DUCHASTEL, J.; ARMONY, V. (1991). " Étude d'un corpus de dossiers de la cour juvénile de Winnipeg à l'aide du système d'analyse de textes par ordinateur (SATO) ". In *Actes du colloque 'Journées internationales d'analyse statistique de données textuelles'*. Barcelone: Universitat Politècnica de Catalunya, 1991: 89-108.

ÉTUDE D'UN CORPUS DE DOSSIERS DE LA COUR JUVÉNILE DE WINNIPEG À L'AIDE DU *SYSTÈME D'ANALYSE DE TEXTES PAR ORDINATEUR* (SATO)§

Jules Duchastel, professeur

Victor Armony, étudiant au doctorat

Centre d'analyse de textes par ordinateur

Université du Québec à Montréal

Montréal, Canada

Cet article expose divers aspects de l'application du Système d'analyse de textes par ordinateur (SATO) dans le cadre d'une recherche portant sur le rôle des professions extra-légales dans le traitement judiciaire des personnes mineures. La recherche a été menée autour du cas de la Cour juvénile de Winnipeg (Canada) pendant la période 1930-1960, le corpus empirique étant constitué par un ensemble de 261 dossiers sociaux et psychiatriques. Au moyen de SATO, on a appliqué des procédures telles que le triage de lexiques, le blocage de locutions, la catégorisation, la segmentation, le calcul de distances statistiques et le repérage de concordances. L'article vise à montrer, par le biais de quelques exemples, l'utilité d'un logiciel de ce genre dans une démarche de lecture sociologique de données textuelles.

Mots-clés : SOCIOLOGIE, ANALYSE-DE-TEXTES-PAR-ORDINATEUR, SATO, CONTENU, JUSTICE, JUVÉNILE, CANADA

1. Le Système d'Analyse de Textes par Ordinateur (SATO)

SATO est un environnement informatique conçu pour assister des utilisateurs non-informaticiens dans l'analyse de données textuelles. Il a été développé par François Daoust du Centre d'Analyse de Textes par Ordinateur de l'Université du Québec à Montréal. SATO se veut une sorte de "boÎte à outils" donnant accès immédiat au texte lui-même et à son vocabulaire et permettant au lecteur d'effectuer, à l'aide de l'ordinateur, des tâches d'annotation, de lecture intensive et de comparaison. Le logiciel fournit un ensemble de procédures servant à la manipulation et à la consultation de l'information

---

§ A paraÎtre dans les Actes du colloque *Jornadas Internacionales de AnÁlisis de Datos Textuales*, Universitat Politècnica de Catalunya, Barcelona (Espagne), 1991.

écrite contenue dans de larges bases de données textuelles en vue de son analyse thématique, stylistique, de contenu ou de discours.

SATO peut exécuter des opérations complexes de classification, de segmentation, de fouille et de dénombrement sur la base de décisions heuristiques prises par l'utilisateur, soit à l'étape de la modélisation, soit dans le processus même de l'analyse des données. Il est important de comprendre que SATO, tout en conservant toujours une trace du texte original, permet au chercheur d'effectuer, en interactivité, des opérations sur celui-ci qui peuvent, en tout temps, être reprises et corrigées. Le chercheur dispose d'une machine permettant de contrôler, pas à pas, les diverses étapes de traitement et ainsi de pouvoir intervenir sur le cours de l'analyse.

Tout texte édité et enregistré sur support informatique en format ASCII est susceptible d'être soumis à SATO. La démarche globale de traitement comporte deux phases distinctes : premièrement, la "génération" des fichiers SATO et, deuxièment, leur "interrogation". Le texte original doit cependant être préalablement édité selon certains critères minimaux : remplacement des caractères réservés à SATO (*, \), ajout du caractère \ pour conserver certaines majuscules (notamment dans le cas des noms propres), définition des alphabets de travail (français, anglais, espagnol, etc.) et identification (facultative) des pages, documents et partitions du corpus. Le fichier ainsi préparé peut être traité par le module SATOGEN (SATO-GÉNération). Celui-ci lit automatiquement le texte et en reconnaÎt les composantes principales : mots, ponctuations, paragraphes, références de pagination. Au terme de ce traitement, SATO a construit, à partir des mots du texte, l'index de toutes les formes lexicales du texte - le lexique du texte -, c'est-à-dire le catalogue de l'ensemble des formes contenues dans le texte et de leur référence exacte.

