Une stratégie intégrée de recherche en sciences humaines dans le Portail ATO-MCD

Jules Duchastel[1], Francis J. Lacoste[2], François Pizarro Noël [3]

[1]Titulaire de la Chaire de recherche du Canada en Mondialisation, Citoyenneté et Démocratie, Professeur titulaire, département de sociologie, Université du Québec à Montréal
[2]Chercheur-programmeur à la Chaire MCD – francis@contre.com
[3] Responsable méthodologique à la Chaire MCD – francois@contre.com

Abstract

This paper presents the ATO-MCD Portal, a Web-based cooperative environment for research projects in content and discourse analysis. The paper is a demonstration of this environment. As such, it showcases the functionalities of the Portal, which are designed to work on corpus and develop analytical grids. It also showcases statistical tools for the analysis of frequency tables. The paper explains how those functionalities and tools are embedded in a Project Book that facilitates coordination between researchers and helps to document the research process. We will present its functionalities designed for the treatment of the corpus and the construction of analytical frameworks. We'll then present the statistical tools meant for the analysis of huge frequencies tabs. The presentation of the paper will be a presentation of this research environment.

Résumé

Cet article présente le Portail ATO-MCD, un environnement Web coopératif pour des projets de recherche en analyse de contenu. Nous présentons les fonctionnalités du Portail pour le traitement des corpus, la construction de grilles d'analyse et les outils statistiques pour l'analyse des tableaux de fréquence. Nous montrons comment ces fonctionnalités sont intégrées autour d'un Cahier de projet qui facilite la coordination entre les chercheurs et favorise la documentation du processus de recherche. La présentation consistera en une démonstration de l'environnement.

Mots-clés : analyse de texte par ordinateur, analyse de contenu, analyse factorielle des correspondances.

1. Introduction

Le Portail ATO-MCD est un environnement coopératif de recherche réalisé par la Chaire de recherche du Canada en mondialisation, citoyenneté et démocratie, le GRADiP (Groupe de recherche en analyse du discours politique) et le centre ATO (Centre d'analyse de texte par ordinateur) dans le cadre du projet ATO-MCD (http://ato.chaire-mcd.ca/). Le projet comporte deux volets. Le premier volet documentaire met à la disposition du grand public et de la communauté de chercheurs un grand nombre de corpus reliés à l'univers du discours politique. La banque de données initiale est composée des corpus constitués au cours des années dans les recherches du GRADIP. En effet, les corpus DNL (Discours Néo-Libéral : documents provenant des gouvernements élus, des syndicats, des conseils patronaux, des Églises et des organismes gouvernementaux), EDM (Espace Délibératif Mondial : comportant des textes des grandes organisations internationales) et CCC (Corpus Constitutionnel Canadien : discours des Premiers ministres fédéraux et provinciaux lors des conférences constitutionnelles, 1940-1992) sont dores et déjà disponibles sur le Portail. À la différence d'une banque tex-

tuelle classique, non seulement le texte des documents constituant les corpus sont accessibles, mais les grilles d'analyse et les notes des chercheurs sont parties intégrantes de ces banques.

Outre ce volet « accessibilité aux fruits de la recherche », le projet ATO-MCD possède aussi un volet « instruments de recherche ». En ce sens, il se veut un environnement coopératif qui intègre à l'intérieur d'une même interface un ensemble d'outils méthodologiques pour l'ATO. Les outils méthodologiques intégrés dans le projet correspondent à la perspective d'« analyse de contenu » privilégiée dans les recherches de la Chaire[1].

Nous insisterons ici sur la dimension « environnement coopératif de recherche ». Nous aborderons tout d'abord les principes généraux qui ont orienté la conception de cet environnement. Afin d'accroître l'aspect coopératif, nous avons renforcé les fonctionnalités favorisant une bonne documentation du processus de recherche. Les fonctionnalités reliées à la gestion des documents et des corpus seront ensuite présentées avant d'aborder les différents outils pour le traitement des données et la production d'analyses. Nous terminerons en esquissant quelques orientations futures pour le développement du Portail.

Pour illustrer notre démarche, nous présenterons l'utilisation que nous avons faite du Portail pour mener à bien un projet de recherche portant sur les discours d'ouverture des sessions législatives par les différents Premiers ministres du Québec entre 1960 et 2003. Le but de cette recherche était d'émettre une « opinion informée » dans la polémique sur la reprise de la « Révolution tranquille » telle qu'évoquée par le nouveau Premier ministre libéral, M. Charest, dans son discours d'ouverture de la session législative à l'automne 2003.

