BOURQUE, G.; DUCHASTEL, J.; BEAUCHEMIN, J. (1994). " Annexe 1 : Méthodologie ". In *La société libérale duplessiste, 1944-1960*. Montréal: Les Presses de l'Université de Montréal, 1994: 337-361.

Annexe méthodologique

Introduction

La présente annexe fournit l'essentiel des informations nécessaires à la compréhension de la démarche méthodologique 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. Nous procéderons plutôt ici à une description avant tout technique des différentes démarches et procédures de recherche. Cela ne nous dispensera cependant pas de situer notre démarche dans le cadre plus général de l'analyse du discours telle que nous l'avons conçue depuis le début de nos recherches sur le discours politique, ni de donner sens à la démarche interprétative qui est la nôtre dans ce livre. Nous décrirons donc succinctement la perspective d'analyse que nous avons privilégiée avant même de présenter les divers choix méthodologiques concernant le corpus, la catégorisation et la stratégie informatique de traitement des données.

Analyse du discours politique

La perspective privilégiée dans ce livre est l'analyse de la dimension socio-sémantique du discours politique duplessiste. Afin de préciser cette approche, nous reviendrons rapidement sur notre conception de la nature du discours et de la pluralité des lectures possibles avant d'insister sur la continuité de notre approche globale.

Il serait facile de caricaturer et de dire que le discours est ou bien le reflet d'une réalité qui lui est sous-jacente et qui l'explique entièrement ou bien la réalité elle-même, en tant qu'elle se construit dans le processus de son énonciation. Dans le premier cas, on aurait affaire à une perspective d'analyse de contenu classique qui considère le discours comme surface reflétante d'une réalité matérielle quelconque (psychologique, sociologique, politique,...). Dans le second, il s'agirait d'une approche strictement constructionniste qui ne suppose aucun fondement axiomatique à la réalité en dehors de sa mise en discours. Il s'agit bien là de deux conceptions opposées du discours comme pure représentation ou comme pure construction. Nous avons adopté depuis le début de nos travaux une position intermédiaire qui consiste à reconnaÎtre dans le discours une double fonction de représentation et de production de la réalité. Nous faisons l'hypothèse que la réalité n'est pas pure représentation mais, en tant même qu'elle existe, comprend toujours une dimension symbolique. Le discours contribue à sa production et à sa reproduction sans pour autant en épuiser toute la substance.

Ceci étant dit, comment caractériser l'objet discours? Nous dirons qu'il est à la fois objet et processus. En tant qu'objet, il a une certaine existence matérielle qui se manifeste à la fois dans des formes et des contenus, chacun pouvant donner lieu à des analyses très variées. En tant que processus, il s'inscrit dans des conditions de production/énonciation qui doivent être définies et dont l'articulation avec le discours doit être pensée. C'est à ces impératifs que nous

avons tenté de répondre en proposant notre démarche méthodologique. Nous inscrivons cette démarche dans la perspective de l'analyse du discours plus que dans celle de l'analyse du contenu pour diverses raisons. Bien que nous ayons privilégié l'analyse de la dimension socio-sémantique du discours politique duplessiste dans le présent ouvrage, notre approche s'inscrit tout de même dans la perspective générale d'analyse du discours de trois points de vue principaux.

D'abord, la définition théorique du discours politique que nous proposons dans ce livre et l'importance théorique accordée à la construction du corpus situe notre démarche dans la tradition française d'analyse du discours. En analyse de contenu classique, les discours sont uniquement rapportés à des locuteurs et à des conditions de production définie en pure extériorité. Aucune théorie de l'articulation des discours et de leurs conditions de production n'est proposée. Au contraire, notre théorie du discours politique implique qu'il soit pensé comme participant aux procès d'institutionnalisation caractéristiques des sociétés modernes et contribuant, en particulier, à produire un bloc social spécifique. Il émane, en ce sens, de la pluralité des institutions privées et publiques au sein desquelles s'organisent les rapports de pouvoir. Cette perspective structurante du discours politique appartient, en propre, à la tradition d'analyse du discours. Comme nous le verrons plus loin, le corpus n'a donc pas été constitué à partir des beaux morceaux, à forte teneur thématique, de quelques acteurs politiques, mais bien sur la base du discours tel qu'il est produit et travaille à l'intérieur de certaines institutions des sphères privée et publique.

Le choix des opérations que nous avons appliquées dans le processus de l'analyse situe notre démarche à mi-chemin entre l'analyse de contenu et l'analyse du discours. L'analyse du discours se distingue de la tradition de l'analyse de contenu classique par son travail de description et de repérage des traces des divers fonctionnements linguistiques, para-linguistiques ou même extra-linguistiques. Il nous semble cependant que cette opposition entre les deux traditions, pour fondée qu'elle puisse avoir été historiquement, est de moins en moins évidente. Il est peu d'analyse de contenu qui ignore complètement la texture du discours et ne retienne que les mots dans leur nudité ou que les thèmes dans leur idéalité. La construction des dimensions paradigmatique et syntagmatique distingue notre approche de certaines formes d'analyse de contenu classique qui puise directement le sens dans les mots (lexicométrie) ou les énoncés (analyse qualitative et thématique). Sur le plan paradigmatique, nous avons choisi de procéder à une catégorisation socio-sémantique systématique. L'idée de la catégorisation appartient en propre à la tradition d'analyse de contenu, mais alors que dans cette dernière le chercheur se voit contraint à choisir entre le texte et les catégories, le texte catégorisé peut, dans notre cas, être fouillé par des modèles d'exploration qui portent aussi bien sur les mots que sur les catégories. Dun autre côté, bien que nous ayons renoncé dans ce livre à toute description des structures syntagmatiques ou textuelles, cette dimension est tout de même prise en compte dans la stratégie d'analyse que nous avons privilégiée. L'essentiel de la démonstration s'appuie, en effet, sur la mise en relation des diverses parties du discours par l'analyse des cooccurrences et l'examen des concordances. La méthode informatisée que nous avons utilisée permet une exploration systématique des relations de proximité sémantique entre les mots du texte. L'accès immédiat aux concordances (dans notre cas, les phrase contenant les éléments cooccurrents) permet de valider, par une lecture en contexte et systématique, les résultats ainsi obtenus. Si l'on considère donc l'ensemble du modèle opératoire mis en oeuvre, le primat accordé à la fonction référentielle nous rattache d'un côté à la tradition d'analyse de contenu alors que, de l'autre, la structuration des axes

paradigmatiques et syntagmatique nous rapproche de la philosophie même de toute analyse du discours.

