La sémiométrie : des mots sans texte

Ludovic Lebart[1], Jean-François Steiner[2]

[1] Télécom-ParisTech – ludovic@lebart.org
[2] Semiometrica – steiner@noos.fr

Abstract

Semiometry designates a specific technique for lifestyle and values sample surveys. The corresponding questionnaire consists of a simple list of words that must be scored by the respondents according to the pleasant or unpleasant effects of these words. A Principal Component Analysis of the table (respondents x scores) produces then several stable principal axes spanning a *semiometric space* that can be observed in most western countries. We stress here the links between Semiometry and some problems involved in textual data analysis: Semantic networks, spontaneous semiometry and statistical processing of open-ended questions. Some results updating our book published in 2003 (La Sémiométrie, *in French*) are mentioned. They will be also presented in the forthcoming English version of the book (*in press*).

Résumé

La sémiométrie est une technique d'enquête de type « style de vie et valeurs » reposant sur un questionnaire qui se réduit à une liste de mots. Ces mots doivent être notés par les répondants en fonction de l'effet agréable ou désagréable qu'ils produisent spontanément. Une analyse en composantes principales de ces notes montre l'existence de plusieurs dimensions stables engendrant un *espace sémiométrique* observable dans la plupart des pays occidentaux. On insiste ici sur les liens entre la sémiométrie et les analyses statistiques de textes, en évoquant notamment la sémiométrie spontanée. Certains résultats présentés ici mettent à jour notre ouvrage de 2003 (La Sémiométrie).

Mots-clés : sémiométrie, socio-linguistique, lexicométrie, questions ouvertes, styles de vie

1. Introduction

Sous le nom de *Sémiométrie*, J-F. Steiner a proposé à la fin des années 80 de remplacer les questionnaires des enquêtes par sondage sur le thème « Valeurs, Style de vie » par une simple liste de mots (Benzécri, 1989 ; Steiner et Auliard, 1992). Pour divers points de vue sur les enquêtes du type *valeurs, styles de vie*, on pourra se référer à Fabre et al. (1981), Valette-Florence (1994), Bréchon (2000). Après diverses tentatives et expérimentations, le questionnaire sémiométrique comprend 210 mots à noter (échelle à 7 positions, de -3 à +3) par des individus (selon que ces mots leur évoquent quelque chose de désagréable [note -3] ou d'agréable [note +3]). Les réponses à ces questionnaires pour des échantillons représentatifs de la population soumises à des analyses en composantes principales produisent plusieurs axes stables (dans le temps : périodes différentes; dans l'espace : pays différents). Ces axes engendrent un sous-espace qualifié d'*espace sémiométrique.* Bien que formulée indépendamment et bien qu'elle utilise des outils différents, l'approche sémiométrique a une parenté certaine avec l'approche de sémantique différentielle d'Osgood (1965). L'espace sémiométrique apparaît comme une structure cohérente et multidimensionnelle qui peut servir de grille d'analyse et trouver des applications dans de nombreux domaines (marketing, sociologie, sciences politiques, estimations de réponses manquantes, méthodologie d'enquêtes, etc.). L'ouvrage « *La sémiométrie, Essai de statistique structurale* » (Lebart, Steiner et Piron, 2003) [librement téléchargeable à partir du site *www.dtm-vic.com*] présente

de façon détaillée la technique et ses applications : choix de la liste de mots retenus, différents codages possibles des notes, comparaisons internationales, etc. On en résumera quelques grandes lignes tout en présentant des résultats plus récents. On étudie successivement la stabilité de la structure sémiométrique (section 2), les rapports entre corrélations sémiométriques et liens sémantiques (section 3), et la sémiométrie spontanée (section 4).

2. Stabilité de la structure sémiométrique

La figure 1 représente une esquisse du plan sémiométrique principal (axes 2 et 3 de l'analyse en composantes principales : 210 mots notés par 16 582 personnes interrogées par la SOFRES entre 1990 et 2002). L'axe 1 est un axe méthodologique de notation (facteur de taille, de participation) auquel est consacré tout un chapitre du livre précité. Il n'est pas pris en compte dans l'espace sémiométrique.

