UNIVERSITÉ DU QUÉBEC À MONTRÉAL

LES MÉTHODES DE FORAGE DE TEXTE POUR L'ANALYSE STRUCTURALE DES REPRÉSENTATIONS SOCIALES

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN SOCIOLOGIE

PAR

JEAN-FRANÇOIS CHARTIER

FÉVRIER 2010

REMERCIEMENTS

Je veux en premier lieu remercier chaleureusement mon directeur Jean-Guy Meunier du département de philosophie de l'UQAM pour sa générosité sans bornes, pour m'avoir initié aux sciences cognitives, aux méthodes d'analyse de textes assistées par ordinateur et au métier de scientifique en général. Grâce à lui, j'ai bénéficié d'un environnement d'étude et de recherche, et surtout d'un climat social et humain, de très grande qualité. Je veux également remercier mon codirecteur Louis Jacob du département de sociologie de l'UQAM, pour avoir accepté de diriger un mémoire de sociologie atypique, de m'avoir fait confiance et de m'avoir laissé toute la latitude que cela nécessitait.

Cette recherche est le résultat de plusieurs collaborations sans lesquelles elle n'aurait probablement jamais pu être menée à bien. D'abord d'étroites collaborations avec mes collègues du laboratoire le LANCI de l'UQAM, professeurs et étudiants avec qui j'ai eu d'innombrables échanges, et en particulier Nicolas Payette, qui a fait l'implémentation informatique des méthodes développées dans ce mémoire. J'ai également bénéficié de collaborations déterminantes à l'Institut des sciences cognitives de l'UQAM via le groupe de recherche interdisciplinaire sur la catégorisation. Ce groupe de recherche m'a permis d'apprécier toute la complexité des phénomènes cognitifs autant du point de vue des sciences sociales que de la linguistique, de la psychologie, de l'informatique et de la philosophie. Et finalement, j'ai eu l'occasion d'entamer des collaborations avec Johanne Saint-Charles et Pierre Mongeau du département de communication, qui m'ont permit de m'initié à l'étude des réseaux sociaux et avec qui nous avons mis sur pied un groupe de recherche sur les réseaux sociaux et sémantiques.

Par ailleurs, j'aimerais aussi remercier ma famille et mes amis pour m'avoir soutenu moralement dans les moments difficiles, par des encouragements, des « mises en perspectives » ou en me changeant simplement les idées lorsque j'étais trop submergé par des préoccupations scientifiques.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES ANNEXES

RÉSUMÉ

Dans cette recherche, nous explorons la pertinence des méthodes de forage de texte pour l'étude des représentations sociales diffusées dans un corpus de presse. Nous nous questionnons à savoir si les méthodes de forage de textes permettent d'atteindre les trois niveaux d'analyse d'une méthode générale d'étude des représentations sociales. Nous émettons l'hypothèse que c'est effectivement le cas, mais seulement dans la mesure où des choix d'opérationnalisation spécifiques sont faits pour certaines étapes et opérations impliquées dans une chaîne de forage de texte.

Pour évaluer cette hypothèse, nous articulons notre démarche autour de quatre (4) objectifs. Premièrement, nous présentons le cadre théorique des représentations sociales en insistant sur quatre concepts principaux: ceux de consonance sociocognitive, d'ancrage, d'objectivation et d'architecture. Deuxièmement, nous présentons le cadre méthodologique du forage de texte sous la forme d'une chaîne de traitement composée de six étapes: préparation du corpus, segmentation, indexation, vectorisation, réduction dimensionnelle de la matrice et classification automatique. Troisièmement, nous menons une analyse critique des apports du forage de texte jusqu'à maintenant dans le domaine d'étude des représentations sociales, en soulignant notamment plusieurs limites dans les choix d'opérationnalisation du logiciel principalement utilisé par les psychosociologues. Finalement, nous menons une expérimentation sur différents choix d'opérationnalisation de forage de texte, mieux adaptés à la théorie des représentations sociales. Nous développons deux techniques d'analyse des structures des représentations sociales, l'une basée sur une modélisation en termes de réseau cognitif de proximité sémantique, l'autre basée sur une modélisation en termes de réseau sociocognitif. Les résultats de l'expérimentation sont illustrés à partir d'une étude de cas: la représentation sociale des accommodements raisonnables dans trois journaux québécois.

MOTS-CLÉS : Représentation sociale, Analyse de textes assistée par ordinateur, Analyse de réseaux, Méthodologie, Accommodements raisonnables.

INTRODUCTION

L'ORIGINE DURKHEIMIENNE DES APPROCHES SOCIOLOGIQUES DE LA COGNITION

Comprendre et expliquer la pensée a été l'objet de plusieurs programmes de recherches hétéroclites et situés dans des traditions intellectuelles parfois très différentes les unes des autres. On pense notamment à certains programmes en philosophie, en psychologie, en linguistique, en biologie, en anthropologie, en intelligence artificielle ou dans les neurosciences. Ces programmes forment aujourd'hui ce qu'on appelle les sciences cognitives (Andler 2004). La sociologie a contribué à la compréhension et à l'explication de la cognition. Cette discipline a proposé plusieurs modèles, hypothèses et concepts, qui forment aujourd'hui ce que certains appellent la « sociologie cognitive ».[1]

Parmi ces modèles sociologiques, il y a une thèse générale commune à plusieurs d'entre eux, dont nous pouvons en retracer l'origine dans l'œuvre fondatrice du sociologue français Émile Durkheim. Cette thèse s'articule autour d'un concept pôle, celui de représentation collective.

Émile Durkheim a développé une thèse sociologique de la cognition (Durkheim 1903, 1912, 1914, 1955), qui a marqué par la suite de nombreux chercheurs des sciences sociales intéressés au problème. Ceux-ci l'ont reprise à leur compte, parfois en la critiquant, parfois en la modifiant et l'améliorant au fil de recherches tant théoriques qu'empiriques.

---

[1] L'expression « sociologie cognitive » est relativement récente et a été introduite par Aaron Cicourel (1979). La sociologie cognitive a autrefois porté une autre étiquette maintenant un peu moins à la mode de « sociologie de la connaissance » (Zerubavel 1997).

1.1 Récupération de la thèse durkheimienne de la cognition

La thèse cognitive de Durkheim, et son concept central de représentation collective ont été récupérés et réinterprétés de multiples manières tout au long du 20e siècle. De nombreux programmes de recherche en ont proposé chacun leur version, à commencer par les contemporains de Durkheim lui-même, tel Mauss (1923-24/2002) ou Halbwachs (1925/1994), puis du côté étasunien avec le structuro-fonctionnalisme de Merton (1947), et la tradition allemande autour de Scheler (1993) ou Stark (1958), la sociologie de la science avec par exemple Fleck (1935/2005) et Bloor (1976), la sociologie phénoménologique de Berger et Luckmann (2003), enfin l'anthropologie de Douglas (2004).

Plus récemment, on retrouve les mêmes intuitions de la thèse durkheimienne dans la psychosociologie initiée par S. Moscovici et son concept de représentation sociale (Moscovici 1976), dans l'anthropologie cognitive étasunienne de D' Andrade et son concept de modèle culturel (D'Andrade 1995), l'anthropologie cognitive de Hutchins et son concept de cognition socialement distribuée (Hutchins 1995), la sociologie cognitive de E. Zerubavel (Zerubavel 1997), ou encore le modèle de la contagion de Sperber (Sperber 1996) parmi bien d'autres.

1.2 Le cognitif et le social

Les thèses sociologiques de la cognition sont aujourd'hui très nombreuses. Au-delà des particularités de chacune, la sociologie cognitive s'articule essentiellement autour de l'intuition fondamentale de Durkheim, soit de problématiser la cognition en relation avec le social.

Un bref survol des traditions françaises, allemandes ou étasuniennes nous montre que ces deux concepts — le cognitif et le social — ont été exprimés de plusieurs manières dans la littérature. Ces expressions ont cependant en commun de rendre manifeste leur interdépendance. Pour Durkheim le cognitif était conceptualisé à l'aide de la notion de représentation collective et le social avec celle de solidarité. Pour Halbwachs (1925/1994), le cognitif était conceptualisé par la notion de mémoire collective et le social par le concept de morphologie sociale, pour Fleck (1935/2005) c'était par le style de pensé et le collectif de

pensée, pour Mannheim (1985) l'idéologie et les classes sociales, pour Gurvitch (1966) les formes de connaissance et de cadre sociaux, pour Goffman (1991) les cadres de l'expérience et les rituels d'interaction, enfin pour Bourdieu (1979) le cognitif était problématisé via les concepts d'habitus et de schème et le social via ceux de champs et de position.

Récemment, de nouveaux programmes de recherche ont repris cette thèse durkheimienne classique et redéploient ses concepts — le cognitif et le social — sous des notions comme celle de représentation sociale et de position sociale (Doise *et al.* 1992), de schéma cognitif et de culture (D'Andrade 1995; Shore 1996), de cognition distribuée et de rituel computationnel (Hutchins 1995), de représentation publique et de chaîne causale cognitive culturelle (Sperber 2000), de cognition située et de situation (Quéré 1999; Lave 1988), de mème et de mèmeplexe (Blackmore 2000), de croyance et de marché cognitif (Bronner 2003), d'information et de réseaux sociaux (Monge et Contractor 2003).

1.3 Le concept de représentation sociale

Parmi ces différents concepts, celui de représentation sociale est aujourd'hui l'un des descendants et représentants les plus visibles et actifs issus de la thèse durkheimienne. Le concept de représentation sociale (RS) est aujourd'hui un modèle explicatif important, reconnu et utilisé par différentes communautés et programmes de recherche, qui abordent la cognition du point de vue original de la sociologie et des sciences sociales (Jodelet 1989).

Ce sont en partie les travaux de Durkheim (couplés à ceux de Piaget) qui ont été la matrice originelle à partir de laquelle la théorie des RS a été construite (Moscovici 1989). L'intuition fondamentale qu'avait eue Durkheim a été récupérée et précisée. Les raffinements sont à la fois dans la modélisation théorique et dans le développement méthodologique.

1.4 La représentation sociale: modélisation du social et du cognitif

Des psychosociologues inspirés des travaux de Moscovici ont notamment insisté sur la dimension structurale de la cognition (Guimelli 1994) et sur les dimensions interactionnelle et communicationnelle des processus de genèse et de construction des RS :

> Ce ne sont pas les substrats, mais les interactions qui comptent. D'où la remarque parfaitement exacte que « ce qui permet de qualifier de sociales les représentations, ce sont moins leurs supports individuels ou groupaux que le fait qu'elles soient élaborées au cours de processus d'échange et d'interactions (Mocovici 1989 : 99)

La modélisation des RS n'a cessé de se raffiner. Bien qu'il en existe plusieurs interprétations possibles, une manière générale de la définir est qu'elle est devenue un modèle pour décrire statistiquement un ensemble de cognitions à propos d'un enjeu social important pour un groupe. Ces cognitions sont organisées selon des structures spécifiques qui sont à la fois intra et inter psychique.

Premièrement, chez un individu, certaines cognitions sont organisées selon une structure cognitive intra psychique spécifique, caractérisée par une hiérarchie. Deuxièmement, ces cognitions sont organisées de manière spécifique entre individus. Il s'agit dans ce cas de structures écologiques ou sociales qui traduisent les réseaux d'interactions dans un groupe. Ce genre de structure peut être vu comme une distribution sociale des cognitions parmi les membres d'un groupe.

D'autre part, ces structures que cherche à modéliser le concept de RS, sont elles-mêmes le résultat de différents processus sociocognitifs de construction collective des connaissances de sens commun. Ces processus sont ceux de la pensée sociale et sont déployés par les individus dans les interactions et les communications sociales. Le domaine d'étude des RS porte à la fois sur l'étude des processus de la pensée sociale et sur les structures qui en émergent.

1.5 La pensée sociale: logiques sociales et logiques cognitives

Pour des chercheurs contemporains à l'interface des sciences sociales et des sciences cognitives, la pensée sociale est comprise comme une modalité particulière de la cognition. La pensée sociale renvoie à différents processus sociocognitifs, qui obéissent à des logiques qui leur sont propres, liées en priorité à la cohérence relationnelle entre les individus. Guimelli résume cette particularité de la pensée sociale en la distinguant clairement de la pensée rationnelle, car selon lui, les sciences sociales et cognitives ont souvent réduit la première à une version « biaisée » de la seconde :

> Le mode de raisonnement mis en œuvre [dans la pensée sociale] n'est pas «biaisé», c'est-à-dire en fait, déficient. Il s'agit plutôt d'une stratégie sociocognitive, qui semble mieux adaptée au problème traité. Ce qui importe ici pour le sujet, ce n'est pas la validité de ses inférences, mais leur légitimité sociale; c'est qu'en définitive, elles puissent être rattachées à une expérience relationnelle positive. (Guimelli 1999 : 22)

L'important, nous dit Guimelli, est que, dans certaines conditions sociales, ce n'est pas tant la validité épistémique qui guide la logique des cognitions que la validité sociale, c'est-à-dire ce qui permet de maintenir les liens sociaux, l'interaction et la communication.

Dans les études sur la pensée sociale et les RS, la cognition est toujours étudiée dans sa relation de dépendance envers les conditions sociales dans lesquelles elle est actualisée :

> [...] les conditions sociales dans lesquelles sont construites les représentations sont susceptibles de générer des règles spécifiques conduisant ainsi à une logique interne « sociocognitive », qui permet l'organisation générale des cognitions. (Guimelli 1999: 80)

Les logiques de la pensée sociale sont à la fois cognitives et sociales. La modélisation de l'organisation des cognitions de la pensée sociale via le concept de RS veut rendre compte de cette logique double:

> Les représentations sociales ont donc cette caractéristique spécifique, qui rend d'ailleurs leur analyse difficile, qu'elles sont soumises à une double logique: la logique cognitive et la logique sociale. Elles peuvent être définies comme des constructions sociocognitives, régies par des règles propres. (Abric 1994b: 14)

Les travaux sur les RS ont contribué à apporter quelques éléments de réponse à ce vaste champ d'études de la pensée sociale. Depuis plus de trente ans, des énergies considérables ont été déployées pour concevoir notamment des méthodes d'analyse des RS. C'est dans cet horizon qu'émerge notre problématique de recherche.

CHAPITRE I

PROBLÉMATIQUE DE RECHERCHE

Introduction

Les acquis théoriques dans le domaine d'étude des RS sont considérables, mais c'est également du côté du développement des méthodes et techniques d'analyse (Moliner *et al.* 2003; Abric 2003) que l'on remarque les meilleures avancées dans l'élaboration d'un programme de recherche fécond et productif, qui a plus de trente ans de pratique, autant en psychologie sociale, en sociologie, en histoire qu'en anthropologie. Au-delà des considérations théoriques, le problème qui nous intéresse plus spécifiquement est de nature méthodologique.

Dans ce chapitre, nous présentons le cadre de notre recherche. Nous commençons par introduire son contexte, qui porte sur les trois niveaux d'analyse exigés par une méthodologie générale d'étude des RS. Dans un deuxième temps, nous présentons notre problématique, qui concerne plus spécifiquement les méthodes d'étude des RS à partir de corpus de presse. Troisièmement, nous formulons notre question de recherche à propos de nouvelles méthodes de forage de texte (*Text Mining*) pour l'étude des RS. Finalement, nous formulons une hypothèse de recherche et proposons quatre (4) objectifs de recherche qui nous permettront de la corroborer ou de la réfuter.

1. Contexte général de la recherche: les trois niveaux d'analyse des représentations sociales

Selon Abric (1994b: 60; 2003c: 376), les méthodes et techniques d'analyse des RS peuvent être classées en trois catégories selon le niveau d'analyse qu'elles permettent. Ces niveaux sont: l'identification du contenu de la RS, l'identification des structures de la RS, et l'identification du noyau central de la RS. Chacun de ces niveaux requiert des méthodes et

techniques spécifiques. Cependant, pour Abric, une méthodologie générale d'étude des RS doit permettre d'atteindre ces trois niveaux d'analyse.

Le premier niveau est l'identification du contenu, c'est-à-dire des éléments formant une RS donnée. Ces éléments sont considérés comme des cognitions. Dans la littérature, une terminologie très diversifiée est utilisée pour les désigner. On utilise, souvent avec peu de distinction, des notions comme celles de schème, de grille de lecture, de schéma, d'attitude, de thème, de croyance, de catégorie, de cognème, de script, de cadre, de stéréotype, de concept, etc. Cependant, on peut, à la suite à Flament et Rouquette (2003), conserver une terminologie générale et parler modestement d'élément cognitif ou de cognition en général, sans en préciser la nature. Par conséquent, nous pouvons définir de manière minimale et générale le contenu d'une RS comme un ensemble d'éléments cognitifs relatifs à un objet social important pour les membres d'un groupe social.

Pour le premier niveau d'analyse d'une méthodologie d'étude des RS il faut donc des méthodes et techniques qui permettent de recueillir ces éléments cognitifs composant une RS particulière.

Le deuxième niveau est l'identification des relations entre éléments cognitifs de la RS. Il s'agit d'identifier les structures de la RS étudiée. Nous vu que la théorie des RS modélise deux types de structure: des structures dites cognitives et des structures sociales ou parfois appelées écologiques.

La structure cognitive est le système de relations entre éléments d'une RS. Ces relations sont modélisées essentiellement de deux manières: par des relations de similitude ou par des relations de schème cognitif de base. Dans les deux cas, l'objectif est de révéler l'organisation cognitive des éléments cognitifs d'une RS chez les membres d'un groupe. Plus spécifiquement, on cherche à modéliser statistiquement comment, en moyenne, les membres d'un groupe organisent leur pensée à propos d'un objet social important pour eux et elles.

Les structures sociales sont les formes que prend la distribution sociale des éléments de la RS parmi les membres d'un groupe. Cette structure est quant à elle modélisée soit en termes de consensus social, et on cherche alors les éléments socialement partagés dans le groupe, ou en

termes de prise de position, et on cherche alors les principes organisateurs de la différenciation des membres d'un groupe.

Pour le deuxième niveau d'analyse d'une méthodologie d'étude des RS il faut donc des méthodes et techniques qui permettent d'identifier les structures d'une RS.

Le troisième niveau d'analyse des RS est l'identification de son noyau central. Le noyau central est composé d'un ou de quelques éléments cognitifs centraux dans les structures de la RS. Les propriétés de centralité des éléments sont nombreuses et dépendent du type de structure retenue pour l'analyse. Certaines propriétés de centralité sont liées à la structure interne de la RS: les éléments centraux ont une forte connexité, une forte saillance, une forte valeur symbolique (Moliner 1994). D'autres sont liés à la structure sociale de la RS: les éléments centraux font l'objet d'un fort consensus social.

Pour le troisième niveau d'analyse des RS il faut donc des méthodes et techniques qui permettent d'identifier la centralité des éléments de la RS étudiée.

1.1 Du contenu au noyau de la représentation sociale

Ce ne sont pas toutes les méthodes et techniques d'analyse des RS qui traitent ces trois niveaux. Certaines ont pour objectif l'analyse du contenu, d'autres celui de certaines structures, d'autre encore celui de l'analyse du noyau. Les deux derniers objectifs — l'analyse de la structure et du noyau — distinguent les approches structurales des approches orientées uniquement sur l'analyse de contenu. Le troisième objectif, quant à lui, distingue les approches structurales de l'école d'Aix-en-Provence des approches structurales de l'école de Genève, ces dernières orientées sur l'identification des « principes organisateurs » de prise de position de la RS (Doise *et al.* 1992).

Selon Guimelli (1994: 18-19) les approches dites des « principes organisateurs » et du « noyau central », bien que toutes deux des approches structurales, présentent des divergences importantes que Flament résume de la manière suivante:

> À côté de nombreux points de convergence, l'opposition majeure entre ces deux approches peut se schématiser ainsi: l'école de Genève insiste sur les « prises de

> position individuelles » et leurs « principes organisateurs » (qui sont généralement sociaux), tandis que l'école d'Aix, à partir de la « théorie du noyau central » de la représentation, insiste sur le « consensus ». (Flament 1999 : 201)

Pour l'école d'Aix-en-Provence, certains éléments de la RS font l'objet d'un consensus entre les membres. Un ou quelques éléments sont partagés de manière relativement consensuelle entre membres d'un groupe. Ces mêmes éléments ont, en théorie, des propriétés spécifiques de centralité cognitive comme la saillance, la connexité ou la valeur symbolique.

Pour l'école de Genève, les éléments de la RS sont distribués de manière inégale entre les membres, traduisant des prises de position de la part des membres par rapport à l'enjeu que véhicule la RS. L'organisation de cette distribution sociale est décrite en termes de « principe organisateur ». Pour l'école de Genève, il n'y a pas de consensus entre les membres, donc pas de noyau. On ne s'intéresse pas aux éléments cognitifs les plus consensuels dans le groupe, mais à la variation des éléments entre membres d'un groupe.

1.2 Les méthodes classiques d'analyse

Depuis trente ans, la méthodologie d'étude des RS s'est beaucoup développée. Elle s'est diversifiée constamment et s'est adaptée aux différents matériaux empiriques et à l'évolution de la théorie. À chacun des niveaux d'analyse — contenu, structure, noyau — correspondent aujourd'hui plusieurs méthodes ou techniques.[2]

Du côté des méthodes d'identification des éléments d'une RS ou de son contenu, on trouve des techniques de cueillette de données classiques comme l'entretien, le questionnaire ouvert ou l'ethnographie, associées à des méthodes d'analyse de contenu et d'analyse de corpus de texte. Les méthodes les plus utilisées sont cependant les méthodes associatives, par exemple les associations libres, les cartes associatives les réseaux associatifs. Selon Moliner (Moliner *et al.* 2002), ces méthodes associatives sont basées sur un principe commun: l'inférence d'éléments à partir d'un inducteur.

---

[2] Des ouvrages récents sont consacrés à ces méthodes (Abric 2003; Moliner et al. 2002). Nous renvoyons le lecteur à ceux-ci pour une description détaillée.

> Ces méthodes reposent toutes sur le même principe. Il s'agit de présenter aux sujets un mot, une courte phrase ou une expression, puis de leur demander d'y associer d'autres mots (réponses ou mots induits). Pour chaque mot inducteur et pour chaque sujet, on obtient un, deux, trois ou quatre mots induits reportés sur une fiche. Ces mots peuvent être des adjectifs, des expressions, des substantifs. (Moliner *et al.* 2002 : 70)

Les inducteurs et les induits peuvent être des mots, expressions, images, etc., dont le chercheur fait l'hypothèse qu'ils sont l'expression des éléments d'une RS. Le chercheur dresse ainsi une sorte de liste des éléments susceptibles de faire partie de la RS étudiée.

Du côté des méthodes d'identification de la structure, les méthodes varient selon qu'on s'intéresse aux structures cognitives de la RS ou à ses structures sociales. Pour les premières, on retrouve essentiellement des méthodes de caractérisation des éléments de la RS étudiée. Là encore, ces méthodes sont nombreuses, mais selon Moliner *et al.* (2002), celles-ci aussi sont basées sur un principe commun: la mise en relation par les sujets des éléments, et l'évaluation de la saillance de chaque élément par chaque sujet :

> Ces techniques, assez variées, sont toutes fondées sur le même principe: demander au sujet lui-même d'effectuer un travail de classement, de comparaison ou de hiérarchisation. L'objectif principal poursuivi par ces techniques concerne la mise en évidences des liens que les sujets établissent entre les divers éléments de la représentation étudiée. (Moliner *et al.* 2002 : 119)

Par ces méthodes, on cherche à révéler la manière dont les individus d'un groupe organisent, en moyenne, les éléments cognitifs qui composent la RS.

Pour ceux qui s'intéressent à la structure sociale, on retrouve essentiellement les méthodes d'analyse factorielle, qui mettent en évidence les variations interindividuelles ou inter-groupes :

> Montrer que les représentations sociales sont aussi des principes organisateurs des différences entre des prises de position individuelles est peut-être l'apport le plus important de l'utilisation raisonnée des analyses de type factoriel. (Doise *et al.* 1992 : 18)

Toutes ces méthodes permettent de mettre en évidence l'organisation des éléments cognitifs de la RS ou l'organisation des éléments de la RS parmi les individus.

Finalement, du côté des méthodes d'identification de la centralité des éléments de la RS, certaines renvoient à l'étude des propriétés de centralité cognitive et d'autres à l'étude des patrons de distribution sociale. Concernant les méthodes qui cherchent à identifier la centralité des éléments de la RS en fonction de leur propriété cognitive, les plus utilisées sont les méthodes d'analyse de similitude (Bouriche 2003), les méthodes de mise en cause et les méthodes d'induction par scénario ambigu (Moliner 1994). Concernant les méthodes d'étude de la centralité en fonction de ses formes de distribution sociale, nous n'en avons trouvé qu'une seule dans la littérature, à savoir l'analyse booléenne de questionnaire développé par Flament (1996; 1999).

Pour ces chercheurs, ces méthodes d'identification de la centralité des éléments dans la structure de la RS sont fondamentales. Elles sont même l'étape la plus importante de la méthodologie, car, selon leur cadre théorique, le noyau est l'identité absolue de la RS: deux RS sont considérées différentes dans la mesure où elles ont un noyau différent.

> Le noyau d'une RS [...] est son identifiant absolu. Cette définition, la plus courte qui soit, a pour corollaire qu'une représentation diffère d'une autre si et seulement si leurs noyaux diffèrent. De même, deux RS sont identiques dès lors que leurs noyaux le sont. On a autrement affaire à des variations contextuelles [...] éventuellement importantes, mais non structurales. La détermination du noyau constitue donc, un objectif essentiel pour toute étude de représentation. (Flament et Rouquette 2003 : 102)

Une RS est composée de plusieurs éléments cognitifs, dont certains sont moins stables que d'autres. Aussi, nous savons que les éléments périphériques de la RS peuvent changer rapidement en fonction des contextes d'actualisation et de leur ancrage. Par conséquent, l'étude structurale des RS orientée vers l'identification du noyau accorde davantage d'importance à la centralité. Il s'agit en quelque sorte d'un travail de tri, de filtre et de hiérarchisation des éléments, afin de distinguer ce qui est secondaire de ce qui est indispensable pour comprendre une RS donnée.

2. La problématique: l'étude du noyau central d'une RS à partir d'un corpus de presse

Les méthodes développées pour le repérage du noyau central s'inscrivent toutes dans une démarche expérimentale ou par enquête, où l'on fait intervenir les sujets d'étude à chaque niveau de la méthodologie: par exemple, lors des associations libres pour l'identification des

éléments constitutifs de la représentation, durant un questionnaire de caractérisation pour l'identification de la structure ou durant l'analyse de similitude pour l'identification du noyau.

Cependant, ce ne sont pas toutes les RS qui peuvent être étudiées dans le cadre de ces méthodes. C'est le cas, notamment, des RS diffusées par des médias de masse comme la presse écrite. Les RS véhiculées par la presse ne sont pas étudiées grâce à un questionnaire; les chercheurs doivent plutôt analyser le contenu des articles de presse. C'est grâce au texte de ces articles (parfois à leurs images) que les chercheurs accèdent aux RS qu'ils véhiculent.

Pour la théorie des RS, les communications de masse sont considérées comme un lieu de diffusion et de fabrication de première importance. La relation entre médias de masse et RS a été comprise de plusieurs manières. Pour certains, comme Flament et Rouquette, ces médias sont compris comme une source où la population s'approprie ses représentations:

> [...] les RS s'élaborent et se transmettent dans le cadre des communications, tout particulièrement les communications de masse, fréquemment reprises ou reflétées dans les échanges inter-personnels. (Flament et Rouquette 2003 : 146)

Pour Moscovici, de tels médias sont plus que des lieux de diffusion: ce sont des lieux privilégiés de « fabrication » des RS (Moscovici 1989: 100) et souvent des leviers de première importance dans la genèse et l'évolution de RS. La presse, écrit-il, « nous fournit régulièrement des expressions de représentations en émergence » (Moliner *et al.* 2002: 45).

On va même parfois jusqu'à affirmer que les médias de masse sont une condition de possibilité d'émergence de RS : « Ainsi, la communication sociale, sous ses aspects interindividuels, institutionnels et médiatiques apparaît-elle comme une condition de possibilité et de détermination des représentations et de la pensée sociale. » (Jodelet 1989: 64)

Dans d'autres circonstances, les médias de masse sont compris comme le reflet plus ou moins distordu des représentations qui circulent dans la société (Rouquette 1998 : 16-17). En somme, si la compréhension des relations entre médias de masse et RS varie, nombreux sont

ceux qui s'entendent sur son importance et sur la pertinence de son investigation pour la théorie des RS.

L'un des défis de l'étude des RS diffusées via des médias de masse, en l'occurrence la presse écrite, est de nature méthodologique. Étudier les RS fabriquées et diffusées par un média de masse comme la presse oblige à travailler à partir d'un matériel empirique atypique, à savoir le corpus de presse composé d'articles de journaux. Il existe très peu de méthodes d'analyse des RS à partir de corpus de presse.

2.1 Les méthodes d'analyse des corpus de texte dans le domaine des représentations sociales

On ne retrouve dans la littérature que peu d'études sur les RS à partir de corpus de textes de nature similaire à un corpus de presse. Dans son étude princeps de la RS de la *psychanalyse*, Moscovici (Moscovici 1976) a procédé à une analyse de la presse française. L'auteur a recueilli au total 1640 coupures de journaux parues sur une période de quatre ans dans 230 journaux différents (Moliner *et al.* 2002: 45). Lalhou (Lalhou 1995a; 1995b; 1998; 2003) a mené une étude de la RS de l'*acte de manger* à partir du dictionnaire le Grand Robert. L'auteur a retenu pour son analyse 544 définitions du dictionnaire liées de près ou de loin à la définition de *l'acte de manger*.

2.1.1 L'analyse de contenu classique vs le forage de texte

Ces deux études incarnent deux types d'analyses de corpus textuel utilisés dans le domaine des RS. Moscovici a procédé à une analyse de contenu classique, alors que Lalhou a utilisé des méthodes basées sur la statistique textuelle, que l'on appelle *forage de texte* (parfois appelée *lexicométrie*, *fouille de texte* ou en anglais *Text Mining*).

L'analyse de contenu est une démarche classique dans les sciences humaines et sociales (Bardin 1977). C'est, la plupart du temps, une analyse *à la main* d'annotation des textes au terme de laquelle le chercheur classe ses documents textuels par catégories qu'il aura lui-même définies. Cette analyse est parfois assistée informatiquement par un logiciel d'aide à l'annotation et à la catégorisation comme N-Vivo

(http://www.qsrinternational.com/products_nvivo.aspx), Prospero (http://prospero.dyndns.org:9673/prospero), Atlas (http://www.atlasti.com/index.html) ou QDA Miner (http://www.provalisresearch.com/QDAMiner/QDAMinerDesc.html).

Quant à elle, l'analyse statistique de données textuelles se fait de manière automatique ou algorithmique. L'analyse se fait de manière assistée informatiquement, mais ici, l'informatique est plus qu'un outil d'aide à l'annotation. On fait appel à des modèles mathématiques complexes pour la manipulation des données textuelles. On parle dans ce cas de méthode de *forage de texte*. Alors que dans les méthodes d'analyse de contenu classique, la classification est manuelle, ces méthodes de forage de texte permettent une classification automatique de textes à partir de leurs propriétés statistiques.

2.1.2 Avantages et inconvénients

Ces deux méthodes — l'analyse de contenu et le forage de texte — ont leurs avantages et inconvénients. L'analyse de contenu est souvent un travail très ardu, long, dépendant de l'expérience et du talent du chercheur. Cependant, aucune méthode de forage de texte basée sur la statistique textuelle ne peut égaler la richesse herméneutique d'une analyse de contenu. La statistique textuelle est aveugle au sens. Elle permet essentiellement de révéler des *patterns* de redondance et c'est a postériori que le chercheur interprète ces redondances. Aussi, la statistique textuelle doit être appliquée sur des corpus de taille importante, alors que c'est justement une des limites majeures de l'analyse de contenu classique, dans le cadre de laquelle on est souvent obligé de procéder par échantillonnage.

Outre l'étude de Moscovici, très peu de recherches sur les RS utilisent aujourd'hui l'analyse de contenu sur des corpus de grande envergure (Moliner *et al.* 2002: 45). Pour plusieurs raisons, notamment celles du temps, de la difficulté de traiter de manière exhaustive les données parce que trop nombreuse et de la place importante accordée à la subjectivité de l'analyste, on juge l'analyse de contenu insuffisante pour l'étude des RS :

> Ceci étant, une fois le matériel de recherche recueilli, et notamment quand celui-ci est particulièrement volumineux, nous nous rendons compte que son analyse risque de devenir une tâche particulièrement laborieuse et que la seule analyse de contenu

> classique ne nous suffira pas, malgré son indiscutable efficacité. (Kalampalikis 2003: 148)

Par ailleurs, depuis quelques années, on utilise de plus en plus des méthodes informatiques de forage de texte. Dans le domaine d'étude des RS, ces méthodes ont été utilisées à l'aide d'une opérationnalisation particulière, celle du logiciel ALCESTE. Elles ont donné des premiers résultats encourageants dans le domaine des RS (Lalhou 1995a; 1995b; 1998; 2003).

2.2 L'analyse des représentations sociales et le forage de texte

Nous observons, depuis environ une dizaine d'années, un intérêt grandissant de la part de la communauté des psychosociologues pour les méthodes de forage de textes dans l'étude des RS. Cette pratique a été initiée par Lahlou et reprise ensuite par d'autres (Kalampalikis 2003; Dany et Apostolidis 2002; Kalampalis et Moscovici 2005; De Alba 2004; Baugnet et Fouquet 2005; Masson et Moscovici 1997). L'étude des RS via de telles méthodes est encore à l'état exploratoire. La compatibilité entre la théorie des RS et le forage de texte reste à démontrer et les possibilités d'application sont encore peu exploitées.

Aussi, nous remarquons que les psychosociologues utilisant ce genre de méthodes se sont jusqu'à maintenant limités à l'utilisation d'une opérationnalisation particulière. Cette opérationnalisation est celle du logiciel ALCESTE, élaboré à l'origine par Max Reinert (Reinert 1993) et commercialisé par la compagnie Image (http://www.image-zafar.com/index.htm). Or, bien que dans la littérature on fasse souvent référence à la « méthode ALCESTE », ALCESTE n'est pas une méthode, mais un logiciel, c'est-à-dire une implémentation et une opérationnalisation particulière d'une méthodologie générale de forage de texte.

D'ailleurs, d'autres logiciels du même genre sont nombreux qui offrent des opérationnalisations alternatives, mais basées sur une méthodologie commune. Parmi ceux-ci, il y a d'autres logiciels commerciaux comme Wordstat de PROVALIS (http://www.provalisresearch.com/), PolyAnalys de MEGAPUTER (http://www.megaputer.com/polyanalyst.php), il y a des plates-formes *open source* comme T2K (http://alg.ncsa.uiuc.edu/do/home), WEKA (http://www.cs.waikato.ac.nz/ml/weka/), KNIME (http://www.knime.org/introduction/features) et de nombreux prototypes issus du

milieu académique tels que NUMEXCO et GRAMEXCO (Biskri et Meunier 2002; Meunier *et al.* 2006).

3. Question de recherche

La problématique de l'analyse des RS à partir de corpus de presse soulève plusieurs questions de recherche. Parmi celles-ci la possibilité, pour les méthodes de forage de texte, de satisfaire les exigences d'une méthodologie générale d'étude des RS.

Lorsqu'un chercheur étudie une RS à travers un corpus d'articles de presse, ou n'importe quel autre corpus du même genre comme des archives, des pages web, des blogues, des documents historiques, il n'a accès qu'aux *traces empirico-sémiotiques* laissées par le travail cognitif de la population étudiée. Ces traces sont des données textuelles.

Cela représente un défi méthodologique important, car le chercheur doit se doter de méthodes d'analyse de texte, qui lui permettront d'atteindre les trois niveaux d'analyse des RS, c'est-à-dire: identifier dans ces textes les indicateurs qui sont l'expression des éléments cognitifs spécifiques à la RS qu'il veut étudier; identifier les structures qui organisent la RS étudiée; et identifier la centralité des éléments cognitifs qui composent la représentation.

Notre question de recherche est donc avant tout de nature méthodologique. Nous nous questionnons à savoir si les méthodes de forage de texte, appliquées à un corpus de presse, permettent d'atteindre ces trois niveaux d'analyse d'une méthodologie générale d'étude des RS.

4. Hypothèse de recherche et objectifs

Notre objectif général de recherche est l'évaluation de la pertinence et de la compatibilité des méthodes de forage de texte pour le domaine d'étude des RS. Notre hypothèse est que le forage de texte est effectivement approprié à l'étude des RS, qu'il peut être appliqué en tant que méthode d'étude des RS et qu'il atteint les trois niveaux d'analyse d'une méthodologie générale des RS. Il nécessite cependant des choix d'opérationnalisation spécifiques, qui doivent être effectués en fonction de l'interprétation particulière que la théorie des RS offre du forage de texte.

Pour valider cette hypothèse, nous procédons en quatre étapes qui correspondent à nos quatre objectifs concrets de recherche. Premièrement, il nous faut mieux comprendre la théorie des RS, les concepts en jeu, notamment celui d'architecture et de noyau central. De plus, il faut mieux comprendre le lien que cette théorie fait avec le langage, celui-ci occupant un rôle de premier plan dans la méthode des RS. En effet, le langage est la principale source de données empiriques du chercheur, que ce soit sous sa forme écrite ou sa forme orale.

Nous analysons les principaux concepts de la théorie des RS que sont la consonance sociocognitive, l'ancrage, l'objectivation et l'architecture des RS. Les méthodes et techniques d'analyse classique des RS sont fondées théoriquement sur ces concepts. Si nous voulons évaluer les techniques d'analyse des RS à partir de forage de texte, il nous faut faire de même.

Deuxièmement, il nous faut mieux comprendre les étapes et opérations qui forment une méthodologie de forage de texte. Ces étapes sont communes à plusieurs logiciels de forage de texte. Lorsqu'une recherche scientifique est basée sur une méthodologie faisant appel à des méthodes informatiques, il est important que la technologie utilisée pour le traitement des données ne devienne pas ce qu'on appelle une « boîte noire ». Cette exigence a été rappelée par des psychosociologues, qui se sont initiés à l'étude des RS via le forage de texte:

> Mener une réflexion sur l'outil informatique consiste à sortir de ce béhaviorisme numérique où on se contente d'entrer d'un côté, sous une forme quelconque, des données et d'attendre impatiemment de l'autre côté que les résultats déjà exploitables sortent. [...] il nous faut cesser de considérer l'ordinateur comme une nouvelle boîte noire et prendre conscience de ce qui se passe à l'intérieur. Il nous faut donc affronter la diversité des logiciels et ce qu'elle recouvre, et notamment pour l'analyse textuelle. Face à la convivialité grandissante de l'univers informatique qui agit comme une véritable poudre aux yeux [...] on oublie facilement que derrière tout logiciel informatique se cache une conception plus ou moins spécialisée du monde. (Buschini et Kalampalikis 2002: 189)

Pour ces psychosociologues, et pour les chercheurs des sciences sociales et humaines en général, une méthodologie faisant appel à l'informatique ne sera pertinente que si les étapes et opérations en jeu sont bien comprises et les plus transparentes possible.

Nous identifierons, parmi les différents logiciels de forage de texte, les étapes et opérations qui leur sont communes et qui forment une chaîne de traitement générale de forage de texte. Ces étapes sont (1) la construction du corpus, (2) le choix d'un domaine d'information et la segmentation, (3) le choix d'une unité d'information et l'indexation, (4) le choix d'une valeur de pondération et la vectorisation, (5) la construction de la matrice et sa réduction dimensionnelle et (6) la classification automatique.

Troisièmement, il nous faut évaluer de manière critique les apports du forage de texte pour l'étude des RS. Jusqu'à maintenant, les études de RS à l'aide de méthodes de forage de texte sont relativement circonscrites. C'est-à-dire que le forage de texte, bien qu'il soit utilisé depuis plus d'une dizaine d'années dans le domaine des RS, n'a fait l'objet que de très peu d'innovation au niveau de son opérationnalisation. Ces études ont toutes utilisé le même logiciel: ALCESTE. Celui-ci ne représente pourtant qu'une opérationnalisation parmi plusieurs possibles. Ces études sont également toutes très proches du type d'analyse à l'origine associé à ce logiciel et à son auteur, à savoir l'analyse des « mondes lexicaux » de Reinert (Reinert 1993; voir aussi Lalhou 1995b: 139). Or, il n'est pas certain qu'ALCESTE soit approprié à l'étude des RS.

Nous menons une analyse critique de la compatibilité du forage de texte via le logiciel ALCESTE pour l'étude des RS. Deux limites importantes du logiciel ALCESTE sont analysées: le type de domaine d'information utilisé et le type de classification automatique. Une troisième limite sera soulignée, mais qui concerne cette fois le forage de texte en général. Cette limite concerne la difficulté de repérer, avec une telle chaîne de traitement, les structures et le noyau de la RS étudiée dans le corpus de presse.

Quatrièmement, nous menons une expérimentation afin d'évaluer différents choix d'opérationnalisation de forage de texte mieux adaptés que ceux d'ALCESTE, selon nous, à l'étude des RS. Ces choix d'opérationnalisation nous permettent de développer des techniques, qui atteignent les trois niveaux d'analyse d'une méthodologie générale d'étude des RS.

Le premier niveau d'analyse est atteint grâce à la sélection d'algorithmes mieux adaptés aux étapes des choix de la segmentation et de la classification. Le deuxième niveau d'analyse est

atteint en développant des techniques d'analyse de la classification, inspirées de l'analyse de similitude et de l'analyse du consensus social. Le troisième niveau d'analyse est atteint en développant des techniques de repérage de la centralité dans les structures cognitives et sociales de la RS.

Les choix d'opérationnalisation et les techniques d'analyse développées sont finalement illustrés à partir d'un cas empirique dont nous avons déjà eu l'occasion d'entamer l'analyse ailleurs (Chartier et al., 2008). Il s'agit de la RS des *accommodements raisonnables* construite via trois journaux québécois: Le Devoir, La Presse et Le Soleil. Nous terminons en évaluant la validité des choix d'opérationnalisation et des techniques développées par des tests de triangulation dans les résultats générés.

CHAPITRE II

CADRE THÉORIQUE DES REPRÉSENTATIONS SOCIALES

Introduction

Avant d'aborder le forage de texte et d'évaluer sa pertinence pour l'étude des RS, nous présentons, dans ce chapitre, quatre concepts importants pour comprendre la théorie des RS. Ce chapitre est divisé en quatre sections où chacune traite l'un des concepts en question. Les trois premières sections présentent trois concepts importants pour le domaine d'étude de la pensée sociale. Ces concepts permettent de comprendre la diffusion, la transformation et la communication des éléments cognitifs formant une RS. Ces processus sociocognitifs ont lieu dans les interactions et les communications sociales. Ces trois concepts sont la consonance sociocognitive, l'ancrage et l'objectivation.

La quatrième section présente le concept d'architecture de la RS. Ce concept général porte à la fois sur les éléments d'une RS et sur les structures qui les organisent. Nous verrons que l'étude des éléments cognitifs composant une RS passe principalement par la médiation du langage. Nous verrons également que les structures sont de nature cognitive et sociale et forment un système central et un système périphérique.

1. La consonance sociocognitive

Une RS est un phénomène émergeant de différents processus sociocognitifs de la pensée sociale ou de la cognition sociale. Ces processus sont impliqués dans les interactions et les communications sociales entre membres d'un groupe ou d'une société. Plusieurs concepts ont été développés dans différents courants théoriques pour décrire et expliquer ces processus

sociocognitifs[3]: par exemple la théorie de l'attribution (Försterling 2001), la théorie des stéréotypes et des attitudes (Kunda 1999), la théorie des heuristiques de jugement (Bless *et al.* 2004), etc.

Parmi ces théories, un de leurs concepts est particulièrement important pour la théorie des RS, soit celui de consonance sociocognitive (aussi appelé équilibre sociocognitif, principe d'homéostasie, balance de Heider, entre autres). On peut retrouver les premières ébauches de ce concept dans les écrits classiques de Marcel Mauss et dans son étude anthropologique sur le don. Mais le concept a surtout été formalisé par la psychologie sociale dans des travaux comme ceux de Festinger (Festinger 1957), de Heider (Heider 1958) ou Watzlawick (Watzlawick *et al.* 1967).[4]

Le processus sociocognitif dont veut rendre compte ce concept de consonance sociocognitive est relativement simple et peut être résumé sous la forme d'une heuristique (règle) de la manière suivante: deux individus qui entretiennent des relations sociales consonantes — ex.: ils perçoivent de manière consonante leur relation de membre à un même groupe — auront également tendance à rechercher la consonance quant à leur cognition sur un tiers objet — ex.: leur interprétation d'une pratique sociale, d'une coutume, d'un rituel, d'un symbole: « Heider's balance theory posited that if two individuals were friends, they should have similar evaluation of an object. » (Monge et Contractor 2003: 204)

Autrement dit, ce processus de recherche de consonance sociocognitive s'articule autour de trois variables: une relation sociale entre deux ou plusieurs individus, un objet de leur attention conjointe et un processus cognitif de catégorisation de cet objet. Il y a consonance sociocognitive lorsque ces individus sont liés socialement (par une relation d'amitié, une relation de membre, une relation familiale, etc.) et que l'objet de l'attention conjointe est catégorisé de manière consonante entre eux. Lorsque ces conditions ne sont pas remplies, on dira que les individus impliqués dans cette situation sont en dissonance.

---

[3] L'étude de la pensée sociale est associée à la tradition française de la psychosociologie cognitive. Elle trouve un correspondant dans la tradition anglo-saxonne autour de l'étude de la cognition sociale.

[4] Pour un résumé, voir Guimelli (Guimelli 1999 : 40 et suivantes).

La consonance sociocognitive peut être décrite comme une heuristique pour l'action dont les effets sont à la base d'une forme de diffusion sociale des cognitions d'une RS et d'une modalité de socialisation cognitive parmi les membres d'un groupe. Il s'agit d'un processus de coordination intercognitive entre membres d'un groupe, menant au partage relativement consensuel de certaines cognitions importantes pour le groupe.

1.1 Une heuristique pour l'action et le jugement

Un individu qui se retrouve dans une situation de dissonance sociocognitive vivra ce qu'on appelle un inconfort ou une tension, plus ou moins grande, qui peut être maintenue, mais à un coût psychique et social: la rupture éventuelle d'une interaction, d'une amitié, d'un groupe, etc. Phénoménologiquement, la consonance sociocognitive est vécue comme une heuristique directionnelle et motivationnelle pour les individus. L'état interne senti en situation de dissonance sert de guide pour diriger l'individu et motiver le choix des actions appropriées (Bless *et al.* 2004: 83).

Selon cette heuristique, les agents cherchent à résoudre les problèmes de dissonance sociocognitive. Ils le font de plusieurs manières. La dissonance entre deux agents peut être réduite en modifiant l'une des trois variables de la règle de la consonance: soit que l'un des agents modifie sa catégorisation (cognition) en adoptant celle de l'autre; soit que l'agent modifie l'objet de dissonance; ou encore, que l'agent modifie le statut ou supprime la relation sociale qu'il entretient avec l'autre individu.

Cette heuristique est une règle spécifique à la pensée sociale, actualisée partout où une situation implique à la fois de la cognition, de l'interaction sociale et de l'attention conjointe.[5] Lorsque les membres d'un groupe social se rencontrent, interagissent et communiquent à propos d'un objet important pour eux (une pratique, une valeur, un principe, etc.), chacun identifie les cognitions des autres membres à propos de cet objet et évalue leur

[5] Par ailleurs, la pensée sociale ne disparaît pas lorsqu'un individu met sa blouse blanche de scientifique. Les sciences sociales ont montré, depuis Ludwick Fleck (1935/2005) et son concept de « collectif de pensée », que les individus membres d'une communauté scientifique actualisent également ces processus sociocognitifs. Ceci fait également partie de la théorie des changements de paradigme de Thomas Kuhn (1970).

degré de consonance. Si un membre se rend compte qu'il est dissonant par rapport au reste du groupe, il aura tendance à adapter ses propres cognitions à celles des autres, de manière à maintenir le lien social. L'hypothèse derrière ce concept est que le maintien du lien social est plus important que le maintien d'une croyance, d'une opinion, d'une interprétation, bref de certains éléments cognitifs.

La recherche de consonance sociocognitive durant les interactions entre individus liés socialement les entraînent dans une coordination intercognitive les uns avec les autres. Cette coordination peut être comprise de plusieurs manières. Elle est, entre autres, un processus d'adaptation mutuel entre membres. Elle mène à une forme de diffusion sociale ou ce que d'autres appellent de la socialisation cognitive dans le groupe.

1.2 Diffusion sociale et socialisation cognitive

Les effets de la recherche de consonance sociocognitive peuvent être interprétés comme des formes spécifiques de diffusion sociale d'élément cognitif. Les membres d'un groupe, en s'adaptant mutuellement aux cognitions des uns et des autres durant les interactions, sont à l'origine de la circulation des cognitions dans le groupe (Monge et Contractor 2003: 205).[6]

La diffusion sociale d'éléments cognitifs dans le groupe est importante pour la construction des RS. La diffusion sociale stabilise des structures écologiques de distribution sociale des cognitions, typiques à celle des RS. En effet, la distribution sociale des éléments cognitifs formant une RS, est organisée selon des structures écologiques spécifiques. L'une de ces structures est caractérisée par une forte cohésion cognitive intragroupe, c'est-à-dire par un fort consensus entre membres sur quelques éléments cognitifs fondamentaux pour eux.

---

[6] Il faut apporter une nuance ici. Les cognitions ne « circulent » pas vraiment parmi des individus. Ce qui circule ce sont des ondes sonores, des rayons lumineux, du sémiotique en général, qui sont produits lors de la communication. On fait l'hypothèse que ce sémiotique exprime certains éléments cognitifs. C'est par un jeu de métaphore que certains disent que les idées (Morin 1991), les croyances (Bronner 2003), les représentations (Sperber 1996), les mèmes (Blackmore 2000), se diffusent dans la population. Nous y reviendrons, en discutant du concept d'objectivation.

Les effets de la recherche de consonance sociocognitive peuvent également être interprétés comme un processus de socialisation cognitive (Zerubavel 1997; Berger et Luckmann 2003). Par socialisation cognitive, on entend la nécessité pour un groupe ou une société, de partager certains types d'éléments cognitifs importants :

> « It is the process of cognitive socialization that allows us to enter the social, intersubjective world. Becoming social implies learning not only how to act but also how to think in a social manner. As we become socialized and learn to see the world through the mental lenses of particular thought communities, we come to assign to objects the same meaning that they have for others around us, to both ignore and remember the same things that they do, and to laugh at the same things that they find funny. » (Zerubavel 1997 : 15)

La socialisation cognitive est vue comme un processus essentiel à l'intégration des membres, au maintien des interactions entre individus et à la cohésion du groupe.[7]

Diffusion sociale d'éléments cognitifs et socialisation cognitive des membres du groupe sont tous (en partie) les effets d'un processus typique de la pensée sociale, qui consiste pour chaque individu à rechercher dans ses interactions sociales et ses communications de la consonance sociocognitive. Il s'agit d'un processus à la base du partage social d'éléments cognitifs importants pour un groupe, condition indispensable à la construction collective des RS. Ce sont ces éléments cognitifs, relativement consensuels entre membres d'un groupe, qui sont susceptibles de former le noyau central d'une RS.

D'autres processus typiques de la pensée sociale contribuent à la construction collective des RS. C'est le cas de l'ancrage et de l'objectivation. Le concept d'ancrage est utilisé pour décrire non pas les processus de partage consensuel d'éléments cognitifs dans un groupe, mais pour décrire les changements et l'évolution de ces cognitions dans un groupe. Quant à l'objectivation, c'est un concept pour décrire la mise en forme nécessaire des éléments cognitifs pour leur communication et leur diffusion efficace dans le groupe.

---

[7] Sur la nécessité d'une consonance sociocognitive entre individus dans une situation d'interaction sociale, voir les travaux classiques de Erving Goffman (1974 ; 1991). Pour Goffman, une situation d'interaction sociale se maintiendra dans la mesure où les interactants utilisent le même *cadre*, c'est-à-dire une même forme de cognition, pour comprendre la situation.

