Analyse exploratoire d'entrevues de groupe : quand ALCESTE, DTM, LEXICO et SATO se donnent la main

François Daoust[1], Gaëlle Dobrowolski[2], Monique Dufresne[3], Claire Gélinas-Chebat[4]

[1]Informaticien au Centre ATO – UQAM, doctorant U. de Franche-Comté, Besançon – France

2Étudiante Licence 3, IUP SID, Université Paul Sabatier, Toulouse,

[3]Professeure, Études françaises, Université Queen's – Kingston – Canada ; chercheure associée au Centre ATO – UQAM – Montréal - Canada

4Professeure, DLDL, UQAM – Montréal – Canada

Abstract

The topic of this presentation is to show how the use of various computer assisted text analysis (ALCESTE, DTM and LEXICO3) could be useful to complete and validate findings of a previous one, named SATO. The data provided was gathered from teenagers' discussion groups about tobacco usage. In a first step, the data was transcribed and submitted to SATO where a construct based on iterative process, leads to database building and hypothesis-testing statistical analysis comparing between lexical data from sub-categorization. Then, bridges based in a XML format were developed in the ATONET web. At last, further analyses were performed with the other programs. Not only each program permits to pinpoint the main results but corroborate and invalidate in such case the results obtained with SATO.

Résumé

Cet article montre comment on peut combiner plusieurs logiciels de lexicométrie (ALCESTE, DTM et LEXICO3) pour valider et compléter une analyse SATO, analyse sémantico-statistique basée sur une construction itérative d'une grille catégorielle. Le corpus analysé est constitué de transcriptions d'entrevues de groupe sur l'usage du tabac. Des passerelles faisant appel à un format pivot en XML ont été développées par le réseau ATONET. Ces chaînes de traitement ont permis de reconfigurer les données de telle sorte qu'il a été possible de faire appel aux points forts de chaque logiciel. La combinaison des méthodes d'analyse a permis d'augmenter la fiabilité des conclusions en donnant des moyens de corroborer ou d'infirmer les hypothèses de départs et les conclusions obtenues à partir d'un seul logiciel.

Mots-clés : analyse de textes, entrevues de groupes, jeunes, tabagisme, catégorisation socio-sémantique, approche inductive et itérative, analyse lexicale, logiciels, XML

1. Introduction

Faisant suite à une première recherche exploratoire, dont les résultats furent présentés aux JADT en 2004 (Gélinas-Chebat et al., 2004), nous présentons ici une démarche méthodologique qui combine plusieurs outils de statistique textuelle et qui vise à valider la grille catégorielle traduisant notre interprétation des données par rapport à notre problématique de recherche.

L'analyse exploratoire, réalisée en 2004 grâce au logiciel SATO (Daoust, 1996, 2004) portait sur un corpus d'entretiens avec des groupes de jeunes sur l'usage du tabac et l'influence de messages antitabac. L'article illustrait la construction d'un système de catégories lexicales basé sur une démarche itérative qui contrastait les données partitionnées selon diverses combinaisons de variables sociologiques décrivant les participants et le moment de leurs

interventions. L'objectif de la démarche était donc de s'appuyer sur le profil sociologique des sujets afin de proposer une grille catégorielle permettant de comprendre, à travers les interventions de chacun, l'influence de messages d'avertissement sur l'usage du tabac.

Dans le cadre des groupes de travail du JADT et du réseau ATONET (Duchastel et al, 2005), nous avons voulu valider cette démarche par l'utilisation combinée de divers logiciels de statistique textuelle : ALCESTE (Reinert), DTM (Lebart), LEXICO3 (Salem) et SATO (Daoust).

Dans un premier temps, nous décrirons le contexte de construction du corpus d'entrevues sur l'attitude des adolescents à l'égard du tabagisme et des messages dissuasifs. Nous rappellerons ensuite les conclusions de l'analyse exploratoire et surtout la méthodologie de construction de la grille catégorielle. Dans les sections suivantes, nous décrirons les stratégies que nous avons déployées pour valider cette première analyse en la confrontant aux résultats produits par diverses méthodes d'analyse statistique. Enfin, nous terminerons par une courte discussion sur l'enjeu méthodologique de la combinaison de ces méthodes.

2. Description du corpus

Le corpus analysé, constitué par Karine Gallopel[1], comprend neuf entrevues sur le tabagisme chez les jeunes et leur perception de la publicité antitabac. Elles ont été réalisées à Rennes en 2000 auprès de 48 jeunes Français qui, pour la plupart, fréquentent une institution scolaire et qui sont âgés de 15 à 25 ans. Chacune des séances réunit cinq à six jeunes (fumeurs ou non-fumeurs, hommes et femmes) et un intervenant, et se divise en deux parties. La première partie se déroule après que l'intervenant a posé quelques questions pour amorcer la discussion et la seconde se caractérise par l'introduction d'une brochure. Au début de chaque transcription, les données sociologiques des personnes qui participent à l'entrevue sont précisées : âge, sexe, fumeur/non-fumeur.

