6.3. Pluralisme et pluralité des interprétations

6.3.1. *Introduction*

Nous distinguerons dans un premier temps pluralisme et pluralité des interprétations. Par la suite, nous nous en tiendrons à la pluralité des interprétations qui se déploie dans la chaîne des opérations d'analyse de matériaux discursifs à l'aide de l'ordinateur. Nous montrerons à travers deux exemples comment des ressources informatiques permettent de valider les interprétations à l'oeuvre dans différentes opérations de la recherche. Nous nous arrêterons d'abord au processus de catégorisation à partir de deux expériences de recherche[1] ayant utilisé le logiciel SATO. Nous nous intéresserons ensuite à la validation de l'interprétation des résultats en triangulant les méthodes proposées par divers logiciels, SATO, ALCESTE, LEXICO3 ET DTM[2]

6.3.1. *Diversité des interprétations*

La diversité des interprétations peut se concevoir selon deux axes. Le premier, horizontal, renvoie au problème de la multiplicité des points de vue possible dans l'observation et l'analyse d'un même objet. Nous parlons alors de pluralisme des interprétations. Le second, vertical, désigne la pluralité des choix herméneutiques qui s'imposent à toutes les étapes de la recherche empirique conduite d'un même point de vue. Le pluralisme des interprétations s'explique du fait que non seulement les objets empiriques sont découpés en fonction des diverses disciplines et problématiques de recherche, mais qu'ils sont surdéterminés par des points de vue épistémologiques, méthodologiques et herméneutiques [DUC 99A]. Le pluralisme des interprétations s'explique par la complexité des choix possibles que le chercheur doit effectuer à ces divers plans.

La pluralité des interprétations renvoie plutôt à la dimension interprétative de l'ensemble des opérations de connaissance produites dans une démarche de recherche [DUC 99B]. Nous distinguons l'interprétation globale qui survient au terme d'une telle démarche et les interprétations locales qui sont autant de choix méthodologiques

---

[1] La première expérience porte sur l'analyse exploratoire d'un corpus d'entrevues sur l'usage du tabac et l'influence de messages antitabac. [DAO 06] [GEL 04] La seconde expérience est tirée d'une étude exhaustive de la représentation de la communauté politique dans les discours des Premiers Ministres du fédéral et des provinces dans le cadre des conférences constitutionnelles au Canada de 1941 à 1992. [BOU 96]

[2] Voir la référence des logiciels en bibliographie.

produits à toutes les étapes de la recherche. Illustrons ce point en montrant le chemin parcouru dans un processus d'analyse de texte par ordinateur. Analyser un texte à l'aide de ressources informatiques implique une transformation successive des états de ce texte [MEU 90][3]. Le passage entre deux états du texte implique des opérations de transformation. Le texte *initial* est le discours lui-même en tant qu'il mobilise des ressources cognitives, linguistiques et culturelles et qu'il s'inscrit dans des conditions de production et de réception qui lui sont propres. Le texte discours est un objet complexe et son analyse implique une série de réductions. La première se réalise dans le passage du discours au texte *manuscrit*. Cette transcription du discours implique des choix quant à l'information qui sera retenue. Doit-on ne retenir que le dire du discours ou doit-on garder la trace de ses formes expressives. Doit-on conserver les marques de l'énonciation et les éléments para-linguistiques qui l'accompagnent ? Ces choix non seulement déterminent l'espace à venir de l'interprétation, mais sont déjà une interprétation donnée d'un premier texte. Le deuxième passage s'effectue entre le texte *manuscrit* et le texte *électronique*. Des opérations de saisie (manuelle ou automatisée), de normalisation (orthographique, syntaxique, sémantique, lemmatique), de traitement (de l'information péritextuelle ou textuelle), de gestion (base de données, partitions) sont autant de lieux où des choix sont effectués sur la base de ce qui apparaît important ou indispensable de conserver. La perte d'information est de nouveau le fruit d'une interprétation de ce qui est valable et une restriction de l'espace des interprétations possibles dans les phases ultérieures de la recherche. Le texte édité devient le matériau à partir duquel les opérations subséquentes de la recherche vont s'appliquer. Sur la base de cette matrice, un ensemble de transformations successives de ce texte produiront des versions enrichies du même texte. Le texte *représentation* se présente donc comme des versions successivement enrichies sur la base de la description de ses unités à l'aide d'un ou de plusieurs systèmes de catégories (morpho-syntaxique, sémantique, rhétorique, etc.) rattachés à des théories plus ou moins descriptives ou explicatives. L'attribution d'une catégorie à un mot ou à un segment textuel sur la base de telles théories implique un choix interprétatif qui aura également un impact sur les interprétations ultérieures.

