REFERENCE

Citation

Duchastel, Jules, François Daoust et Dimitri della Faille (2004) "SATO-XML : une plateforme Internet ouverte pour l'analyse de texte assistée par ordinateur", dans Heiden, Serge et Bénédicte Pincemin (sous la dir.), 7es Journées d'analyse de données textuelles, Louvain-la-Neuve, Presses Universitaires de Louvain, pp. 353-363.

RIS (EndNote)

TY - CONF
AU - Duchastel, Jules
AU - Daoust, François
AU - Della Faille, Dimitri
PY - 2004
TI - SATO-XML : une plateforme Internet ouverte pour l'analyse de texte assistée par ordinateur
T2 - Le poids des mots. Actes des 7es Journées d'analyse de données textuelles
SP - 353
EP - 363
CY - Louvain-la-Neuve
PB - Presses Universitaires de Louvain
A2 - Purnelle, Gérald
A2 - Fairon, Cédrick
A2 - Dister, Anne
ER -

SATO-XML : une plateforme Internet ouverte pour l'analyse de texte assistée par ordinateur

Jules Duchastel[1], François Daoust[2], Dimitri della Faille[3]

[1]Professeur au département de sociologie, UQAM – Montréal – Canada

[2]Informaticien au Centre ATO, UQAM – Montréal – Canada ; doctorant en sciences du langage à l'Université de Franche-Comté – Besançon – France
daoust.francois@uqam.ca

[3]Doctorant au département de sociologie, UQAM – Montréal – Canada
della_faille_de_leverghem.dimitri@courrier.uqam.ca

Abstract

In this contribution, we present a computer-based infrastructure available on the Internet, which allows the manipulation and analysis of text corpora. By the way of an HTML interface the researcher is given access to a personal workspace, a text library, some lexical resources, as well as software applications and procedures for a collaborative work respectful of everyone's data and specific analysis' strategies. The SATO software, available in a client-server mode, allows the categorization of data and the iterative construction of protocols of analysis. XML gives the opportunity to save and exchange data in a standard format. Thus, the described data can be either imported from or exported to other software applications for statistical, linguistic or graphic treatments. The interface available on the Internet includes modes of simplified access to large documented corpora, in particular those of interest for Professor Jules Duchastel's Canada Research Chair in Globalization, Citizenship and Democracy. In this contribution, we are presenting a few exploratory analyses as examples of the possibilities of this computer-based infrastructure.

Résumé

Cet article présente une infrastructure informatique, accessible par le Web, qui permet de manipuler et d'analyser des corpus de textes. Une interface HTML donne au chercheur l'accès à un espace de travail personnel et à des librairies de textes, de ressources lexicales, de programmes et de procédures permettant d'envisager un travail coopératif qui respecte les stratégies d'analyse et les données de chacun. Au niveau des traitements, le logiciel SATO, accessible en mode « client-serveur » permet de catégoriser les données et de construire des protocoles d'analyse de façon itérative. La normalisation XML permet une conservation et un échange des données dans un format standard. Ainsi, les données décrites peuvent être importées ou exportées pour être traitées par divers logiciels statistiques, linguistiques ou graphiques. L'interface Web comprend aussi des modes simplifiés d'accès à de grands corpus documentés, en particulier ceux faisant partie des axes de recherche de la Chaire de recherche du Canada en Mondialisation, citoyenneté et démocratie du professeur Jules Duchastel. Dans cet article, quelques analyses exploratoires illustrent l'utilisation de cette infrastructure logicielle.

Mots-clés : analyse de texte par ordinateur, SATO-XML, interface HTML, corpus sur le Web.

1. Introduction

Le développement d'une infrastructure de recherche au profit de la communauté des chercheurs en analyse de texte vise à rendre accessible sur Internet des *corpus vivants*, c'est-à-dire analysables en ligne en fonction des stratégies spécifiques de chaque chercheur. Nous présentons ici une architecture développée autour du logiciel SATO (Système d'analyse de texte par ordinateur ; Daoust, 1996), mais qui permet également de rassembler divers modules d'analyse statistique, linguistique, etc.

La section deux situe cette architecture dans le contexte du développement du portail ATO-MCD relié à la Chaire de recherche du Canada en Mondialisation, citoyenneté et démocratie du professeur Jules Duchastel (2001). Elle introduit également les principes méthodologiques qui fondent cette architecture. La troisième section est consacrée à la présentation de l'architecture et du modèle SATO. Enfin, dans une quatrième section, nous présentons des exemples d'utilisation de l'infrastructure logicielle accessible par le Web.

2. Contexte et principes méthodologiques

C'est au printemps de 2001 qu'ont débuté les travaux de développement d'une infrastructure de recherche élargie en analyse de texte par ordinateur dans le cadre de la Chaire de recherche du Canada en Mondialisation, citoyenneté et démocratie. Le projet vise à intégrer des acquis développés au cours des années mais qui restent encore trop dispersés (Duchastel, 1993 ; Duchastel et Armony, 1996).

