Jules Duchastel et Victor Armony
sociologues, département de sociologie, UQAM

(1995)

“La catégorisation socio-sémantique.”

Un document produit en version numérique par Jean-Marie Tremblay, bénévole,
professeur de sociologie retraité du Cégep de Chicoutimi
Courriel: jean-marie_tremblay@uqac.ca
Site web pédagogique : http://www.uqac.ca/jmt-sociologue/

Dans le cadre de: "Les classiques des sciences sociales"
Une bibliothèque numérique fondée et dirigée par Jean-Marie Tremblay,
professeur de sociologie au Cégep de Chicoutimi
Site web: http://classiques.uqac.ca/

Une collection développée en collaboration avec la Bibliothèque
Paul-Émile-Boulet de l'Université du Québec à Chicoutimi
Site web: http://bibliotheque.uqac.ca/

Politique d'utilisation de la bibliothèque des Classiques

Cette édition électronique a été réalisée par Jean-Marie Tremblay, bénévole, professeur de sociologie au Cégep de Chicoutimi à partir de :

Jules Duchastel et Victor Armony
sociologues, département de sociologie, UQAM

“La catégorisation socio-sémantique.”

Un texte publié dans l'ouvrage collectif, ***Actes des Troisièmes journées internationales d’analyse statistique de données textuelles***. Rome : CISU, 1995, pp. 193-200.

http://www.chaire-mcd.uqam.ca/

M Jules Duchastel, sociologue, professeur au département de sociologie de l’UQÀM, nous a accordé le 5 janvier 2005 son autorisation de diffuser électroniquement toutes ses oeuvres.

Courriel : duchastel.jules@uqam.ca

Polices de caractères utilisée :

Pour le texte: Times New Roman, 14 points.
Pour les notes de bas de page : Times New Roman, 12 points.

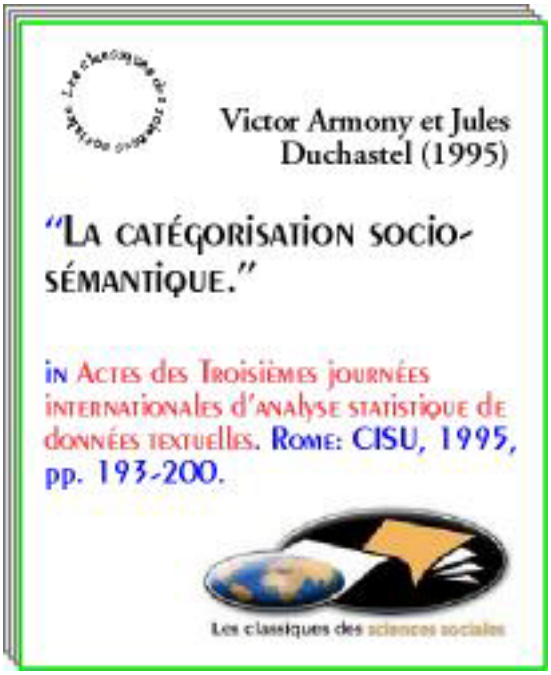
Édition électronique réalisée avec le traitement de textes Microsoft Word 2008 pour Macintosh.

Mise en page sur papier format : LETTRE US, 8.5’’ x 11’’.

Édition numérique réalisée le 11 octobre 2012 à Chicoutimi,
Ville de Saguenay, Québec.

Jules Duchastel et Victor Armony
sociologues, département de sociologie, UQAM

“La catégorisation socio-sémantique.”



Un texte publié dans l'ouvrage collectif, ***Actes des Troisièmes journées internationales d’analyse statistique de données textuelles***. Rome : CISU, 1995, pp. 193-200.

Jules Duchastel et Victor Armony
sociologues, département de sociologie, UQAM

“La catégorisation socio-sémantique.”

Un texte publié dans l'ouvrage collectif, ***Actes des Troisièmes journées internationales d’analyse statistique de données textuelles***. Rome : CISU, 1995, pp. 193-200.

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

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

Plan de l'article

1. Introduction

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

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

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

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

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

De manière générale, nous définissons la catégorisation des données textuelles comme l'ensemble des procédures visant à superposer aux unités d'enregistrement une ou plusieurs grilles de codage à valeur descriptive et analytique. La catégorisation sert à caractériser les éléments du corpus en leur attribuant de l'information de type extra ou péri-textuel (renseignements sur le locuteur, les circonstances de l'énonciation, etc.) et/ou en les classifiant selon des principes d'homogénéité (fonctionnelle, sémiotique, topique, etc.). Chaque unité du corpus reçoit alors des « étiquettes » qui la spécifient vis-à-vis d'un certain nombre de règles taxinomiques établies par l'analyste. Dans le cas particulier de la catégorisation socio-sémantique, telle que nous la concevons, on vise à classer - de manière exhaustive et exclusive - les mots à valence référentielle (noms et adjectifs) en fonction d'un système de catégories thématiques.