Enfin, l'idée-même de réfléchir au processus interprétatif nous situe davantage dans une perspective d'analyse du discours. En analyse de contenu, l'interprétation n'est conçue que comme processus inférentiel. Le texte est une surface indicielle d'un autre niveau de réalité qui lui est extérieur. Nous reprenons l'idée de Molino selon laquelle tout texte est constitué de trois matérialités inséparables: sa production, sa textualité et sa réception. Comme production, le discours politique doit être vu, comme nous l'écrivons ci-haut, dans son rapport aux différents procès d'institutionnalisation de la société. En tant que matérialité textuelle, il doit être travaillé sur les axes paradigmatiques et syntagmatiques. Reste l'interprétation qui renvoie au problème de la réception. Lorsqu'on pense à la réception, on est tenté de regarder du côté des destinataires naturels du discours. Dans le cas de discours historiques, il est toujours difficile de considérer systématiquement la réception et on se trouve le plus souvent obligé de prendre pour acquis que le discours a produit ses effets. Mais la réception peut aussi se concevoir à travers le processus-même de l'interprétation. L'analyste se trouve, en effet, en position de récepteur lorsqu'il lit, travaille et interprète un texte. En d'autres mots, le sens qui ressort de ces processus est également le produit de la lecture interprétative. Ricoeur va également dans le sens de Molino lorsqu'il insiste sur l'indépendance du texte comme objet d'étude, ce qui rend possible son objectivation plus ou moins poussée. Mais il ajoute que quelque soit le degré de sophistication des modèles de description des données discursives, arrive toujours le moment de l'interprétation. Nous avons développé ailleurs l'idée que ce moment doit être repoussé le plus possible afin d'accroÎtre l'espace de la description objective et des inférences explicites. C'est pourquoi nous tentons de distinguer une interprétation intimement associée au raisonnement scientifique, répondant donc à des règles explicites (inférentielle, par exemple) et une interprétation externe qui échappe nécessairement au plan expérimental et se réfère alors à la connaissance du monde et à la représentation théorique que nous en proposons. De là un retour nécessaire à la problématique du discours comme objet historique et social.

Ainsi, nous avons évité la perspective herméneutique naïve que l'on retrouve souvent en analyse de contenu classique. Nous avons tenter de construire progressivement l'interprétation dans le cadre d'une démarche analytique explicite. Mais, au terme de cette démarche, s'est imposée la nécessité d'une interprétation sociologique générale du duplessisme vu à travers la production du discours politique. C'est encore à ce niveau que l'analyse que nous produisons ne peut être assimilée à une simple analyse de contenu. Le cadre interprétatif tient compte de la théorie du discours comme des hypothèses socio-historiques qui ne peuvent être formulées qu'à un niveau général qui excède celui de l'empirie des discours analysés.

Approche heuristique

Dans Restons traditionnels et progressifs, nous avons qualifié notre approche d'empirico-constructiviste. Cette dénomination ne visait pas à situer notre démarche dans le courant constructionniste ou déconstructionniste, mais à en souligner le caractère itératif qui permet l'aller-retour entre l'examen empirique d'un grand nombre de données construites et la consolidation d'hypothèses et de modèles. Ce mouvement entre données empiriques et interprétation analytique s'appuie sur la souplesse d'une stratégie d'analyse assistée par ordinateur qui rend possible aussi bien la projection d'hypothèses sur le texte que la remontée de la surface du texte vers des généralisations théoriques. Cette stratégie non-déterministe

permet de réviser aussi souvent que jugé nécessaire les descriptions appliquées au texte et d'en reproduire l'exploration en vertu de nouvelles hypothèses. Cet arsenal méthodologique fut proposé en réponse au problème de la surdétermination des résultats par les procédures d'analyse. Rappelons la critique de Gardin concernant les limites de toute analyse du discours dont les procédures contribuent à contraindre à l'avance la nature des résultats. Nous jugions alors que le modèle informatique proposé permettait la construction-déconstruction progressive des données en suscitant la découverte de résultats inattendus.

Nous avons essentiellement conservé la même perspective. En analyse de contenu, il est courant de distinguer démarche heuristique et administration de la preuve. Ces deux démarches sont conçues comme des étapes successives d'un même processus analytique. D'une certaine manière, cette opposition renvoie au problème soulevé ci-haut du degré de détermination appliqué aux données. Mais, dans la tradition d'analyse de contenu, l'étape heuristique est souvent conçue comme un moment qui échappe à proprement parler à la méthode. Il s'agit de la lecture “flottante”, “intuitive” proposée comme point de départ à la construction d'un protocole d'observation. En ce qui nous concerne, notre méthode permet, en quelque sorte, d'intégrer la démarche heuristique à l'ensemble du processus analytique. Nous verrons dans les sections suivantes comment la souplesse du logiciel SATO autorise l'application de procédures de description, d'exploration et d'analyse aux différents moments du processus global. Cela nous a permis d'adopter une approche progressive de co-construction du sens produit par le discours. Les descriptions insuffisantes ont pu être reprises, les explorations sans issues abandonnées, celles qui promettaient poursuivies dans de nouvelles directions, les analyses improductives laissées pour compte, alors que celles qui convergeaient vers une interprétation congruente du discours duplessiste achevées.