Les annotations éditiques ont été remplacées par un balisage symbolique conforme à la syntaxe de SATO. Un astérisque introduit les balises, aussi appelées *propriétés*, et celles de notre corpus sont les suivantes : *locuteur, *sexe, *fumeur, *page et *pub. En voici un exemple :

> ***page=gallo02/11***
> ***pub=brochure *locuteur=s36 *fumeur=non *sexe=h** *Bah, la brochure là, elle nous présente ce qui nous attend si on fume. Mais c'est très... quoi, moi j'ai lu ça, mais je ne sais pas je ne suis pas fumeur, donc je ne ressens peut-être pas ça de la même façon. À la limite on passe dessus comme ça, ça apporte quelques chiffres.*

Ce type de transcription constitue une forme de balisage *pré-XML* facile à lire pour un humain, mais n'offrant pas l'universalité des formats d'échange normés. La plupart des logiciels d'analyse textuelle utilisent un *format propriétaire* impliquant un travail de conversion non trivial pour faire circuler les données d'un logiciel à l'autre. C'est dans ce contexte que plusieurs développeurs de logiciels se sont réunis pour élaborer une stratégie de conversion des formats de données entre les logiciels. Cette proposition, qui fait l'objet d'une communication aux JADT 2006 (Daoust et Marcoux, 2005), est basée sur un format pivot en XML développé à partir de recommandations du *Text Encoding Initiative*. Des programmes *Perl* (passerelles) permettent de convertir les données des formats propriétaires vers le format XML-TEI et du format XML-TEI vers les formats propriétaires. Ainsi pour réaliser les

[1] Nous remercions Karine Gallopel, maître de conférences à l'Université de Rennes en France, qui nous a permis d'analyser son corpus d'entrevues.

travaux dont nous rendons compte dans cet article, nous avons utilisé le corpus en format SATO et nous avons eu recours au logiciel pour reconfigurer le corpus aux fins d'exploitation dans les logiciels ALCESTE, LEXICO et DTM. Ces nouvelles versions du corpus, après exportation par SATO en format XML-TEI, seront converties par les passerelles PERL vers les formats propriétaires ALCESTE, LEXICO et DTM. Voici la liste de ces diverses reconfigurations du corpus.

1- Corpus *Initial*. Ce corpus contient la transcription des entrevues dans leur découpage original en interventions. Pour l'analyse, on exclut les interventions des intervenants.
2- Corpus *Avant*. Ce corpus contient la partie du corpus *Initial* précédant l'exposition au message antitabac.
3- Corpus *Après*. Ce corpus contient la partie du corpus *Initial* suivant l'exposition au message antitabac.
4- Corpus P*articipant*. Ce corpus est le résultat d'une reconfiguration du corpus original. Il rassemble de manière continue l'ensemble des interventions de chaque participant identifié par un nom résumant son profil et suffixé par *a* ou *b* pour identifier le discours du participant *avant* et *après* le message antitabac. Il est à noter que nous avons éliminé du corpus les participants dont le profil sociologique est incomplet afin de concentrer l'analyse sur les variables principales.
5- Participant *catégorisé*. Ce corpus reprend les données du corpus *Participant*, mais dans lequel les unités lexicales thématisées au cours de l'analyse SATO sont remplacées par la valeur de la propriété catégorielle *thème*. Les mots non thématisés restent inchangés.
6- Participant *réduit*. Ce corpus reprend les données du corpus *Participant*, mais en remplaçant tous les mots par les catégories qui leur ont été attribuées lors de l'analyse avec SATO. Les mots qui n'avaient pas été catégorisés seront remplacés par la catégorie vide *x*.

3. Une construction itérative des catégories

Rappelons d'abord la démarche utilisée dans l'analyse initiale avec SATO. Le principe de base de la démarche consistait à comparer, avec des indices statistiques simples, les lexiques associés à des sous-textes découpés d'après les variables établies au départ : sexe, fumeur/non-fumeur, avant/après le message antitabac. Cette comparaison, appliquée au lexique brut, a fourni les premiers indices pour construire un lexique catégorisé reflétant les points d'ancrage de notre chaîne interprétative.

Pour comparer les lexiques, nous avons 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. Il ne s'agit pas d'un test statistique, mais simplement de l'utilisation d'une métrique appliquée à l'espace ou à un sous-espace lexical. 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. L'algorithme peut être appliqué aux formes lexicales elles-mêmes ou aux valeurs de propriétés correspondant à notre catégorisation lexicale. L'approche était essentiellement dichotomique : nous avons comparé un sous-texte à un autre, via leur lexique respectif. On peut aussi avoir recours à un *algorithme de participation* qui calculera 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 se fondait sur un va-et-vient interactif entre ce que révélait l'analyse lexicale et les contextes d'utilisation des mots mis en évidence par les algorithmes de distance et de participation. Au premier niveau de l'analyse, nous avons travaillé sur les

données brutes, c'est-à-dire les formes lexicales elles-mêmes. Nous nous sommes ainsi donné la possibilité de voir apparaître des différenciations portées par la morphologie des mots en termes de nombre, genre, personne, temps. De plus, nous nous sommes intéressés tout autant, sinon davantage, à des marqueurs d'énonciation, comme les pronoms personnels, les marqueurs phatiques, les marques de la négation, de l'interrogation, les verbes épistémiques (croire, penser...), etc. qu'aux mots sémantiquement pleins. C'est en s'appuyant sur l'analyse lexicale de ces données brutes que nous avons élaboré notre grille catégorielle sémantique.

Le retour constant aux énoncés, ne serait-ce que par un parcours rapide des contextes courts de type KWIC (*key words in context*), s'est avéré essentiel pour ébaucher nos hypothèses et inscrire les unités lexicales dans des systèmes de catégories sémantiques et énonciatives susceptibles de traduire, dans le discours même, ce que nous cherchions à comprendre, à savoir 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 nos données textuelles. Elle correspondait un peu à la procédure de codage de l'analyse qualitative à cette différence qu'elle s'appuyait sur des procédures d'analyse lexicale qui permettent de tenir compte de l'ensemble des données et sur l'examen de phénomènes discursifs difficilement repérables par une simple lecture linéaire.

Le parcours rapide des contextes a été la première stratégie de vérification. Ainsi, l'observation d'un ensemble de mots semblait renvoyer à des préoccupations des jeunes concernant leur apparence, le mot *clair* semblait de la même classe sémantique que *jaune* (*doigts jaunes, dents jaunes*). Le retour aux contextes a montré que le mot *clair* n'avait rien à voir avec l'apparence, mais était plutôt utilisé comme marque phatique : *C'est clair, c'est évident*.