La construction de la grille de catégories suit une logique « constructiviste », c'est-à-dire qu'il s'agit d'une démarche empirique et itérative à visée interprétative, dont l'application se fait au moyen d'une lecture contextualisée : chaque occurrence est codée eu égard à sa signification dans la phrase. Cette perspective privilégie donc l'aspect *paradigmatique* mais *localisé* des unités du discours : le mot « droit »,

par exemple, ne sera catégorisé comme « domaine juridique » que si le sens de l'énoncé le justifie, car outre l'idée de « ce qui est conforme à une règle », il peut aussi signifier « redevance » (domaine économique) ou encore être utilisé dans une locution adverbiale comme « à bon droit ». Le diagramme suivant illustre cette logique : un même mot peut appartenir à deux catégories différentes (cas I et III), deux mots différents peuvent appartenir à une même catégorie (cas II) [2].

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

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

On sait que la catégorisation est une dimension-clé de l'analyse de contenu. Comme l'a dit Bernard Berelson (1952) : *« content analysis*

---

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

rises or faite by its content catégories ». C'est au moyen des catégories de contenu que l'information véhiculée par un message est réduite et uniformisée dans le but d'en produire, selon la célèbre formule, une « description objective, systématique et quantitative » [3]. Or, l'analyse de contenu est fortement associée à l'utilisation de « dictionnaires généraux » (par exemple, le *Laswell Value Dictionary*, le *Harvard Psychosocial Dictionary).* La stratégie des dictionnaires généraux se caractérise par l'utilisation d'un nombre limité de catégories (environ 60 à 150), la discrimination des homographes à partir de normes de désambiguisation et le traitement des locutions, ainsi que par le fait que la plupart des mots du texte sont codés, que chaque catégorie comporte un nom *(tag)* et une définition de ses règles d'application et que les mots ambigus peuvent être exclus de la catégorisation (la catégorisation multiple étant déconseillée dans ce type d'approche) (Weber, 1984). On privilégie donc dans cette stratégie les schèmes de codage *a priori* plutôt que *a posteriori* (Wood, 1980).

L'analyse de contenu se veut une analyse quantitative du langage, ou plus précisément, *une quantification des données qualitatives* (Roberts & Popping, 1993). Or, depuis déjà une trentaine d'années, un courant se développe au sein de la sociologie et de l'anthropologie qui se penche lui aussi sur les données « non-numériques » mais avec une approche justement « qualitative » (fondée en grande partie sur la *grounded theory).* Cette analyse proprement qualitative vise fondamentalement à décrire et à comprendre la culture et le comportement des individus et de leurs groupes du point de vue de ceux qui sont l'objet d'étude (Bryman, 1988). Les matériaux exploités sont souvent des entrevues ou des notes de terrain ; le chercheur tente de capturer la

---

[3] De là que la *fiabilité* de la catégorisation soit un problème névralgique dans toute démarche de ce genre. D'un point de vue conceptuel, la catégorisation consiste à regrouper des objets selon un ou plusieurs critères, en acceptant de négliger toutes les autres différences (Matalon, 1988). Il est alors évident qu'il faut optimiser la qualité du travail de codage, autant sur le plan de la définition des principes d'équivalence et de distinction (la construction de la grille ou du dictionnaire) que sur celui de leur application empirique (l'adéquation des catégories attribuées aux unités d'enregistrement). Selon Robert Philip Weber (1985), les trois types de fiabilité de la catégorisation sont : (1) la *stabilité* (les mêmes catégories aux mêmes unités), (2) la *reproductibilité* (cohérence entre les décisions des différents codeurs) et (3) la *précision* (par rapport à un standard).

complexité des phénomènes sociaux en faisant émerger du texte lui-même les concepts qui structureront sa théorie (Strauss, 1987). Naturellement, des ressources informatiques sont souvent mises à contribution pour gérer les masses de données que ce genre d'approche génère. Les logiciels les plus répandus dans ce domaine sont ceux de codage-repérage *(çode-and-retrieve programs).* Ils permettent de diviser le texte en séquences de mots et de leur attacher des codes pour pouvoir par la suite afficher toutes les parties qui ont reçu le même code ou combinaison de codes ; certains logiciels de ce type facilitent aussi la formulation de relations entre les catégories de façon à développer des classifications conceptuelles de grande complexité (Weitzman & Miles, 1995).