Analyse du discours à base lexicale en contexte

Comment alors qualifier l'analyse que nous avons conduite dans ce livre? Nous avons retenu l'appellation d'"analyse du discours à base lexicale en contexte". Un examen plus serré des différents éléments de cette locution donnera un aperçu général de la méthode avant que nous en présentions systématiquement les éléments. D'abord, notre approche privilégie la fonction référentielle du discours. Nous traitons les unités sémantiques et leurs combinaisons. 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. Enfin, la dimension contextuelle renvoie à l'axe syntagmatique. Nous ne proposons pas, comme ce fut le cas dans d'autres travaux, d'analyse des relations fonctionnelles entre éléments de la phrase. Ce que nous observons ici, ce sont les relations de cooccurrence dans le contexte de la phrase, sur la base à la fois du lexique des cooccurrents et de la phraséologie elle-même sous forme de concordances. Nous avons donc un triple accès au sens du texte à travers les mots, leur catégorie et le contexte où ils émergent.

Corpus

Aperçu du corpus global

Avant de présenter les sous-corpus que nous avons retenus pour la présente étude, nous évoquerons succinctement les principes qui ont guidé la constitution d'un corpus global

représentant le discours politique duplessiste et nous donnerons un aperçu de cet ensemble discursif.

Nous avons déjà discuté, dans Restons traditionnels et progressifs, de l'influence des dimensions théorique et méthodologique de notre approche sur la définition de l'univers des discours à partir duquel nous avons constitué notre corpus. La théorie du discours politique que nous avons proposée comporte deux aspects qui ont orienté notre travail sur le corpus. La description des divers traits qui caractérisent, selon nous, le discours politique nous a conduit à définir l'espace-même des discours que nous considérions de nature politique La définition théorique que nous donnons du discours politique et l'hypothèse socio-historique que nous avons formulée à propos du régime Duplessis ont également contribué à délimiter plus concrètement le corpus. Enfin, la méthode d'analyse assistée par ordinateur, en nous offrant la capacité de traiter et d'analyser de grands ensembles textuels, nous a conduit à définir les limites quantitatives du corpus. Voyons succinctement comment chacun de ces facteurs a déterminé le choix final de notre corpus.

Le discours politique n'est pas avant tout un discours théorique ou doctrinaire. Il est discours de masse qui circule dans les interstices des sphères privée et publique. La cohérence et l'homogénéité ne constituent pas sa caractéristique première. Il est au contraire hétérogène et polémique, point de vue parmi d'autres points de vue dans le jeu des échanges discursifs. Il traite de questions particulières dans des lieux distincts, tout en ayant tendance à élargir l'espace du questionnement et des institutions qui participent au débat. Deux conséquences découlent de cette conception: d'une part, nous avons renoncé aux énoncés doctrinaires ou programmatiques au profit des discours de la pratique politique tels qu'ils émergent dans le cadre de différentes institutions; d'autre part, nous avons élargi la définition des institutions qui contribuent de manière significative au débat politique en nous intéressant aussi bien aux institutions de la sphère privée qu'à celles de la sphère publique.

Nous considérons que le discours politique contribue de manière prévalante à la représentation de l'espace, de la communauté, des rapports sociaux et du rapport de l'individu à la société (l'éthique). Cette définition a permis de préciser, dans le contexte socio-historique qui nous intéressait, les principales institutions et forces sociales émettrices de discours politiques. Ainsi avons-nous défini un univers de discours politique pouvant rendre compte de l'ensemble du travail discursif politique sous le régime duplessiste. Cet ensemble, à partir duquel nous avons procédé à la constitution du corpus, se compose de deux sous-ensembles principaux. Nous avons retenu les discours correspondant aux activités fondamentales de la sphère publique dans la société moderne. Le discours électoral, en tant qu'exercice du processus démocratique, s'intéresse aux enjeux politiques définis par les différents partis politiques. Les discours du trône et du budget posent les orientations politiques fondamentales des gouvernements élus. Le discours législatif définit les domaines d'intervention privilégiés et les réponses apportées par le parti au pouvoir. Le discours constitutionnel, enfin, met en jeu les dimensions fondamentales de la régulation politique et de la forme du régime.

Il existe un second ensemble de discours que nous avons considérés comme étant de nature politique. Il s'agit du discours sur les rapports sociaux émis par diverses institutions de la sphère privée. Nous avons retenu du monde économique le discours de certaines organisations patronales et syndicales. Pour le discours patronal, nous avons identifié une série de publications périodiques disponibles sur toute la période, soit celles de la Chambre de commerce de Montréal et du Mouvement coopératif Desjardins. De même, nous avons retenu trois séries de publications syndicales, deux du monde ouvrier et une de la classe agricole. De

plus, comme nous nous en expliquons abondamment dans ce livre, nous avons considéré l'Église comme une institution majeure dans le procès de production politique de la société québécoise. Nous avons donc constitué un ensemble de discours à orientation politique formé de deux parties. Le discours des mandements des Évêques y représente les positions officielles de l'Église catholique, alors que les publications étudiante, ouvrière et agricole rendent compte des mouvements d'action catholique.

Les facilités de traitement informatique des données nous ont incité, dans la construction du corpus global, à fixer des bornes temporelles larges et à favoriser la multiplication des locuteurs ce qui autorisait ainsi la comparaison diachronique selon des diverses périodes du régime entre 1936 et 1960 et synchronique entre les divers locuteurs (Union nationale et parti libéral, patrons, ouvriers et agriculteurs, Évêques, jeunesses catholiques).

