BOURQUE, G.; DUCHASTEL, J.; avec la coll. de V. ARMONY (1996). " Annexe II: Méthodologie". In *L'identité fragmentée. Nation et citoyenneté dans les débats constitutionnels canadiens, 1941-1992*. Montréal: Fides, 1996: 343-350.

La présente annexe fournit l'essentiel des informations nécessaires à la compréhension de la démarche mise en oeuvre pour le traitement et l'analyse du discours politique dont les résultats font la trame de ce livre. On pourra trouver ailleurs un exposé plus élaboré des fondements théoriques et méthodologiques de notre travail (Duchastel, 1995, 1992; Duchastel et Armony, 1994, 1993). Nous procéderons plutôt ici à une description des divers choix méthodologiques concernant le corpus, la catégorisation et la stratégie de traitement des données.

La mise en forme du corpus

Dans le type de démarche que nous privilégions, la construction et la mise en forme du corpus s'avère une phase névralgique": même si celle-ci comporte des opérations de nature surtout technique, leur exécution rigoureuse et systématique est essentielle à la réussite des procédures analytiques subséquentes. Le corpus colligé -"un ensemble de documents photocopiés"- devra devenir une *base de données textuelles* susceptible d'être traitée à l'aide de l'ordinateur.

Une fois que les documents ont été sélectionnés, leur contenu est "numérisé"", c'est-à-dire qu'ils sont convertis en fichiers informatiques. Cela s'effectue en trois étapes": d'abord, on effectue la saisie optique de chaque page (au moyen d'un *scanner*)"; par la suite, ces images sont soumises à un logiciel de "reconnaissance" qui traduit les graphismes en caractères"; enfin, les textes obtenus de cette manière sont édités afin de corriger les erreurs de reconnaissance, enlever les éléments qui ne sont pas pertinents pour l'analyse du discours (tables de matières, diagrammes, etc.) et uniformiser le format d'écriture (simplification des styles typographiques, standardisation des noms propres, etc.).

Par la suite, ces archives informatisées doivent subir trois autres manipulations": le blocage des locutions, la classification morpho-syntaxique et l'introduction des repères contextuels. La première consiste en l'identification des multi-termes. Il s'agit des unités lexicales composées de plusieurs mots": par exemple, "Banque du Canada". L'objectif est de pouvoir les traiter en tant qu'entrées uniques du lexique (ainsi, lorsque le terme "Canada" apparaÎt dans la formule "Banque du Canada", il n'est pas compté parmi les occurrences de "Canada"). Cette approche permet de mieux dépister les référents du discours. La deuxième opération vise à regrouper les mots en fonction des principales classes grammaticales": noms communs, verbes, pronoms, etc. Cela est fait au moyen d'un "dictionnaire" informatisé qui attribue à chaque vocable du corpus une "étiquette" morpho-syntaxique. Cette classification est nécessaire pour déterminer les candidats à la catégorisation socio-sémantique. En effet, 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 auraient nécessité un traitement trop spécialisé. Enfin, les documents font l'objet d'un codage fondé sur leur "coordonnées de production"": date, lieu et identité du locuteur (nom et province représentée). Des repères sont inscrits dans les fichiers eux-mêmes, de sorte que toute interrogation de la base des données peut s'effectuer de manière ciblée : par exemple, quels sont les termes

caractéristiques du discours des premiers ministres des provinces maritimes durant la conférence de 1964? quelles sont les phrases où Pierre Elliott Trudeau parle de "bilinguisme" entre 1978 et 1982?

La catégorisation socio-sémantique

Nous définissons la catégorisation socio-sémantique comme un ensemble de procédures visant à superposer aux unités lexicales une grille de codage à valeur descriptive et analytique d'un point de vue sociologique. La catégorisation du corpus est clé 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 avons effectué une catégorisation "en contexte"": chaque occurrence est soumise à une décision. Le catégorisateur 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 attribuer une "étiquette" informatique.