2. L'ancrage

Le partage d'éléments cognitifs au sein des membres d'un groupe est toujours partiel et temporaire. En même temps que l'économie de certains éléments cognitifs d'un groupe tend vers le consensus, d'autres processus de la pensée sociale sont à la base d'une tendance contraire, mais complémentaire. Cette tendance est celle du changement et de la différenciation. En même temps qu'il y a stabilisation de ce que Durkheim appelait « l'assiette mentale » d'un groupe (Durkheim 1912), il y a évolution et transformation des éléments constituant son économie cognitive. Ces transformations sont fonction de l'idiosyncrasie des positions sociales des individus. Le concept d'ancrage est utilisé pour décrire ce processus.

Le concept d'ancrage est très près de celui de la consonance sociocognitive. Il se distingue cependant, car il décrit un processus spécifique d'appropriation des cognitions d'un groupe en fonction des appartenances multiples de chaque individu. L'ancrage est un concept pour décrire le processus par lequel un individu, membre d'un groupe social, ajuste les cognitions qui circulent dans le groupe en fonction de son idiosyncrasie positionnelle, c'est-à-dire de ses appartenances simultanées à plusieurs groupes.

L'ancrage veut rendre compte du fait que la recherche de consonance sociocognitive d'un individu n'est jamais exclusive à un seul groupe. Un individu cherche la consonance avec des individus membres de groupe différent. Il doit gérer parfois des contractions et ajuster ses cognitions en fonction de plusieurs relations sociales. L'ancrage implique donc deux autres concepts : celui de position sociale et celui d'intégration.

2.1 Position sociale : relation d'appartenance multiple et connaissance d'arrière-plan

Chaque individu occupe dans la société une position sociale unique. En nous inspirant des écrits classiques de Simmel sur les cercles sociaux (Simmel 1908), nous pouvons définir la position sociale d'un individu par l'extension de ses appartenances à différents groupes sociaux (Simmel parle de « cercles sociaux »). Selon Simmel, les appartenances de groupes d'un individu sont toujours multiples, c'est-à-dire que chaque individu est toujours inscrit dans plus d'un groupe social à la fois (famille, ami-e-s, travail, université, associations

diverses, etc.).[8] De plus, la configuration des appartenances est unique à chaque individu, c'est-à-dire que chacun est inscrit dans une combinaison idiosyncrasique de relations sociales. Cette configuration d'appartenances multiples incarne sa position sociale, ou ce que Simmel appelle son « système de coordonnées » social (1908 : 416).

La position sociale d'un individu lui sert de connaissance d'arrière-plan (*background*). Elle est à la fois du capital et de la contrainte pour lui. Un individu est toujours l'actualisation idiosyncrasique du capital social et cognitif de plusieurs groupes à la fois. En même temps, ce *background* est contraignant, car l'individu doit maintenir avec lui une certaine consonance sociocognitive.[9]

C'est-à-dire que chaque individu communique et interagit avec différents individus, eux-mêmes membres de différents groupes. Par conséquent, chaque individu recherche la consonance sociocognitive avec plus d'un groupe social à la fois. Il doit maintenir un équilibre complexe entre plusieurs liens sociaux et doit ajuster ses cognitions en conséquence, ou rompre le lien social.

Dans ses interactions avec les membres d'un groupe donné, un individu est mis en contact avec les éléments cognitifs qui circulent dans ce groupe. Cet individu s'approprie ces éléments par socialisation, mais en fonction de son réseau complexe d'appartenances multiples. Les éléments cognitifs qui circulent dans un groupe peuvent parfois être en dissonance avec d'autres cognitions véhiculées dans un autre groupe auquel est membre ce même individu. Il doit gérer ces contradictions possibles. Il doit ancrer les éléments cognitifs du groupe en fonction de sa position sociale.

Alors que le concept de consonance sociocognitive ne tenait compte que d'un lien social à la fois, le concept d'ancrage tient compte de la position sociale de l'individu dans le réseau des appartenances multiples. Le premier permettait d'expliquer comment un individu s'adaptait aux cognitions véhiculées par un groupe. Le deuxième permet d'expliquer comment un

---

[8] Pour Simmel, les appartenances multiples sont une structure sociale typique de la modernité.

[9] C'est par exemple, mais dans d'autres mots, la thèse développée par Bourdieu dans *La distinction* (1979).

individu, tout en cherchant à s'adapter, est malgré tout obligé d'ajuster les éléments cognitifs du groupe aux contraintes de sa position sociale. Alors que le premier processus est source de diffusion consensuelle des cognitions dans le groupe, le processus d'ancrage est source d'innovation et de changement dans le groupe.

2.2 Intégration et innovation

Dans un groupe, certains éléments cognitifs tendent à être relativement bien partagés par les membres, mais d'autres sont dynamiques et se transforment en fonction des rapports sociaux et des communications. Chaque individu ancre et ajuste les éléments cognitifs qui circulent dans un groupe social en fonction de sa position sociale. Ce processus d'ajustement est un travail d'intégration et d'innovation.

L'ancrage est un concept développé à l'origine par Moscovici (Moscovici 1976), repris et précisé par plusieurs (Doise *et al.* 1986; Jodelet 1989; Guimelli 1999). L'ancrage est un double processus d'intégration d'un élément cognitif par un individu, et d'innovation.

L'ancrage est d'abord un processus d'intégration, d'appropriation ou d'acquisition d'un élément cognitif lors des interactions et des communications sociales entre individus. L'ancrage est la capacité des individus à intégrer des éléments cognitifs véhiculés par un groupe, en harmonie avec sa position sociale et les cognitions qu'il partage déjà avec d'autres groupes. Il s'agit d'un travail d'enracinement et de familiarisation en référence au capital social et cognitif de sa position sociale :

> L'ancrage va donc consister, entre autres, dans l'intégration d'élément de connaissance nouveau dans un réseau de catégories plus familières, déjà signifiantes pour le groupe. (Guimelli 1999 : 67)
>
> Le processus d'ancrage consiste en l'incorporation de nouveaux éléments de savoir dans un réseau de catégories plus familières. (Doise *et al.* 1992 : 14-15)
>
> [...] l'ancrage enracine la représentation et son objet dans un réseau de significations qui permet de les situer en regard des valeurs sociales et de leur donner cohérence. (Jodelet 1989 : 73)

L'ancrage est également un processus de transformation et d'innovation. C'est un processus par lequel un individu modifie un élément cognitif partagé avec les autres membres d'un groupe, en fonction de son idiosyncrasie positionnelle.

Dit autrement, un agent cognitif ne fait pas que copier les éléments cognitifs qui circulent dans les groupes auxquels il est membre, il les ajuste et les transforme, car ils ne sont pas tous nécessairement consonants entre eux. L'ancrage, par un individu, des éléments cognitifs véhiculés par un groupe procède toujours en fonction de son propre « système de coordonnées ». Ce « système de coordonnées » fonctionne comme un système de contraintes sur l'individu, c'est-à-dire qui le force à ancrer les cognitions du groupe dans celles qu'il partage déjà avec d'autres individus et des groupes différents.

Cette contrainte devient aussi un moteur pour l'innovation et le changement. En effet, en cherchant à ancrer ou à intégrer les éléments cognitifs d'un groupe, un individu doit malgré tout chercher à préserver les relations de consonance sociocognitive qu'il entretient avec d'autres individus et d'autres groupes. Cette double tension est une caractéristique essentielle de l'ancrage qui peut mener à de l'innovation, c'est-à-dire à transformer, parfois de manière minime, parfois de manière radicale, certains éléments cognitifs et parfois une RS dans sa totalité.

L'objectivation est un autre processus typique de la pensée sociale indispensable à la construction des RS. L'objectivation est un concept pour décrire un processus de *mise en forme* des éléments cognitifs communiqués par les individus dans les interactions sociales.

3. L'objectivation

Durant les interactions sociales, seulement une petite partie des cognitions des individus sont communiquées, c'est-à-dire sont exprimées publiquement via un système sémiotique ou un signe, par exemple le langage, l'icône, le texte, etc. Un anthropologue comme Dan Sperber (1989, 1996) parlera d'ailleurs de représentation publique :

> Une représentation peut exister à l'intérieur même de l'utilisateur; il s'agit alors d'une représentation mentale. [...] une représentation peut aussi exister dans l'environnement de l'utilisateur par exemple le texte qui est sous vos yeux; il s'agit

> alors d'une représentation publique. Une représentation publique est généralement un moyen de communication entre un producteur et un utilisateur distincts l'un de l'autre. (Sperber 1989 : 133)

Ce sont seulement ces éléments cognitifs qui peuvent faire l'objet d'un partage social et sur lesquels un travail de recherche de consonance et d'ancrage est opéré par les individus. Rendre un élément cognitif public, c'est le communiquer. C'est via la médiation de différentes formes sémiotiques, que l'on puisse considérer que les cognitions circulent ou se diffusent entre individus.

La communication est évidemment à la base du partage collectif des cognitions menant à la construction des RS. La communication des cognitions obéies à différentes formes de diffusion sociale (Roger 2003) ou ce que d'autres appellent de la contagion (Sperber 1996; Monge et Contractor 2003). Ces communications sont favorisées, entre autres, par des structures écologiques spécifiques, c'est-à-dire des configurations relationnelles et communicationnelles entre individus, qui favorisent la diffusion. Ces structures écologiques sont déterminantes pour l'émergence des RS (Rouquette 1998).

Cependant, la communication des cognitions dans les interactions nécessite davantage que des *patterns* relationnels appropriés entre individus. Les cognitions doivent être communiquées dans un format spécifique, qui implique différentes opérations de mise en forme. Il s'agit d'un processus typique de la pensée sociale.

L'objectivation est un concept pour décrire le processus par lequel des individus expriment dans un format public leur cognition, afin de les communiquer aux autres. L'objectivation est généralement décrite comme un processus de sélection et de schématisation que les agents opèrent sur leur cognition afin de les rendre plus intelligibles et faciles à appréhender et à mémoriser. Ceci leur permet une meilleure communication dans le groupe. Doise et ses collaborateurs préciseront que : « L'objectivation rend concret ce qui est abstrait [...] elle facilite la communication, ce qui est de la plus grande importance pour le tissage du lien social » (Doise *et al.* 1992 : 14).

L'objectivation rend concret ce qui est abstrait, c'est-à-dire qu'il s'agit d'un processus par lequel un individu formate des idées, des principes, des concepts généraux, etc., afin de les

exprimer via des notions familières et tangibles, comme des expressions, des maximes, des proverbes, etc. Ceci les rend plus faciles à appréhender par les congénères (Sales-Wuillemin 2005 : 187).

On décrit également le processus de l'objectivation par ses opérations de sélection et de schématisation sur les cognitions communiquées et par ses effets de stabilisation sur l'économie cognitive du groupe.

3.1 La sélection

On dira que communiquer c'est sélectionner. Chaque individu possède des cognitions à propos du monde qui lui sont idiosyncrasiques. Le capital cognitif de chaque individu est unique compte tenu de son contexte, son vécu, son histoire biographique, sa position sociale, ses capacités cognitives, etc. Si cet individu veut rendre publique une de ses cognitions, il doit sélectionner dans ce capital ce qui est pertinent, c'est-à-dire ce qui est le plus susceptible d'être intelligible pour les autres. Ce processus de sélection est une dimension importante de l'objectivation :

> Sorte de filtrage par lequel sont éliminées certaines informations ambiguës ou problématiques [...] Elle se fait grâce à une condensation des informations et un rejet des aspects conflictuels. Il reste à ce niveau que ce qui est homogène et qui fait sens pour le groupe. (Sales-Wuillemin 2005 : 187-188)

Cette sélection, ou filtrage, est fait selon plusieurs paramètres, notamment les objectifs de la communication, les valeurs du groupe, le contexte, la pertinence et les hypothèses que le locuteur émet sur les attentes de l'auditeur (Sperber et Wilson 1989; Guimelli 1999 : 65).

L'objectivation est le corollaire de l'ancrage. Si chaque individu ancre, c'est-à-dire personnalise les cognitions partagées par le groupe durant sa socialisation cognitive, lorsqu'il communique à son tour ces cognitions, il a tendance à s'exprimer via des lieux communs, généralement des expressions connues et redondantes.

3.2 La schématisation

La schématisation est à la fois une simplification et une décontextualisation des éléments cognitifs que l'individu veut communiquer. Les agents cognitifs ne font pas seulement sélectionner l'information qu'ils communiquent, ils la simplifient, la généralise, de manière à la rendre plus facile à comprendre : « Ce processus relève très directement de la pensée sociale qui simplifie les éléments de l'information à grand trait à partir d'une logique qui reste interne au groupe. » (Guimelli 1999 : 65)

La décontextualisation est une forme d'abstraction des éléments cognitifs idiosyncrasiques à un individu. Dans la communication, chaque individu exprime ses cognitions dans un format décontextualisé de manière à les rendre plus facilement réinterprétables en fonction des attentes du groupe. La décontextualisation des cognitions leur permet d'être ensuite récupérées par d'autres et réutilisées dans des contextes très différents :

> [...] les éléments d'information sélectionnés sont détachés du corpus initial sans tenir compte du contexte dans lequel ils se trouvaient. Ainsi, ils peuvent prendre une signification globale plus proche des attentes du groupe. (Guimelli 1999 : 65)

L'objectivation est un processus essentiel de mise en forme des cognitions pour la communication. La sélection et la schématisation permettent de transformer les cognitions de manière à les rendre intelligibles pour tous et pertinentes dans plusieurs contextes d'utilisation.

3.3 La stabilisation

L'objectivation est finalement à la base de la stabilisation ou de la cristallisation de certains éléments cognitifs dans le groupe. Certains parlent dans ce cas d'institutionnalisation (Berger et Luckmann 2003; De Munk 1999). Trois effets de l'objectivation des cognitions sont particulièrement importants pour la stabilisation des cognitions dans un groupe : l'extériorisation par le signe, la contrainte du système de signe, la pérennité du signe.

Premièrement, en exprimant dans un format public une cognition — via un mot, une icône, un artefact, un signe en général — celle-ci acquiert une dimension d'extériorité. L'objectivation est à la base de la stabilisation de l'économie cognitive, car elle permet ce

que Durkheim appelait la « communion des esprits ». En effet, cette extériorité signifie que le signe devient à son tour le représentant externe de la cognition. Il l'incarne dans un artefact perceptible à son tour par les autres individus. Ce signe, pour Durkheim, était compris en termes de symbolisation :

> [...] des représentations collectives ne peuvent se constituer qu'en s'incarnant dans des objets matériels, choses, êtres de toutes sortes, figures, mouvements, sons, mots, etc., qui les figurent extérieurement et les symbolisent, car c'est seulement en exprimant leurs sentiments, en les traduisant par un signe, en les symbolisant extérieurement que les consciences individuelles, naturellement closes les unes aux autres, peuvent sentir qu'elles communient et sont à l'unisson. (Durkheim 1914)

Par exemple, le drapeau incarne ou symbolise le concept de patriotisme, le crucifix celui de la chrétienté, etc. L'objectivation cependant, passe avant tout par le langage, qui est le système de signes le plus important et le plus sophistiqué des sociétés humaines (Berger et Lunckmann 2003 : 55). Dans les communications, chaque individu rend une partie de ses cognitions accessibles, c'est-à-dire perceptibles à l'autre, par des signes externes et via lesquels l'interprétation de l'élément cognitif peut être négociée ou éventuellement uniformisée.

Deuxièmement, le signe sert de modèle à la cognition. Il stabilise le flux des pensées dans un ensemble de formes défini et relativement clos. L'un sert de modèle à l'autre, ou pour reprendre à nouveau une expression de Durkheim, le langage est « l'armature » dans laquelle se déploie la pensée : « Le langage n'est pas seulement le revêtement extérieur de la pensée; c'en est l'armature interne. Il ne se borne pas à la traduire au-dehors une fois qu'elle est formée; il sert à la faire. » (Durkheim 1912)

On affirmera par conséquent que le signe, et particulièrement le langage, est une coercition pour la cognition :

> En tant que système de signes, le langage possède la qualité d'objectivité. Je le rencontre en tant qu'artifice [artefact] extérieur à moi-même et il m'apparaît coercitif. Le langage me contraint à ses propres modèles. (Berger et Luckmann 2003 : 57)

L'objectivation, en tant que processus d'expression publique d'un élément cognitif dans un signe, est vue comme source de stabilisation des cognitions, pas seulement parce qu'il lui

confère une extériorité, mais parce que le signe est en soi un système de contraintes sur la pensée.

Troisièmement, le signe est une mémoire externe pour le groupe, c'est-à-dire qu'il est un support indépendant dans lequel sont déposés des éléments importants de manière parfois plus durable que par la mémoire des individus. Durkheim à nouveau mentionnera :

> [Lorsque les sentiments collectifs] viennent s'inscrire sur des choses qui durent, ils deviennent eux-mêmes durables. Ces choses les rappellent sans cesse aux esprits et les tiennent perpétuellement en éveil; c'est comme si la cause initiale qui les a suscités continuait à agir. (Durkheim 1912)

Le texte ou l'icône est souvent considéré comme des exemples canoniques des mémoires externes des groupes (Goody 1979). L'objectivation est source de stabilisation des cognitions dans un groupe, car la pérennité de la matérialité du signe contribue à celle de l'élément cognitif.

4. Architecture de la représentation sociale : éléments, structures et noyau

Les trois premiers concepts que nous avons vus sont des concepts pour décrire des processus de la pensée sociale menant à la diffusion, la transformation et la communication des éléments cognitifs formant une RS. Ces processus montrent que les RS ne sont jamais totalement fixées, cristallisées ou instituées. Elles sont en constante évolution : de nouvelles apparaissent, plusieurs versions d'une même RS peuvent entrer en compétition à un certains moments, et finalement, certaines mourront et seront remplacées.

Pour Moliner et ses collaborateurs (Moliner *et al.* 2002), ceci correspond au cycle de vie des RS, composé de trois étapes soit l'émergence, la stabilité et la transformation :

> Les représentations ne sont pas des structures statiques. Elles naissent, elles se transforment et parfois disparaissent au gré des évolutions de l'environnement social. Dans un effort d'adaptation constant, les groupes construisent, transforment ou abandonnent leurs représentations du monde. [...] il est possible de repérer trois périodes importantes dans l'histoire d'une représentation sociale : la phase

> d'émergence, la phase de stabilité et celle de transformation. (Moliner *et al.* 2002 : 32)[10]

Pour certains chercheurs et chercheuse, on observe dans ce cycle de vie des phénomènes de stabilisation. L'organisation cognitive et sociale des cognitions, dans certains groupes et à certaines périodes, se stabilise et il est possible d'en donner à ce moment une description. Le concept d'architecture décrit les structures des RS caractérisées par une certaine stabilité dans un groupe. Du point de vue de l'étude empirique des RS, c'est surtout lorsqu'une RS a atteint un plateau de stabilité qu'il est possible d'en repérer et d'en analyser les structures.[11]

Nous pouvons, à la suite de Flament, définir de manière générale une RS comme un ensemble organisé d'éléments cognitifs à propos d'un objet important pour un groupe social et que l'on retrouve présent de manière saillante dans les communications sociales (Flament 1994b : 37). Deux dimensions sont donc essentielles à définir dans le concept d'architecture des RS : les éléments cognitifs qui composent la RS et les structures de leur organisation, tout spécialement son système central.

Cette définition du concept de RS correspond à l'analyse structurale « aixoise » des RS. Elle a pour origine la thèse du noyau central de Abric (Abric 1994a) selon laquelle le concept de RS décrit un système composé d'éléments cognitifs organisé par des structures formées d'un système central, aussi appelé un noyau, et d'une périphérie. Les éléments centraux et périphériques sont donc structurellement différenciés.

L'objectif premier d'une analyse structurale (aixoise) des RS, est le repérage de ce noyau de la RS, car on considère que c'est le noyau qui détermine la RS. Autrement dit, c'est le noyau

---

[10] Il existe évidemment dans la littérature d'autres manières de schématiser ce cycle de vie. Pour une conception de la société où l'évolution des connaissances obéit à un espace darwinien, voir Popper (Popper 1998). Pour une conception de la société où les croyances naissent et meurent selon un marché cognitif, voir Bronner (Bronner 2003). Voir également la noosphère de Morin (Morin 1991) ou le modèle épidémiologique de Sperber (Sperber 1996).

[11] Ceci ne signifie pas que les RS dans les phases d'émergence et de transformation ne sont pas organisées selon différentes structures spécifiques. Ces structures sont cependant plus difficiles à analyser. Elles nécessitent de nouveaux concepts et de nouvelles méthodes (Moliner 2001).

qui forme l'identité de la RS. Deux RS avec des noyaux différents seront considérées également comme différentes. Ainsi, une partie importante du travail du psychosociologue consiste à repérer dans l'ensemble des éléments cognitifs formant une RS, ceux qui sont les plus susceptibles de former son noyau :

> Tous les éléments de la représentation n'ont pas la même importance. Certains sont essentiels, d'autres importants, d'autres, enfin, secondaires. [...] étudier une représentation sociale, c'est d'abord, et avant toute chose, chercher les constituants de son noyau central (Abric 2003 : 59-60)

Nous présentons quelques notions importantes pour comprendre le concept d'architecture des RS, à commencer par les propriétés générales des éléments cognitifs formant une RS. Nous verrons qu'il y a dans la littérature un flou théorique autour de la définition de ces éléments. C'est pour cette raison d'ailleurs que nous avons jusqu'ici utilisé le terme générique *d'élément cognitif* pour les désigner. Ensuite, nous aborderons les propriétés des structures des RS. Nous nous attarderons surtout aux propriétés du système central, mais nous soulignerons également quelques points importants à propos du système périphérique.

4.1 Éléments cognitifs

Dans le domaine d'étude des RS, les cognitions formant une RS sont comprises dans un sens très large : des processus et des structures de connaissance « ordinaire » ou de « sens commun ». Ces cognitions sont modélisées de plusieurs manières et font référence à des notions comme celles de schème, de concept, de catégorie, de schémata, de script, de croyance, de grille de lecture, de thème, etc.

Malgré ces imprécisions dans la littérature, nous pouvons identifier quelques propriétés spécifiques aux éléments cognitifs formant une RS, notamment leurs relations avec le langage.

4.1.1 Propriétés des éléments cognitifs d'une représentation sociale

Ce ne sont pas n'importe quels concepts, schèmes, scripts, etc., qui peuvent prétendre former une RS. Pour faire partie d'une RS, ces cognitions doivent remplir des conditions minimales

(Flament et Roquette 2003 : 32) liées à leur saillance, à la division du travail et aux pratiques sociales.

Parmi ces conditions, l'élément cognitif en question doit avoir une forte saillance sociocognitive, c'est-à-dire avoir ce qu'on appelle une « fonction de concept » et être impliqué de manière importante (qualitativement et quantitativement) dans les interactions et les communications. Il s'agit généralement d'un élément cognitif à propos d'un objet, qui incarne un enjeu, un problème ou un conflit d'interprétation parmi les membres d'un groupe (Flament et Rouquette 2003 : 32).

Flament et Rouquette proposent également de voir les problèmes de la pensée sociale comme étant d'un type particulier. Il s'agirait de problèmes auxquels on ne peut pas appliquer de manière satisfaisante un jugement de vérité. Les réponses apportées aux problèmes de la pensée sociale ne sont pas vraies ou fausses, elles sont plus ou moins heureuses dans les conséquences qu'elles impliquent :

> Les problèmes que doit résoudre la pensée sociale [...] appartiennent à la catégorie de ceux que l'on appelle « mal défini » ou « mal structuré ». Il s'agit de tous ceux qui ne satisfont pas au critère de Minsky, que l'on peut ainsi formuler : un problème est bien défini lorsqu'on peut démontrer de toute proposition de solution qu'elle est juste ou fausse. À défaut, on peut seulement estimer ou montrer qu'une solution proposée ou effectivement mise en œuvre est plus ou moins efficace, plus ou moins coûteuse, plus ou moins heureuse, plus ou moins douloureuse, et ainsi de suite. Ces problèmes [...] correspondent à l'immense majorité des situations rencontrées dans la vie quotidienne [...] (Flamente et Rouquette 2003 : 143)

Ceci vient peut-être simplement du fait que les problèmes de la pensée sociale sont fortement imprégnés de considérations axiologiques. Qu'est-ce qu'un groupe idéal d'ami-es? Qu'est-ce que l'intelligence? Qu'est-ce qu'une race? Qu'est-ce que le travail? Qu'est-ce que les droits de l'Homme? Les réponses à ce genre de problèmes, qui ont toutes été l'objet d'étude de RS dans la littérature, font appel aux valeurs du groupe et sont fondées *in fine* sur la force des convictions des membres. [12]

---

[12] Un courant important de la sociologie de la connaissance, incarné à l'origine par Karl Mannheim (Mannheim 1985), puis d'autres comme Berger et Luckmann (2003), Bloor (1976), Gurvitch (1966), etc., pose comme

Pour rependre l'un des exemples de Flament et Rouquette, il serait surprenant que des éléments cognitifs à propos du dentifrice puissent être considérés comme formant une RS, mais c'est certainement le cas des éléments cognitif à propos de l'hygiène.

Ensuite, les éléments cognitifs d'une RS doivent aussi être inscrits dans des rapports sociaux de division du travail, c'est-à-dire à la fois de différenciation et de complémentation sociale entre groupes.

> Il est juste d'ajouter que si nos représentations sont sociales, ce n'est pas seulement à cause de leur objet commun, ou du fait qu'elles sont partagées. Cela tient également à ce qu'elles sont le produit d'une division du travail qui les marque d'une certaine autonomie. (Moscovici 1989 : 100)

Les RS sont le fruit d'une division du travail cognitif. Par exemple, si certains éléments cognitifs d'une RS de l'hygiène font parfois l'objet d'un consensus dans un groupe, ils sont également différenciés à travers les groupes d'une société. Chaque groupe est spécialisé dans la production d'une version spécifique. D'ailleurs, l'étude empirique d'une RS est souvent de nature comparative entre groupes.

Troisièmement, les cognitions formant une RS doivent être liées à des pratiques sociales dans la population étudiée (Abric 1994b). Une pratique sociale est définie de manière générale comme un comportement d'interaction sociale ou une modalité de socialité. Ce sont des cognitions, par exemple, à propos des pratiques religieuses, à propos des pratiques politiques, des pratiques économiques, des pratiques sexuelles, etc. Les pratiques sociales sont des comportements sociaux motivés et dirigés par l'actualisation d'un réseau d'attentes et d'obligations, ou de droits et de devoirs, spécifiques aux cognitions de la RS. C'est durant ces pratiques sociales que les cognitions sont confrontées ou testées. Les pratiques sociales sont donc importantes, elles sont des moments de confrontation des cognitions de la RS avec la réalité.

---

postulat de départ, que ces problèmes ne peuvent être traités via des catégories du vraie ou du faux et que leur étude scientifique ne peut être que de nature comparative. Le domaine d'étude des RS reprend intégralement ce postulat.

En somme, ce n'est pas n'importe quel type de cognition qui compose une RS. D'autre part, nous verrons maintenant que ce ne sont pas tous les éléments cognitifs composant une RS qui sont accessibles à l'étude empirique du chercheur.

4.1.2 Le biais du contexte d'actualisation et d'expression des cognitions d'une RS

Les éléments cognitifs d'une RS ne sont pas énumérables ni dénombrables. C'est-à-dire qu'il est méthodologiquement impossible, pour une RS donnée, de dresser un inventaire exhaustif des éléments cognitifs qui la composent.

En pratique, le nombre d'éléments retenu pour l'analyse varie considérablement d'une étude à l'autre et dépend beaucoup de la méthode utilisée. Ce nombre peut varier à l'intérieur d'une fourchette d'environ 10 à 60 éléments, mais certaines études retiennent jusqu'à 80 éléments (Aïssani *et al.* 1990). Certaines méthodes sont destinées à repérer immédiatement les éléments centraux, par exemple les techniques de mise en cause (Moliner *et al.* 2002 : 134). Dans ce cas, seulement un petit nombre d'éléments est identifié et suffisant. D'autres méthodes ne parviennent à repérer le noyau seulement qu'après avoir identifié les relations qu'il entretient avec les autres éléments de la périphérie. C'est le cas notamment de l'analyse de similitude basée sur des mesures de connexité des éléments (Moliner *et al.* 2002 : 123). Dans ce cas, un nombre plus important est nécessaire.

Les éléments cognitifs d'une RS ne sont pas indexables pour plusieurs raisons. Deux biais méthodologiques sont déterminants : le contexte d'actualisation et la médiation via le discours (Flament et Rouquette 2003 : 21). Pour Abric (Abric 2003 : 61), ces biais sont source d'une « zone muette » de la représentation, c'est-à-dire qu'il y a toujours des éléments non verbalisés (ou non verbalisables). Or, c'est via ces expressions, généralement via le langage, que sont étudiées les RS.

La cognition est un inobservable et le chercheur ne peut l'étudier que par l'intermédiaire de ses actualisations ou expressions chez les sujets étudiés. Nous en avons discuté précédemment avec le concept d'objectivation. L'objectivation est un processus de mise en forme, qui sélectionne et schématise. Le chercheur doit en tenir compte dans son étude, car ces éléments cognitifs actualisés ou exprimés sont souvent les seules traces empiriques à

partir desquelles il peut étudier une RS. Ces actualisations et expressions sont soumises à plusieurs facteurs contextuels et varient d'un sujet à l'autre. Ces facteurs contextuels ne peuvent pas être entièrement contrôlés par le chercheur :

> En effet, toutes RS sont nécessairement actualisées en situation, c'est-à-dire sous l'influence de facteurs contingents, intervenant pour des individus et des groupes particuliers, ici et maintenant. Il faut insister sur ce point : la même représentation, structurellement parlant, va se trouver circonstanciellement mobilisée, évoquée ou convoquée dans des conditions définies, concrètes, singulières mêmes, qui en affecteront plus ou moins profondément l'expression. Ce sont ces modifications d'expression que l'on appellera génériquement « effet de contexte ». (Flament et Rouquette 2003 : 118)

L'effet de contexte dont parlent Flament et Rouquette n'est pas sans rappeler un autre concept important pour la psychologie sociale cognitive, à savoir celui de *l'effet priming*. Le *priming* est un concept qui cherche à rendre compte du processus par lequel sont activés, en mémoire, certains éléments cognitifs, tandis que d'autres seront inhibés, ceci en fonction du contexte social :

> « Priming can be referred to as information activation— rendering it more likely to be used in further processing. [...] One fundamental assumption underlying the influence of priming holds that individuals do not search for all potential categories that could be relevant for the encoding of a given stimulus input. Because of capacity constraints, individuals are not able to check all potentially applicable categories and consequently need to truncate the search process. » (Bless *et al.* 2004 : 38-39)

En outre, le *priming* propose de distinguer entre ressources cognitives disponibles et ressources cognitives accessibles :

> « Although the terms are often used like synonyms, it is useful to distinguish between the availability of information in memory (i.e., whether some representation exists at all) and accessibility (i.e., whether that representation is ready to be retrieved at the moment). » (Bless *et al.* 2004 : 60)

Dans un contexte social donné, un agent cognitif n'a toujours accès qu'à une quantité limitée de ressources cognitives, en l'occurrence, celles activées. Le chercheur est tributaire de ce processus mnémonique.

Par conséquent, le chercheur n'a toujours accès qu'à des traces empiriques limitées et contextuelles d'une RS et déterminées par les processus mnémoniques propres aux sujets.

Ces biais méthodologiques interdisent au chercheur d'affirmer que ces traces empiriques sont l'expression de l'ensemble des cognitions impliquées dans la RS. [13]

4.1.3 L'étude des représentations sociales via la médiation du langage

La très grande majorité des études empiriques de RS passent par l'analyse du discours, c'est-à-dire l'analyse d'un matériel langagier en général, qu'il soit sous la forme d'une analyse de réponses à un questionnaire, d'une analyse d'association libre, d'une analyse d'entrevue, d'une analyse de corpus de presse, ou autre. C'est à partir de ces matériaux que les chercheurs repèrent les éléments cognitifs de la RS :

> [...] dans la plupart des cas, ce sont des productions discursives qui permettent d'accéder aux représentations. (Abric 1994b : 15)
>
> Les études sur les RS portent sur du matériel langagier (réponses à des questionnaires, associations libres, entretiens, etc.) (Doise *et al.* 1992 : 25)
>
> Les éléments d'une RS sont le plus souvent indexés par des termes ou des groupes de termes que l'on extrait d'entretiens, de réponses à des questionnaires, des discussions de groupe, de déclarations écrites ou encore, éventuellement, d'un corpus de presse. On a donc affaire, opérationnellement en tout cas, à des mots. (Flament et Rouquette 2003 : 57)

Or, entre langage et cognition, il n'y a pas de relation d'équivalence. Les éléments linguistiques du langage ne correspondent pas nécessairement aux éléments cognitifs d'une RS. On ne peut pas affirmer, par exemple, que l'inventaire du lexique est isomorphe à l'inventaire des éléments cognitifs d'une RS (Flament et Rouquette 2003 : 81). Néanmoins, le langage est utilisé comme média, comme un indicateur ou un « truchement » par lequel on accède aux cognitions formant une RS :

> [...] il s'agit d'étudier, par le biais d'un truchement technique, l'organisation même de la connaissance, dont les mots fournissent seulement des indicateurs. Quelle que soit son importance, le moyen d'accès n'est pas le but, tout comme la route n'est pas la destination. Autrement dit, on ne s'intéresse pas à l'organisation du lexique en tant que

[13] Par ailleurs, cette difficulté méthodologique ne semble pas vraiment gêner ces chercheurs, étant donné que leur approche est avant tout descriptive et comparative (Flament et Rouquette 2003).

> tel, mais à l'organisation du savoir commun. Le langage n'est pas l'objet. (Flament et Rouquette 2003 : 58)
>
> On doit donc répudier toute idée de recensement ou de comptage pour caractériser l'ensemble d'éléments cognitifs constituant une représentation à un moment donné. D'autres propriétés, d'ailleurs plus intéressantes [...] peuvent être prises en compte. (Flament et Rouquette 2003 : 23).

Ce qui ne signifie évidemment pas que le langage n'intervient pas dans la construction des RS. Nous avons vu avec Durkheim que le langage, en tant que système de signes, stabilise ou cristallise les cognitions dans le groupe. Mais, poursuit Durkheim, le langage : « a une nature qui lui est propre, et, par suite, des lois qui ne sont pas celles de la pensée. Puisque donc il contribue à l'élaborer, il ne peut manquer de lui faire violence en quelque mesure et de la déformer. » (Durkheim 1912)

Langage et cognition, bien qu'intriqués par une relation jugée centrale, ne sont pas équivalents.[14] Ceci pose des problèmes importants pour l'étude des RS, notamment, celui du choix des éléments linguistiques dans lesquels sont susceptible d'être exprimés les éléments cognitifs d'une RS. Ces éléments linguistiques sont parfois le mot, la phrase, etc. Dans le cas d'une étude à l'aide de méthodes associatives, on considère généralement les lexèmes induits comme l'expression des éléments de la RS (Flament et Rouquette 2003; De Rosa 2003). Dans le cas d'une analyse de contenu classique, on considère généralement les thèmes comme l'expression des cognitions (Moliner *et al.* 2002). Dans le cas d'une analyse par forage de texte avec le logiciel ALCESTE, ce sont les « mondes lexicaux », c'est-à-dire des isotopies de lexèmes cooccurrents, qui sont retenus comme indicateurs ou expressions des éléments cognitifs d'une RS (Kalampalikis 2003). Il s'agit sans aucun doute d'un problème de fond pour la théorie des RS, non résolu de manière satisfaisante jusqu'à maintenant.

Ceci appelle également une dernière remarque importante sur les éléments cognitifs de la RS et leur relation avec le langage. Puisque les cognitions étudiées sont essentiellement celles exprimées dans le langage, ou du moins dans une forme sémiotique, on considère que les

---

[14] Pour être plus précis, on juge majeure sinon indispensable l'interaction entre cognitions et des formes sémiotiques générales (Meunier 2006), dont le langage naturel est certes l'exemple canonique, mais il en existe d'autres : l'iconographie, la gestuelle, etc.

éléments cognitifs d'une RS sont essentiellement de la connaissance déclarative. Rouquette affirmera que l'étude empirique des RS : « [...] vise que la connaissance réfléchie, c'est-à-dire celle qui passe par un traitement symbolique [...] on se limite enfin aux seules connaissances déclaratives. » (Rouquette 1994: 154)

Par *connaissance déclarative*, il faut se rapporter à sa définition technique pour la psychologie cognitive. La connaissance déclarative fait référence à la mémoire sémantique. On y retrouve essentiellement les connaissances propositionnelles, encyclopédiques et épisodiques (Fortin et Rousseau 2005 : 367-369). On oppose souvent ces connaissances aux connaissances procédurales ou tacites, qui sont considérées comme plus difficiles à exprimer dans le langage.

Ces remarques, à la fois sur le contexte et le langage, soulignent de manière importante les limites du territoire cognitif pris en compte par la théorie des RS.

4.2 Structures : les modalités de relations dans les structures de la représentation sociale

Une RS n'est pas seulement un amoncellement de cognitions, c'est également des structures cognitives et sociales spécifiques. C'est-à-dire qu'une RS ne peut pas être seulement définie par son contenu, elle le sera également par ses relations : « Une RS formant, par hypothèse définitoire, un système, les divers éléments qui la composent sont reliés entre eux. » (Flament et Rouquette 2003 : 85). Autrement dit, une RS est caractérisée par des structures dans lesquelles les éléments cognitifs sont hiérarchisés. Les éléments cognitifs d'une RS ne sont pas tous structurellement équivalents : certains éléments peuvent être centraux, c'est-à-dire déterminants, alors que ces mêmes éléments peuvent être secondaires et périphériques pour une autre RS. Abric affirme que deux RS au contenu identique, c'est-à-dire composées des mêmes éléments cognitifs, peuvent correspondre à deux RS différentes, si celles-ci se différencient au niveau de leurs structures (Abric 2003b : 60).

Les structures d'une RS stable (cristallisée, instituée) sont à la fois cognitives et sociales. Les structures cognitives sont caractérisées par un système central, aussi appelé un noyau, et par un système périphérique. Les structures sociales sont d'un côté des formes de partage social consensuel d'éléments cognitifs parmi les membres et de l'autre des positions sociales

auxquelles sont associées des prises de position, c'est-à-dire des structures sociales de différenciation par rapport aux autres membres du groupe. Selon la théorie des RS, ce sont les éléments composant le noyau d'une RS, qui sont également l'objet d'un consensus, alors que les éléments périphériques sont plutôt l'objet de prise de position.[15]

Les relations cognitives entre éléments d'une représentation sont particulièrement importantes pour la théorie des RS. Elles ont été l'objet de deux types de modélisation, auquel correspondent deux types de méthodes pour les identifier : l'analyse des schèmes cognitifs de base et l'analyse de similitude.

4.2.1. Les schèmes cognitifs de base

Le modèle des schèmes cognitifs de base propose qu'il soit possible d'identifier parmi les relations entre éléments cognitifs d'une RS, différentes modalités :

> Le modèle suggère qu'entre un élément de connaissance A (un mot ou une expression qui opérationnalise la notion de « cognème ») [...] et un élément de connaissance B (un autre « cognème »), il existe une relation R qui peut prendre plusieurs états (on peut dire aussi : plusieurs modalités). (Guimelli 2003 : 120)

Ces modalités de relation correspondent à 28 connecteurs, qui sont eux-mêmes regroupés en cinq schèmes cognitifs de base : de lexique (3 connecteurs), de voisinage (3 connecteurs), de composition (3 connecteurs), de praxie (12 connecteurs) et d'attribution (7 connecteurs). Les trois premiers schèmes cognitifs de base sont décrits comme des fonctions de description entre deux éléments. Le schème appelé « praxie » est décrit comme une fonction de prescription liée à des pratiques et le schème appelé « attribution » est décrit comme une fonction évaluative liée à du jugement (voir le tableau I, inspiré de Rouquette (1994) et Abric (2003c : 389).

---

[15] Ceci laisse d'ailleurs entendre, d'une part, que la recherche de consonance sociocognitive est beaucoup plus importante pour les cognitions centrales et d'autre part, que l'ancrage est davantage possible sur les cognitions périphériques de la RS.

Tableau 1: Les schèmes cognitifs de base comme modalité de relation entre éléments cognitifs d'une représentation sociale

Méta-schème	*Schème Cognitif de Base*	*Connecteurs et expression des relations*
Descriptif	Lexique :	A signifie la même chose que B
		A peut être défini comme B
		A est le contraire de B
	Voisinage :	A fait partie, est inclus, est un exemple de B
		A a pour exemple, comprend, inclut B
		A appartient à la même catégorie générale de B
	Composition :	A est une composante de B
		A a pour composante B
		A et B, sont tous les deux les constituants de la même chose, du même objet
Praxéologique, lié à la prescription	Praxie :	A fait B
		A a une action sur B
		A utilise B
		C'est B qui fait A
		A est une action qui a pour objet, s'applique à B
		Pour faire A on utilise B
		B est quelqu'un qui agit sur A
		B désigne une action que l'on peut faire sur A
		B est un outil que l'on utilise sur A
		A est utilisé par B
		On utilise A pour faire B
		A est un outil que l'on peut utiliser pour B
Évaluation, lié au jugement	Attribution :	A est toujours caractérisé par B
		A est souvent caractérisé par B
		A est parfois caractérisé par B
		A doit avoir la qualité de B
		B évalue A
		A a pour effet B
		A a pour cause B

4.2.2 Les relations de similitude

Un deuxième modèle est celui de l'analyse de similitude. Dans ce modèle, on ne cherche pas à décrire dans le détail les différentes modalités de relation entre éléments cognitifs d'une RS. Pour des raisons souvent de nature méthodologique, il est parfois impossible de procéder à une analyse fine des schèmes cognitifs de base. On cherche plutôt dans ces cas à décrire la forme générale de l'organisation cognitive des éléments de la RS :

> Autant il peut être utile de qualifier avec précision les relations entre certains éléments de la RS [...] autant il peut être utile d'expliciter l'essentiel de la structure d'association qui existe sur l'ensemble des éléments. On ne sait pas faire les deux à la fois, et il se révèle commode à l'usage de ne pas toujours chercher à discerner la qualité spécifique de chaque association [...] et de ne considérer que la notion de *similitude* entre éléments, en la quantifiant sur une population donnée. (Flament et Rouquette 2003 : 85-86)

L'analyse de similitude ramène toutes les différentes modalités relationnelles à une seule : nommée une relation de similitude. Une relation de similitude entre deux éléments d'une RS décrit le fait que — pour différentes raisons qui ne sont pas retenues dans l'analyse — les individus d'un groupe ont, de manière générale, tendance à les associer, à considérer qu'ils vont ensemble ou qu'ils sont proches l'un de l'autre.

De plus, les relations de similitude sont considérées comme étant à la fois symétriques et non transitives. Si l'élément A est considéré similaire à l'élément B, B est aussi considéré similaire à A. Si A est considéré similaire à B et que B est considéré similaire à C, les éléments A et C ne seront pas nécessairement considérés similaires.

Lorsque la structure de la RS est modélisée par des relations de similitude entre éléments cognitifs, on la représente souvent sous la forme d'un réseau, c'est-à-dire un graphe, où les nœuds sont les éléments cognitifs de la RS et les liens des relations de similitude moyenne entre chaque pair de cognition. La figure 1, inspirée partiellement de Flament et Rouquette (Flament et Rouquette 2003 : 88), montre un graphe complet, valué et non orienté :

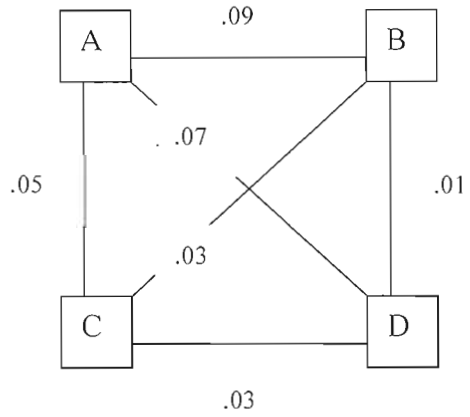


Figure 1 : Exemple de graphe complet valué et non orienté

La figure 1 est simple, mais typique du genre de graphe utilisé pour représenter la structure cognitive de la RS en fonction de ses relations de similitude. Dans cet exemple, sont affichés quatre éléments fictifs de l'économie cognitive d'un groupe social également fictif. Ces quatre éléments, soit A, B, C et D, sont tous reliés par une relation de similitude caractérisée par les membres du groupe fictif, qui varie entre 0.1 et 0.9.

Les mesures de similarité utilisées sont très nombreuses : corrélation, distance euclidienne, indice D, etc. (Moliner *et al.* 2002 : 149; Flament et Rouquette 2003 : 87; Bouriche 2003). Les valeurs de similitude entre chaque pair de nœuds du graphe traduisent une proximité, où, plus un lien possède une valeur élevée, plus les nœuds qu'il lie sont similaires. Dans l'exemple de la figure 1, les éléments A et B sont considérés comme plus similaires que B et D.

La représentation en réseau de la structure de la RS est particulièrement intéressante, car elle permet ensuite de considérer les propriétés mathématiques du graphe (Wasserman et Faust 1994) comme des indicateurs des propriétés cognitives de la structure de la RS. L'une des premières propriétés mathématiques révélées est celle dite de « l'arbre maximum », qui consiste à n'extraire du graphe complet que les relations les plus fortes, ou ce que Flament et Rouquette appellent le « squelette » du réseau :

> L'analyse de similitude veut ne conserver qu'un « squelette » de cet ensemble, mais qui est le plus signifiant possible. (Flament et Rouquette 2003 : 88)

> [...] c'est aussi une manière de dire qu'on a retenu le plus de similitudes de l'ensemble avec un graphe aussi simple que possible [...] en ne retenant que ce qui, à un certain point de vue théorique, est essentiel et signifiant. (Flament et Rouquette 2003 : 89)

Dans la figure 2 suivante est extrait l'arbre maximum du graphe complet de la figure 1 :

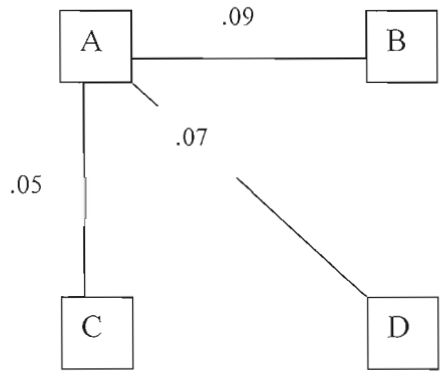


Figure 2 : Exemple d'un arbre maximum

Un arbre maximum est un sous-graphe connexe sans cycle, extrait d'un graphe complet valué, et dont la somme des valeurs des liens est plus élevée que tout autre arbre connexe sans cycle possible dans le graphe. Un graphe connexe signifie que tous les nœuds du graphe sont liés à au moins à un autre nœud. Alors que « sans cycle » signifie l'absence de boucle dans le graphe. On dira également que l'arbre maximum est le sous-graphe connexe sans cycle le plus « lourd », qui cherche à traduire le pattern relationnel le plus important du réseau, en l'occurrence le pattern relationnel le plus important de la structure cognitive de la RS.

L'interprétation des propriétés mathématiques comme indicateur sur les propriétés cognitives de la RS doit cependant demeurer prudente. Car, si les affinités entre structure de la RS et structure mathématique du graphe sont heuristiques, il n'y a pas entre les deux d'isomorphisme (Flament 1996b : 145).

> Un graphe est une structure mathématique; une RS a aussi une structure (disons sociocognitive), il est tentant de rapprocher ces deux structures et il peut y avoir du

> sens à le faire. Mais une identification mécanique de l'une à l'autre peut être source de graves erreurs d'interprétation : toutes propriétés structurales mathématiques ne sont pas nécessairement révélatrices d'une propriété structurale des réponses humaines. (Flament et Rouquette 2003 : 90)

> Les liens géométriques et statistiques ne doivent pas être confondus avec les liens sémantiques et psychologiques [...] (Doise al. 1992 : 95)

La cognition n'a pas une structure en graphe ni la forme d'un arbre maximum. Les modèles mathématiques utilisés sont des métriques heuristiques permettant d'illustrer des inobservables. Par exemple : « on ne découvre pas que la structure cognitive est euclidienne, mais on découvre la nature de cette structure à travers le filtre de la métrique euclidienne. » (Doise *et al.* 1992 : 95)

L'identification des relations entre éléments cognitifs d'une RS est essentielle, car certaines propriétés des cognitions formant le noyau central d'une RS, sont définies sur la base de leur forte connexité, c'est-à-dire la propriété d'entretenir beaucoup de relations avec les autres éléments cognitifs dans la structure.

4.3 Système central

Les cognitions du système central de la RS sont ce qui lui confère son identité : deux RS avec des noyaux différents sont considérées comme différentes. Ce noyau contient très peu d'éléments, généralement deux ou trois, rarement plus de six (Flament et Rouquette 2003 : 23).

De manière générale, on résume le système central de la RS de la manière suivante :

> C'est lui qui détermine la signification de la représentation (fonction génératrice); son organisation interne (fonction organisatrice); sa stabilité (fonction stabilisatrice). (Flament et Rouquette 2003 : 24)

Flament et Rouquette accordent trois principales fonctions au noyau : une fonction génératrice de signification, car la signification des éléments périphériques est dépendante de celle des éléments du noyau; une fonction d'organisation cognitive des éléments de la RS, c'est-à-dire que c'est en fonction du noyau que sont organisés les éléments de la RS; et une fonction stabilisatrice de la structure, puisque c'est le noyau qui résiste au changement.

Une manière simple de définir les éléments cognitifs composant le noyau d'une RS, poursuivent Flament et Rouquette, est : « de dire qu'ils servent de critères décisoires dans la catégorisation des spécimens rencontrés : leur présence est nécessaire pour que le spécimen soit attribué à sa classe représentationnelle d'appartenance, et leur absence suffisante pour que le spécimen soit exclu de cette classe. » (Flament et Rouquette 2003 : 23-24).

Un exemple classique pour illustrer ceci est l'étude de la RS du *groupe idéal d'amis*. Parmi les éléments importants constituant cette RS, les études recensent les éléments *égalité ou absence de hiérarchie entre individus* et *le partage d'opinion similaire parmi les individus*. Cependant, les recherches montrent aussi qu'il arrive parfois que les agents sociaux considèrent un groupe *sans partage d'opinion similaire* étant malgré tout une instance de la classe *groupe idéal d'amis*, alors qu'un groupe dans lequel il y a de la hiérarchie est invariablement rejeté comme instance de la classe. Ainsi, l'élément *égalité ou absence de hiérarchie* est considéré comme un élément central de la RS, non négociable, alors que l'élément *partage d'opinion* est périphérique.

Il y a d'autres propriétés pour caractériser les éléments cognitifs formant le noyau d'une RS.

4.3.1 Trois propriétés des cognitions centrales

Différents travaux (Moliner 1994 : 206; Moliner et Martos 2005; Bataille 2002) ont proposé quatre propriétés typiques pour caractériser les éléments cognitifs du système central : leur saillance, leur valeur symbolique, leur connexité dans la structure cognitive et leur consensus social.

La saillance est cette propriété évidente et importante qu'ont les cognitions centrales d'être actualisées plus souvent que les autres dans les communications. Ce sont des communications à propos de controverse, d'enjeux d'interprétation, de légitimation, etc. Empiriquement, les indicateurs de la saillance se traduisent généralement par la répétition de mots ou d'expressions, susceptibles d'être des traces langagières des cognitions du système central. Les indicateurs de la saillance sont souvent utilisés pour repérer dans les discours les traces linguistiques — essentiellement lexicales — susceptibles d'exprimer des éléments centraux d'une RS.

Les trois autres propriétés sont plus intéressantes, car on peut les lier aux trois processus de la pensée sociale que sont la recherche de consonance sociocognitive, l'ancrage et l'objectivation.

4.3.1.1 La valeur symbolique et l'objectivation

La valeur symbolique exprime cette propriété du système central selon laquelle il y a des cognitions composant une RS qui sont non négociables, c'est-à-dire que l'on ne peut remettre en cause sans remettre en cause la RS elle-même.