Pour vérifier les hypothèses d'interprétation construites à partir des analyses de distance sur les unités lexicales brutes, la deuxième stratégie a consisté à catégoriser le lexique au moyen d'une propriété lexicale que nous avons nommée *thème* et qui traduit en 28 catégories les ancrages lexicaux de l'interprétation. Le préfixe *soc-* renvoie à un ensemble de catégories référents 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*.

La procédure de catégorisation procédait des mots caractéristiques révélés par l'algorithme de distance vers l'ensemble du vocabulaire. Après les mots caractéristiques, nous avons examiné de façon systématique les mots fréquents et complété la catégorisation en examinant le lexique trié par ordre alphabétique pour catégoriser les variantes flexionnelles pertinentes. Pour confirmer nos intuitions, nous avons repris nos analyses de distance et de participation en les appliquant cette fois sur les valeurs de la propriété *thème*.

La notion d'*apparence* s'est alors confirmée. Les sujets, avant la brochure, abordent les effets superficiels du tabagisme, à savoir, la couleur des dents et des doigts, le teint, l'odeur des vêtements et des cheveux… La notion de plaisir ressortait aussi comme thème avant l'introduction de la brochure, ainsi que les notions de dépendance, santé et éducation.

Après l'introduction de la brochure, la catégorie *concret* (impact et solutions) ressortait. Nous avons aussi vu émerger les notions de *volonté, mort* et *maladie*. D'autre part, nous constations que notre hypothèse sur les pronoms personnels de la première personne (*soc-je*) ne se confirmait pas. L'écart observé avant et après la brochure était spécifique à la forme *j'*.

Nous avons affiné l'analyse en comparant les interventions avant et après l'introduction de la brochure selon le profil sociologique des sujets. Ainsi, nous avons été en mesure de constater que ce sont les non-fumeurs qui semblent le plus touchés par la brochure comme en témoigne la dominance des thèmes relatifs aux effets négatifs du tabagisme : *maladie* et *mort*.

Cette 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, ici des échanges oraux analysés sous forme de transcriptions. 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 textuel des données a aussi l'avantage de produire des données qualifiées qui traduisent la démarche interprétative de l'analyste.

4. Analyse avec ALCESTE

Il est intéressant de noter qu'ALCESTE (Reinert, 2002), qui s'inscrit dans la tradition française d'analyse de données initiée par Benzécri (1973, 1981), répond aussi aux préoccupations de l'analyse de discours en psychologie clinique. Notre corpus d'entretiens n'est pas loin, dans sa facture formelle, de ce type de discours. Les deux approches s'appuient donc sur la notion de discours : davantage dans une perspective sociologique avec SATO, et davantage avec une perspective psychanalytique et sémiotique avec ALCESTE.

D'un certain point de vue, la méthode ALCESTE se situe à l'opposé de celle que permet SATO, qui permet la construction itérative de la grille catégorielle. D'abord, contrairement à SATO, ALCESTE propose une méthode complètement automatique qui vise à faire émerger des *mondes lexicaux*. Pour ce, ALCESTE construit des *énoncés simples* dont l'approximation statistique correspond à des segments de texte de longueur comparable respectant les frontières de l'*unité de contexte élémentaire* (UCE), généralement la phrase. Ainsi, ALCESTE tente de faire émerger la structure du discours par le dépistage de profils de répétition dans les énoncés simples, alors qu'avec SATO, nous sommes partis d'hypothèses structurantes du discours pour *faire parler les données.*

Pour l'analyse du corpus avec ALCESTE, nous avons procédé à quatre expérimentations à partir des premières configurations du corpus décrites à la section.2 : corpus *Initial,* corpus *Avant,* corpus *Après et* corpus *Participant*.

ALCESTE a produit 2042 UCE sur le corpus *Initial*, réparties en deux classes qui comptent respectivement environ le tiers et les deux tiers des UCE. La première classe est fortement caractérisée par des UCE provenant des interventions exprimées après l'exposition au message antitabac (Chi2=33.82). On trouve aussi, mais plus faiblement, une présence significative des UCE des non-fumeurs. La deuxième classe est fortement caractérisée par les UCE provenant des interventions précédant la présentation du message antitabac (Chi2=33.82). On trouve aussi, mais plus faiblement, une présence significative des UCE des fumeurs (Chi2=8.81). Le tableau qui suit montre le vocabulaire le plus caractéristique de ces deux classes.

Formes représentatives de la classe n°1		
Chi2	**u.c.e. dans la classe**	**Formes réduites**
100.00	51	cancer+
93.85	38	image+
83.51	31	choc+
82.20	38	poumon+
81.60	35	choqu+er
73.64	42	preventi+f
61.71	23	routier+
53.58	20	temoign+23
53.47	107	voir.
50.88	19	tele
49.69	39	pub+
46.79	22	femme+
45.39	24	mort+
42.83	23	mourir.
42.16	46	tabac+

Formes représentatives de la classe n°2		
Chi2	**u.c.e. dans la classe**	**Formes réduites**
102.21	446	fum+er
68.65	233	arret+er
28.50	95	commenc+er
28.44	170	fum+eur
22.54	64	essa+yer
21.46	87	envi+e
20.22	69	arrete+
19.36	108	cigarette+
17.11	61	paquet+
16.34	64	volonte+
16.04	68	prendre.

À la lumière de ces résultats, il nous serait possible d'interpréter les classes crées par ALCESTE, comme nous l'avons fait pour le tableau de distance de SATO, en regroupant ce vocabulaire autour de plusieurs axes. 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. Ce discours est davantage exprimé par les non-fumeurs. Pour la deuxième classe, nous sommes à même de voir des verbes et des noms qui semblent renvoyer directement à la consommation de cigarette en termes d'arrêt, d'envie et de volonté. Ce vocabulaire est surtout marqué par les fumeurs.