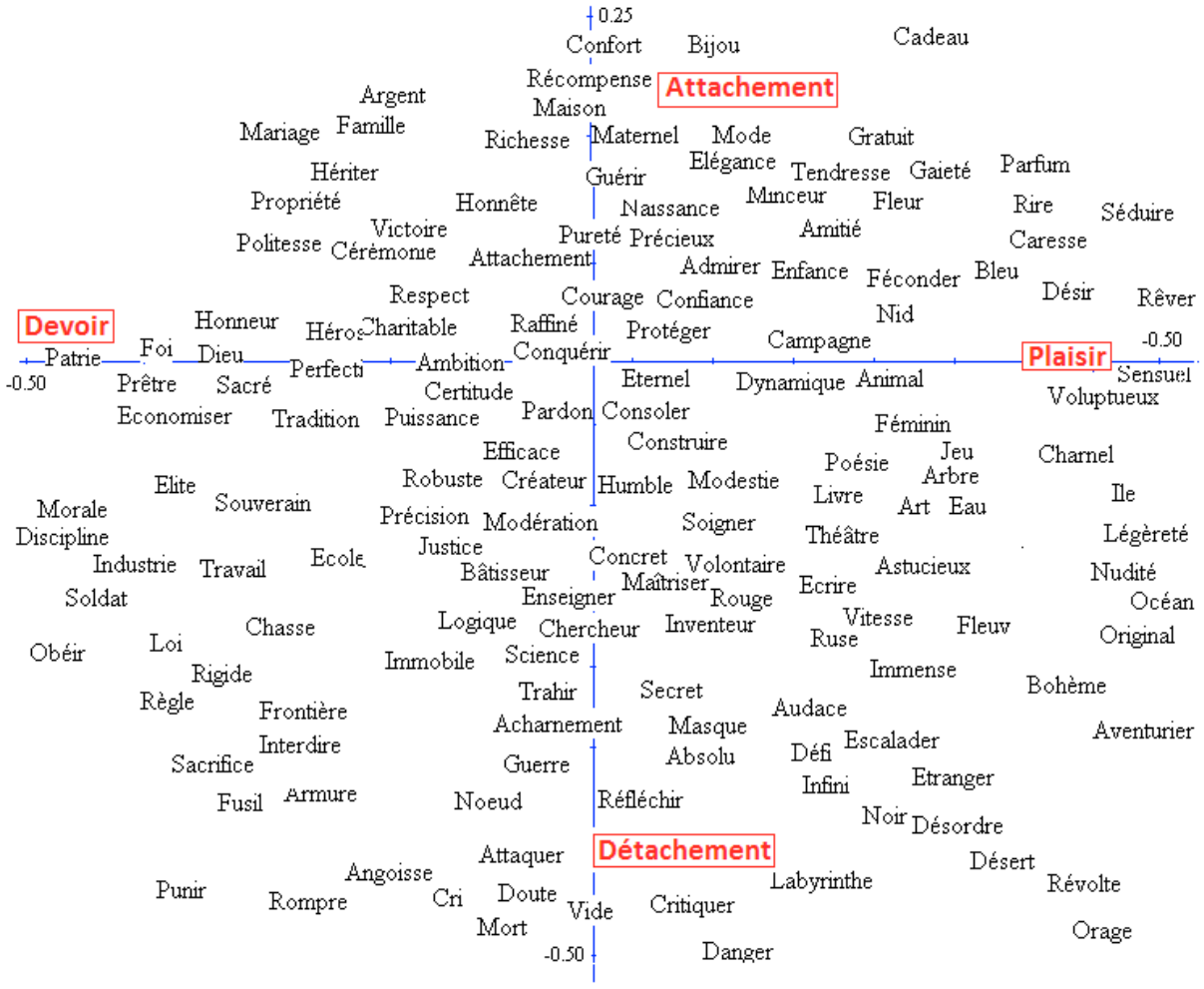

Figure 1. Esquisse du plan sémiométrique principal (axes 2 x 3). Ce plan se retrouve dans tous les pays occidentaux, avec quelque fois des interversions entre les axes 2 et 3. Les noms apposés sur les axes (Devoir – plaisir pour l'axe 2, Détachement – attachement pour l'axe 3) sont conventionnels.

La stabilité des axes principaux a été éprouvée empiriquement sur différents types de populations : les échantillons France entière, mais aussi les sous-populations obtenues en décomposant les échantillons par catégories socio-démographiques (sexe, âge, activité, profession), par pays (expériences menées dans une dizaine de pays occidentaux) et dans le temps (enquêtes menées par la SOFRES sur plusieurs années consécutives en France).

Cette stabilité a conduit à s'interroger sur la portée universelle des structures obtenues, sans ignorer les hypothèses fortes posées, notamment le choix des mots et leur contexte culturel.

Au plan statistique, la stabilité des résultats a été vérifiée par plusieurs méthodes de rééchantillonnage (différentes techniques de *bootstrap*). On a appliqué le *bootstrap* de façon classique sur l'ensemble des individus pour étudier la stabilité de la structure vis-à-vis des fluctuations d'échantillonnage, puis, de façon originale, sur l'ensemble des variables pour s'assurer que les structures ne dépendent pas strictement des mots sélectionnés.

2.1. Le Bootstrap sur variables

On peut supposer que les 210 mots sont choisis par un tirage de type multinomial dans une urne contenant l'ensemble des mots "sémiométrisables" de la langue française (il s'agit de mots pleins -pas de mots outils ou grammaticaux- faisant partie du vocabulaire de base, non consensuels, non ambigu sémantiquement -non polysémiques-, chargés émotionnellement ou axiologiquement - pas de mots neutres du type *table*, ni de mots techniques comme *ordinateur*). Naturellement, un tirage dans cet univers suppose à chaque fois la constitution d'un nouveau questionnaire et un nouveau travail de terrain (dont le coût serait prohibitif). Comme dans le cas des échantillons d'individus, le *bootstrap sur variables* fournit une solution pragmatique et économique à ce problème : tirage avec remise dans l'ensemble fini des mots déjà sélectionnés. Il ressort de cette étude que les premiers axes sont assez indépendants de la composition du questionnaire.