Lors de l'interrogation du texte, au moyen du module SATOINT (SATO-INTerrogation), l'utilisateur travaille sur une représentation du texte actualisant deux plans : l'axe syntagmatique (les items lexicaux en contexte, soit le texte lui-même) et l'axe paradigmatique (les items lexicaux hors contexte, soit le lexique) (Daoust 1990 : 57). Ce module permet le repérage sélectif de chaÎnes lexicales sur ces deux axes. Le chercheur pourra donc étudier les mots ou groupes de mots à partir du lexique ou dans leur contexte d'occurrence. Ce module permet également d'associer aux mots du texte, ou aux formes lexicales, des propriétés aux valeurs numériques ou symboliques, en contexte ou hors contexte. Dans le cas de la catégorisation en contexte, c'est comme si l'on ajoutait une ligne d'information dans le corps du texte, concernant une caractéristique se rapportant à un mot ou à un segment textuel. Dans le cas de la catégorisation hors contexte, il s'agit de l'ajout d'une colonne au catalogue des formes du texte. SATOINT permet également la création et la projection de dictionnaires de catégories et de locutions - unités polylexicales - sur les textes ou sur le lexique.

Ce module permet également de segmenter le corpus en autant de sous-textes à partir des propriétés textuelles déjà définies ou encore de certaines caractéristiques liées aux formes lexicales. Les segments ou les domaines de texte, ainsi définis par le chercheur, serviront de base à la comparaison des divers résultats obtenus par l'application de patrons de fouille. SATOINT tient en effet son nom de sa fonction essentielle d'interrogation des données textuelles et lexicales. Il comporte une syntaxe d'interrogation simple et efficace. Celle-ci admet comme élément de recherche soit l'expression littérale des éléments ou une combinaison de caractères de remplacement permettant notamment des jeux de troncation à

gauche, à droite ou à l'intérieur des chaÎnes de caractères. Les requêtes peuvent être également une combinaison de mots et de descripteurs. Dans tous les cas, ces patron de fouille permettent de produire des lexiques d'occurrences et de cooccurrences et des concordances à contexte variable.

Enfin, SATOINT permet d'effectuer certaines analyse des données : calculs de participation relative, de lisibilité et de distance statistique. SATO permet la programmation de fichiers exécutables et est muni d'une interface afin de pouvoir récupérer les résultats du dépouillement dans des logiciels d'édition de textes, des chiffriers, des bases de données.ou des logiciels d'analyse statistique.

Cet article rend compte, en s'appuyant sur quelques exemples, d'une expérience d'application de SATO pour le dépouillement d'un corpus documentaire relativement important avec l'objectif de produire une interprétation sociologique. Nous décrirons dans un premier temps le contexte de la recherche dont est tiré ce qui suit. Nous présenterons par la suite différentes opérations qui ont été produites sur ce corpus et donnerons des exemples de résultats.

2. Justice juvénile et dossiers psycho-sociaux

Durant la deuxième moitié du dix-neuvième siècle, le mouvement philanthropique des *childsavers* imposait graduellement partout en Occident le dessein d'un traitement différencié pour les jeunes délinquants. Les *childsavers* entreprenaient une campagne de sauvegarde des enfants judiciarisés au nom d'une idéologie humaniste s'appuyant sur les principes de la morale traditionnelle et sur les conceptions naissantes de la criminologie positiviste. Au Canada, la Loi sur les jeunes délinquants de 1908 consolidait dans cette direction la philosophie de la probation, pierre angulaire d'une cour conçue comme clinique de prise en charge "personnalisée". Dès lors, par opposition aux tribunaux criminels où la punition et la protection de la société représentaient les principaux objectifs, la cour juvénile devait servir aux meilleurs intérêts des enfants référés (Hudson et al., 1988 :4).

Le mineur délinquant, déresponsabilisé et en conséquence démuni de volonté criminelle, est devenu alors la cible de l'appareil examinateur de la justice juvénile et désormais pris en charge en tant que manifestant un problème de comportement. Les travailleurs sociaux, les psychiatres et les psychologues s'impliquèrent dans la gestion judiciaire des enfants déviants, entre autres, en produisant des évaluations savantes de leur profil psycho-social. L'identité de chaque jeune délinquant était ainsi traduite sous forme de *dossier psycho-social* . L'analyse historique de ce genre de matériel (très rarement accessible aux chercheurs en raison de son caractère strictement confidentiel) constitue un chapitre inéluctable dans le travail de compréhension des mutations du système judiciaire juvénile au Canada.

La Cour juvénile de Winnipeg (Province du Manitoba), établie en 1909 d'après la législation fédérale, est la première cour juvénile au Canada et elle a servi de modèle pour la mise sur pied de tribunaux semblables dans les autres Provinces canadiennes. Le corpus de cette recherche a été constitué au moyen d'une opération de découpage sur un ensemble disponible de dossiers de mineurs ayant eu des contacts à titre de justiciables avec cette cour depuis sa création et jusqu'aux débuts des années '60. Le corpus comprend le texte intégral de 261 rapports professionnels provenant de 101 dossiers. Il comprend 112 rapports d'enquête sociale (qui constituent le sous-corpus SOC) et 149 rapports d'expertise psychiatrique ou psychologique (sous-corpus PSY). La plupart de ces rapports ont été produits pendant les années '40 (70%), le reste correspondant davantage aux décennies antérieure (8%) et

postérieure (17%). La distribution des rapports selon le sexe du mineur référé montre une légère prédominance masculine (53%). Les 112 rapports SOC ont été produits par 42 travailleurs sociaux, tandis que les 149 rapports PSY ont été rédigés par 15 psychiatres et 6 psychologues.