Le projet « Ouverture » nous aura permis de développer une stratégie intégrée de recherche dans le cadre du Portail ATO-MCD en fonction de nos besoins réels de chercheurs. Cette étude d'ampleur restreinte nous a permis ainsi de couvrir chacune des étapes du processus de recherche. Cela nous a conduit à mieux évaluer l'efficacité et la pertinence des outils que nous souhaitons mettre en œuvre sur le Portail.

Le projet Ouverture porte sur les discours d'ouverture tenus à l'Assemblée Nationale du Québec, de la première législature de Jean Lesage en 1960 jusqu'au discours de Jean Charest en 2003. Le corpus DNL comprenait déjà les discours d'ouverture prononcés entre 70 et 84. Il suffisait de compléter le corpus en y incorporant les discours de la décennie 1960 et ceux de la période 1985-2003.

Le corpus additionnel provient de deux sources documentaires :

1. Avant 1964, les *Journaux de l'Assemblée Législative de la Province de Québec* qui contiennent une reconstitution des débats parlementaires. Le discours d'ouverture, prononcé au nom du gouvernement par le Lieutenant Gouverneur, étant écrit, nous sommes assurés de la fidélité relative du discours rapporté dans cette publication.

2. À partir de 1964, le *Journal des débats* publie le verbatim des débats.

[1] La stratégie de recherche présentée ici correspond à une configuration particulière des outils disponibles dans le Portail ATO-MCD mettant en coopération trois composantes logicielles. On trouvera, dans le chapitre intitulé : « SATO_XML : une plateforme Internet ouverte pour l'analyse de texte assistée par ordinateur », la présentation générale de l'architecture informatique aux divers plans des données, des traitements informatiques et de l'infrastructure matérielle. Ce texte se concentre sur la ressource logicielle principale du projet ATO-MCD, soit le logiciel SATO-XML.

2. Principes pour un environnement coopératif

Le premier objectif d'un environnement coopératif de recherche est de permettre à plusieurs chercheurs de travailler simultanément au même projet de recherche. Pour qu'un tel environnement puisse remplir efficacement cet objectif, il doit offrir des instruments facilitant la coordination des actions des membres d'une même équipe. Afin de rencontrer cet objectif, l'ensemble des outils de traitement et d'analyse offerts s'articulent autour d'un Cahier de projet (Figure 1 en annexe). Les deux éléments principaux de celui-ci sont le Relevé des opérations et le Journal de bord. Le premier est le journal informatique des opérations effectuées sur le site par chacun des chercheurs du projet. Chacune des opérations produites par les chercheurs est consignée dans le Relevé des opérations. Quant au Journal de bord, il permet aux chercheurs de documenter leurs décisions stratégiques et d'en consigner la justification. Chacune des entrées du Journal de bord peut être reliée aux opérations informatiques afférentes du Relevé des opérations. Par exemple, le Journal de bord sera l'endroit où l'on pourra lire que le 10 novembre Guillaume a travaillé à la construction de la grille thématique. Cette entrée est liée aux opérations qui ont permis la construction de cette grille. On voit comment, en plus de favoriser la coordination, l'intégration autour d'un Cahier de projet favorise aussi la documentation et la *traçabilité* des opérations, indispensables pour la réussite d'une démarche de recherche intégrée. Le Cahier de projet est aussi le lieu où sont consignés l'ensemble des artefacts produits par les différents outils de traitement et d'analyse (AFC, tableaux statistiques, corpus, fragments annotés, etc.). Le cahier permet de garder une trace de tout le processus de recherche et de le reconstituer en tout temps. De plus, le moment venu, il permet de présenter l'ensemble du processus et de mieux cerner chacune de ses étapes.

C'est aussi pour permettre le travail coopératif de manière distribuée à plusieurs endroits qu'une interface web a été retenue. Ce choix possède plusieurs avantages pour le volet « diffusion » du projet. Enfin, l'interface web offre la possibilité d'utiliser le logiciel sur n'importe quelle plate-forme informatique (Mac, PC ou Linux).

Le second objectif de l'environnement coopératif est la « collaboration » informatique qui permet de fédérer, dans un environnement de recherche cohérent, un ensemble de composantes informatiques spécialisées. Dans cette première version, trois composantes spécialisées sont intégrées :

1) la composante SATO (http://nouvelle.ato.uqam.ca/) est utilisée pour tout ce qui relève du traitement textuel : établissement de sous-corpus contextuels, constitution de « tableaux lexicaux », catégorisation en contexte, etc. ;

2) la composante Guidexpert (http://fable.ato.uqam.ca/guidexpert/guidexpert.htm) est utilisée pour l'attribution de catégories grammaticales, la lemmatisation et certaines tâches de catégorisation thématique semi-automatiques ;

3) la composante R (http://www.r-project.org/) est utilisée pour tous les traitements statistiques (par exemple, les AFC).