2.2. Le cas de la Chine (Hong-Kong).

Après plus de quinze années d'enquêtes dans plusieurs pays occidentaux (France, Canada, Etats-Unis, Grande Bretagne, Italie, Espagne, Grèce, Allemagne, Finlande, République tchèque) au cours desquelles des espaces sémiométriques similaires furent observés, la première enquête en pays asiatique a conduit à revoir l'hypothèse d'une universalité de la structure, hypothèse qui se trouvait pourtant renforcée à chaque nouvelle enquête. On parle depuis d'universalité pour un contexte culturel donné. Certes l'enquête sémiométrique menée à Hong-Kong ne porte que sur un échantillon modeste (795, à comparer par exemple à 9094 pour l'échantillon des USA). Cependant, la qualité du terrain ne semble pas devoir être mise en cause. De même, l'exercice très difficile de traduction du questionnaire à partir de la version anglaise a été réalisé avec le plus grand soin, après plusieurs itérations de traductions inversées. En bref, on observe dans ce contexte deux axes de notations dominants (axes opposant les notes élevées aux notes faibles, mais aussi les notes modérées aux notes extrêmes), le plan sémiométrique (2,3) se retrouvant au niveau (3,4) avec cependant de nombreuses différences que des spécialistes des cultures orientales pourront peut-être expliquer. Le domaine de recherche reste actuellement ouvert... en l'absence de nouvelles (et coûteuses) expérimentations.

3. Corrélations sémiométriques et liens sémantiques

La question suivante s'est posée d'emblée : *les proximités sémantiques entre mots de la liste ne sont-elles pas responsables de l'essentiel de la structure observée* ? Autrement dit, de façon un peu schématique, *la structure que l'on observe est-elle une structure linguistique plutôt que psychologique ou psycho-sociologique* ?

Deux mots ayant des sens voisins seraient notés de façon similaire et donc corrélés, et le *pattern* observé (sur la figure 1, par exemple, pour les axes 2 et 3) ne serait autre que le reflet de ces liens sémantiques. La stabilité de la structure en découlerait, puisque la langue est

relativement stable dans le temps. Elle est aussi pratiquement la même pour les différentes classes d'âge, le sexe, etc.. Un tel réseau sémantique doit aussi résister - plus ou moins - à une traduction du questionnaire, d'où la relative stabilité observée d'un pays à un autre. On va vérifier que la structure sémiométrique ne se réduit que très partiellement à une structure sémantique. Au niveau local, autour d'un point représentant un mot, on pourra trouver des voisins sémantiques, mais les grandes oppositions responsables d'axes stables ne sont pas observées spontanément à partir d'analyses de simples proximités sémantiques.

3.1. Voisinage sémantique du questionnaire sémiométrique

Dans cette première expérience, chaque mot de la sémiométrie est décrit par ses « synonymes », collationnés à partir d'une source externe. [Dans la version française de l'ouvrage paru en 2003, c'est le "Dictionnaire de synonymes et contraires" de Henri Bertrand du Chazaud (Robert, 1994) qui a été utilisé. Pour la version anglaise (2014), on a travaillé à partir du thésaurus anglais de l'Institut de Sciences Cognitives du CNRS].

Le nombre de voisins sémantiques est très variable. Ce nombre peut être très grand et peut atteindre 22 lignes de texte pour le mot *Changement* ou très court (voire inexistant, ainsi, l'adjectif *Maternel* n'a pas de voisin sémantique dans ce dictionnaire particulier).

A partir de ce nouveau recueil de textes est construit un *tableau croisé* contenant en lignes les 210 mots de base, et en colonnes tous les mots rencontrés dans le recueil, c'est-à-dire tous les voisins sémantiques des 210 mots. Le tableau qui sera soumis à une analyse des correspondances contient, à l'intersection de la ligne i et de la colonne j, la valeur « 1 » si le mot j figure parmi les voisins sémantiques du mot i, et contient la valeur « 0 » sinon.

Le premier résultat obtenu, décevant, mais prévisible, est une séparation entre les noms et les verbes… qui ne peuvent pas être synonymes, et assez rarement voisins sémantiques dans les dictionnaires usuels, alors que noms et verbes figurent parmi les mots de la sémiométrie. On procédera donc à l'analyse des seuls noms et adjectifs, qui, au nombre de 177, sont très majoritaires dans la liste originale des 210 mots du questionnaire.

On n'observe plus des nuages de points réguliers et équilibrés comme celui de la figure 1, mais des grappes de mots qui s'opposent à tous les autres. Ce phénomène semble dû à la *non-transitivité* des similitudes sémantiques, et à l'absence de pertinence de la notion d'*éloignement sémantique*. La non-transitivité peut s'exprimer de la façon suivante : si le mot A a un sens voisin de celui du mot B, et si le mot B a un sens voisin de celui du mot C, alors, A n'a pas forcément un sens voisin de celui du mot C. On peut même assez rapidement aboutir à un antonyme de A. Ainsi, *abandonner* est un voisin sémantique de *donner*, voisin de *partager*, qui est lui-même un voisin de *participer*, à son tour voisin de *se joindre*.

Le graphe sémantique est loin d'être un graphe régulier, ses sommets ont des degrés très variables, certains sommets sont même isolés.

L'observation des liens sémantiques issus d'un dictionnaire avec ceux que l'on peut dériver des notes sémiométriques aide à comprendre la nature du fait statistique et de la structure que l'on observe.

Tout d'abord, il existe, parmi les 210 mots de la sémiométrie, des mots qui n'ont pas de voisins sémantiques communs avec les autres (en fait avec plus de deux autres) comme par exemple les mots : *Lune, Peau, Arbre, Fusil, Ile, Théâtre, Soldat*. Les corrélations observées entre ces mots et les autres ne sont donc pas des artefacts sémantiques, elles relèvent d'un contexte perceptif plus général que nous allons précises dans les sous-sections 3.2 et 3.3.