La grille que nous proposons est avant tout un classement empirique des différents objets du discours politique. Elle est empirique dans la mesure où elle a été progressivement construite à partir de l'observation et de la catégorisation effective des divers sous-corpus, dans le but de rendre compte du contenu socio-sémantique qui s'en dégageait. Cela n'exclut cependant pas qu'elle réfère aux dimensions théoriques de l'analyse du discours politique dans la société moderne. C'est ainsi qu'elle permet d'identifier les principaux acteurs, institutions et valeurs travaillés par ce discours. Cette grille devient donc un outil d'analyse. L'application de catégories aux mots du texte n'a pourtant pas l'effet de faire disparaÎtre le mot sous la catégorie. Le système informatique utilisé permet, en effet, 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.

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

Cette démarche préalable à l'analyse permet de garantir la signification "qualitative" des résultats obtenus par le biais des calculs statistiques": chaque vocable qui apparaÎt dans les tableaux correspond à des occurrences sémantiquement pleines. Par exemple, lorsque le mot *unité* ressort dans un tableau de cooccurrence, on sait qu'il s'agit d'une association statistiquement significative entre le mot-pôle et l'acception axiologique du terme *unité* (car les usages non-axiologiques du mot *unité* -""par exemple, dans l'expression "l'unité de mesure""- ont été écartés lors de la catégorisation socio-sémantique).

La démarche analytique

Notre approche privilégie la fonction référentielle du discours": nous traitons les unités sémantiques (axe paradigmatique) et leurs combinaisons (axe syntagmatique). Le lexique constitue donc la base de notre analyse. Nous y repérons les unités sémantiques à travers les lexèmes mêmes ou des regroupements de lexèmes ayant reçu la même catégorie socio-sémantique, donc sur l'axe paradigmatique. Nous examinons également la dimension contextuelle qui renvoie à l'axe syntagmatique. Sans proposer d'analyse des relations fonctionnelles entre éléments de la phrase comme nous avons pu le faire ailleurs (Duchastel, Paquin et Beauchemin, 1994, 1992"; Bourque et Duchastel, 1988), nous nous intéressons aux relations de cooccurrence dans le contexte de la phrase, à partir du lexique des cooccurrents et des concordances. Nous avons donc un triple accès au sens du texte à travers les mots, leur catégorie et le contexte où ils émergent.

L'analyse a été effectuée à l'aide du logiciel SATO (*Système d'Analyse de Textes par Ordinateurs*). Ce logiciel est défini par son auteur comme un système de base de données textuelles qui permet d'annoter des textes multilingues et de les manipuler de diverses façons: repérage de concordances, construction de lexiques, catégorisation des mots, dénombrements de tout ordre et analyseurs lexicométriques.(Daoust, 1989: 117).

SATO est un environnement informatique que l'on peut représenter comme une boÎte à outils dans laquelle se trouverait un ensemble d'instruments destinés à l'analyse des données textuelles. En plus de la diversité des outils disponibles, l'originalité principale de SATO réside dans le fait qu'il permet à l'usager d'entretenir un rapport interactif au texte qu'il étudie. Les diverses tâches d'annotation, de production de lexiques, de repérage de concordances ainsi que d'analyse lexicométrique s'effectuent directement à l'écran et peuvent être constamment réitérées ou modifiées. Mais, quelque soient les manipulations et les enrichissements successifs dont le texte est l'objet, le texte original demeure accessible en tout temps.

Les procédures

SATO permet d'abord de produire des lexiques de tous ordres suivant des paramètres fixés par l'utilisateur. On pourra, par exemple, produire le lexique de tous les mots commençant par "constitu" dans les allocutions de la période 1941-1950 dont la fréquence dans le corpus est supérieure à 10 occurrences. Le logiciel produit instantanément un lexique, ventilé par sous-corpus s'ils ont été préalablement définis comme domaines (discours du fédéral, du Québec, des autres provinces), regroupant dans ce cas des mots tels "Constitution", "constitutionnel", "constitutionnelles", etc. Ces lexiques peuvent être ordonnés en fonction de la fréquence des mots qu'ils regroupent ou de leur ordre alphabétique.

SATO facilite également l'étude du covoisinage à travers la production de lexiques de cooccurrences. Le logiciel offre la possibilité de repérer et de dénombrer, suivant une multitude de paramètres possibles, la co-présence de mots. L'analyse s'est ainsi penchée, de diverses manières, sur le fonctionnement discursif du covoisinage, s'intéressant tantôt aux relations qui s'établissent entre des notions afférentes aux valeurs, tantôt aux rapports entre certaines valeurs et des catégories de l'univers social ou encore au réseau que forment les mots eux-mêmes, indépendamment de leur catégorie. Dans ce livre, nous avons privilégié l'étude des cooccurrences en fonction d'un test de signification statistique.