L'intégration de ces composants est réalisée dans un système conçu à l'aide de l'environnement web Plone (http://www.plone.org/). Bien que chacune de ces trois composantes possède une interface propre qui demeure accessible à partir du Portail, l'idée centrale de la stratégie intégrée de recherche est de proposer une interface transversale qui intègre les spécialités de chacune des composantes à des instruments de documentation et de coordination du travail de recherche.

3. Construction et traitement des corpus

Sur le Portail, un corpus est l'ensemble des documents, enrichis des descriptions produites par l'application de grilles de catégories propres à chaque projet. Les chercheurs sont libres de définir leurs propres corpus et grilles analytiques ou de retenir des corpus et des grilles déjà existantes. Par exemple, le corpus CCC est composé d'un ensemble de documents reliés aux conférences constitutionnelles canadiennes et d'une grille de catégorie socio-sémantique. Ces documents et/ou cette grille pourraient être réutilisées dans un nouveau contexte de recherche. Dans la mesure où un chercheur désire constituer un nouveau corpus, il lui suffit d'envoyer vers le Portail la copie numérique des documents qu'il souhaite analyser.

Dans le cas du projet Ouverture, nous avons réutilisé des textes déjà disponibles dans le corpus (DNL) et nous avons numérisé les documents additionnels pour répondre à de nouvelles questions de recherche. Nous avons réuni ces documents en format *texte simple* (.txt) et nous les avons téléchargés sur le Portail dans le dossier que nous avions créé à cet effet. Lors du téléchargement des documents, en plus des informations de nature référentielle telles que le titre du document ou la référence bibliographique, certaines variables qui serviront à partitionner le corpus (par exemple, le moment de l'énonciation ou le type de locuteur) ont été spécifiées à l'aide de balises dans le texte numérisé. Dans le projet Ouverture, chaque document utilisait comme variable l'année, la session, la législature, le parti et le Premier ministre. Il aurait été possible d'ajouter d'autres variables et de cumuler plusieurs modalités d'une variable sur un même segment textuel. L'ensemble des variables ne recevront pas nécessairement de traitement statistique au moment de l'analyse. Enfin, des variables supplémentaires peuvent toujours être définies à une étape ultérieure de la recherche. Les outils servant à la construction de ces variables seront abordés dans la suite de l'exposé. L'ensemble des variables et de leurs modalités définies sur un corpus est accessible dans le Cahier de projet.

Les transformations dans la représentation informatique des corpus ainsi que les procédures d'indexation nécessaires pour leur traitement par les composantes Guidexpert ou SATO sont transparentes pour les chercheurs. Dans le cas où les fonctionnalités des différentes composantes utilisées par le Portail ne seraient pas jugées suffisantes, les chercheurs pourront exporter leur corpus dans un format qui permettra son traitement selon l'ensemble des fonctionnalités propres à chaque logiciel. En fait, à partir des informations fournies lors du téléchargement des documents du corpus, des versions du corpus adaptées aux différents logiciels sont automatiquement générées. Des versions des corpus utilisables directement dans SATO, Guidexpert ou d'autres logiciels fédérés sur le Portail sont toujours disponibles. Les usagers peuvent donc utiliser ces corpus adaptés aux logiciels de leur choix à leur gré pour réaliser des opérations supplémentaires.

Une fois le corpus complété, les chercheurs peuvent amorcer l'analyse en utilisant les divers outils pour construire différentes grilles descriptives ou produire des analyses statistiques à partir des tableaux générés à l'aide du système.

4. Outils pour la construction de grilles analytiques

À l'intérieur du Portail, une « grille descriptive» réfère à une représentation particulière des données lexicales ou textuelles sur laquelle des analyses statistiques pourront être effectuées. L'analyse la plus simple porte sur la distribution des formes lexicales dans différentes partitions du corpus. Il est possible cependant de travailler sur les formes lemmatisées produites à l'aide de Guidexpert. Les recherches du GRADIP ont plutôt privilégié les catégories socio-sémantiques appliquées aux noms communs et aux adjectifs (Duchastel et Armony, 1995).

Dans tous les cas, c'est la distribution des mots catégorisés ou non qui fera l'objet d'analyses statistiques.

La composante Guidexpert comporte des connaissances linguistiques et sémantiques du français et de l'anglais permettant de traiter une version lemmatisée du lexique et d'assigner automatiquement les fonctions syntaxiques aux formes lexicales. Cette description morpho-syntaxique peut-être utile pour limiter l'étendue de la catégorisation en contexte ou pour établir des sous-ensembles textuels, par exemple, pour une analyse de la distribution des adjectifs par locuteur. Dans le projet Ouverture, nous avons eu recours à Guidexpert afin de restreindre l'analyse aux noms et aux adjectifs. Nous avons utilisé les capacités linguistiques du logiciel Guidexpert pour assigner automatiquement une catégorie syntaxique à chaque mot du corpus. Par la suite, nous avons transféré ces informations dans la version SATO du corpus (Figure 2 en annexe).