À partir du texte édité ou des diverses versions du texte *représentation*, il sera possible de produire autant de sous-textes combinant les partitions du texte et les diverses descriptions. Ces sous-textes peuvent aussi bien provenir d'explorations des données décrites que de lectures guidées par des hypothèses a priori. Mais dans les deux cas, les hypothèses induites ou déduites formulent des interprétations possibles du texte. Enfin, le dernier passage est celui qui va de l'ensemble des sous-textes prenant la forme de *résultats* vers le texte reconstruit comme *interprétation* globale. Il

[3] Nous reprenons l'idée de Jean-Guy Meunier selon laquelle les traitements successifs du texte initial produiraient autant de nouvelles versions du texte : du texte discours au texte *manuscrit*, puis au texte *électronique*, au texte *représentation*, aux sous-textes *résultats*, enfin au texte *interprétatif* final.

y a, au terme de l'opération, réappropriation de l'objet à connaître dans un texte nouveau dont la complexité est restaurée mais qui diffère du discours initial. Ce texte final est celui de l'expert en sciences sociales. C'est ici qu'il y a possiblement hiatus entre les données et le sens qui leur est restitué. Autant les interprétations locales peuvent être maîtrisées dans la mesure où on est conscient de les produire à travers chaque choix méthodique, autant les interprétations globales sollicitent de nouveaux aspects cognitifs, linguistiques, culturels et des conditions de production et de réception qui sont propres à un nouveau discours.

6.3.2. *Interprétations et dispositifs expérimentaux.*

Ce qui distingue l'analyse textuelle assistée par ordinateur du simple commentaire interprétatif, c'est la construction de dispositifs expérimentaux qui visent à construire des faits qui soutiennent l'interprétation. Selon Benoît Habert, le dispositif expérimental est un « montage d'instruments, d'outils et de ressources destinés à produire des « faits » dont la reproductivité et le statut (l'interprétation) font l'objet de controverses » [HAB 05]. Dans sa définition du dispositif expérimental, Habert indique qu'un instrument, c'est « un dispositif expérimental qui a réussi ». L'instrument est donc un dispositif stabilisé dont le mode d'emploi et l'interprétation des résultats produits fait l'objet d'un certain consensus. Quand on procède à une nouvelle analyse mobilisant de nouvelles questions de recherche ou visant un discours dont le fonctionnement reste à expliciter, on doit élaborer un dispositif expérimental original adapté à des hypothèses nouvelles, ou à tout le moins, à des hypothèses qui se déploient dans des fonctionnements discursifs spécifiques.

D'un point de vue technique, le dispositif expérimental se matérialise par des procédures de calculs transparentes et reproductibles, et par des procédures assistées de catégorisation dont la trace doit être explicite. Cela signifie que les critères de cette catégorisation sont clairement exprimés et qu'il est possible de retourner au corpus pour repérer les balises qui sont la marque physique de la codification. Ainsi, la controverse de l'interprétation pourra s'appuyer sur la discussion serrée des procédures de constitution des faits sur lesquels elle s'appuie.

L'utilisation des procédures informatisées a aussi pour objectif de permettre la coexistence de plusieurs dispositifs expérimentaux construits sur un même corpus. Ces dispositifs, matérialisant divers points de vue et perspectives théoriques, peuvent soutenir à la fois la complémentarité des points de vue et la multiplicité des parcours interprétatifs correspondant à la nature plurielle intrinsèque de la lecture.