Considérant maintenant le thésaurus anglais, il existe aussi des couples de mots, comme *Irony* et *Humour*, qui ont un grand nombre de voisins sémantiques communs (*mind, caustic, malice, wit, witty, funny, mischievous, malicious, sharp, satirical, mocking, projection, satire*) et qui sont donc très proches sémantiquement alors que l'espace sémiométrique les sépare de façon significative. Sur la figure 1, *Irony* (Ironie) se situe près de l'étiquette DETACHEMENT et *Humour* au voisinage de l'étiquette PLAISIR.

Humour est doux (mots corrélés : *Softness, Tenderness*); *Irony* est abrasif (*Savage, Disorder*). *Humour* crée des liens (*Friendship, Gift, Trust, Caress*), *Irony* romps les liens (*Critique, Revolt, Attack, Detachment, Change, Ruse*). *Humour* est un jeu (*Laughter, Game, Seduce*), *Irony* un drame (*Danger, Mystery, Adventure, Challenge, Black*).

La matrice des correlations des 210 variables (mots) contient 21,945 coefficients. On peut donc répéter l'analyse précédente (limitée ici à la paire : *Irony* - *Humour*) pour toutes les paires de mots. Le halo sémantique autour de chaque mot est en fait multidimensionnel. A la notion binaire de synonymie est substituée un espace beaucoup plus riche incluant des aspects psychologiques, contextuels, sociologiques, pratique, émotionel.

En conclusion de ce paragraphe, on retiendra que :
- la structure sémantique de la liste sémiométrique telle qu'elle est décrite (grossièrement, il est vrai) par un dictionnaire de synonymes ne donne pas lieu à des axes (ou directions principales) stables.
- localement, on retrouve des associations sémantiques sur les cartes sémiométriques, mais il existe aussi des exceptions notables.
- Les grandes oppositions stables observées dans le champ sémiométrique ont un caractère psycho-sociologique, voire socio-démographique marqué mais ne relèvent pas ou peu du registre synonymie-antonymie.

3.2. Les « champs sémantiques internes » : dénotation et connotation

On appellera champ sémantique interne pour un mot donné l'ensemble des mots (pris dans une liste de mots fixée *a priori*, d'où le qualificatif d'*interne*) qui lui sont corrélés. La distance utilisée est d'autant plus petite que le coefficient de corrélation entre les deux mots est élevé. Cette dénomination est justifiée par l'interprétation *a posteriori* des proximités observées. Il n'était pas évident que des notes fondées seulement sur l'agrément ou le désagrément engendrent des proximités sémantiques. Cette section montre au contraire la cohérence et la finesse des proximités observées.

Certains mots possèdent un champ sémantique interne riche et dense défini par des corrélations élevées avec de nombreux autres mots de la liste. C'est le cas pour les mots *Efficace* et *Courage*. Des mots comme *Montagne, Voluptueux, Mystère,* proches d'un petit nombre de mots, ont chacun un champ plus restreint.

La sémiotique distingue parfois deux types de sens pour les mots, le sens dénotatif (neutre et objectif, donné par les dictionnaires de la langue) et le sens connotatif (affectif et subjectif, contenu dans les évocations du mot). Exemple : le mot *mer dénote* une vaste étendue d'eau salée tandis qu'il *connote* l'immensité, la liberté, l'aventure, la profondeur, la tempête, le naufrage, etc. Parce qu'elle mesure la charge émotionnelle contenue dans les mots, la sémiométrie repose sur le sens connotatif des mots et non sur leur sens dénotatif. Ceci explique que l'espace reconstruit à partir d'un dictionnaire des synonymes (constitués à partir du sens dénotatif des mots) n'a que peu de similitude avec celui de la sémiométrie.

Il faut aussi distinguer plusieurs types de connotations : d'une part les connotations positives, dont l'évocation produit des sensations agréables, et les connotations négatives, dont l'évocation produit des sensations désagréables. On peut encore distinguer les connotations collectives, qui proviennent du contexte culturel dans lequel un groupe d'individus donné baigne, et les connotations individuelles, qui sont fonction de l'expérience de chaque individu. Ainsi, pour reprendre l'exemple du mot "mer" : 1/ si la *liberté* est une notion en général positive, le *naufrage* en est une négative, 2/ si pour la plupart des hommes la mer connote la *profondeur*, pour un marin-pêcheur elle connotera, en plus, la *subsistance*, et pour un plaisancier, la *détente*. D'autre part, si le *blanc* connote, en général, la *pureté* pour les occidentaux, il connote le *deuil* pour les Chinois (cf. sous-section 2.2 ci-dessus). Les espaces de connotations sont plus multidimensionnels que les espaces de dénotations...

4. La sémiométrie spontanée

Pour se libérer de la liste fixe des mots, une expérience a consisté à incorporer les deux *questions ouvertes* suivantes dans une enquête par sondage auprès de la population générale :

« Sans qu'on puisse expliquer pourquoi, peut-être à cause de ce qu'ils évoquent, certains mots nous sont agréables, d'autres désagréables.
En ce qui vous concerne personnellement, quels sont les mots qui vous sont le plus agréable ? (Citez le plus de mots possible).
Quels sont maintenant les mots qui vous sont le plus désagréable ? (Citez le plus de mots possible) ».