Au niveau méthodologique, cette intégration vise à soutenir une démarche d'analyse de discours. Comme l'écrivait Michel Pêcheux,

> L'analyse de discours ne prétend pas s'instituer en spécialiste de l'interprétation maîtrisant « le » sens des textes, mais seulement construire des procédures exposant le regard-lecteur à des *niveaux opaques à l'action stratégique d'un sujet* [...]. L'enjeu crucial est de *construire des interprétations* sans jamais les neutraliser ni dans le « n'importe quoi » d'un discours sur le discours, ni dans un espace logique stabilisé à prétention universelle (Pêcheux, 1984 ; cité par Maingueneau, 1997).

Comme l'indique Maingueneau,

> Étant donné le statut de l'analyse de discours, on ne peut pas se contenter d'« appliquer » de manière aveugle des protocoles méthodologiques à des corpus. À chaque fois, il faut mener une réflexion spécifique pour construire, de manière interactive, le corpus et son mode d'investigation (Maingueneau, 1997).

L'architecture informatique que nous proposons vise à faciliter cette construction dans un processus itératif et contrôlé dont la trace est explicite.

Au niveau des données, le projet vise l'accueil, la conservation et l'exploitation scientifique de corpus de textes numérisés provenant de la communauté canadienne et internationale des chercheurs en analyse du discours. L'exportation et l'importation des données selon un format XML apparaissent comme une condition pour faciliter la conservation, l'échange et le traitement des corpus et des données lexicales. XML, rappelons-le, est un langage général de balisage des documents électroniques qui permet de publier, conserver, annoter et transformer des textes selon un protocole indépendant des formats propriétaires. Faisant l'objet de concertations (The TEI Consortium, 2001), les protocoles de balisage XML facilitent le transfert des données et des résultats entre logiciels. Signalons que, si le projet vise tout particulièrement les données textuelles en langue française portant sur le discours politique, la plateforme est extensible aux autres domaines de recherche en sciences sociales et en lettres.

Au niveau des traitements informatiques, l'objectif est de fournir un environnement flexible, entièrement accessible via Internet, et permettant au chercheur de déployer ses propres stratégies d'annotation, d'exploration et d'analyse de corpus collectifs ou personnels. Au cœur de la plateforme logicielle, on retrouve le logiciel SATO, augmenté de fonctionnalités permettant l'accueil et l'exportation de données en format XML.

Cette technologie permet d'envisager un véritable travail coopératif jumelant un espace de travail personnel avec des ressources partagées : corpus, bases de données lexicales, documentation et guides méthodologiques. Il sera dès lors envisageable de transformer les collaborations fondées sur le partage de résultats en projets de recherche coopératifs durables voués à

la coexploitation de la base de données et au partage des corpus, lexiques et des savoirs socio-sémantiques. Puisque ce poste de travail électronique utilise une technologie Web standard, il est facilement modifiable et documentable par des tutoriels, manuels, bulles d'aides et guides méthodologiques. Il est également aisé d'implémenter des versions multilingues.

Au niveau matériel, le projet privilégie une approche souple faisant appel à de l'équipement standard rassemblé en îlots de traitement rassemblant plusieurs ordinateurs en réseau. La plateforme logicielle peut donc être déployée dans une variété de configurations allant de l'ordinateur personnel à un réseau élaboré d'ordinateurs se partageant les données et les traitements.

Pour comprendre les motifs à la base de cette stratégie de développement, il faut rappeler les grandes étapes d'évolution des technologies informatiques. On a connu la période des ordinateurs centraux basés, d'une part, sur un traitement centralisé et, d'autre part, sur un accès décentralisé aux données et aux traitements par le biais des terminaux accessibles par modem. Par la suite, on a assisté au triomphe de la micro-informatique qui a démocratisé l'accès au traitement des données sur le poste de travail de l'utilisateur devenu plus puissant que les ordinateurs centraux d'autrefois et à un coût qui dépasse à peine celui des terminaux de jadis. La troisième *révolution* informatique a trait à la généralisation de la réseautique via Internet et l'intégration multimédia et hypertextuelle que permettent le Web et le langage HTML.

HTML est un dialecte issu de la norme SGML. Après avoir connu un développement accéléré et un peu anarchique d'HTML avec la concurrence effrénée dans le développement des navigateurs, le W3C qui arbitre le développement du WEB a décidé d'arrêter l'évolution d'HTML pour promouvoir XML, un langage de balisage issu d'une simplification de SGML et qui intègre la notion d'*extensibilité*. Ce retour à plus de rigueur dans la normalisation des formats des données a pour toile de fond l'impératif de l'échange des données sur Internet dans la perspectives de services Web permettant à des ordinateurs d'échanger des données en vue de les traiter.

Par ailleurs, l'ordinateur personnel est devenu à lui seul un véritable centre de calcul dont l'entretien dépasse souvent les capacités de l'utilisateur, en particulier en ce qui concerne les mises à jour des logiciels et des chaînes de traitement. De plus, dans le domaine de la recherche, nous faisons face à des produits en évolution qui ne disposent pas toujours du même niveau de support que les logiciels commerciaux ou grand public. De là, la nécessité d'aller vers des solutions mixtes qui concentrent des ressources de traitement accessibles par le Web et qui extensionnent le bureau de travail personnel du poste local vers des îlots de traitement distants. De là, aussi, la nécessité du travail coopératif, au-delà du simple échange de publications scientifiques, de telle sorte que l'on puisse échanger des données, en ce qui nous concerne les corpus de textes, les bases de données lexicales, les procédures informatiques et les méthodologies. L'accès à des traitements via le WEB, et la normalisation XML des données à des fins d'échanges entre plateformes informatiques, sont donc des tendances en développement. Outre le projet ATO-MCD, citons, à titre d'exemples d'infrastructure Web dans le domaine de l'analyse des données textuelles à des fins de recherche et d'enseignement, les projets Tapor et Weblex.