6.3.3. *Analyse exploratoire et construction itérative de grilles de catégories*

Pour illustrer le processus de construction itérative d'une grille catégorielle basée sur une analyse exploratoire de corpus, nous utiliserons le cas d'une recherche portant sur un corpus d'entrevues sur l'usage du tabac et l'influence de messages antitabac [GÉL 04]. Les entrevues ont été réalisées en 2000 auprès de 48 jeunes Français répartis en neuf groupes. Le point de départ de la démarche consistait à comparer, à l'aide d'indices statistiques simples, les lexiques associés à des sous-textes découpés d'après les variables de stratification établies au départ : sexe, fumeur/non-fumeur, avant/après le message antitabac. L'établissement du corpus se situe donc dès le départ dans un contexte interprétatif qui vise à répondre à un certain nombre de questions de recherche ayant trait à l'influence de messages antitabac sur un public cible. C'est ainsi que chaque entrevue se déroule en deux temps. On a d'abord un premier échange amorcé par quelques questions de l'intervenant. Ensuite, l'intervenant introduit une brochure antitabac et la discussion se poursuit suite à la présentation de ce message dissuasif.

Pour comparer les lexiques selon le profil des locuteurs et les étapes de l'entrevue, les chercheurs ont utilisé un algorithme de distance lexicale basé sur la distance du Chi2. La mesure évalue l'écart dans l'utilisation d'un vocabulaire donné entre deux sous-ensembles du corpus. Les formes lexicales sont triées par ordre décroissant de contribution à la mesure de distance, ce qui permet d'identifier, par ordre d'importance, les spécificités de chaque sous-texte. Ils ont également utilisé un algorithme de participation qui calcule les moyennes normalisées d'un ensemble de formes lexicales, correspondant généralement à une catégorie lexicale, pour chacun des sous-textes constitués en cours d'analyse. Cette démarche exploratoire réalisée à l'aide du logiciel SATO [DAO 05] se fonde sur un va-et-vient interactif entre ce que révèle l'analyse lexicale et les contextes d'utilisation des mots mis en évidence par les algorithmes de distance et de participation.

Ainsi, l'application de l'algorithme de distance sur les fréquences lexicales associées aux interventions avant et après l'introduction de la brochure anti-tabac ont permis de constater que les mots qui décrivent l'apparence physique et la santé en général sont ceux qui caractérisent le plus vocabulaire avant l'introduction de la brochure. On trouve aussi des mots évoquant le plaisir et la dépendance. À l'inverse, après l'introduction de la brochure, on retrouve en tête de liste les mots : *témoignage, concret, solution, chiffres, mort.*

Le passage du lexique des formes non décrites au lexique catégorisé, fruit d'un dispositif expérimental explicite, autorise un niveau d'interprétation ayant une portée plus générale. Par exemple, le Tableau I permet de visualiser la sur représentation et la sous représentation d'une catégorie de la grille dans divers sous-textes rassemblant les interventions des locuteurs d'après leur profil sociologique. Ici, l'analyseur

PARTICIPATION calcule la fréquence relative de la catégorie *mort*. *A* et *B* désignent avant et après la brochure. Nous avons aussi les particules *fu* et *nf* pour fumeur et non-fumeur, ainsi que *ho* et *fe* pour homme et femme. On constatera que le thème de la mort ressort plus après l'introduction de la brochure et que cette saillance positive est caractéristique des non-fumeurs des deux sexes, et des femmes, fumeuses ou pas.

Tableau I : Analyseur PARTICIPATION (sujet=mort)