4.1. Mise en place de l'expérience

L'enquête a été réalisée par la SOFRES en 1995 auprès de 1 191 répondants. Les réponses correspondantes ont produit un « texte artificiel » long de 41 547 *occurrences*, à partir de 7 170 mots distincts cités. Beaucoup de mots n'ont été cités qu'une fois, et donc ne jouent aucun rôle dans le calcul des distances entre répondants. Parmi les 7 170 mots distincts, il y en a 1 466, soit 20.4 %, qui sont cités quatre fois ou plus. Si l'on se restreint au texte artificiel formé par ces 1 466 mots, il reste 33 950 occurrences, ce qui constitue 82 % du corpus initial.

Répondant a:
printemps soleil santé joie bonheur enfant voyage vacances argent maison cérémonie restaurant promenade ami fleur jardin beauté compagnie bricoler tricot loisirs magasin fête visite découverte lecture mer montagne destinée
*hiver *froid *neige *verglas *maladie *pollution *secte *drogue *solitude *inactivité *laideur *mort *guerre *souffrance *pauvreté *misère *corruption *attente
Répondant b:
amour fleur joie gentillesse tendre gaie aimable jolie souriant nature forêt montagne caresse beauté aide chocolat cerise lit sommeil pain vin femme enfant vie voyage famille ami
*méchanceté *tuerie *drame *hypocrite *égoïste *viol

Tableau 1. Exemples de réponses libres

Le tableau 1 donne les réponses de deux répondants. Les mots précédés d'un astérisque sont les mots cités comme désagréables. En effet, certains mots, comme *feu*, *argent, pluie*, *ordre*, peuvent figurer (pour des individus différents) avec ou sans astérisques, c'est-à-dire être cités comme désagréables par les uns, et comme agréables par les autres. Le tableau 2 donne la liste des mots apparaissant plus de 136 fois. Les quatre mots les plus fréquemment cités sont absents du questionnaire sémiométrique. Il s'agit de mots «consensuels » éliminés, par

construction, de la liste des 210 mots. Puis apparaissent les mots **guerre, *mort, amitié, enfant, famille*, *fleur,* figurant tous les six *(Enfance* pour *enfant*) dans le questionnaire sémiométrique. Il y a évidemment beaucoup de redondances ou de relations d'implication dans les mots cités spontanément comme (*santé, *maladie, *cancer, *sida*) ou encore (*amitié, ami*), (**mort*, **décès*), *(*vol, *voleur*), *(*crime, *meurtre*), (*amour, aimer*).

mots	*freq*	*mots*	*freq*	*mots*	*freq*
amour	731	musique	277	*racisme	165
soleil	628	*chômage	265	*haine	165
vacances	498	*accident	263	paix	156
*maladie	475	joie	258	montagne	156
*guerre	439	nature	209	*cancer	156
*mort	421	*violence	191	rire	153
amitié	379	santé	189	douceur	151
enfant	366	tendresse	184	*froid	148
famille	358	beauté	180	maison	147
fleur	328	*méchanceté	173	sourire	145
bonheur	305	argent	171	fête	143
voyage	292	*mensonge	167	bébé	141
mer	286	liberté	167	travail	137

Tableau 2. Mots cités spontanément les plus fréquents

On constate aussi que les substantifs sont majoritaires. Les verbes et les adjectifs apparaissent d'abord sous des formes ambiguës ; *rire* (153) et *sourire* (145), sont probablement cités la plupart du temps comme des substantifs compte tenu de leur contexte – nous verrons que les catégories grammaticales apparaissent souvent par séquences dans une même réponse. On note également des thèmes liés à l'actualité immédiate *(*pédophilie, *impôts* (95)*, *drogue* (133)*, *attentat* (59)*, *secte* (77)), exclus *a priori* du questionnaire sémiométrique.

On note l'absence de citation spontanée de certains mots du questionnaire sémiométrique fermé. Pour cet échantillon de 1 191 répondants, 173 mots (sur les 210 que compte le questionnaire sémiométrique) sont cités spontanément comme agréables ou désagréables. Si l'on ne retient que les 600 premiers répondants, seulement la moitié des mots du questionnaire sémiométrique sont cités spontanément par les personnes interrogées (apparaissent plus de 200 fois : *fleur, *guerre, *mort, amitié*).

absolu	*conquérir*	*humble*	*modération*	*réfléchir*
acharnement	*défi détachement*	*interroger*	*muraille*	*règle*
armure astucieux	*élite*	*inventeur*	*noeud*	*rigide*
attachement	*enseigner*	*labyrinthe*	*or*	*robuste*
bâtisseur	*escalader*	*logique*	*produire*	*sacré*
cérémonie concret	*féconder*	*magie*	*question*	*utilitaire*
	fermeté	*masque*	*recueillement*	*viril*

Tableau 3. Les 37 mots du questionnaire sémiométrique non cités spontanément

Il y a dans les réponses une grande fréquence de mots consensuels, et aussi une extrême dispersion sur les mots relativement rares ou idiosyncratiques (exemple : *chèvrefeuille, clafoutis, candélabre*) tenant parfois au caractère ludique du questionnaire. La probabilité de trouver des mots assez neutres (comme les mots du questionnaire sémiométrique : *enseigner, interroger, question, utilitaire*) est alors très faible. Le nombre d'hapax dans ces réponses libres est considérable (4 266, pour 7 170 mots distincts).