Le portail Weblex de l'École normale supérieure Lettres et Sciences humaines de Lyon vise à fournir un accès par Internet à des outils d'analyse textuelle. Encore en développement, le portail permettra un accès aux outils lexicométriques développés depuis des années dans des équipes de recherche dont la tradition remonte au Centre de lexicologie politique de Saint-Cloud. Outre l'accès à des outils d'analyse quantitative des données textuelles aux fonctionnalités apparentées à celles de Lexico (Salem) et Hyperbase (Brunet), le logiciel Weblex entend fournir une édition hypertexte du document et un moteur de recherche très complet

(Heiden, 2002). Le Centre ATO de l'UQAM collabore avec l'équipe de Lyon depuis plusieurs données et la convergence vers des protocoles XML devrait faciliter le transfert des données et des traitements entre les deux groupes.

Au Canada, on retrouve un autre projet de développement d'un portail pour l'analyse textuelle. Il s'agit du Text-Analysis Portal for Research (TAPoR) : « TAPoR permettra l'établissement d'une infrastructure de chercheurs et de ressources informatiques pour l'analyse des textes à travers le pays par la mise sur pied de six centres régionaux afin de former un portail national pour l'analyse des textes » (Rockwell *et al.*, 2002).

Pour sa part, SATO, dans sa version HTML, est offert depuis plusieurs années déjà en accès libre au Centre ATO de l'UQAM à l'adresse http://www.ling.uqam.ca/ato. Tout comme la version DOS qui la précédait, SATO-HTML donne la priorité aux fonctions d'annotation et de catégorisation lexicale et contextuelle ainsi qu'aux stratégies d'analyse personnalisées (scénarios) accompagnées de mécanismes de trace de l'exploration. Comme la plupart des logiciels d'analyse textuelle, on retrouve dans SATO les fonctionnalités classiques de concordance et de fréquences lexicales, mais augmentées de dispositifs de catégorisation. Au niveau des fonctions statistiques, seules les fonctions de base sont directement intégrées au logiciel. En contre-partie, le logiciel permet de produire à loisir des matrices d'occurrences destinées à être traitées par des analyseurs statistiques externes, par exemple des analyses factorielles de correspondance[1].

La section suivante décrit l'architecture du système et ses perspectives de développement futur.

3. Architecture de la plateforme SATO-XML

On pourrait qualifier le logiciel SATO de *tableur textuel*. Le système permet d'accueillir un corpus brut ou déjà annoté ; il permet de l'annoter ou de changer l'annotation déjà présente, de catégoriser le corpus selon des grilles définies par l'analyste et une fois décrit, de l'exploiter de multiples manières. SATO permet de garder une trace complète du processus de description et d'analyse du corpus. Le logiciel offre aussi la possibilité de programmer des dispositifs de *lecture électronique* (Daoust, 2002) et, donc, d'établir des protocoles d'analyse personnalisés et adaptés à chaque type de discours.

SATO, dans ses versions 3 et 4, est un logiciel destiné à supporter une variété de stratégies d'analyse textuelle. Il repose sur une reconfiguration du texte linéaire (chaîne de caractères) sous la forme d'un plan lexique/occurrences. L'axe lexical répertorie l'ensemble des chaînes de caractères constituant les mots, ponctuations, et toutes chaînes de caractères admissibles à un alphabet défini par l'utilisateur. L'axe des occurrences représente l'ordonnancement des unités lexicales suivant l'ordre naturel du texte (de gauche à droite et de bas en haut pour les langues latines).

L'objectif de cette reconfiguration est de faire émerger la dimension lexicale du texte. Il est à noter qu'à part quelques normalisations éditiques mineures, cette reconfiguration est non destructrice, c'est-à-dire qu'elle permet à tout moment de reconstituer le texte original dans sa forme linéaire. Cette reconstitution à la volée permet de produire des éditions sur mesure avec mise en évidence des mots (couleur, soulignement, etc.) selon des critères définis par l'analyste. Il est possible d'exporter ces éditions dans des formats facilitant leur traitement par d'autres logiciels. Aussi, derrière chaque forme lexicale et chaque occurrence, on retrouve un hyperlien donnant accès à diverses fonctions de catégorisation et de parcours.

[1] On trouvera dans le chapitre intitulé « Une stratégie intégrée de recherche en sciences humaines dans le Portail ATO-MCD » un exemple d'intégration de diverses composantes logicielles pour le traitement en chaîne d'un corpus de discours politique.

L'émergence de la dimension lexicale du texte dans le plan lexique/occurrences permettra de distinguer la catégorisation hors contexte, qui appartient au lexique de la langue ou du domaine, de la catégorisation contextuelle, qui appartient davantage à l'énoncé et à la structure discursive. Dans SATO, les systèmes de catégorisation sont appelés *propriétés*. Exception faite de quelques propriétés prédéfinies par le logiciel, l'utilisateur définit lui-même ses propriétés selon les besoins de son analyse.