Propriété	**Couverture**	**Lexèmes**	**Occurrences**	**Cote Z**
Fréqtot	78703/78703 (100.00%)	9/3985 (0.23%)	80/78703 (0.10%)	0.00
A	23544/78703 (29.91%)	4/2087 (0.19%)	19/235440 (0.8%)	-1.01
B	28074/7870335 (67%)	6/2351 (0.26%)	47/28074 (0.17%)	3.46
Afu	13758/78703 (17.48%)	4/1580 (0.25%)	13/13758 (0.09%)	-0.26
Bfu	15923/78703 (20.23%)	6/17490.(34%)	24/15923 (0.15%)	1.94
Anf	9786/7870312.(43%)	2/1240 (0.16%)	6/9786 (0.06%)	-1.25
Bnf	11898/78703 (15.12%)	3/1425 (0.21%)	23/11898 (0.19%)	3.14
Aho	14468/78703 (18.38%)	4/1634 (0.24%)	8/14468 (0.06%)	-1.75
Bho	16010/78703 (20.34%)	4/1797 (0.22 %)	21/16010 (0.13%)	1.17
Afe	9076/78703 (11.53%)	2/1153 (0.17%)	11/9076 (0.12%)	0.58
Bfe	11811/78703 (15.01%)	5/1379 (0.36%)	26/1181 (0.22%)	4.04

C'est en s'appuyant sur l'analyse lexicale de ces données brutes que les chercheurs ont pu élaborer des modèles interprétatifs permettant d'inscrire les points d'ancrage lexicaux dans des systèmes de catégories sémantiques[4] et énonciatives susceptibles de traduire, dans le discours même, l'attitude des jeunes par rapport au tabagisme et l'influence de publicités dissuasives. La catégorisation visait donc à établir le pont entre la problématique de recherche et les données textuelles en comparant les interventions avant et après l'introduction de la brochure selon le profil sociologique des sujets. Suite à la catégorisation des mots caractéristiques, les chercheurs ont

[4] La grille traduit, en 28 catégories, les ancrages lexicaux de l'interprétation (le préfixe soc- renvoie à un ensemble de catégories référant aux rapports sociaux identifiés par les jeunes) : apparence, arrêt, négation, concret, danger, dépendance, soc-je, maladie, mort, plaisir, publicité, tabac, nicotine, drogue, interdiction, fumeur, soc-ami, soc-famille, soc-gens, liberté, envie, conscience, volonté, soc-jeune, coûts, début, santé, éducation, prévention.

procédé à un examen systématique des mots triés par ordre alphabétique afin de repérer les variantes flexionnelles pertinentes. Par exemple, l'analyseur DISTANCE a montré une présence importante du pronom « j »' après l'introduction de la brochure, ce qui suggérait que la brochure ait pu provoquer une plus grande implication personnelle de nos sujets. Pour valider cette hypothèse, il fallait ajouter à la catégorie tous les pronoms renvoyant à la première personne sous l'hypothèse qu'il s'agirait là d'une marque de prise en charge de l'énoncé par l'énonciateur. Pour confirmer leurs intuitions, les chercheurs ont repris les analyses de distance et de participation en les appliquant cette fois sur les catégories de la grille et leur fréquence. Ainsi, pour la catégorie *soc-je*, il s'est avéré que la différence de fréquence observée pour la forme *j'* ne valait plus pour la catégorie. L'interprétation du phénomène *j'* devra donc être trouvée ailleurs.

La construction d'une grille catégorielle s'appuie sur un protocole d'analyse de corpus qui se veut à la fois transparent et respectueux de la spécificité du contexte d'énonciation. C'est une démarche itérative qui combine l'approche inductive, souvent associée aux méthodes qualitatives, l'utilisation d'outils simples de statistique lexicale, et une approche plus sensible à la pragmatique textuelle. Ce traitement a l'avantage de produire des données qualifiées qui traduisent la démarche interprétative de l'analyste en fournissant des points d'ancrage importants pour l'ensemble de la chaîne interprétative verticale. Une fois établie, la valeur interprétative de la catégorisation doit encore être validée au-delà du corpus témoin par l'application extensive et méthodique de la grille sur un corpus élargi, comme l'illustre la section suivante.