Le matériel empirique recueilli consiste en un ensemble de photocopies des dossiers. Au moyen d'un logiciel d'édition, le texte de chaque dossier a été transcrit et enregistré sous forme de fichier informatique. En vue de l'analyse à effectuer, on a déposé sur les fichiers des marques de segmentation afin d'être en mesure d'interroger le matériel de façon sélective. Ainsi, le texte fut préparé pour être décortiqué selon les divers types de rapport, leurs dates de rédaction, leur position dans le dossier d'origine, les institutions productrices, les noms des professionnels qui les ont signés et le sexe des mineurs référés.

L'étendue totale du corpus est de 129,693 mots. Le sous-texte SOC englobe 100,923 mots (77.8%) et le sous-texte PSY englobe 28,770 (22.2%). SATO permet de traiter les deux sous-textes soit comme un seul ensemble ou séparément.

3. Le lexique : fréquences, mots-thèmes, locutions

Ainsi que nous l'avons suggéré plus haut, SATO peut être considéré comme une boÎte à outils par opposition à l'image informatique conventionnelle de "boÎte noire". En effet, l'utilisateur de ce logiciel dispose d'une série de procédures capables de produire des opérations dont le choix, la séquence d'application et l'articulation relèvent de la nature spécifique de chaque recherche. Soit à partir de l'exécution d'un protocole d'analyse prédéfini, soit selon une stratégie exploratoire et ouverte, le chercheur "manoeuvre" ces procédures selon ses besoins propres. Dans les pages qui suivent, nous exposerons, de manière extrêmement simplifiée, le cheminement effectué dans l'étude des dossiers psycho-sociaux de la Cour juvénile de Winnipeg, tout en montrant successivement les fonctionalités de SATO qui ont permis, à chaque étape de la démarche, d'obtenir des résultats pertinents en regard de notre problématique théorique.

L'un des premiers pas de toute analyse à entrée lexicale est bien sûr l'examen du vocabulaire. Au moyen de la commande Écrire lexique de SATO, la liste complète des formes lexicales dans le texte (ou dans les sous-textes) peut être générée de façon immédiate à l'écran ou acheminée sur un fichier externe en format ASCII. Le patron de rangement (alphabétique, par ordre de fréquences, par longueur des lexèmes) et les restrictions (par exemple : n'admettre dans le lexique que les lexèmes correspondant à un rang de fréquence donné) sont fixés dans chaque cas par l'utilisateur.

Étant donné que l'un des buts centraux de l'analyse de données textuelles consiste à réduire de manière optimale la quantité d'information à interpréter, la stratégie la plus habituelle pour aborder l'étude du vocabulaire est la déploiement de la liste des lexèmes rangés par ordre de fréquences décroissantes. Depuis les premières analyses lexico-statistiques, les chercheurs ont toujours constaté la forte asymétrie des distributions de fréquences lexicales; environ la moitié des lexèmes de tout texte de grande taille n'y sont utilisés qu'une seule fois. En revanche, un nombre très réduit de lexèmes englobe normalement une grande partie des occurrences. L'observation des lexèmes les plus fréquents dans un texte fournit donc un aperçu du contenu prédominant, en partant de la prémisse que la répétition régulière de certains choix lexicaux chez un locuteur (ou un ensemble de locuteurs) suggère l'existence d'un nombre restreint de thèmes du discours. Nous utiliserons alors le concept de "mots-thèmes" (Guiraud, 1953 : 155)

pour désigner les formes lexicales qui caractérisent un texte (ou un sous-texte) à cause de leur très haute fréquence (le seuil étant évidemment fixé de façon arbitraire).

Le tableau 1 représente les lexiques SOC et PSY triés par SATO selon une hiérarchie de fréquences jusqu'au rang 80. Des 6,888 lexèmes composant le texte SOC, les 80 plus fréquents (les mots-thèmes SOC) représentent ensemble 53.5% des occurrences lexicales; des 3,625 lexèmes composant le texte PSY, les 80 plus fréquents (les mots-thèmes PSY) en représentent 50%.