Mais, qu'ALCESTE confirme que la variable *avant/après* le message antitabac représente le premier élément de structuration du corpus constitue pour nous le résultat le plus significatif. Cette dominance est telle qu'elle empêche le repérage d'autres classes. Soulignons de plus la présence de l'opposition *fumeur/non-fumeur* qui est la deuxième variable prise en compte dans l'analyse SATO.

Pour neutraliser l'effet dominant du profil *avant/après*, on a présenté à ALCESTE deux autres corpus contenant respectivement les interventions avant et après le message antitabac.

Pour le corpus *Avant*, on obtient trois classes représentant respectivement 36 %, 48 % et 16 % des 674 UCE retenues. La première classe est marquée par une dominance des fumeurs, la seconde par les non-fumeurs et les hommes tandis que les femmes dominent la troisième. L'analyse du corpus *Après* crée deux classes représentant respectivement 28 % et 72 % des 1168 UCE retenues. La première classe est marquée par une dominance des non-fumeurs. La seconde par les fumeurs.

Enfin, nous avons soumis à ALCESTE le corpus *Participant*. Pour ce corpus, nous avons rassemblé toutes les interventions de chaque participant, ce qui peut avoir pour effet de réunir des énoncés interrompus dans le corpus d'origine. Mais surtout, nous avons éliminé les participants dont le profil sociologique n'était pas connu. C'est sans doute ce qui aura permis de contraster davantage les données de telle sorte qu'ALCESTE produit immédiatement trois

classes rassemblant 64 %, 21 % et 15 % des 877 UCE retenues. La première classe est d'abord caractérisée par les énoncés avant le message antitabac et secondairement par les UCE des fumeurs. La deuxième classe est caractérisée par les non-fumeurs après le message antitabac tandis que la troisième classe est très caractéristique des énoncés après le message et teintée par le discours des femmes.

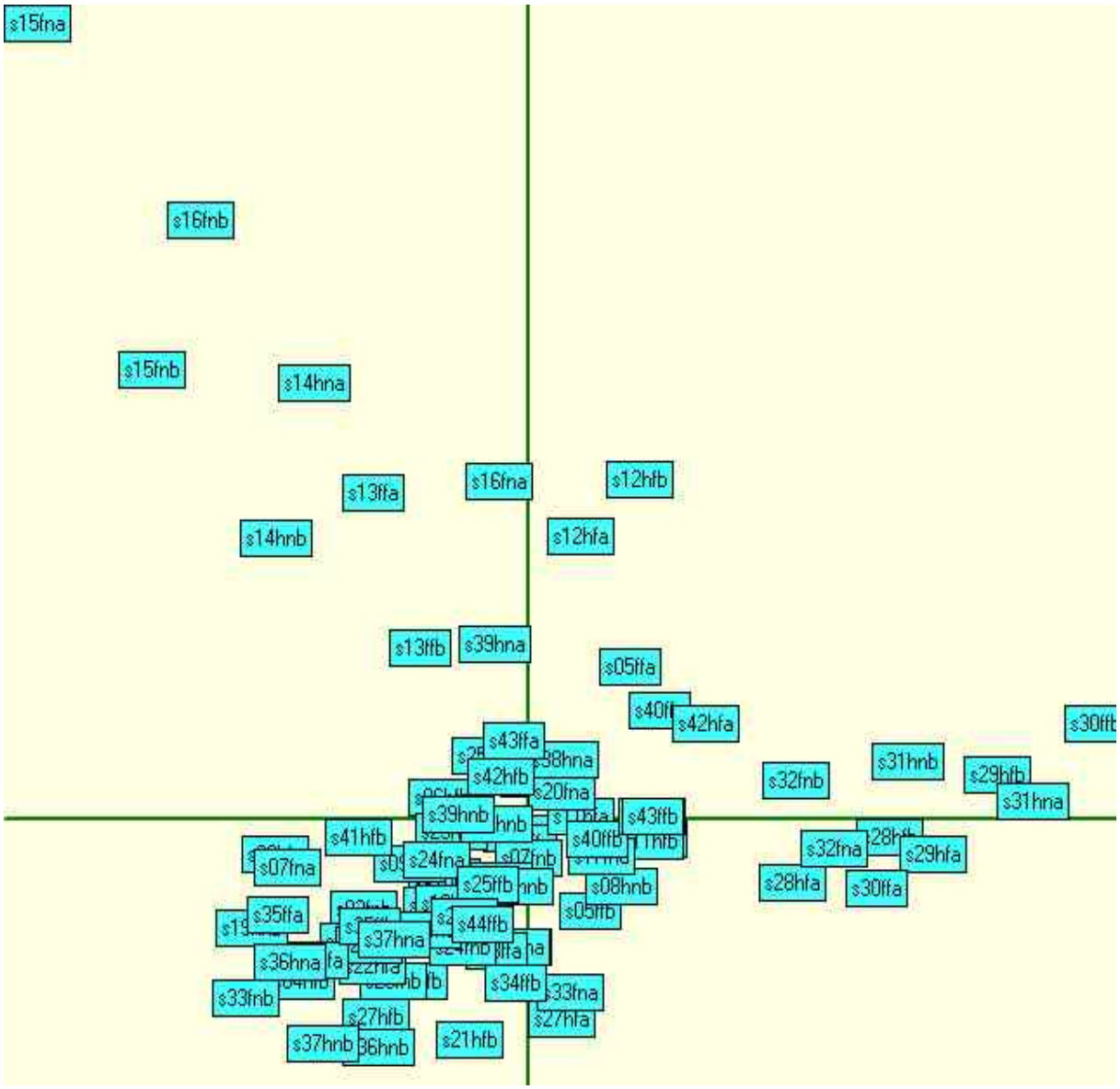
Signalons que le calcul des segments répétés, qu'on trouve aussi dans les autres logiciels statistiques, fait ressortir l'expression *sécurité routière* qui renvoie à une discussion sur une campagne sur la sécurité routière. On aurait donc eu avantage à figer cette expression dans SATO pour la classer sous la catégorie *publicité.*

ALCESTE a donc retrouvé, à partir des énoncés simples, ce que nous avions observé lors de la comparaison de lexiques construits sur la base d'un découpage global du corpus sur la base de profils. Ce point de rencontre entre les approches ascendantes et descendantes est un outil important de validation de l'interprétation.