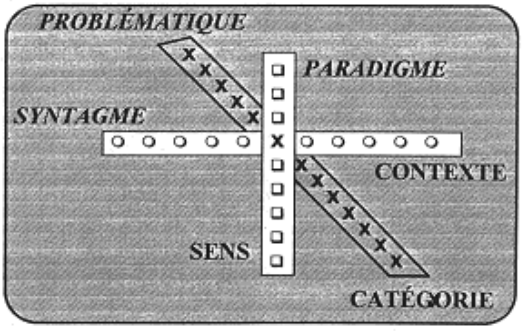
6.3.4 Processus de catégorisation à partir d'une grille

Nous donnerons l'exemple d'un processus de catégorisation qui a été effectuée dans le cadre d'une recherche sur le discours constitutionnel canadien [BOU 96]. La catégorisation s'est effectuée à partir d'une grille de catégories socio-sémantiques élaborée dans le cadre de recherches antérieures sur des corpus de discours politique au Québec et au Canada et affinée au stade exploratoire de cette recherche. Deux questions se posaient alors aux chercheurs : quelle est la nature de l'acte de catégoriser et comment assurer la stabilité et la fiabilité d'une telle opération? L'attribution d'une catégorie à une unité du discours consiste, d'une part, à distinguer cette unité d'autres unités dont le sens diffère et, d'autre part, à regrouper cette unité avec d'autres unités dans une classe d'objets sur la base d'une communauté de sens. Il s'agit bien pour le chercheur de produire une interprétation locale quant au sens à donner à cette unité. Les critères traditionnels de méthode exigent que cette opération de catégorisation soit à la fois valide, fiable et stable. La validité mesure le degré d'adéquation entre une problématique théorique et la description des objets d'observation. La fiabilité et la stabilité visent à assurer que cette mesure d'adéquation soit constante dans le temps et dans l'espace. Cela implique qu'un

même codeur à divers moments et plusieurs codeurs à un même moment effectuent les mêmes choix sur un même objet.

Il est utile de décomposer l'acte de catégoriser selon trois axes afin de mieux en saisir la complexité et, par la suite, de définir des stratégies informatiques susceptibles d'accroître la stabilité des choix interprétatifs. Lorsque le chercheur se trouve devant une unité de discours à catégoriser, il dispose de trois critères que l'on peut représenter selon trois axes : syntagmatique, paradigmatique, problématique (voir Schéma). Chaque unité de codage se trouve d'abord sur un axe syntagmatique (la phrase, le paragraphe ou toute unité de contexte décidée par le chercheur). Le sens prêté à une unité de codage tiendra compte de ce contexte. En effet, la catégorisation socio-sémantique doit le plus souvent recourir au contexte en raison de la variation des sens possible de chaque unité. Le second axe est l'axe paradigmatique, c'est-à-dire l'ensemble des sens possibles qu'une unité peut revêtir. Le codeur choisira le sens correspondant à son usage en contexte. Dans une recherche sociologique, il ne suffit pas de connaître l'ensemble des significations que peut emprunter une même unité, il faut rapporter ce mot à une grille de catégories correspondant à la problématique de la recherche. Le codeur choisira donc sur la base d'une signification possible (axe paradigmatique) dans un contexte donné (axe syntagmatique) une catégorie correspondant à une grille (axe problématique) pour l'appliquer à une unité du discours (mot ou segment).

Critères pour l'identification du sens des mots à catégoriser



Nous n'entrerons pas ici dans tous les détails de la catégorisation socio-sémantique du corpus constitutionnel. Les candidats à la catégorisation étaient limités aux substantifs et aux adjectifs identifiés suite à l'application d'un dictionnaire morpho-syntaxique (BDL en SATO) et désambiguïsés en contexte. Un second dictionnaire de mots jugés a priori non pertinents en relation avec la théorie socio-politique retenue fut également projeté sur le lexique. Enfin, un troisième dictionnaire de mots dont le sens était jugé invariable nous a permis de réduire de quelque vingt pour cent le nombre de candidats à la catégorisation. Les noms et les adjectifs