Description du corpus étudié

Le corpus global que nous avons décrit très succinctement a donné lieu à des analyses déjà publiées dont la principale est certainement Restons Traditionnels et progressifs qui porte exclusivement sur le discours du Budget. Nous avons dû, dans le présent livre, limiter nos ambitions à certains sous-ensembles du corpus pour des raisons d'espace et de cohérence analytique. En effet, la prise en compte de tous les discours et de toutes les périodes délimitant notre corpus global constitue un projet d'une ampleur beaucoup plus grande que celle qui est autorisée dans un seul livre. Nous avons donc choisi de retenir deux sous-corpus particuliers au cours d'une partie seulement de la période du gouvernement de Maurice Duplessis. Nous avons retenu celle qui, de 1944 à 1960, couvre le retour au pouvoir de l'Union nationale après la deuxième guerre mondiale. Des analyses préalables nous ont permis de déterminer que l'après guerre constitue une entité cohérente. D'une part, le Québec se retrouve non seulement dans une période de mutation économique accélérée, mais également confronté à un projet de transformation du mode de régulation politique. D'autre part, la thématique du discours duplessiste se resserre autour d'axes dominants qui diffèrent de la période 1936-1939.

Les deux sous-ensembles retenus seront donc le corpus des discours de l'Union nationale dans les institutions de la sphère publique, à l'exception du discours du Budget déjà largement étudié et le discours des Mandements des Évêques, dans la sphère privée. Dans le premier cas, il s'agit de concentrer notre attention sur le discours du parti de Maurice Duplessis afin d'en saisir tous les ressorts. Dans le second, nous nous penchons sur l'une des principales institutions qui, dans la sphère privée, occupe une place stratégique dans la régulation libérale instituée par l'Union nationale. Ensemble, ils constituent deux des axes principaux du discours politique dominant durant la période. Nous référons, dans ce livre, au concept de discours duplessiste pour désigner l'ensemble de ces deux discours. Dans le cas où nous les considérons indépendamment, nous distinguons le discours de l'Union nationale et celui des Mandements des Évêques.

Corpus de l'Union nationale

Nous présenterons donc plus en détail le sous-corpus de l'Union nationale en distinguant le discours du Trône, le discours législatif, le discours électoral et le discours constitutionnel. Nous nous attarderons principalement à décrire la source ainsi que les principes de sélection,

d'édition et d'échantillonnage retenus pour chaque sous-ensemble. Il faut noter que seul le discours des élus de l'Union nationale a été retenu pour ce livre.

Le discours du Trône

Le sous-corpus du discours du Trône est constitué de tous les discours du Trône prononcés par l'Union nationale de 1944 à 1960 inclusivement. Le discours du Trône inaugure la session parlementaire. Il est habituellement l'occasion de rappeler les réalisations du gouvernement et constitue un énoncé d'intentions générales qui correspond à l'essentiel du programme législatif de la session à venir.

Le texte de tous les discours du Trône de la période est publié par l'imprimeur officiel de la province de Québec. Le travail de constitution de ce sous-corpus n'a posé aucun problème particulier étant donnée l'accessibilité de cette publication officielle du Québec. En raison de leur caractère relativement succinct, les textes n'ont pas été échantillonnés et ont été saisis presque intégralement sur support magnétique. Seule, une procédure d'édition visant à éliminer les titres et sous-titres a été appliquée afin d'éviter le gonflement artificiel du lexique.

Le discours législatif

La composition du sous-corpus du discours législatif a présenté davantage de difficultés. Il faut, en effet, rappeler qu'il n'existe pas de journal des débats durant la période. Cet ensemble de textes réunit donc les comptes rendus des débats entourant les lectures et sanctions des principales lois de la 22e à la 25e législature du Québec tels que rapportés par le journal Le Devoir. La couverture journalistique du Devoir nous a semblé la plus rigoureuse et la plus exhaustive en l'absence de comptes rendus officiels.

Six domaines permettant de rendre compte des composantes essentielles de l'activité législative et de révéler l'état de la société et de ses transformations, ont servi à circonscrire le sous-corpus.

1 - Agriculture et colonisation

2 - Richesses naturelles

3 - Économie: industries, commerce, finance et coopérative

4 - Législation ouvrière

5 - Domaine social: santé, éducation, habitation, famille

6 - Relations fédérales-provinciales

La première étape de la sélection a consisté en un relevé exhaustif des lois recoupant les six domaines législatifs définis. De ce bassin, n'ont alors été retenues que les lois inédites et donc soumises pour la première fois à la législature. Les modifications ou amendements à des lois existantes ont donc été écartés. La deuxième étape de la sélection a consisté à compiler les articles du Devoir traitant des débats entourant la sanction de ces projets de lois. Seuls les articles faisant la manchette et se présentant comme compte rendu des débats entourant l'adoption d'une loi ont été retenus. Les éditoriaux et les articles de fond ont donc été éliminés parce qu'ils représentaient l'expression d'un point de vue extérieur aux débats en chambre.

Cet ensemble fut l'objet d'un échantillonnage aléatoire systématique, en raison de sa taille imposante. Pour chacune des quatre législatures au cours desquelles les projets de lois ont été débattus, nous avons retenu cinq articles. D'abord, pour chaque législature, les articles ont été numérotés suivant la date de leur parution. Ensuite, le nombre total d'articles trouvés dans le cadre de chaque législature fut divisé par cinq, afin d'obtenir la séquence de sélection des articles. Enfin, pour chaque législature, un premier article fut retenu au hasard dans le premier segment et les quatre autres en fonction de la séquence définie plus haut.