Il n'existe pas de protocole universel de création ou d'application de ces grilles. Plusieurs approches existent. Une des possibilités est d'appliquer un dictionnaire de catégories (BDL, Banque de données lexicales) et de désambiguïser en contexte les lexèmes recevant plus d'une catégorie à l'aide de SATO. On peut aussi utiliser l'interface de Guidexpert pour explorer le corpus et construire de manière inductive une grille thématique à partir des champs sémantiques les plus fréquents. Pour ce faire, on crée une grille de thématisation qui associe à des modalités d'une variable thématique un ensemble de formes lexicales et de champs sémantiques larges ou restreints qui sont repérés automatiquement par Guidexpert. Quelle que soit la méthode utilisée pour construire cette thématisation, elle sera documentée dans le Cahier de projet et appliquée à partir du Portail. Cette thématisation pourra faire l'objet de traitements statistiques ultérieurs, si la variable répond aux propriétés requises. Dans la suite de l'exposé, nous ne reviendrons pas sur les fonctionnalités de thématisation sémantique de Guidexpert faute d'espace.

5. Outils pour l'analyse statistique

SATO permet de construire des tableaux représentant la distribution des unités lexicales (formes lexicales ou lemmatisées) ou d'une variable (par exemple, des catégories socio-sémantiques) à travers une partition donnée du corpus. Par exemple, on peut obtenir le tableau de la distribution des formes lexicales par locuteur ou le tableau de la distribution des catégories socio-sémantiques par période. Dans le cas du projet Ouverture, nous avons construit à l'aide de SATO le tableau de la distribution des lexèmes distingués selon leur fonction grammaticale par législatures (Figure 3 en annexe).

Le principe de partition que nous avons retenu pour ce tableau est la variable législature, c'est-à-dire la période d'exercice d'un gouvernement d'une élection à l'autre. La stricte chronologie annuelle ne permettait pas une analyse en fonction des partis ou des locuteurs, puisque, certaines années, deux Premiers ministres provenant de partis différents ont prononcé un discours d'ouverture. La variable Premier ministre ne permettait pas de distinguer les différents mandats d'un même Premier ministre. La variable Parti taisait l'aspect diachronique que nous privilégiions en analysant un corpus recouvrant plus de 40 ans. Évidemment, la variable législature nous permettait de retenir à la fois toutes les informations concernant le Premier ministre, le Parti et l'année.

Dans l'environnement SATO, ces tableaux peuvent être construits aussi bien sur l'ensemble du corpus que sur la distribution d'une unité ou d'une variable dans un sous-ensemble du corpus. La forme que peut prendre chaque tableau est variable. Par exemple, il est fréquent de s'intéresser à la distribution des « cooccurrences » ou, plus précisément, aux diverses formes

lexicales qui apparaissent dans le contexte (une ou plusieurs phrases, une borne numérique, etc.) d'une forme particulière.

Les tableaux peuvent être filtrés selon divers critères paramétrables par les chercheurs afin, par exemple, d'éliminer les hapax ou de ne retenir que les formes dont la fréquence totale est supérieure à la médiane. Tous ces tableaux peuvent être sauvegardés dans le Cahier de projet et peuvent être annotés. Dans le projet Ouverture, nous avons filtré le tableau en ne retenant que les noms et adjectifs dont la fréquence était supérieure à 10. Ainsi, nous avons expurgé les textes des mots outils pour ne garder que les mots présentant un fort contenu référentiel. En ce sens, nous sommes restés fidèles à la méthodologie du GRADIP : « 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. » (Duchastel et Armony, 1995). En ne retenant que les noms et les adjectifs présentant plus de dix occurrences dans l'ensemble du corpus, nous sommes passés de 13034 lexèmes à 1042 lexèmes. Au terme de toutes ces manipulations, nous traitons 74 % des adjectifs et des noms communs, soit 19% du corpus total.

L'instrument d'analyse statistique privilégié pour ces volumineux tableaux est l'analyse factorielle des correspondances (Lebart et Salem, 1994). Cette analyse permet de visualiser les variations dans la distribution des unités entre les parties du corpus (locuteurs, périodes, etc.) Outre les vertus heuristiques de cette visualisation, le choix privilégié de cette méthode d'analyse se justifie par ses multiples propriétés qui la rendent robuste vis-à-vis des perturbations dans les données. (Viprey 2003 ; *Lebart et al.*, 2000). Pour le projet Ouverture, cette méthode était toute indiquée puisque les discours varient beaucoup en taille. Les AFC ont été réalisées à l'aide de R sur le tableau construit précédemment.