5. Analyse avec LEXICO

LEXICO (Salem et al, 2003) permet de calculer les spécificités lexicométriques de parties d'un corpus d'après un modèle probabiliste basé sur la loi hypergéométrique (cf. Lebart & Salem, 1994). Il rend aussi possibles des analyses factorielles de correspondances (AFC) sur un corpus partitionné. Les unités décomptées sont exclusivement constituées à partir de la liste des délimiteurs fournie par l'utilisateur (ponctuations et autres caractères non alphanumériques), sans recours à des ressources dictionnairiques extérieures.

Nous avons soumis à LEXICO le corpus *Participant* qui délimite le corpus en participants avant et après la présentation du message antitabac.



Corpus Participant : individus sur le plan des 2 premiers axes de l'AFC

Comme on le constate, il est difficile de tirer des conclusions claires à partir de ce graphique qui situe les individus dans l'espace vectoriel des lexèmes. Certes, les noms des individus décrivent leur profil sociologique, mais on ne voit pas ici de patrons clairs. On reviendra sur ce type d'analyse avec DTM qui permet de tracer les modalités des variables catégorielles sur le plan de l'AFC. Avec LEXICO, on a plutôt utilisé le calcul des spécificités pour repérer les particularités associées à des groupes d'individus. On s'est demandé jusqu'à quel point une mesure probabiliste comme les spécificités pouvait converger ou s'écarter de la simple métrique du Chi2 utilisée par l'analyseur DISTANCE de SATO. On a donc demandé à LEXICO de calculer les spécificités associées aux participants avant l'exposition au message antitabac. Nous avons reporté ces résultats sur la sortie de l'analyseur DISTANCE de SATO appliqué au lexique avant et après le message.

Fréqtot	avant	après	explique	cumul	
0.08	0.15	0.03	0.55	0.55	clair * (lexico 6)
0.05	0.00	0.09	0.50	1.05	brochure
0.25	0.37	0.17	0.49	1.54	aussi * (lexico 6)
0.46	0.60	0.36	0.40	1.94	t' * (lexico 5)
0.07	0.12	0.03	0.39	2.33	santé * (lexico 5)
0.77	0.95	0.64	0.39	2.72	ouais * (lexico 3)
0.02	0.04	0.00	0.32	3.03	appelle * (lexico 4)
0.05	0.01	0.09	0.31	3.35	risques (lexico -5)
0.06	0.10	0.03	0.31	3.66	dépendance * (lexico 5)
0.06	0.10	0.03	0.31	3.96	plaisir * (lexico 5)
1.65	1.88	1.49	0.30	4.26	je * (lexico 3)
0.02	0.05	0.00	0.28	4.54	doigts * (lexico 4)
0.01	0.03	0.00	0.26	4.80	odeur * (lexico 4)
0.16	0.09	0.21	0.25	5.05	elle (lexico -5)
0.11	0.06	0.15	0.24	5.30	beaucoup (lexico -3)
0.03	0.00	0.05	0.24	5.53	lire (lexico -4)
0.13	0.18	0.09	0.23	5.76	toi * (lexico 4)
0.01	0.03	0.00	0.23	5.99	3ème * (lexico 3)
0.03	0.00	0.04	0.23	6.22	témoignage
0.05	0.09	0.03	0.22	6.44	grave * (lexico 3)
0.42	0.32	0.49	0.22	6.66	!
0.26	0.33	0.20	0.22	6.88	ben * (lexico 3)
0.08	0.04	0.11	0.21	7.09	"
0.61	0.49	0.69	0.21	7.30	-
0.44	0.34	0.51	0.21	7.50	peut (lexico -3)
0.02	0.03	0.00	0.20	7.70	caractère * (lexico 3)
0.28	0.20	0.34	0.19	7.90	te (lexico -4)
0.27	0.34	0.21	0.19	8.09	suis * (lexico 3)
0.02	0.04	0.01	0.18	8.26	jaunes * (lexico 3)
0.01	0.02	0.00	0.17	8.44	choqué *(lexico 3)
0.01	0.02	0.00	0.17	8.61	conséquence * (lexico 3)
0.01	0.02	0.00	0.17	8.78	influencé * (lexico 3)
0.01	0.02	0.00	0.17	8.95	net * (lexico 3)
0.01	0.02	0.00	0.17	9.12	publicitaire * (lexico 3)
0.02	0.00	0.03	0.17	9.30	solution
0.01	0.03	0.00	0.17	9.46	aimes * (lexico 3)

Comparaison entre les spécificités et la distance du Chi2

Ce tableau est constitué à partir d'une sortie de l'analyseur *distance* de SATO. La première colonne contient la fréquence d'une forme lexicale dans l'ensemble du corpus. Les deuxième et troisième colonnes indiquent la fréquence de la forme dans l'ensemble des interventions des participants avant et après le message antitabac. La colonne *explique* donne la contribution de la forme lexicale à la mesure de distance entre les parties *avant* et *après* le message antitabac. La colonne *cumul* contient la somme partielle de ces contributions. Suit ensuite la forme lexicale elle-même à laquelle nous avons ajouté manuellement les calculs de spécificité de LEXICO accompagné d'un exposant qui rend compte du degré de signification de l'écart constaté. Un exposant négatif est la marque d'une sous-représentation significative de l'entrée lexicale.