résiduels ont dû être catégorisés en contexte. Afin d'assurer la fiabilité et la stabilité, nous nous sommes appuyés sur des fonctionnalités de SATO. Chaque mot différent, candidat à la catégorisation (dans ce cas, les substantifs et les adjectifs n'ayant pas été éliminés dans les phases précédentes) fut catégorisé dans une même opération. Nous obtenons, par exemple, une concordance du mot Québec, ce qui nous permet d'examiner tous les contextes où le mot apparaît. Cette méthode accroît considérablement la stabilité des décisions du fait même de la synchronicité de la procédure. Le mot Québec peut avoir différentes significations selon le contexte. Il peut s'agir de la Province du Québec en tant que territoire, du Gouvernement du Québec (sous la forme : Le Québec demande...), de la ville de Québec ou, encore, du peuple du Québec (Le Québec souverain). Dans chaque cas, la catégorie variera. Deux outils additionnels nous permettent d'accéder à l'information paradigmatique et problématique. Devant la tâche d'appliquer une catégorie à une occurrence de Québec, le codeur peut être intéressé à connaître quels sont les choix de catégories s'offrant à lui.

Dans la mesure où une pareille grille de catégories a été appliquée sur un autre corpus ou dans le cas où des décisions ont déjà été prises sur des occurrences du mot Québec dans le corpus soumis à l'étude, il est possible d'obtenir instantanément la liste des catégories déjà appliquées au mot Québec (voir Tableau II).

Tableau II : Lexique des catégories du mot Québec

Fréqtot	**socio**	**FED**	**QUE**	**PHQ**	**AUT**	**P-FED**	**P-QUE**	**P-PHQ**	**P-AUT**	
13	nil	0	8	5	0	0	0.02	0.00	0	québec
36	us7b8	15	7	14	0	0.02	0.02	0.01	0	québec
127	et2b2	7	85	31	4	0.01	0.22	0.01	0.03	québec
2	us7b3	0	1	1	0	0	0.00	0.00	0	québec
335	us7b2	18	165	137	15	0.03	0.43	0.05	0.10	québec
1	(et2b2,u*	0	1	0	0	0	0.00	0	0	québec
7	us2c4	0	7	0	0	0	0.02	0	0	québec

Une seconde information est aussi disponible. Il s'agit de la liste des mots qui ont reçu une catégorie donnée. Ainsi, le codeur peut vérifier l'univers de sens qui est couvert par l'une ou l'autre catégorie (voir le tableau III). Ici, on demande au logiciel de nous indiquer les mots qui ont reçu la catégorie espace territorial.

Tableau III : Lexique des mots ayant reçus la catégorie « us7b2 »

Fréqtot	socio	FED	QUE	PHQ	AUT	P-FED	P-QUE	P-PHQ	P-AUT	
2	us7b2	0	0	2	0	0	0	0.00	0	est
10	us7b2	1	3	6	0	0.00	0.01	0.00	0	province
335	us7b2	18	165	137	15	0.03	0.43	0.05	0.10	québec
7	us7b2	0	4	3	0	0	0.01	0.00	0	québécois
4	us7b2	0	4	0	0	0	0.01	0	0	québécoise
2	us7b2	0	0	2	0	0	0	0.00	0	territoriaux

L'exercice de catégorisation est l'exemple paradigmatique de la dimension herméneutique liée à toute opération de recherche. L'utilisation des fonctionnalités de SATO a rendu possible l'explication des procédures impliquant des choix interprétatifs et a permis d'accroître la fiabilité et la stabilité du processus dans son ensemble.

6.3.4 Validation par triangularisation des méthodes

Pour paraphraser le terme triangulation des données utilisé en analyse qualitative, on pourrait utiliser l'expression triangulation des méthodes pour désigner la nécessité de vérifier jusqu'à quel point les procédures de traitement influent sur la stabilité des faits construits sur lesquels s'appuie l'interprétation. Pour illustrer ce processus, nous reprendrons l'exemple du corpus d'entrevues sur l'usage du tabac [DAO 06] Dans cette expérience, les chercheurs ont mis à contribution des méthodes et des logiciels inspirés de la tradition française d'analyse des données[5] : ALCESTE, DTM, LEXICO.