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

Notre grille de catégorisation est avant tout une classification empirique (mais conceptuellement fondée) des différents référents du discours politique. Son application aux items lexicaux n'a pourtant pas

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

3. Principes et procédures de la catégorisation socio-sémantique

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

Une catégorisation morpho-syntaxique préalable, inspirée de la grammaire de base du français, vise à déterminer si le mot est un nom, un verbe, un adjectif, une préposition, etc. Cette catégorisation est né-

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

cessaire pour déterminer les candidats à la catégorisation socio-sémantique car nous n'avons retenu à cette fin que les noms et les adjectifs. Les formes fonctionnelles ont été exclues en raison de leur faible potentiel sémantique et les verbes ignorés parce qu'ils appartiennent à une sémantique particulière qui nous éloigne de notre visée interprétative.

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

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

Nous envisageons la catégorisation comme un processus itératif : au fur et à mesure qu'il se développe, une dynamique d'aller-retour fait en sorte qu'il soit possible de : (1) détecter des régularités dans les décisions qui n'étaient pas prévues (ou « conscientes ») ; (2) détecter des inconsistances dans l'application de la grille. On peut donc dire qu'il s'agit d'un double processus d'apprentissage (sur la base de l'accumulation de décisions correctes) et de correction d'erreurs (sur la base de l'identification des décisions incorrectes). Deux documents d'appui à la catégorisation ont été créés à cet égard. Le premier regroupe, pour chaque catégorie de la grille, l'ensemble de termes du corpus qui l'ont

reçue. On parle alors de « l'éventail lexical » des catégories : cette information sert à compléter la définition de chaque catégorie et permet de vérifier sa consistance interne. Le second document est l'envers du premier : il est l'index alphabétique de toutes les formes avec mention des catégories qui leur ont été affectées dans les diverses sections du corpus. Il est alors possible d'observer les différents « usages » d'un même terme. Ces documents sont mis à jour régulièrement (chaque fois que de nouveaux textes sont catégorisés) et servent à expliciter et à formaliser les critères de catégorisation ainsi qu'à effectuer un contrôle périodique de sa fiabilité (stabilité, reproductibilité et précision).

4. L'analyse des données lexicales catégorisées

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

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

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

et action de servir », ce mot renvoie à diverses modalités d'interaction entre des acteurs sociaux.

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

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

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

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

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

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

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

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

Disons pour finir qu'il est important de remarquer que *Youtput* de ce type de procédure est, dans une certaine mesure, tributaire des décisions (pré-)analytiques prises au moment de la catégorisation. Nous voulons signaler ici que les résultats obtenus montrent en même temps :

(a) la validité de la catégorisation, c'est à dire le fait que nous avons bien classifié les usages du mot « service(s) » ; ceci est très important pour une entreprise comme la nôtre qui vise à superposer aux données textuelles brutes un système de repères servant à réaliser d'autres
fouilles lexicométriques mais aussi hyper-textuelles ;

(b) la possibilité d'identifier certains traits des « représentations » de divers domaines de l'activité sociale ; la catégorisation des usages du mot « service(s) », fondée sur une « observation sociologique » - objective mais non pas « neutre » - a permis de circonscrire trois espaces lexicaux différenciés.

5. Conclusion

Nous avons essayé de montrer dans cette communication l'utilité d'une catégorisation socio-sémantique quand on entreprend l'étude d'un corpus à pertinence sociologique. Nous avons indiqué ailleurs la valeur heuristique d'une analyse purement lexicométrique réalisée sur des données textuelles « brutes » (Armony & Duchastel, 1995) ; ce type d'approche, bien que très fructueux à l'étape exploratoire et apte à produire des descriptions quantitatives tout à fait intéressantes, reste trop sommaire lorsqu'on vise à générer des fouilles ciblées sur des référents précis du discours social. Nous avons vu qu'une même forme lexicale peut être l'indicateur socio-sémantique de différents objets de l'univers politique et qu'il est possible d'en tenir compte au moyen de

codes attribués en contexte d'occurrence. La catégorisation que nous proposons permet également de calculer la cooccurrence globale de, par exemple, l'ensemble de toutes les notions qui renvoient au domaine des pratiques juridiques (le mot « droit » lorsqu'utilisé dans son sens de « prérogative » plus le mot « justice » au sens de « légalité », etc.) et, de cette manière, d'articuler l'analyse du discours au cadre interprétatif du chercheur. Bref, nous croyons que ce type de démarche s'avère essentiel si l'on vise à décortiquer, en sociologues, la parole des acteurs afin d'y trouver la façon dont ils conçoivent leur monde.

RÉFÉRENCES

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

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

Baudelot, Ch. (1994). Préface. In Lebart, L. & Salem, A. *Statistique textuelle.* Paris : Dunod, pp. v-vi.

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

Bryman, A. (1988). *Quantity and Quality in Social Research.* London : UnwinHyman.

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

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

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

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

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

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

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

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

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

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

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

Fin du texte