On observe qu'il y a un très large recouvrement entre les formes lexicales qui contribuent le plus à la distance et les spécificités calculées par LEXICO. Parmi les mots manquants, il y a les ponctuations qui, apparemment, ne sont pas prises en compte par LEXICO, de même que les formes absentes dans le corpus *Avant.* Par ailleurs, on retrouve ces formes absentes des spécificités du corpus *Après : brochure* (9), *témoignage* (4), *solution* (4), *image* (3). LEXICO fait ressortir un plus grand nombre de spécificités que celles qui font partie du tableau. Il faut dire que l'analyseur DISTANCE de SATO se limite arbitrairement aux 50 premières formes

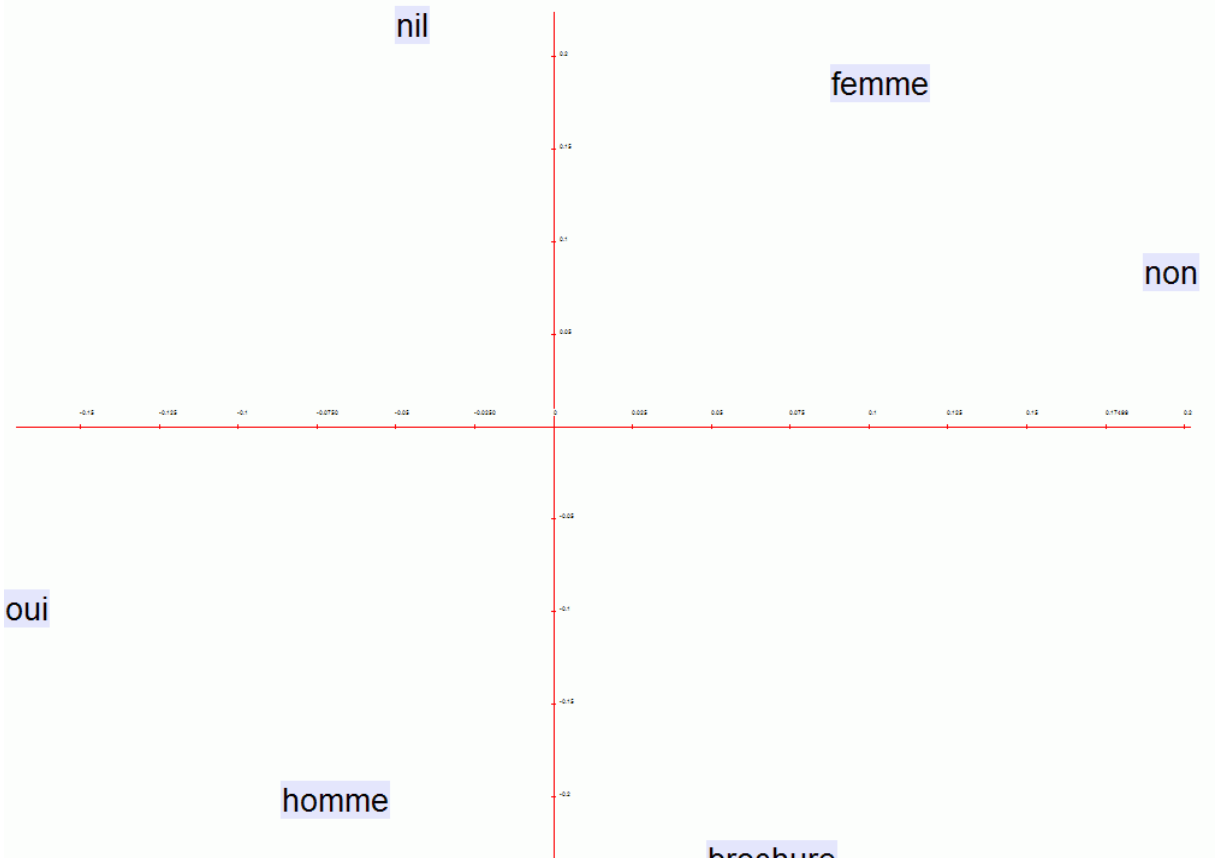
lexicales dans l'ordre de leur contribution à la distance. On voit ici que ce nombre aurait avantage à être augmenté et que la mesure de spécificité de LEXICO fournit un bon complément à la DISTANCE du Chi2 par l'ajout d'un seuil de spécificité.

6. DTM

Le logiciel DTM (Lebart, 2005) 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. Les questions fermées produisent soit des données directement numériques (poids, âge, etc.) ou des données catégorielles qui peuvent être codées par leur numéro d'ordre dans une liste fermée. Les réponses aux questions ouvertes produisent des données qualitatives, du texte brut dont les mots peuvent être comptés produisant ainsi des variables représentant le nombre d'occurrences du mot.

Nous avons utilisé ce modèle de couplage des questions ouvertes et fermées pour l'analyse du corpus P*articipant.* On considère le corpus comme un ensemble de 87 individus. Le profil sociologique est enregistré comme autant de réponses catégorielles à des questions fermées : pub (nil, brochure), sexe (homme, femme) et fumeur (non, oui). La question ouverte est unique et la *réponse* est composée de l'ensemble des interventions du participant, l'*avant* et l'*après* message antitabac étant considérés comme deux *questionnaires* distincts.