Récemment, Moliner et Martos (Moliner et Martos 2005), à la suite de Bataille (Bataille 2002) et revisitant les travaux classiques de Moscovici (1976), proposent également de voir la valeur symbolique d'un élément cognitif par ce qu'ils appellent son « potentiel sémantique ». Le potentiel sémantique renvoie à cette propriété qu'ont les éléments centraux d'une RS, d'être à la fois polysémique et vide de sens. Les éléments centraux sont polysémiques, car les individus les interprètent de plusieurs manières selon les contextes d'actualisation, comparativement aux éléments périphériques, qui sont plus univoques et précis. Ils sont aussi vides de sens, car les individus ont toujours de la difficulté à les définir. Empiriquement, Kalampalikis et Moscovici (2005) ont montré que ces cognitions s'expriment dans un langage pauvre, ce qu'ils appellent une faible « température informationnelle », c'est-à-dire un lexique redondant, peu diversifié et imprécis : « Elle produit du consensus et se manifeste par la redondance lexicale qu'elle engendre. Redondance qui se traduit par l'usage de répertoires lexicaux communs, des formules admises et socialement actives. » (Kalampalikis et Moscovici 2005 : 22)

Ces cognitions sont actualisées dans des communications, des contextes et des situations d'interactions très différentes. Alors qu'elles semblent constamment entourées d'un flou sémantique susceptible de créer de la confusion dans les communications, leur interprétation au contraire, ne pose que rarement problème pour les individus.

On voit pourquoi on peut lier de manière privilégiée cette propriété du système central avec le processus d'objectivation. L'objectivation consiste précisément à simplifier, schématiser, abstraire, décontextualiser, cristalliser, un élément cognitif. Vidé de leur précision

sémantique, ceci augmente leur pouvoir symbolique (Moliner et Martos 2005 : 90). Redondant, le format des expressions langagières de ces éléments cognitifs facilite leur communication (Kalampalikis et Moscovici 2005 : 22).

4.3.1.2 La connexité et l'ancrage

La connexité est la propriété des cognitions centrales d'être fortement impliquées ou intégrées dans la structure cognitive de la RS, c'est-à-dire en entretenant un grand nombre de relations significatives avec les autres cognitions de la RS.

La connexité a longtemps été considérée comme « la voie privilégiée » (Moliner *et al.* 2002 : 122) pour le repérage de la centralité des éléments cognitifs dans la structure de la RS. Cette propriété propose de voir la structure cognitive de la RS sous la forme d'un réseau, tel que nous l'avons vu précédemment (figures 1 et 2).[16] La connexité des cognitions est mesurée à l'aide de mesures de centralité des nœuds dans un graphe modélisant la structure de la RS. Ces mesures de centralité sont très nombreuses (Freeman 1979). Parmi les plus utilisées, on retrouve les calculs de centralité de degré, de proximité ou d'intermédiarité, le filtrant des cliques. Ces différents algorithmes ont été développés dans différents programmes de recherche, certains issus de l'étude des RS (Flament 1981; Bouriche 2003), d'autres de l'étude des réseaux sociaux (Wasserman et Faust 1994) ou l'étude des réseaux sémantiques (Popping 2000). Ces programmes de recherche ont en commun de travailler sur la théorie mathématique des graphes.

Si nous revenons rapidement à l'exemple du graphe de la figure 1, on peut par exemple calculer sur ce graphe la centralité de degré des nœuds. La centralité de degré d'un nœud mesure la proportion de relations qu'un nœud entretient avec les autres nœuds du graphe. Dans sa version la plus simple, elle se calcul de la manière suivante : la somme des liens

---

[16] Cette idée de voir la structure des cognitions comme un réseau a pour origine contemporaine les travaux de Mynsky en intelligence artificielle. Cette idée fut reprise par les sciences sociales et humaines, notamment par l'anthropologie cognitive (Rice 1980; Casson 1983; D'Andrade 1995: 138). La métaphore du réseau a également servi pour modéliser les RS. Le modèle du réseau a été formalisé par l'une des méthodes d'analyse les plus populaires dans le champ des RS, à savoir l'analyse de similitude.

qu'un nœud *i* entretient avec les autres nœuds *j* du graphe, divisé par le nombre de relations possibles dans le graphe (soit $g - 1$, où g correspond au nombre de nœuds dans le graphe) :

$$\text{Centralité de degré} = \frac{\sum_i x_{ij}}{(g-1)}$$

Appliqué à un graphe valué, $\sum_i x_{ij}$ est égal à la somme de la valeur des liens entre *i* et *j*. Dans notre exemple, la centralité de degré des nœuds A, B, C et D est respectivement de 0,7, 0,43, 0,37, 0,37. L'élément cognitif représenté par le nœud A est dans ce cas considéré comme plus connexe, parce que plus central dans la structure.

À nouveau, il faut être prudent dans l'interprétation des propriétés de centralité cognitive via des propriétés de centralité mathématique. La centralité graphique n'est jamais le simple reflet ni le miroir d'une centralité cognitive :

> Il est [...] souvent considéré que la propriété mathématique de centralité dans un graphe peut conduire à considérer les éléments centraux comme des éléments du noyau central d'une représentation sociale. Cela n'est pas aussi évident. (Bouriche 2003 : 250)
>
> Il ne faut pas utiliser les propriétés mathématiques du graphe sans retour critique. Les critères statistiques de centralité, qu'ils soient calculés sur l'arbre maximum ou sur la totalité des relations, ne donnent pas directement le noyau central d'une représentation sociale. (Bouriche 2003 : 251)

Ces différentes propriétés de centralité mathématique sont seulement des indicateurs sur la connexité des éléments cognitifs dans la structure d'une RS.

La connexité d'un élément cognitif, comprise en termes de proximité ou de similarité avec les autres éléments de la RS, peut être considérée comme le résultat d'un processus d'intégration cognitive par les membres du groupe étudié. Plus une idée, un concept, un schème, est en moyenne, jugée par les membres d'un groupe comme proche, similaire ou en harmonie avec tous les autres éléments d'une RS, plus on peut considérer cet élément comme ayant été l'objet d'un travail d'appropriation. C'est en ce sens que l'on peut considérer la forte connexité du noyau central d'une RS comme un effet du processus d'ancrage. Effectivement, l'ancrage est le processus par lequel les agents sociaux s'approprient les cognitions

véhiculées par leurs groupes d'appartenance en les intégrant à leur version idiosyncrasique de la RS.

4.3.1.3 Le consensus social et la recherche de consonance sociocognitive

Les propriétés des éléments centraux d'une RS vues jusqu'à maintenant sont des propriétés essentiellement cognitives (connexité) et communicationnelles (valeur symbolique), qui concernent, dans leur définition traditionnelle, des propriétés attribuables à un individu. Toutefois, une propriété supplémentaire est indispensable pour caractériser les cognitions du système central. Cette propriété est de type social ou écologique, c'est-à-dire qu'elle concerne la distribution sociale des cognitions d'une RS parmi les membres d'une population. Les cognitions du système central sont socialement partagées dans le groupe, selon un consensus important. Ce partage social est important pour la construction des RS, car il permet un type particulier et élémentaire de lien social entre individus que Durkheim a souvent souligné :

> [...] il existe une solidarité sociale qui vient de ce qu'un certain nombre d'états de conscience sont communs à tous les membres de la même société. (Durkheim 1893)
>
> Tout le monde sait, en effet, qu'il y a une cohésion sociale dont la cause est dans une certaine conformité de toutes les consciences particulières à un type commun qui n'est autre que le type psychique de la société. (Durkheim 1893)

Le partage de certaines cognitions communes est une condition de possibilité de l'identité du groupe, la cohésion, la socialité et la communication. Guimelli le résume ainsi : « le système central constitue la base commune, collectivement partagée, des représentations sociales. » (Guimelli 1999 : 81)

D'autres recherches (Flament 1996; Flament 1999; Gaymard 2002; Flament et Rouquette 2003) ont montré que les structures de partage social des cognitions centrales d'une RS sont plus complexes qu'elles ne semblaient le supposer dans le modèle durkheimien classique. Ces travaux ne remettent pas en cause la nécessité d'un consensus dans un groupe sur les cognitions centrales de la RS, mais proposent une nouvelle définition de ce consensus. Cette nouvelle définition s'éloigne de celle de Durkheim pour se rapprocher davantage d'une définition en termes d'équilibre.

Pour Durkheim en effet, ce consensus est pensé en termes de « substrat » homogène et commun à l'ensemble du groupe, voire de la société. Les cognitions centrales sont celles communes à l'ensemble des membres. La nouvelle définition qui émerge en particulier dans les travaux de Flament, remets en question la nécessité de ce substrat, même de la possibilité qu'il y en ait un.

On propose plutôt de voir l'adhésion des membres aux cognitions centrales d'une RS comme un processus plus élastique, où il y a place à de la négociation. Cette proposition de voir le consensus en termes d'équilibre est étroitement liée à celle de voir le partage social comme étant — en partie — l'aboutissement d'un processus de recherche de consonance sociocognitive entre membres d'un groupe. Elle propose de voir l'individu comme un *tacticien motivé* (Bless *et al.* 2004 : 5), doté de différentes stratégies de négociation de son rapport à l'ordre social.[17]

Par exemple, il se peut que pour une RS donnée, deux éléments A et B soient très centraux, sans pour autant qu'ils soient partagés par la totalité des membres du groupe. Les membres du groupe peuvent dans une certaine mesure, sélectionner lequel des éléments, le A ou le B, ils retiendront dans leur version idiosyncrasique de la RS (par exemple parce qu'il est plus consonant avec leur position sociale). Cette négociation reste cependant confinée dans un horizon déterminé. Si les membres peuvent négocier leur adhésion, ils ne peuvent pas rejeter en bloc à la fois l'élément A et B, ils doivent au moins intégrer fortement l'un ou l'autre à leur RS.

En somme, chaque propriété des cognitions centrales de la RS est étudiée via des méthodes d'analyse qui leur sont spécifiques. On s'attend également à ce que, pour une RS stable ou cristallisée, ces différentes propriétés convergent vers l'identification d'une ou d'un même petit ensemble de cognitions centrales. Cette convergence reste cependant difficile à étudier étant donné que la plupart des recherches se concentrent sur l'analyse d'une seule propriété parmi plusieurs.

---

[17] Par ailleurs, cette conception de l'individu ou de l'acteur est très similaire à celle des approches ethnométhodologiques et goffmaniennes.

4.4 Système périphérique

L'étude du système central a été l'objet de la majorité des recherches sur les RS. Mais on découvre, notamment depuis Flament (Flament 1989; 1994a), que le système périphérique est composé également d'éléments cognitifs importants pour la pensée sociale. Peut-être l'une des raisons pour lesquelles le système périphérique a longtemps été considéré secondaire est qu'on l'a défini d'abord et avant tout par sa relation de dépendance envers le noyau central de la représentation (Flament 1989 : 229).

Cette relation de dépendance de la périphérie envers le système central est ce qui caractérise en premier lieu les cognitions du système périphérique. Mais cela l'a parfois réduite à être considérée par les chercheurs et chercheuses comme épiphénomène du noyau.

Cependant, les cognitions composant la périphérie ont des propriétés qui lui sont spécifiques : ce sont des cognitions liées à l'expérience personnelle, aux pratiques sociales et à l'ancrage d'élément nouveau. Ce sont aussi des cognitions liées à l'idiosyncrasie biographique des agents cognitifs. Le tableau 2, repris de Guimelli (Guimelli 1999 : 86), résume succinctement quelques propriétés des éléments cognitifs de la périphérie en comparaison de celles du noyau.

Tableau 2 : Caractéristiques distinctives entre éléments du système central et du système périphérique de la représentation sociale

Système central	Système périphérique
▪ Mémoire collective et identité du groupe (enjeu inter-groupe)	▪ Mémoire biographique (enjeu interindividu) et expérience individuelle
▪ Commun au groupe (partagé socialement de manière consensuelle)	▪ Distribué socialement de manière inégale (différencié dans le groupe)
▪ Fonction de concept (détermine le sens de la RS)	▪ Fonction normative pour les pratiques sociales
▪ Fonction organisatrice (cohérent, rigide et stable)	▪ Fonction d'intégration (supporte des anomalies et des contradictions temporaires)
▪ Résistance au changement et insensibilité aux contextes	▪ Dynamique et varient en fonction des contextes d'actualisation

Les cognitions périphériques sont, contrairement au noyau, beaucoup moins stables. Elles sont modifiées par les individus en fonction de leurs expériences de la réalité sociale, de leur

position sociale, des contextes et des communications. Flament (Flament 1989) suggère d'imaginer la périphérie comme une zone tampon entre le noyau et la réalité sociale. Si la périphérie est modifiée plus souvent que le système central, c'est que ce sont les cognitions de ce système qui, dans un premier temps, seront transformées si les agents doivent s'adapter à une nouvelle réalité sociale :

> En fait, la périphérie de la représentation sert de zone tampon entre une réalité qui la met en cause, et un noyau central qui ne doit pas changer facilement. Les désaccords de la réalité sont absorbés par les schèmes périphériques, qui, ainsi, assurent la stabilité (relative) de la représentation. (Flament 1989 : 230)

De plus, le système périphérique permet d'intégrer dans la RS des cognitions idiosyncrasiques liées à la biographie et aux expériences de chaque individu. La périphérie permet une telle intégration sans remettre en cause (la plupart du temps) l'organisation des cognitions centrales.

> Le système périphérique permet alors l'intégration dans la représentation du sujet de ces variations individuelles directement liées à ses expériences personnelles, à son degré d'implication par rapport à l'objet, bref, à la situation dans laquelle il se trouve personnellement. Le système périphérique autorise ainsi la construction de représentations sociales individualisées, organisées néanmoins autour d'un noyau central commun. (Guimelli 1999 : 85-86)

C'est via ce système que chaque individu s'approprie la RS. En effet, si on peut définir le système central comme étant un ensemble organisé de cognition socialement partagée selon un consensus, le système périphérique lui, est caractérisé par une plus grande variation interindividuelle. On comprend par conséquent que la périphérie est une structure beaucoup plus dynamique et c'est la première à être modifiée par les processus d'ancrage expliqués précédemment. En tant que zone tampon, la périphérie permet d'accumuler les particularités individuelles jusqu'à une sorte de seuil de tolérance. La périphérie permet alors des transformations graduelles, qui sauvegardent à la fois l'organisation cognitive chez l'individu et le tissu social du groupe.

Retour sur le premier objectif

Ce chapitre était consacré à notre premier objectif de recherche, qui consistait à présenter le domaine d'étude des RS d'un point de vue théorique. Les principaux concepts ont été

présentés et surtout, le rôle que cette théorie accorde au langage a été traité, ce qui sera indispensable pour l'évaluation du forage de texte plus loin.

Les trois premières parties présentaient trois concepts importants pour décrire certains processus de la pensée sociale, qui permettent de comprendre la construction des RS dans les interactions et les communications sociales. Ces trois concepts sont la consonance sociocognitive, l'ancrage et l'objectivation.

Le concept de consonance sociocognitive décrit un processus qui a lieu dans les interactions et les communications sociales, où les membres d'un groupe recherchent l'équilibre entre eux. C'est-à-dire que des individus qui partagent un lien social chercheront également à partager des cognitions similaires. Ce concept propose de voir la sauvegarde du lien social généralement plus coercitive que le maintien d'une cognition par un agent. La recherche de consonance sociocognitive est un processus à la base de la diffusion consensuelle et de la socialisation des cognitions dans un groupe.

Le concept d'ancrage décrit un processus de différenciation des cognitions distribuées dans un groupe en fonction de leur position sociale dans la société. C'est-à-dire qu'il y a toujours des cognitions véhiculées par les groupes, auxquels sont membres les individus, qui ne sont pas compatibles ou consonantes entre elles. Chaque individu doit pourtant intégrer ces dissonances. Il y parvient, mais en modifiant, de manière minime ou importante, les cognitions propres à chaque groupe. L'ancrage est un processus à la base de l'évolution et des changements dans l'économie des cognitions d'un groupe.

Le concept d'objectivation décrit un processus de mise en forme, de sélection et de schématisation, et généralement via le langage, des cognitions durant la communication. L'objectivation est la dimension publique de la RS. L'objectivation est également un processus à la base de la stabilisation et de cristallisation des éléments cognitifs dans le groupe.

La quatrième partie a présenté le concept d'architecture de la RS. Cette partie était consacrée à la théorie structurale du noyau central des RS. Les éléments composant une RS sont repérés et étudiés via des indicateurs linguistiques. Méthodologiquement, il est impossible de faire

l'inventaire de tous les éléments cognitifs susceptibles de composer une RS. Par ailleurs, certains éléments d'une RS sont structurellement plus importants que d'autres, c'est-à-dire qu'ils sont plus centraux dans les structures caractérisant une RS et l'objectif premier d'une recherche empirique sur les RS est précisément le repérage de ce noyau.

Une représentation sociale est une modélisation

Nous terminons ce chapitre sur un dernier point important à souligner et souvent négligé dans la littérature. La théorie des RS est une modélisation descriptive et non ontologique de l'organisation des éléments cognitifs dans un groupe ou une population déterminée. Ce point est particulièrement important à souligner lorsque vient le temps de comprendre les structures cognitives de la RS. Lorsqu'on analyse une RS donnée, par exemple celle à propos du *travail,* en réalité, ce qu'on cherche à réaliser c'est une modélisation statistique de ce qu'est, en moyenne, cette RS chez les membres du groupe.

Chaque membre du groupe possède une RS du travail qui lui est idiosyncrasique, et dans un groupe, il y a autant de *versions* de la RS du *travail* qu'il y a d'individus. Ce sera, par exemple, l'objet de recherche d'un programme comme « l'épidémiologie des représentations » d'étudier la distribution sociale des versions d'une même représentation dans une population (Sperber 1996). Le programme de recherche de la théorie des RS est différent. La modélisation qui résulte de l'analyse du ou de la psychosociologue ne correspond à aucune RS du *travail* en particulier parmi les membres de la population étudiée. Il s'agit plutôt d'une abstraction de ces différentes versions. Cette abstraction sera faite parfois sous le mode d'un idéaltype, d'autres fois par des calculs de moyenne, par le repérage d'intersections entre représentations individuelles.

CHAPITRE III

LE CADRE MÉTHODOLOGIQUE DU FORAGE DE TEXTE

Introduction

Une méthodologie d'étude des RS doit comprendre différentes méthodes et techniques correspondants à trois niveaux d'analyse : l'identification des éléments de la RS, l'identification de ses structures et l'identification de son noyau central (Abric 1994b : 60; 2003c : 376). Les méthodes qui conviennent à chacun de ces niveaux ont pour la plupart été élaborées dans le cadre de démarches expérimentales ou par enquêtes. Elles ont en commun de faire intervenir des sujets à toutes les étapes de l'enquête, généralement par l'intermédiaire d'un questionnaire ou d'une méthode associative.

Lorsque des RS sont étudiées par l'analyse de corpus de presse, un cadre méthodologique différent est nécessaire : soit des approches qualitatives classiques d'analyse de contenu, soit des approches quantitatives d'analyse statistique des données textuelles qui font appel à des méthodes de forage de texte.

Dans ce chapitre, est abordé le deuxième objectif de recherche, qui consiste présenter un cadre méthodologique général de forage de texte. Ce cadre est commun à plusieurs logiciels utilisés dans les sciences humaines et sociales.[18] Tout d'abord, est présentée brièvement la thèse linguistique sur laquelle est basée cette méthode : la sémantique distributionnelle. Cette thèse appartient à la statistique textuelle et a été récupérée par plusieurs courants de recherche, notamment en informatique et plus particulièrement dans les pratiques de lecture et d'analyse de texte assistée par ordinateur (LATAO).

---

[18] À titre d'exemple, voir ceux nommés dans la section 2.2 du premier chapitre.

Deuxièmement, sont présentées ces pratiques de LATAO. Il s'agit d'un domaine spécifique où la statistique textuelle s'est développée. Troisièmement, est présenté le forage de texte en tant que méthode particulière de LATAO. Il s'agit d'un ensemble de méthodes informatiques, qui ont pour objectif l'organisation et l'extraction d'information dans un corpus de texte, à l'aide de différents algorithmes. Ces algorithmes viennent aussi bien de la linguistique computationnelle, du forage de données (*Data Mining*) que de l'apprentissage machine et la reconnaissance de forme. Ces algorithmes sont utilisés généralement dans le cadre d'une méthode, qui se déploie de manière relativement séquentielle sur plusieurs étapes de traitement des données textuelles. Enfin, ce chapitre se termine par la présentation dans le détail des étapes d'une méthode générale de forage de texte.

1. La statistique textuelle et la thèse distributionnaliste

Une thèse importante pour la statistique textuelle trouve ses origines dans les travaux classiques de Harris (Harris 1968) et de Benzécri (Benzécri 1981). Cette thèse, et les méthodes d'analyse quantitative de données textuelles qu'elle permet sont nées de la rencontre de plusieurs disciplines comme la linguistique, la statistique, l'analyse de discours et l'informatique (Lebart et Salem 1994). Aujourd'hui, ses domaines d'application sont nombreux, aussi bien en sociologie qu'en psychologie, dans le traitement des enquêtes, en histoire, etc.

Cette thèse est le distributionnalisme et, plus spécifiquement en ce qui nous concerne, la sémantique distributionnelle. Nous pouvons la résumer brièvement de la manière suivante : la sémantique d'un mot cible dans un texte dépend de son usage, c'est-à-dire des relations de cooccurrence avec les autres mots présents dans le texte. Cette thèse prévoit que deux mots caractérisés par des patterns de cooccurrence similaires sont généralement proches sémantiquement l'un de l'autre, soit par synonymie, par hyponymie ou encore par méronymie, métonymie :

> « The analysis of a given environment has to be made in respect to all the other environments of the word in question, so that the given one is obtained from the others by the regular processes. » (Harris 1991: 11).

> « Meaning is more easily stated as a property of word combinations (or of words in combination) than of words by themselves. [...] Once we see that the meaning of a word in a particular occurrence depends on its environment, the categorization of meaning-ranges can be replaced by a categorization of the environing words. Note that when we judge the meaning of a word-occurrence by what can replace it in the text, we are really judging the meaning by textual environment. » (Harris 1991 : 325).

Dans la version harrissienne en particulier, on insiste sur le fait que la signification est le résultat de l'usage socioculturel des mots, et non l'inverse. La distribution de l'information (des mots) dans la parole ou dans le texte est le résultat de pratiques et d'habitudes historiquement situées et socialement partagées par les membres d'une communauté linguistique :

> « [...] with Sapir [Harris] saw that socially instituted form or pattern, especially that in language, constitutes aspects of meaning that could not exist without it. [...] Harris saw, as Sapir had seen, that people have historically used language to build up, over generations, types of meaning that did not previously exist and in fact could not exist without language. This linguistic information is conventional, and it is objective to the degree that it is public, transmitted and shared by means of language, as distinct from the private, subjective world of perceptions apart from language. » (Nevin 1993 : 360)

À titre d'exemple (simpliste, mais qui sert d'illustration), dans les phrases suivantes les mots *maison* et *demeure* sont utilisés de manière similaire, c'est-à-dire qu'ils ont des patterns relationnels similaires. Selon la thèse distributionnaliste, ils seront par conséquent considérés comme proche sémantiquement :

> Cette *maison* sur la colline est la mienne.
>
> Cette *demeure* sur la colline est la mienne.

Ces deux mots sont considérés sémantiquement proches, non pas en fonction de leur contenu ou de leur référent, mais en fonction de leur usage, c'est-à-dire leur position relationnelle dans le texte (ici, la phrase). Il s'agit donc d'une approche empirique du langage, inductive et probabiliste : statistiquement, on observe dans un corpus de texte donné (ici deux phrases seulement) que l'usage des mots *maison* et *demeure* est similaire (ici à 100%).

Ceci peut également s'appliquer à des domaines de texte plus grands que les mots. On peut traduire statistiquement la sémantique d'un segment de texte, tels une phrase, un paragraphe, une page, etc. Si on prend un paragraphe par exemple, celui-ci peut être décrit statistiquement par sa distribution informationnelle, par exemple, par la distribution de ses lexèmes. Dans cet exemple, la distribution de lexème sert à décrire l'information présente dans le paragraphe. Mais ce pourrait aussi être une distribution de mots composés, de lettres, de n-grammes, d'une catégorie linguistique quelconque.

Selon la thèse distributionnaliste, deux paragraphes caractérisés par des distributions lexicales similaires ont de fortes probabilités d'avoir des significations similaires. À titre d'exemple, on peut prendre le segment de texte suivant composé de deux phrases :

> *segment$_{(1)}$: Cette maison est la mienne. C'est la maison où j'habite depuis toujours et c'est la maison à vendre.*

On peut traduire ce segment$_{(1)}$ par sa distribution informationnelle sous la forme d'une courbe (courbe rose dans le graphique de la figure 3). Dans cette courbe, l'axe des abscisses représente les lexèmes, et l'axe des ordonnées leur fréquence :

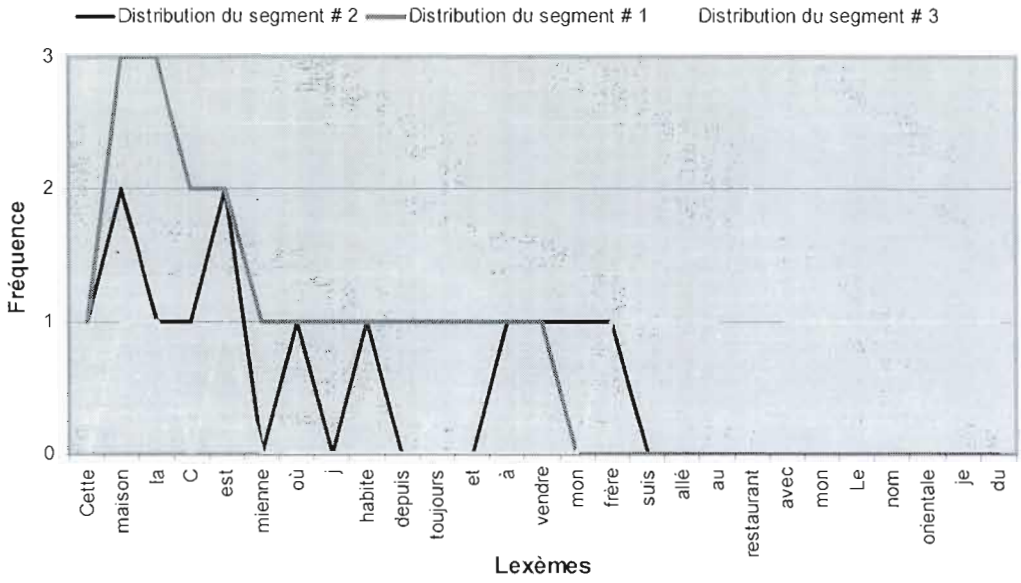


Figure 3 : Exemple de courbe de distribution informationnelle de trois segments de texte

Un deuxième segment, composé de deux phrases (courbe bleue), pourrait être par exemple le suivant :

segment$_{(2)}$: Cette maison est à vendre. C'est la maison où mon frère habite.

Un troisième segment, composé lui aussi de deux phrases (courbe jaune), pourrait être par exemple le suivant :

segment$_{(3)}$: Je suis allé au restaurant avec mon frère. Le nom du restaurant était la Maison orientale.

Avec la thèse distributionnaliste comme base théorique, on explore des dimensions sémantiques d'un texte à l'aide de différents calculs statistiques. Pour décrire la structure organisant l'information dans un texte, on fait appel aux outils généraux de la statistique. Certains de ces outils sont très simples, comme le calcul de fréquence, le TTP (type/token

ratio)[19], d'autres sont plus complexes, comme le calcul de distance entre deux courbes et la classification mathématique.

Pour terminer rapidement avec notre exemple précédent, par un simple coup d'œil, on peut faire l'hypothèse que les segment$_{(1)}$ et segment$_{(2)}$ sont beaucoup plus proches sémantiquement l'un de l'autre que le segment$_{(3)}$. Un simple calcul de distance sémantique en termes de mot commun corrobore cette observation : la distance entre les segment$_{(1)}$ et segment$_{(2)}$ est de 0,67 (1- (9/27), alors que la distance entre les segment$_{(1)}$ et segment$_{(3)}$ est de 0,89 (1- (3/27), et que la distance entre les segment$_{(2)}$ et segment$_{(3)}$ est de 0,85 (1- (4/27). Nous verrons plus loin que la représentation vectorielle des données textuelles permet d'appliquer des algorithmes de calculs de distance plus sophistiqués, comme la distance euclidienne et ses variantes.

Par ailleurs, le développement de la statistique textuelle a été fortement dépendant du développement informatique. En fait, la statistique textuelle s'inscrit dans le courant plus large des pratiques de lecture et d'analyse de textes assistées par ordinateur (LATAO), qui elles-mêmes datent d'environ cinquante ans.

2. La lecture et l'analyse de texte assistée par ordinateur

La LATAO est une pratique qui a aujourd'hui pénétré la plupart des sciences sociales et des humanités (Hockey 2001). On la retrouve en philosophie (Meunier et Forest 2000), en sociologie (Popping 2000; Demazière *et al.* 2006), en littérature (Brunet 1986), et ailleurs.

Bien qu'elle utilise parfois des méthodes algorithmiques similaires à d'autres programmes de recherche — comme l'intelligence artificielle (Hobbs 1990), la recherche d'informations et la bibliothéconomie (Salton 1989), la linguistique computationnelle (Mitkov 2003) — elle se distingue par ses objectifs, ses domaines d'application et l'interprétation des résultats (Meunier *et al.* 2005). La LATAO s'inscrit généralement dans une recherche plus générale : pour le philosophe, ce peut être l'analyse conceptuelle, pour le littéraire, l'analyse stylistique,

[19] Le type/token ratio est le rapport d'une unité d'information spécifique sur l'ensemble des unités d'un texte, par exemple le nombre de lexèmes sur le nombre d'occurrences de mot dans un texte. C'est une mesure statistique classique de la diversité lexicale d'un texte.

pour le sociologue, l'analyse de discours. Comme le souligne Meunier et ses collaborateurs (Meunier *et al.* 2005), la LATAO se présente rarement comme l'objet de recherche en soi, mais plutôt comme une méthode pour mener à bien une recherche typique à chaque discipline :

> « In this field of application, these techniques cannot be an end in their self. No text reader and analyzer would find it very useful to submit his corpus to a classification and a categorization process just for the sake of it. But we believe that are useful if they are linked to the dynamic process of text interpretation. » (Meunier *et al.* 2005 : 963)

Elle permet cependant d'aborder et de traiter des données empiriques de nature textuelle de nouvelles manières, susceptibles de générer de nouvelles observations et conclusions parfois difficiles à obtenir autrement, notamment à cause de la grande quantité de données à traiter et de leur complexité.

Le type d'analyse qu'a rendu possible la pratique LATAO s'est rapidement développé, offrant des outils de plus en plus sophistiqués. Meunier et ses collaborateurs (Meunier *et al.* 2005) proposent de voir l'évolution de ceux-ci sur quatre générations.

Une première génération, de 1950-1970, a débuté avec les nouvelles possibilités du texte électronique. On retrouve dans cette période l'apparition des premières techniques de numérisation du texte. La deuxième génération, de 1970-1980, a vu l'apparition des premiers outils élémentaires de description et de manipulation des données textuelles, par exemple les calculs de fréquence, l'indexation du lexique, la concordance, la lemmatisation. La troisième génération, de 1980-1995, a été surtout caractérisée par des initiatives de standardisation des formats électroniques (SGML, HTML, RDF, XML, etc.) et par l'étiquetage et le traitement linguistique des textes. La quatrième génération, depuis environ 1995, a consisté en l'application de modèles mathématiques sophistiqués pour la description et la manipulation de données textuelles. On fait référence ici à la classification et la catégorisation automatique des textes (sur la base de modèle vectoriel). Cette génération a donné lieu à des méthodes statistiques puissantes appelées aujourd'hui du forage de texte.

3. Le forage de texte

Le forage de texte représente un ensemble de méthodes informatiques consacrées à l'extraction et l'organisation de l'information et de la connaissance dans un corpus de données textuelles. Ces méthodes sont utilisées dans le cadre d'une méthode de recherche à la croisée entre la LATAO et le forage de donnée. Elle se différencie cependant du forage de données en général, car elle porte plus spécifiquement sur un type particulier de données, à savoir le texte. Le forage de texte est donc un ensemble de méthodes informatiques issues de la linguistique, de la statistique textuelle, de l'apprentissage machine, etc., qui a pour objectif l'analyse de texte :

> « Text mining or knowledge discovery from text (KDT) [...] deals with the machine supported analysis of text. It uses techniques from information retrieval, information extraction as well as natural language processing (NLP) and connects them with the algorithms and methods of KDD [knowledge discovery from data], data mining, machine learning and statistics. Thus, one selects a similar procedure as with the KDD process, whereby not data in general, but text documents are in focus of the analysis. » (Hotho *et al.* 2005 : 4)

Le texte est donc considéré comme un type particulier de données. Parmi ses particularités, le fait qu'on considère les données textuelles comme « non structurées » (Weiss *et al.* 2005), c'est-à-dire non organisées par une structure externe aux corpus, par exemple sous la forme d'une base de données. Une part importante des objectifs des méthodes de forage de texte est de parvenir, de manière assistée informatiquement, à faire émerger d'un corpus de données textuelles brutes, des structures susceptibles de révéler des éléments de connaissance intéressants pour l'analyste :

> « Text mining can be also defined — similar to data mining — as the application of algorithms and methods from the fields machine learning and statistics to texts with the goal of finding useful patterns. [...] Thus, our focus is on methods that extract useful patterns from texts in order to, e.g., categorize or structure text collections or to extract useful information. » (Hotho *et al.* 2005 : 4-5)

Meunier et ses collaborateurs (Meunier *et al.* 2005, 2006) proposent de distinguer les différentes méthodes de forage de texte en deux familles. En premier lieu, on fait généralement appel à des méthodes ascendantes (*bottom-up*), comme la classification automatique des textes (Jain *et al.* 1999; Kaufman et Rousseeuw 1990), qui permet de

regrouper, sous forme de classe, des données textuelles partageant certains attributs similaires. En deuxième lieu, d'autres méthodes de nature plus descendante (*top-down*), comme la catégorisation automatique (Sébastiani 2002), permettent quant à elles de repérer, à partir d'un patron préalablement identifié, des données textuelles d'un corpus qui lui correspondent.

Pour les sciences humaines et sociales, l'approche ascendante est particulièrement intéressante et adaptée. En effet, que ce soit pour le sociologue devant un corpus d'article de journal, ou le littéraire devant le corpus de Shakespeare, leur démarche consiste souvent à organiser leurs données, les structurer, les hiérarchiser, les trier, les classer, etc., de manière à en faire émerger des formes qui auront du sens pour lui en fonction de son cadre théorique.

Ces méthodes ascendantes, auxquelles correspondent différentes opérations algorithmiques, peuvent être regroupées sous une méthode générale qui se déploie de manière relativement séquentielle sur six étapes.

4. Les étapes d'une méthode de forage de texte

Dans cette section, nous présentons une méthode générale de forage de texte. L'objectif de cette méthode est de générer, à partir d'un corpus de données textuelles brutes comme un corpus d'articles de presse, une structure de classes. Autrement dit, on procède à une classification automatique de segments de texte extraits d'un corpus. Nous proposons de présenter cette méthode sous la forme d'une chaîne de traitement informatique générique commune à plusieurs logiciels (mais souvent gardée cachée pour différentes raisons).[20]

Nous allons présenter dans le détail : (4.1) la construction et la préparation du corpus de texte à analyser, (4.2) le choix du domaine d'information (DOMIF) et la segmentation du corpus de texte, (4.3) le choix de l'unité d'information (UNIF) et l'indexation, (4.4) le choix d'une valeur de pondération et la vectorisation des segments de texte, (4.5) la construction de la

[20] Les raisons pouvant mener à la dissimulation des véritables opérations en jeu dans un logiciel sont nombreuses. Dans le cas de logiciels propriétaires comme ALCESTE, PROVALIS et d'autres, c'est surtout une question de protection contre la récupération ou le calquage.

matrice et la sélection de ses dimensions, et (4.6) le choix de l'algorithme de classification et la validation de la classification.

	Étapes	Choix d'opérationnalisations
1	Constitution du corpus	Articles de presse, œuvres, pages web, verbatims, archives, etc.
2	Segmentation (DOMIF)	Document, page, paragraphe, concordance, phrase, etc.
3	Indexation (UNIF)	Lexème, lemme, n-gramme, catégories morphosyntaxiques, mot-composé, etc.
4	Vectorisation	Binaire, TF, TF-IDF, fréquence standardisée, BM25, etc.
5	Réduction dimensionnelle	Analyse en composante principale, LSA, lemmatisation, antidictionnaire, etc.
6	Classification automatique	K-Means, ART1, SOM, Classification hiérarchique, CLARANS, DBSCAN, EM, etc.

Figure 4 : Diagramme séquentiel des opérations d'une chaîne de forage de texte

Ce cadre général a été appliqué aussi bien à l'analyse des RS, tel que pratiqué à l'aide du logiciel ALCESTE, que pour d'autres types d'analyses de textes assistées par ordinateur utilisant d'autres logiciels, comme l'analyse thématique en sciences de l'information (Forest 2006), l'analyse conceptuelle en philosophie (Meunier et Forest à paraître 2009), l'analyse des attracteurs sémantiques en anthropologie (Gagnon 2004), l'analyse des mondes lexicaux en littérature (Reinert 1993), et l'analyse documentaire en général (Desjardins 2006). Ces différents types de recherches se ressemblent dans la mesure où elles procèdent toutes par ces six étapes. Ce qui les distingue est l'opérationnalisation de ces différentes étapes, c'est-à-dire le choix d'un algorithme particulier à chaque étape et surtout, le choix d'un cadre d'analyse et d'interprétation.

4.1 La sélection et la préparation du corpus de texte à analyser

La première étape d'une méthode de forage de texte est la sélection et la préparation du corpus. Un corpus est une collection de textes réunis par le chercheur à des fins d'expérimentations et d'analyse. Cette étape, évidente, est néanmoins essentielle, car en amont de tout le reste de la démarche : « Clearly, the first step in text mining is to collect the data (i.e., the relevant documents) » (Weiss *et al.* 2005 : 15). Cependant, il n'y a pas de

protocole standard pour cette première étape. C'est-à-dire qu'il n'y a pas de guide de procédures qui fasse l'unanimité et qui garantisse la validité de l'étape, et qui permettrait de dire : ce corpus est valide, car nous avons fait telle et telle chose. Nous pouvons néanmoins, à partir de la pratique issue autant du forage de texte (Hotho *et al.* 2005) que de l'analyse de contenu classique (Bardin 1977), identifier des règles de bonne pratique qui vont assurer, dans une certaine mesure, la qualité de l'analyse.

Issues de l'analyse de contenu, on retrouve quatre règles de bonne pratique. Premièrement, l'identification de l'objet de recherche : des objectifs différents nécessiteront des opérationnalisations également différentes de la méthode. Ce n'est pas la même chose que de faire du forage de texte pour de la veille documentaire, de l'analyse thématique ou de l'analyse de RS.

Deuxièmement, la sélection des types des textes selon un critère de pertinence pour l'objectif de recherche. La construction du corpus se fait en rassemblant, dans une même collection, plusieurs textes jugés pertinents par le chercheur compte tenu de son objectif de recherche. Le chercheur ou la chercheuse doit choisir s'il veut travailler sur des articles de journaux de presse, des recettes de cuisine, des archives, des courriels échangés dans une entreprise, des blogues sur le web, etc.

Un autre critère important est l'exhaustivité. Le chercheur doit rassembler systématiquement l'ensemble des textes qui respectent les critères qu'il a identifiés précédemment. Si cela est impossible, il devra entamer une démarche afin que son corpus, bien que non exhaustif, soit le plus représentatif possible.

Finalement, les textes formant le corpus doivent également être caractérisés par une certaine homogénéité, c'est-à-dire partager un certain nombre d'attributs communs. Par exemple, ils doivent traiter d'un même sujet, à propos d'un même objet ou avoir le même auteur, être dans la même langue, etc. Ces critères d'homogénéité varient selon l'objectif de recherche.

Du côté du forage de texte, on retrouve aussi quatre principales règles. Premièrement, on retrouve la nécessité de constituer un corpus qui soit significatif statistiquement. Puisque le forage de texte est une chaîne de traitement informatique de nature mathématique, le corpus

doit être suffisamment important, en termes de taille, de nombre de caractère, de mots, etc., pour que les différentes opérations (algorithmes) de la méthode puissent repérer des redondances.

Ensuite, le format électronique du texte doit être compatible avec celui du système :

> « To mine text, we first need to process it into a form that data-mining procedures can use. » (Weiss *et al.* 2005 : 15)

> « The software tools need to read data just in one format, and not in the many different formats they came in originally. » (Weiss *et al.* 2005 : 19)

Autrement dit, il faut soumettre à la chaîne de traitement un format qu'elle peut traiter. Ceci peut impliquer de formater tous les textes en .XML, en format .TXT, ou .RTF. Autres exemples : parfois il faut changer la ponctuation, uniformiser le nombre de retours de chariot entre paragraphes, le nombre d'espaces entre les mots.

Il faut également nettoyer les textes de différentes « coquilles » comme les erreurs de frappe ou les caractères non reconnus. Il faut également filtrer les métadonnées non pertinentes pour l'objectif de recherche. Ces métadonnées peuvent être le numéro de page, des liens hypertextes (si notre corpus est composé de pages web), les codes ou numéros de documents, etc.

Finalement, il faut s'assurer que les textes sont d'une bonne qualité orthographique. Ceci est important, car les algorithmes qui forment la méthode sont aveugles à la sémantique. Ils ne font des calculs que sur des chaînes de caractères et non sur des mots.[21] Dans un tel cadre, les chaînes de caractères *apartenir* et *appartenir* seront considérées comme différentes. Un corpus avec trop d'erreurs orthographiques entraînera du bruit dans le traitement, réduisant par la suite la qualité de l'analyse.

[21] Un mot est une interprétation *humaine* d'une série de caractères.

4.2 Le choix du domaine d'information et la segmentation du corpus

La deuxième étape de cette méthode générale de forage de texte est l'identification du type de domaines d'information (DOMIF) pertinent pour le chercheur et la segmentation du corpus de texte en fonction de ce type. L'objectif de cette étape est d'extraire du corpus de texte toutes les instances ($domif_{(i)}$) du DOMIF sélectionné.

Dans le forage de texte, un DOMIF est un type de segment de texte. Un DOMIF peut être la phrase, le paragraphe, la page, un nombre de ligne, une séquence de n-mots, ou tout autre type pertinent (Meunier *et al.* 2005). Une fois le DOMIF sélectionné, un algorithme extrait du corpus toutes ses $domif_{(i)}$. Si le DOMIF est la phrase, on extrait toutes les phrases du corpus. Si c'est la page ou le paragraphe, on extrait toutes les instances du type *page* ou toutes les instances du type *paragraphe* dans le corpus.

La sélection du DOMIF se fait en fonction de l'objectif de recherche et des caractéristiques du corpus. Les instances du DOMIF sont les segments de textes, qui seront par la suite organisés par une opération de classification automatique. Classer des pages ou des phrases ou des paragraphes donnera évidemment des classes différentes. Cette étape dépend également de la taille du corpus. En effet, on segmentera différemment un corpus s'il est composé de dix pages que s'il en contient mille. Dans le premier cas, une segmentation par phrase sera probablement plus intéressante, alors que dans le deuxième cas, le paragraphe ou la page sera un DOMIF plus approprié.

4.3 Le choix de l'unité d'information et l'indexation du corpus

La troisième étape est l'identification d'un type d'unités d'informations (UNIF) et l'indexation de toutes les instances ($unif_{(i)}$) de l'UNIF sélectionnée. L'objectif de cette étape est de sélectionner un type d'information — un type de variable — par lequel le chercheur veut caractériser ou décrire le DOMIF sélectionné précédemment. Cette étape est donc dépendante de la précédente.

Nous avons vu qu'un DOMIF est généralement un type de segment de texte aux proportions variables. Ce type de segment de texte est caractérisé par différents types d'informations. Ce sont généralement ses propriétés linguistiques, dont la plus utilisée est le lexème.

Cette étape consiste à choisir un UNIF qui servira à décrire un DOMIF. Par extension, on décrit tous les $domifs_{(i)}$ par les $unifs_{(i)}$ qu'ils contiennent. Autrement dit, il s'agit de sélectionner le type d'information que l'on veut retenir pour l'analyse. Un DOMIF peut par exemple être décrit par sa distribution de lexèmes, mais il peut également être décrit par ses catégories morpho-syntaxiques, ses n-grammes, ses mot-grammes[22], etc.

Cette étape requiert des algorithmes d'extraction automatique des instances de l'UNIF sélectionnée, à partir desquels on construit un index. On parle parfois d'un processus de « tokenization » (Hotho *et al.* 2005). Si par exemple l'UNIF est le lexème et le DOMIF la phrase, on extrait tous les lexèmes différents de toutes les phrases composant le corpus. Le choix d'une UNIF dépend des objectifs de recherche, du DOMIF, parfois de la langue du corpus.

4.4 Pondération et vectorisation

La quatrième étape est la pondération de chaque instance de l'UNIF dans chaque instance de DOMIF. Ceci est ensuite traduit dans un espace vectoriel. L'objectif de cette étape est de représenter un corpus de texte sous la forme d'une matrice de vecteurs (Salton et McGill 1983; Memmi 2000) qui permet par la suite des manipulations mathématiques utiles à partir desquelles on pourra faire émerger des structures dans les données :

> « Here, each segment with its information units is seen as a vector where the informational content is translated as a property of this vector. [...] Thus the whole text is transformed into a vectorial space. This allows the use of all the mathematical tools that can be applied on this space. This is the force of the classification and

[22] Un n-gramme est une séquence arbitraire de n-caractères. Par exemple, le lexème *forage* contiendrait les tri-grammes suivants for, ora, rag, age. Les n-grammes sont un type d'UNIF très intéressant, car performant et indépendant de la langue du texte (Damashek 1989; Cavnar et Trenkle 1994). La même logique peut aussi s'appliquer aux mots sous la forme de mot-gramme (Choueka 1988).

> categorization approach. The text is not parsed linguistically but mathematically. » (Meunier *et al.* 2005 : 966)
>
> « Despite of its simple data structure without using any explicit semantic information, the vector space model enables very efficient analysis of huge document collections. » (Hotho *et al.* 2005)

Dans cet espace vectoriel, chaque $domif_{(i)}$ représente un vecteur à n-dimensions. Les dimensions de chaque vecteur correspondent aux $unif_{(i)}$ présentes dans le $domif_{(i)}$. Par exemple, pour le $domif_{(1)}$ suivant (un domaine d'information de deux phrases) :

> $domif_{(1)}$: *Cette maison est la mienne. C'est la maison où j'habite depuis toujours et c'est la maison à vendre.*

On peut représenter l'information dans ce $domif_{(1)}$ par différente UNIF, par exemple le lexème. On extrait du $domif_{(1)}$ toutes les instances du type lexème. Ce $domif_{(1)}$ contient les unités d'information suivantes :

> $domif_{(1)}$ = [CETTE, MAISON, EST, LA, MIENNE, C, OÙ, J, HABITE, DEPUIS, TOUJOURS, ET, À, VENDRE].

On accorde ensuite à chaque $unif_{(i)}$ une valeur de pondération qui traduit son poids dans le $domif_{(1)}$. Le choix d'une mesure de pondération est une étape importante et plusieurs possibilités d'encodages de cette valeur sont possibles, selon les objectifs de recherche. Pour Hotho *et al.* : « The main task of the vector space representation of documents is to find an appropriate encoding of the feature vector. » (Hotho *et al.* 2005)

Il existe plusieurs algorithmes de calcul de pondération d'une $unif_{(i)}$ dans un $domif_{(i)}$. Le plus simple est une valeur binaire : une $unif_{(i)}$ a une valeur de 0 ou 1 selon qu'elle est présente ou absente du $domif_{(i)}$.

On utilise également la fréquence relative (TF) : une $unif_{(i)}$ a une valeur égale à sa fréquence dans le $domif_{(i)}$. Pour l'exemple précédent, le $domif_{(1)}$ peut être représenté par ses lexèmes et pondéré par leurs TF de la manière suivante :

$domif_{(1)}$ = [(CETTE, 1), (MAISON, 3), (EST, 3), (LA, 3), (MIENNE, 1), (C, 2), (OÙ, 1), (J, 1), (HABITE, 1), (DEPUIS, 1), (TOUJOURS, 1), (ET, 1), (À, 1), (VENDRE, 1)]

Chaque $domif_{(i)}$ est ainsi vectorisé en fonction de ses $unifs_{(i)}$ et d'une valeur de pondération. On projette ensuite chaque vecteur dans un espace vectoriel à n-dimensions où chaque vecteur correspond à une coordonnée.

Il y a beaucoup d'autres mesures de pondération dans la littérature. Ceci est un domaine de recherche en soi. Par exemple, la *fréquence documentaire inverse* (IDF), ou le TF*IDF, l'entropie, le chi^2, etc. (Harman 2005).

Sous toute cette diversité de techniques de pondération, on peut souligner deux grands principes qui leur sont communs. Le premier est un principe de représentativité. La valeur de pondération d'une $unif_{(i)}$ dans un $domif_{(i)}$ sera d'autant plus élevée que l'$unif_{(i)}$ est très représentative du $domif_{(i)}$. La pondération par le TF est une manière de traduire ce critère de représentativité. Par exemple, dans le $domif_{(1)}$ précédent, les lexèmes très fréquents seront considérés comme plus représentatifs du $domif_{(1)}$ que les lexèmes moins fréquents. Dans ce cas, MAISON est une unité d'information plus représentative que DEPUIS.

Le deuxième est un principe de discrimination. Dans ce cas, la valeur de pondération d'une $unif_{(i)}$ pour un $domif_{(i)}$ sera d'autant plus élevé que cette $unif_{(i)}$ permet de discriminer ce $domif_{(i)}$ des autres $domifs_{(j)}$ formant le corpus. C'est ce critère dont veut rendre compte par exemple la pondération par le IDF : la valeur IDF d'une $unif_{(i)}$ sera élevé pour un $domif_{(i)}$ si cette $unif_{(i)}$ est présente dans ce $domif_{(i)}$, mais absente dans les autres $domifs_{(j)}$ du corpus. Si nous revenons à nos trois exemples précédents :

$domif_{(1)}$: *Cette maison est la mienne. C'est la maison où j'habite depuis toujours et c'est la maison à vendre.*

$domif_{(2)}$: *Cette maison est à vendre. C'est la maison où mon frère habite.*

$domif_{(3)}$: *Je suis allé au restaurant avec mon frère. Le nom du restaurant était la Maison orientale.*

Dans ce cas, RESTAURANT aura une valeur de pondération IDF élevée, car présent seulement dans le $domif_{(3)}$, alors que MAISON aura une valeur nulle, car présent dans tous les domaines d'information.

Un calcul de pondération comme le TF*IDF veut quant à lui rendre compte à la fois du principe de représentativité et de discrimination en combinant un calcul de fréquence et de fréquence documentaire inverse.

