produit les lexiques de cooccurrence des trois usages du mot « service(s) » afin de pouvoir observer sommairement leurs covoisinages respectifs (tableau 1) [6].

Tableau 1

On constate que la catégorisation a effectivement donné lieu à un découpage sociologique intéressant. Outre les cooccurrents attendus (à cause de leur proximité thématique mais aussi, soulignons-le, en tant qu'effets de la catégorisation elle-même), comme « sociaux », « secteur » et « membres », on voit ressortir trois lexiques différents, chacun ayant une cohérence interne assez évidente. Notons, par exemple, certains termes qui renvoient au contexte actuel de rigueur budgétaire : les *coupures* dans les services publics (domaine social), la *précarité* dans le secteur des services (domaine économique), les *coûts* des services aux travailleurs (domaine institutionnel). A partir des lexiques obtenus, nous pouvons revenir, par le biais de concordances, aux contextes syntagmatiques et ainsi voir comment les centrales syndicales rappellent « les responsabilités de l'État en matière de

services sociaux, de santé et d'éducation » (domaine social), dénoncent « la dégradation de la durée et de la qualité des produits et des services » (domaine économique) et s'affairent à « donner [à nos] services une efficacité beaucoup plus grande » (domaine institutionnel).

Disons pour finir qu'il est important de remarquer que l'*output* de ce type de procédure est, dans une certaine mesure, tributaire des décisions (pré-)analytiques prises au moment de la catégorisation. Nous voulons signaler ici que les résultats obtenus montrent en même temps :
(a) la validité de la catégorisation, c'est à dire le fait que nous avons bien classifié les usages du mot « service(s) » ; ceci est très important pour une entreprise comme la nôtre qui vise à superposer aux données textuelles brutes un système de repères servant à réaliser d'autres fouilles lexicométriques mais aussi hyper-textuelles ;
(b) la possibilité d'identifier certains traits des « représentations » de divers domaines de l'activité sociale ; la catégorisation des usages du mot « service(s) », fondée sur une « observation sociologique » – objective mais non pas « neutre » – a permis de circonscrire trois espaces lexicaux différenciés.

5. Conclusion

Nous avons essayé de montrer dans cette communication l'utilité d'une catégorisation socio-sémantique quand on entreprend l'étude d'un corpus à pertinence sociologique. Nous avons indiqué ailleurs la valeur heuristique d'une analyse purement lexicométrique réalisée sur des données textuelles « brutes » (Armony & Duchastel, 1995) ; ce type d'approche, bien que très fructueux à l'étape exploratoire et apte à produire des descriptions quantitatives tout à fait intéressantes, reste trop sommaire lorsqu'on vise à générer des fouilles ciblées sur des référents précis du discours social. Nous avons vu qu'une même forme lexicale peut être l'indicateur socio-sémantique de différents objets de l'univers politique et qu'il est possible d'en tenir compte au moyen de codes attribués en contexte d'occurrence. La catégorisation que nous proposons permet également de calculer la cooccurrence globale de, par exemple, l'ensemble de toutes les notions qui renvoient au domaine des pratiques juridiques (le mot « droit » lorsqu'utilisé dans son sens de « prérogative » plus le mot « justice » au sens de « légalité », etc.) et, de cette manière, d'articuler l'analyse du discours au cadre interprétatif du chercheur. Bref, nous croyons que ce type de démarche s'avère essentiel si l'on vise à décortiquer, en sociologues, la parole des acteurs afin d'y trouver la façon dont ils conçoivent leur monde.

RÉFÉRENCES

Achard, Pierre (1986). Analyse du discours et sociologie du langage, Langage et société, no 37, pp. 5-60.

Armony, V. & Jules Duchastel (1995). Some computer-aided heuristic procedures for political discourse analysis. American Sociological Association Annual Meeting, Washington D.C.

Baudelot, Ch. (1994). Préface. In Lebart, L. & Salem, A. Statistique textuelle. Paris : Dunod,

pp. v-vi.