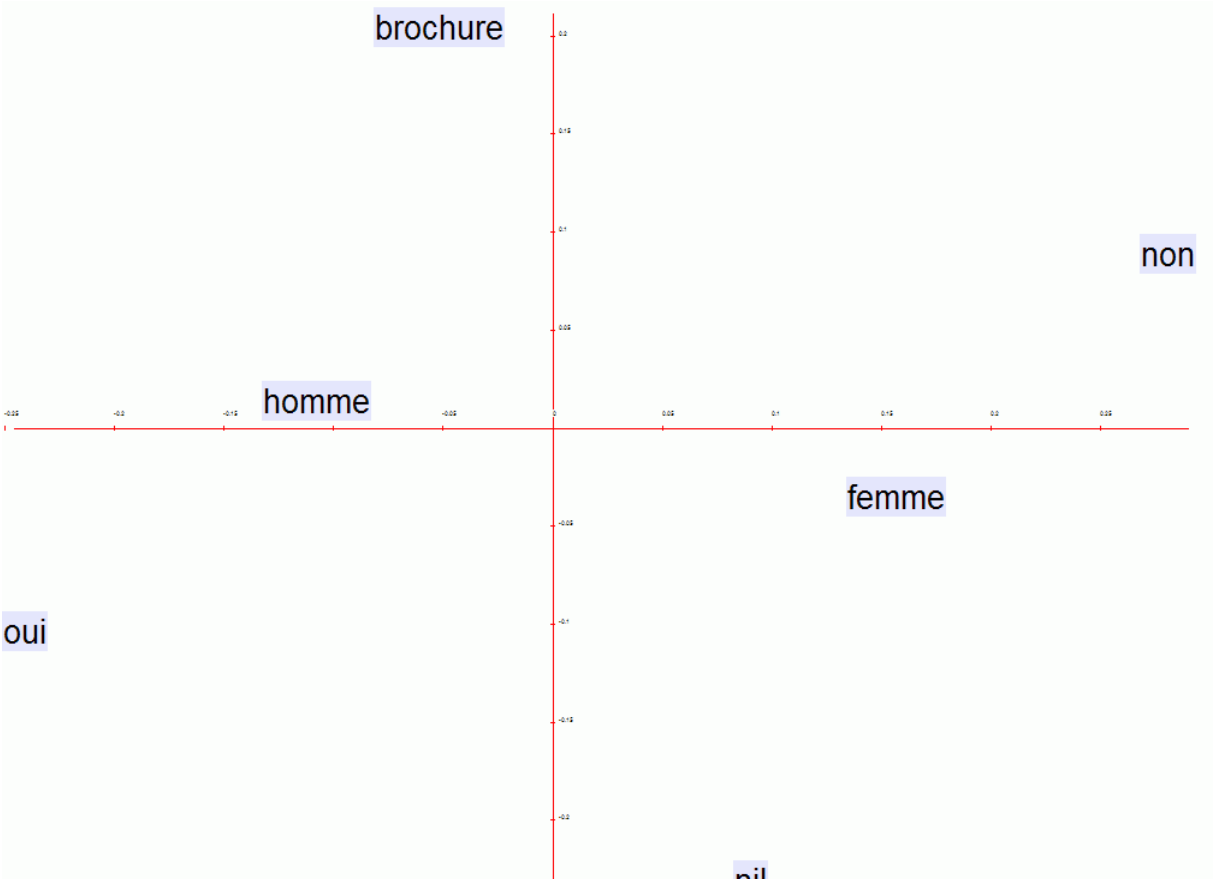
DTM procède à une analyse factorielle des correspondances croisant ces 87 individus et les 903 formes lexicales dont la fréquence est supérieure à quatre. Ensuite DTM trace les variables catégorielles dans l'espace de l'AFC. Les oppositions entre les diverses modalités de nos variables sociologiques apparaissent sur les trois premiers axes de l'AFC. Voici le plan tracé par les deux premiers axes.



Corpus Participant : variables catégorielles sur le plan des 2 premiers axes de l'AFC

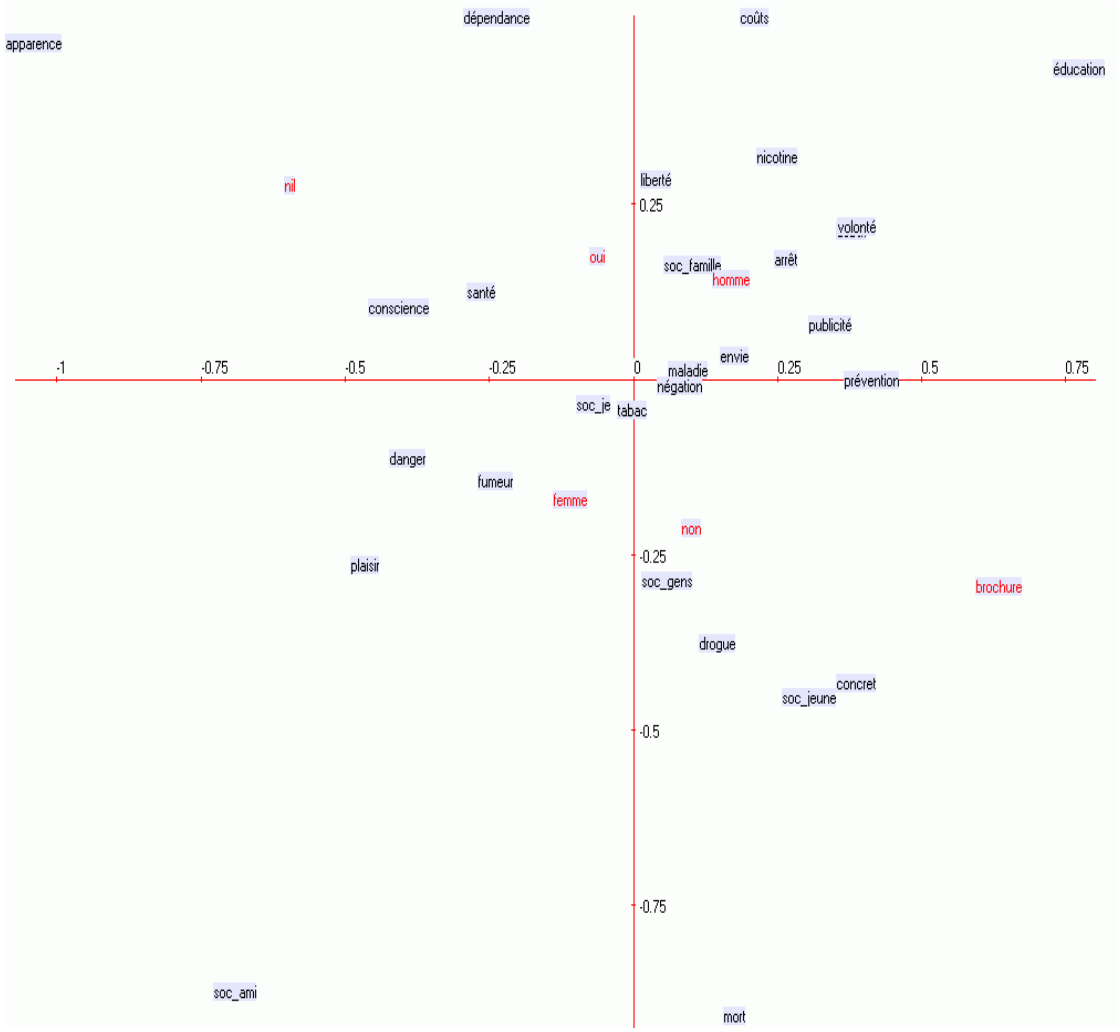
L'influence du message antitabac et de nos variables catégorielles sur la structure du discours semble confirmé. Mais, l'objectif de l'analyse catégorielle avec SATO va au-delà de cette constatation et vise, par la construction d'une grille de catégories lexicales, à interpréter les objets du discours. 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 y écarte des unités lexicales jugées trop circonstancielles et on y ajoute d'autres unités contribuant à une de nos catégories socio sémantiques. Il n'est pas assuré que la classification établie d'après ces critères interprétatifs soit performante d'un point de vue statistique, même si c'est ce que nous souhaitons. Le bon ajustement statistique sera considéré comme un critère de validation de notre grille.