La catégorisation des formes lexicales ou des occurrences au moyen de ces propriétés peut se faire par manipulation directe à l'écran, précodage sur le texte ou par divers dispositifs algorithmiques : dictionnaires, patrons morphologiques ou filtres sur les propriétés, patrons de cooccurrences positionnelles ou booléennes. Le logiciel permet de constituer ses propres dictionnaires. Des dispositifs d'héritage permettent de définir des propriétés textuelles projetées à partir du lexique ou des propriétés lexicales condensées à partir des occurrences. Le *filtre* est un patron syntaxique permettant de désigner et de rassembler un ensemble de formes lexicales ou d'occurrences par des contraintes sur les caractères de la chaîne ou ses valeurs de propriété.

La définition des contextes pour les concordances, cooccurrences ou segments calculés s'effectue à la volée selon les besoins de l'analyse. On peut aussi définir au besoin des sous-textes et leurs lexiques associés. Le logiciel fournit des dispositifs de comptage permettant de produire diverses matrices d'occurrences dans les segments ainsi constitués. Des mesures statistiques simples permettent de révéler ou de contraster la distribution des fréquences associées aux occurrences spécifiées par un filtre SATO. Les matrices produites par le logiciel peuvent servir de données pour des logiciels d'analyse statistique.

La trace de toutes les manipulations effectuées sur un corpus est enregistrée dans un journal cumulatif daté. On peut, par simple copier-coller des commandes ainsi tracées, constituer des fichiers de commandes appelées *scénarios*. Ces scénarios permettent d'automatiser des fonctions d'analyse et de traitement qui pourront par la suite être appliquées sur divers corpus.

SATO fonctionne en mode client-serveur au moyen d'une interface HTML standard. Le logiciel est accompagné d'un environnement de gestion HTML permettant de définir des comptes d'usagers, d'ouvrir des sessions qui pourront être servies en parallèle. Le système permet de constituer des banques de textes ainsi que des librairies de scénarios et de dictionnaires. L'interface HTML est modifiable à loisir pour créer des applications particulières dans diverses langues. Cette interface permet de jumeler SATO avec d'autres logiciels, des pages HTML et d'utiliser toute la puissance des langages de scriptage comme Perl, PHP, Python, etc.

Les requêtes envoyées par l'utilisateur à partir de son navigateur Web sont d'abord reçues par un programme général, une passerelle, qui gère le dialogue avec une application. Donc, la même passerelle qui dialogue avec SATO peut servir d'interface à tout autre programme qui lit un fichier de commandes et génère un fichier de résultats. Il est donc facile de rassembler autour de SATO une variété de modules informatiques qui seront déployés à la demande de l'utilisateur à partir de son navigateur Web. Ainsi, nous avons déjà mis au point une chaîne de traitement faisant appel au logiciel *Guidexpert* (Plante *et al.*, 2003) pour réaliser une description linguistique et sémantique de corpus. De même, nous prévoyons intégrer des logiciels statistiques et des systèmes de visualisation des résultats commandés par le chercheur dans son espace de travail privé à partir de son navigateur Web.

L'implantation du logiciel dans une architecture Web (SATO-HTML) a permis le développement d'une expertise dans le domaine des interfaces HTML et CGI (*common gateway interface)*. La nouvelle implantation SATO-XML a permis de produire une deuxième version de l'interface qui en augmente l'utilisabilité et qui supporte des interfaces multilingues. Aussi,

toute la partie qui consiste à donner accès au *bureau de travail* de l'usager sur le serveur a été complétée et revue de façon à la distinguer de l'usage du logiciel SATO lui-même. D'autres développements sont à prévoir afin d'exploiter les potentiels de filtrage et de transformation des textes en format XML.

Du point de vue interne au logiciel, la différence la plus importante entre SATO-XML et SATO-HTML sera le passage au jeu de caractères UNICODE, ce qui implique des filtres de conversion permettant de récupérer les données antérieures. Aussi, l'abandon du code hérité de la version DOS sera l'occasion d'augmenter diverses limites de traitement : dimension maximale des corpus, nombre de propriétés, attributs d'affichage et d'hyperliens, etc. Du point de vue de la syntaxe externe des corpus importés et exportés, la nouveauté a trait à l'utilisation de formats XML s'ajoutant au format propriétaire défini avant l'apparition des normes XML et SGML.

On pourrait qualifier la phase actuelle de développement du logiciel de phase de consolidation permettant de passer aux nouvelles normes XML et UNICODE. Ce passage se réalise dans le contexte d'une plateforme de type client-serveur basée sur une technologie Web standard facilitant le traitement coopératif entre logiciels indépendants s'échangeant des fichiers de données dans des formats standardisés. L'étape suivante consistera à ajouter un formalisme et des dispositifs de traitement permettant d'exploiter les relations structurelles tissées par le texte. Les relations les plus immédiates concernent la macrostructure de présentation du texte en sections emboîtées avec titres et renvois. Mais, elles concernent aussi les diverses constructions syntaxiques et stylistiques, les structures argumentaires, rhétoriques, dialogiques, et les divers liens marquant la cohérence textuelle. Ces dispositifs, étudiés par la linguistique textuelle (Adam, 1990), ainsi que la reconnaissance de la *macrostructure sémantique* des textes exigent des dispositifs informatiques de *catégorisation structurelle,* par analogie à la catégorisation simple que nous pratiquons actuellement. L'objectif à plus long terme est donc d'exploiter pleinement les relations entre les segments textuels dans un tracé de *lecture-explicitation* ou dans des analyses lexicales sensibles aux marques de structure.