Berelson, B. (1952). Content Analysis in Communication Research. New York, Illinois University Press.

Bryman, A. (1988). Quantity and Quality in Social Research. London: Unwin Hyman.

Duchastel, J. (1995). Texte, discours et idéologies, Revue Belge de Philologie et d'Histoire, vol. 73, no 3.

Duchastel, J. & Armony, V. (1993). Un protocole de description de discours politiques, in Actes des Secondes journées internationales d'analyse statistique de données textuelles. Paris : Télécom, pp. 159-183.

Lebart, L. & Salem, A. (1994). Statistique textuelle. Paris : Dunod.

Maingueneau, D. (1987). Nouvelles tendances en analyse du discours. Paris : Hachette.

Matalon, B. (1988). Décrire, expliquer, prévoir : démarches expérimentales et terrain. Paris : A. Colin.

Roberts, C. & Popping, R. (1993). Computer-supported Content Analysis : Some Recent Developments, Social Science Computer Review, vol. 11, no 3, pp. 283-291.

Strauss, A. L. (1987). Qualitative Analysis for Social Science. Cambridge: Cambridge University Press.

Weber, R. Ph. (1985). Basic Content Analysis. Beverly Hills : Sage.

Weber, R. Ph. (1984). Computer-Aided Content Analysis : A Short Primer, Qualitative Sociology, vol. 7, no 1/2, pp. 127-147.

Weitzman, E. A. & Miles, M. B. (1995). Computer Programs for Qualitative Data Analysis : A Software Sourcebook. Thousand Oaks: Sage.

Wood, M. (1980). Alternatives and Options in Computer Content Analysis, Social Science Research, vol. 9, no 3, pp. 273-286.

NOTES

1. « Le discours politique néo-libéral et les transformations actuelles de l'État (Québec, Canada, 1980-1990) », projet dirigé par Gilles Bourque et Jules Duchastel et subventionné par le Conseil de recherches en sciences humaines (CRSH) du Canada. [Retour au texte]

2. Nous employons ici des catégories générales. Notre grille permet d'opérer une classification beaucoup plus nuancée de ces termes. [Retour au texte]

3. De là que la *fiabilité* de la catégorisation soit un problème névralgique dans toute démarche de ce genre. D'un point de vue conceptuel, la catégorisation consiste à regrouper des objets selon un ou plusieurs critères, en acceptant de négliger toutes les autres différences (Matalon, 1988). Il est alors évident qu'il faut optimiser la qualité du travail de codage, autant sur le plan de la définition des principes d'équivalence et de distinction (la

construction de la grille ou du dictionnaire) que sur celui de leur application empirique (l'adéquation des catégories attribuées aux unités d'enregistrement). Selon Robert Philip Weber (1985), les trois types de fiabilité de la catégorisation sont : (1) la *stabilité* (les mêmes catégories aux mêmes unités), (2) la *reproductibilité* (cohérence entre les décisions des différents codeurs) et (3) la *précision* (par rapport à un standard). [Retour au texte]

4. Ce logiciel a été développé par François Daoust, Centre ATO, Université du Québec à Montréal. [Retour au texte]

5. Il s'agit d'un exemple de type I (voir diagramme ci-haut). Signalons, avant de continuer, que les « analyses » qui suivent n'ont pour but que d'illustrer schématiquement (avec des catégories simplifiées) la démarche d'investigation que nous proposons. Une véritable étude doit bien évidemment se fonder sur le traitement extensif d'un ensemble de notions-clés, orienté par des protocoles exploratoires et des hypothèses de travail. [Retour au texte]

6. Nous avons appliqué un algorithme développé par Guy Cucumel, professeur à l'Université du Québec à Montréal. Dans le tableau, on indique la fréquence de cooccurrence (Fc) et la probabilité de l'association (P). [Retour au texte]

- Commentaires: visib@corpus.ato.uqam.ca -