4.2. Premières explorations des réponses

Les analyses qui suivent permettent de mieux comprendre la nature des réponses à ces questions ouvertes. La figure 2 présente une « carte auto-organisée de Kohonen » (Kohonen, 1989). Il s'agit d'une description graphique des associations entre mots (positifs, sans astérisque) apparaissant plus de 25 fois dans l'ensemble des réponses. Deux mots appartenant à une même case sont souvent cités simultanément par la même personne. Ceci reste encore vrai, mais dans une moindre mesure, s'ils sont situés dans des cases voisines. Si les cases sont très éloignées, les mots correspondants sont, au contraire, rarement cités simultanément.

égalité justice fraternité franchise	respect intelligence honnêteté bonté	politesse gentillesse entente affection	harmonie		joli beau		pardon merci bonjour
solidarité sincérité	générosité fidélité compréhensio	tendresse	sérénité douceur calme		bon		
tolérance humour courage	partage paix liberté confiance	nature espoir confort	vivre tranquillité plaisir	île sieste manger femme dormir			heureux aimer
jeunesse gaieté convivialité	joie beauté amitié	vie rire détente amour	repos chaleur	rêve lit	étoile oiseau eau couleur	vert caresse bleu	doux câlin
bien-être	réussite passion chance bonheur	fête cadeau	sourire loisir enfant chanson		sable lumière livre fleur ciel	parfum baiser	rose papa maman bébé
	propreté	naissance mariage jeux argent anniversaire	été vacances	voyage sport soleil repas rencontre printemps plage musique	mer campagne	chat ami	gâteau chocolat Noël
gentil aimable		travail santé cadeaux	week-end loisirs famille	voiture promenade maison champagne animaux	danse	neige chien	théâtre peinture
agréable		voyages petits-enfan enfants amis	fleurs	télévision restaurant	vélo randonnée montagne lecture cinéma	pluie oiseaux jardin cuisine bateau	vin verdure fruit forêt chant arbre

Figure 2. Carte de Kohonen représentant les associations entre mots dans les réponses libres

Des regroupements comme (*été, vacances*), (*papa, maman*), (*gâteau, chocolat*), (*pardon, merci, bonjour*) sont l'indice d'un remplissage un peu automatique du questionnaire.

Il est clair que le questionnaire fermé a le mérite d'éviter ces dérives et les surpondérations accidentelles de certains thèmes.

Cette tendance à l'association des mots énoncés consécutivement concerne aussi la catégorie grammaticale des mots : un adjectif (ou un nom, ou un verbe) aura tendance à être suivi par des adjectifs (ou des noms, ou des verbes).

En conclusion de cette première exploration des réponses aux deux questions ouvertes, on doit relever le caractère extrêmement *bruité* des données recueillies de cette façon, par comparaison à un recueil de notes attribuées à une liste de mots identiques pour toutes les personnes interrogées. La distance entre individus va dépendre en fait du petit nombre de mots qu'ils peuvent avoir en commun, et les individus n'ayant aucun mot en commun auront des distances indifférenciées.

4.3. Choix spontané et caractéristiques des répondants

Des regroupements *a priori* des répondants sont réalisés à partir de quelques unes de leurs caractéristiques disponibles. Ce sera l'occasion de voir que le sexe et l'âge, considérés isolément, ou mieux encore simultanément, ne sont pas indépendants des mots cités comme agréables ou désagréables.

Les mots caractéristiques des jeunes *(plaisir, manger, dormir,…)* et des personnes plus âgées (*politesse, courage, fraternité,...*) ne sont pas sans rappeler les deux extrémités de l'axe horizontal de la figure 1 (axe deux de la sémiométrie).

La figure 3 représente, sous forme de proximités graphiques, une synthèse des liens existant entre les six catégories (3 classes d'âge x 2 sexes : deux catégories proches ont des profils lexicaux communs) et entre les mots (deux mots proches ont des profils socio-démographiques similaires).

Mots caractéristiques	Valeurs-test	Probabilités
Homme moins de 30 ans		
1 dormir	3.24	.001
2 plaisir	3.03	.001
3 manger	2.84	.002
4 loisir	2.29	.011
Homme plus de 55 ans		
1 courage	3.59	.000
2 fraternité	2.80	.003
3 propreté	2.68	.004
4 santé	2.24	.012
Femme moins de 30 ans		
1 chocolat	3.06	.001
2 bébé	3.02	.001
3 animaux	2.46	.007
4 maman	2.26	.012
5 câlin	2.13	.016
6 été	2.12	.017
Femme plus de 50 ans		
1 merci	3.00	.001
2 affection	2.83	.002
3 politesse	2.53	.006
4 bonjour	2.14	.016

Tableau 4. Mots caractéristiques de quatre catégories sexe-âge

La représentation obtenue confirme la complémentarité, et même l'additivité des effets de ces deux variables de base. En effet, ces deux variables se déploient selon des directions orthogonales, l'âge, horizontalement, opposant les catégories les plus âgées à droite, aux plus jeunes à gauche; le sexe, selon une direction verticale, opposant les hommes, en bas, aux femmes, en haut. Rien, dans les six colonnes du tableau d'entrée soumis à l'analyse, n'indique qu'elles proviennent du croisement de deux variables. La seule information qui décide de la position des points-colonnes sur le graphique est le profil lexical de ces colonnes.