4. Exemples d'utilisation de la plateforme

Les paragraphes qui suivent illustrent quelques moments d'une analyse réalisée à l'aide de SATO-XML dans son état actuel de développement. Dans notre exemple, nous avons choisi les communiqués de presse en langue anglaise produits par trois groupes de défense des animaux : *World Wildlife Fund*, *Sea Shepherd* et *Greenpeace*. Ces communiqués concernent la levée du moratoire sur la pêche à la morue par l'Union Européenne (en décembre 2002) et l'annonce de la reprise de la pêche à la baleine par l'Islande (en août 2003). Comme ces communiqués ont été émis durant la même période par des groupes différents, mais s'adressant aux mêmes personnes (les membres des groupes, le public en général, les médias ainsi que les organisations mises en cause), ils permettent au chercheur de supposer les groupes assis autour d'une même table installée dans un espace délibératif à l'échelle mondiale, un espace où la production textuelle joue un rôle de premier plan.

Nous avons sélectionné un texte par groupe et par thème (baleines et poissons), soit six textes au total. Il existe pour l'utilisateur deux façons d'envoyer ses textes vers l'espace disque qui lui est réservé sur le serveur : soit à l'aide d'un formulaire disponible dans l'interface du bureau Web de SATO, soit par FTP (*File Transfer Protocol*). L'accès aux textes demeure privé, c'est-à-dire que ces derniers ne sont accessibles qu'à leur propriétaire qui pourra cependant décider de les partager en mode lecture avec d'autres membres de son groupe ou en autoriser le dépôt dans une librairie publique accessible à tous.

Les textes retenus résidant sur le serveur, nous pouvons créer un corpus à l'aide d'un formulaire HTML. Le contenu du corpus sera déterminé par une référence à chacun des six fichiers

contenant les communiqués de presse. SATO en produira alors une représentation sous la forme d'un plan lexique-occurrences. Le logiciel tiendra compte des annotations du chercheur distinguant, par exemple, les diverses parties constitutives des textes : auteurs, titres, sections, etc. L'image 1 (cf. annexe 1) illustre la procédure de soumission d'un corpus.

Cette photo d'écran donne un aperçu de l'interface du bureau sur le serveur. À gauche se trouve le menu. Si on clique sur un item suivi d'un +, on développe les sous-items. Dans cet exemple, on clique sur l'item *Soumission.* La section centrale de l'écran présente la partie supérieure du formulaire de soumission d'un corpus. La section du bas est la bannière d'identification. Pour faire suite à la soumission du formulaire, SATO génère le corpus et passe dans la section analyse du logiciel. Pour les sessions ultérieures, on entrera directement dans la section analyse en choisissant l'item *Corpus personnel* sur le bureau.

Une première manière d'explorer le corpus est d'afficher le lexique des formes lexicales. Dans l'illustration qui suit, nous présentons un lexique ventilé par organisme et par thème. L'image 2 présente l'interface de commande et le formulaire d'affichage du lexique. À gauche, on retrouve le menu de commandes de SATO. En cliquant sur l'item principal *lexique* suivi d'un +, on obtient le formulaire *Afficher* dans la section centrale de l'écran. Le champ *filtre* reçoit alors le patron *$*Freqtot<50>5* qui indique que seuls les items dont la fréquence totale est inférieure à 50 et supérieure à 5 seront retenus. Dans le champ *Tri*, la propriété *Fréqtot* est sélectionnée afin d'ordonner le lexique par la fréquence totale dans le corpus.

L'image 3 (cf annexe 1) présente le résultat de la soumission du formulaire précédent. Outre la colonne indiquant la fréquence totale (Fréqtot), on peut voir une colonne pour chacun des groupes (WWF pour *World Wildlife Fund*, SEA pour *Sea Sheperd* et GRE pour *Greenpeace*), ainsi que la distribution lexicale selon les deux thématiques concernant la pêche à la baleine (BALEINES) et la pêche à la morue (POISSONS). Dans la dernière colonne se trouve la forme lexicale. Dans la partie inférieure de la photo d'écran, on a le journal qui garde la trace de toutes les opérations effectuées durant la session de travail. De plus, la trace cumulative de chaque session est conservée dans le journal associé au corpus. On peut, par simple copier-coller de commandes reproduites dans le journal, construire des scénarios de commandes qui pourront être appliqués à loisir sur le même corpus ou tout autre corpus.

Si on clique sur une forme lexicale, on dévoile dans la fenêtre du bas un menu de catégorisation que nous retrouvons dans l'image 4 de l'annexe 1. La partie droite de la photo d'écran révèle chacune des propriétés associées au mot retenu, ici le nom propre *Lieberman.* On y trouve notamment la propriété *nature_entité* ajoutée en cours d'analyse pour décrire la nature des acteurs sociaux : *technocrates*, *animal*, *protecteurs*, *public*, *autres en faveur des animaux*, *médias*, *scientifiques* et *pêcheurs*. La partie gauche de l'écran de catégorisation contient un menu permettant d'accéder aux contextes courts (KWIC) du mot cliqué, de le catégoriser, de sauvegarder les annotations, etc.