Le Portail offre plusieurs outils pour l'aide à l'interprétation des AFC. Outre la possibilité de visualiser dans un même plan deux axes de l'AFC (on s'intéressera la plupart du temps aux diverses combinaisons des premiers axes), plusieurs fonctionnalités permettent de pallier la difficulté de visualiser la projection de plusieurs centaines d'éléments dans le plan. Par exemple, on pourra limiter la représentation des points-lignes à celles qui répondent à certains critères comme la fréquence. Pour affiner l'interprétation, on pourra aussi afficher dans une teinte différente les éléments qui sont particulièrement bien représentés dans le plan (seuil paramétrable applicable au cosinus carré) ou ceux qui ont une importante contribution dans la construction d'un des axes du plan. Ce sont ces stratégies que nous avons adoptées pour le projet Ouverture. Nous avons limité l'affichage des points-lignes aux « meilleurs » de chacun des axes projetés dans le plan, c'est-à-dire les points lignes ayant la meilleur position sur l'axe (COS2) jusqu'à ce que l'ensemble sélectionné représente 60% des contributions à l'axe. Dans le graphique présenté en annexe (Figure 4), les points-lignes affichés sont l'union des ensembles liés à chaque axe projeté sur le plan.

Encore une fois, les différentes visualisations produites peuvent être sauvegardées dans le Cahier de projet. Comme tous les éléments sauvegardés dans le Cahier de projet, la visualisation particulière reste aussi liée aux objets dont elle est issue afin de pouvoir retracer l'origine de sa construction**.**

6. Conclusion

Le Portail ATO-MCD se veut un environnement permettant la recherche coopérative. L'élaboration de la stratégie de recherche proposée est axée sur l'intégration des différents outils de construction des corpus, de traitement des données et d'analyse statistique autour d'un Cahier de projet qui permet la coordination, mais surtout la documentation du processus

de recherche. En ce sens, le Portail propose de rendre possible et/ou de faciliter l'application de méthodologies « courantes » dans le domaine de l'analyse de texte par ordinateur. Les concepts, méthodes et outils qu'il rend ainsi accessibles sont bien connus des praticiens de l'ATO. Ainsi, le Portail est un environnement qui intègre ces méthodes éprouvées de manière cohérente, favorisant ainsi le développement d'un « communauté argumentative » de chercheurs en ATO. De cette volonté découle le choix d'organiser l'interface usager autour du Cahier de projet, qui facilite la documentation et la justification des décisions inhérentes à tout processus de recherche. Le Portail permet donc le développement d'outils de coopération sur deux plans. Tout d'abord, il permet la coopération et le suivi des opérations par plusieurs chercheurs sur un même projet. Ensuite, il permet l'intégration des fonctionnalités de plusieurs logiciels d'analyse de texte et de traitement des données.

Dans l'avenir, les développements du Portail porteront sur l'intégration d'autres types d'outils analytiques éprouvés. La priorité sera donnée à l'intégration d'analyses statistiques qui complémenteront l'AFC. La classification hiérarchique et le calcul des spécificités sont des exemples d'instruments statistiques qui pourraient enrichir l'éventail des outils disponibles sur le Portail (Lebart et Salem, 1994). Une attention particulière sera aussi portée à l'intégration des méthodes pouvant être utilisées sur des structures de graphes pour l'étude des cooccurrences (Batagelj et Mrvar, 2003).

Références

Batagelj V. et Mrvar A. (2003). Developing Pajek - Exploratory analysis of networks. Analyse des Données Relationnelles / EHESS-INED. INED. http://vlado.fmf.uni-lj.si/pub/networks/doc/seminar/paris03.pdf.

Duchastel J. et Armony V. (1995). La catégorisation socio-sémantique. In *Actes des JADT 1995 :*193-200.

Lebart L. et Salem A.. (1994). *Statistique textuelle*. Dunod.

Lebart L., Morineau A. et Piron M. (2000) *Statistique exploratoire multidimensionnelle. 3e édition.* Dunod.

Viprey J.-M. (2003). *Morneille, Colière et messieurs Labbé*. http://laseldi.univ-fcomte.fr/morneille.htm

Viprey J.-M. (2002) : *Analyses textuelles et hypertextuelles des Fleurs du Mal* [texte intégral et moteur de recherche sur CD-Rom; exploration lexicale, morpho-syntaxique, prosodique, phonématique], Champion, Lettres Numériques n°5.