Dans leur approche initiale avec SATO, les chercheurs ont contrasté le lexique global en découpant le corpus selon des variables externes au texte et qui ont trait à certaines caractéristiques sociales des locuteurs. La méthode ALCESTE procède de façon inverse. Le logiciel construit une classification des énoncés, dont l'approximation statistique correspond à des segments de texte de longueur comparable. Ainsi, ALCESTE tente de faire émerger la structure du discours par le dépistage de profils de répétition dans les énoncés simples. Il est ensuite possible de juxtaposer les variables externes sur les classes d'énoncés et sur leur vocabulaire caractéristique. Appliqué au corpus d'entrevues, ALCESTE produit deux classes. La première classe est fortement caractérisée par des interventions exprimées après

[5] Voir le chapitre 5.2 de Christophe Lejeune

l'exposition au message antitabac. On trouve aussi, mais plus faiblement, une présence significative des interventions des non-fumeurs. La deuxième classe est fortement caractérisée par des interventions précédant la présentation du message antitabac. On trouve aussi, mais plus faiblement, une présence significative des interventions des fumeurs

Il serait possible d'interpréter les classes créées par ALCESTE en tentant de caractériser le vocabulaire caractéristique des classes. Par exemple, la première classe fait ressortir des thèmes tels que la prise de conscience (*voir, choquer, choc, image, témoignage*), la mort et la maladie (*cancer, poumon, mort*), la médiatisation (*pub, télé, spot, prévention routière*). Les interventions après le message antitabac touchent des thèmes plus graves et marquent une réaction par rapport aux campagnes de publicité. Mais, ce qui attire surtout l'attention ici, c'est qu'ALCESTE confirme que la variable avant/après le message antitabac représente le premier élément de structuration du corpus confirmant ainsi l'hypothèses a priori utilisée avec SATO.

Pour construire leur grille, les chercheurs avaient utilisé l'algorithme de distance pour contraster le vocabulaire. Ils ont voulu vérifier la convergence de cet indice avec le test statistique utilisé par LEXICO [SAL 03]. Il s'agit du calcul des spécificités qui fait appel au modèle de la loi hypergéométrique. Les chercheurs ont constaté qu'il y avait un très large recouvrement entre les formes lexicales qui contribuent le plus à la distance et les spécificités calculées par LEXICO.

Le logiciel DTM [LEB 05] se présente comme un outil dédié à l'analyse exploratoire de données numériques multivariées et de données textuelles. L'exemple type de données admissibles au logiciel est la compilation de sondages comprenant à la fois des réponses à des questions fermées et à des questions ouvertes. Pour l'analyse des entrevues, les chercheurs ont utilisé ce modèle de couplage des questions ouvertes et fermées. Ils ont considéré le corpus comme un ensemble de 87 individus. La question ouverte est unique et sa réponse est l'ensemble des interventions du participant, l'avant et l'après message antitabac étant considérés comme deux questionnaires distincts. Les questions fermées traduisent le profil sociologique de l'individu (sexe et usage du tabac) et les conditions d'énonciation (avant ou après la brochure).

DTM procède à une analyse factorielle des correspondances croisant ces 87 individus et les 903 formes lexicales dont la fréquence est supérieure à quatre. Ensuite il trace les variables sociologiques dans l'espace construit par l'analyse factorielle des correspondances (AFC). Le graphique permet de constater que la répartition dans l'espace lexical des caractéristiques sociologiques des jeunes reprend les oppositions repérées par l'algorithme de distance. On obtient donc les mêmes clivages que l'on parte, avec SATO, des variables sociologiques pour contraster le lexique, ou que l'on parte, avec ALCESTE, des énoncés pour construire des mondes lexicaux et y placer

les traits sociologiques, ou, finalement, que l'on parte, avec DTM, du texte découpé en individus pour contraster les traits sociologiques. Trois méthodes différentes confirment la stabilité du modèle interprétatif.