Cette catégorisation sémantique permet de visualiser la fréquence des différents types d'acteurs et leur répartition dans les divers textes. L'image 5 de l'annexe 1 montre un affichage du texte avec mise en évidence des mots correspondants à des acteurs sociaux. À l'écran, les mots sont de couleur différente en fonction de chacune des valeurs de la propriété *nature_entité.* Ces valeurs décrivent les différents acteurs qui s'opposent et s'allient dans le discours des groupes de défense des animaux pour les deux thèmes choisis. Les catégories ont été établies à partir du lexique, mais elles peuvent aussi être désambiguïsées à la lecture du mot en contexte (KWIC). Par exemple, un même acteur peut être considéré, selon le contexte, comme un scientifique ou comme un protecteur.

Un affichage du lexique des catégories d'acteurs (non illustré ici), trié en fonction des fréquences cumulées par groupe et pondéré par la taille du texte, montre que les catégories

d'acteurs qui distinguent le plus les groupes entre eux sont celles d'*autres en faveur des animaux* et de *public* (sur-représenté dans les textes de GRE), de *protecteurs* (sur-représenté dans les textes de WWF) et de *technocrates* (sous-représenté dans les textes de WWF).

Il apparaît, après l'affichage du texte catégorisé et coloré (non illustré ici) en fonction des différentes valeurs de la propriété *nature_entité*, que le discours de *Greenpeace* met en scène le plus grand nombre d'acteurs, correspondant à l'ensemble des valeurs de la propriété et répartis également dans le sous-corpus. Quant au *World Wildlife Fund*, reconnu comme le moins radical des trois groupes, il insiste dans son texte sur la morue, sur les *protecteurs* des animaux, la présence des autres acteurs n'étant que suggérée alors que le texte concernant les baleines mentionne le *public* (*community, people, consumers*). *Sea Shepherd* n'évoque, dans son texte sur les baleines, que les *pêcheurs* (*fleet*, *whalers*) confrontés à l'action directe du groupe. Le texte sur la morue est moins menaçant et moins direct et le nombre d'acteurs mentionnés s'accroît. L'opposition mise en évidence par le groupe se trouve cette fois entre le *public* et les *gouvernements*. Ces données confirment le radicalisme reconnu du groupe qui n'hésite pas à saborder les baleiniers.

Divers lexiques d'occurrences ou de cooccurrences peuvent être générés en fonction des critères de partition du corpus. Plusieurs tableaux ont été produits par des analyseurs statistiques simples appliqués sur le corpus. Les limites de cet article ne permettent pas de les reproduire ici. Il s'agissait plutôt d'illustrer quelques moments d'une analyse et de mettre en lumière les possibilités d'une plateforme Internet ouverte. On pourra, par ailleurs, consulter une démonstration en ligne sur le site web de la chaire MCD et du Centre ATO de l'UQAM.

Références

Adam J.-M. (1990). *Éléments de linguistique textuelle, Théorie et pratique de l'analyse textuelle*. Mardaga.

Brunet Ét. *Logiciel HYPERBASE (version 2.3*), http://ancilla.unice.fr/~brunet/pub/hyperbase.html site visité le 7 janvier 2004.

Daoust F. (1996). *SATO 4, Manuel de référence*. Centre d'analyse de texte par ordinateur, UQAM.

Daoust F. (2002). L'analyse de texte assistée par ordinateur, lunette de lecture des textes électroniques. *Communication présentée au colloque Publications et lectures numériques : problématiques et enjeux, 70ième congrès de l'ACFAS*, EBSI, Montréal. http://www.ebsi.umontreal.ca/rech/acfas2002/daoust.pdf

Duchastel J. (1993). Discours et informatique : des objets sociologiques ? *Sociologie et sociétés*, vol. (25/2) : 157-170.

Duchastel J. (2001). *Présentation du projet ATO-MCD.* http://ato.chaire-mcd.ca/presentation/

Duchastel J. et Armony V. (1996). Textual Analysis in Canada: An Interdisciplinary Approach to Qualitative Data. *Current Sociology*, vol. (44/3) : 259-278.

Heiden S. (2002). *Weblex, Manuel Utilisateur, version 4.1 intermédiaire*, http://lexico.ens-lsh.fr/doc/weblex.pdf https://weblex.ens-lsh.fr/ sites visités le 7 janvier 2004.

Maingueneau D. (1997). *L'Analyse du Discours*. Hachette.

Pêcheux M. (1994). Sur les contextes épistémologiques de l'analyse du discours. *Mots*, vol. (9). Presses de la Fondation nationale des sciences politiques.

Plante P. *et al.* (2003). Guidexpert ATO. http://fable.ato.uqam.ca/guidexpert/guidexpert-ato-wp.htm

Rastier F. (1989). *Sens et textualité*. Hachette.

Rockwell *et al.* (2002). *TAPoR: Text-Analysis POrtal for Research*. http://huco.ualberta.ca/Tapor/ site visité le 7 janvier 2004.

Salem A., Lamalle C., Martinez W. et Fleury S. *Lexico 3*. http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/ site visité le 7 janvier 2004.

The TEI Consortium (2001). *Text Encoding Initiative, The XML Version of the TEI Guidelines*. In Sperberg-McQueen C.M. et Burnard L (Eds).