4.5 La construction de la matrice et la sélection des dimensions

La cinquième étape est l'organisation des vecteurs représentant un corpus sous la forme d'une matrice (Manning and Schütze 1999) avec, en rangée, les $domifs_{(i)}$ et, en colonne, les $unifs_{(i)}$. Tous les vecteurs sont standardisés, c'est-à-dire ramenés au même nombre de dimensions. Dans la matrice, il y a autant de rangées qu'il y a de $domif_{(i)}$ issus du corpus analysé et autant de colonnes qu'il y a d'$unif_{(i)}$ retenues pour décrire l'information dans les $domifs_{(i)}$. La figure 5 illustre le genre de matrice que l'on peut obtenir :

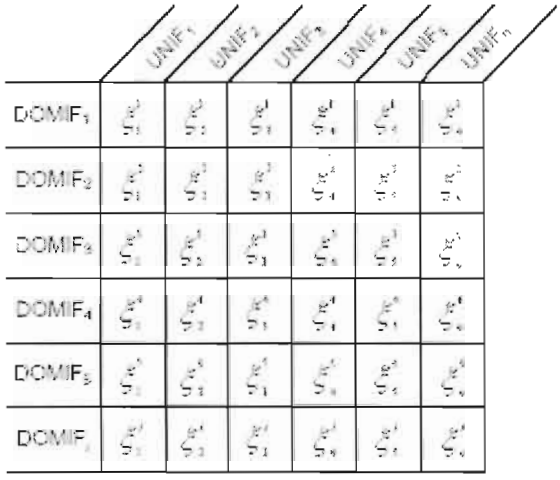


Figure 5 : Matrice $domif_{(i)}$ * $unif_{(i)}$

Un problème important rencontré à cette étape est le nombre de dimensions de la matrice, c'est-à-dire le nombre de colonnes représentant chaque $unif_{(i)}$. Si le chercheur ou la chercheuse analyse un corpus de plusieurs centaines de pages où, par exemple, l'information est représentée via une distribution lexicale, cette matrice peut facilement avoir plus de dix

mille dimensions, c'est-à-dire plus de dix mille lexèmes différents. On appelle ce phénomène la « malédiction dimensionnelle ».

De plus, dans le cas de matrice très large, on se retrouve souvent avec plusieurs valeurs nulles, c'est-à-dire où les vecteurs de la matrice sont caractérisés majoritairement par des zéros.[23] Or, ceci peut causer des problèmes de performance, autant du point de vue du traitement informatique que de la qualité des résultats de la classification de l'étape suivante (Weiss *et al.* 2005). D'autre part, certaines dimensions de la matrice, c'est-à-dire certaines $unif_{(i)}$, peuvent ne pas être pertinentes pour l'analyse. Ce peut être le cas par exemple des $unifs_{(i)}$ présentes dans un seul $domif_{(i)}$ du corpus. Il est donc important à cette étape de réduire les dimensions de la matrice.

Il existe plusieurs opérations algorithmiques possibles pour cette étape, certaines sont de nature linguistique, d'autres statistiques. La plupart sont utilisées lorsque l'UNIF retenue pour décrire le DOMIF est le lexème. Dans les opérations de nature linguistique, on retrouve des algorithmes classiques de lemmatisation, qui consiste à ramener différents lexèmes à leur racine commune.[24] Ensuite, on procède généralement à la suppression des « mots fonctionnels » ou « mots outils » à l'aide d'un antidictionnaire (*stop-list*). Un antidictionnaire est composé d'articles, de propositions, de nombres, de quelques verbes auxiliaires, quelques adverbes, etc. On utilise aussi parfois un dictionnaire de synonymes comme WORDNET (Hotho *et al.* 2003) pour ramener à une seule $unif_{(i)}$ plusieurs lexèmes synonymes.

Parmi les opérations de nature statistique, on retrouve des algorithmes qui permettent de ne conserver que les $unifs_{(i)}$ les plus représentatives et les plus discriminantes de la matrice. On

---

[23] Ceci est une conséquence de la standardisation de la longueur des vecteurs. En effet, si nous avons une matrice de dix mille dimensions et que les vecteurs représentent, par exemple, des paragraphes, avec en moyenne cent à deux cents lexèmes différents chacun, une fois organisés sous la forme d'une matrice, la plupart des cases de la matrice seront égales à zéro.

[24] Le lemme et le stemme sont des flexions du lexème, qui ramènent les différentes variantes d'un lexème à sa racine, par exemple le pluriel au singulier, le féminin au masculin, etc. La lemmatisation procède généralement en ramenant le lexème à son infinitif, par exemple, le lexème *marcherait* est ramené au lemme *marcher*. Le *stemming* procède quant à lui par troncature où le lexème *marcherait* est ramené à le stemme *march*.

supprime généralement les hapax.[25] Aussi, on peut supprimer les $unifs_{(i)}$ distribuées de manière trop équiprobable parmi les $domifs_{(i)}$.[26] Les méthodes possibles pour la réduction dimensionnelle sont très nombreuses. Certains appliqueront sur la matrice des méthodes plus sophistiquées, comme les analyses en composantes principales (Jain *et al.* 1999 : 270) ou les analyses sémantiques latentes (Landauer *et al.* 1998).

Autre point important à souligner, la réduction de la matrice ne doit pas aboutir à une matrice dite « creuse », c'est-à-dire où l'une des rangées n'aurait que des zéros comme valeur, car ceci signifierait qu'un $domif_{(i)}$ ne serait représenté par aucune de ses $unifs_{(i)}$, donc impossible à classer par la suite.

Il existe de nombreuses stratégies.[27] Ceci constitue un domaine de recherche. En dernière instance, la plupart de ces stratégies dépendent des objectifs de recherche. Néanmoins, leur base commune est de ne conserver dans la matrice que l'information la plus pertinente pour l'analyse : « The aim of this step is to produce a set of most relevant information unit. » (Meunier *et al.* 2005 : 965).

4.6 La classification automatique

La sixième étape est la classification automatique des vecteurs formant la matrice, c'est-à-dire la classification des $domifs_{(i)}$ extraits du corpus analysé. Il s'agit d'une opération de classification automatique et non supervisée, aussi appelée « clustering » (Jain *et al.* 1999; Hotho *et al.* 2005; Weiss *et al.* 2005). La classification a pour objectif d'organiser un ensemble de données — ici les instances du DOMIF extraites du corpus — selon une structure en classe inhérente aux données. À partir des propriétés mathématiques des

---

[25] C'est-à-dire les $unifs_{(i)}$ à une seule occurrence ou les $unifs_{(i)}$ trop singulières à quelque $domif_{(i)}$. Par exemple, on supprime les $unifs_{(i)}$ présentes dans moins de 1% des $domifs_{(i)}$ du corpus.

[26] Par exemple, on supprime les $unifs_{(i)}$ présentes dans plus de 40% des $domifs_{(i)}$ du corpus.

[27] En réalité, plusieurs de ces stratégies, pour des raisons pratiques d'implémentation, sont réalisées avant la construction de la matrice. C'est généralement le cas de la lemmatisation et du *stemming*. Pour faciliter la présentation de la méthode, nous les avons toutes regroupées dans une même étape.

domifs$_{(i)}$, traduites sous la forme de vecteurs, un algorithme de classification regroupe ces domifs$_{(i)}$ selon la similarité de la distribution des unifs$_{(i)}$ qui les composent.

La classification produit des groupes de domifs$_{(i)}$ similaires, c'est-à-dire qui partage une distribution d'information semblable. La classification organise donc les domifs$_{(i)}$ de manière à ce qu'ils soient très semblables à l'intérieur d'une classe et différenciés entre les classes.

> « Clustering method can be used in order to find groups of documents with similar content. The result of clustering is typically a partition (also called) clustering P, a set of clusters *P*. Each cluster consists of a number of documents *d*. Objects — in our case documents — of a cluster should be similar and dissimilar to documents of other clusters. Usually the quality of clusterings is considered better if the contents of the documents within one cluster are more similar and between the clusters more dissimilar. Clustering methods group the documents only by considering their distribution in document space (for example, a *n*-dimensional space if we use the vector space model for text documents). » (Hotho *et al.* 2005)

Sur un même corpus de texte, le nombre de classes obtenues peut varier, et ce en fonction de plusieurs critères : la nature du corpus traité, le DOMIF et l'UNIF retenus, le nombre d'instances, le type de classification employé, son paramétrage, etc.

Les types de classifications non-supervisées sont très nombreux. On retrouve des types de classifications hiérarchiques descendantes, tel qu'utilisés dans le logiciel ALCESTE (Reinert 1993), ou ascendantes, tel qu'utilisés dans le logiciel de PROVALIS. Les techniques d'analyse factorielle sont également un type élémentaire de classification (Benzecri 1891). D'autres types de classifications plus sophistiquées forment par exemple les méthodes statistiques de classification par partitions itératives, tels les *K-Means* et leurs variantes (MacQueen 1967), ou encore les méthodes neuronales, tel le *Self Organizing Maps* (Kohonen 2001). Certaines méthodes sont des classifications rigides (*hard-clustering*), où un domif$_{(i)}$ ne peut être attribué qu'à une seule classe, d'autres sont des classifications souples (*soft-clustering*), où un domif$_{(i)}$ peut être attribué à plus d'une classe à la fois.[28]

---

[28] Les types de méthodes de classification sont bien plus nombreux que ce qui est présenté ici. On peut retrouver chez Jain *et al.* (1999 : 274) et Desjardins (2006) une description technique détaillée.

Comme le rappel Meunier *et al.* (2005 : 967) et Forest (2006 : 55), le choix d'un type de classification dépend de plusieurs facteurs et chacun a ses limites et avantages compte tenu d'un objectif de recherche. Lorsque le forage de texte est utilisé comme méthode de traitement et d'analyse de données textuelles dans les sciences sociales et humaines (ce que certains appellent les « digital humanities » (Hockey 2001; MacCarty 2005)), deux facteurs sont particulièrement importants, à savoir la performance de l'algorithme et sa facilité d'interprétation. Souvent, la recherche consiste à faire un compromis entre ces deux contraintes.[29] Par performance on entend la qualité de la classification, alors que par facilité d'interprétation on entend la rétractibilité ou la transparence du calcul de l'algorithme.

Les méthodes de classification hiérarchique sont probablement les plus faciles à interpréter, particulièrement dans leur représentation en dendrogramme. Par contre, elles font partie des méthodes les moins performantes lorsqu'appliquées sur des corpus de taille importante (Hotho *et al.* 2005; Jain *et al.* 1999; Weiss *et al.* 2005). Les approches neuronales sont reconnues pour la qualité de leurs résultats (Desjardins 2006), mais peuvent gêner le chercheur, car la complexité de leur calcul est telle qu'elles deviennent rapidement des « boîtes noires ». Les approches statistiques semblent un compromis entre les deux. D'ailleurs, le K-Means reste aujourd'hui l'une des techniques de classification non-supervisées les plus utilisées dans le forage de texte (Hotho *et al.* 2005).

Il est également important de noter que les méthodes de classification non-supervisées sont des méthodes exploratoires et descriptives. Elles permettent de regrouper des $domifs_{(i)}$ selon différents critères de similarité, mais une partition de $domifs_{(i)}$ n'est toujours qu'*une parmi plusieurs possibilités*. Il est toujours possible d'organiser d'une manière différente des $domifs_{(i)}$ issus d'un corpus. Il n'existe pas de bonne classification à priori. Il existe malgré tout des indicateurs statistiques pour évaluer la qualité d'une classification.

Ces indicateurs statistiques sont des stratégies servant à fournir au chercheur ou à la chercheuse des indices susceptibles de refléter la qualité de la classification. Ces indicateurs

[29] Le facteur de performance est aussi associé à la performance de calcul de l'algorithme, par exemple à son temps de traitement. Dans le domaine des sciences humaines et sociales, le facteur de performance associé au temps de traitement est beaucoup moins important. Nous ne le retenons pas ici parce qu'il n'est pas pertinent.

sont très nombreux dans la littérature et, encore une fois, le développement de ces algorithmes constitue un objet de recherche en soi. Les plus utilisés, selon Hotho *et al.* (2005) sont le *coefficient d'erreur moyenne au carré* et le *coefficient silhouette* (voir aussi Kaufman et Rousseeuw 1990). Dans la tradition française, on utilise également le coefficient du chi².[30] Le principe commun de ces indicateurs est le suivant : une classification sera de bonne qualité si les $domifs_{(i)}$ d'une classe sont très similaires entre eux, et très différents des autres $domifs_{(j)}$ issus du corpus. La représentation des $domifs_{(i)}$ dans un espace vectoriel facilite ce genre de calcul : plus la distance entre vecteurs est petite intra-classe et grande inter-classe, plus la classification est considérée de bonne qualité.

Retour sur le deuxième objectif

Ce chapitre était consacré à notre deuxième objectif de recherche, qui consistait à présenter, de manière la plus transparente possible, le forage de texte et son cadre méthodologique.

Dans une première section a été présentée la thèse linguistique *distributionnaliste* sur laquelle est basé le forage de texte. Selon cette thèse, la sémantique des textes, essentiellement via ses dimensions lexicales, est analysée de manière statistique et probabiliste, notamment à l'aide d'un modèle mathématique vectoriel.

Dans les sections deux et trois, les méthodes de forage de texte ont rapidement été recadrées dans le contexte historique des pratiques d'analyses de texte assistées par ordinateur en sciences humaines et sociales et distinguées des méthodes de forage de données en informatique.

Dans la dernière section ont été présentées six étapes qui composent une chaîne de traitement générale de forage de textes. Cette chaîne représente le tronc commun d'une méthode

[30] C'est le cas du logiciel ALCESTE.

commune à plusieurs logiciels. Elle mène à une classification, c'est-à-dire une partition des différents $domifs_{(i)}$ extraits du corpus, selon une structure en classes.[31]

Les étapes d'une chaîne de forage de texte sont : premièrement, la préparation et la constitution d'un corpus de textes à l'aide de règles de bonne pratique issues à la fois de l'analyse de contenu et du forage de texte; deuxièmement, la sélection d'un DOMIF et la segmentation du corpus; troisièmement, la sélection d'une UNIF et son indexation pour chaque instance du DOMIF; quatrièmement, la pondération des $unifs_{(i)}$ pour chaque $domifs_{(i)}$ et la vectorisation des $domifs_{(i)}$; cinquièmement, la construction de la matrice et la sélection des dimensions pertinentes pour la classification; et sixièmement, la classification automatique des $domif_{(i)}$ vectorisés et sa validation.

Une telle méthode est de plus en plus appliquée dans différents domaines d'études des humanités et des sciences sociales qui travaillent sur du matériel empirique de nature textuelle (Meunier *et al.* 2006). Et l'application de ce genre de méthode à l'étude des RS se pratique depuis une dizaine d'années. Autant dans le domaine d'étude des RS que dans les humanités et les sciences sociales, la production d'une segmentation, d'un index ou d'une classification n'est jamais l'objectif en soi de la recherche. Chaque étape de la chaîne est l'objet d'une interprétation en fonction du cadre théorique d'analyse du chercheur. De la constitution du corpus à la classification, chaque étape nécessite des choix d'opérationnalisations, c'est-à-dire des choix d'algorithmes compatibles avec les objectifs de recherche.

[31] D'autres organisations selon d'autres types de structures sont possibles, comme les treillis de Gallois (Priss 2006; Wille 1992).

CHAPITRE IV

ANALYSE CRITIQUE DU FORAGE DE TEXTE POUR L'ÉTUDE DES REPRÉSENTATIONS SOCIALES : LE CAS ALCESTE

Introduction

L'objectif général de cette recherche est l'évaluation des méthodes de forage de texte pour l'étude des RS, plus particulièrement pour le repérage et l'analyse du noyau central. Dans le deuxième chapitre ont été présentés les principaux concepts de la théorie des RS et dans le troisième les principales étapes et opérations en jeu dans une méthode générale de forage de texte.

L'objectif spécifique de ce chapitre est d'évaluer dans un premier temps la manière dont le forage de texte a été utilisé jusqu'à maintenant pour l'étude des RS : c'est-à-dire comment chacune des étapes et opérations ont été interprétées en fonction du cadre théorique des RS. Il sera démontré que chaque étape et opération est interprétée de manière originale et implique par conséquent des choix d'opérationnalisations adaptés, c'est-à-dire des algorithmes compatibles avec le cadre théorique des RS.

L'utilisation du forage de texte pour l'étude des RS a déjà donné des premiers résultats intéressants et encourageants. Ces études ont été faites via le logiciel ALCESTE, qui est une opérationnalisation particulière d'une méthode de forage de texte. Il s'agit du logiciel majoritairement utilisé par la communauté des psychosociologues. Dans la deuxième section de ce chapitre, la pertinence et la portée des choix d'opérationnalisations du logiciel ALCESTE seront évaluées.

1. Interprétation et choix d'opérationnalisation

Une partie importante du travail méthodologique du chercheur consiste précisément à définir l'interprétation que peut proposer un cadre théorique pour chaque étape de forage de texte, et de sélectionner les algorithmes qui lui sont les plus pertinents. L'interprétation que propose la théorie des RS du forage de textes se déploie autour de questions telles : que représente le texte pour le psychosociologue? Comment sélectionner dans les textes formant un corpus, les données qui sont spécifiques à l'expression d'une seule RS? Quelles sont les propriétés, ou les unités linguistiques des textes, susceptibles de caractériser une RS? Comment repérer dans les textes les éléments cognitifs formant une RS? Comment identifier les structures de la représentation et la centralité de ses éléments?

1.1 Le texte comme trace sémiotico-empirique d'un travail cognitif

La constitution du corpus dépend de ce que les textes sélectionnés par le chercheur ou la chercheuse représentent, compte tenu de son cadre théorique et de ses objectifs de recherche. C'est-à-dire que l'on doit identifier de *quoi* le texte est considéré comme étant une trace sémiotico-empirique. Le texte est-il considéré comme un indice à propos de l'idéologie du groupe, ses rapports de pourvoir, ses institutions, le style littéraire de l'époque ou encore la maîtrise de la langue?

Dans le cas de l'étude des RS, le texte est considéré par le chercheur comme la trace sémiotico-empirique du travail cognitif des membres d'un groupe ou d'une population déterminée. Il est le « terrain » à partir duquel on va ensuite recueillir des données pertinentes pour l'analyse. Le corpus de textes a un statut analogue à l'entretien, l'observation, le questionnaire, etc., bien qu'il a évidemment ses particularités.

1.2 Cibler les données empiriques spécifiques à la représentation sociale étudiée

Le choix du DOMIF correspond au type de données que l'on veut extraire du corpus : phrase, énoncé, paragraphe, ou autres, et sont les traces sémiotico-empiriques de l'objet d'étude La somme des $domifs_{(i)}$ extraits du corpus correspond à la somme des données recueillies. L'étape du choix du DOMIF et de la segmentation est interprétée comme étant le travail de

cueillette des données pertinentes pour l'objet étudié. Pour les psychosociologues, cette étape nécessite des algorithmes qui permettent d'extraire des segments de texte du corpus qui peuvent être considérés comme les traces sémiotico-empiriques du travail cognitif à propos de la RS qu'ils veulent étudier.

De plus, le critère d'exhaustivité est ici important. On sait en effet qu'une RS est toujours actualisée de manière partielle, due à l'effet de contexte. Par conséquent, une instance de DOMIF n'est jamais suffisante à elle seule pour décrire une RS. Comme l'a explicité Lalhou, chaque $domif_{(i)}$ extrait d'un corpus doit être considéré comme l'expression idiosyncrasique, partielle et contextuelle d'une RS par un individu :

> [...] on suppose que les différentes descriptions empiriques [autrement dit, les $domifs_{(i)}$] sont des formulations incomplètes, et éventuellement partiellement inexactes, d'un type idéal constitué d'un assemblage de traits, déduits par induction des combinaisons descriptives observées. (Lalhou 1995b : 145)

Chaque instance est considérée comme un point de vue sur la RS, qui est fonction d'une position sociale, d'un travail de consonance, d'ancrage et d'objectivation de la part d'un individu. Chaque $domif_{(i)}$ issu du corpus est considéré comme un lieu d'expression d'une RS, mais qui ne dévoile qu'un ou quelques-uns de ses aspects. Chaque instance d'un DOMIF est l'expression d'un ou de quelques éléments cognitifs susceptibles de composer la RS étudiée. Les chercheurs doivent donc extraire du corpus le plus de $domifs_{(i)}$ possible afin de neutraliser le biais de l'effet de contexte.

Par ailleurs, cette étape implique également un travail de triage, ou de filtrage, des données textuelles dans le corpus qui ne relève pas de la RS que l'on veut étudier. C'est-à-dire que tout le texte formant un corpus n'est pas nécessairement lié à la RS étudiée. C'est le cas, par exemple, d'un corpus d'articles de journaux. Un même article peut être la trace sémiotico-empirique d'un travail cognitif à propos de plusieurs RS. Par exemple, on pourrait facilement retrouver dans un tel document un premier paragraphe qui traiterait de la violence dans le sport et un dernier à propos des accommodements raisonnables.

Dans l'analyse de contenu classique, on procède généralement par segmentation et triage manuel des textes du corpus, selon l'évaluation qualitative qu'en fait le chercheur. Le triage

est fait en fonction d'un concept, d'un thème, d'une catégorie, d'une représentation sociale ou tout autre pivot pertinent. Traduit dans le cadre d'une méthode comme le forage de texte, ceci exige des algorithmes qui permettront aux chercheurs d'extraire d'un corpus *seulement* les paragraphes, énoncés, lignes, ou mots qui sont les expressions spécifiques de la RS étudiée.

Meunier et Forest (2009) ont abordé ce problème en distinguant la *lecture et l'analyse de texte assisté par ordinateur* (LATAO) de la *lecture et l'analyse conceptuelle de texte assistée par ordinateur* (LACTAO). Alors que la LATAO est une analyse qui porte sur l'ensemble des données textuelles formant un corpus, la LACTAO tente de cibler, dans un corpus, l'analyse sur les données textuelles spécifiques à un seul concept particulier. La différence entre les deux approches est notamment dans l'échelle d'analyse. La LATAO est une analyse macrotextuelle et la LACTAO une analyse microtextuelle. Alors que la LATAO porte sur l'ensemble des thèmes, concepts représentations ou catégories présents dans un corpus, la LACTAO est une analyse fine des différentes dimensions d'un objet particulier. L'étude des RS via du forage de texte nécessite une approche LACTAO.

1.3 Le type d'information susceptible d'exprimer les représentations sociales

Les $unifs_{(i)}$ sont les propriétés retenues par le chercheur pour caractériser les $domifs_{(i)}$. Dans le cadre d'une analyse des RS, une fois sélectionné un type de DOMIF et l'extraction faite de ses instances, les psychosociologues doivent utiliser des algorithmes qui leur permettront de sélectionner les propriétés de ces $domifs_{(i)}$ pertinents pour la théorie des RS.

Ces propriétés sont généralement les mots. C'est par les mots que les psychosociologues étudient les RS. Dans le chapitre 2, il a été démontré que l'organisation des mots est le moyen par lequel ces chercheurs accèdent à l'organisation de la connaissance.

C'est le cas notamment lorsque l'on procède par des méthodes *associatives*. À un inducteur donné, des sujets associent des induits. L'organisation des associations entre inducteur et induits est le « truchement », c'est-à-dire l'indicateur par lequel on accède indirectement à l'organisation des éléments cognitifs de la RS. Flament et Rouquette insistent constamment sur ce point pour une analyse structurale des RS : « les contenus ne peuvent être autre chose

que des instanciations des catégories, et le discours n'est que l'un des truchements de la cognition [...] » (Flament et Rouquette 2003 : 74)

Interprété dans le cadre du forage de texte, chaque $domif_{(i)}$ est un inducteur, c'est-à-dire une instanciation d'un élément cognitif d'une RS et les propriétés de chacun de ces éléments sont caractériser par une distribution de lexèmes.

1.4 Le poids des mots

L'étape de la vectorisation consiste à évaluer le poids de chaque $unif_{(i)}$ pour chaque $domif_{(i)}$. Dans le cadre de la théorie des RS, elle peut être interprétée de la manière suivante : les $unifs_{(i)}$ sont les mots dans un segment de texte utiliser par des individus pour caractériser une RS. Les mots dans chacun de ces segments ne sont pas tous de même importance : pour des raisons statistiques et linguistiques certains sont plus révélateurs des éléments cognitifs formant la RS.

Dans le cadre d'une méthode associative par exemple, les mots induits en premier et le plus souvent par les sujets sont généralement considérés comme plus révélateurs ou plus représentatifs de la RS. Dans le cadre d'une méthode de forage de texte, on utilise différentes techniques de filtrage des mots. Statistiquement, les mots utilisés par tous ne sont pas discriminants, tandis que ceux qui sont trop idiosyncrasiques ne permettent pas la comparaison et sont donc filtrés. Linguistiquement, les mots sémantiquement significatifs, aussi appelés « mots pleins », comme les noms, les adjectifs, quelques verbes et adverbes sont plus pertinents que les pronoms ou les articles. Les psychosociologues doivent appliquer un filtrage à l'aide d'un antidictionnaire.

1.5 Cartographier la représentation sociale dans les textes

Nous avons vu que la classification permet de regrouper les $domifs_{(i)}$ issus d'un corpus qui partagent des propriétés communes, c'est-à-dire qui partagent une distribution informationnelle similaire. Dans le cadre de la théorie des RS, la classification est interprétée comme une opération de *cartographie* des traces sémiotico-empiriques à propos de la RS étudiée. C'est-à-dire qu'elle permet de regrouper les différents segments de texte, qui sont

susceptibles d'exprimer un même aspect ou un même élément de la RS étudiée. Selon Lalhou : « En d'autres termes, nous cherchons à repérer dans le discours des classes d'énoncés analogues, qui peuvent être considérées comme des expressions d'un même noyau de sens. » (Lalhou 1995b : 139-140)

On interprète les différents $domifs_{(i)}$ d'une classe comme étant des traces sémiotico-empiriques contextuelles ou des « points de vue partiels », mais complémentaires, sur un ou quelques éléments cognitifs de la RS étudiée :

> Pour cela, l'analyste va considérer [les $domifs_{(i)}$] comme une succession de vues partielles du Monde, et supposer que l'ensemble de ces vues renvoie à un ensemble plus petit d'idées. Donc, que plusieurs vues partielles renvoient à une même idée. (Lalhou 1995a : 3)

Autrement dit, on considère chaque classe obtenue comme un répertoire de traces empiriques suffisamment similaires entre elles pour être considérées comme les expressions textuelles d'un même élément cognitif de la RS étudiée : « On considère les classes obtenues comme étant les éléments de base constitutifs de la représentation sociale [...] » (Lalhou 2003 : 46)

Dans le cadre d'analyse des RS, les algorithmes de classification utilisés doivent permettre de classer les $domifs_{(i)}$ de texte en fonction des éléments cognitifs qui leur sont communs. La classification doit permettre de produire autant de classe qu'il y a d'éléments cognitifs importants composant la RS étudiée. Ceci afin de permettre une sorte de cartographie des éléments cognitifs formant la RS étudiée.

En somme, chaque étape d'une méthode de forage de texte peut être mise en correspondance avec la cadre théorique d'analyse des RS. Ceci demande cependant des choix d'opérationnalisations spécifiques, afin de rendre le forage de texte compatible avec la théorie des RS.

2. ALCESTE et les limites du forage de texte pour l'étude des représentations sociales

Une méthode de forage de texte est une chaîne de traitement dans laquelle les chercheurs doivent faire des choix d'opérationnalisations afin de la rendre compatible avec la théorie des RS et les objectifs empiriques de recherche. Cette réflexion nécessaire sur ces choix

d'opérationnalisations a pourtant fait l'objet de peu de travaux dans le domaine d'étude des RS. ALCESTE reste, pour l'instant, la seule option explorée.

Nous sommes en partie d'accord avec l'affirmation suivante de Buschini et Kalampalikis qui affirment que : « Le problème [...] est de savoir s'il y a une adéquation parfaite des logiciels lexicométriques à l'étude des représentations sociales », et poursuivent en ajoutant que : « Sans nier leur utilité et leur apport certain, [ils pensent] néanmoins que ce n'est pas le cas. » (Buschini et Kalampalikis 2002 : 190)

Nous précisons que ce n'est pas la méthode de forage de texte dans son ensemble qui rencontre des limites insurmontables, mais des opérationnalisations particulières de celle-ci. D'ailleurs, lorsque ces auteurs font référence à des « logiciels lexicométriques » ils parlent implicitement d'ALCESTE.

En outre, nous sommes d'accord que l'opérationnalisation qu'offre ALCESTE rencontre des limites importantes. Nous présentons dans cette section les choix d'opérationnalisations offerts par le logiciel ALCESTE. Ensuite, nous abordons trois limites importantes d'ALCESTE pour le repérage et l'analyse du noyau central d'une RS dans un corpus d'article de presse.

2.1 L'opérationnalisation proposée par ALCESTE

Bien qu'il soit toujours possible d'interpréter de plusieurs manières les résultats offerts par ALCESTE, ceux-ci restent relativement confinés dans un même horizon. Cet horizon est déterminé en amont. Le chercheur (utilisateur du logiciel) reste tributaire des choix d'algorithmes par le concepteur. Indirectement, il reste également tributaire du cadre théorique du concepteur qui a motivé ces choix d'algorithmes. Dans le cas d'ALCESTE, ce cadre théorique est celui des « mondes lexicaux » (Reinert 1993, 1997).[32]

---

[32] Une recherche en-soi, qui n'est pas la nôtre ici, consisterait d'ailleurs à évaluer la compatibilité de la théorie des mondes lexicaux de Reinert avec celle des RS. Ce fut en partie l'objectif de Lalhou dans sa thèse de doctorat (1995b).

Les choix d'opérationnalisations dans ALCESTE sont nombreux, certains laissent plus de flexibilité que d'autres. Nous en présentons rapidement quelques-uns, notamment ceux qui présentent certaines limites.[33]

Le DOMIF retenu dans le logiciel ALCESTE est ce qu'on appelle une « unité de contexte élémentaire » (UCE). Une UCE correspond, nous dit Reinert, à l'idée de phrase ou d'énoncé. Il s'agit d'un segment de texte, circonscrit par une ponctuation, d'un maximum d'environ 200 mots (qui varie selon les versions du logiciel). Un UCE peut également correspondre à l'idée du paragraphe circonscrit par le retour de chariot et à quelques alternatives supplémentaires circonscrites par un marqueur indiqué manuellement par le chercheur-utilisateur.

Le type d'UNIF choisi par Alceste est le lexème. Chaque $domif_{(i)}$ est décrit par la distribution de ces lexèmes.[34] La pondération pour la vectorisation se fait exclusivement par une valeur binaire : présence ou absence d'une $unif_{(i)}$ dans un $domif_{(i)}$. La réduction des dimensions de la matrice procède à l'aide de stratégies classiques de lemmatisation, de stemmatisation et un filtrage par des antidictionnaires de mots fonctionnels et des hapax.

La classification utilise un algorithme de classification descendante hiérarchique. La meilleure manière de comprendre la logique de cet algorithme est à l'aide de sa représentation en dendrogramme. La procédure de classification commence à partir d'une classe « parent », qui contient la totalité des $domifs_{(i)}$ extraits du corpus. Une première itération consiste à diviser cette classe « parent » en deux sous-classes « enfant » les plus différentes possible.

Une deuxième itération consiste ensuite à prendre, parmi les deux nouvelles sous-classes « enfant », celle qui est la plus grande (en nombre de $domifs_{(i)}$) et la diviser à nouveau en deux nouvelles sous-classes « enfant ». On répète l'itération jusqu'à ce que l'une des trois conditions suivantes soit atteinte : soit que l'on a autant de classes que de $domifs_{(i)}$, donc

---

[33] Reinert a présenté à plusieurs endroits son logiciel, notamment (Reinert 1993, 1997, 2008).

[34] Il existe ici une certaine flexibilité, notamment dans la possibilité de faire de la troncature pour obtenir des mot-composés.

aucune classe supplémentaire à diviser; soit que l'on a effectué les 15 itérations permises par le logiciel; ou soit que la division d'une classe en deux nouvelles sous-classes « enfant » réduit la qualité de la classification, comprise en termes d'optimalité à partir d'un coefficient d'association de chi².

L'exemple de la figure 6 montre une classification descendante hiérarchique fictive en sept itérations sur 16 segments de texte.

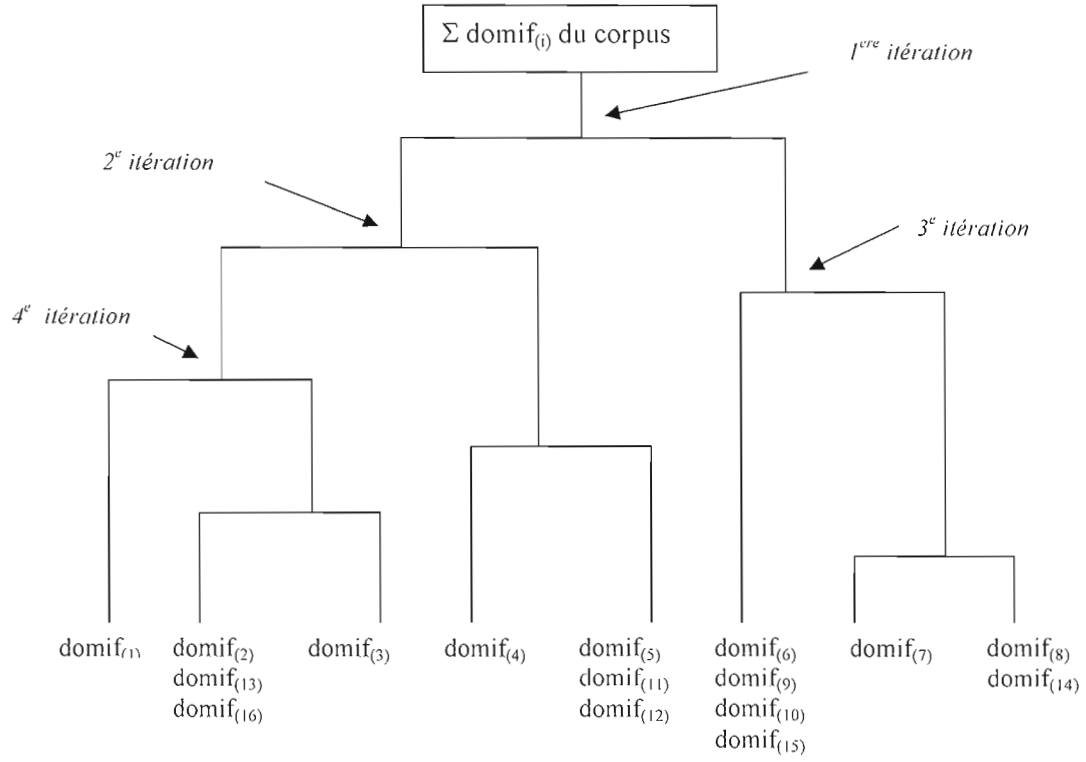


Figure 6 : Dendogramme représentant une classification descendante hiérarchique

Nous ne discutons pas du détail du calcul.[35] Notons simplement que la division d'une classe « parent » en deux sous-classes « enfant » se fait sur la base du coefficient d'association du chi². La procédure peut se résumer de la manière suivante : pour chaque classe « parent », on produit les *n* divisions possibles en deux sous-classes « enfant » ($n = 2^{d-1} - 1$, où *d* est le

[35] Les manuels d'utilisation du logiciel offrent ce genre d'information.

nombre de $domifs_{(i)}$ dans la classe « parent »). Pour chaque *n* divisions candidates, on calcul le chi^2 de chaque $domif_{(i)}$ associé à une classe « enfant ». On retient, parmi les *n* divisions candidates, celle qui maximise la somme des chi^2. L'objectif est d'obtenir à la fin un nombre de classes dans lesquelles les $domifs_{(i)}$ ont les plus hautes probabilités d'appartenance.

2.2 Limites du forage de texte en général et d'ALCESTE en particulier

L'utilisation d'ALCESTE pour l'étude des RS comporte de nombreux avantages, dont le plus important est certainement la fiabilité du logiciel (qui est rendu à sa 5^e version). Son utilisation est facile et assistée et la plupart des opérations sont transparentes pour l'utilisateur. De plus, il a été appliqué avec succès à une grande variété de domaines.

Cependant, pour l'étude des RS dans un corpus d'articles de presse, il rencontre certaines difficultés qui concernent la compatibilité des algorithmes sélectionnés avec les objectifs d'une analyse des RS orientée vers le repérage de son noyau central. Deux de ces limites sont particulièrement importantes dans ALCESTE : le type de DOMIF retenu et l'algorithme de classification. Une troisième limite concerne de manière plus générale la méthode que nous avons présentée précédemment, mais dont ALCESTE est également tributaire. Cette troisième limite est l'insuffisance des résultats d'une classification. Autrement dit, la cartographie des éléments composant la RS étudiée n'est pas suffisante pour un cadre théorique structural des RS. Cette opération de cartographie permet seulement d'atteindre le premier niveau d'analyse des RS, celui de l'identification des éléments cognitifs d'une RS. Or, une analyse des RS implique également le repérage des structures et du noyau central de la RS.

2.2.1 L'hétérogénéité du domaine d'information

La première limite que nous soulignons est le type de DOMIF retenu par ALCESTE : l'unité de contexte élémentaire (UCE). Les $domifs_{(i)}$ extraits à partir de ce type de DOMIF sont trop hétérogènes entre eux et ne permettent pas, dans un corpus d'articles de journaux, de cibler les traces sémiotico-empiriques spécifiques à une RS en particulier.

À titre d'illustration, supposons que notre corpus soit constitué d'un seul article de presse, par exemple celui-ci, tiré de La Presse du 1er décembre 2007, par Martin Croteau :

$domif_{(1)}$. La scène est disgracieuse. Des joueurs de hockey qui se tapent dessus au centre de la patinoire. Des coéquipiers qui sautent la clôture pour se joindre à la mêlée. Les entraîneurs qui les regardent franchir la bande sans faire le moindre geste pour les en empêcher. Ces images, diffusées cette semaine par la chaîne The Score, ont fait le tour du Canada. Car les bagarreurs ont 8 ans. Tout a commencé lors d'un match novice disputé à l'occasion du 36e Guelph Powerplay Tournament, une compétition annuelle qui rassemble 125 équipes de partout en Ontario. Les Duffield Devils de North York venaient d'écraser le Thunder de Niagara Falls par la marque de 8-1. La tension a grimpé à la fin de la rencontre lorsqu'un joueur de Niagara Falls a terrassé un adversaire avec une mise en échec. La foule, composée de parents, applaudit le geste. À la sirène, des joueurs commencent à se pousser. Avant longtemps, un joueur des Devils s'accroche à un adversaire, le projette sur la glace et commence à le rouer de coups de poing. Les autres se jettent dans la mêlée. Les coups fusent de partout.

$domif_{(2)}$: En arrière-plan, on aperçoit un joueur des Devils étendu sur la glace. Il vient de sauter la bande et il a perdu pied. Sitôt relevé, il échange un regard avec son entraîneur puis patine à toute vitesse pour se joindre à la bagarre. Les arbitres, à peu près deux fois plus grands que les pugilistes, se tiennent debout au centre de la mêlée, ne sachant où donner de la tête. Un autre joueur des Devils tente de franchir la clôture. C'est alors que l'entraîneur adjoint de Niagara Falls intervient et pousse le gamin sur son banc. Lorsque les enfants cessent enfin de se chamailler, c'est au tour des adultes de jeter de l'huile sur le feu. Les entraîneurs se crient des injures. Une mère quitte les gradins pour venir frapper l'un d'eux. Pendant cette nouvelle altercation, l'entraîneur adjoint du Thunder, Randy Brant, aurait craché au visage de l'entraîneur des Devils. Le lendemain, l'Association de hockey mineur de Niagara Falls a suspendu les entraîneurs du Thunder et s'est excusée pour la conduite des jeunes joueurs. "Il est clair que les joueurs de Duffield étaient plus nombreux que ceux de Niagara Falls lorsque la bagarre a éclaté, a-t-elle toutefois écrit dans son communiqué. Il est évident que les joueurs ne faisaient que se défendre. " L'organisation des Devils n'a émis aucun commentaire à la suite de l'incident.

$domif_{(3)}$: La police de Guelph a ouvert une enquête pour déterminer si le comportement des entraîneurs était criminel. Mais jeudi, elle a annoncé qu'aucune accusation ne serait portée contre eux. L'Association ontarienne de hockey mineur pourrait également imposer des sanctions. Chose certaine, cette bagarre vaut un nouvel œil au beurre noir à notre sport national. En 2004, l'attaquant Todd Bertuzzi, des Canucks de Vancouver, avait sauvagement frappé un adversaire par derrière. Steve Moore, de l'Avalanche du Colorado, a dû prendre sa retraite à cause de ses blessures au cou. Dans une entrevue avec La Presse Canadienne, Émile Thérien, père d'un joueur professionnel et ancien président du Conseil du Canada pour la sécurité, s'est indigné de voir des enfants si jeunes jouer dans un environnement aussi compétitif. "Cet incident empeste de tout ce qui cloche avec le hockey dans ce pays", a-t-il asséné. M. Thérien estime que seuls les enfants âgés de plus de 12 ans devraient être autorisés à participer à des tournois compétitifs.

> $domif_{(4)}$: Toujours dans le sport, un incident survenu récemment à Calgary montre que le débat sur les accommodements raisonnables dépasse largement les frontières du Québec. Une adolescente de 14 ans a été expulsée d'un tournoi de soccer pour avoir porté un hidjab. Safaa Menhem est arrivée en retard pour le match qui opposait son équipe au Chinook Phantom, samedi dernier. À son arrivée sur le terrain, l'arbitre a averti l'entraîneur que le foulard de la jeune fille n'était pas conforme aux règlements de sécurité. Elle avait pourtant participé à cinq autres matchs avec son hidjab sans recevoir d'avertissement.
>
> $domif_{(5)}$: L'Association de soccer mineur de Calgary et le Conseil musulman local se sont plaints de cette décision, affirmant qu'elle nuira à l'accessibilité du soccer, de loin le sport le plus pratiqué au pays. Mais contrairement à ce que l'on observe au Québec, le gouvernement albertain a choisi de se mouiller. Cette semaine, le ministre du Loisir, Hector Goudreau, s'est prononcé en faveur de l'expulsion de la jeune joueuse. "À long terme, cependant, il faudra étudier l'impact de cette décision, a-t-il indiqué au Calgary Herald. Il faudra voir comment on peut trouver un accommodement. "

Dans cet article, ALCESTE nous permettrait d'extraire toutes les instances d'un DOMIF qui correspond, par exemple, au paragraphe. Nous aurions ici cinq segments, ou UCE, de texte. Si ce que nous voulons analyser c'est le contenu de l'article dans sa totalité, une segmentation de la sorte est une bonne option d'opérationnalisation. Une segmentation par phrase serait également intéressante à explorer. Ceci correspond à une stratégie LATAO classique définie précédemment.

Par contre, si notre objectif est d'analyser une RS spécifique dans ce corpus, il est probable que ce ne soit pas la totalité du corpus qui soit pertinent à retenir. Par exemple, si nous nous intéressons à la RS des accommodements raisonnables, comme ce sera le cas dans le dernier chapitre, les $domif_{(1)}$, $domif_{(2)}$, et $domif_{(3)}$ sont peu pertinents pour comprendre cette RS, on y traite plutôt de la violence dans le hockey chez les mineurs.

Or, nous avons vu que l'homogénéité du corpus par rapport à l'objet d'étude est une condition jugée importante par l'analyse de contenu, pour assurer la pertinence de l'analyse. De plus, dans un cadre de recherche comme celui des RS, le choix du DOMIF correspond à l'étape de la cueillette des données sur la RS. Cette étape est jugée cruciale :

> L'étude des représentations sociales pose deux problèmes méthodologiques redoutables : celui du recueil des représentations et celui de l'analyse des données obtenue. [...] Mais en amont de l'analyse des données, *la méthodologie du recueil apparaît comme un point clé déterminant prioritairement la valeur des études sur les représentations* [nous soulignons]. Quels que soient l'intérêt et la puissance d'une

> méthode d'analyse, il est bien évident que le type d'informations recueillies, leur qualité et leur pertinence déterminent directement la validité des résultats obtenus et des analyses réalisées. Dès lors, la première question qui se pose au chercheur sur les représentations sociales concerne les outils qu'il va choisir et utiliser pour appréhender son objet. (Abric 1994b : 59)

En transposant ce commentaire pour le forage de texte, le choix du DOMIF détermine prioritairement la pertinence de la classification que nous obtiendrons. Autrement dit, plus les $domifs_{(i)}$ sont étrangers à la RS étudiée, plus la cartographie des éléments cognitifs qui la composent sera également étrangère et non pertinente.

L'opérationnalisation de cette étape doit offrir aux psychosociologues des algorithmes qui leurs permettront d'extraire les $domifs_{(i)}$ les plus représentatifs de la RS étudiée. Il faut par conséquent adopter une stratégie LACTAO, ce que ne permet pas ALCESTE.

2.2.2 Les limites de la classification descendante hiérarchique

La deuxième limite du forage de texte pour l'étude des RS via ALCESTE est l'utilisation de la classification descendante hiérarchique pour la cartographie des éléments cognitifs de la RS étudiée. La limite que nous soulignons est que ce type de classification ne permet pas de repérer un nombre suffisamment grand et différencié de classes, susceptibles de correspondre aux éléments cognitifs de la RS étudiée.

Cette limite est causée par les choix d'opérationnalisations dans ALCESTE. En effet, dans ce logiciel, le nombre de classes que l'on obtient dépend essentiellement de trois conditions : soit que l'on a obtenu autant de classes qu'il a de $domifs_{(i)}$ issus du corpus; soit que l'on a atteint les 15 itérations permises; ou soit que l'on a obtenu une classification optimale (à partir du chi^2).

À moins d'appliquer la classification sur un corpus très petit et très hétérogène, la première condition n'est jamais atteinte. La deuxième cependant est une limite importante, car 15 itérations dans le cadre d'une classification descendante hiérarchique permettent un maximum de $n+1$ classes (où n = le nombre d'itérations), c'est-à-dire un maximum de 16 classes. Lorsque nous interprétons cette étape de classification comme une opération de cartographie des éléments cognitifs formant la RS étudiée, ce choix d'opérationnalisation

dans ALCESTE contraint à priori le repérage d'un maximum de 16 éléments ou moins. Ce nombre arbitraire relativement petit est gênant pour un cadre théorique comme celui des RS. En effet, nous avons vu, dans le chapitre 2, que les psychosociologues retiennent pour leur analyse généralement entre 20 et 60 éléments.

De plus, si le nombre maximal d'itérations dans ALCESTE est fixé ainsi c'est qu'en pratique, les 15 itérations permises sont rarement atteintes.[36] Dans la pratique, le nombre de classes que l'on obtient varie habituellement entre 3 et 9, ce qui est encore plus problématique pour la théorie des RS. Cette limite empirique est d'ailleurs causée par une bonne raison : le calcul d'optimisation des classes à l'aide du chi² fait en sorte que, généralement, plus le nombre de classes augmente, plus la somme des chi² diminue, donc plus les classes sont considérées de mauvaise qualité.

D'ailleurs, pour ces raisons et pour d'autres,[37] liées en général aux mauvais résultats qu'elle génère, la classification descendante hiérarchique n'est pratiquement jamais utilisée dans le domaine du forage de texte : « In practice the divisive procedure [la classification descendante hiérarchique] is almost of no importance due to its generally bad result » (Hotho *et al.* 2005 : 18)

De plus, il semble que la classification descendante hiérarchique est surtout adaptée à la classification de petit corpus, c'est-à-dire qui contient un petit nombre de $domifs_{(i)}$ (Weiss *et*

---

[36] Nous n'avons recensé aucune recherche utilisant ALCESTE où la classification descendante hiérarchique aurait produit les 16 classes possibles.

[37] En fait, bien que nous n'abordions pas ce point technique ici, la principale raison invoquée pour le rejet de la classification descendante hiérarchique pour du forage de texte est son *coût computationnel* exorbitant à chaque itération. En effet, tel que nous l'avons rapidement mentionné, à chaque itération, la classification descendante hiérarchique doit calculer les 2^{d-1} -1 divisions possibles d'une classe « parent » en deux sous-classes « enfant ». En fonction d'un critère statistique d'inertie (pour ALCESTE le chi²), elle retient parmi celles-ci la division optimale. Imaginons que l'on veut diviser une classe « parent » de 500 $domifs_{(i)}$: l'algorithme doit calculer les 2^{500-1} -1 divisions possibles pour une *seule* itération. On comprend ainsi qu'ALCESTE se limite à un maximum de 15 itérations!

al. 2005).[38] Or, nous verrons dans le prochain chapitre que nous cherchons à appliquer la classification automatique sur plusieurs centaines de $domifs_{(i)}$.

Ces limites d'ALCESTE se manifestent, notamment, dans les travaux de Lalhou (Lalhou 1995a; 1995b; 1998; 2003). Pour les résumer rapidement, cet auteur a cherché à repérer, à partir du dictionnaire Le Grand Robert, la RS de *l'acte de manger*. À partir du dictionnaire, il a extrait plusieurs milliers de $domifs_{(i)}$ correspondant à 544 définitions liées à la définition de *manger*. Il a utilisé ALCESTE afin de produire une classification, qu'il interprète comme une cartographie des éléments cognitifs formant la RS étudiée. ALCESTE a produit 6 classes, que l'auteur a interprétées comme correspondant aux éléments DÉSIR, VIVRE, PRENDRE, REMPLIR, NOURRITURE, REPAS.

L'une des conséquences possibles d'une mauvaise classification pour l'identification des éléments cognitifs formant la RS étudiée est de se retrouver avec des classes très grandes, de plusieurs centaines de $domifs_{(i)}$. Le risque possible dans ce cas est d'avoir des classes dans lesquelles les $domifs_{(i)}$ sont trop hétérogènes et parmi lesquelles on retrouve difficilement le ou les quelques éléments cognitifs saillants qui leur sont communs. Selon nous, c'est précisément ce qui est arrivé aux recherches de Lalhou, qui a dû se contenter d'identifier des éléments cognitifs si généraux qu'ils ont été considérés par l'auteur lui-même comme un peu triviaux (Lalhou 2003 : 52).

2.2.3 Les limites quant à l'identification du noyau central

La troisième limite que nous soulignons est l'insuffisance des résultats d'une classification pour les objectifs d'analyse structurale d'une RS. La classification, interprétée comme une opération de cartographie des éléments cognitifs formant une RS, permet d'atteindre seulement l'un des trois niveaux d'analyse que doit remplir une méthode générale d'étude des RS.

---

[38] En fait, le développement des algorithmes de classification non supervisée évolue de manière si rapide qu'une affirmation aussi tranchée que celle de Weiss *et al.* est à nuancer. Voir par exemple Zhao et Karypis (2002, 2005) pour une implémentation de la classification descendante hiérarchique relativement performante sur de grands corpus.

Nous avons vu en effet que le deuxième niveau d'analyse est l'identification des structures organisant les éléments cognitifs de la RS, et que le troisième niveau est l'identification des éléments formant son noyau central. Le cadre de la méthode de forage de texte que nous avons présentée, qui a pour dernière étape la production d'une classification, est, en ce sens, insuffisante et ne permet pas d'atteindre les deux derniers niveaux d'analyse.

À notre connaissance, seule l'étude de Kalampalikis et Moscovici (Kalampalikis et Moscovici 2005) est parvenue partiellement à développer une technique d'analyse à partir de forage de texte, qui permet de déterminer la position structurale des éléments cognitifs identifiés par la classification automatique. Selon ces auteurs, la valeur symbolique des éléments cognitifs d'une RS peut être mesurée à partir de différents indicateurs de redondance, qui traduit leurs degrés d'objectivation. Parmi ces indicateurs, ils utilisent un calcul de diversité lexicale classique en lexicométrie : le ratio type/token. Le ratio type/token mesure la proportion de lexèmes (ou de ses formes réduites) différents par rapport au nombre d'occurrences totales de lexèmes dans le corpus : plus le corpus est diversifié, plus le ratio augmente.

Une mesure de redondance lexicale est la différence 1-(type/token). Plus le nombre de lexèmes différents augmente par rapport au nombre total d'occurrences, plus on considère que le corpus est pauvre, redondant, répétitif et marqué par peu de variation. La redondance lexicale est considérée comme un indicateur des effets de l'objectivation, un processus de la pensée sociale qui consiste à sélectionner, schématiser, simplifier et stabiliser dans le langage l'expression des cognitions centrales d'une RS. Ces cognitions sont par exemple actualisées via des expressions populaires, des maximes et des stéréotypes qui sont des formats langagiers facilitant leur communication et leur diffusion.

Par conséquent, plus un élément cognitif est actualisé dans un langage redondant, plus on le considère objectivé, et plus il est considéré comme candidat potentiel pour composer le noyau central dans la RS. C'est ce qu'ont montré Kalampalikis et Moscovici à propos de la RS du *marxisme*. Ils ont montré que les classes qui ont un contenu lexical redondant regroupent les $domifs_{(i)}$ qui sont l'expression, dans le langage, des éléments cognitifs centraux de la RS.