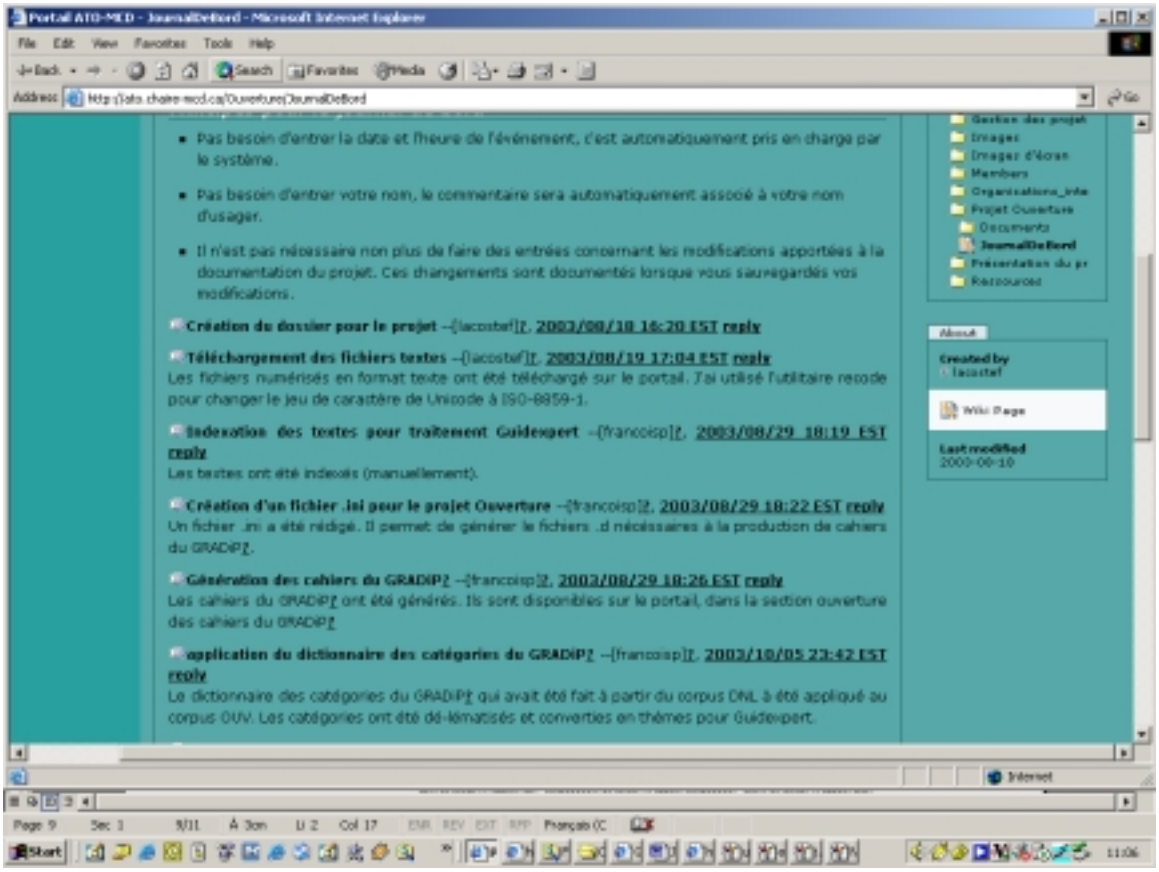

Figure 1. Cahier de Projet

```
1  *en=al960¶
2  *legis=26¶
3  *parti=plq¶
4  *pm=lesage¶
5  *session=2¶
6  *Phrase=3 Première^Cs=|adj|^Lemme="premier" tâche^Cs=|nc|^Lemme="tâche" du^Cs=|prep|^Lemme="du"
   gouvernement^Cs=|nc|^Lemme="gouvernement" a été de^Cs=|prep|^Lemme="de" déterminer^Cs=|vinf|^Lemme="déterminer" les
   besoins^Cs=|nc|^Lemme="besoin" les plus urgents^Cs=|adj|^Lemme="urgent"¶
7  de^Cs=|prep|^Lemme="de" la province^Cs=|nc|^Lemme="province".^Cs=|ponct|^Lemme="." *Phrase=4 Il a
   déjà^Cs=|adv|^Lemme="déjà" adopté^Cs=|vppa|^Lemme="adopter" des^Cs=|prep|^Lemme="des"
   initiatives^Cs=|nc|^Lemme="initiative" concrètes^Cs=|adj|^Lemme="concret".^Cs=|ponct|^Lemme="." *Phrase=5
   Au^Cs=|prep|^Lemme="au" cours^Cs=|prep|^Lemme="cours" de^Cs=|prep|^Lemme="de" la¶
8  session^Cs=|nc|^Lemme="session" présente^Cs=|adj|^Lemme="présent", il se propose^Cs=|vfle|^Lemme="proposer"
   de^Cs=|prep|^Lemme="de" soumettre^Cs=|vinf|^Lemme="soumettre" aux^Cs=|prep|^Lemme="aux"
   chambres^Cs=|nc|^Lemme="chambre" un^Cs=|num|^Lemme="un" programme^Cs=|nc|^Lemme="programme" de^Cs=|prep|^Lemme="de"¶
9  législation^Cs=|nc|^Lemme="législation" visant^Cs=|vppa|^Lemme="viser" à^Cs=|prep|^Lemme="à"
   répondre^Cs=|vinf|^Lemme="répondre" aux^Cs=|prep|^Lemme="aux" exigences^Cs=|nc|^Lemme="exigence"
   collectives^Cs=|adj|^Lemme="collectif" les plus pressantes^Cs=|adj|^Lemme="pressant" de^Cs=|prep|^Lemme="de" la¶
10 population^Cs=|nc|^Lemme="population", à^Cs=|prep|^Lemme="à" élargir^Cs=|vinf|^Lemme="élargir" le
   champ^Cs=|nc|^Lemme="champ" d'^Cs=|prep|^Lemme="d'"action^Cs=|nc|^Lemme="action" et