Ce sous-corpus a dû être édité en raison de la nature même des comptes rendus journalistiques. Les comptes rendus des débats en chambre sont, en effet, truffés d'interventions du journaliste visant à situer le contexte, à décrire l'atmosphère, à relier la discussion actuelle à un débat antérieur ou encore à commenter des attitudes, des comportements susceptibles d'éclairer le lecteurs sur la psychologie des protagonistes. Ces aspects du reportage et tous ceux qui d'une manière ou d'une autre renvoient au traitement journalistique de l'information ont été éliminés. Un patient travail d'édition a alors consisté au repérage, puis à l'éradication du discours proprement journalistique afin de ne conserver que les propos tenus à l'assemblée législative tels que rapportés par le journal. Cette épuration du texte ne garantit pas que l'objectivité du discours ait pu être restituée, mais sans doute s'agit-il là de la meilleure façon de reproduire empiriquement la substance du débat.

L'édition du texte aura ensuite nécessité que soit reconstitué le style direct de l'énonciation. En effet, le compte rendu journalistique implique par définition que le discours soit rapporté sous la forme du "il". Des propos sont imputés à un locuteur qui les a lui-même énoncés sous la forme du "je". Le travail a consisté à transposer un discours tenu à la troisième personne du singulier à la première en fonction de règles de transformation strictes. C'est sous cette forme que ce sous-corpus a été finalement saisi sur support magnétique.

Pour les fins du présent livre, nous avons aussi éliminé toute intervention qui n'était pas attribuable à un élu de l'Union nationale. Le texte correspondant aux interventions des membres de l'opposition officielle et des tiers partis a ainsi été systématiquement éliminé. Le discours législatif représente donc un condensé épuré et serré des positions duplessistes sur le plan législatif durant la période.

Le discours électoral

Le sous-corpus du discours électoral est composé des interventions publiques des candidats de l'Union nationale durant les campagnes de 1944, 1948, 1952, 1956 et 1960. De la même façon que pour le discours législatif, la constitution de ce sous-corpus a été effectuée à partir de comptes rendus journalistiques. Le journal Le Devoir a servi de source à la cueillette des textes correspondant à toutes les campagnes électorales sauf celle de 1956 pour laquelle nous avons dû recourir au journal La Presse, en raison du boycott par l'Union nationale dont faisait l'objet Le Devoir à cause de ses positions trop ouvertement critiques vis-à-vis du régime. Les positions du Devoir incitèrent les dirigeants de l'Union nationale à interdire aux journalistes de ce journal l'accès à ses assemblées politiques.

Les articles retenus aux fins de l'analyse sont ceux qui faisaient la manchette. Il s'agit plus précisément des comptes rendus de discours prononcés par les candidats de l'Union nationale lors d'assemblées publiques. La plage temporelle à l'intérieur de laquelle s'est effectuée la collecte est délimitée par les dates correspondant à l'annonce de la tenue de l'élection jusqu'au jour du scrutin.

Tous les articles présentés en manchette et portant sur les propos tenus par les candidats unionistes durant la campagne ont d'abord été réunis. L'échantillonnage a été réalisé suivant les même principes que ceux qui ont été appliqués au discours législatif. Les articles colligés furent regroupés en fonction des campagnes électorales auxquelles ils correspondaient, avant d'être l'objet d'un échantillonnage aléatoire systématique. Encore ici, le critère d'échantillonnage résidait dans l'obligation de retenir cinq articles par campagne. Le sous-corpus définitif fut obtenu au terme de l'opération décrite plus haut.

L'édition du discours électoral présentait le même genre de difficultés que celles que nous avons rencontrées pour le discours législatif. Le traitement journalistique de même que les traits de mise en pages et d'édition (titres et sous-titres) ont été éliminés. Comme pour le législatif, la transcription du discours électoral a été modifiée de manière à le ramener à la forme du discours direct.

Le discours constitutionnel

Ce sous-corpus a été constitué à partir des publications gouvernementales regroupant les procès-verbaux des Conférences fédérales-provinciales des premiers ministres. Les textes des conférences de 1945-46, 1950, 1955 et 1957 ont été retenus. C'est dire que toutes les conférences fédérales-provinciales tenues durant la période étudiée ont été considérées.

Pour les fins du présent livre, seuls les interventions des représentants du gouvernement du Québec, le plus souvent de Maurice Duplessis lui-même, furent conservées. Plus précisément, le sous-corpus réunit à la fois les interventions majeures des représentants québécois et les discussions surgissant spontanément dans le feu du débat. S'ajoutent à cela les mémoires déposés par la province de Québec qui énoncent les positions du Québec sur certains aspects particuliers des réformes constitutionnelles proposées par le gouvernement fédéral.

Le texte original n'a subi que peu de transformations. Nous avons éliminé les tableaux statistiques et les passages à caractère exclusivement technique ou protocolaire (heure de reprise des discussions, numéro de salle de rencontre, convocations à des banquets, etc). Le texte a été saisi sur support magnétique sous cette forme à peine modifiée.

Le tableau suivant résume quantitativement les caractéristiques du corpus de l'union nationale. On retrouvera, pour chacun des sous-corpus du quadruplet, le nombre total de mots qu'il contient et, parmi ces derniers, le nombre de mots auxquels a été attribuée une catégorie socio-sémantique et, enfin, la participation de chaque sous-corpus au corpus de l'Union nationale en fonction du nombre de mots.