Ces résultats de recherche sont encourageants quant à notre objectif général d'évaluer la pertinence et la compatibilité du forage de texte pour l'étude des RS. Ils sont cependant insuffisants. Comme l'affirment Flament et Rouquette, la redondance lexicale « n'est pas la preuve suffisante qu'il existe une RS cristallisée dans le groupe considéré, c'est-à-dire une RS possédant un noyau central [...]. Il s'agit en somme d'un indice d'organisation, qu'il est indispensable de corroborer par d'autres. » (Flament et Rouquette 2003 : 62)

Bien que la redondance dans le langage soit un indicateur important sur la centralité des éléments cognitifs d'une RS, ce dernier doit être corroboré par d'autres indicateurs à propos d'autres propriétés de centralité, tels la connexité et le consensus social.[39]

Retour sur le troisième objectif

Dans ce chapitre, l'objectif était d'évaluer la contribution du forage de texte jusqu'à maintenant pour l'étude des RS. Dans une première section a été explicitée la manière dont les étapes d'une méthode de forage de texte sont interprétées dans le cadre théorique d'analyse des RS.

Pour les psychosociologues, les textes formant le corpus étudié sont avant tout considérés comme des traces sémiotico-empiriques à propos du *travail cognitif* des membres d'un groupe. L'étape du choix du DOMIF et de la segmentation correspond à une opération de cueillette des données dans le corpus à propos de la RS que l'on veut étudier. Pour la théorie des RS, un $domif_{(i)}$ correspond à un point de vue sur l'objet étudié, une description partielle et contextuelle de la RS. Il faut aux chercheurs des algorithmes qui leurs permettront d'extraire les $domifs_{(i)}$ spécifiques à l'objet d'étude, en l'occurrence une RS donnée. Le type d'UNIF retenu pour décrire les $domifs_{(i)}$ est le lexème et ceux-ci doivent être filtrés en fonction de leur pertinence. Finalement, la classification automatique est interprétée comme une opération de cartographie des éléments cognitifs composant la RS étudiée. Chaque classe obtenue est considérée comme une classe d'équivalence, d'énoncés ou de phrases analogues, qui sont

[39] ALCESTE permet d'appliquer des analyses factorielles aux classes obtenues, qui sont une manière heuristique de visualiser la structure qui les organise. Cependant, on ne voit pas comment il est possible, à partir des différents plans factoriels, d'identifier la centralité des classes. Le problème reste donc entier.

considérés comme des traces sémiotico-empiriques, ou langagières, d'un même élément cognitif.

Dans la deuxième section, trois limites importantes dans les choix d'opérationnalisations d'ALCESTE et de leur incompatibilité avec la théorie des RS ont été soulignées, dont deux spécifiques à ALCESTE et une troisième qui concerne la chaîne générale de traitement du forage de texte.

Les deux limites spécifiques à ALCESTE sont le type de DOMIF et le type de classification utilisé. Le type de DOMIF utilisé est trop général, pour permettre de cibler les données textuelles spécifiques à *une* RS étudiée. Quant au type de classification, il est mal adapté à une opération de cartographie des éléments formant une RS. La classification descendante hiérarchique génère un nombre trop petit de classes, trop générales et pas assez différenciées entre elles.

La troisième limite est l'insuffisance des résultats de la classification pour une analyse des RS orientée vers le repérage du noyau central. La classification ne permet que d'atteindre le premier niveau d'analyse des RS. Des travaux récents ont montré que l'analyse de la redondance dans le langage peut être un indicateur de centralité des cognitions. Mais ceux-ci doivent être corroborés avec d'autres techniques d'analyse qui tiennent compte d'autres propriétés de centralité.

Limite liée à une utilisation non-modulaire d'une méthode de forage de texte

En forage de texte, une part importante du travail méthodologique du chercheur consiste précisément à choisir les algorithmes les plus appropriés à ses objectifs de recherche et son cadre théorique d'analyse. Ce travail est souvent négligé dans la recherche. La discussion et la justification autour du choix de tel ou tel algorithme sont souvent escamotées. On passe directement à la présentation des résultats. Or, chaque étape de cette méthode de forage de texte peut être réalisée à l'aide de différents algorithmes : on n'utilise pas le même algorithme pour une segmentation par ligne ou une segmentation par paragraphe, de même que ce n'est pas la même chose de classer à l'aide d'un algorithme de classification descendante hiérarchique, un algorithme statistique par partition itérative ou un réseau de neurones

dynamique. Certains algorithmes sont mieux adaptés que d'autres à certains cadres théoriques.

Dans la littérature, cette négligence n'est ni triviale ni fortuite. Elle est symptomatique de l'attitude des chercheurs par rapport à la technologie informatique. Les chercheurs préfèrent la plupart du temps utiliser un logiciel « clé en main », facile d'utilisation et ergonomique. Les scénarios d'utilisation de ces logiciels sont plutôt rigides et fortement assistés, réduisant les égarements ou les tâtonnements possibles. Dans ces logiciels, la plupart des choix d'opérationnalisations ont déjà été faits en amont par le concepteur, ou sont réduits au minimum, et l'utilisateur en reste tributaire.

Dans ces logiciels, la chaîne de traitement est *non-modulaire*, souvent peu manipulable, d'autant plus lorsqu'il s'agit d'un produit propriétaire sans accès au code source : « [...] the technology is often a closed one; it is proprietary, rigid, non-modular and, as often underlined, not fully adapted [...] (Meunier *et al.* 2006 : 129)

L'utilisation de ce genre de logiciels a certes plusieurs avantages. Ce sont habituellement des logiciels commerciaux, connus par la communauté des pairs, robustes et qui ont fait leurs preuves quant à la qualité de leur traitement informatique. Leur simple évocation dans une démonstration peut parfois renforcer la crédibilité de l'analyse des données.

Cependant, le revers de la médaille est la fixité de la chaîne de traitement et le rétrécissement de la méthode à *une version* opérationnalisée parmi plusieurs possibles. Ce rétrécissement peut avoir des conséquences importantes dans la démarche du chercheur : au lieu d'adapter le logiciel au cadre théorique du chercheur, ce dernier peut être contraint d'adapter le cadre théorique au logiciel.

CHAPITRE V

EXPÉRIMENTATION ET VALIDATION DE DIFFÉRENTS CHOIX D'OPÉRATIONNALISATIONS : UNE ILLUSTRATION À PARTIR DE LA REPRÉSENTATION SOCIALE DES ACCOMMODEMENTS RAISONNABLES DANS LA PRESSE QUÉBÉCOISE

Introduction

Dans les chapitres précédents ont été présentés les trois niveaux d'analyse des RS, les principaux concepts du cadre théorique des RS, les étapes d'une chaîne de forage de texte et certaines limites du forage de texte pour l'étude des RS. Ces imites concernes en particulier la nécessité de développer de nouvelles techniques d'analyse des structures et de la centralité des éléments composant une RS.

Dans ce cinquième chapitre est abordé le quatrième objectif de recherche. Celui-ci est le cœur de notre démarche d'évaluation de la pertinence et de la compatibilité du forage de texte pour l'étude des RS. Il sera présenté une opérationnalisation originale de la chaîne de forage de texte mieux adaptée au cadre théorique d'analyse des RS. Nous développons ensuite de nouvelles techniques d'analyse, qui permettent d'atteindre les trois niveaux d'analyse attendus par une méthodologie générale des RS.

Le premier niveau d'analyse est atteint avec des choix d'opérationnalisations spécifiques aux étapes de la sélection du DOMIF et du choix d'un algorithme de classification. Ceci permet de produire une cartographie des éléments cognitifs de la RS étudiée qui ne rencontre pas les inconvénients soulignés à propos du logiciel ALCESTE.

Le deuxième niveau d'analyse est atteint en développant des techniques d'analyse sur les résultats issus de la classification. Ces techniques permettent une modélisation originale des structures cognitives et sociales de la RS étudiée. Une première technique, inspirée de

l'analyse de similitude, permet une modélisation de la structure cognitive via un *réseau cognitif de proximité sémantique*. Une deuxième technique, inspirée de l'analyse de consensus, permet une modélisation de la structure sociale via un *réseau sociocognitif de distribution sociale* des éléments cognitifs d'une RS. Le troisième niveau d'analyse est atteint en développant des techniques de mesure de la centralité des éléments cognitifs dans les différents réseaux.

Dans la dernière section, une analyse comparative est menée sur les techniques de mesure de la centralité développées dans cette recherche avec celle proposée par Kalampalikis et Moscovici (Kalampalikis et Moscovici 2005). Une approche multiméthodologique est aujourd'hui considérée comme indispensable pour l'étude des RS. La stratégie d'Apostolidis (Apostolidis 2003), qui consiste à procéder par triangulation des résultats issus des différentes techniques de mesure de la centralité, sera adoptée. Une corrélation significative entre résultats serait un indicateur encourageant quant à la validité de nos techniques d'analyse et, *a fortiori*, quant à la pertinence et à la compatibilité du forage de texte pour l'étude des RS.

Chaque choix d'opérationnalisations et d'expérimentations est illustré à l'aide d'un cas empirique : la RS des *accommodements raisonnables*[40], telle que conceptualisée dans trois journaux québécois : Le Devoir, La Presse et Le Soleil.

1. Une opérationnalisation alternative du forage de texte pour la cartographie des éléments cognitifs composant une représentation sociale

Dans cette section sont présentés nos choix d'opérationnalisations de forage de texte, c'est-à-dire les algorithmes particuliers retenus pour chaque étape de la chaîne de traitement. L'opérationnalisation que proposée suit les étapes d'une méthode classique de forage de texte. Une attention particulière est consacrée aux étapes 2 et 6, où nos choix

[40] En fait, la question de savoir si les accommodements raisonnables sont effectivement l'objet d'une RS n'est pas considérée ici. Nous prenons pour acquis que c'est le cas à des fins d'expérimentation et d'illustration de notre analyse du forage de texte pour l'étude des RS. Des recherches plus approfondies sont nécessaires pour vérifier cette question, qui est hors du cadre notre recherche. Cependant, nos expérimentations plus bas nous indiquent, indirectement, que nous avons de bonnes raisons de croire que c'est le cas.

d'opérationnalisation sont déterminants. Ces décisions font suite aux limites d'ALCESTE et du forage de texte en général, soulignées au chapitre précédent, concernant l'hétérogénéité des $domifs_{(i)}$ et la classification descendante hiérarchique.

Dans un premier temps, nous présentons le corpus de presse à partir duquel sont menées et illustrées nos expérimentations. Ensuite, nous opérationnalisons l'étape du choix et de l'extraction du DOMIF à l'aide d'un algorithme de *concordance*. Les troisième, quatrième et cinquième étapes, respectivement la sélection de l'UNIF, la vectorisation et la réduction dimensionnelle, sont opérationnalisées par des algorithmes classiques comme un index de lexème, une pondération par fréquence relative, un filtrage par la lemmatisation, par l'utilisation d'un antidictionnaire et par le IDF. La sixième étape de classification est opérationnalisée à l'aide de l'algorithme K-Means, qui permet une cartographie plus fine des éléments composant la RS étudiée. La validation statistique de la qualité de la classification est obtenue à l'aide du coefficient silhouette simplifié.

1.1 Présentation du corpus

Le corpus d'expérimentation est composé en tout de 3434 articles de presse, issus des journaux La Presse, Le Devoir et Le Soleil, soit respectivement 1450 pour le premier, 1192 pour de deuxième et 792 pour le troisième. La période de publication de ces articles de presse est de 20 ans, soit du 14 avril 1988 au 31 janvier 2008. Ils ont été récupérés à partir de la base de données Biblio Branchée (Eureka). La sélection s'est faite à l'aide d'une procédure de recherche par mots-clés. Pour le choix des mots-clés, nous nous sommes basés sur un critère de saillance, c'est-à-dire « la désignation la plus commune de l'objet de représentation dans la population considérée » (Flament et Rouquette 2003 : 81). Ces mots-clés sont les expressions *accommodement* et/ou *accommodements*. Par conséquent, tous les articles des trois journaux contenant au moins une occurrence de l'une de ces expressions ont été retenus pour former notre corpus.

Les expressions *accommodement* et *accommodements* (pour la suite, *accommodement(s)*) apparaissent pour la première fois dans ces journaux en 1988. Mais c'est surtout à partir de 1993 que le nombre d'articles dans lesquels nous les retrouvons augmente. Ce nombre reste relativement stable jusqu'à la fin de 2006 à partir de quand il va se décupler. Le graphique de

la figure 7 illustre le nombre d'articles par année formant notre corpus. La série bleue est le nombre d'articles par année cumulé pour les trois journaux; la série jaune est leur moyenne.

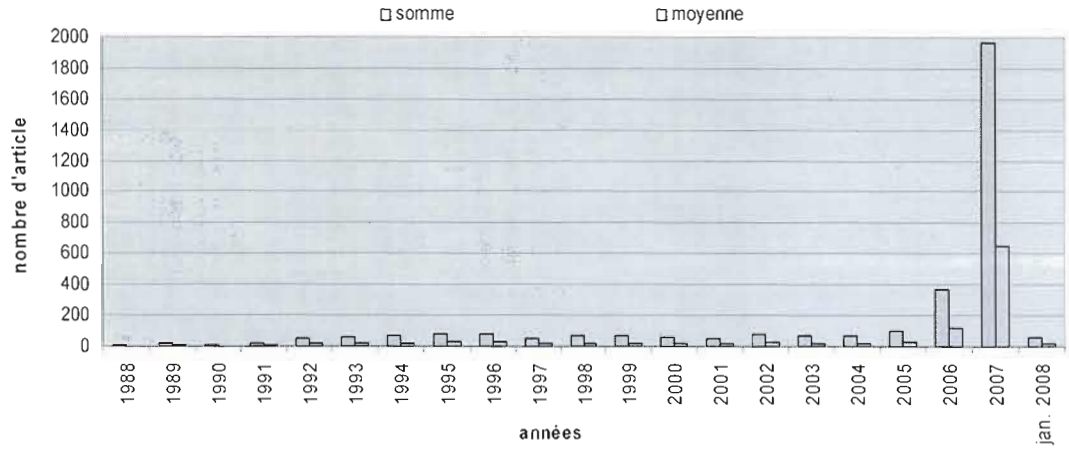


Figure 7 : Distribution du nombre d'articles par année

Une première recherche peut porter sur l'ensemble de ces 3434 articles de presse, afin de repérer et d'analyser sur ces trois journaux pris ensemble, la RS des *accommodements raisonnables* (AR). Bien que cette possibilité puisse être intéressante pour une première approche exploratoire, cela génère des résultats très généraux, difficiles à interpréter pour la théorie des RS (Chartier *et al.* 2008). Par conséquent, il est intéressant de découper ce corpus selon des groupes et des périodes significatives.

1.1.1 Trois périodes de saillance et trois communautés épistémiques

Selon la théorie des RS, la fréquence d'utilisation d'éléments linguistiques, notamment lexicaux, peut être vue comme un indicateur de la saillance des éléments cognitifs associés à ces expressions. Ainsi, on peut voir la fréquence d'utilisation des expressions *accommodement(s)* comme un indicateur de la saillance des éléments cognitifs qui lui sont associés. Suivre l'évolution de cet indicateur de saillance permet de repérer des périodes dans l'évolution du travail cognitif à propos des AR.

Le graphique de la figure 8 montre la fréquence d'utilisation de l'expression *accommodement(s)* pour les trois journaux cumulés et leur moyenne.

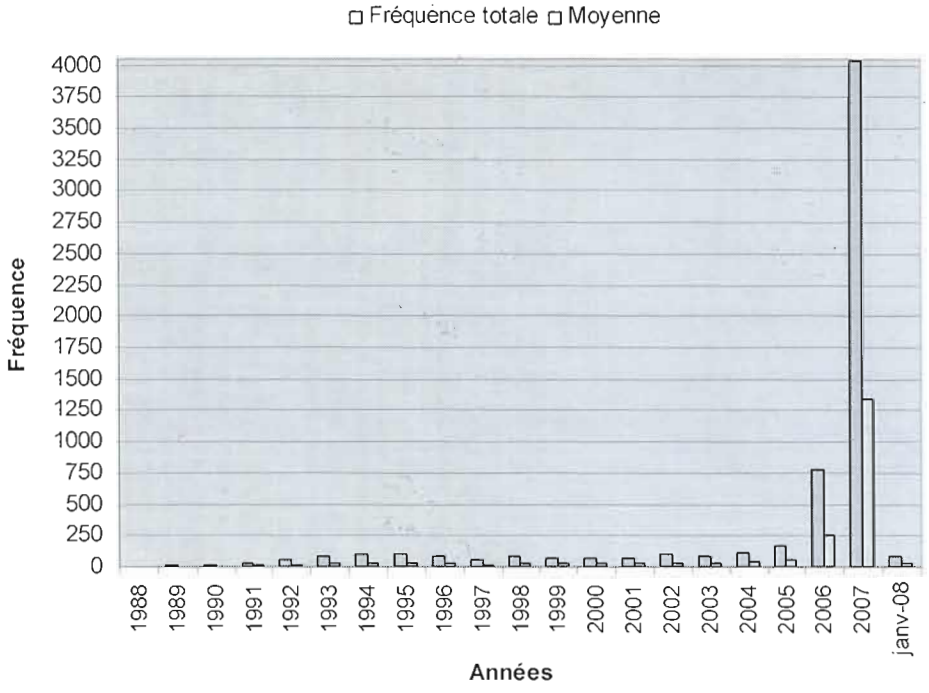


Figure 8 : Fréquence des expressions *accommodement(s)* par année

Un premier regard sur cette distribution indique au moins deux périodes dans l'évolution de l'utilisation de l'expression *accommodement(s)* : une première, qui commence autour de 1988, mais surtout à partir de 1993; et une deuxième, qui commence en 2006.

En changeant l'échelle d'observation, tel qu'illustré dans le graphique de la figure 9, nous remarquons que l'année 2007 est elle-même marquée par deux périodes : l'une commençant fin 2006 et qui se termine en juillet 2007; l'autre commençant en août 2007 et qui se termine en janvier 2008.

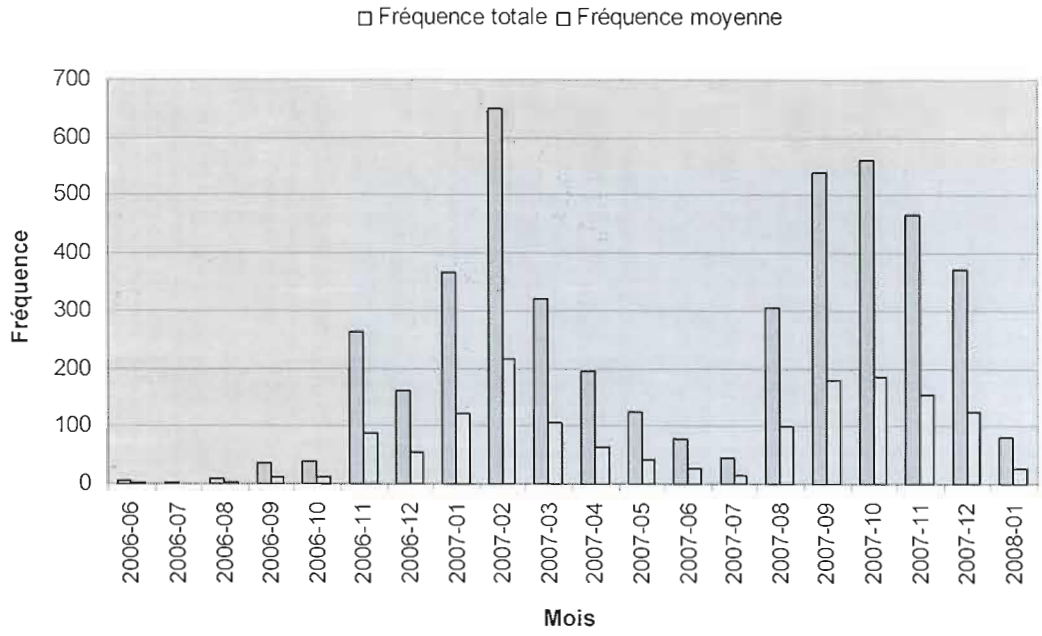


Figure 9 : Fréquence des expressions *accommodement(s)* par mois

En somme, en fonction d'un indicateur de saillance basé sur la fréquence d'occurrence de l'expression *accommodement(s)*, nous identifions trois périodes : de 1988 à octobre 2006, de novembre 2006 à juillet 2007 et d'août 2007 à janvier 2008. Ces périodes dans l'évolution de la saillance ne sont pas fortuites : elles correspondent à des évènements socio-juridico-politiques ayant fortement contribué à orienter le travail cognitif des journalistes à propos de la RS des AR.

La première période débute avec les premières décisions des tribunaux en matière de pratiques d'accommodements raisonnables au Canada. Mais au Québec, ce sera surtout à partir de 1993, avec l'affaire *Smart c. T. Eaton Ltée*, que les expressions *accommodement(s)* commencent à être utilisées par les journalistes. Cette période est ponctuée régulièrement par

différents évènements liés à des pratiques d'accommodements, surtout en milieu de travail.[41] La deuxième période débute en 2006 avec le déclenchement d'élections provinciales où le thème des AR devient un enjeu politique et médiatique de première importance. Finalement, la troisième période débute avec l'ouverture de la *Commission de consultation sur les pratiques d'accommodement reliées aux différences culturelles*, coprésidée par Gérald Bouchard et Charles Taylor.

D'autre part, nous proposons de considérer les trois journaux La Presse, Le Devoir et Le Soleil comme trois communautés épistémiques, ou trois groupes différents, dont certains de leurs membres, c'est-à-dire les auteur-e-s ayant écrit un article parmi ceux formant notre corpus, ont été impliqué-e-s dans un processus de travail cognitif collectif menant à la construction de la RS des AR. Ces membres sont nombreux, nous avons identifié 608 auteur-e-s différent-e-s en tout, 198 pour Le Devoir, 319 pour La Presse et 160 pour Le Soleil[42]. Certains d'entre eux n'ont écrit qu'un seul article formant notre corpus, d'autres sont à l'origine de plus d'une trentaine d'entre eux.

En somme, notre corpus de 3434 articles est divisé en 9 sous-corpus pour chaque journal et période. Le tableau 3 indique le nombre d'articles pour chacun de ces sous-corpus.

---

[41] Notre but n'est pas de relater l'historique des pratiques d'accommodements au Québec et au Canada. On peut retrouver ce genre d'information dans le rapport de la Commission Bouchard-Taylor (Bouchard et Taylor 2008). Par ailleurs, les auteurs du rapport ont une segmentation des périodes de l'évolution des AR très similaire à la nôtre.

[42] Ce qui indique qu'une cinquantaine de journalistes ont écrit dans plus d'un journal.

Tableau 3 : Nombre d'articles par période/journal

	Période 1	Période 2	Période 3	Total
Le Devoir	D1= 485	D2= 337	D3= 370	1192
La Presse	P1= 497	P2= 462	P3= 491	1450
Le Soleil	S1= 248	S2= 281	S3= 263	792
Total	1231	1082	1127	3434

1.2 Le DOMIF et la concordance

Pris tels quels, les articles de presse, bien qu'ils aient été sélectionnés parce qu'ils véhiculaient une expression commune et saillante, sont trop généraux pour que nous puissions y mener une analyse spécifique à la RS des AR. Tel que discuté dans le chapitre précédent, l'ensemble des données textuelles de ces articles est trop hétérogène et non spécifique à la RS étudiée. Par conséquent, il faut extraire de ces articles de presse les $domifs_{(i)}$ porteurs des traces sémiotico-empiriques spécifiques à la RS des AR. Nous proposons d'opérationnaliser cette étape à l'aide d'un algorithme de concordance (Princemin *et al.* 2006; McCarthy 2004; Rockwell 2003; Lebart et Salem 1994 : 21).

L'algorithme d'une concordance peut être décrit de la manière suivante : il s'agit d'extraire, des textes formant un corpus, tous les contextes — phrases ou énoncés — autour d'une *forme-pôle*. Une forme-pôle est habituellement un mot, et son contexte les phrases avant et après ce mot. Une concordance est la somme des contextes d'une forme-pôle. Par conséquent, chaque contexte de la concordance représente un $domif_{(i)}$ et il y a autant de $domifs_{(i)}$ qu'il y a d'occurrences de la forme-pôle dans le corpus.

Le choix de la forme-pôle est déterminant. Le chercheur doit sélectionner une forme qui soit considérée comme une instanciation saillante et canonique dans le langage de son objet d'étude.[43] Nos formes-pôles sont les expressions *accommodement* ou *accommodements*. La

[43] Par exemple, lorsque Meunier et Forest (2009) ont voulu extraire d'un corpus composé de l'œuvre de Pierce les segments de texte à propos du concept ESPRIT, ils ont utilisé une concordance dont la forme-pôle était le

taille de leur contexte est de cinq phrases, soit la phrase dans laquelle on retrouve l'occurrence de la forme-pôle, les deux phrases précédentes et les deux phrases suivantes, tel qu'illustré dans l'exemple du tableau 4 (La numérotation des phrases est entre crochets) :

Tableau 4 : Exemple d'un contexte de 5 phrases autour des formes-pôles *accommodement(s)*. Le Devoir, 1993-08-30.

Contexte avant	Phrase centrale	Contexte après
[1] Dans nos sociétés de plus en plus tolérantes, outillées désormais de chartes des droits, la discrimination directe est interdite et nous apprenons de mieux en mieux à réprimer la discrimination indirecte, c'est-à-dire les effets pervers, pour les minorités, de lois conçues pour la majorité. [2] Les tribunaux obligent ainsi les diverses autorités à des «accommodements raisonnables».	[3] On ne pourrait à peu près plus, par exemple, congédier quelqu'un dont l'horaire de travail entre en conflit avec les prescriptions normales de sa religion, à moins de prouver qu'aucun «**accommodement** raisonnable» n'est possible.	[4] Appuyé sur cette jurisprudence, le Conseil des communautés culturelles nous propose de transposer la sagesse des juges dans nos rapports quotidiens et il en codifie la manière: partenaires de diverses cultures, voici les principes et les moyens des «accommodements raisonnables». [5] L'idée de s'entendre plutôt que de se tirer des roches n'est pas mauvaise

On peut considérer ce type de DOMIF comme un *contexte d'inférences* à partir d'un inducteur. Tous les $domifs_{(i)}$ sont considérés comme des contextes d'inférence à propos de la RS des AR, induits à partir des formes-pôles *accommodement(s)*. Ce choix d'opérationnalisation cherche à traduire la logique des techniques d'association typique à l'étude des RS où l'on identifie, à partir d'un inducteur, le travail cognitif, c'est-à-dire les induits, effectués par les individus.

Pour les 9 sous-corpus, on extrait toutes les instances de ce DOMIF, c'est-à-dire tous les contextes d'inférence des formes-pôles *accommodement(s)*. Le tableau 5 indique le nombre de $domifs_{(i)}$ pour chaque période et journal.

---

lexème MIND. Si une étude s'intéresse au concept de MAISON, elle utilisera sûrement comme forme-pôle le lexème MAISON, mais peut-être aussi DEMEURE. L'enjeu autour du choix d'une forme-pôle consiste à sélectionner un signe qui pourra représenter adéquatement notre objet d'étude.

Tableau 5 : Nombre de $domifs_{(i)}$ par période/journal

	Période 1	Période 2	Période 3	Total
Le Devoir	D1= 644	D2= 780	D3= 807	2231
La Presse	P1= 718	P2= 912	P3= 1018	2648
Le Soleil	S1= 366	S2= 521	S3= 506	1333
Total	1668	2213	1331	6212

1.3 Étape 3, 4, 5 : UNIF, vectorisation et réduction dimensionnelle

Les étapes 3, 4 et 5 font appel à des algorithmes classiques pour le forage de texte, dont plusieurs sont également utilisés dans ALCESTE et d'autres logiciels. Le type d'UNIF choisi est le lexème. La pondération de chaque $unif_{(i)}$ dans chaque $domif_{(i)}$ est la fréquence relative, c'est-à-dire la fréquence de chaque lexème dans chaque contexte d'inférence issue de la concordance. La réduction dimensionnelle de la matrice est avant tout effectuée à l'aide d'une lemmatisation des lexèmes,[44] c'est-à-dire une réduction à leur racine par troncature.[45] Les $unifs_{(i)}$ de la matrice, qui étaient des lexèmes, sont par conséquent remplacées par leur racine, ce qui réduit environ de moitié la matrice.

Ensuite est appliqué un filtre de mot fonctionnel, afin de ne conserver que les mots pleins. Notre liste de mots fonctionnels est formée de 567 lexèmes différents, qui correspondent à 455 lemmes différents (voir annexe I pour la liste de mots fonctionnels utilisée). Nous avons finalement filtré les dimensions de la matrice non discriminantes à l'aide d'un calcul IDF : les

---

[44] Tel que nous l'avons explicité dans le chapitre 3, pour des raisons d'implémentation, en réalité la lemmatisation est effectuée immédiatement après avoir extrait l'index des lexèmes. La pondération est également effectuée sur l'index des lemmes et non sur l'index des lexèmes. Mais pour des raisons de présentation, il est plus simple et clair de présenter l'étape de réduction des dimensions de la matrice après l'étape du choix de l'UNIF et de la vectorisation.

[45] Pour la lemmatisation, nous avons utilisé l'algorithme de Porter (Porter 1980; Willett 2006). Nous avons récupéré le code source d'un algorithme de lemmatisation pour le français à partir de la librairie en ligne Snowball (dirigée par M.F. Porter) : http://snowball.tartarus.org.

$unifs_{(i)}$, présentent dans moins de 1% et dans plus de 40% des $domifs_{(i)}$, sont filtrées de la matrice et non retenues pour la classification. Ceci a, par exemple, filtré les lexèmes « accommodement » (« accommod » pour le lemme) et « raisonnable » (« raison » pour le lemme), présents dans beaucoup trop de $domifs_{(i)}$ pour être pertinents pour une classification automatique.

1.4 Étape 6 : Cartographier une représentation sociale avec K-Means

L'étape six consiste à sélectionner un algorithme de classification pour la cartographie des éléments formant la RS étudiée. Nous proposons d'opérationnaliser cette étape à l'aide de l'algorithme K-Means. Il s'agit d'un algorithme classique dans le domaine de la classification automatique non-supervisée qui est encore aujourd'hui l'un des algorithmes les plus utilisés, robustes et efficaces dans l'identification de classe (Wu *et al.* 2008). Son principal avantage, comparé la classification descendante hiérarchique telle qu'implémentée dans ALCESTE, est sa capacité à optimiser les résultats de classification peu importe le nombre de classes. En effet, le K-Means fonctionne par amélioration itérative d'une partition initiale de k classes, déterminées a priori par le chercheur ou la chercheuse (Memmi 2000).

Le K-Means est calculé dans un espace vectoriel. On peut présenter la logique de l'algorithme de la manière suivante : premièrement, le chercheur détermine le nombre k de classes qu'il souhaite obtenir. Les k classes sont initialisés, c'est-à-dire que chaque classe se fait attribuer une coordonnée dans l'espace vectoriel qui correspond au vecteur initial de celle-ci.[46]

---

[46] Il existe de nombreuses techniques d'initialisation des classes. La technique classique consiste à sélectionner aléatoirement k vecteurs parmi les $domifs_{(i)}$ que l'on veut classer Cependant, cette technique a l'inconvénient de rendre la qualité de la classification très dépendante de l'initialisation de départ. En effet, si par hasard, l'initialisation retient les vecteurs les plus similaires parmi tous les $domifs_{(i)}$, la classification sera de très mauvaise qualité. Par conséquent, nous avons opté ici pour une heuristique simple qui consiste à prendre seulement un premier vecteur au hasard à partir des $domifs_{(i)}$ disponibles pour initialiser la première classe. L'initialisation du deuxième cluster se fait en prenant pour coordonnée le vecteur le plus éloigné du premier, l'initialisation de la troisième classe en prenant le vecteur le plus éloigné de la classe la plus proche, et ainsi de suite jusqu'à ce que le nombre k de classes soit obtenu (Bradley et Fayyad 1998).

Ensuite, dans une première itération, on prend un $domif_{(i)}$ vectorisé et on l'associe à la classe (le vecteur initial) la plus proche. La distance entre deux vecteurs est calculée à partir du *normalise DOT product* (un calcul d'angle cosinus).[47] On calcul ensuite leur vecteur moyen (la moyenne entre le $domif_{(i)}$ vectorisé et le vecteur initial de la classe), ce qui donne ce qu'on appelle le « centroïde » de la classe. À chaque fois qu'un $domif_{(i)}$ vectorisé est attribué à une classe, on calcul à nouveau son centroïde. On répète l'opération pour les n $domifs_{(i)}$ que l'on veut classer. Dans une deuxième itération, on reprend tous les $domifs_{(i)}$ vectorisés et on redistribue chacun vers le centroïde le plus proche. On répète l'itération jusqu'à ce que plus aucun $domif_{(i)}$ vectorisé ne soit attribué à un nouveau centroïde, c'est-à-dire à une nouvelle classe. Autrement dit, on répète l'itération jusqu'à ce que la valeur des k centroïdes ne change plus (Hotho *et al.* 2005 : 19). Une fois qu'il y a stabilisation, la classification est optimale (compte tenu du nombre k de classes et de l'initialisation de départ).

Le principal problème avec K-Means est le choix à priori du nombre k de classes, qui peut varier entre 2 et n (où n est égale au nombre de $domifs_{(i)}$ que l'on veut classer). Le choix du nombre k de classes dépend de plusieurs facteurs : des facteurs qualitatifs, comme la pertinence de k compte tenu du cadre théorique et des objectifs de recherche; des facteurs quantitatifs, comme un calcul d'optimisation de la proximité intra-classe et de la distance inter-classes.

Pour nos neuf sous-corpus, n varie entre 366 et 1018. Nous avons mis 40 comme limite qualitative supérieure à k. C'est-à-dire que nous limitons à 40 classes la cartographie de la RS avec K-Means que nous supposons correspondre à 40 éléments cognitifs formant la RS des

---

[47] Cette étape implique un calcul de distance entre deux vecteurs, celui du $domif_{(i)}$ et celui de la classe. K-Means peut fonctionner avec différentes mesures de distance : euclidienne, Jaccard, etc. Selon Brunet (2003), il semble d'ailleurs avoir convergence entre ces différentes mesures. Nous avons opté pour l'un des calculs de distance les plus utilisés en forage de texte, le normalise DOT product (un calcul d'angle cosinus) (Weiss et al. 2005 . 91).

AR. Cette limite a été choisie empiriquement et en fonction d'un survol de la littérature sur les RS.[48]

Nous avons vu au chapitre précédent qu'il existe de nombreuses stratégies quantitatives pour évaluer la qualité d'une classification. Nous avons utilisé le *coefficient silhouette simplifié* (Hruschka et Covoes 2005). Ce coefficient varie entre 0 et 1 : plus sa valeur se rapproche de 1, plus la qualité de la classification est jugée bonne. Dans le graphique de la figure 10 sont illustrés les valeurs du coefficient silhouette simplifié pour les 9 sous-corpus, pour 10 à 60 classes par partition.

---

[48] Nous avons vu au chapitre 2 que les éléments d'une RS ne sont pas dénombrables. L'analyse peut porter sur un très petit nombre d'éléments, deux ou trois, ou un très grand nombre, plus de quatre-vingts. Il n'y a pas de seuil à priori et, mal menée, autant une étude sur deux que sur quatre-vingts éléments peut mener à des résultats triviaux.

Quarante est un nombre moyen pour une étude de RS. Évidemment, il est toujours possible avec K-Means d'augmenter le nombre k initial, ce qui peut permettre éventuellement une meilleure différenciation des éléments cognitifs formant la RS étudiée. La contrepartie est la perte d'intelligibilité des résultats et de l'analyse. En effet, en fixant la limite supérieure de k à 40, pour 9 sous-corpus, il est possible que l'analyse doive éventuellement porter sur un total de 360 classes, ce qui est déjà lourd comme traitement. Si notre analyse avait porté seulement sur une période pour un journal, nous aurions pu augmenter k et par le fait même augmenter potentiellement le détail de l'analyse.

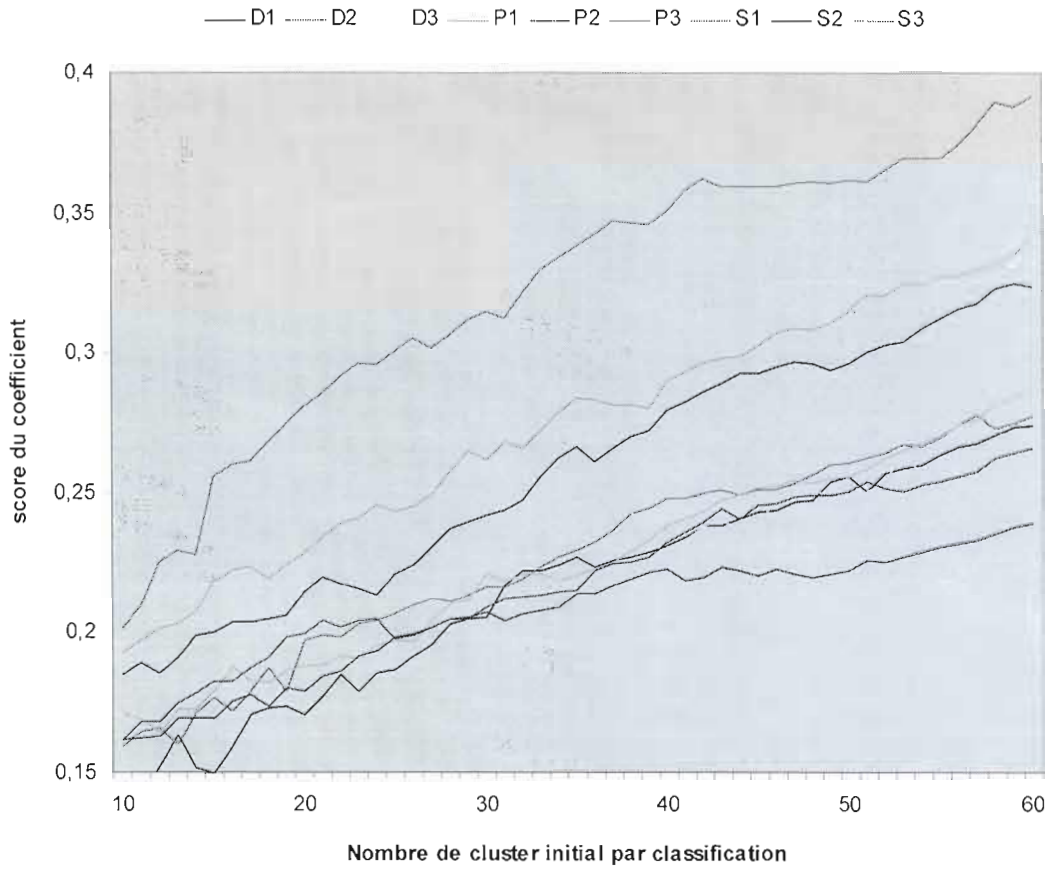


Figure 10 : Distribution des scores du coefficient silhouette simplifié pour les partitions variant entre 10 à 60 classes

Ce graphique montre que, l'augmentation du nombre *k* de classes initiales est relativement proportionnel à l'augmentation de la qualité de la classification pour les 9 sous-corpus. En fixant à 40 la limite qualitative de *k*, la classification optimale pour les 9 sous-corpus varie entre 38 et 40 (voir tableau 6). Les scores du coefficient silhouette simplifié varient quant à eux entre 0,223 et 0,351.[49]

[49] Par ailleurs, si on se fit à Hotho *et al.* (2005), ces scores sont plutôt faibles. Cependant, ceci n'est pas le résultat d'une mauvaise classification, mais plutôt des contraintes inhérentes à nos sous-corpus. En effet, ceux-ci contiennent des $domifs_{(i)}$ très homogènes, car générés à partir d'une concordance. Plus les $domifs_{(i)}$ soumis à une

Tableau 6 : Nombre de classes par période/journal

	Période 1	Période 2	Période 3	Total
Le Devoir	D1= 40	D2= 40	D3= 38	118
La Presse	P1= 40	P2= 40	P3= 40	120
Le Soleil	S1= 40	S2= 40	S3= 40	120
Total	120	120	118	358

1.4.1 La catégorisation et l'interprétation des classes

La classification permet d'atteindre le premier niveau d'analyse des RS : l'identification des éléments cognitifs de la RS. Une fois la classification terminée, le travail du psychosociologue consiste dans un premier temps à catégoriser chacune des classes à l'aide d'une étiquette linguistique traduisant le ou les quelques éléments cognitifs sous-jacents aux $domifs_{(i)}$ regroupés ensemble.

La catégorisation des classes peut se faire selon différentes techniques. On peut faire appel aux techniques classiques d'analyse de contenu. Cela nécessite de lire les $domifs_{(i)}$ de chaque classe. Bien que cette démarche soit des plus riches, elle devient rapidement impraticable lorsqu'on a affaire à plusieurs milliers de $domifs_{(i)}$. Pour assister davantage le processus de catégorisation des classes, différentes techniques statistiques sont possibles. Elles permettent de ne retenir pour chaque classe que l'information la plus représentative nécessaire à sa catégorisation.

Une technique souvent utilisée consiste à extraire l'isotopie de chaque classe, c'est-à-dire de repérer sur la base de critère statistique les lexèmes (ou leur forme réduite) les plus représentatifs de chacun.[50] Forest (2006), par exemple, a montré que ces isotopies peuvent être repérées via un calcul de TF*IDF (vue au chapitre 3). À titre d'illustration, pour la classe

opération de classification automatique sont homogènes quant à la distribution de leurs $unifs_{(i)}$, plus il est difficile de les regrouper dans des classes très différenciés les uns des autres.

[50] Dans ALCESTE par exemple, ces isotopies sont appelées « mondes lexicaux » et se calculent à partir du chi^2.

D3-30, on peut résumer l'information présente en retenant son isotopie (sous forme de lemme) pour catégoriser son contenu. Cette technique nous permet d'observer que le ou les quelques éléments cognitifs saillants dans la classe D3-30, s'actualisent dans le langage via les thèmes suivants : CITOYENNET, QUÉBÉCOIS, VALEUR, SOCIET, CONSTITU, IDENT, CHART, RELI, INSTITU.

Autre exemple, le classe D3-35 est quant à elle caractérisée par l'isotopie suivante : RELIGION, MOUV, CATHOL, LAÏQU, PUBLIC, CULTUREL, JOU, LIEUX, SCOLAIR, ÉCO. Selon Forest (2006), ceci nous indique les principaux thèmes présents dans une classe.

Pour certaines recherches, l'extraction d'isotopie est une technique de catégorisation trop générale et ne permet pas de révéler les aspects les plus intéressants pour une analyse fine des classes (Meunier et Forest 2009). Avec K-Means, d'autres techniques statistiques sont possibles pour assister le processus de catégorisation.

Il est possible d'extraire d'une classe ses $domifs_{(i)}$ les plus représentatifs. En effet, on peut considérer que, dans une classe, les $domifs_{(i)}$ les plus proches du centroïde sont également les plus représentatifs, car ce sont eux qui minimisent la distance avec tous les autres $domifs_{(i)}$ dans la classe.

À titre d'illustration, nous avons extrait de la classe D3-30 le $domif_{(i)}$ le plus proche du centroïde, c'est-à-dire le plus proche du vecteur moyen de la classe. Il s'agit du $domif_{(3803)}$. Autrement dit, on peut considérer le $domif_{(3803)}$ comme étant le contexte d'inférence à propos de la RS des AR qui représente le mieux, statistiquement, tous les autres contextes d'inférence de cette classe.

> $domif_{(3803)}$: Un bon nombre de questions que se pose la commission Bouchard-Taylor ne disparaîtraient pas du jour au lendemain à la suite de l'accession du Québec à la souveraineté. En effet, la discrimination systémique, les inégalités économiques, l'inégalité dans la participation et la représentation politique des Québécois de diverses origines, la xénophobie continuerait à interpeller toutes les composantes de la société québécoise. De la même manière, le caractère pluriel de la société québécoise continuerait à soulever des débats sur les accommodements raisonnables, la laïcité, les politiques d'inclusion et de reconnaissance de la diversité.

> Toutefois, la souveraineté nous donnera les moyens de consolider et de faire jaillir un cadre référentiel nettement plus clair pour tous les Québécois. Il serait ainsi possible d'affirmer que les principaux fondements éthiques, politiques, culturels et juridiques du cadre civique commun sont les institutions démocratiques, la Charte québécoise des droits de la personne, la Charte de la langue française et la politique québécoise de l'interculturalisme sans que chacun de ces éléments soit remis en question ou balisé, comme c'est actuellement le cas, par le fait que le Québec continue d'être inféodé au cadre et aux normes canadiens.

Nous avons fait la même chose dans la classe D3-35, où cette fois-ci c'est le $domif_{(4066)}$ qui est statistiquement le plus représentatif :

> $domif_{(4066)}$ Maniant avec dextérité le sophisme à la manière d'un jésuite défroqué, Marc Ouellet est entré dans le débat public à la faveur d'un concept que l'actualité lui a fourni sur un plateau d'argent en tant qu'instrument de son combat réactionnaire : les accommodements raisonnables. Il faut discriminer en fonction de la religion catholique et protestante quant à l'enseignement de la religion à l'école.
>
> La majorité québécoise, présupposée catholique ou protestante, a aussi un droit à des accommodements raisonnables. Sa dernière trouvaille en la matière n'est pas des moindres : « Le cours d'État d'éthique et de culture religieuse ». Nouveau sophisme, car il y a eu, encore récemment, un véritable « cours d'État » de religion et de morale dans les écoles du Québec puisqu'il était protégé par la clause constitutionnelle « nonobstant ».

Par ces deux exemples simples, on voit notamment que ce ne sont pas les mêmes dimensions de la RS des AR qui sont traitées. Dans la classe D3-30, les AR sont pensés dans un cadre politique et identitaire. On présente les AR comme un cas particulier d'un problème plus général qu'est la diversité dans la société québécoise. Une meilleure définition de l'identité québécoise, aux sens politique et citoyen du terme, par exemple via l'ascension à la souveraineté, est amenée comme une solution politique possible. Tandis que dans la classe D3-35, les AR sont pensés dans le cadre de la religion et de la laïcité. À nouveau, on présente les AR comme une instanciation particulière d'un problème de société plus générale. Dans ce cas, les accommodements religieux dans les écoles sont un cas concret illustrant les luttes idéologiques en général dans la société québécoise entre catholicisme et laïcité.

Pour des raisons de démonstration, nous nous limitons ici à la catégorisation et l'interprétation de ces deux classes seulement. Cette interprétation a été assistée par l'extraction de leur isotopie et par une lecture du contexte statistiquement le plus représentatif

de chaque classe. Une recherche plus large porterait sur l'ensemble des 38 classes, et sur les 40 classes des autres sous-corpus. Cependant, ce n'est pas l'objet de notre recherche d'analyser dans le détail la RS des AR. Au-delà de l'analyse du contenu de la RS des AR — par exemple via son évolution ou sa variation entre journaux — nous voulons démontrer qu'opérationnalisé de manière adéquate, le forage de texte offre des techniques d'analyse des RS qui atteignent les trois niveaux d'analyse attendus par une méthodologie des RS.

2. L'identification des structures organisant une représentation sociale par l'analyse de réseaux

Jusqu'à maintenant, la plupart des recherches sur les RS à partir de méthodes de forage de texte se terminent à l'étape précédente : une fois effectuée la classification, on tente d'analyser et de catégoriser le contenu de chaque classe, soit par des techniques classiques d'analyse de contenu, soit par des techniques d'extraction de leur isotopie, ou encore par une technique d'échantillonnage des $domifs_{(i)}$ les plus représentatifs.[51]

Dans une analyse structurelle des RS, ceci est insuffisant. L'objectif de recherche d'une approche structurelle n'est pas tant l'analyse fine de chaque élément de la RS, mais plutôt l'analyse de leur *position structurelle*. On cherche à repérer dans les structures les éléments qui occupent une position centrale. Traduit dans le cadre du forage de texte, l'objectif est de repérer parmi les 38 ou 40 classes de chaque sous-corpus celle la plus centrale, c'est-à-dire celle la plus susceptible de regrouper les données textuelles —les contextes d'inférence de la concordance — qui sont l'expression dans le langage des éléments cognitifs centraux de la RS étudiée.

Nous avons développé deux techniques d'analyse de la classification, qui permettent de générer deux types de structures parmi les classes. La première technique cherche à modéliser la structure cognitive de la RS, alors que la deuxième cherche à modéliser la structure sociale de la RS. Les deux techniques utilisent la théorie mathématique des graphes.

---

[51] Ce ne sont évidemment pas les seules techniques possibles d'analyse et de catégorisation du contenu des classes. Avec ALCESTE on peut notamment utiliser la technique des segments répétés (Lebart et Salem 1994).

Dans le premier cas, il s'agit d'un graphe à un mode où les nœuds correspondent aux classes et les liens aux relations entre classes. Autrement dit, on considère les nœuds comme les éléments de la RS et les arcs comme des modalités de relation cognitive entre éléments. Une première structure est ainsi obtenue sous la forme d'un réseau cognitif. Dans le deuxième cas, il s'agit d'un graphe à deux modes, où deux types de nœuds correspondent respectivement aux classes et aux individus tandis que les liens correspondent aux relations entre individus et classes. On considère chaque lien entre deux nœuds comme une relation d'association entre un journaliste et un élément cognitif. Une deuxième structure est ainsi obtenue sous la forme d'un réseau sociocognitif.

2.1 Réseau cognitif de proximité sémantique entre éléments d'une représentation sociale

Dans le domaine d'étude des RS, la modélisation de la structure cognitive d'une RS sous la forme d'un réseau a déjà fait l'objet de plusieurs réflexions et travaux. C'est le cas d'une technique classique d'analyse structurale des RS, appelée « l'analyse de similitude ». La modélisation que nous proposons sous la forme d'un réseau de proximité sémantique veut reproduire la logique de cette technique.

Rappelons que l'analyse de similitude est basée sur le repérage du noyau d'une RS à partir des propriétés de connexité des éléments cognitifs dans la structure. Dans sa forme canonique, l'analyse de similitude procède en trois étapes : premièrement, à l'aide d'une technique associative, le chercheur demande à une population de générer, à partir d'un inducteur, une série d'induits (mots ou courtes phrases) qui sont ensuite considérés comme les formes linguistiques des éléments cognitifs composant la RS étudiée. Deuxièmement, à l'aide d'un questionnaire de caractérisation, le chercheur ou la chercheuse demande aux sujets d'évaluer les relations de similitude entre les différentes formes linguistiques retenues pour exprimer la RS étudiée. Troisièmement, les relations de similitudes entre éléments sont représentées sous la forme d'un graphe et différentes mesures de centralité dans le graphe sont utilisées pour repérer les nœuds représentant les éléments les plus connexes dans la structure de la RS. Ces éléments sont considérés comme formant le noyau central de la RS étudiée.

Il est possible de modéliser un réseau cognitif autrement qu'avec des relations de similitude. La théorie des RS utilise parfois des relations à plusieurs modalités, comme les schèmes cognitifs de base. Par ailleurs, selon Popping (2000 : 99), les relations dans un réseau cognitif — ce qu'il appelle un réseau de concept — peuvent être caractérisées par quatre propriétés formelles : sa signification sémantique, sa direction, la valeur de la relation et son signe (positif ou négatif). Dans l'analyse de similitude, la signification sémantique des relations est la similitude ou la proximité entre deux éléments, les relations sont bidirectionnelles, la valeur correspond à celle de la similitude et le signe est positif.

Généralement, l'analyse de similitude utilise un indice de similitude appelé « indice D » (Bouriche 2003) qui est appliqué sur les résultats du questionnaire de caractérisation. Mais, selon Flament et Rouquette (2003), plusieurs indices de similitude sont possibles : « [...] est en fait acceptable tout indice $s(XY)$ qui augmente lorsque la ressemblance entre X et Y augmente, selon des critères dont le psychosociologue est juge. » (Flament et Rouquette 2003 : 87)

Bien que cette technique soit utilisée dans un cadre de recherche très différent, nous proposons de traduire la logique de l'analyse de similitude dans le cadre du forage de texte. Le défi, d'un point de vue méthodologique, consiste entre autres à sélectionner un indice de similitude qui soit opérationalisable sous la forme d'un algorithme de forage de texte. Plus précisément, il nous faut un indice de similitude qui augmente lorsque la similitude entre deux classes augmente également.