Nous présenterons brièvement le fondement mathématique de ce test. Nous nous intéressons donc à la cooccurrence d'un mot particulier, le mot pôle, avec l'ensemble des mots qui

apparaissent avec lui dans un segment donné. Le but de la méthode est d'obtenir la liste des mots cooccurrant avec le mot pôle, pour lesquels la cooccurrence est statistiquement significative, aussi bien lorsque la cooccurrence est surabondante ou au contraire lorsqu'elle est rare. Nous avons retenu la phrase comme segment de référence, considérant que celle-ci représente une unité "naturelle" de sens. Nous considérons donc que l'ensemble des phrases du corpus analysé constitue l'échantillon de référence. Pour une cooccurrence particulière, l'observation est donc la phrase et la variable étudiée (que nous appelons x) est le nombre de phrases contenant cette cooccurrence. La mise en évidence de la significativité de la cooccurrence s'effectuera alors par l'intermédiaire du test statistique suivant":

Soit n le nombre de phrases du corpus.

Soit f_p le nombre de phrases contenant le mot pôle.

Soit f_c le nombre de phrases contenant le mot cooccurrent dans le corpus.

Soit f_{pc} le nombre de phrases contenant le mot pôle et le mot cooccurrent dans le corpus.

Si la présence des deux mots dans une phrase est due au hasard, la fréquence espérée des phrases contenant les deux mots est :

$$e = \frac{f_p \times f_c}{n}$$

et la variable X suit une loi binomiale de paramètres n et e/n.

Formellement, on peut alors tester l'hypothèse nulle que la proportion de phrases contenant le mot pôle et le mot cooccurrent est de e/n sachant qu'on en a fpc dans l'échantillon, l'hypothèse alternative étant l'hypothèse contraire. Si $f_{pc} \geq e$, on calcule la probabilité que X soit supérieur à fpc si la proportion de phrases contenant le mot pôle et le mot cooccurrent est de e/n. Si cette probabilité est excessivement petite, on en conclue que e/n ne peut pas être la proportion de phrases contenant le mot pôle et le mot cooccurrent et on rejette l'hypothèse nulle. Il suffit de fixer un seuil de probabilité en dessous duquel on considère la probabilité comme étant trop faible, par exemple 0.01 (soit 99"% de confiance).

Un autre type de test statistique nous a permis de détecter les notions et les catégories (ensemble de notions sémantiquement proches) nodales -"par période"- et distinctives -"par locuteurs"- du corpus. Pour une période donnée, les notions ou catégories nodales sont celles dont la répartition entre les locuteurs est jugée aléatoire. Autrement dit, ces notions ou catégories ne sont spécifiques à aucun locuteur en particulier. Cela veut dire concrètement que les fréquences relatives sont similaires dans tous les sous-corpus. A l'inverse, mais selon le même principe statistique, les notions ou catégories distinctives sont celles qui apparaissent plutôt concentrées dans l'un des sous-corpus.

Enfin, en raison même de la structure de représentation des données propre à SATO, l'utilisateur peut retourner au texte à tous moments de l'investigation. Il est ainsi possible de retrouver la phrase correspondant aux occurrences ou cooccurrences que l'analyse lexicale aura mises à jour. SATO permet le repérage instantané des concordances à partir de critères

de sélection fixés par l'utilisateur. On peut définir la concordance comme la chaÎne syntagmatique (ou la liste des chaÎnes syntagmatiques) comportant l'occurrence d'une ou la cooccurrence de plusieurs formes lexicales sélectionnées. Les mots du texte apparaissent alors en contexte (dont les limites sont définies par l'utilisateur) et peuvent faire l'objet de nouvelles manipulations (nouvelle catégorisation, sous-catégorisation, désambiguïsation, etc.) dont le résultat pourra, par la suite, être réinvestigué. Les concordances sont utilisées lors des opérations de catégorisation, mais aussi afin de valider l'interprétation des lexiques produits par nos modèles de fouille. Nous avons reproduit, dans ce livre, un choix de ces concordances afin d'illustrer la pertinence de ces analyses.