TABLEAU 1

Description statistique du corpus de l'Union nationale

TrôneÉlect.Légis.Const.Total

Nbr. de mots224092717375465397801648 27

Nbr. de mots 5091499612915644629448

catégorisés

% de mots22.7%18.4%17.1%16.2%17.86%

catégorisés

Participation 13.6%16.5%45.8%24.1%100%

au corpus

Corpus religieux

Nous n'avons retenu du corpus religieux que celui constitué par les Mandements des évêques. Aussi nous faut-il définir ce que sont ces mandements. Ce sous-corpus regroupe les textes publiés par les diocèses, sous le titre: “Mandements, lettres pastorales et circulaires des Évêques”. Nous avons retenu les Mandements des diocèses de Montréal, Québec, Sherbrooke, Hauterive et Chicoutimi en raison de leur représentation du territoire québécois. Il faut ajouter que ces Mandements constituent un ensemble discursif hétérogène. Ils sont principalement constitués de trois types de documents: les documents de régie interne (53.42% du total des documents), les textes théologiques et dogmatiques (20.59%) et les énoncés de doctrine sociale (25.99%). Nous nous sommes intéressé exclusivement aux derniers. Les textes de régie interne sont, en effet, des directives de l'archevêché à l'adresse du clergé et les documents dogmatiques ou théologiques débattent de questions avant tout religieuses. Par contre, le discours social de l'Église s'adresse à la population et aux fidèles par le biais du clergé. Leur format peut varier d'une lettre pastorale collective, à une lettre épiscopale, à une allocution radiophonique, etc., mais ces discours ont tous en commun d'exposer la doctrine sociale de l'Église sur les rapports sociaux et les relations de l'individu à la société.

Bien que n'ayant retenu que les discours qui porte directement sur les rapports sociaux que l'on retrouve dans le sous-corpus de doctrine sociale, les mandements représentaient un ensemble trop vaste qu'il a donc fallu échantillonner. Après avoir fixé une limite approximative de quarante pages de texte par année tout au long de la période, nous avons adopté un premier principe qualitatif pour l'échantillonnage. Les textes retenus devaient apparaÎtre dans aux moins trois des cinq diocèses et, ce faisant, constituaient le plus souvent des lettres pastorales signées par un grand nombre d'Évêques. Ce principe a suffi à déterminer le corpus pour la plupart des années de la période retenue. Dans les cas où cette méthode générait un nombre largement supérieur à quarante page, nous avons procédé à un échantillonnage aléatoire simple.

Le corpus religieux n'a pas été l'objet de procédures complexes d'édition. Nous n'avons éliminé que les titres et les sous-titres, ainsi que les formules conventionnelles utilisées à la fin des lettres, précisant le lieu premier de la publication, la date et la liste œs signataires. Le tableau 2 résume quantitativement les caractéristiques du corpus religieux.

TABLEAU 2

Description statistique du corpus des Mandements des Évêques pour la période 1944-1960

Nbr. de mots185724

Nbr. de mots29711

catégorisés

% de mots16%

catégorisés

Catégorisation

L'analyse effectuée dans ce livre s'appuie sur le corpus global que nous venons de décrire, enrichi d'une description morpho-syntaxique et socio-sémantique. L'ensemble des opérations produites sur ce corpus a été abondamment présent é dans Restons traditionnels et progressifs. Nous ne reviendrons pas sur la description morpho-syntaxique puisque celle-ci n'aura servi dans la présente démarche qu'à identifier les candidats à la catégorisation socio-sémantique. Quant à cette dernière, il ne saurait être question de reprendre extensivement la description qui en a déjà été faite dans l'ouvrage cité. Nous ne présenterons ici que les informations essentielles à la compréhension de cette opération, en nous limitant à l'utilisation particulière que nous avons faite de ce système de catégories dans la présente analyse.

Sens de la catégorisation

La catégorisation est au fondement de toute connaissance. Elle permet de rapprocher ce qui, à certains égards, est semblable et de distinguer ce qui, d'un même point de vue, est différent. Elle consiste fondamentalement à attribuer à un objet du monde une catégorie parmi l'ensemble des catégories appartenant à un système donné. Ces systèmes peuvent être de nature plus théorique ou empirique, selon le degré de formalisation des relations que l'on peut décrire entre les éléments qui le constituent. Il est donc nécessaire de définir d'abord la nature de l'objet à catégoriser et des systèmes de catégories qui lui seront appliqués.

Nous avons choisi de retenir le mot comme unité à catégoriser, plutôt qu'un segment textuel plus large. Formellement, ce choix est beaucoup plus facile, car il ne comporte aucune décision sur la définition matérielle du mot. L'unité lexicale présente également l'avantage de pouvoir être traitée avec grande facilité au plan informatique. Le choix d'un segment plus large poserait à la fois des problèmes méthodologiques et théoriques. Seul un segment équivalant à la phrase serait repérable sur la base formelle de la ponctuation. Tous les autres cas demanderaient soit une analyse syntaxique (par exemple, dans le cas des segments propositionnels), soit une analyse qualitative (par exemple, dans le cas des segments thématiques). De toute manière, la catégorisation d'objets complexes devient problématique sur le plan théorique puisqu'elle présuppose un travail de description préalable (implicite ou explicite). Pour toutes ces raisons, nous nous en sommes tenus aux mots. Il en découle que l'analyse mise en oeuvre est fondée sur la présence et la récurrence plus ou moins forte de mots dans le discours, même si cette présence est mise en relation avec le contexte d'apparition.

Nous avons retenu deux systèmes de catégories, l'un morpho-syntaxique et l'autre socio-sémantique. Le premier système, inspiré 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 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 la production d'une sémantique particulière.

Qu'en est-il alors de cette grille de catégories socio-sémantiques? S'il est possible de dire que le système de catégories morpho-syntaxiques a un fondement théorique, il est plus difficile de l'affirmer en ce qui concerne le système socio-sémantique que nous proposons. En effet, les catégories syntaxiques correspondent peu ou prou aux unités de base qui seront activées par le système de règles que représente toute théorie syntaxique. Au contraire, malgré leurs prétentions, les sémantiques générales, linguistiques ou conceptuelles, n'arrivent pas à définir de tels systèmes de règles qui permettraient de rendre compte du fonctionnement et des interrelations qui caractérisent leurs unités. À plus forte raison, toute sémantique appliquée à des domaines de discours ne saurait prétendre former de tels systèmes théoriques. La grille que nous proposons est donc 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. Comme nous l'avons fait valoir plus haut, l'application de catégories aux mots du texte n'a pourtant 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. Les études lexicométriques classiques comportent des limites sérieuses en ce qu'elles obligent à accéder au sens à partir des mots pris isolément. L'application de catégories socio-sémantiques aux mots présente au contraire l'avantage de regrouper dans des classes d'équivalence des mots renvoyant à une même “signification”. Nous pouvons alors observer des régularités 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 doit être validée. 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.