L'avantage des analyses multivariées, c'est qu'elles tiennent compte de l'ensemble des données pour confirmer l'existence d'un profilage sociologique au sein même du discours. En contrepartie, l'interprétation de ces profils est difficile puisque la construction de cette représentation est purement algébrique. À l'opposé, si elle s'appuie dans un premier temps sur des mots saillants repérés par la distance du Chi2, la grille catégorielle s'élabore sur des bases sémantiques. On écarte des unités lexicales jugées trop circonstancielles et on y ajoute d'autres unités contribuant à l'une ou l'autre des catégories socio-sémantiques. Pour vérifier la stabilité des résultats avec le lexique catégorisé, on construit un corpus dans lequel chacun des mots est remplacé par sa catégorie, y compris les mots non catégorisés représentés par un symbole arbitraire unique. Le système d'axes devient alors plus facilement interprétable puisqu'il se base maintenant sur un vocabulaire réduit à 29 catégories. On dispose ainsi d'un très bel outil de validation de la construction de la grille de catégories socio-sémantiques. Ainsi, on voit alors s'étaler aux quatre points cardinaux les catégories les plus excentriques et donc les plus signifiantes de notre système interprétatif : *apparence, dépendance, coûts, éducation, mort et soc-ami*. À l'inverse, on voit apparaître au centre du plan les catégories banales qui constituent les référents communs du discours.

6.3.5 Conclusion

Allant au-delà de l'observation descriptive et du commentaire, la démarche illustrée ici montre comment l'interprétation peut s'appuyer sur des méthodologies transparentes et explicites rendues possibles par l'utilisation de programmes informatiques. Ainsi, la construction de grilles catégorielles peut s'appuyer sur des indices statistiques qui tiennent compte de la globalité du corpus. L'application de la grille gagne en systématicité et en précision. Enfin, par la combinaison des méthodes d'analyse, on augmente la fiabilité des conclusions en fournissant des moyens de corroborer ou d'infirmer des hypothèses et des conclusions.

Bibliographie

[BOU 96] Bourque, G. et DUCHASTEL, J., (avec la collaboration de V. ARMONY). L'identité fragmentée. Nation et citoyenneté dans les débats constitutionnels canadiens, 1941-1992. Montréal: Fides, 383 pages, 1996.

[DAO 05] Daoust, F. Système d'analyse de texte par ordinateur, Version 4.2, Centre ATO, UQAM, http://www.ling.uqam.ca/sato/satoman-fr.html

[DAO 06] Daoust, F., Dobrowolski, G., Dufresne, M., Gélinas-Chebat, C., Analyse exploratoire d'entrevues de groupe : quand ALCESTE, DTM, LEXICO et SATO se donnent la main. Actes des JADT-2006, vol. 1, pp- 313-326, Les Cahiers de la MSH Ledoux no. 3, Presses universitaires de Franche-Comté, 2006.

[DUC 99A] Duchastel, J. et Laberge, D. « La recherche comme espace de médiation interdisciplinaire », Sociologie et Sociétés, vol. XXXI, no.1, pp. 63-76, 1999.

[DUC 99B] DUCHASTEL, J. et Laberge, D. « Des interprétations locales aux interprétations globales : combler le hiatus", Ramognino, N et Houle, G., Sociologie et normativité scientifique, Toulouse, Presses Universitaires du Mirail, pp. 51-72, 1999.

[GÉL 04] Gélinas-Chebat, C., Daoust, F., Dufresne, M., Gallopel, K., Lebel, M., « Analyse exploratoire d'entrevues de groupe: les jeunes Français et le tabac », Le poids des mots, Actes des JADT-2004, vol. 1, pp- 479-487, Presses universitaires de Louvain, 2004.

[HAB 05] Habert, B., Instruments et ressources électroniques pour le français Ophrys Paris ISBN 2-7080-1119-7 p.2., 2005.

[LEB 05] Lebart, L (2005). Data and Text Mining. École nationale supérieure de télécommunications, Paris. http://www.enst.fr/egsh/lebart/, 2005.

[MEU 90] Meunier, J.-G., "Le traitement et l'analyse informatique des textes", Gestion de l'information textuelle, ICO, vol. 2, III, p. 9-18, sept 1990.

[SAL 03] Salem, A., Lamaille, C., Martinez, W. et Fleury, S. (Manuel Lexico 3, version 3.41. http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/team.htm, 2003.

[REI 02] Reinert, M, Alceste, Manuel de référence, Université de Saint-Quentin-en-Yvelines, CNRS, 2002.