Une liste élémentaire de ce genre est en mesure d'offrir déjà une diversité de pistes pour l'analyse. Il y a d'abord ce qui pouvait être attendu d'après la nature du corpus : les rangs 1 à 20 sont dominés par les formes fonctionnelles usuelles de l'anglais écrit standard (*"the"*, *"and"*, *"to"*, etc.) et par le nom anonymisé du mineur référé (*"XXX"*). On y trouve, toutefois, plusieurs éléments fort significatifs; par exemple, le fait que le lexique SOC semble être caractérisé largement par des substantifs renvoyant au foyer (*"home"*, *"mother"*, *"Mrs. XXX"*, *"family"*, *"father"*, etc.), alors qu'on pourrait penser qu'un discours de ce type devrait référer à l'enfant judiciarisé plutôt qu'à son entourage familial. D'autre part, les nombreuses formes nominales référant à la temporalité (*"time"*, *"years"*, *"months"*, etc.) reflètent l'importance du principe d'agencement chronologique de l'information; les références à l'école (*"school"*) et au travail (*"work"*) démarquent les sphères d'activité privilégiées dans le discours des enquêteurs sociaux. Quant au lexique PSY, on observe sa perspective disciplinaire spécialisée, surtout au niveau de l'évaluation de l'aptitude intellectuelle (*"intelligence"*, *"I.Q."*); d'ailleurs, la haute fréquence du pronom personnel *"I"*, ainsi que la présence de certains verbes de modalisation (*"think"*, *"find"*), mettent en évidence la prise en charge subjective du discours par les psychiatres.

Or, pour représenter les lexiques, SATO a découpé automatiquement la linéarité syntagmatique du texte suivant le critère : "une chaÎne graphique de caractères entre deux espaces ou délimiteurs = un lexème". Par contre, on repère dans le lexique certaines formes (*"Court"*, *"Home"*, *"School"*, *"History"*) qui pourraient faire partie de locutions, celles-ci étant des expressions figées formées par deux ou plusieurs mots. Dans le corpus examiné ici, il peut s'agir de noms d'institutions (*"Juvenile Court"*) ou de termes polylexicaux (où la valeur sémantique des composantes n'est pas équivalente à celle de leur agrégation; par exemple : *"Social History"* n'est pas seulement une "histoire à caractère social", mais un concept ayant un sens spécifique dans le discours professionnel du *caseworker*). Afin de pouvoir traiter les locutions comme des lexèmes unitaires, SATO admet la projection d'un dictionnaire externe des expressions composées. Ce dictionnaire, préparé par l'utilisateur avec un éditeur de textes, est géré par LOCUTION, un programme auxiliaire de SATO. Suite à la projection du dictionnaire, les locutions sont bloquées dans le texte au moyen d'un code (_) formant ainsi de nouvelles formes lexicales qui seront autant de nouvelles entrées du lexique (par exemple, *"Hugh_John_McDonald_School"* devient un lexème unitaire du vocabulaire).

Dans le cas de l'étude des dossiers psycho-sociaux, le traitement des locutions comporte encore un intérêt supplémentaire. Outre les locutions relatives aux noms propres et aux termes sémantiquement complexes, un type particulier de locutions a retenu aussi notre attention : les rubriques d'information. Les rapports sociaux et psychiatriques sont rédigés selon une organisation topique, une "rubrique" étant, dans ce genre de documents, une sorte de sous-titre précédant un segment thématique (couramment un paragraphe). En voilà un exemple :

Child's History : Mother states child was born in General_Hospital after full term pregnancy. The conditions of his birth were normal; he began teething at the age of nine months, walked at the age of 16 months; he was a breast fed child and was weaned at the age of 15 months.

Il nous a semblé avantageux de traiter ces rubriques comme étant autant de locutions puisqu'elles constituent un système "naturel" de repères thématiques à l'intérieur du corpus. En effet, lorsque le rédacteur du rapport a posé une rubrique et lui a associé un segment textuel (ou bien a associé ce même segment textuel à une formule prescrite d'un formulaire imprimé), il a découpé une zone de la réalité à laquelle réfère son discours en plaçant les données sous une "étiquette". Voilà un extrait de la liste de rubriques bloquées comportant la forme *"History"*:

Child'_s_History	*Personal_History*
Child'_s_History_and_Character	*Personal_History_and_Character*
Family_History	*Personal_History_of_Patient*
Family_History_and_Character	*Placements_and_Work_History*
Family_History_and_Home_Conditions	*Previous_Court_History*
Health_History	*School_and_Work_History*
History_and_Character_of_the_Child	*School_History*
History_of_School_Progress	*Sexual_History*
Medical_History	*Social_History_and_Reactions*
Personal_and_Development_History	*Work_History*

Après la projection du dictionnaire de locutions sur le texte, le lexique subit certaines modifications : un nombre de formes qui étaient libres deviennent des composantes de locutions, ce qui entraÎne, d'une part, la disparition de quelques formes du lexique (*"History"* n'existe plus en tant que lexème) et, d'autre part, l'apparition de nouvelles formes (*"Child'_s_History"*, etc.). L'opération de blocage de locutions a donc produit une transformation des "données brutes" visant leur adéquation à la stratégie de recherche.