Processus de catégorisation

La catégorisation a été effectuée sur l'ensemble du corpus par une équipe de codeurs sous la supervision constante d'un coordonnateur. Les dispositifs informatiques dont nous disposions n'ont pu que partiellement alléger le travail que représente la catégorisation. Dans le cas de la catégorisation morpho-syntaxique, un logiciel de catégorisation par projection de dictionnaires et application de règles locales (CBSF) nous a permis de catégoriser automatiquement près de 85% des occurrences sans référence au contexte. Les 15% résiduels ont dû être désambiguïsés avec l'assistance d'un système informatique donnant accès au contexte et permettant l'apposition manuelle de catégories. La catégorisation socio-sémantique a, inversement, été effectuée hors contexte pour à peine 15% des occurrences. Plus de 85% des mots ont reçu leur catégorie par l'application de la procédure de catégorisation manuelle avec visionnement du contexte. Cela s'explique facilement du fait que l'appartenance inconditionnelle à une classe socio-sémantique est un phénomène très rare et qu'il est difficile d'imaginer des règles susceptibles de départager automatiquement la polysémie des mots.

Les codeurs étaient donc appelés à choisir parmi les différentes appartenances socio-sémantiques possibles d'un mot, celle qui était 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.

Contenu de la grille

On trouvera à la fin de la présente annexe la liste des catégories et leur définition. Nous nous contenterons ici d'en donner une description très générale. La grille de catégories est constituée de cinq familles. Les trois premières renvoient à des institutions particulières et aux objets qu'elles génèrent. Les catégories économiques rendent compte à la fois des concepts et notions générales de l'activité économique, de certains acteurs économiques et des principales institutions de l'économie marchande. Les catégories politiques désignent les principales institutions de la sphère publique, les concepts et notions de l'activité politique ainsi que les acteurs qui s'y adonnent. La troisième famille regroupe les institutions sociales qui échappent aux deux premières. Elle porte sur des institutions de l'espace public et de l'espace privé ainsi que sur les acteurs qui y correspondent. L'univers social regroupe les représentations de l'espace et du temps ainsi que les dimensions démographique, sociale et communautaire de l'identité. Enfin, les catégories renvoyant aux valeurs forment un cinquième ensemble. Cette famille se subdivise en divers sous-groupes représentant des dispositifs différents du système de référence ontologique de la société. Ainsi, y retrouve-t-on les valeurs traditionnelles, les valeurs émancipatrices, les valeurs de contrôle social, les valeurs éthiques et les valeurs existentielles.

Utilisation des catégories dans l'analyse

Dans le présent livre, nous avons privilégié avant tout l'étude des valeurs. L'analyse nous a souvent conduit à examiner bien d'autres catégories, mais nous avons construit le propos sur la base d'un examen approfondi de l'univers des valeurs dans les différents discours analysés. Il importe donc de définir plus avant chacun des sous-ensembles de valeurs afin d'en préciser la portée. Il faut d'abord dire que le concept même de valeur est difficile à définir. Il est, dans l'ensemble de la littérature, posé comme une évidence dont la définition n'a pas besoin d'être précisée. Les valeurs sont des notions qui cimentent le discours et orientent l'action. Contrairement aux autres catégories de la grille, elles ne désignent pas des objets ou des relations entre ces objets. Elles constituent des référents idéaux.

Il nous est apparu, au cours du long processus de la recherche, que ces valeurs marquaient profondément la nature même du discours politique. Certes, celui-ci contribue à la production d'un bloc social, articulant de manière spécifique des institutions et des forces sociales, définissant les frontières spatiales et temporelles d'une communauté. C'est donc dans l'examen de l'ensemble des autres catégories que nous avons pu retrouver ce travail. Mais, c'est aussi un discours qui se particularise par sa définition de référents qui soutiennent l'organisation des rapports sociaux et définissent la relation entre l'individu et la société. Nous avons donc privilégié l'observation des dispositifs de valeurs.

Les valeurs traditionnelles et les valeurs émancipatrices représentent les deux dispositifs alternatifs qui permettent de définir la référence ontologique de la société. C'est donc à travers l'étude du comportement de ces ensembles de valeurs que nous avons pu réfléchir sur la véritable nature, traditionnelle ou moderne, du duplessisme. Les valeurs de contrôle social

permettent, par ailleurs, d'étudier les dispositifs idéologiques qui sont mis en oeuvre pour gérer les rapports entre classes et forces sociales. Enfin, les valeurs éthiques permettent de voir comment la relation de l'individu à la société est définie. Ces deux derniers ensembles constituent, selon nous, la base du discours disciplinaire qui est promu par le duplessisme. Enfin, un ensemble résiduel définit les valeurs existentielles qui renvoient à l'univers du bonheur tel que le définit Aristote.

Informatique

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.

Génération d'une base de données lexicales.

La démarche globale de traitement en SATO comporte deux phases de traitement distinctes: la génération des fichiers SATO et leur interrogation. Le texte original devra être préalablement édité selon certains critères minimaux. La première phase de traitement (SATOGEN) consiste à récupérer le texte, ainsi édité, et à le transformer en un lexique de formes lexicales. Celles-ci sont définies dans SATO comme des chaÎnes de caractères situées entre deux espaces ou délimiteurs. SATO reconnaÎt, de cette façon, les entités du texte, mots, ponctuations, paragraphes et références de pagination, auxquelles il attribue une adresse informatique à partir de laquelle le texte pourra être recomposé et interrogé suivant diverses combinaisons de ces variables. La structure des données qu'adopte SATO permet alors de consulter les données hors-contexte du point de vue du lexique, ou, en contexte, dans des segments textuels de longueur variable.