Pour procéder à cette validation, on demandera à SATO de substituer la catégorie aux unités lexicales qu'elle décrit. Ce corpus artificiel, nommé *Participant catégorisé*, suivra la chaîne de traitement habituel : exportation en XML-TEI avec sélection des variables pertinentes, conversion dans le format DTM en utilisant la passerelle ATONET et soumission à DTM. L'AFC est alors calculée en croisant les 87 participants avec 702 variables textuelles composées des formes lexicales non catégorisées et des catégories sémantiques remplaçant les formes lexicales catégorisées sémantiquement (propriété *thème*). Cette substitution recouvre 12,26 % des occurrences. Ici encore, comme on peut le voir, la projection des variables sociologiques sur le plan factoriel suit le même jeu d'oppositions.



Corpus Participant catégorisé variables catégorielles sur le plan des 2 premiers axes de l'AFC

Nous ferons un pas de plus pour confronter notre modèle catégoriel en réduisant toutes les formes lexicales qui ne font pas partie de la grille à une catégorie vide nommée arbitrairement *x*. Le corpus, que nous appellerons *Participant réduit*, contiendra encore le même nombre d'occurrences, mais avec un lexique de 29 unités lexicales seulement se substituant à l'ensemble des occurrences du corpus *Initial* sachant que les 28 catégories utiles représentent un peu plus de 12% des occurrences. Le corpus, ainsi réduit à notre grille socio-sémantique, permet-il toujours de faire les ressortir les variables externes? Voici le graphique.



Corpus Participant réduit : variables catégorielles et lexique sur le plan des 2 premiers axes de l'AFC

Comme l'espace des variables se réduit aux catégories, il est possible de visualiser correctement à la fois le lexique et les modalités des *questions fermées*. On dispose ainsi d'un très bel outil de validation de la construction de la grille de catégories lexicales. Présentée de cette façon, la visualisation des catégories sémantiques dans le plan factoriel ouvre aussi de nouvelles fenêtres d'investigation pour revenir aux contextes et affiner la grille si nécessaire. Il est quand même assez remarquable de voir s'étaler aux quatre points cardinaux les catégories les plus excentriques : *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.

Cette cartographie traduit l'intention même de l'analyse du discours qui vise à faire ressortir les positions sociales à l'intérieur même du langage. 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.

7. Conclusion

Cette première utilisation combinée de logiciels d'analyse textuelle a été grandement facilitée par les protocoles d'échange de données réalisées par le réseau ATONET. Il est ainsi possible de créer de multiples chaînes de traitement qui permettent de reconfigurer les données et de faire appel aux points forts de chaque logiciel. Mais il y a plus. 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. C'est ainsi qu'on peut aller au-delà des impressions et des commentaires descriptifs pour produire des représentations de discours sociaux susceptibles d'agir comme modèles.

Références

Benzécri, J.-P. (1973). *L'Analyse des Données* (tome 1 et 2), DUNOD, Paris.

Benzécri, J.-P. (1981). *Pratique de l'Analyse des Données : linguistique et lexicologie*. DUNOD, Paris.

Daoust, F. (1996, 2004). *SATO 4, Manuel de référence*, Centre ATO, UQAM, Montréal.

Daoust, F. et Marcoux, Y. (2006). Logiciels d'analyse textuelle : vers un format XML-TEI pour l'échange de corpus annotés, *JADT 2006*.

Duchastel, J. et al. (2005). *ATONET, Réseau pour l'échange de ressources et de méthodologies en analyse de texte assistée par ordinateur*. http://www.atonet.net

Gélinas-Chebat, C., Daoust, F., Dufresne, M., Gallopel, K. et Lebel, M.-H. (2004). Analyse exploratoire d'entrevues de groupe : les jeunes Français et le tabac. In Purnelle, G., Fairon, C. et Dister, A. éditeurs, I, *Actes des JADT-2004* : 479-487.

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

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

Reinert, M. (2002). *Alceste, Manuel de référence*. Université de Saint-Quentin-en-Yvelines, CNRS.

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