4. Catégorisation : mots-outils, rubriques thématiques

Tout traitement quantitatif de données comporte une activité de codage ayant pour effet la réduction de la diversité et la concentration des traits de la population examinée. Dans le cadre de l'analyse de textes, la catégorisation est une procédure de classification de l'information lexicale par laquelle chaque unité peut être classée par rapport à des variables ou "propriétés" syntaxiques, sémantiques, thématiques, contextuelles, etc. La commande Propriété en environnement SATO permet d'appliquer des grilles complexes de catégories et

d'effectuer le repérage d'information à partir d'elles. La catégorisation est réalisée par l'attribution de valeurs de propriétés aux lexèmes en contexte ou hors contexte. Par exemple, la forme lexicale *"boy"* pourrait être catégorisée hors contexte (dans le lexique) et recevoir la valeur "nom commun" de la propriété "classe-grammaticale" tandis qu'en même temps, ses occurrences en contexte (dans le texte) pourraient recevoir, soit la valeur "sujet" (*"The boy has been running away"*), soit la valeur "objet" (*"Mrs. XXX dislikes the boy"*) de la propriété "position". Un décompte des seuls lexèmes "noms communs / position-sujet" pourrait alors être produit.

Dans le cadre de l'étude des dossiers judiciaires, l'une des premières préoccupations a consisté à catégoriser tous les mots-outils (par opposition aux mots-pleins) afin de les exclure des fouilles hors contexte. La classe des mots-outils inclut les mots qui ont un rôle fonctionnel dans la langue (articles, conjonctions, prépositions, verbes auxiliaires, etc.) et qui, par conséquent, sont relativement "vides" de contenu sémantique. SATO permet, au moyen de la commande Dictionnaire, de gérer des thésaurus externes comportant des valeurs de propriétés (de façon similaire à l'application de dictionnaires de locutions). Voilà un extrait du dictionnaire des mots-outils compilé pour cette recherche :

a*mot=outil	along*mot=outil	anybody*mot=outil
about*mot=outil	also*mot=outil	anymore*mot=outil
above*mot=outil	although*mot=outil	anyone*mot=outil
across*mot=outil	altogether*mot=outil	anything*mot=outil
after*mot=outil	always*mot=outil	anyway*mot=outil
again*mot=outil	am*mot=outil	anywhere*mot=outil
against*mot=outil	among*mot=outil	are*mot=outil
ago*mot=outil	an*mot=outil	around*mot=outil
ahead*mot=outil	and*mot=outil	as*mot=outil
all*mot=outil	another*mot=outil	at*mot=outil
almost*mot=outil	any*mot=outil	away*mot=outil

Dans la section précédente, il a été question du blocage des rubriques d'information. En liant les composantes des titres de rubriques, nous avons constitué une famille d'expressions polylexicales qui permettent de découper le texte en segments thématiques. Dans ce sens, nous avons procédé au regroupement des rubriques du texte en fonction de leur appartenance thématique. Le mineur en probation est la cible de trois stratégies d'intervention qui donnent lieu à quatre catégories thématiques. L'enfant et sa famille constituent d'abord un "cas social" aux yeux du travail social, mais également un "cas judiciaire" pour les instances légales. Il est enfin un "cas médical" du point de vue de l'appareil clinique. Définissant une propriété "Rubrique", nous avons donc attribué une des quatre catégories suivantes à chaque occurrence de ces rubriques : "cas-social-famille" (rubrique =fam), "cas-social-enfant" (rubrique=enf),

"cas-judiciaire" (rubrique=jud) ou "cas-médical" (rubrique=méd). Voilà la liste des principales rubriques du texte SOC, avec leur fréquence et leur catégorie :

F	Cat.	Rubrique	F	Cat.	Rubrique
91	fam	*Mother*	19	enf	*School_and_Employment_Record*
89	fam	*Father*	18	enf	*Personal_History_and_Character*
66	fam	*Family_History*	17	med	*Physical_and_Mental_Condition*
41	enf	*School*	17	enf	*Social_Interests_and_Companions*
38	fam	*Siblings*	16	enf	*Personal_History*
30	fam	*Home_Conditions*	15	enf	*Education*
25	enf	*Personality*	15	enf	*Habits*
21	jud	*Cause_of_Delinquency*	15	fam	*Home_and_Home_Condition*

En termes généraux, les rubriques portant sur la famille établissent une histoire familiale et un aperçu des conditions du foyer; les rubriques sur l'enfant touchent à sa performance scolaire, sa personnalité et son histoire individuelle; les rubriques du "cas-judiciaire" réfèrent principalement aux causes du problème, aux raisons de l'intervention et à l'histoire judiciaire du mineur; enfin, les rubriques du "cas-médical" décrivent les conditions physiques et mentales de l'enfant concerné ou réfèrent à lui en termes de "patient" et de "maladie".