Partition du corpus

Cette phase de mise en forme complétée, le texte peut être interrogé à l'aide du module d'interrogation (SATOINT). Celui-ci regroupe un ensemble de commandes donnant accès à plusieurs fonctionnalités du logiciel. Une première commande "Domaine" permet de partitionner le texte afin de centrer l'analyse sur des segments textuels particuliers ou de comparer le fonctionnement de plusieurs segments en fonction d'un ou d'un ensemble d'indicateurs. Cette partition peut s'effectuer à partir de certaines bornes inhérentes à la structure du texte (lignes, paragraphes ou documents composant le corpus) ou d'autres principes délimiteurs définis par l'utilisateur. Dans ce livre, la principale partition correspond

aux cinq sous-corpus que nous avons définis. Comme on l'a vu, il nous a été possible d'interroger le corpus dans son entier, un sous-corpus particulier ou encore la combinaison de sous-ensembles particuliers. Nous avons aussi créé de nouvelles partitions en vertu de la combinaison de diverses variables (par exemple, les Mandements selon deux périodes différentes) ou encore sur la base du partage d'un même trait (par exemple, la présence d'un mot ou d'une catégorie dans des segments donnés du texte). Ces domaines sont définis à l'écran au moment de la description ou de l'exploration des données. Ils peuvent aussi être redéfinis en tout temps suivant la progression de la démarche.

Module de catégorisation

Une des fonctionnalités les plus importantes de SATO est certes son module de catégorisation. Les mots ou des entités textuelles plus larges peuvent recevoir ou voir modifier des catégories qui serviront, par la suite, à structurer la recherche d'information. Ce module permet d'associer aux mots du texte, ou aux formes lexicales, des propriétés numériques ou symboliques, en contexte ou hors contexte. La catégorisation hors contexte est la plus simple. Il s'agit d'attribuer, indépendamment du contexte, une valeur de propriété à une mot ou une forme lexicale. Toutes les occurrences de ce mot ou de cette forme posséderont cette valeur. Pour chaque propriété, c'est comme si l'on ajoutait une colonne au catalogue des formes du texte devenant ainsi interrogeable selon cette dimension. Dans le cas de la catégorisation en contexte, c'est une ligne d'information que l'on ajoute dans le corps du texte, concernant une caractéristique se rapportant à une occurrence d'un mot ou d'un segment textuel. La valeur de propriété ainsi accolée à cette occurrence peut être transférée dans le lexique qui reçoit ainsi une nouvelle entrée. En effet, le mot peuple ayant reçu la valeur de propriété “communauté” apparaÎtra comme une entrée différente du mot peuple n'ayant pas reçu cette valeur.

Patrons de fouilles

Une fois catégorisé et les marques de segmentation déposées, le texte peut alors être interrogé suivant plusieurs modalités prévues par SATO. Une syntaxe simple et efficace admet comme élément de recherche soit l'expression littérale d'un mot, soit une combinaison de caractères de remplacement permettant notamment des jeux de troncation à gauche, à droite ou à l'intérieur des chaÎnes de caractères. Les requêtes peuvent être également une combinaison de mots et/ou de descripteurs (catégories). Dans tous les cas, ces patrons de fouille permettent de produire des lexiques d'occurrences et de cooccurrences ainsi que des concordances à contexte variable.

Lexiques d'occurrences

Le logiciel 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 "Canad" dans les discours du Trône et législatif dont la fréquence dans le texte est supérieure à 12 occurrences. SATO produit instantanément un lexique, ventilé par sous-corpus s'ils ont été préalablement définis comme domaines, regroupant dans ce cas des mots tels "Canada", "canadien" , "canadienne", ... Ces lexiques peuvent être ordonnés en fonction de la fréquence des mots qu'ils regroupent, de leur ordre alphabétique, de leur longueur, etc.

Lexiques de cooccurrences

SATO facilite également l'étude du covoisinage à travers la production de lexiques de cooccurrence. Le logiciel offre la possibilité de repérer et de dénombrer, suivant une multitude de paramètres possibles, la co-présence de formes lexicales. 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.

Concordances

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 le texte 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,...) 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.

Analyses de données

SATOINT contient enfin certaines fonctionnalités d'analyse de données: calculs de participation relative, de lisibilité, de distance statistique entre sous-ensembles textuels,... Sato est également muni d'une interface permettant de récupérer les résultats du dépouillement pour traitement statistique ultérieur. l'essentiel du raisonnement que nous avons poursuivi dans ce livre porte sur les lexiques de cooccurrences. Ces lexiques présentent la liste des mots ordonnés en fonction de la signification de leur association au mot pôle. Aucune mesure de signification n'étant disponible dans SATO, nous avons développé avec l'aide précieuse de Guy Cucumel un test de significativité des cooccurrences. Ce test nous permet d'attester la signification d'une association forte ou faible entre un mot cooccurrant avec le mot pôle au delà d'un certain seuil statistique. De manière générale, ce seuil a été fixé à 95%. Nous avons retenu le seuil de 90% dans de rares cas, en mentionnant que l'association avait moins de chance d'être significative.

Fondement mathématique du test de cooccurrence

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 0.05.

Si $f_{pc} \leq e$, on calcule la probabilité que X soit inférieur 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. Dans ce deuxième cas, c'est la rareté de la cooccurrence qui est significative.

Ce calcul pouvant être répété pour tous les mots cooccurrents avec le mot pôle, on peut obtenir un classement des cooccurrences les plus significatives. La cooccurrence la plus significative étant celle associée à la probabilité la plus faible. Enfin, notre méthode se généralise aisément à des p-occurrences, soit des cooccurrences muliples, avec p aussi grand que l'on veut.