Annexe 1 : Photos d'écran

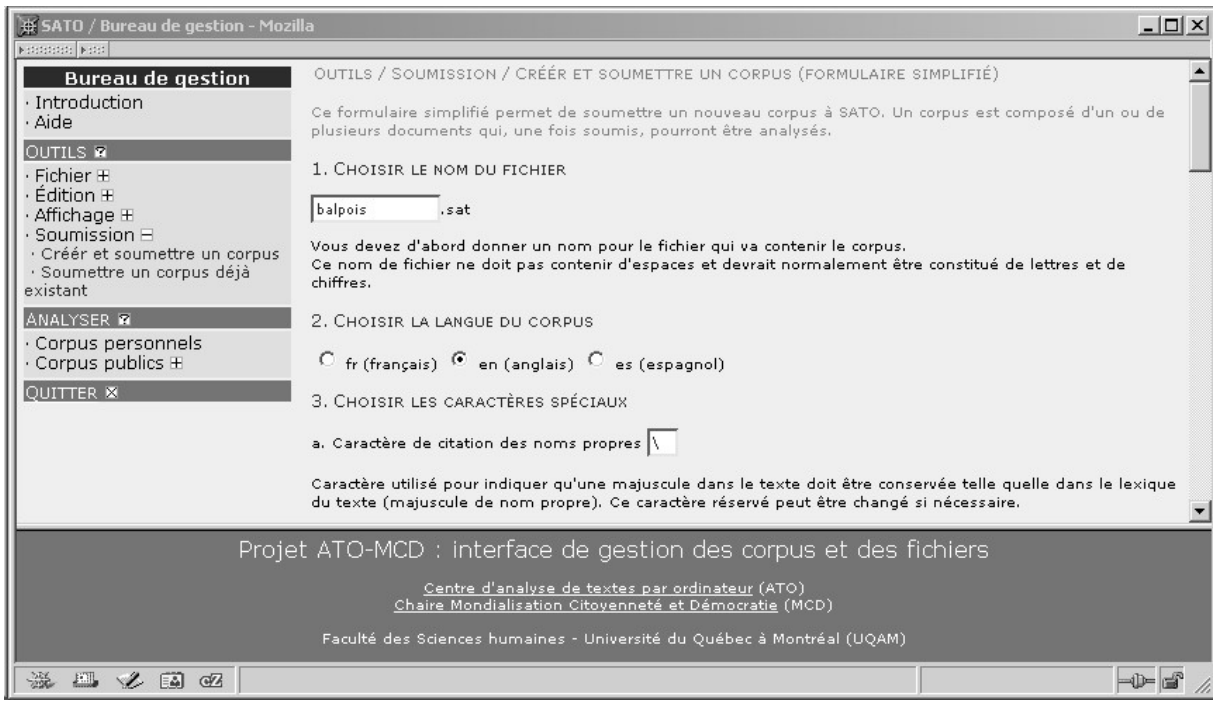

Image 1. Soumission d'un corpus

Image 2. Formulaire d'affichage du lexique

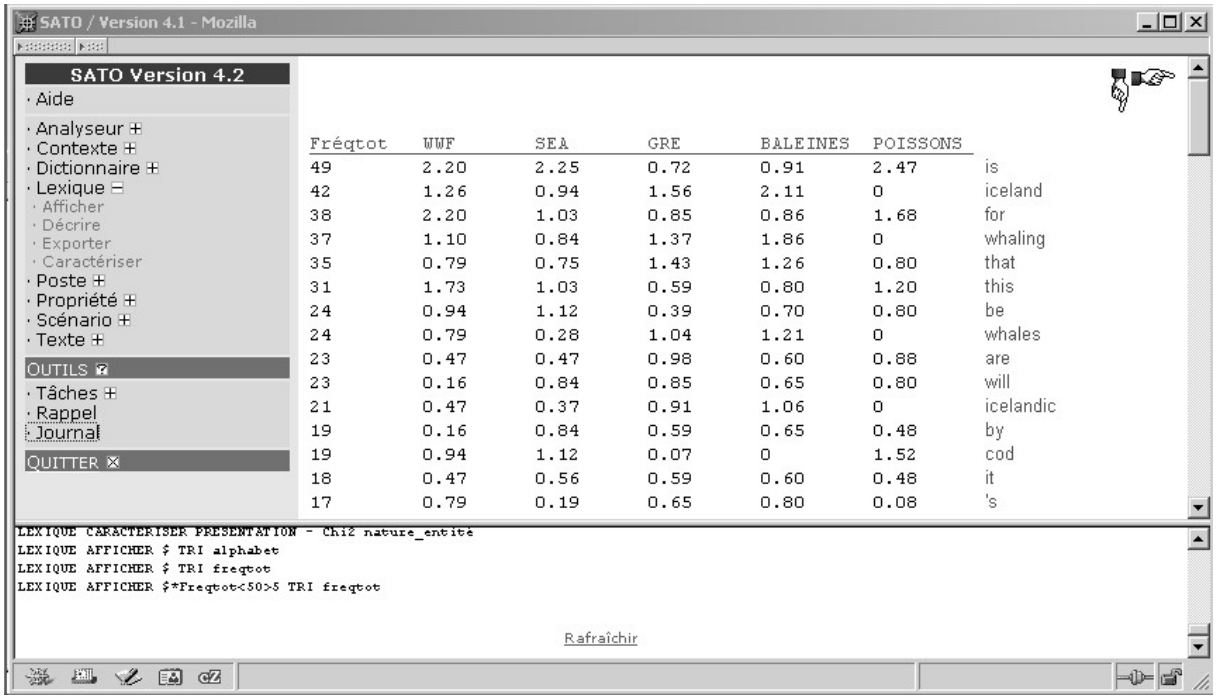

Image 3. Affichage du lexique et du journal