2.1.1 Relations de similitude dans un espace vectoriel

Il est possible de mesurer la similitude entre deux classes dans un espace vectoriel comme celui utilisé dans notre méthode de forage de texte. Une telle mesure calcule la similarité entre les distributions d'$unifs_{(i)}$ de deux classes, c'est-à-dire la similitude entre leurs patterns de cooccurrence lexicale. La thèse de la sémantique distributionnaliste que nous avons vue au chapitre 3 nous suggère dans ce cas que nous pouvons parler de proximité sémantique entre deux classes.

Nous proposons de mesurer cette proximité sémantique entre classes à partir de leur centroïde. Par conséquent, deux classes caractérisées par des centroïdes proches l'un de l'autre dans un espace vectoriel sont également considérées comme proche sémantiquement. L'indice de similitude est calculé à partir du *normalise DOT product*.[52] Les valeurs de proximité sémantique varient donc entre 0 et 1. Plus la similitude entre deux classes augmente, plus sa valeur se rapproche de 1. Deux classes liées par une relation de proximité sémantique de 1 sont considérées comme identiques sémantiquement, alors que deux classes liées par une relation de proximité de 0 sont considérées complètement distinctes.[53]

La figure 11 représente le réseau de proximité sémantique entre classes du sous-corpus D3.[54] Il s'agit d'un réseau composé de 38 nœuds et 703 relations valuées entre 0,012 et 0,631 (voir les annexes II à X pour l'ensemble des réseaux de proximité sémantique pour les 9 sous-corpus).

---

[52] Le normalise DOT product est une mesure de distance entre deux vecteurs. Par conséquent, pour calculer la valeur de la proximité entre les centroïdes de deux clusters, il suffit de faire la différence : 1- (normalise DOT product).

[53] Plus précisément, lorsque la valeur de proximité entre deux classes est égale à 1, ceci signifie que les vecteurs correspondants au centroïde des deux classes sont identiques. Lorsque la valeur est de 0, c'est que les deux centroïdes sont complètement différents.

[54] Pour la représentation graphique des réseaux, nous avons utilisé le logiciel UCINET 6 et NetDraw (Borgatti et al. 2002).

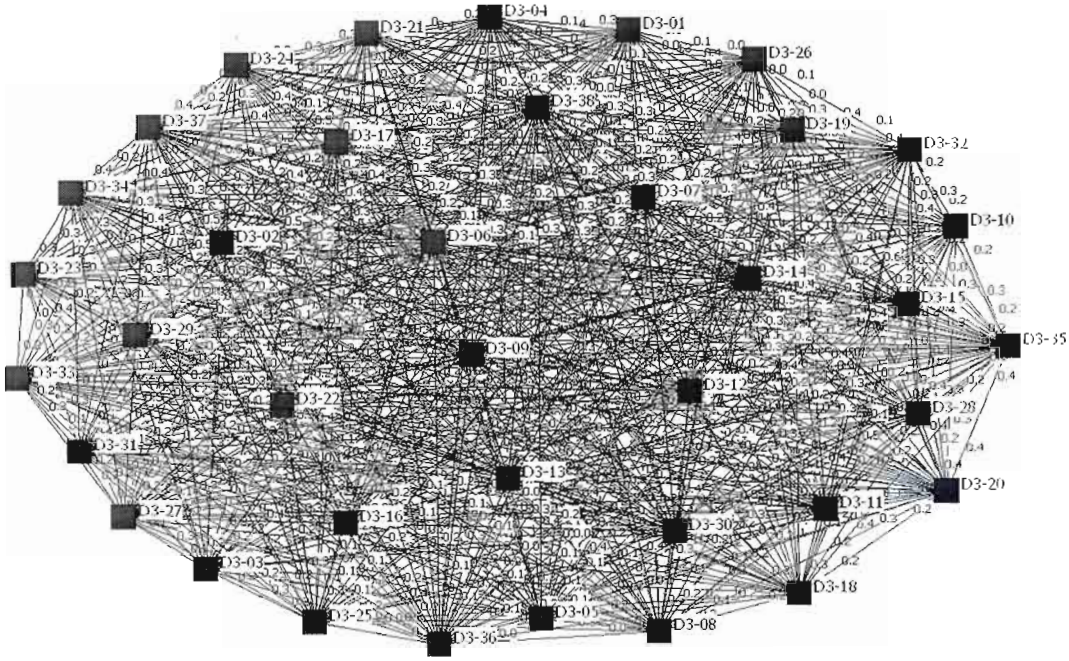


Figure 11 : Réseau de proximité sémantique de D3

Ce réseau, compte tenu de sa taille, est pratiquement impossible à lire tel quel. Il peut être utile dans ce cas de chercher à ne garder que les patterns les plus forts de la RS. Une première manière de faire est d'appliquer un filtre au seuil (Bouriche 2003), afin de ne conserver que les relations les plus fortes. Dans la figure 12, nous n'avons conservé que les relations de proximité sémantique supérieures à 0,5.

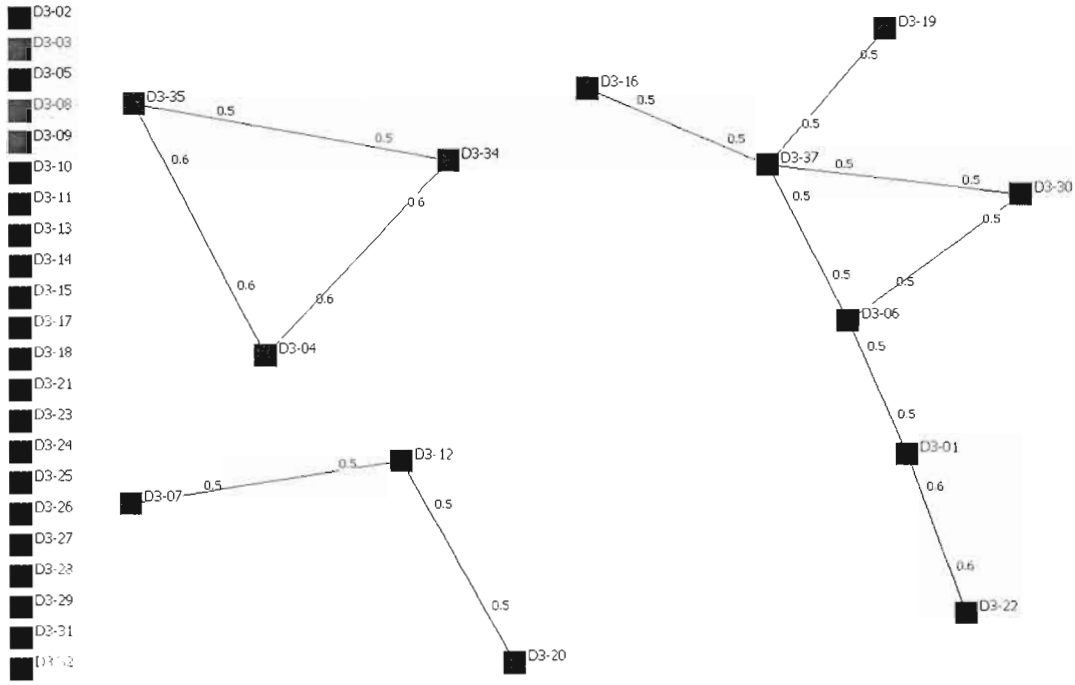


Figure 12 : Principales composantes du réseau cognitif de D3

Trois composantes apparaissent : une première composée des nœuds D3-35, D3-34 et D3-04; une deuxième composée des nœuds D3-07, D3-12 et D3-20; et une troisième, plus grande, composée des nœuds D3-16, D3-19, D3-37, D3-30, D3-06, D3-01 et D3-22. Ceci est déjà un résultat d'analyse intéressant, qui nous révèle des lieux de très forte connexité dans la RS étudiée. Ces composantes sont certainement des candidates potentielles pour former le système central de la RS.

Nous avons également vu (chapitre 2) que dans le domaine d'étude des RS, l'arbre maximum est souvent utilisé pour simplifier des réseaux cognitifs trop complexes. La figure 13 est l'arbre maximum du réseau cognitif de D3.

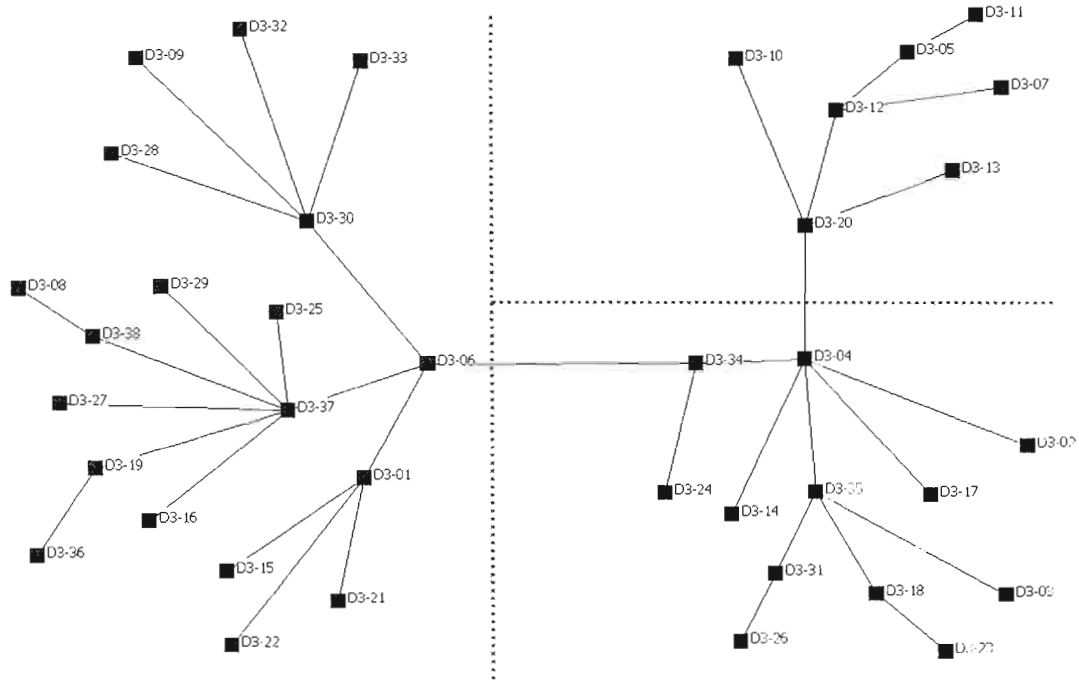


Figure 13 : Arbre maximum extrait du réseau de proximité sémantique de D3

L'arbre maximum est intéressant, car c'est le plus petit graphe connexe sans cycle que l'on peut extraire du graphe complet en ne retenant que les relations les plus fortes. Il nous permet de visualiser à nouveau les trois composantes identifiées précédemment, mais en les recadrant avec les autres éléments du réseau. Ceci nous permet de constater que ces trois composantes sont des centres de gravité importants autour desquels s'organise le reste des éléments.

Le réseau de proximité sémantique de la figure 12 est considéré comme une modélisation de la structure de la RS des AR véhiculée dans le sous-corpus D3. Chaque point représente une classe regroupant les données textuelles — contextes d'inférence — qui sont l'expression dans le langage d'un ou de quelques éléments cognitifs saillants. Les relations entre classes sont des relations de proximités sémantiques mesurées en termes de proximité entre vecteurs. Nous proposons ici qu'elles soient interprétées comme une modélisation formelle des relations de similitude entre éléments cognitifs.

L'arbre maximum est aussi une modélisation de la structure de la RS, mais simplifiée. On dira parfois qu'elle permet de ne garder que le « squelette » de la RS, ce qui est le plus important.

La logique de l'analyse de similitude est ainsi reproduite, mais adaptée au forage de texte. Son principe général reste : plus deux éléments cognitifs sont similaires, plus ils sont liés par une relation forte.

Pour l'étude des RS, cette modélisation est utile dans la mesure où elle permet par la suite de repérer les éléments les plus centraux de la RS, c'est-à-dire les éléments les plus connexes. Mais elle peut également être intéressante pour de nombreux programmes de recherche en sciences sociales et humaines qui s'intéressent aux représentations collectives, notamment l'anthropologie cognitive étasunienne (D'Andrade 1995; Shore 1996) dont les objectifs de recherche sont très similaires au programme français d'étude des RS.

2.2 Réseau sociocognitif de distribution sociale des éléments cognitifs de la représentation sociale

Dans le domaine d'étude des RS, la modélisation de la structure de distribution sociale des cognitions a été faite essentiellement de deux manières. Nous avons vu que l'école de Genève, qui s'intéresse aux variations interindividuelles de prises de position et à ses principes organisateurs, a essentiellement utilisé les analyses factorielles (Doise *et al.* 1992). L'école d'Aix-en-Provence, qui s'intéresse au consensus, a quant à elle utilisé l'analyse booléenne de questionnaire (Flament 1996b, 1999; Gaymard 2002).

L'analyse du consensus, contrairement à celui des prises de position, est un indicateur pour le repérage du noyau central d'une RS. Évidemment, l'analyse de questionnaire ne peut être utilisée ici, mais d'autres techniques peuvent être développées pour analyser le partage social. C'est le cas de l'analyse de réseau sociocognitif. À notre connaissance, dans le domaine d'étude des RS, l'analyse de réseaux sociocognitifs n'a pas été utilisée pour la modélisation du partage social. Par ailleurs, d'autres chercheurs issus d'autres domaines des sciences sociales ont récemment exploré cette possibilité et produisent des résultats de recherche des

plus intéressants pour notre propos (Roth 2005; Monge et Contractor, 2003; Carley 1986; Diesner et Carley 2005).

Nous proposons de construire un réseau sociocognitif à partir des résultats de la classification. Dans ce réseau, un individu, c'est-à-dire un journaliste, est lié à un élément cognitif, c'est-à-dire une classe, lorsqu'au moins l'un de ses contextes d'inférence se retrouve dans la classe. Ainsi, on considère que deux individus partagent un même élément cognitif lorsqu'au moins un de leurs contextes d'inférence à chacun est suffisamment similaire a celui de l'autre pour être classés ensemble.

Par exemple, dans la classe D3-30 on retrouve 12 individus différents qui ont produit en tout 43 contextes d'inférence. On considère par conséquent que ces 12 individus partagent un élément cognitif. C'est le cas des contextes d'inférence suivants ($domif_{(4347)}$ et $domif_{(3877)}$), produits par deux auteurs différents.

$domif_{(4347)}$: L'idée d'instituer une citoyenneté québécoise fait l'objet de débats depuis plusieurs années déjà. Proposée en 2001 par la commission Larose sur la situation du français, elle avait été rejetée par le Parti québécois, qui craignait que ce ne soit qu'un «gadget symbolique». Depuis, il y a eu la reconnaissance du Québec comme nation par la Chambre des communes, le débat sur les accommodements raisonnables, puis l'ADQ qui en a fait la proposition en campagne électorale. Et voilà que la nouvelle chef péquiste, Pauline Marois, a déposé à l'Assemblée nationale jeudi un projet de loi sur l'identité reprenant cette idée.

Si le contexte est davantage propice qu'il y a six ans à une citoyenneté québécoise, la difficulté demeure la même : comment, dans le contexte constitutionnel actuel, donner à une citoyenneté québécoise une valeur telle que les nouveaux arrivants la demanderont et consentiront à travers ce geste à adhérer aux valeurs qui font l'identité québécoise?

$domif_{(3877)}$: Constitution

Hier, la Commission de consultation sur les pratiques d'accommodement reliées aux différences culturelles a également entendu parler des mérites d'adopter une constitution québécoise, qui énoncerait les valeurs de la société québécoise.

Cela contribuerait à faciliter l'intégration des immigrants et à juger des demandes d'accommodement raisonnable, a-t-on plaidé devant la commission.

Ainsi, Mme Line Chaloux, de l'organisme Le Coffret qui œuvre auprès des immigrants en région, a suggéré qu'une telle constitution québécoise énonce des valeurs comme l'égalité homme-femme, la laïcité de l'espace public, de l'enseignement et des services de santé.

Nous avions souligné précédemment que dans la classe D3-30, les AR étaient pensés dans un cadre politique et identitaire où une meilleure définition de l'identité politique des Québécois était présentée comme une solution. Nous retrouvons clairement cette idée dans ces deux contextes d'inférence $domif_{(4347)}$ et $domif_{(3877)}$, à la différence près qu'ici, on insiste sur l'adoption d'une constitution québécoise comme solution politique et identitaire concrète. Leur contenu langagier étant similaire (toujours d'un point de vue distributionnaliste), on considère que ces deux individus partagent au moins un élément cognitif commun.

Dans le sous-corpus D3, il y a en tout 59 individus différents qui se partagent (au moins) 38 éléments cognitifs identifiés durant la classification. Chaque élément cognitif est partagé en moyenne par 6,29 individus. La figure 14 illustre son réseau sociocognitif (voir les annexes XI à XIX pour tous les réseaux sociocognitifs). Dans ce réseau, les points rouges sont les 59 individus et les carrés bleus sont les éléments cognitifs partagés.

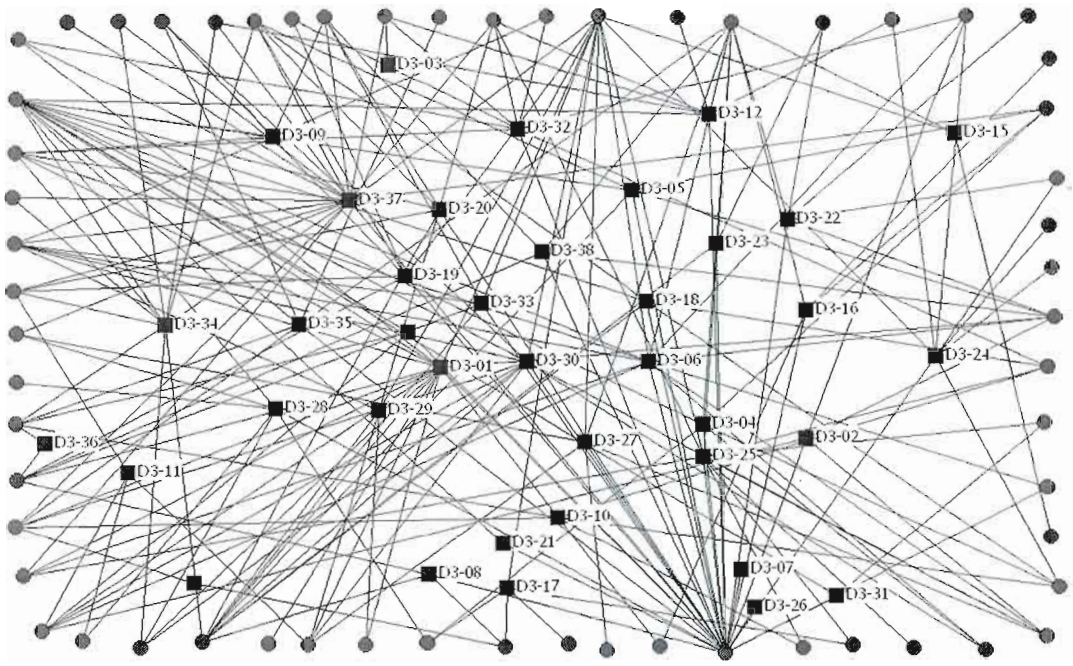


Figure 14 : Réseau sociocognitif pour le sous-corpus D3

Cette technique d'analyse du partage social des éléments cognitifs est particulièrement intéressante pour le domaine d'étude des RS. Elle permet de représenter formellement les

structures de distribution sociale d'une RS parmi les membres d'un groupe ou d'une population. En effet, la modélisation en réseau, autant pour la structure cognitive que pour la structure sociale de la RS, permet de faire appel aux outils mathématiques de la théorie des graphes pour décrire formellement ces structures (Wasserman et Faust 1994).

Cette technique d'analyse nous permet de reproduire le principe de l'analyse du consensus, mais en l'adaptant au forage de texte. Plus un élément cognitif fait l'objet d'un consensus dans le groupe, plus il est lié à un nombre élevé d'individus.

À titre d'exemple, l'analyse de ces réseaux sociocognitifs peut être utile pour l'étude de la cohésion d'un groupe. On remarque que dans D3, aucun individu n'est isolé du reste de la communauté, tous sont liés, alors que dans D1 (annexe XI), certains individus sont coupés du reste de la communauté, car ils ne partagent aucun élément cognitif avec elle. Un calcul de densité (Monge et Contractor 2003 : 196) sur ces réseaux sociocognitifs peut offrir également des indices au chercheur sur la cohésion du groupe. Dans le journal Le Devoir, on remarque d'ailleurs que cette cohésion a doublé entre la première et la troisième période, passant de 0,06 pour D1 à 0,11 pour D3.

Autre exemple : l'analyse de ces réseaux peut permettre de repérer les acteurs clés impliqués dans le travail cognitif de la RS étudiée. Autrement dit, on peut cibler qui sont les individus ayant le plus contribué, ce qu'on appelle également des *leaders d'opinion*. Dans D3, ces leaders sont les journalistes Stéphane Baillargeons, Antoine Robitaille et Marie-Andrée Chouinard (figure 15).

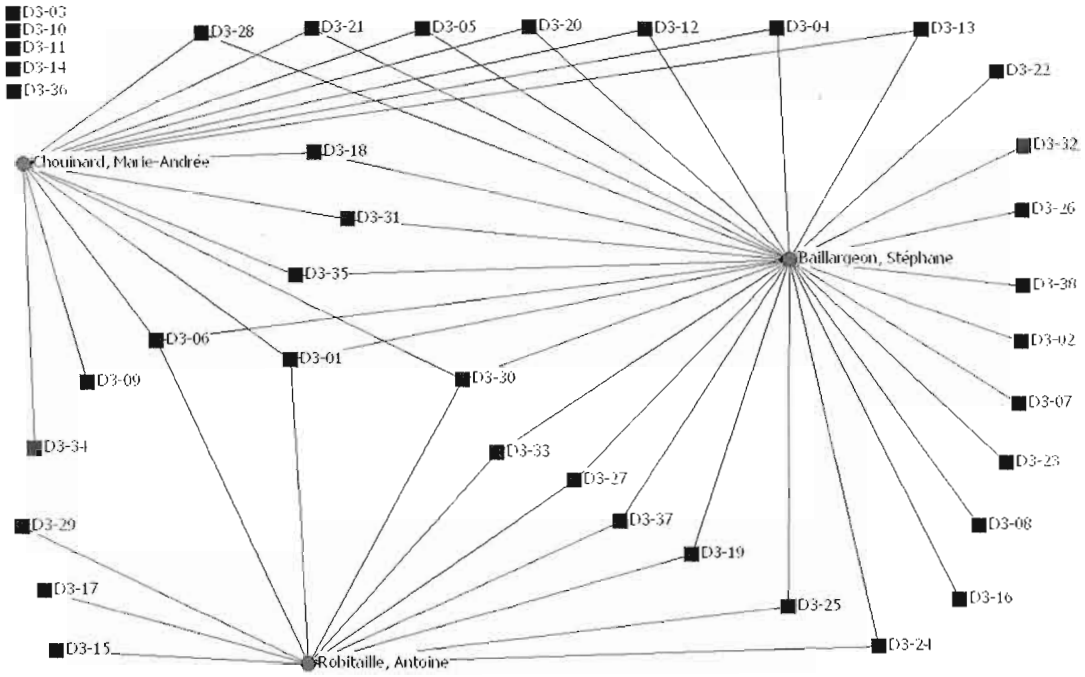


Figure 15 : Les leaders d'opinion dans le réseau sociocognitif de D3

D'autre part, l'analyse de la distribution sociale de la connaissance est un objet d'étude classique pour la sociologie (Berger et Luckmann 2003). L'analyse des réseaux sociocognitifs se révèle une technique des plus pertinentes pour plusieurs programmes de recherche comme l'étude de la diffusion des connaissances (Roger 2003), l'épidémiologie des représentations (Sperber 1996), la memetique (Blackmore 2000) ou la contagion sociale (Monge et Contractor 2003).

3. L'identification du noyau central par la centralité dans les réseaux

Nous avons vu que selon la théorie des RS le noyau central est caractérisé par plusieurs propriétés, notamment une forte valeur symbolique, une forte connexité et un fort consensus social. Ces propriétés sont des effets de différents processus de la pensée sociale : l'objectivation, l'ancrage et la recherche de consonance sociocognitive.

La valeur symbolique d'un élément cognitif d'une RS correspond à son niveau d'objectivation ou d'institutionnalisation. Plus l'élément est objectivé, plus son expression

publique, principalement via le langage, est cristallisée et fixée, c'est-à-dire par rapport à laquelle il y a peu de variation inter-individuelle.

La connexité d'un élément d'une RS correspond à son niveau d'implication dans la RS. Plus l'élément est connexe, plus il est intégré avec les autres éléments de la RS, plus on considère qu'il est « enraciné » ou qu'il a été l'objet d'une appropriation, d'une acculturation par les membres de groupe.

Le consensus social sur un élément cognitif correspond à son niveau de partage social dans la population. Plus l'élément est socialement distribué, plus on le considère important pour le groupe, dans la mesure où il est au centre de la socialisation cognitive et du maintien du lien social entre les membres.

3.1 Trois techniques de repérage de la centralité des éléments cognitifs d'une représentation sociale

Chacune de ces propriétés du noyau central peut être étudiée par des techniques d'analyse spécifique. Kalampalikis et Moscovici (Kalampalikis et Moscovici 2005) ont déjà démontré que la valeur symbolique des éléments formant une RS peut être mesurée à partir d'indicateurs de redondance dans le langage, ce qu'ils appellent la « température informationnelle ». Ils ont développé une technique de repérage de la centralité des éléments cognitifs d'une RS basée sur différents indicateurs de redondance lexicale.

Suivant leur démarche, on peut par exemple mesurer le degré d'objectivation à l'aide de la redondance des lexèmes (réduits à leurs lemmes) utilisés dans chaque classe. On utilise la différence du ratio type/token (1- (type/token) : plus une classe est caractérisée par de la redondance, plus on considère que les éléments de la RS représentée par la classe sont objectivés. Si nous prenons à nouveau l'exemple du sous-corpus D3, on peut représenter chaque classe par une valeur d'objectivation. La figure 16 illustre les scores d'objectivation pour les 38 classes de D3. Les scores varient entre 0,15 et 0,84, indiquant une hiérarchie dans les résultats, des plus redondants aux plus diversifiés.

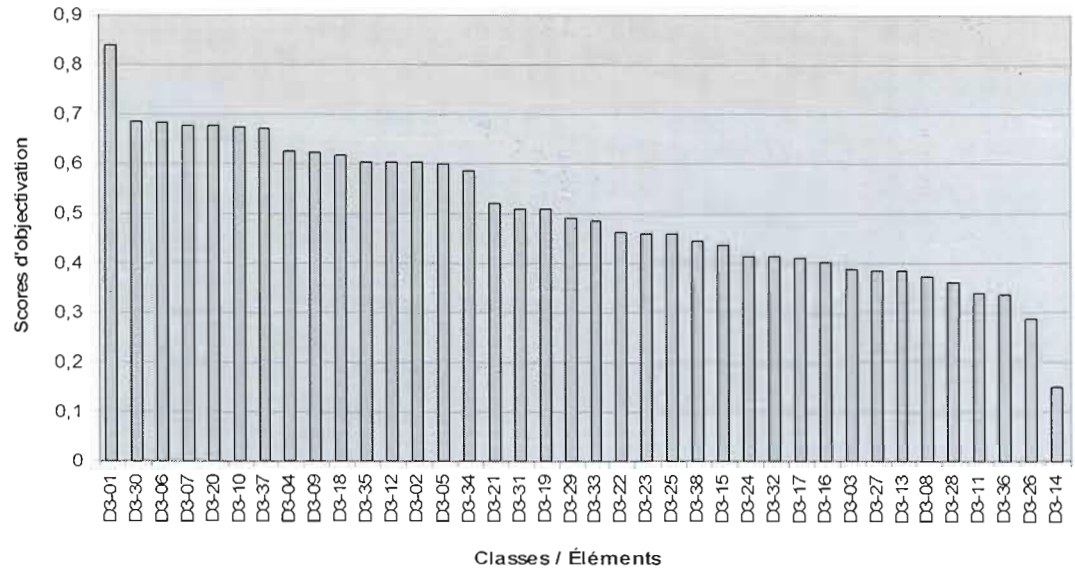


Figure 16 : Distribution des scores d'objectivation pour D3

Selon la technique de Kalampalikis et Moscovici, nous remarquons que dans ce sous-corpus D3 c'est la classe D3-01 qui possède le plus haute valeur d'objectivation, soit 0,84. Il s'agit de loin de la classe qui contient les contextes d'inférences les plus redondants, indice que les éléments cognitifs exprimés dans ces contextes sont particulièrement centraux dans la RS étudiée.

Nous savons cependant que cette technique est insuffisante et doit être corrélée à d'autres qui tiennent compte de différentes propriétés du noyau central. En modélisant les structures de la RS via des réseaux cognitifs de proximité sémantique et des réseaux sociocognitifs de distribution sociale des éléments cognitifs, on peut développer deux techniques supplémentaires en faisant appel à des algorithmes de calcul de centralité de degré dans les graphes.

3.1.1 Centralité dans le réseau cognitif de proximité sémantique

Dans le réseau cognitif de proximité sémantique, certaines mesures de connexité, comme le filtrant des cliques, s'appliquent spécifiquement à l'arbre maximum alors que d'autres, comme la centralité de proximité, s'appliquent plutôt au réseau complet. La centralité de

degré peut être mesurée soit à partir de réseau complet ou de l'arbre maximum. Pour la démonstration, nous proposons de l'appliquer au réseau complet de proximité sémantique.

Appliquée à un réseau valué de proximité sémantique, la centralité de degré permet de repérer le nœud dont la somme des valeurs de ses liens avec les autres nœuds est la plus petite du réseau. Autrement dit, le nœud le plus central est celui dont la valeur moyenne de ses liens avec les autres nœuds est la plus petite.

Dans le sous-corpus D3, chaque élément se fait attribuer une valeur de connexité en fonction de sa centralité de degré dans le réseau cognitif. Le graphique de la figure 17 montre que ces scores de connexité varient entre 0,05 et 0,32, indiquant à nouveau une hiérarchie dans les éléments des plus connexes aux plus excentriques.

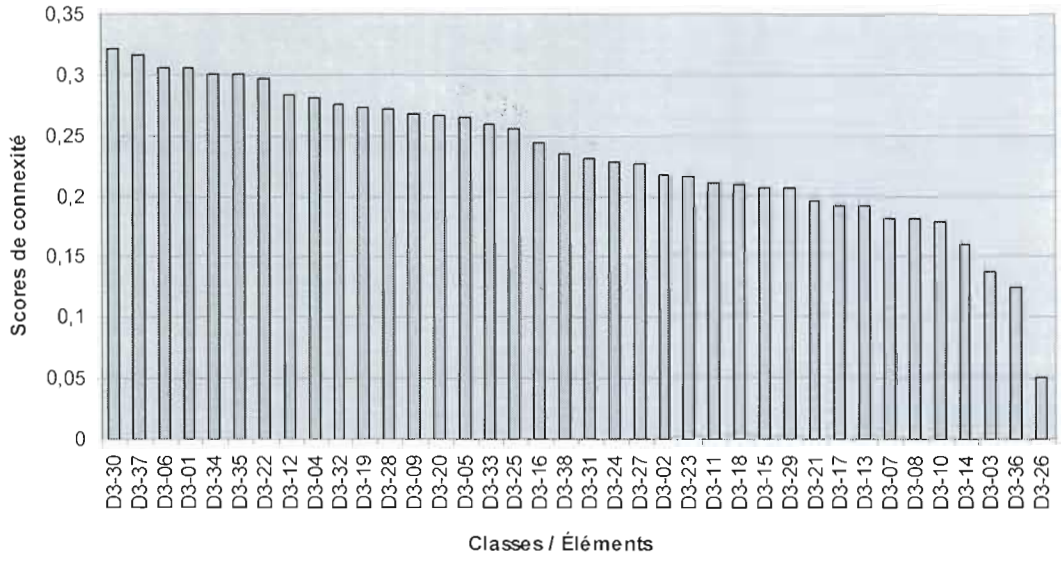


Figure 17 : Distribution des scores de connexité pour D3

Pour le sous-corpus D3, nous remarquons cette fois-ci que c'est la classe D3-30 qui a la plus haute valeur de connexité. Il représente le point dans le graphe qui minimise la distance sémantique moyenne avec tous les autres points du réseau. Selon cette technique d'analyse, nous pouvons considérer que la classe D3-30 regroupe les données textuelles — contextes

d'inférences — qui sont l'expression dans le langage des éléments cognitifs les plus centraux de la RS des AR, parce que les mieux intégrés ou ancrés dans le groupe.

Nous savons que le repérage du noyau central à l'aide d'indicateurs de centralité basés sur la connexité a longtemps été considéré comme « la voie privilégiée ». Mais, comme c'est le cas pour la technique d'analyse de Kalampalikis et Moscovici basée sur la « température informationnelle », cette stratégie aussi est aujourd'hui considérée comme insuffisante (Moliner *et al.* 2002 : 123).

3.1.2 Centralité dans le réseau sociocognitif

L'analyse de réseau sociocognitif nous permet de poursuivre l'analyse et de mesurer cette fois le consensus social sur les éléments de la RS étudiée. On considère dans ce cas que plus un élément est socialement partagé parmi les membres d'un groupe, plus il est l'objet d'un consensus. Autrement dit, plus une classe regroupe un nombre élevé d'individus différents, plus celle-ci est l'expression dans le langage des éléments cognitifs consensuels de la RS.

Dans le réseau sociocognitif, une mesure de centralité de degré peut être à nouveau appliquée. Plus un nœud *i*, qui correspond à une classe, est lié à un nombre élevé de nœuds *j*, qui correspond aux individus, plus *i* est considéré socialement partagé.

Dans le sous-corpus D3, chaque élément se fait attribuer un score de consensus social en fonction de sa centralité de degré dans le réseau sociocognitif. La figure 18, montre que ces scores de consensus varient entre 0,02 et 0,32, indiquant à nouveau une hiérarchie parmi les éléments, des plus consensuels aux plus marginaux.

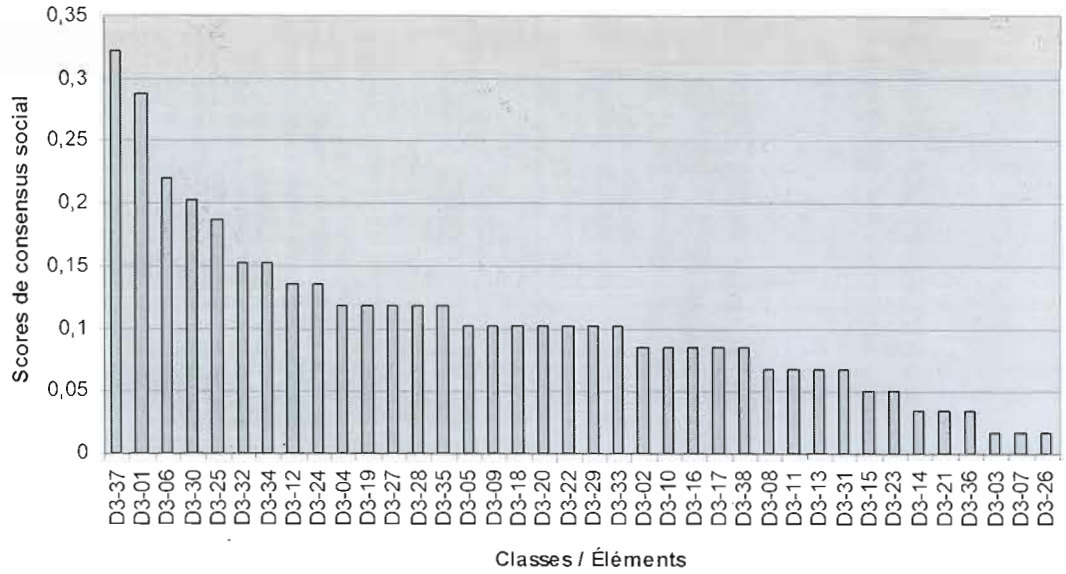


Figure 18 : Distribution des scores de consensus social pour D3

Cette troisième technique d'analyse de la centralité nous indique cette fois que c'est la classe D3-37 qui a la plus haute valeur de consensus social. Les éléments cognitifs exprimés dans cette classe sont partagés par 32 % des individus. Selon la théorie des RS, ces indicateurs sur le consensus social des éléments nous dénotent qu'ils sont probablement à la base du lien social et qu'il s'agit, par conséquent, d'éléments qui ont été l'objet de recherche de consonance sociocognitive entre membres.

3.2 Triangulation des techniques d'analyse de la centralité

On insiste de plus en plus, dans le domaine d'étude des RS, pour une approche « multiméthodologique » (Abric 2003). Nous l'avons constaté, les différentes techniques d'analyse, de la redondance dans le langage à l'analyse de similitude, sont toutes considérées individuellement comme insuffisantes et doivent plutôt se compléter les unes les autres. On demande de plus en plus, par conséquent, des analyses comparatives sur les différentes méthodes et techniques. Ce besoin apparaît, en tout cas, dans l'exemple qui nous a servi d'illustration.

En effet, on remarque que, selon les trois techniques de repérage du noyau central, la classe la plus centrale retenue par chacune varie. Selon la technique de Kalampalikis et Moscovici, c'est la classe D3-01 qui est la plus centrale, selon notre technique basée sur le réseau cognitif de proximité sémantique, c'est la classe D3-30 et finalement selon notre technique basée sur le réseau sociocognitif, c'est la classe D3-37. À première vue, ceci nous indique que selon la technique d'analyse employée, le noyau de la RS change, ce qui serait contraire aux hypothèses fondamentales de la théorie des RS.

Afin de valider la valeur épistémologique des différentes techniques d'analyse des RS, Apostolidis (Apostolidis 2003) a récemment proposé une stratégie de validation par « triangulation » :

> L'idée de triangulation repose sur un principe de validation des résultats par la combinaison des différentes méthodes visant à vérifier l'exactitude et la stabilité des observations. [...] la triangulation a été conçue comme une procédure pour vérifier une hypothèse, mise à l'épreuve dans des différentes opérations méthodologiques pour tester si oui ou non les résultats corroborent entre eux. (Apostolidis 2003 : 14-15)

Il s'agit d'une stratégie épistémologique de validation basée sur le principe simple du croisement des résultats : si on peut démontrer que différentes méthodes et techniques produisent, à partir des mêmes données, des résultats qui convergent de manière significative, ceci nous permet de valider positivement les méthodes et techniques en questions.

Nous proposons deux tests de triangulation. Le premier est basé sur le rang moyen des différents scores de centralité. Il est illustré à nouveau sur le sous-corpus D3. Le deuxième test est basé sur des calculs de corrélation entre les différentes mesures de centralité. Ce test est effectué pour les 9 sous-corpus.

3.2.1 Triangulation par le rang moyen

Ce test nous permettra de déterminer le noyau central de la RS des AR dans D3. Bien que nous ayons constaté que la classe la plus centrale varie d'une technique à l'autre, cette divergence n'est qu'apparente. Une triangulation par rang moyen semble plutôt nous indiquer que, dans le cas de D3, cette divergence de surface cache la complexité du noyau.

En effet, nous savons que le noyau d'une RS peut être composé de plus d'un élément (c'est en fait le cas, la plupart du temps). En faisant la moyenne des rangs de centralité, selon les trois techniques vues, nous pouvons constater une convergence nette dans un petit nombre de classes très centrales. Comme on peut le constater dans le graphique de la figure 19, les classes D3-01, D3-30, D3-37 et D3-06 sont celles qui occupent en moyenne les premiers rangs, loin devant les autres.

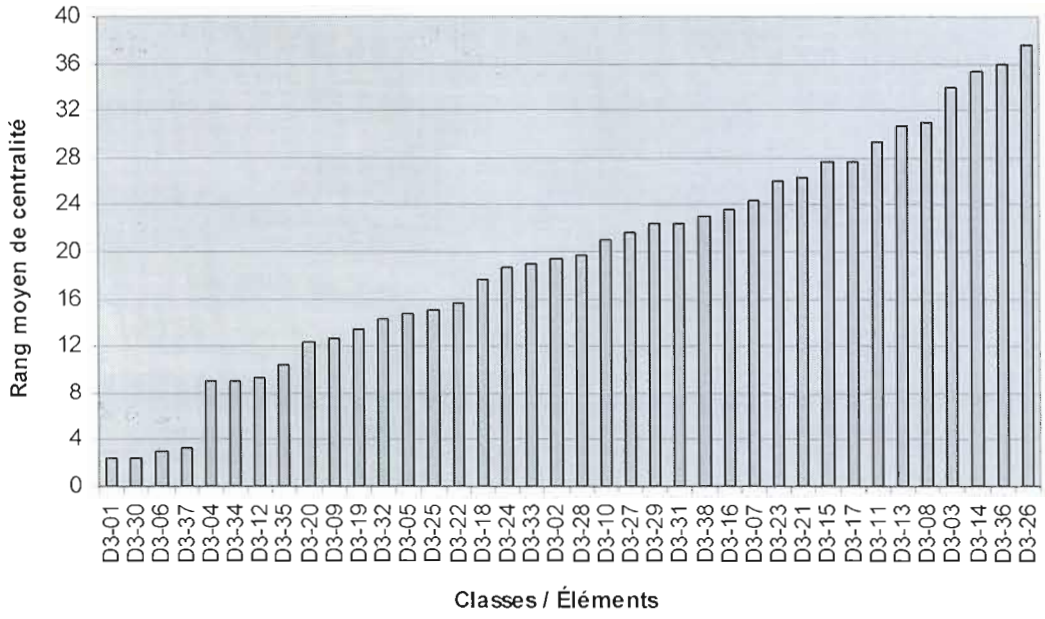


Figure 19 : Rang moyen de centralité des classes pour D3

Avec ces quatre classes, on retrouve l'une des composantes que nous avions identifiées précédemment avec le filtre au seuil. Ces résultats sur le sous-corpus D3, nous démontre qu'une application croisée des trois techniques d'analyse de la centralité des éléments de la RS produit des résultats convergents et permettent de trianguler un petit nombre d'éléments à la fois très objectivés dans le langage, très connexes entre eux dans la structure cognitive et largement partagés socialement parmi les membres.

Ces résultats pour D3 sont encourageants pour notre objectif et notre hypothèse de recherche. Il reste à vérifier si ce phénomène de convergence s'observe dans les autres sous-corpus. Des

calculs de corrélation (la corrélation de Pearson) entre scores de centralité sur les autres sous-corpus laissent entrevoir que c'est effectivement le cas.

3.2.2 Triangulation par les corrélations

Pour les 9 sous-corpus, nous avons calculé les trois différents scores de centralité (annexes XX à XXVIII). Le calcul de la corrélation de Pearson entre les scores de centralité — d'objectivation, de connexité et de consensus social — est une autre manière pour nous de valider par triangulation les résultats produits par notre méthode de forage de texte et les techniques développées pour le repérage du noyau central.

Dans le cas du sous-corpus D3, la centralité des éléments de la RS des AR en fonction de leur connexité est significativement corrélée à la centralité des éléments déterminée en fonction de leur degré d'objectivation. Ils sont corrélés à 0,6.

Ce phénomène s'observe pour l'ensemble des trois journaux et pour les trois périodes retenues. Le tableau 7 suivant est une matrice de corrélation entre les scores de connexité et les scores d'objectivation pour les 9 sous-corpus. Les corrélations sont moyennement fortes et varient entre 0,51 et 0,81. La corrélation moyenne est de 0,61 avec un écart-type de 0,1.

Tableau 7 : Corrélations entre les scores de connexité et d'objectivation

	Période 1	**Période 2**	**Période 3**	**Moyenne**
Le Devoir	D1 = 0,81	D2 = 0,73	D3 = 0,6	0,71
La Presse	P1 = 0,66	P2 = 0,6	P3 = 0,66	0,64
Le Soleil	S1 = 0,53	S2 = 0,51	S3 = 0,43	0,49
Moyenne	0,67	0,61	0,56	0,61

($\sigma = 0,1$)

C'est un résultat attendu par la théorie des RS que plus un élément est connexe dans la structure de la RS, plus il est objectivé, c'est-à-dire caractérisé par une forte valeur symbolique. Les résultats du tableau 7 sont en ce sens significatifs et nous indiquent que les

deux techniques d'analyse, bien qu'elles portent sur des propriétés différentes du noyau, se corroborent les unes les autres.

Dans le tableau 8, la matrice de corrélation est cette fois-ci entre les scores de connexité et les scores de consensus social. Les corrélations sont légèrement supérieures aux précédentes et varient entre 0,61 et 0,76. La corrélation moyenne est de 0,69 avec un écart-type très petit de 0,05.

Tableau 8 : Corrélations entre les scores de connexité et de consensus social

	Période 1	**Période 2**	**Période 3**	**Moyenne**
Le Devoir	D1 = 0,65	D2 = 0,75	D3 = 0,76	0,72
La Presse	P1 = 0,74	P2 = 0,71	P3 = 0,62	0,69
Le Soleil	S1 = 0,68	S2 = 0,61	S3 = 0,66	0,65
Moyenne	0,69	0,69	0,68	0,69

σ =0,05

Ces résultats sont à nouveau significatifs. Ils nous indiquent que les résultats générés par une technique d'analyse de la centralité basée sur la connexité sont corrélés aux résultats d'analyse de la centralité générés par une technique basée sur le consensus social. À nouveau, la triangulation des deux techniques illustre qu'elles se corroborent réciproquement.

Dans le tableau 9, nous avons finalement une matrice de corrélation entre les scores d'objectivation et les scores de consensus social. On remarque une légère diminution dans la force des corrélations, qui varient ici entre 0,25 et 0,75. La corrélation moyenne est de 0,54 avec l'écart-type le plus élevé de 0,16.

Tableau 9 : Corrélations entre les scores d'objectivation et de consensus social

	Période 1	Période 2	Période 3	Moyenne
Le Devoir	D1 = 0,75	D2 = 0,54	D3 = 0,57	0,62
La Presse	P1 = 0,74	P2 = 0,5	P3 = 0,67	0,64
Le Soleil	S1 = 0,43	S2 = 0,25	S3 = 0,38	0,35
Moyenne	0,64	0,43	0,54	0,54

($\sigma = 0,16$)

La convergence dans les résultats issus des deux techniques est malgré tout positive : lorsqu'une technique mesure une centralité forte sur une classe, l'autre technique permet d'observer des résultats similaires. Il semble que c'est surtout dans le journal Le Soleil que la force des corrélations a diminué, particulièrement dans la deuxième période (qui correspond aux élections provinciales de 2006-2007). Compte tenu de la force des corrélations observées ailleurs dans les deux autres journaux, cette diminution dans la convergence des résultats laisse supposer qu'elle ne soit pas nécessairement due à des limites dans les techniques développées, mais plutôt aux réalités empiriques de la RS des AR dans ce journal.

En effet, une faible corrélation dans les scores de centralité, comme c'est le cas pour S2, peut être un indicateur du degré de structuration de la RS. Peut-être avons-nous affaire dans ce cas à une RS en phase de transformation. Peut-être que le travail cognitif à propos des AR dans ce groupe et à cette période n'est carrément pas organisé à la façon des RS. Une RS émerge de conditions sociales et cognitives spécifiques, liées à différents processus de la pensée sociale qui n'ont peut-être pas eu lieu dans S2. Ce sont des pistes de recherche qu'il faut explorer empiriquement et que nous ne faisons que suggérer ici.

De manière générale, les résultats de corrélation sont très encourageants. Les trois techniques d'analyse de la centralité produisent des résultats convergeant de façon significative. Ce sont des résultats attendus par la théorie des RS : le noyau central d'une RS est composé à la fois d'éléments fortement objectivés, très connexes et qui font l'objet d'un consensus dans le groupe. Compte tenu de la robustesse de ces résultats, nous affirmons que la triangulation des

techniques d'analyse développées est une validation positive du forage de texte pour l'étude des RS. Ces résultats permettent également la corroboration d'une hypothèse fondamentale de la théorie des RS, selon laquelle une RS est structurée par un noyau central à la fois connexe, consensuel et objectivé.

Retour sur le quatrième objectif

Le quatrième objectif de cette recherche était de proposer des choix d'opérationnalisations de forage de texte mieux adaptés à la théorie des RS et à partir desquels nous pouvions développer différentes techniques d'analyse qui atteignent les trois niveaux d'analyse attendus par une méthode générale d'étude des RS. Cet objectif est considéré atteint.

Deux choix d'opérationnalisations en particulier ont permis cela : la concordance et le K-Means. Le K-Means, appliqué sur les résultats de la concordance, a permis de produire des partitions pour 9 sous-corpus, que l'on peut interpréter comme autant de cartographies de la RS. Un premier niveau d'analyse de la RS peut être mené sur cette cartographie. Il consiste à catégoriser et interpréter le contenu des classes. Ce travail peut être fait de plusieurs manières, par exemple en faisant appel aux techniques classiques de l'analyse de contenu. Cependant, lorsque nous avons affaire à plusieurs milliers de $domif_{(i)}$, l'analyse de contenu peut se révéler impraticable et il est utile dans ce cas de faire appel à d'autres techniques statistiques issues du forage de texte. Nous avons souligné deux d'entre elles : l'extraction de l'isotopie de chaque classe et l'extraction des $domif_{(i)}$ moyens.

Le deuxième niveau d'analyse, qui consiste à identifier les structures organisant les éléments de la RS, est atteint en développant deux techniques d'analyse basées sur la modélisation en réseau des résultats de la classification. Une première technique, inspirée de l'analyse de similitude, permet la modélisation de l'organisation des éléments de la RS sous la forme d'un réseau cognitif de proximité sémantique. Une deuxième technique, inspirée de l'analyse du consensus, permet la modélisation de l'organisation des éléments sous la forme d'un réseau sociocognitif de distribution sociale de la RS.

Le troisième niveau d'analyse, qui consiste à identifier le noyau central d'une RS, a été atteint via un calcul de centralité de degré dans les réseaux cognitifs et sociocognitifs. Les

résultats d'analyse de la centralité des éléments de la RS ont été validés par une démarche de triangulation des résultats. Cette stratégie avait pour objectif de tester la convergence dans les résultats des différentes techniques d'analyse. Un premier test de triangulation a été fait par l'analyse du rang moyen de centralité et un second test a été mené par l'analyse des corrélations dans les résultats. Ces deux tests sont une validation positive des techniques d'analyse développée et en amont, des choix d'opérationnalisations effectués sur la chaîne de forage de texte.

Par ailleurs, bien que notre opérationnalisation du forage de texte ait généré des résultats des plus intéressants pour l'étude des RS, nos choix ne sont pas les seuls possibles et d'autres sont probablement meilleurs encore. De futures recherches devront réfléchir et expérimenter différentes opérationnalisations et les évaluer.

CONCLUSION

Cette recherche explorait la pertinence et la compatibilité méthodologique entre le forage de texte et l'étude des RS. Notre question de recherche portait sur la possibilité pour les méthodes de forage de texte, d'atteindre les trois niveaux d'analyse des RS que doit permettre une méthode générale d'étude des RS. Nous avions émis l'hypothèse que le forage de texte peut effectivement permettre l'analyse multi-niveaux des RS, mais dans la mesure où des choix d'opérationnalisations spécifiques étaient faits.

L'étude des implications de cette hypothèse et de sa corroboration ou de sa réfutation s'est faites autour de 4 objectifs de recherche. Le premier (chapitre 2) était l'analyse du cadre théorique des RS et des principaux concepts en jeu. Cette étape était importante, car c'est à partir de ce cadre théorique, que doit être jugée la pertinence du forage de texte. Pour reprendre une expression de Berthelot (1990), les méthodes de forage de texte ne seront vraiment intéressantes pour les psychosociologues des RS que si elles respectent et s'adaptent aux exigences des « schèmes d'intelligibilités » de la théorie.