5. Domaines et distances

L'un des principes méthodologiques de base dans l'analyse de textes est la comparaison. Étant donné un corpus, le contraste inter-textuel ou intra-textuel constitue une activité essentielle de la démarche. En ce qui concerne l'analyse intra-textuelle (comparaison de segments à l'intérieur du corpus), les critères de partition du texte doivent être fixés par le chercheur, deux stratégies complémentaires étant viables à ce propos. D'un côté, il y a ce qu'on peut nommer la "partition naturelle du texte"; cela comprend, par exemple, les sections d'un document, les unités qui composent le corpus (des articles de journaux, des tracts, etc.). De l'autre côté, il y la partition imposée par le chercheur à partir de l'application de critères de recherche; par exemple, la différenciation des locuteurs dans un dialogue ou des paragraphes thématiques dans un écrit.

Pour ce qui est de la "partition naturelle", il s'agit habituellement d'une segmentation exhaustive et linéaire du texte : Texte X = segment A + segment B + ... + segment N. SATO, par le biais de la commande Segmenter, rend possible l'application de divers types de segmentation : par lignes, par paragraphes, par documents ou à partir de marques de segmentation déposées au préalable par l'utilisateur lors de la préparation du matériel. La partition du deuxième type comporte la délimitation discontinue d'un ou de plusieurs

domaines de données extraites à différents endroits du texte. SATO permet, avec la commande Domaine, de concentrer l'analyse sur un domaine défini par l'utilisateur, de façon que toutes les procédures ultérieures (par exemple, le calcul de fréquences, la fouille de concordances) n'agissent que sur la partition pertinente. Un domaine quelconque peut être activé ou désactivé par l'utilisateur selon les besoins de la démarche analytique.

Prenons un exemple. Tel qu'on l'a déjà fait remarquer, l'investigation sur la famille du mineur semble occuper une place cruciale dans l'enquête menée par le tribunal. On a donc décidé de constituer un domaine textuel avec l'ensemble des contenus des rubriques catégorisées sous la valeur "cas-social-famille". A partir d'une commande appropriée, SATO procède à une fouille du texte, et chaque fois qu'il détecte l'occurrence d'une rubrique de ce type, il "ramasse" le contenu qui lui est associé et le range dans le domaine "Famille" (domaine A).

Le domaine étant constitué, on s'intéressera au dépistage systématique des contenus qui le caractérisent. Plus particulièrement on identifiera les formes lexicales qui marquent la spécificité du vocabulaire correspondant aux rubriques sur la famille par rapport aux autres rubriques des dossiers sociaux. En effet, SATO inclut un algorithme qui permet d'évaluer l'originalité lexicale comparative de deux domaines. Il s'agit de la commande Distance, basée sur la mesure de distance du Chi-carré, qui donne comme résultat un tableau comportant les 50 formes lexicales qui contribuent le plus à l'écart de deux domaines du point de vue de leur vocabulaire. Pour appliquer la procédure de distance, l'ensemble des contenus des rubriques restantes ("cas-social-enfant" + "cas-judiciaire" + "cas-médical") est pris à titre de second sous-texte (domaine B) aux fins de la comparaison; le texte SOC au complet est pris comme source de pondération dans le calcul de la distance du Chi-carré. Le tableau 2 affiche le résultat de l'opération. Il permet de distinguer les lexèmes qui marquent davantage l'originalité de chaque domaine en termes comparatifs. La première colonne affiche le poids relatif (pourcentage) de chaque lexème par rapport au vocabulaire plein du texte (par exemple, la première ligne nous indique que les occurrences du lexème *"born"* représentent 0.74% des mots-pleins du texte SOC). Les deuxième et troisième colonnes font de même par rapport aux vocabulaires des domaines considérés (les occurrences de *"born"* représentent 1.54% des mots-pleins du domaine Famille, et 0.28% des items pleins du domaine "autres"). La quatrième colonne traduit en pourcentage la contribution relative de chaque lexème à la mesure de distance (*"born"* explique 1.96% de la valeur de distance totale entre les deux domaines). Enfin, la dernière colonne affiche les pourcentages cumulés de cette contribution.Parmi les 50 lexèmes signalés par le logiciel, il y a ceux qui appartiennent davantage au premier sous-texte plutôt qu'au second; ces lexèmes apparaissent accompagnés d'un astérisque (par exemple, *"born"* est utilisé davantage, en termes statistiques, dans le domaine Famille que dans l'ensemble des autres domaines : 1.54% versus 0.28%). Ce sont les lexèmes que nous prendrons comme étant les "mots-clés" du domaine Famille. L'algorithme de SATO nous a indiqué qu'il existe une utilisation significativement différente de ce groupe de mots en termes de fréquences comparatives entre le domaine Famille et le reste du texte SOC. Le domaine Famille est donc caractérisé par 42 mots-clés (ou lexèmes qui expriment l'originalité relative de ce domaine-ci), leurs occurrences représentant 17% du total de mots-pleins du domaine :