Finalement, cette figure décrit, dans le cadre d'un continuum nuancé, les oppositions entre sexes pour une classe d'âge donnée, et l'étendue des variations lexicales internes à chaque sexe, en fonction de l'âge (Notons que les axes 2 et 3 de la sémiométrie (figure 1) opposent respectivement, de la même manière, les jeunes (à droite sur la figure 1) aux personnes plus âgées et les hommes aux femmes).

La figure 3 montre aussi les ellipses de confiance *bootstrap* des points-catégories (petites ellipses) : il est clair que le *pattern* observé est stable, malgré la taille modérée de l'échantillon. Les ellipses de confiance relatives aux mots sont beaucoup plus grandes (les ellipses sélectionnées à titre d'exemple concernent les mots *pardon, livre, maman, manger*).

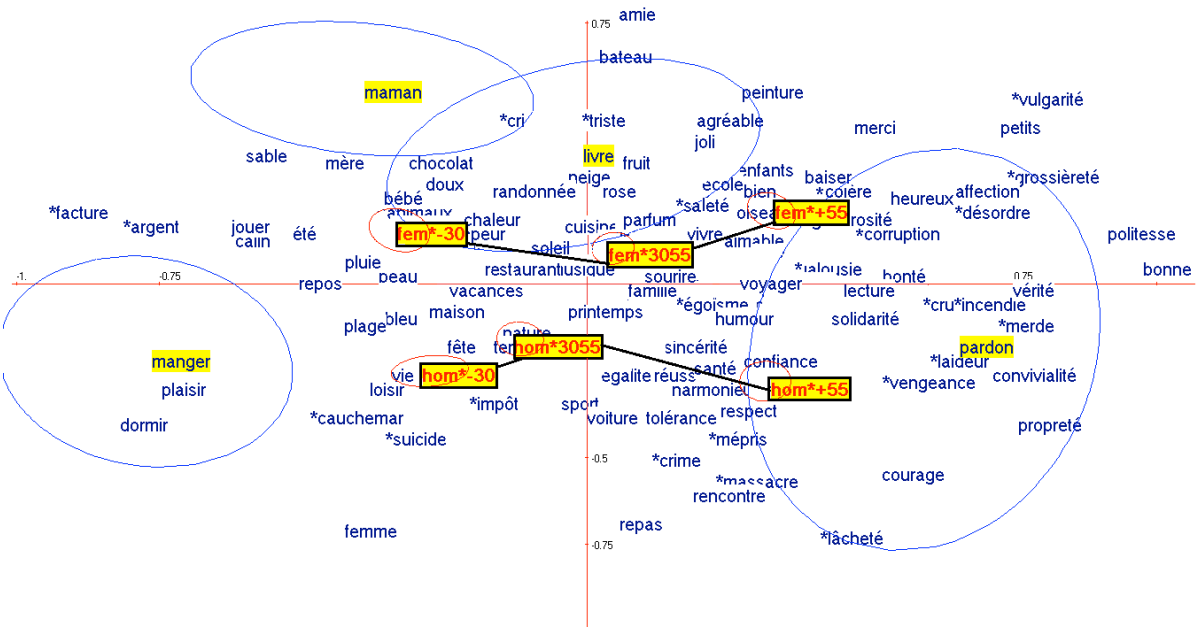

Figure 3. Plan principal de l'analyse des correspondances de la table (mots x catégories). Ellipses de confiance des catégories et de mots (les mots précédés d'un astérisque sont les mots cités spontanément comme étant les plus désagréables)

La grande taille de ces ellipses ne modifie pas l'interprétation des proximités. En bas à gauche du graphique, le mot *manger* reste caractéristique des hommes jeunes, quelle que soit sa place à l'intérieur de son ellipse de confiance, de même, à droite, le mot *pardon* reste caractéristique des personnes âgées ; en haut, les mots *livre* et *maman* caractéristiques des femmes.

4.4. Rapprochement entre sémiométrie et questions ouvertes

Nous avons pu disposer des notes sémiométriques pour 335 personnes parmi les 1 191 répondants aux deux questions ouvertes invitant à citer spontanément des mots agréables ou désagréables. Cela nous autorise à répondre à la question : quels sont les mots cités spontanément qui caractérisent les premiers axes sémiométriques. Les mots cités

spontanément vont être considérés comme des variables nominales supplémentaires (au même titre que le sexe ou l'âge) et seront donc projetés sur les axes principaux.

Alors que les coordonnées des mots de la sémiométrie sur les axes sont leurs coefficients de corrélation avec les axes, les coordonnées des mots cités spontanément projetés comme des catégories supplémentaires seront converties en valeurs-test (variables normales centrées réduites qui prennent en compte l'effectif concerné (fréquence du mot).
La position des mots spontanés sur le premier axe (axe méthodologique de *participation à l'enquête*) est intéressante. Cet axe oppose, en effet, les individus utilisant pleinement l'échelle proposée pour les notes à des individus qui n'utilisent que la partie centrale de l'échelle.
Les mots *de la liste sémiométrique* les plus caractéristiques des individus qui utilisent pleinement l'échelle des notes sont, pour l'analyse sémiométrique des 335 individus : *Courage, Politesse, Héros, Honneur, Protéger, Robuste, Tradition, Dynamique, Raffiné, Elégance, Honnête…*, avec des coefficients de corrélation avec l'axe variant de –0.59 à –0.49.
Les mots caractérisant ceux qui n'utilisent que la partie centrale de l'échelle sont : *Trahir, Angoisse, Révolte, Faute, Danger, Désordre, Mort…,* dont les coefficients de corrélation avec l'axe, nettement plus faibles, varient de 0.29 à 0.17.