Le deuxième objectif (chapitre 3) était de présenter les différentes étapes et opérations impliquées dans une chaîne informatique de forage de texte. Cet objectif était nécessaire afin de rendre la plus transparente possible la nature du traitement des données textuelles. Nous avons, à la suite de Meunier *et al.* (2005, 2006), présenté cette chaîne de manière séquentielle sur six étapes : constitution du corpus, segmentation, indexation, vectorisation, réduction dimensionnelle et classification automatique.

Le troisième objectif était d'évaluer de manière critique les apports jusqu'à maintenant du forage de texte pour l'étude des RS. Depuis une dizaine d'années, l'application de méthodes de forage de texte dans ce domaine s'est essentiellement fait à l'aide d'une opérationnalisation et d'une implémentation particulière : celle du logiciel ALCSTE. L'apport de ce logiciel s'est cependant révélé limité dans sa capacité à générer des techniques d'analyse qui atteignent les trois niveaux d'analyse des RS.

Suite à ces limites, le quatrième objectif était de mener des expérimentations sur différents choix d'opérationnalisations du forage de texte, de manière à développer des techniques d'analyse des RS, qui permette à la fois d'identifier les éléments cognitifs de la RS étudiée, les structures cognitives et sociales qui l'organisent et de repérer son noyau central. Des choix d'opérationnalisations spécifiques aux étapes d'extraction des domaines d'information et de la classification, via la concordance et K-Means, ont permis d'atteindre le premier niveau d'analyse. Nous avons développé une première technique d'analyse de la classification qui a permis de modéliser formellement la structure cognitive des RS sous la forme d'un réseau cognitif de proximité sémantique. Nous avons ensuite développé une technique de modélisation de la structure sociale de la RS sous la forme d'un réseau sociocognitif de distribution sociale des éléments cognitifs d'une RS. Finalement, nous avons proposé des techniques pour le repérage du noyau central des RS à partir de mesure de centralité de degré dans les graphes. Finalement, les techniques développées ont été évaluées à l'aide de deux tests de triangulation des résultats issus d'une étude de cas empirique : la RS des accommodements raisonnables dans La Presse, Le Devoir et Le Soleil. Cette validation s'est révélée très encourageante compte tenu de notre objectif de recherche.

Au terme de ces quatre objectifs, nous sommes en mesure de corroborer notre hypothèse de départ. Il est possible de trouver des adéquations entre les « schèmes d'intelligibilité » de la théorie des RS et le forage de texte. Cependant, ceci demande une attitude particulière du chercheur par rapport à la technologie informatique, ce que certains (Meunier *et al.* 2006) ont appelé une utilisation « modulaire » de la technologie, où le chercheur peut intervenir à chaque étape via des choix d'opérationnalisations et même d'implémentation.

Cette recherche a également permis de corroborer une hypothèse fondamentale pour la théorie des RS selon laquelle le noyau est à la fois constitué des éléments les plus connexes dans la structure cognitive, les plus objectivés dans les communications et les plus consensuels parmi les membres du groupe social.

Les résultats obtenus dans cette recherche nous invitent à poursuivre dans cette voie. Dans de futures recherches, il faudra premièrement pousser plus loin l'analyse d'un cas empirique comme celui des accommodements raisonnables. Nous nous sommes limités ici à une analyse

minimale et démonstrative, qui était suffisante compte tenu de nos objectifs, mais mener une étude de cas complète serait un « test de validation » nécessaire et très heuristique.

Limites et ouvertures

Tout au long de cette recherche, plusieurs questions, limites et hypothèses ont émergé sans que nous puissions les aborder. Outre, selon nous, la nécessité de mener à bien une étude empirique de plus grande envergure, nous aimerions terminer avec deux commentaires généraux susceptibles d'orienter nos futurs travaux. Un premier concerne la théorie des RS en tant que telle, tandis que le deuxième concerne un choix d'opérationnalisation que nous souhaitons explorer.

Les éléments cognitifs

Tout au long de cette recherche, nous nous sommes contentés, suite à une suggestion de Flament et Rouquette (2003), de parler modestement « d'élément cognitif » pour décrire les cognitions composant une RS. Bien que pratique et prudent, ceci s'avère rapidement insuffisant. Il est nécessaire, selon nous, de préciser davantage ce que les théories de la pensée sociale en général, et des RS en particulier, entendent par « cognition ». En effet, on peut supposer que des notions comme celles de *schème, de prototype, de schémata, de grille de lecture, de stéréotype, de cadre, de cognème*, et bien d'autres, qui sont toutes des notions interpellées, ne renvoient pas cependant à des cognitions de même nature.

Il faut, selon nous, que la théorie des RS précise davantage son concept de cognition. Une meilleure compréhension du rôle du langage est certainement déterminante dans ce cas. Pourtant, sur ce point également la théorie des RS reste discrète, on se contente souvent de ne le considérer que comme un « truchement méthodologique ».

Peut-être faudra-t-il sortir du champ théorique des RS pour apporter quelques éléments de réponse. L'anthropologie cognitive étasunienne nous semble, dans ce cas, un domaine à explorer en priorité (D'Andrade 1995; Shore 1996; Hutchins 1995; Strauss et Quinn 1998). Il y a également, du côté de la psychologie sociale cognitive, des avancées considérables qui ont été faites (Kunda 1999; Bless *et al.* 2004).

En ce sens, Kunda (Kunda 1999) propose comme dénominateur commun à ces différentes notions pour décrire le concept de cognition, le concept de *concept* : « Because the nature of representation has been a topic of some debate, I have chosen to use the simplest and most familiar of the general terms, concept [...] » (Kunda 1999: 17). Pour Kunda, c'est le concept qui est à la base des processus cognitifs de la pensée sociale : « The many crucial functions of concepts include classification, inferring additional attributes, guiding attention and interpretation, communication, and reasoning. » (Kunda 1999 : 17)

Il est intéressant de se rappeler que Durkheim avait lui-même accordé un rôle de premier plan au processus cognitif de la conceptualisation. Pour Durkheim, le concept est une cognition à la base de la communication et de la vie sociale en général. Ce sont les concepts qui, sous certaines conditions spécifiques, se transforment en représentation collective :

> « Nous ne comprenons qu'à condition de penser par concepts. » (Durkheim 1914)
>
> « [...] la conversation, le commerce intellectuel entre les hommes consistent dans un échange de concepts. » (Durkheim 1912)
>
> « [...] les concepts sont-ils l'instrument par excellence de tout commerce intellectuel. C'est par eux que les esprits communient. » (Durkheim 1914),

Mais, même le concept de *concept* soulève encore aujourd'hui des polémiques théoriques importantes (Meunier 2006) que nous ne pouvons ici traiter. Il s'agit néanmoins d'une piste prometteuse à explorer.

La classification dure

Notre deuxième commentaire, d'ordre méthodologique celui-là, concerne les limites de la classification *dure* pour la cartographie de la RS. Nous avons observé, durant nos expérimentations, que le fait de ne pouvoir classer que de manière exclusive les segments de texte extraits du corpus pouvait être problématique. En effet, il est tout à fait admissible de considérer qu'un même segment de texte puisse être l'expression dans le langage de plus d'un élément cognitif à la fois et, par conséquent, d'être classé dans plus d'un ensemble à la fois.

Notre choix d'opérationnalisation avec K-Means ne permet pas cela (ALCESTE non plus). Malgré que nous ayons obtenu de bons résultats durant la triangulation des résultats, en les analysant plus en détail on constate des observations difficiles à intégrer dans le cadre théorique des RS. Ceci se manifeste en particulier dans la modélisation de la structure sociale sous la forme d'un réseau sociocognitif. Selon la théorie des RS, on devrait s'attendre non seulement à ce que le noyau soit composé d'éléments cognitifs largement partagés socialement par les membres du groupe, mais en plus, on devrait s'attendre à ce que ce noyau couvre la totalité des individus selon le principe des « des majorités alternatives non exclusives » (Flament 1996 : 120; Flament et Rouquette 2003).

Si nous reprenons le sous-corpus D3, qui nous a servi à plusieurs reprises d'illustration, on remarque que, bien que le noyau constitué de quatre éléments soit largement distribué dans la population de journalistes, il ne la couvre pas en entier, mais seulement 59% des individus qui la composent (figure 20).

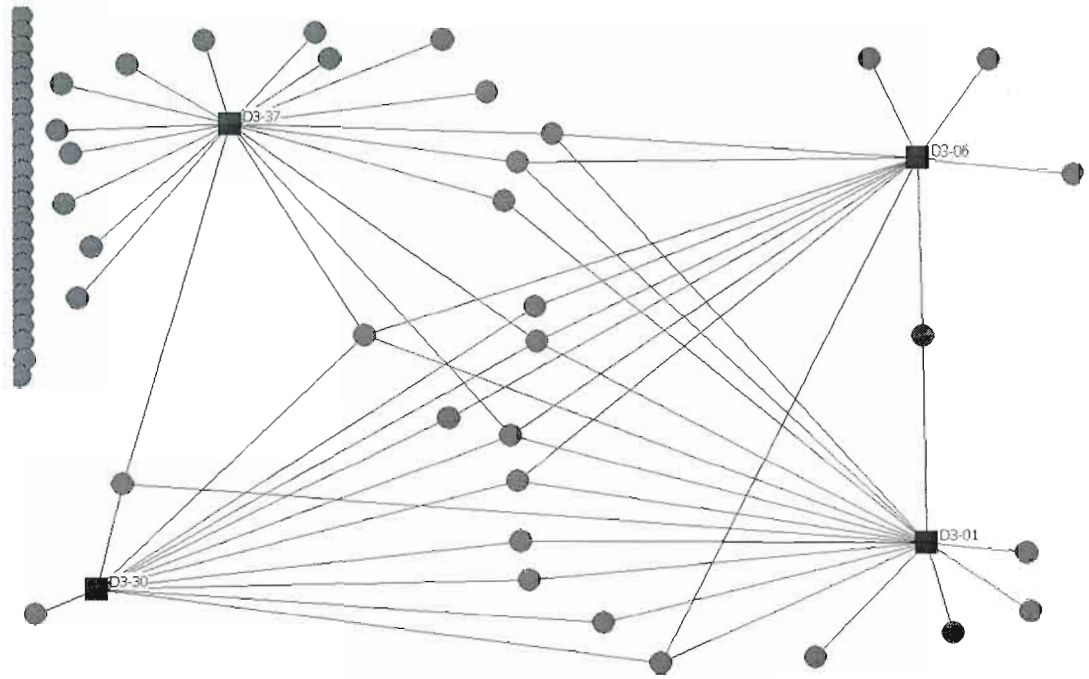


Figure 20 : Réseau sociocognitif du noyau de la représentation sociale des accommodements raisonnables dans D3

Selon nous, ceci est en partie explicable par la rigidité des algorithmes de classification exclusive, comme celui utilisé durant nos expérimentations. Dans de futures expérimentations, il serait intéressant d'observer les résultats que généreraient différents algorithmes de classification plus souples, basés sur la classification probabiliste, c'est-à-dire en termes de degré d'appartenance d'un segment de texte à une classe. Ce genre d'algorithme devrait augmenter la densité des réseaux sociocognitifs produits.

En somme, l'utilisation des méthodes de forage de texte pour l'étude des RS est encore à un stade exploratoire et beaucoup reste à faire. L'horizon des possibilités qu'offrent ces méthodes pour ce domaine d'étude et pour les sciences humaines et sociales en général commence à peine à être entrevu.

ANNEXES

Annexe I : Liste de mots fonctionnels sous forme lemmatisée

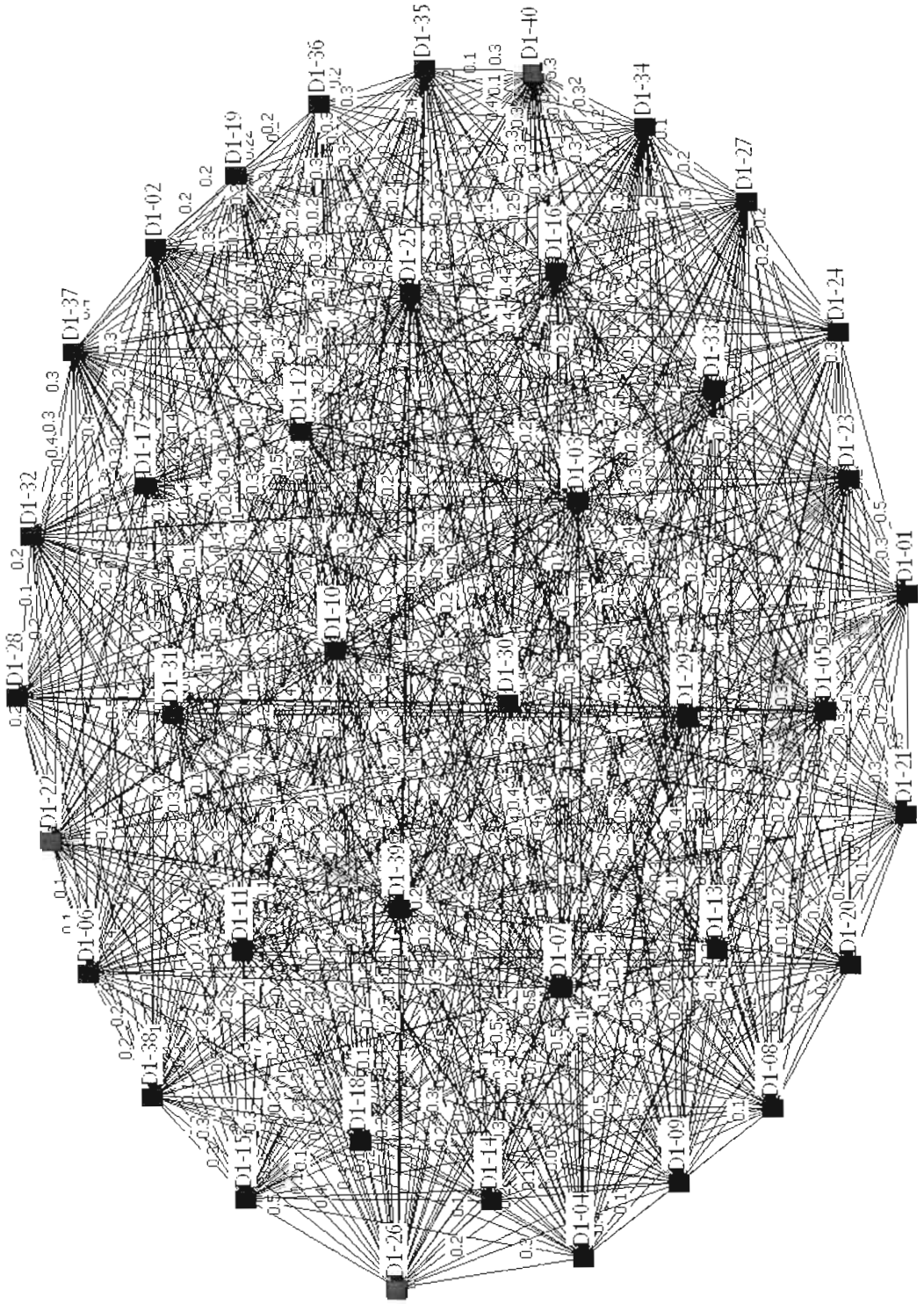
abord	accord	Actuel	affil	afin	ah
ai	aie				
aient	ailleur	Ains	ait	alentour	ali
alor	ap				
apr	apres	Arrier	as	assez	au
aucun	audit				
aujourd	auparav	Aupres	auquel	aur	aurion
auron	auront				
auss	aussitôt	aut	autour	autr	autrefois
aux	auxdit				
auxquel	avaient	avais	avait	avant	avec
avez	avi				
avion	avoir	avon	ayant	ayez	ayon
bah	banco				
bas	bé	beaucoup	ben	bien	bientôt
bis	ca				
ça	çà	cah	cahin	car	ce
céan	cec				
cel	celui	cent	cepend	cert	certain
cet	ceux				
cf	cg	cgr	chacun	chaqu	cher
chez	ci				
cinq	cinqu	cl	clair	cm	cm^2
combien	comm				
comment	consider	contr	contrario	crescendo	dan
davantag	de				

debout	dedan	defaçon	dehor	dejà	déjà
del	delà				
demain	demanier	depuis	derechef	derri	des
desdit	désorm				
desquel	dessous	dessus	deux	dev	dever
dg	die				
dir	direct	dito	diver	divers	dix
dl	dm				
donc	dont	dorénav	douz	du	dud
duquel	dur				
e	égal	eh	elle	emblé	en
encor	encoursd				
eneffet	enfin	enfonct	ensuit	entre	enver
environ	es				
essentiel	est	estpourquoi	et	et/ou	étaient
étais	était				
étant	etc	ête	été	éti	étion
être	eu				
eue	euh	eûm	eurent	eus	euss
eussent	eussion				
eut	eût	eux	évident	expres	extenso
extrem	facil				
facto	flac	fortior	fûm	fur	furent
fus	fuss				
fussent	fussion	fut	fût	général	ghz
gr	gràceà				
gràceau	grosso	guer	ha	han	haut
hé	hein				
hem	heu	hg	hi	hl	hm
hm^3	holà				
hop	hor	horm	horsd	hui	huit

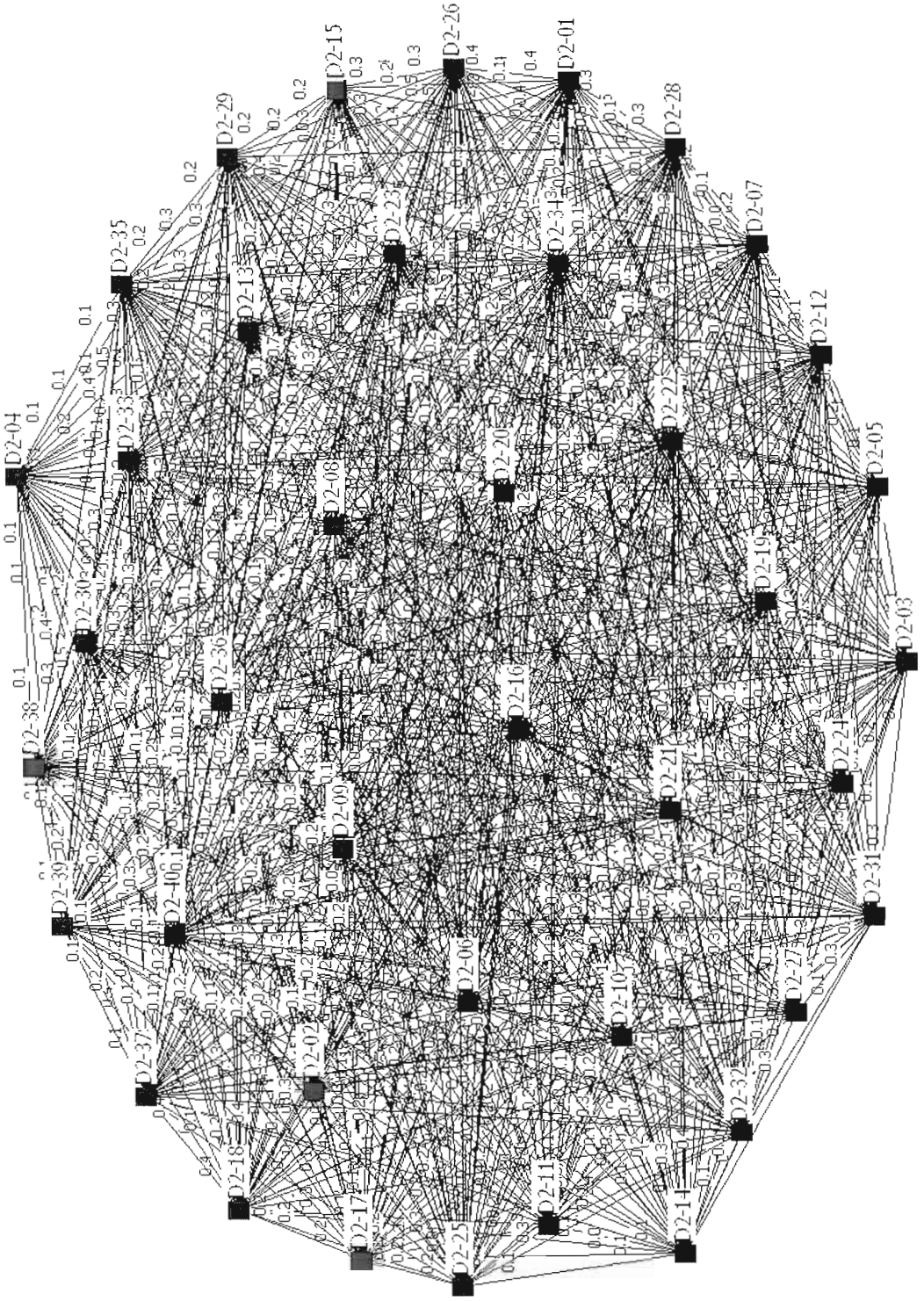
hum	ibidem				
ici	idem	il	illico	in	ipso
item	jad				
jam	je	jusqu	kg	km	km²
la	là				
laquel	le	lequel	les	lesquel	leur
lez	loin				
longtemp	lor	lorsqu	lui	m²	m³
ma	maint				
mainten	mais	malgr	me	mêm	mg
mgr	mhz				
mieux	mil	mill	milliard	million	minim
mis	ml				
mm	mm²	modo	moi	moin	mon
moult	moyen				
mt	naguer	ne	néanmoin	neuf	ni
n°	non				
nonobst	normal	nos	not	notr	nous
nul	null				
octant	oh	on	ont	onze	or
ou	où				
ouais	oui	outr	par	parbleu	parc
parcequ	parexempl				
parfois	parm	particulier	partout	pas	pass
passim	pend				
petto	peu	peut	peuvent	peux	pis
plupart	plus				
plusieur	plutôt	point	posterior	pour	pourquoi
pourt	préalabl				
pres	presqu	primo	prior	probabl	prou
puis	puisqu				

qu	qua	quand	quant	quar	quas
quatorz	quatr				
que	quel	quelqu	quelquefois	qui	quiconqu
quinz	quoi				
quoiqu	rapid	récent	revoic	revoilà	rien
sa	san				
sauf	se	secundo	seiz	selon	sensu
sept	ser				
serion	seron	seront	seul	si	sic
sin	sinon				
sitôt	situ	six	soi	soient	sois
soit	soix				
somm	son	sont	soudain	sous	souvent
soyon	stricto				
suis	suiv	sur	surtout	sus	ta
tacatac	tant				
tantôt	tard	te	tel	temp	ter
toi	ton				
tôt	toujour	tous	tout	toutcfois	treiz
trent	tres				
trois	trop	tu	un	une	usd
ver	vers				
via	vic	vingt	vis	vit	vitro
vivo	voic				
voilà	voir	volonti	vos	votr	vous
zéro					

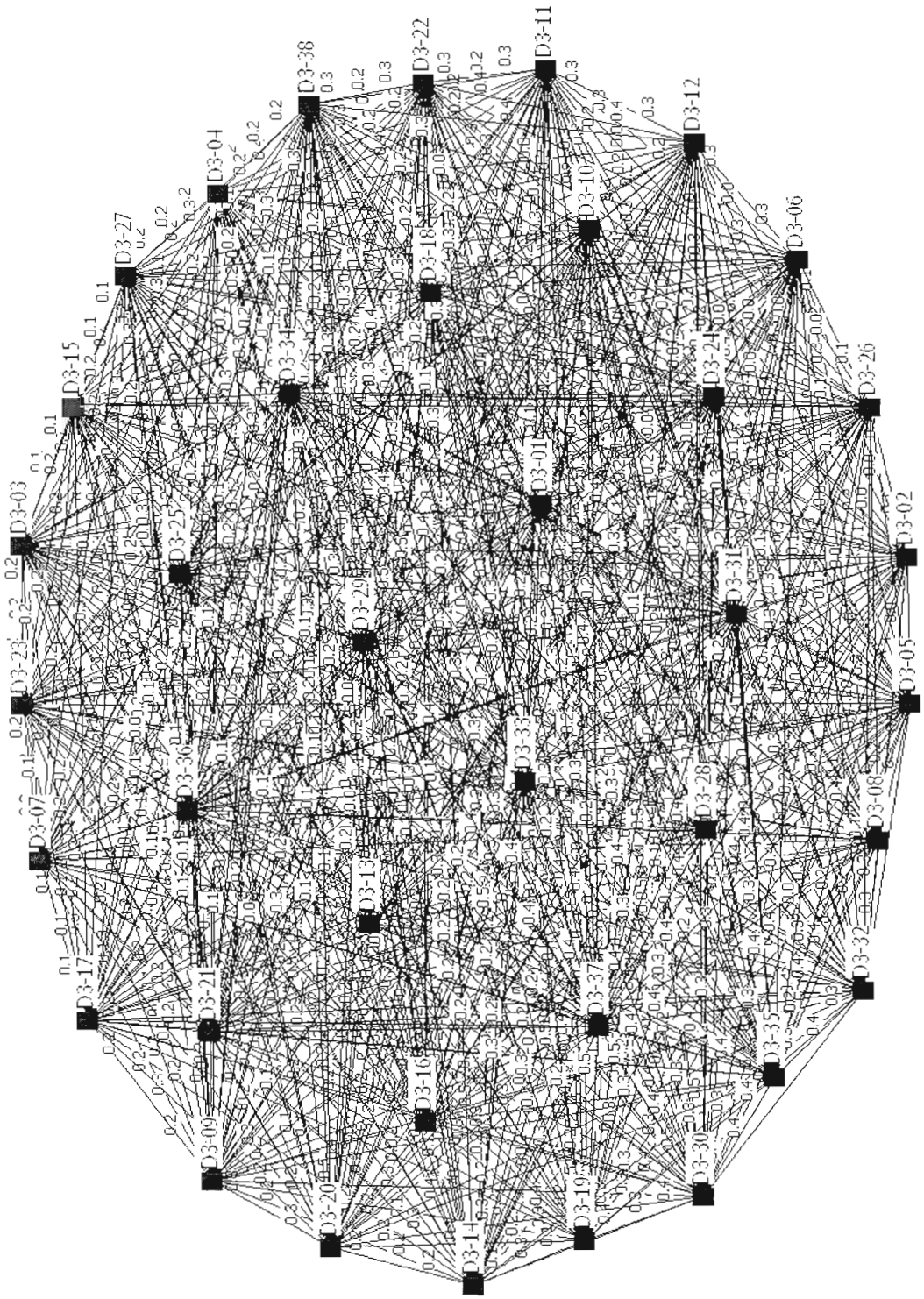
Annexe II : Réseau cognitif de proximité sémantique dans D1



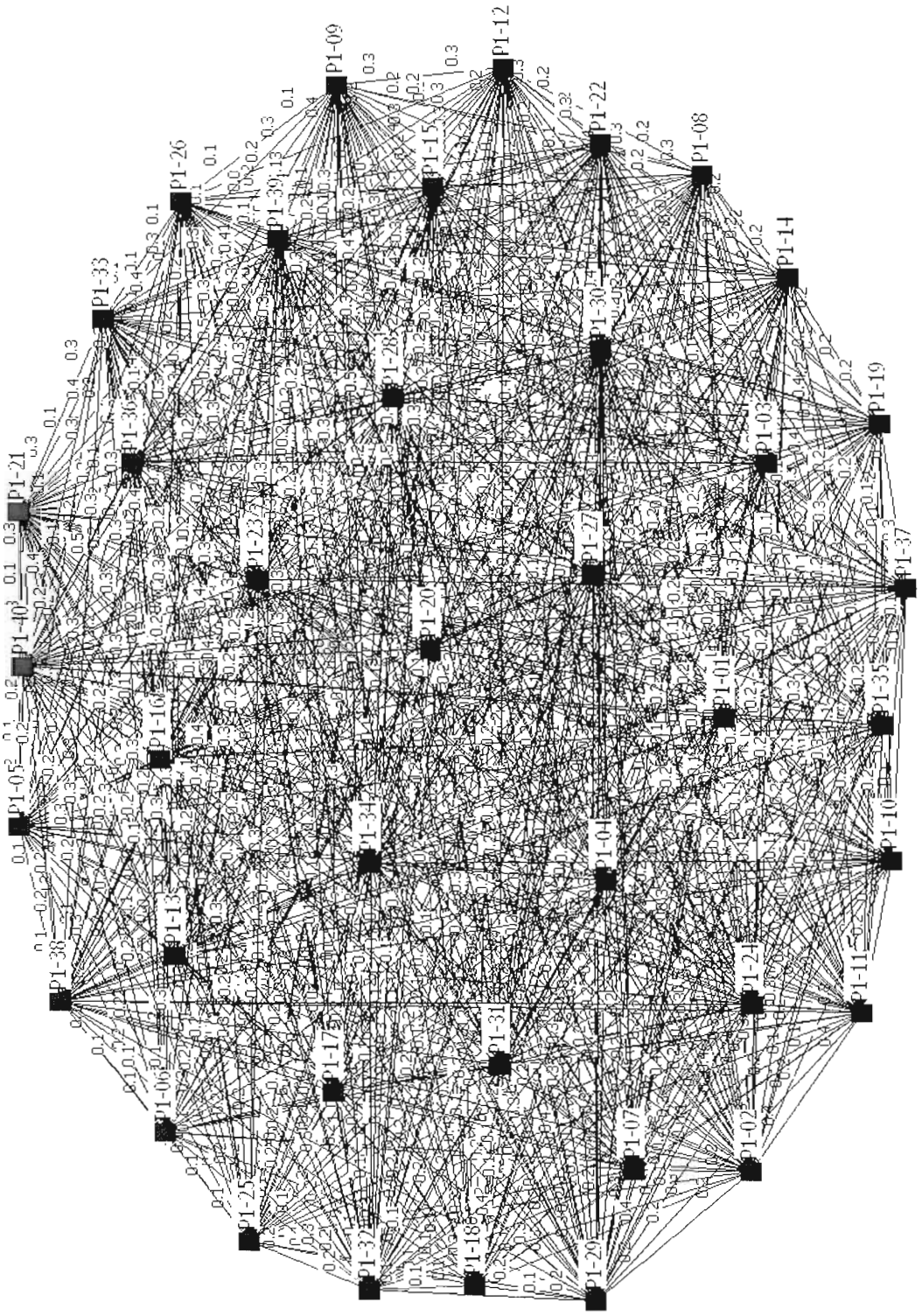
Annexe III : Réseau cognitif de proximité sémantique dans D2



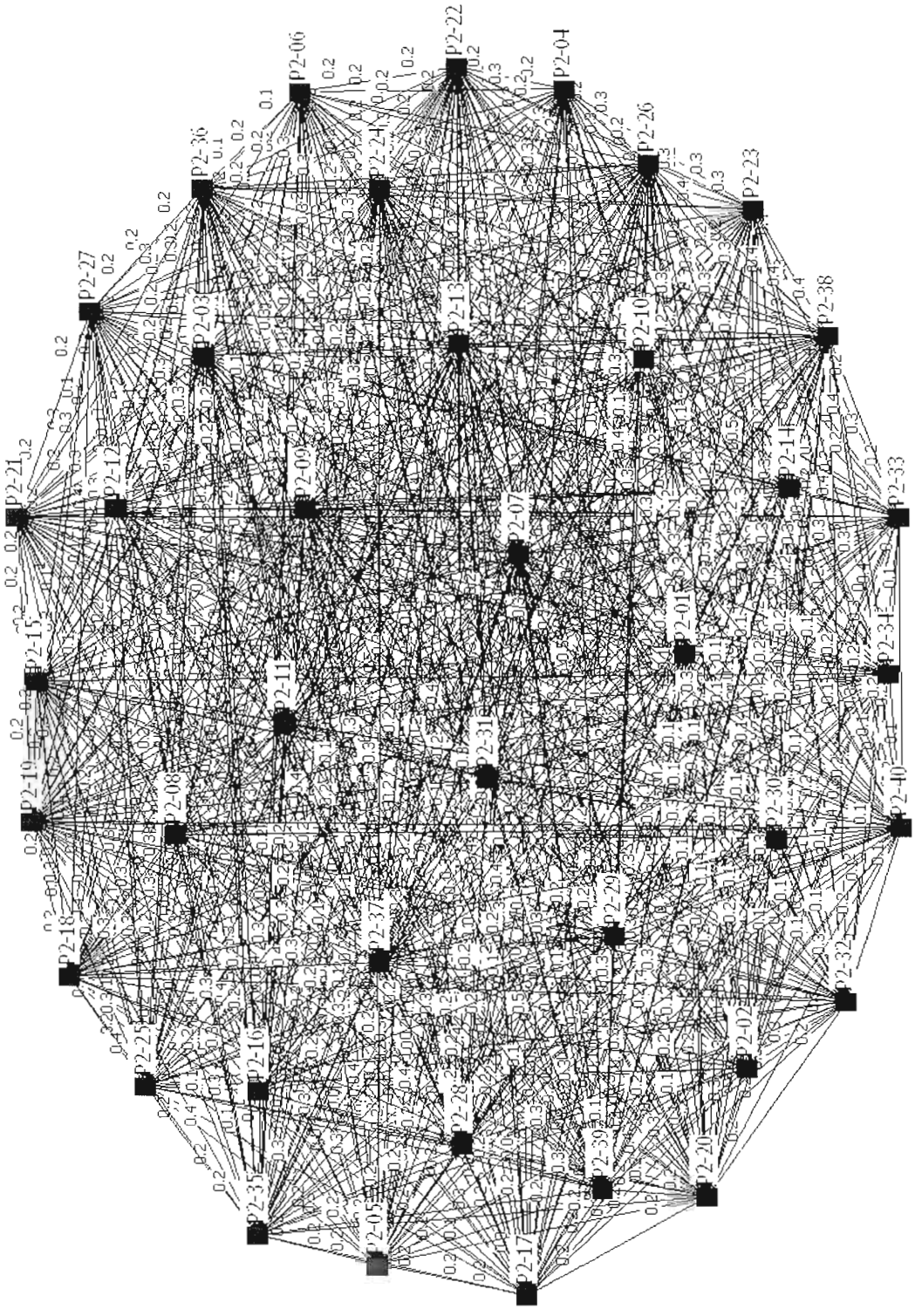
Annexe IV : Réseau cognitif de proximité sémantique dans D3



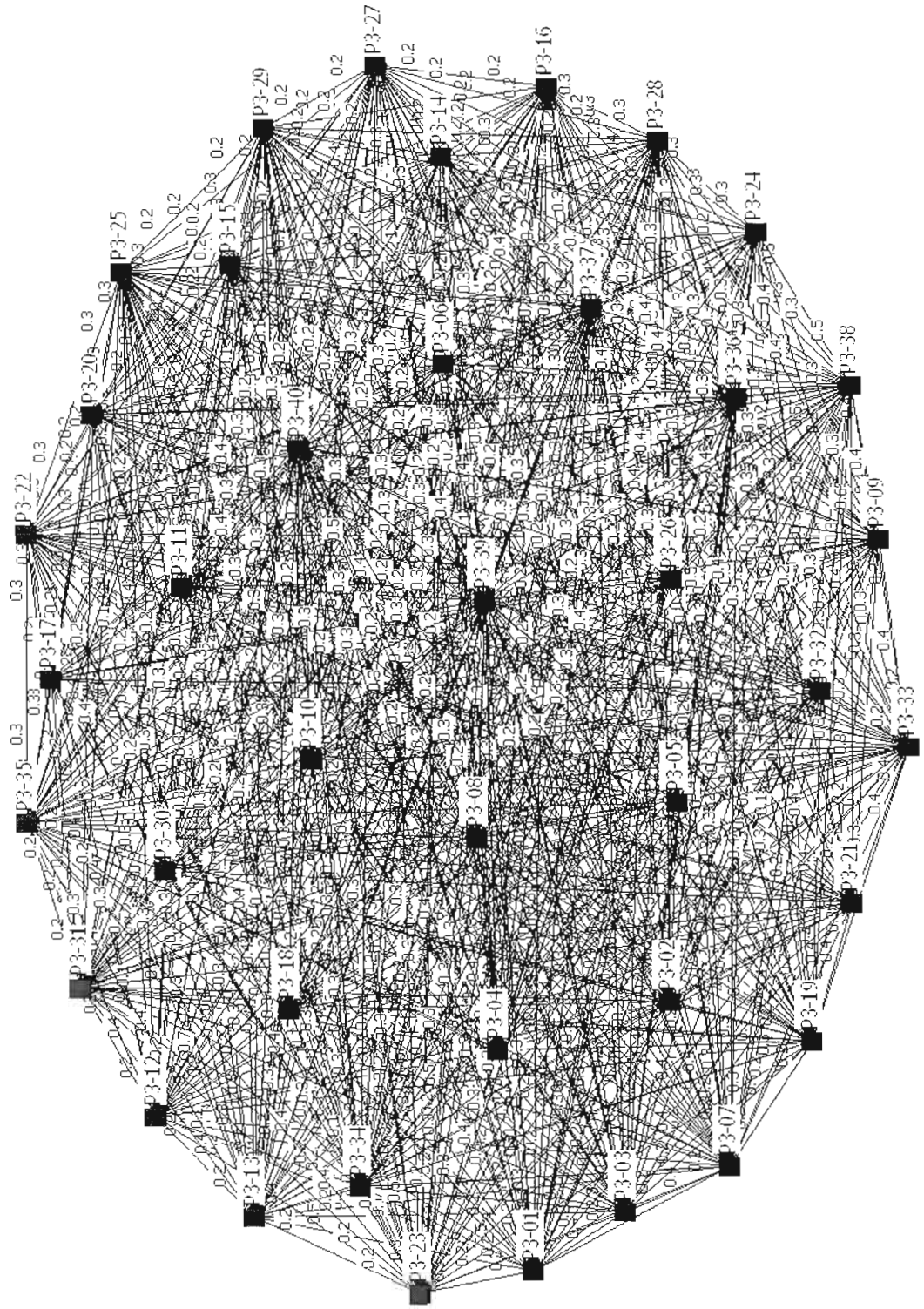
Annexe V : Réseau cognitif de proximité sémantique dans P1



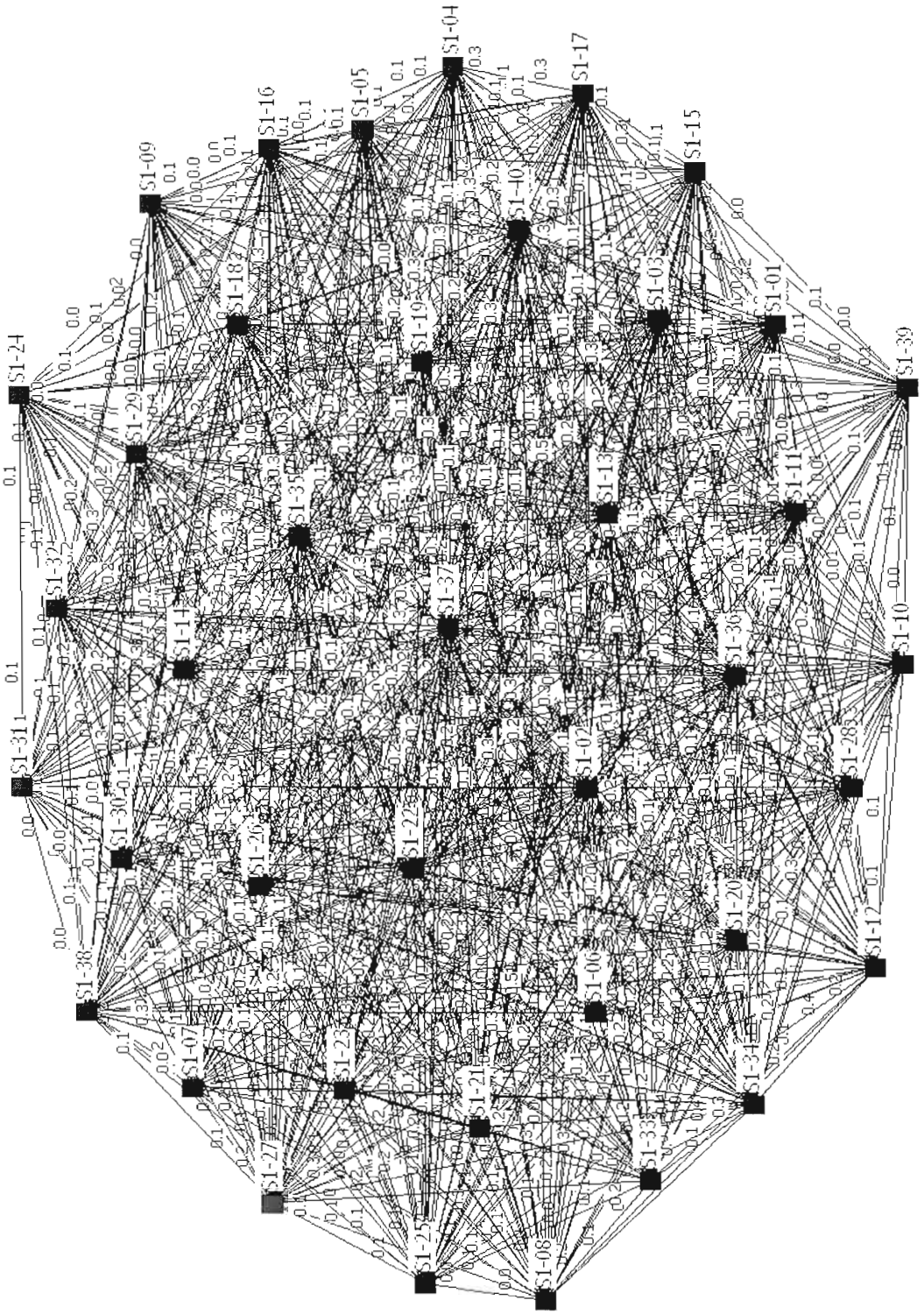
Annexe VI : Réseau cognitif de proximité sémantique dans P2



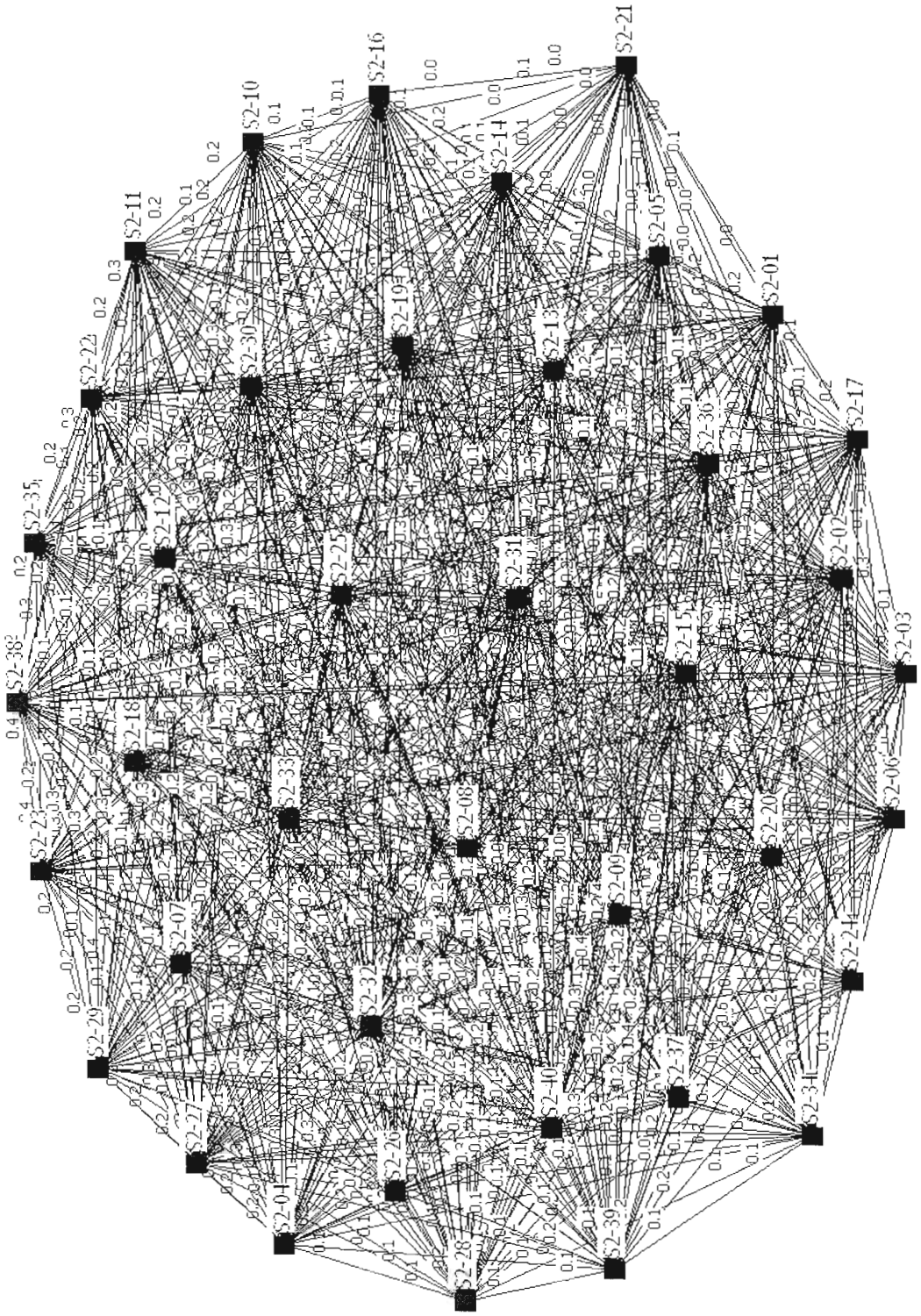
Annexe VII : Réseau cognitif de proximité sémantique dans P3



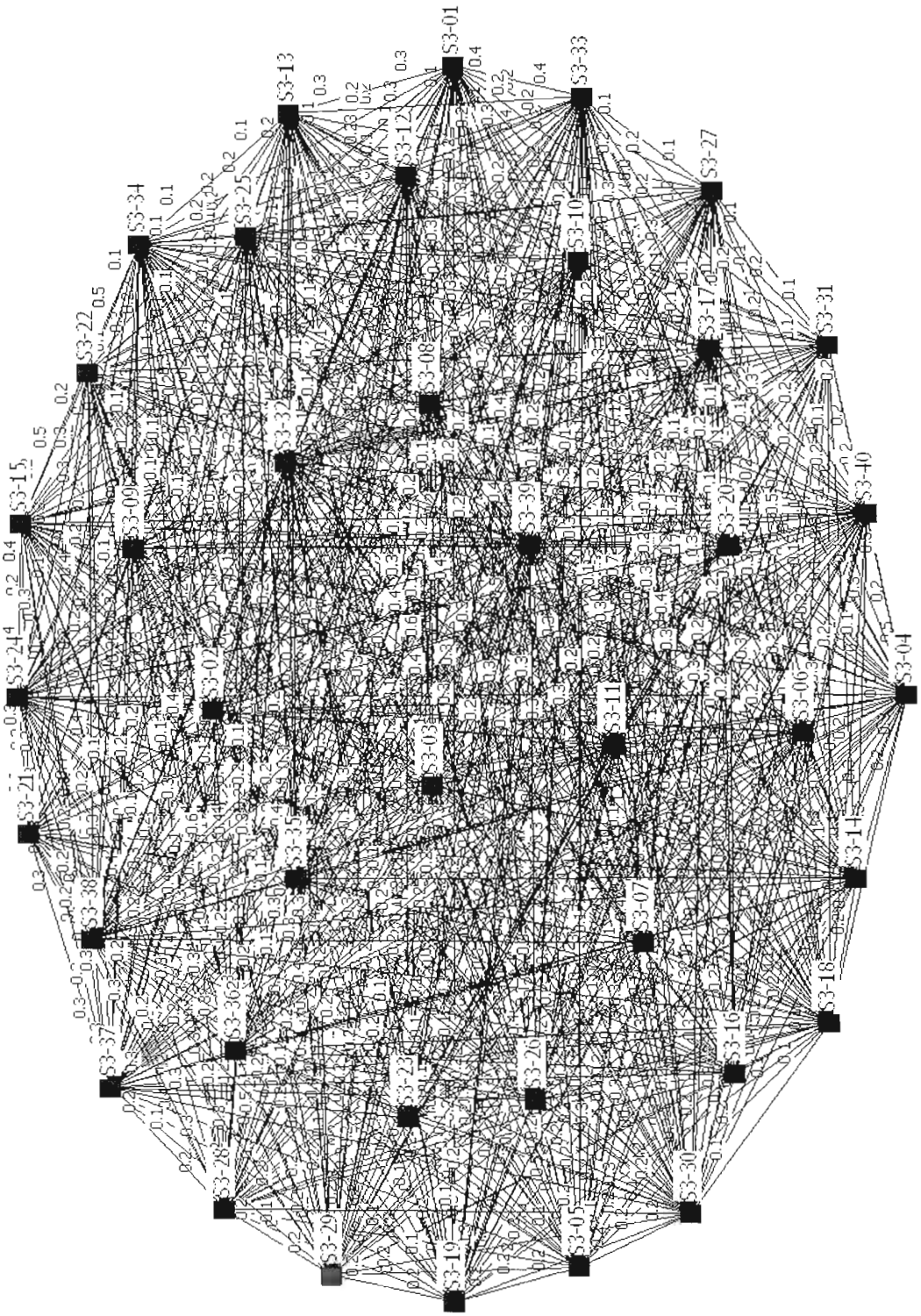
Annexe VIII : Réseau cognitif de proximité sémantique dans S1



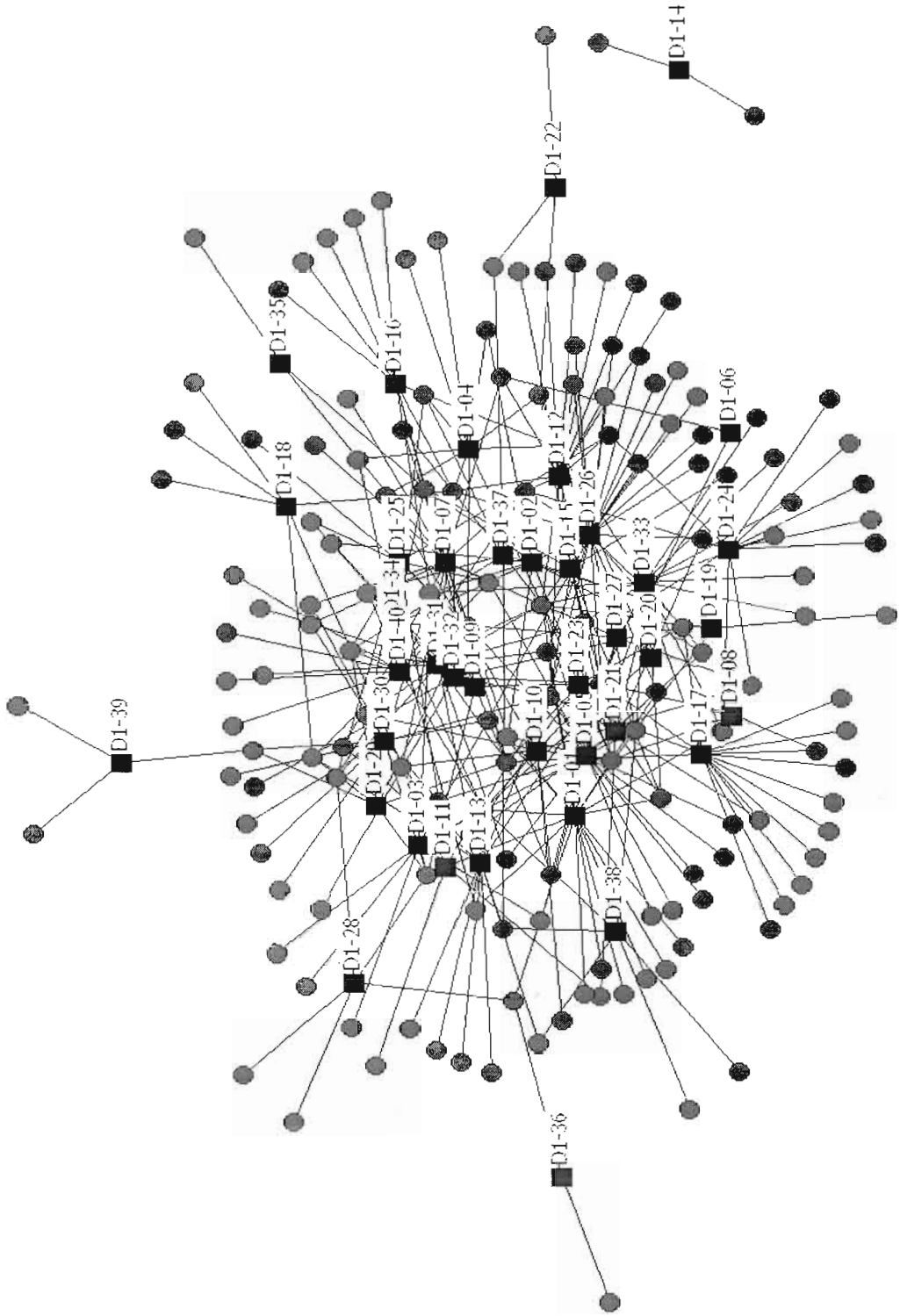
Annexe IX : Réseau cognitif de proximité sémantique dans S2