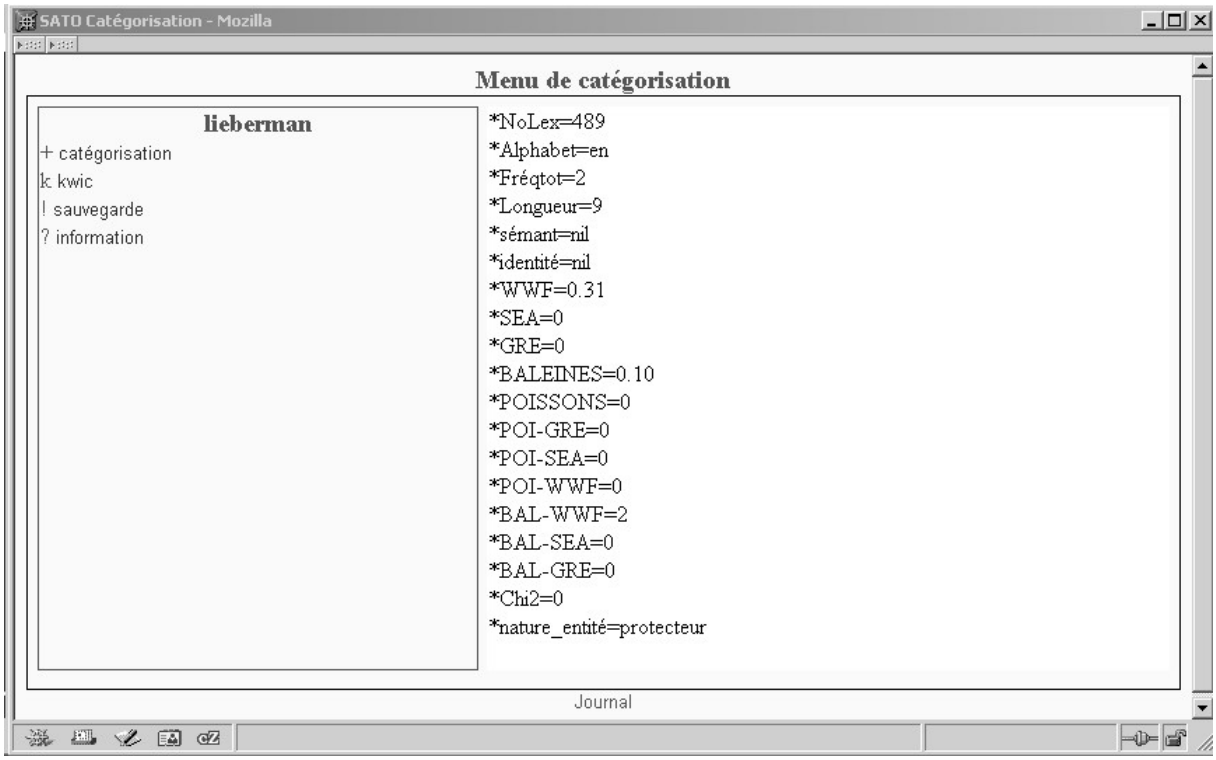

Image 4. Menu de catégorisation avec l'information

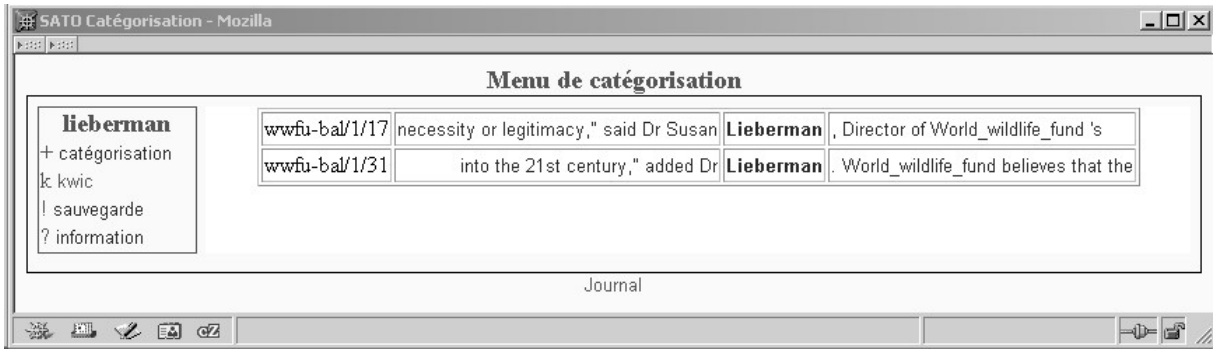

Image 5. Menu de catégorisation avec le KWIC

Image 6. Affichage du texte avec mise en couleur des acteurs sociaux