born, Mr. XXX, married, children, family, living, marriage, Mrs. XXX, died, husband, education, Winnipeg, father, years, lived, wife, health, house, common-law, Ukrainian, Roman_Catholic, rooms, good, moved, unknown, nationality, present, floor, support, drinks,

parentage, presently, descent, French, Protestant, parents, described, live, English, Russia, Detroit, occupy

Outre les mentions génériques de la famille (*"family"*) et de ses membres (notamment les parents : *"Mr. XXX"*, *"Mrs. XXX"*, etc.), on peut y observer clairement divers sujets : naissances (*"born"*), lien marital (*"married"*, *"common-law"*), mort (*"died"*), éducation (*"education"*), habitation (*"living"*, *"house"*, etc.), santé (*"health"*), nationalité (*"Ukrainian"*, etc.) ou origine (*"descent"*), et religion (*"Roman_Catholic"*, etc.). Il y a aussi une référence à l'alcool (*"drinks"*). La grille implicite de description objective de la famille paraÎt couvrir donc davantage les aspects suivants : genre et lieu de logement, statuts d'éducation et marital des parents, leur origine ethnique, leur état de santé, leur religion, leur possible absence du foyer en raison de décès et, enfin, la consommation d'alcool.

6. Concordances : le retour au contexte

L'examen des lexiques et leur comparaison privilégie l'axe paradigmatique (rapports de substitution) de l'organisation textuelle. Les fréquences d'utilisation des lexèmes servent de quantificateurs de contenu; la thèse implicite veut que le plus une forme lexicale est répétée dans un discours donné, le plus elle imprègne ce discours de sa charge sémantique. Les avantages opérationnels d'une telle perspective sont évidents au niveau de la manipulation des données, mais les limites posées sur le plan de l'interprétation sont aussi manifestes dès lors qu'on sait que le sens en discours découle autant du choix des lexèmes que de leur combinatoire. Le fait que, par exemple, *"intelligence"* soit le deuxième substantif le plus fréquemment utilisé dans le texte PSY nous renseigne à propos de l'intention référentielle du locuteur (il *veut* parler de l'intelligence de l'enfant dans son discours), mais ne nous apprend rien sur le régime de co-occurrences qu'il lui associe (adjectivation, prédication, etc.). Le retour au contexte syntagmatique (rapports de succession) devient donc nécessaire et cela peut être réalisé de manière très efficace au moyen de la fouille de concordances.

Une concordance est définie, en termes généraux, comme la chaÎne syntagmatique (ou la liste des chaÎnes syntagmatiques) comportant l'occurrence d'une ou de plusieurs formes lexicales sélectionnées. Cet outil permet d'observer le voisinage des lexèmes à l'intérieur des énoncés : les mots sont saisis "en contexte". Le repérage de concordances en SATO est une opération automatique déclenchée à partir de la définition d'un patron de fouille (protocole de conditions qu'un énoncé doit rencontrer pour être retenu dans la concordance). La procédure Concordance de SATO admet un niveau de précision très élevé lors de la formulation du patron de fouille. En voilà un, en guise d'exemple :

mot : intelligence

typ=(psy1, psy2)

sex~inconnu

Ce patron sélectionnera toutes les concordances où le lexème *"intelligence"* apparaÎt dans un rapport de type psychiatrique (psy1) ou psychologique (psy2) référant à un mineur dont on connaÎt le sexe. Comme on peut le voir, il s'agit d'un patron combinant plusieurs conditions logiques. SATO effectue la fouille et affiche (ou achemine dans le fichier témoin) la liste des concordances satisfaisant les critères établis. La fouille de concordance implique une définition préalable des bornes des chaÎnes syntagmatiques à repérer. Les deux critères les plus souvent utilisés sont l'extension par nombre d'unités à gauche et à droite du pôle et la

délimitation par la ponctuation. Dans le premier cas, on peut fixer, pour une fouille donnée, un contexte de, par exemple, trois lexèmes de chaque côté; dans le second cas, on peut fixer un contexte déterminé par les délimiteurs forts (point, point d'interrogation, etc.). A partir du patron de fouille ci-dessus et pour un contexte de ponctuation forte, 122 phrases au complet avec leurs codes d'identification ont été obtenues. Voilà, par exemple, une des concordances:

110 *page=section9/12/50/3 ... *page=section9/12/53/5
*ins=jcmd*typ=psy1*ord=1*sex=mas*pér=1944*num=19802*sig=bir

This boy intelligence is somewhat below normal and I have classified him as borderline or low normal.