   accroître^Cs=|vinf|^Lemme="accroître" l'efficacité^Cs=|nc|^Lemme="efficacité" du^Cs=|prep|^Lemme="du"
   gouvernement^Cs=|nc|^Lemme="gouvernement"¶
11 par^Cs=|prep|^Lemme="par" la création^Cs=|nc|^Lemme="création" de^Cs=|prep|^Lemme="de"
   nouveaux^Cs=|adj|^Lemme="nouveau" ministères^Cs=|nc|^Lemme="ministère", à^Cs=|prep|^Lemme="à"
   moderniser^Cs=|vinf|^Lemme="moderniser" ou remodeler^Cs=|vinf|^Lemme="remodeler" l'appareil^Cs=|nc|^Lemme="appareil"¶
12 administratif^Cs=|adj|^Lemme="administratif" existant^Cs=|vppa|^Lemme="exister".^Cs=|ponct|^Lemme="." *Phrase=6 Le
   gouvernement^Cs=|nc|^Lemme="gouvernement" vous invitera^Cs=|vfle|^Lemme="inviter" à^Cs=|prep|^Lemme="à"
   étudier^Cs=|vinf|^Lemme="étudier" un^Cs=|num|^Lemme="un" projet^Cs=|nc|^Lemme="projet" de^Cs=|prep|^Lemme="de"
   loi^Cs=|nc|^Lemme="loi"¶
13 pour^Cs=|prep|^Lemme="pour" autoriser^Cs=|vinf|^Lemme="autoriser" la création^Cs=|nc|^Lemme="création"
   d'^Cs=|prep|^Lemme="d'"un^Cs=|num|^Lemme="un" ministère^Cs=|nc|^Lemme="ministère" des^Cs=|prep|^Lemme="des"
   \Affaires^Cs=|np|^Lemme="Affaires" culturelles^Cs=|adj|^Lemme="culturel" qui aura pour^Cs=|prep|^Lemme="pour"¶
14 sa juridiction^Cs=|nc|^Lemme="juridiction", entre autres organismes^Cs=|nc|^Lemme="organisme", un^Cs=|num|^Lemme="un"
   \Office^Cs=|np|^Lemme="Office" de^Cs=|prep|^Lemme="de" la linguistique^Cs=|nc|^Lemme="linguistique",
   un^Cs=|num|^Lemme="un"¶
15 département^Cs=|nc|^Lemme="département" du^Cs=|prep|^Lemme="du" \Canada^Cs=|np|^Lemme="Canada"
```

Figure 2. Transfert des informations sémantiques et lemmatiques de Guidexpert vers SATO