Les mots *cités spontanément* caractérisant les individus qui utilisent pleinement l'échelle des notes sont, suivis de leurs valeurs-test entre parenthèses, *confiance* (–2.9), *aimer* (–2.8), *bonjour* (–2.7), *merci* (–2.4), *courtoisie* (–2.1), *honnête* (–2.1). On retrouve bien les notions de politesse et d'honnêteté. Le mot spontané le plus caractéristique de ces répondants est **contrainte* (considéré donc ici comme un mot particulièrement désagréable, avec une valeur-test de 3.7, qui est la plus élevée de l'ensemble des mots sur l'axe).

On apprend donc qu'il s'agit avant tout de personnes qui rejettent les contraintes (traduction du fait statistique : qui sont caractérisées de façon significatives par la citation spontanée du mot *contrainte* comme mot désagréable). Le rejet du mot *travail* (et probablement des contraintes qu'il représente) va dans le même sens.

4.5. Conclusion de la section 4

a) On ne peut retrouver l'espace sémiométrique par un questionnement ouvert du type proposé dans cette section, c'est à dire sans aucune contrainte. On obtient bien des associations locales, schématisées par la carte de Kohonen de la figure 2, des plans factoriels présentant une parenté assez marquée avec des plans de la sémiométrie (figure 3) mais pas toutes les dimensions latentes stables et interprétables que révèle le questionnaire fermé.

b) La citation spontanée induit une dispersion sans limite du vocabulaire, réduisant de façon corrélative la signification des distances entre individus. De plus, beaucoup de mots riches de sens et de valeurs ne sont, *a priori*, ni agréables, ni désagréables, et donc ont peu de chance d'apparaître dans les réponses spontanées. Notons aussi que les mots consensuels (*amour*, *vacances*, etc.) lestent le recueil sans apporter une information décisive.

c) A propos du rôle des notes par opposition à une simple mention de présence ou d'absence : La note permet de se référer à une note moyenne pour chaque mot, note moyenne qui n'est pas connue des répondants. On peut mal noter un mot, et pourtant le noter au dessus d'une note moyenne que l'on ignore au moment de l'interview. La note a donc une capacité métrologique supérieure au relevé de la simple présence ou absence d'un mot.

d) Incapable de produire seule des axes bipolaires, le corpus des réponses aux questions ouvertes permet cependant d'enrichir *a posteriori* l'interprétation des axes calculés à partir du questionnaire sémiométrique.

Conclusion générale

Dans les cas de réponses à des questions ouvertes, on sait qu'il existe des stratifications socio-linguistiques, influençant le vocabulaire et la syntaxe des réponses, qui induisent une structure de l'ensemble des répondants avant même la prise en compte du contenu de ces réponses. Une nouvelle médiation entre la question et le contenu des réponses est introduite par ce que l'on peut appeler l'aspect hédoniste du vocabulaire que mesure la sémiométrie, induisant de nouvelles structures significativement liées elles aussi à de nombreuses caractéristiques socio-démographiques. Pour les analyses de réponses à des questions ouvertes, la sémiométrie nous convainc un peu plus qu'il n'existe pas de contenu facilement séparable de la langue et des tropismes langagiers des personnes qui répondent.

[Toutes les procédures de calcul utilisées sont implémentées dans le logiciel Dtm-Vic qui peut être librement téléchargé, avec des exemples de données sémiométriques, à partir du site www.dtmvic.com].

Références

Benzécri J.P. (1989). Essai d'analyse des notes attribuées par un ensemble de sujets aux mots d'une liste, *Les Cahiers d'Analyse des Données*, 1, 73-98.

Bréchon P. (2000). *Les valeurs des Français. Evolutions de 1980 à 2000*, Armand Colin, Paris.

Efron B. (1979). Bootstraps methods : another look at the Jackknife, *Ann. Statist*., 7, 1-26.

Efron B. (1982). The Jackknife, the Bootstrap and other resampling plans, *SIAM*, 1982, 116-130.

Fabre J., Morlat G., Pagès J-P et Stemmelen E. (1981). Les Structures de l'Opinion Publique, *Le Progrès Technique*, n° 22-24, ANRT, Paris.

Kohonen T. (1989). *Self-Organization and Associative Memory*, Springer-Verlag, Berlin.

Lebart L., Salem A. et Berry E. (1998). *Exploring Textual Data*, Kluwer Ac. Publisher, Dordrecht.

Lebart L., Piron M. et Steiner J-F. (2003). *La Sémiométrie; Essai de statistique structurale.* Dunod, Paris.

Lebart L., Steiner J-F., Wisdom J. et Piron M. (2014). *The Semiometric Challenge: Words, Lifestyle and Values.* L2C, Rivesaltes [*in press*].

Osgood C. E. (1965). Cross-Cultural Comparability in Attitude Measurement via Multilingual Semantic Differentials, in: Steiner I., Fishbein M. (Eds*): Current Studies in Social Psychology*, Holt, Rinehart and Winston, New York.

Valette-Florence P. (1994*). Les styles de vie. Bilan critique et perspectives*, Nathan, Paris.

Steiner J.F.et Auliard O. (1992). La sémiométrie, In : *La qualité de l'information dans les enquêtes*, Lebart L. (Ed.), Dunod, Paris, 241-274.