Cette concordance illustre le type de contexte syntagmatique que l'on trouve associé de façon régulière au mot *"intelligence"* dans le texte PSY : registre subjectif comme source d'autorité (*"I have classified him as..."*) et qualification sur l'axe de normalité (*"below normal", "borderline", "low normal"*). Il s'agit, en effet, d'un discours dont la valeur de vérité s'appuie sur le savoir du locuteur et où il est question de classer l'objet (l'intelligence du mineur) par rapport à une échelle objective.

Revenons à notre exemple du domaine Famille. Lors de l'analyse de la distance statistique entre le domaine Famille et le reste du texte SOC, on a effectué l'identification des mots-clés (lexèmes marquant l'originalité relative du vocabulaire) et on a produit, sur cette base, une série d'inférences. Dans une seconde étape d'analyse, ces mots-clés peuvent servir de patrons de fouille de concordances afin d'observer leur sens en contexte. A titre d'exemple, nous prendrons ici la forme *"drinks"*; le tableau 3 montre les 10 premières concordances de la liste produite par SATO. On remarque nettement comment l'enquêteur établit une "mesure" du vice (*"a little", "quite a bit", "somewhat", "heavily", "to excess"*) et l'associe à d'autres modèles de déviance (*"smokes", "gambles", "is promiscuous", "is unstable"*). Il est clair que ce type d'information permet d'approfondir notre connaissance du modèle de normalité familiale implicite à l'intervention discursive des travailleurs sociaux.

Or, si l'on répétait cette opération pour les 42 mots-clés du domaine Famille (et aux autres domaines : cas-social-enfant, cas-judiciaire et cas-médical), on obtiendrait comme résultat un document (formé par l'ensemble des listes de concordances) à être analysé en tant que représentation simplifiée et organisée de l'information textuelle spécifique au domaine.

Conclusion

Nous avons voulu illustrer brièvement, à partir d'un exemple de recherche, les principales caractéristiques du logiciel SATO. Ce logiciel comporte un grand nombre de fonctionnalités que l'on retrouve éparpillée dans d'autres systèmes. Comme on a pu s'en rendre compte, SATO n'est pas d'abord orienté vers l'analyse statistique de données, mais plutôt pensé comme boÎte à outils pour la manipulation de données textuelles : annotation, catégorisation, segmentation, blocage des locutions et génération de lexiques et de concordances. Ces fonctionnalités sont associées dans un cadre ergonomique donnant, en tout temps, le contrôle au chercheur. SATO simule, en quelque sorte, les opérations de traitement et d'analyse que tout chercheur produit sur ses corpus, tout en démultipliant sa capacité de travail (traitement de grands corpus, disponibilité immédiate de données qualitatives ou quantitatives, etc.) et en permettant un continuel va et vient entre les données et les opérations mises en oeuvre pour

les analyser (production et correction d'opérations de catégorisation et de segmentation, par exemple).

Dans le cas de notre recherche, SATO aura été un instrument précieux pour aider à l'analyse de documents textuels sur une base comparative. La souplesse des opérations nous aura permis de produire un grand nombre de descriptions de ces données, mais surtout de progresser, par l'expérimentation instantanée, vers des représentations de plus en plus adéquates de leur contenu.

Références

DAOUST, F. (1990): "L'informaticien, le lecteur et le texte. L'approche SATO". *ICO. Intelligence artificielle et sciences cognitives au Québec*, volume 2, numéro 3; 55-60.

GUIRAUD, P. (1953): *Langage et versification d'après l'oeuvre de Paul Valéry*. Klincksieck, Paris

HUDSON, J., HORNICK, P. et BURROWS, B. (Eds.) (1988): *Justice and the Young Offender in Canada*. Wall & Thompson, Toronto.

MEYER, Ph. (1977): *L'enfant et la raison d'État*. Éditions du Seuil, Paris.**Fiche technique de SATO:**

Développé en turbo-Pascal, SATO peut être utilisé sur du matériel de type IBM-PC et compatibles (8086, 80286, 80386) et ne nécessite pas de co-processeur mathématique. Tourne sous DOS et exige 640 KiloOctets de mémoire RAM. Peut utiliser la mémoire étendue (LIM 4.0). Ne demande que 500K de mémoire de masse pour l'installation du logiciel. Limites physiques de SATO pour un corpus (collection de documents):

nombre maximum de lexèmes (entrées au lexique) :	28 000
nombre maximum de caractère dans le lexique :	256 000
nombre maximum de mots et lignes dans le texte:	1 419 264