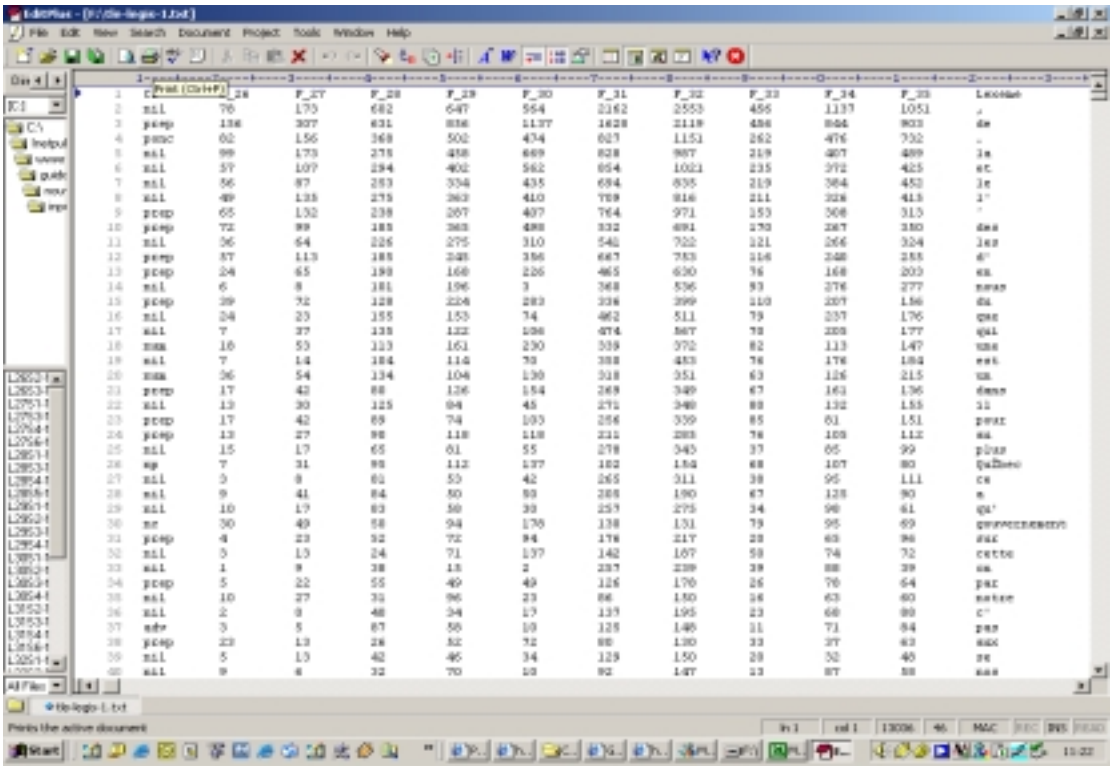

Figure 3. Tableau lexical entier des noms et adjectifs présentant plus de dix occurrences

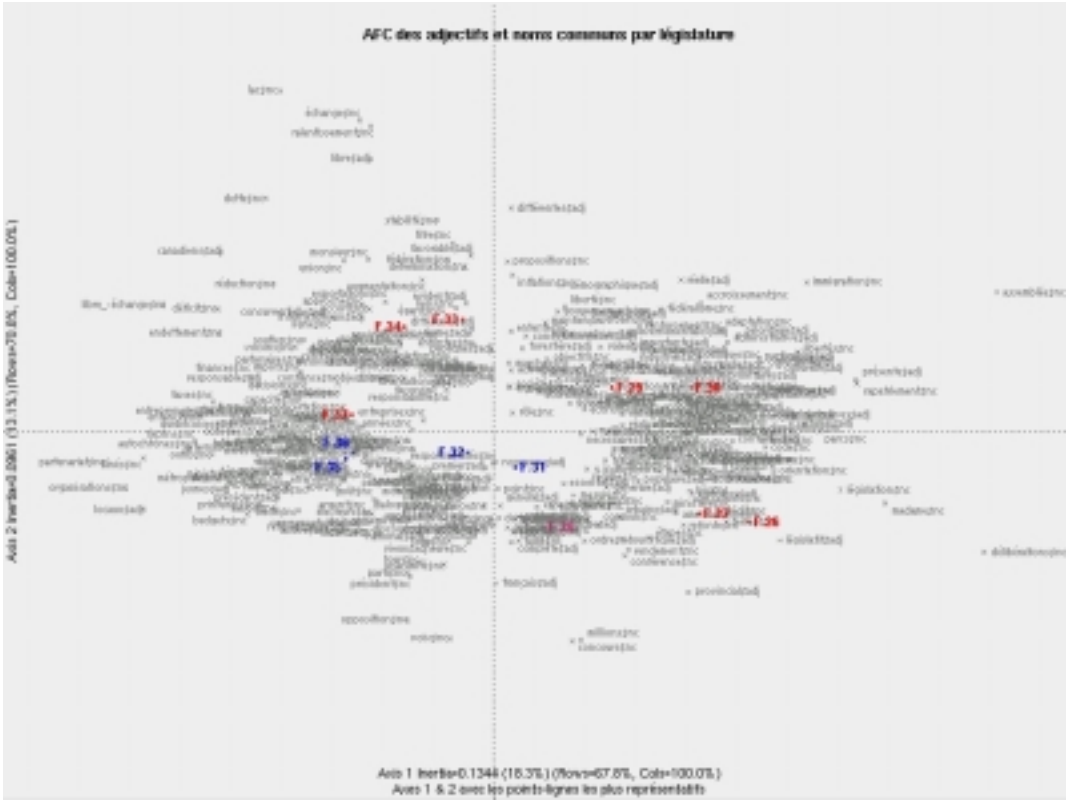

Figure 4. Analyse factorielle des correspondances produite à partir du tableau lexical entier