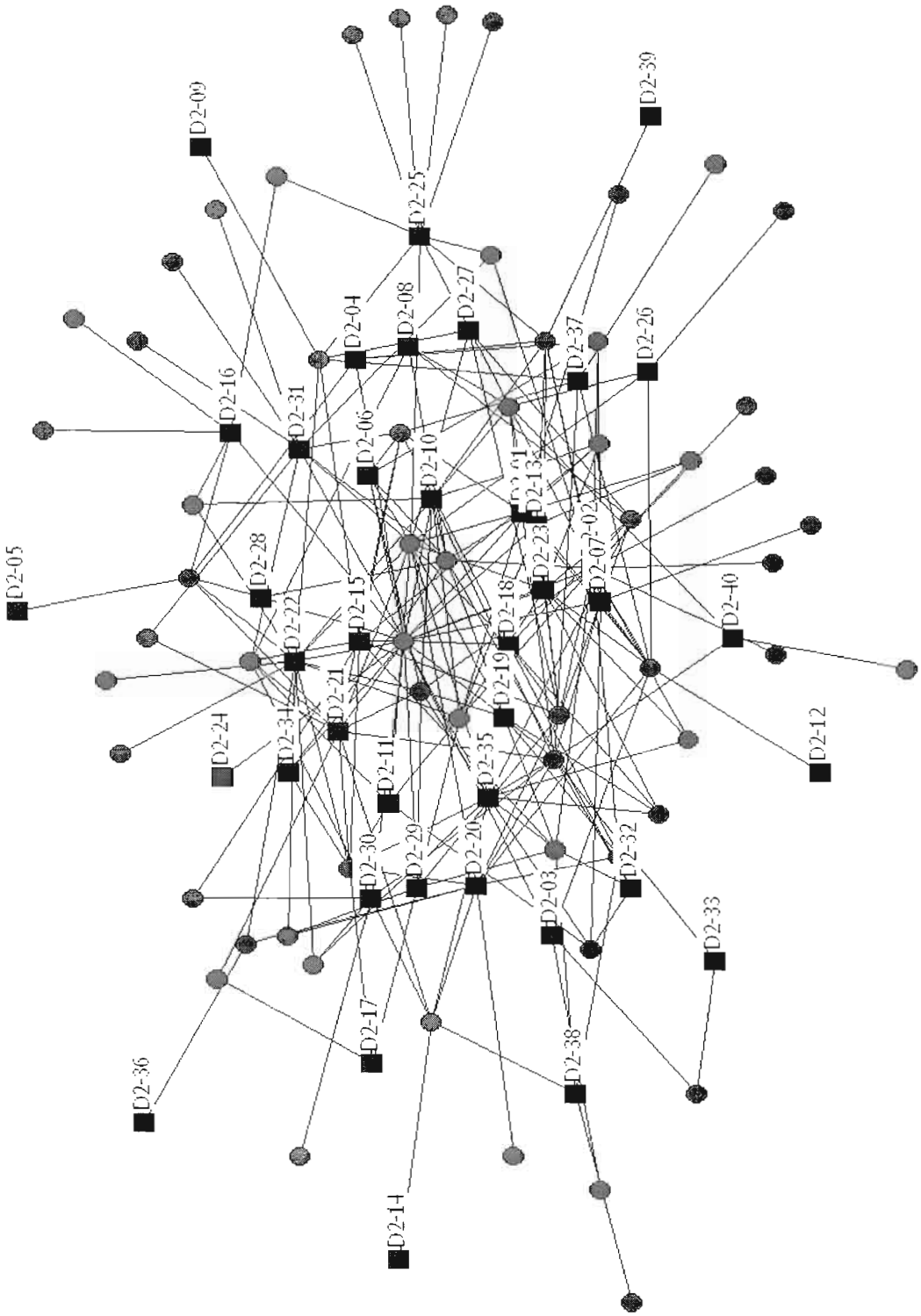
Annexe X : Réseau cognitif de proximité sémantique dans S3



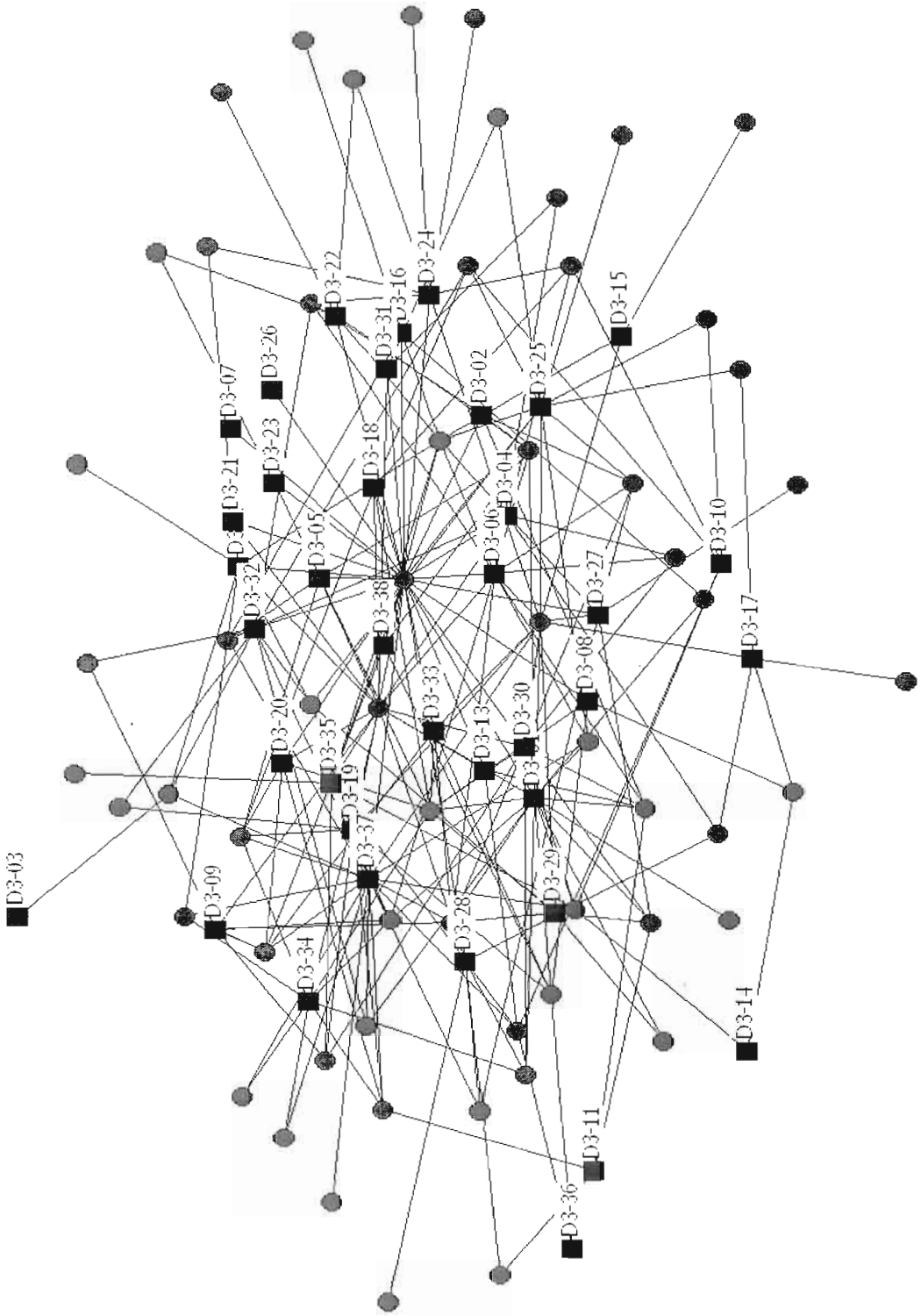
Annexe XI : Réseau sociocognitif dans D1



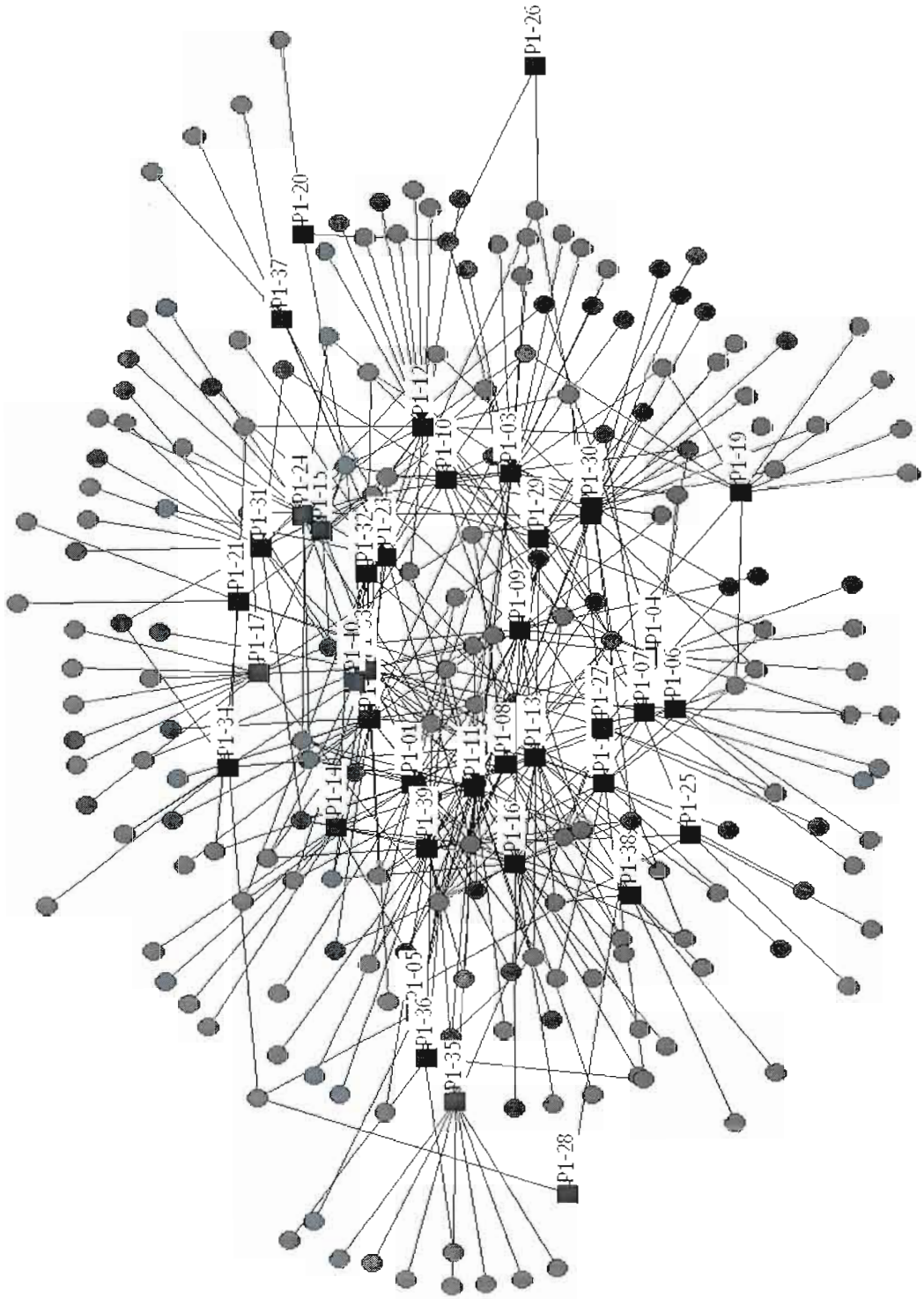
Annexe XII : Réseau sociocognitif dans D2



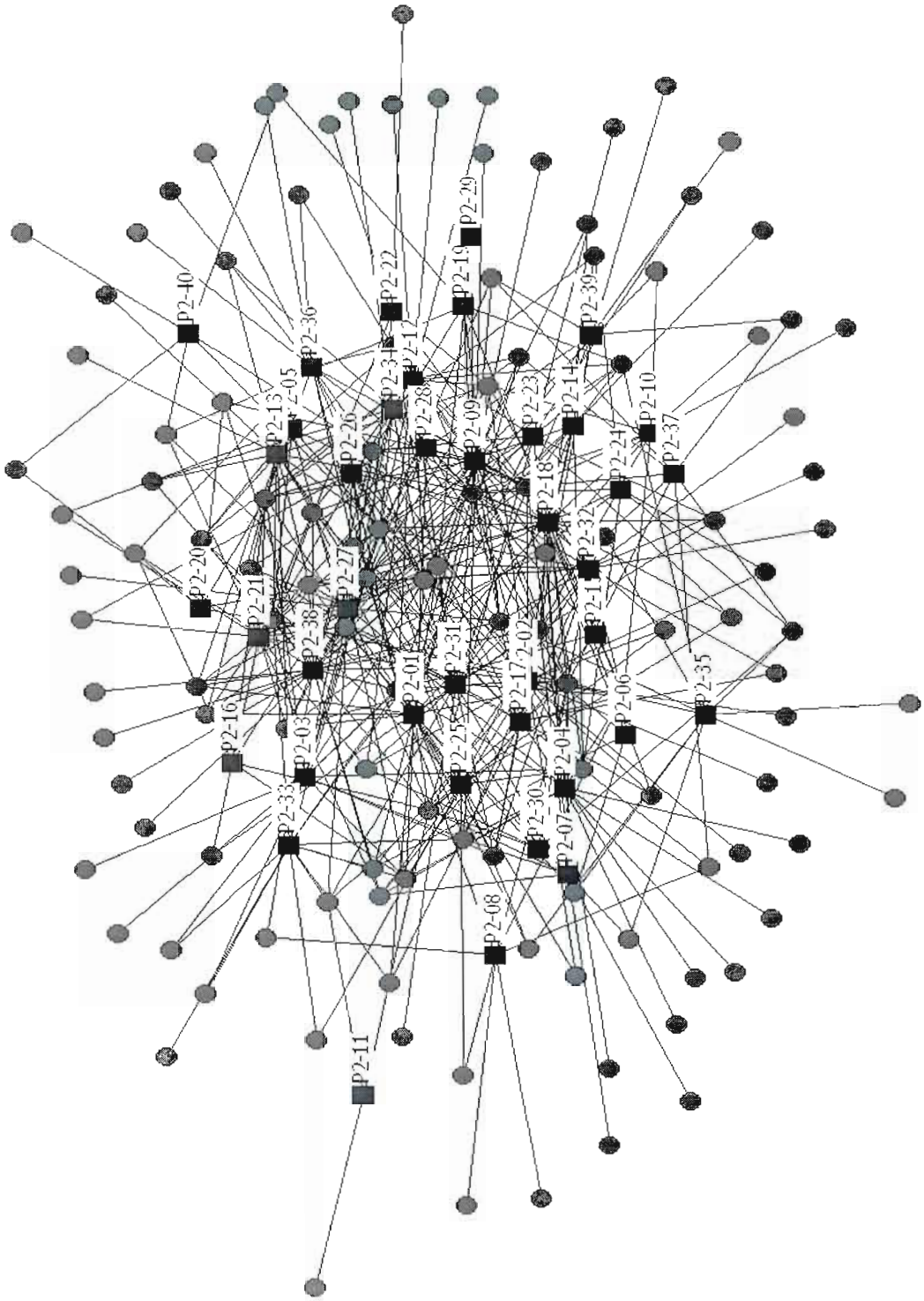
Annexe XIII : Réseau sociocognitif dans D3



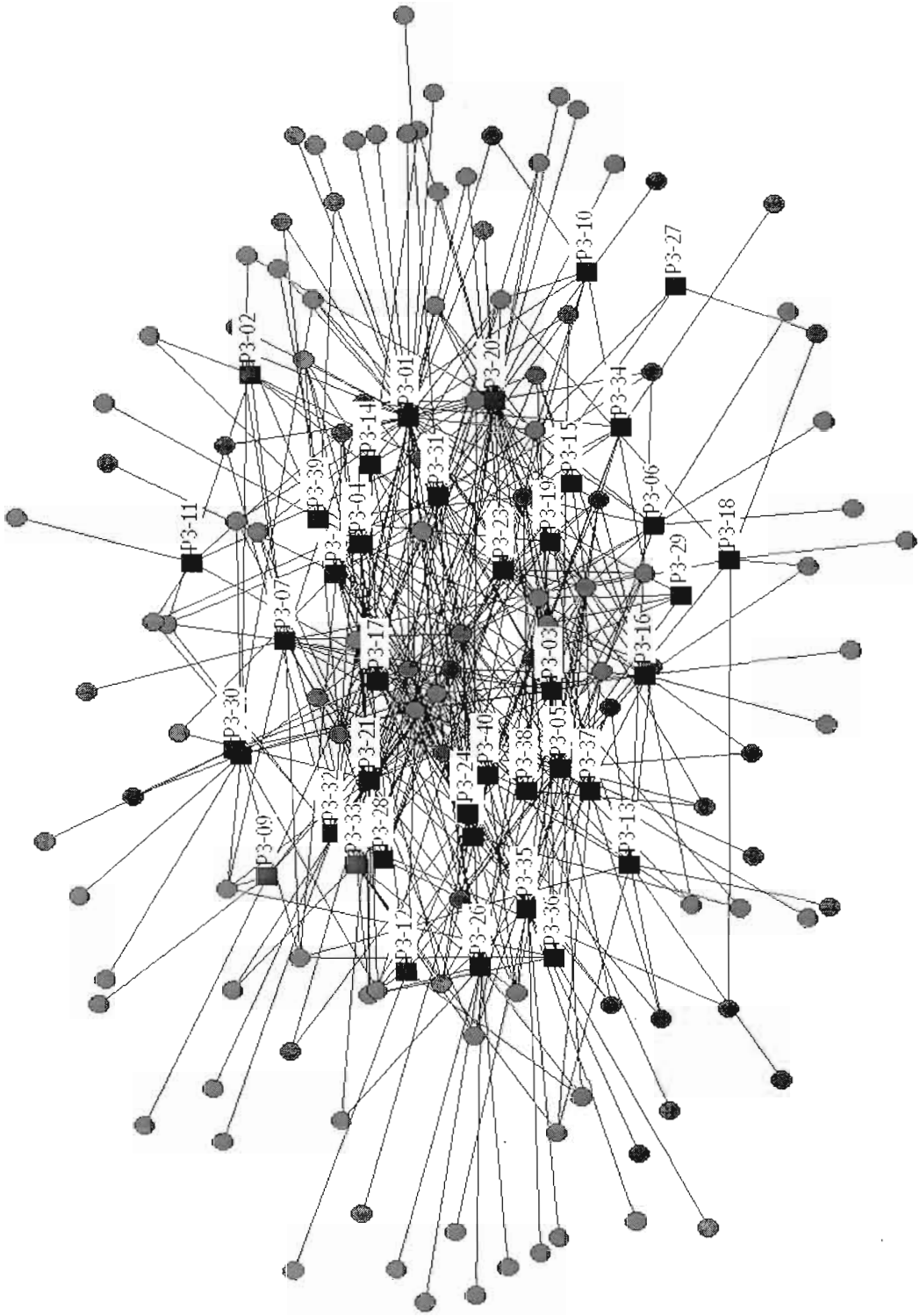
Annexe XIV : Réseau sociocognitif dans P1



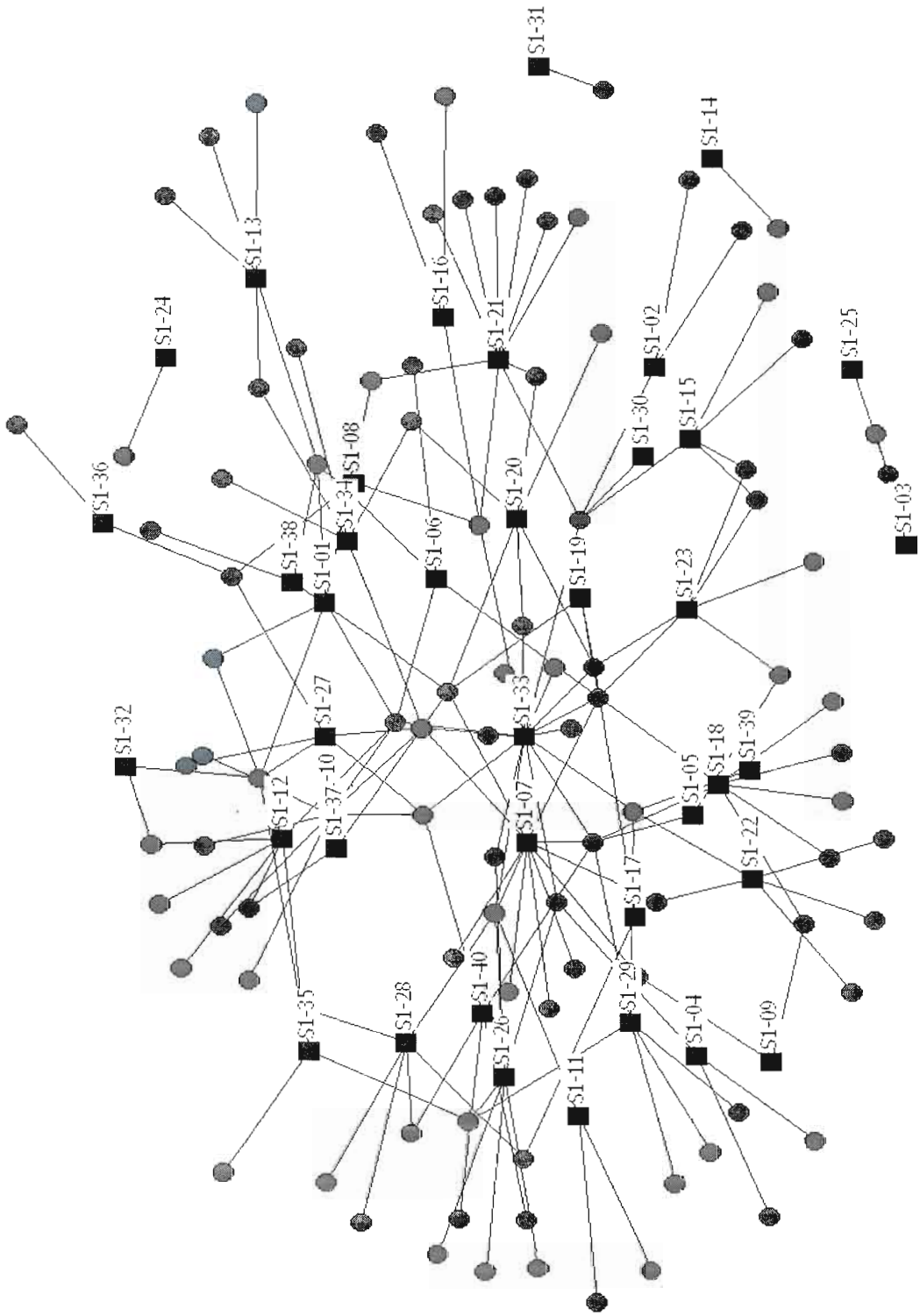
Annexe XV : Réseau sociocognitif dans P2



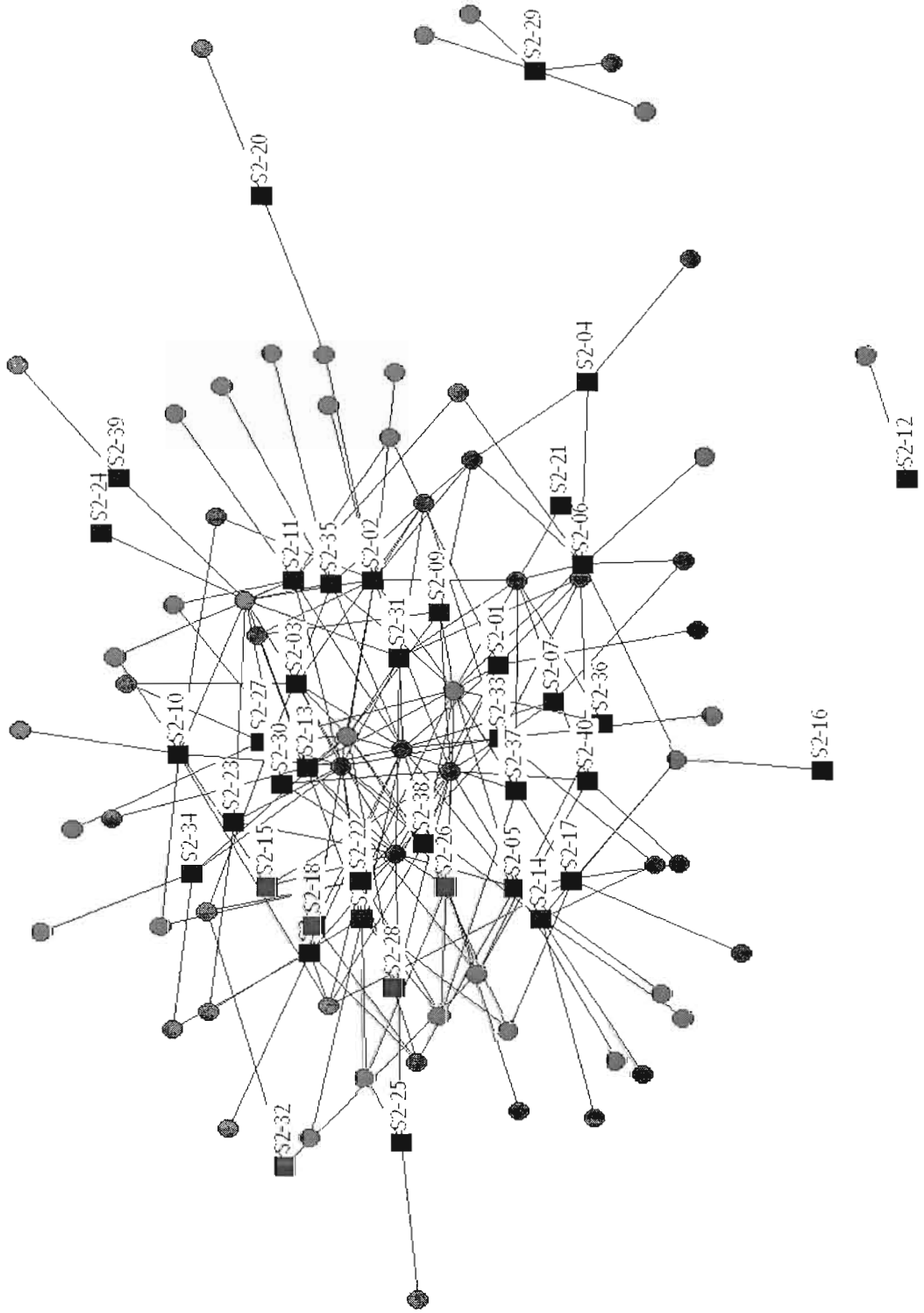
Annexe XVI : Réseau sociocognitif dans P3



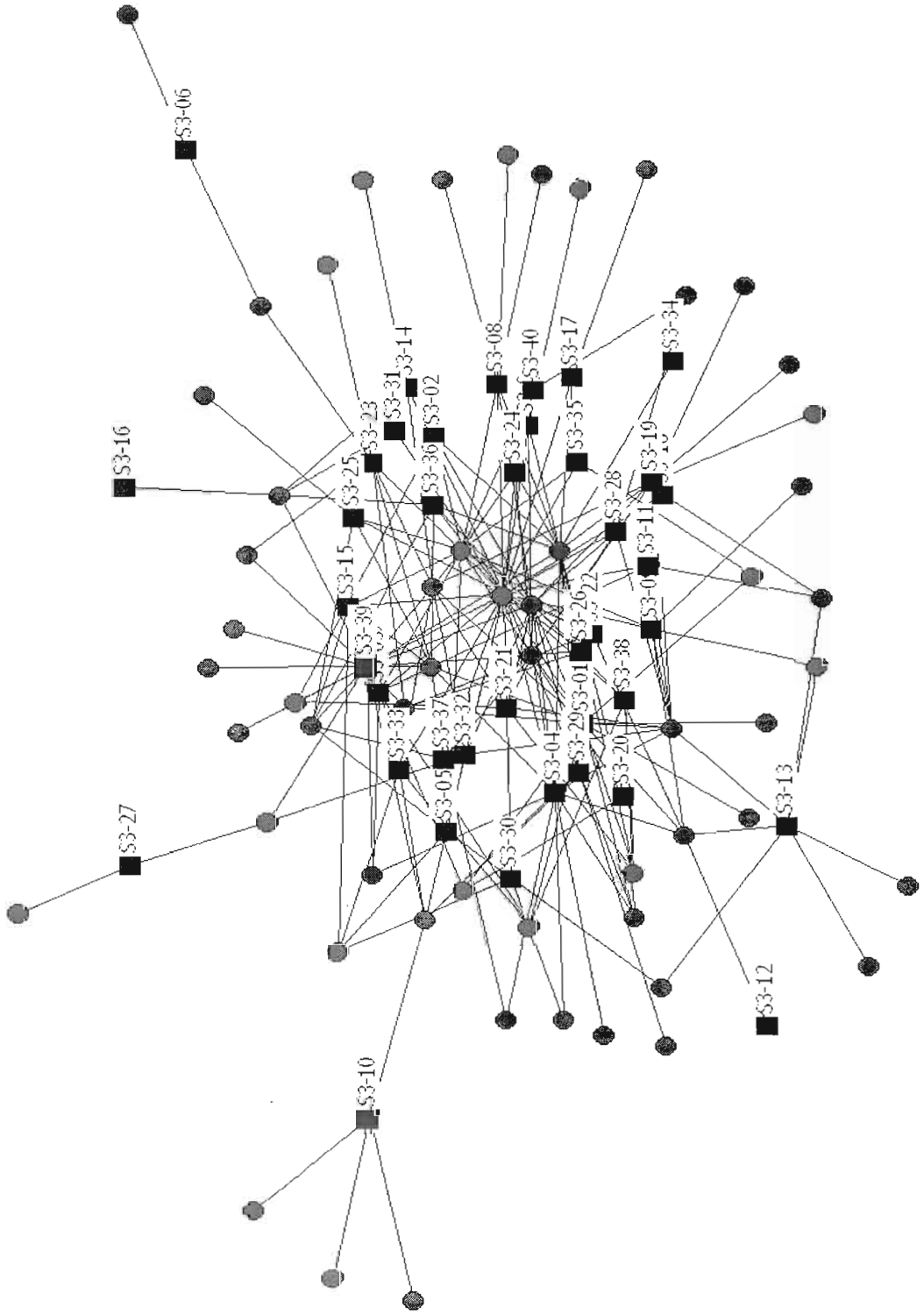
Annexe XVII : Réseau sociocognitif dans S1



Annexe XVIII : Réseau sociocognitif dans S2



Annexe XIX : Réseau sociocognitif dans S3



Annexe XX : Distribution de trois scores de centralité dans D1

Classes	Score de consensus social	Score de connexité	Score d'objectivation
D1-01	0,12658228	0,2997091	0,637019231
D1-02	0,05063291	0,28916246	0,423167849
D1-03	0,05696203	0,28330879	0,423173804
D1-04	0,06962025	0,31952547	0,585872576
D1-05	0,13291139	0,28484797	0,627755102
D1-06	0,01265823	0,16407754	0,177419355
D1-07	0,11392405	0,29578901	0,684536082
D1-08	0,01898734	0,14109237	0,204301075
D1-09	0,03797468	0,2808094	0,551526718
D1-10	0,06329114	0,24466472	0,281746032
D1-11	0,02531646	0,22178639	0,275362319
D1-12	0,09493671	0,29194685	0,687842278
D1-13	0,06962025	0,26672452	0,535893155
D1-14	0,01265823	0,13443845	0,163636364
D1-15	0,06329114	0,31401679	0,618160652
D1-16	0,05696203	0,28808218	0,687068115
D1-17	0,10126582	0,28756609	0,529113924
D1-18	0,03797468	0,18586603	0,28358209
D1-19	0,01898734	0,1640772	0,215053763
D1-20	0,02531646	0,21761181	0,387978142
D1-21	0,05063291	0,19567077	0,303797468
D1-22	0,01898734	0,16637319	0,381188119
D1-23	0,03164557	0,28948138	0,497206704
D1-24	0,06962025	0,30364522	0,386363636
D1-25	0,06329114	0,29411168	0,544839255
D1-26	0,17088608	0,30300545	0,677245105
D1-27	0,03797468	0,189782	0,253521127
D1-28	0,03164557	0,22494995	0,467889908
D1-29	0,03797468	0,3018945	0,520391517
D1-30	0,07594937	0,31507913	0,467336683
D1-31	0,10126582	0,31072625	0,749716232
D1-32	0,0443038	0,27463438	0,484756098
D1-33	0,08860759	0,25709133	0,457399103
D1-34	0,0443038	0,23746034	0,328042328
D1-35	0,01898734	0,26214097	0,32996633
D1-36	0,01265823	0,17269687	0,34
D1-37	0,05063291	0,31062362	0,380597015
D1-38	0,03797468	0,26700606	0,406649616
D1-39	0,01898734	0,18512063	0,204819277
D1-40	0,07594937	0,26893111	0,5

Annexe XXI : Distribution de trois scores de centralité dans D2

Classes	Score de consensus social	Score de connexité	Score d'objectivation
D2-01	0,186440678	0,314922272	0,635338346
D2-02	0,305084746	0,300122936	0,737272727
D2-03	0,06779661	0,205082677	0,307692308
D2-04	0,06779661	0,127379874	0,2
D2-05	0,016949153	0,096586859	0,269230769
D2-06	0,101694915	0,225216097	0,534709193
D2-07	0,169491525	0,296512197	0,652133581
D2-08	0,06779661	0,211871185	0,362204724
D2-09	0,016949153	0,193165523	0,451977401
D2-10	0,13559322	0,268496156	0,489837398
D2-11	0,084745763	0,261893295	0,520697168
D2-12	0,016949153	0,102087113	0,482517483
D2-13	0,101694915	0,218875364	0,413793103
D2-14	0,016949153	0,0927558	0,450704225
D2-15	0,152542373	0,235541015	0,697495183
D2-16	0,101694915	0,187356667	0,399176955
D2-17	0,050847458	0,134187718	0,182926829
D2-18	0,118644068	0,267815977	0,60315534
D2-19	0,13559322	0,296766487	0,735764555
D2-20	0,169491525	0,245216379	0,602836879
D2-21	0,186440678	0,233394831	0,42
D2-22	0,237288136	0,277746846	0,481054366
D2-23	0,186440678	0,304383195	0,7721843
D2-24	0,016949153	0,088440846	0,173076923
D2-25	0,169491525	0,210688395	0,450805009
D2-26	0,06779661	0,249931728	0,451530612
D2-27	0,06779661	0,111582859	0,2
D2-28	0,050847458	0,167448321	0,379032258
D2-29	0,084745763	0,177923379	0,28
D2-30	0,118644068	0,272037359	0,578343949
D2-31	0,169491525	0,2571775	0,499254844
D2-32	0,06779661	0,239616126	0,62962963
D2-33	0,033898305	0,215274803	0,495548961
D2-34	0,050847458	0,164584169	0,409482759
D2-35	0,203389831	0,244390546	0,758017493
D2-36	0,016949153	0,115252441	0,602739726
D2-37	0,084745763	0,254908954	0,642140468
D2-38	0,06779661	0,271921359	0,668079096
D2-39	0,016949153	0,117440736	0,398058252
D2-40	0,050847458	0,183872138	0,522123894

Annexe XXII : Distribution de trois scores de centralité dans D3

Classes	Score de consensus social	Score de connexité	Score d'objectivation
D3-01	0,28813559	0,30591486	0,83940397
D3-02	0,08474576	0,21852108	0,60332542
D3-03	0,01694915	0,13888135	0,38732394
D3-04	0,11864407	0,28151108	0,62512077
D3-05	0,10169492	0,2653027	0,60032103
D3-06	0,22033898	0,30615459	0,68400771
D3-07	0,01694915	0,18245838	0,67706013
D3-08	0,06779661	0,18163405	0,3740458
D3-09	0,10169492	0,26765189	0,62297297
D3-10	0,08474576	0,18024946	0,67323944
D3-11	0,06779661	0,21126405	0,34070796
D3-12	0,13559322	0,28459054	0,60363248
D3-13	0,06779661	0,19276162	0,38666667
D3-14	0,03389831	0,16048568	0,1509434
D3-15	0,05084746	0,20798378	0,43726236
D3-16	0,08474576	0,24492838	0,40248963
D3-17	0,08474576	0,19336162	0,41090909
D3-18	0,10169492	0,21009486	0,6167979
D3-19	0,11864407	0,27291	0,50825083
D3-20	0,10169492	0,26648162	0,67684478
D3-21	0,03389831	0,19683784	0,51879699
D3-22	0,10169492	0,29685919	0,46168959
D3-23	0,05084746	0,21706189	0,45873786
D3-24	0,13559322	0,2282573	0,41504178
D3-25	0,18644068	0,25607135	0,45859873
D3-26	0,01694915	0,05122162	0,28888889
D3-27	0,11864407	0,22706919	0,38679245
D3-28	0,11864407	0,27268703	0,36177474
D3-29	0,10169492	0,20747892	0,49079755
D3-30	0,20338983	0,32292378	0,68475275
D3-31	0,06779661	0,23099432	0,50842697
D3-32	0,15254237	0,27657324	0,41477273
D3-33	0,10169492	0,2595773	0,48630137
D3-34	0,15254237	0,30119405	0,58571429
D3-35	0,11864407	0,30068757	0,60382166
D3-36	0,03389831	0,12599676	0,33673469
D3-37	0,3220339	0,31661811	0,6719057
D3-38	0,08474576	0,23454216	0,44491525

Annexe XXIII : Distribution de trois scores de centralité dans P1

Classes	Score de consensus social	Score de connexité	Score d'objectivation
P1-01	0,10309278	0,302714103	0,721849866
P1-02	0,10824742	0,270309487	0,627757353
P1-03	0,10824742	0,332208462	0,51810585
P1-04	0,05154639	0,227615641	0,501285347
P1-05	0,03092784	0,187099231	0,444126074
P1-06	0,03608247	0,187201026	0,322916667
P1-07	0,04123711	0,182181282	0,352739726
P1-08	0,07216495	0,242860769	0,486486486
P1-09	0,07731959	0,285507949	0,555555556
P1-10	0,02061856	0,13040641	0,180555556
P1-11	0,13917526	0,266476667	0,679338843
P1-12	0,12371134	0,291197179	0,718309859
P1-13	0,09278351	0,218423077	0,567489115
P1-14	0,04639175	0,213871026	0,34741784
P1-15	0,03092784	0,19500359	0,483471074
P1-16	0,08247423	0,260047436	0,662251656
P1-17	0,05670103	0,202024615	0,28708134
P1-18	0,05670103	0,174627179	0,750679964
P1-19	0,04639175	0,229432821	0,348754448
P1-20	0,01546392	0,204779744	0,311881188
P1-21	0,03092784	0,185647436	0,336787565
P1-22	0,04639175	0,191914872	0,362385321
P1-23	0,06185567	0,286073846	0,476808905
P1-24	0,05154639	0,264978205	0,394299287
P1-25	0,0257732	0,224610769	0,27480916
P1-26	0,01030928	0,093801795	0,097560976
P1-27	0,05154639	0,228393077	0,450746269
P1-28	0,01030928	0,198758974	0,429824561
P1-29	0,04639175	0,249456667	0,446650124
P1-30	0,10309278	0,28908359	0,583247156
P1-31	0,07216495	0,257532051	0,37995338
P1-32	0,02061856	0,185514615	0,529824561
P1-33	0,08247423	0,278575128	0,531141869
P1-34	0,03092784	0,195936667	0,428571429
P1-35	0,05154639	0,156219231	0,267716535
P1-36	0,0257732	0,221762051	0,361502347
P1-37	0,0257732	0,193643077	0,270440252
P1-38	0,03092784	0,205295128	0,418410042
P1-39	0,08247423	0,290336923	0,670886076
P1-40	0,04123711	0,280252564	0,432515337

Annexe XXIV : Distribution de trois scores de centralité dans P2

Classes	Score de consensus social	Score de connexité	Score d'objectivation
P2-01	0,211382114	0,321466923	0,6607445
P2-02	0,211382114	0,274062821	0,76912378
P2-03	0,130081301	0,206545897	0,43175487
P2-04	0,130081301	0,251243333	0,5916442
P2-05	0,06504065	0,213152308	0,30672269
P2-06	0,032520325	0,187346667	0,56542056
P2-07	0,06504065	0,211364359	0,4245283
P2-08	0,06504065	0,140377436	0,29896907
P2-09	0,219512195	0,303190256	0,59729448
P2-10	0,06504065	0,218591026	0,40540541
P2-11	0,024390244	0,104345128	0,38297872
P2-12	0,146341463	0,255168974	0,40703518
P2-13	0,146341463	0,220362308	0,48888889
P2-14	0,081300813	0,237258462	0,34628975
P2-15	0,032520325	0,161937179	0,09677419
P2-16	0,040650407	0,093677692	0,27536232
P2-17	0,130081301	0,219942564	0,6313416
P2-18	0,130081301	0,286413333	0,72432432
P2-19	0,056910569	0,238675897	0,54589372
P2-20	0,048780488	0,186036923	0,31372549
P2-21	0,154471545	0,210284615	0,43258427
P2-22	0,048780488	0,214747692	0,32589286
P2-23	0,105691057	0,294696923	0,69361277
P2-24	0,056910569	0,222357436	0,38720539
P2-25	0,138211382	0,271713846	0,51145038
P2-26	0,081300813	0,217520256	0,38697318
P2-27	0,170731707	0,234640769	0,62916667
P2-28	0,06504065	0,240395641	0,65043695
P2-29	0,016260163	0,214918205	0,66180758
P2-30	0,040650407	0,229918462	0,35609756
P2-31	0,18699187	0,313582821	0,68317503
P2-32	0,138211382	0,263892564	0,55023184
P2-33	0,105691057	0,203610256	0,41328413
P2-34	0,105691057	0,272287436	0,50478469
P2-35	0,081300813	0,245608462	0,64184852
P2-36	0,12195122	0,225934359	0,5
P2-37	0,06504065	0,226443846	0,51869159
P2-38	0,25203252	0,302038462	0,57074341
P2-39	0,073170732	0,213994872	0,38888889
P2-40	0,048780488	0,14372641	0,66878981

Annexe XXV : Distribution de trois scores de centralité dans P3

Classes	Score de consensus social	Score de connexité	Score d'objectivation
P3-01	0,37190083	0,339460769	0,77673546
P3-02	0,04132231	0,221573846	0,34375
P3-03	0,09090909	0,1767	0,34375
P3-04	0,09917355	0,269375641	0,505617978
P3-05	0,2231405	0,314724103	0,707411504
P3-06	0,08264463	0,199584615	0,374407583
P3-07	0,1322314	0,280797179	0,613885505
P3-08	0,10743802	0,284858462	0,542857143
P3-09	0,04958678	0,199419744	0,431192661
P3-10	0,04958678	0,237347949	0,458937198
P3-11	0,04132231	0,257897179	0,352272727
P3-12	0,05785124	0,225742821	0,391941392
P3-13	0,09090909	0,296334103	0,654193548
P3-14	0,05785124	0,244993846	0,418079096
P3-15	0,04958678	0,198676154	0,426666667
P3-16	0,10743802	0,270934872	0,428940568
P3-17	0,09917355	0,266137949	0,485961123
P3-18	0,05785124	0,17570359	0,368421053
P3-19	0,09917355	0,255762564	0,617269545
P3-20	0,23966942	0,304355641	0,548825711
P3-21	0,21487603	0,312201795	0,631111111
P3-22	0,1322314	0,298298462	0,613948919
P3-23	0,10743802	0,204878974	0,505862647
P3-24	0,09917355	0,284156154	0,458579882
P3-25	0,09917355	0,273317436	0,529976019
P3-26	0,12396694	0,294707436	0,525862069
P3-27	0,02479339	0,198606154	0,444444444
P3-28	0,04958678	0,28941641	0,571691176
P3-29	0,04132231	0,227668205	0,442028986
P3-30	0,03305785	0,250816923	0,315151515
P3-31	0,1322314	0,271413077	0,585271318
P3-32	0,12396694	0,242561795	0,503546099
P3-33	0,10743802	0,296243077	0,649141631
P3-34	0,05785124	0,303790256	0,623513871
P3-35	0,14876033	0,253106154	0,584018801
P3-36	0,0661157	0,258778974	0,352777778
P3-37	0,08264463	0,280124872	0,464379947
P3-38	0,14876033	0,346968718	0,590769231
P3-39	0,12396694	0,263387692	0,793824701
P3-40	0,10743802	0,288579231	0,406779661

Annexe XXVI : Distribution de trois scores de centralité dans S1

Classes	Score de consensus social	Score de connexité	Score d'objectivation
S1-01	0,051020408	0,21171114	0,50617284
S1-02	0,030612245	0,08916396	0,768656716
S1-03	0,010204082	0,1100827	0,348214286
S1-04	0,030612245	0,1517956	0,408695652
S1-05	0,020408163	0,17655231	0,368181818
S1-06	0,040816327	0,13316251	0,405882353
S1-07	0,112244898	0,20136064	0,373219373
S1-08	0,020408163	0,08852933	0,553846154
S1-09	0,020408163	0,06438556	0,16
S1-10	0,071428571	0,15911062	0,298245614
S1-11	0,030612245	0,13158842	0,202380952
S1-12	0,081632653	0,19311736	0,572972973
S1-13	0,051020408	0,17101853	0,293478261
S1-14	0,010204082	0,17195027	0,292553191
S1-15	0,051020408	0,18305368	0,302197802
S1-16	0,030612245	0,13905504	0,313559322
S1-17	0,020408163	0,13208824	0,285714286
S1-18	0,071428571	0,23310764	0,484313725
S1-19	0,030612245	0,12341853	0,221153846
S1-20	0,06122449	0,22540907	0,366548043
S1-21	0,102040816	0,16362506	0,338709677
S1-22	0,051020408	0,17952381	0,325123153
S1-23	0,06122449	0,15547816	0,340101523
S1-24	0,010204082	0,05393108	0,05
S1-25	0,010204082	0,0891655	0,1875
S1-26	0,071428571	0,17015321	0,338164251
S1-27	0,051020408	0,17117805	0,49
S1-28	0,06122449	0,16682536	0,317948718
S1-29	0,06122449	0,19211921	0,30994152
S1-30	0,010204082	0,08167205	0
S1-31	0,010204082	0,07078487	0,139534884
S1-32	0,020408163	0,19635008	0,444444444
S1-33	0,173469388	0,24199099	0,552631579
S1-34	0,051020408	0,17113624	0,433691756
S1-35	0,030612245	0,16578212	0,590604027
S1-36	0,020408163	0,11183798	0,093023256
S1-37	0,020408163	0,17584943	0,184466019
S1-38	0,030612245	0,11620377	0,215384615
S1-39	0,010204082	0,05045027	0
S1-40	0,051020408	0,19076115	0,353413655

Annexe XXVII : Distribution de trois scores de centralité dans S2

Classes	Score de consensus social	Score de connexité	Score d'objectivation
S2-01	0,11111111	0,200119788	0,45756458
S2-02	0,20634921	0,230738157	0,46764092
S2-03	0,11111111	0,239760023	0,65520206
S2-04	0,04761905	0,181967363	0,39007092
S2-05	0,11111111	0,19536524	0,33823529
S2-06	0,11111111	0,234476204	0,62407862
S2-07	0,04761905	0,169132428	0,47674419
S2-08	0,11111111	0,244394254	0,46536797
S2-09	0,07936508	0,161240478	0,45126354
S2-10	0,0952381	0,161037942	0,30373832
S2-11	0,12698413	0,232492773	0,32763533
S2-12	0,01587302	0,098499773	0,5862069
S2-13	0,12698413	0,306218834	0,56125356
S2-14	0,11111111	0,190764148	0,55450237
S2-15	0,04761905	0,188428457	0,31707317
S2-16	0,01587302	0,088392486	0,18181818
S2-17	0,07936508	0,143459564	0,23478261
S2-18	0,04761905	0,179855547	0,5954023
S2-19	0,14285714	0,227278647	0,37391304
S2-20	0,03174603	0,169962335	0,3
S2-21	0,01587302	0,036162242	0,22222222
S2-22	0,14285714	0,238570824	0,39453125
S2-23	0,06349206	0,274084092	0,60787671
S2-24	0,01587302	0,224717542	0,33333333
S2-25	0,04761905	0,163048718	0,51245552
S2-26	0,14285714	0,257325805	0,53318078
S2-27	0,07936508	0,16590511	0,23770492
S2-28	0,03174603	0,129324261	0,66666667
S2-29	0,06349206	0,159584837	0,22413793
S2-30	0,04761905	0,213253647	0,44444444
S2-31	0,12698413	0,236391988	0,62833914
S2-32	0,03174603	0,165879913	0,36936937
S2-33	0,11111111	0,139304775	0,3483871
S2-34	0,06349206	0,146992426	0,27966102
S2-35	0,11111111	0,198097576	0,43396226
S2-36	0,04761905	0,179364552	0,32768362
S2-37	0,06349206	0,263768503	0,5701107
S2-38	0,0952381	0,197994915	0,42599278
S2-39	0,03174603	0,123383563	0,21428571
S2-40	0,07936508	0,204105688	0,56161137

Annexe XXVIII : Distribution de trois scores de centralité dans S3

Classes	Score de consensus social	Score de connexité	Score d'objectivation
S3-01	0,28571429	0,32175172	0,65682138
S3-02	0,07142857	0,21433998	0,45454545
S3-03	0,08928571	0,23752231	0,53159041
S3-04	0,23214286	0,28297261	0,62390351
S3-05	0,10714286	0,22112208	0,4743083
S3-06	0,03571429	0,12234102	0,21875
S3-07	0,07142857	0,1982515	0,44705882
S3-08	0,125	0,23831621	0,50482315
S3-09	0,10714286	0,18151732	0,40248963
S3-10	0,07142857	0,11481123	0,28712871
S3-11	0,07142857	0,17001778	0,39325843
S3-12	0,01785714	0,08831849	0,68224299
S3-13	0,125	0,17685963	0,30487805
S3-14	0,07142857	0,20107246	0,15384615
S3-15	0,08928571	0,229038	0,56557377
S3-16	0,01785714	0,161082	0,38392857
S3-17	0,05357143	0,19906282	0,16883117
S3-18	0,03571429	0,13852715	0,14583333
S3-19	0,14285714	0,20756896	0,62707838
S3-20	0,05357143	0,2491211	0,6185567
S3-21	0,125	0,26291534	0,38311688
S3-22	0,08928571	0,1973087	0,34673367
S3-23	0,125	0,26688215	0,44471744
S3-24	0,07142857	0,24252343	0,48606811
S3-25	0,08928571	0,21371943	0,475
S3-26	0,10714286	0,22531979	0,34439834
S3-27	0,03571429	0,10692953	0,44578313
S3-28	0,08928571	0,21602291	0,26490066
S3-29	0,10714286	0,16920821	0,39247312
S3-30	0,07142857	0,2373614	0,54812834
S3-31	0,03571429	0,13787902	0,18309859
S3-32	0,10714286	0,18882463	0,3028169
S3-33	0,08928571	0,24991724	0,46822742
S3-34	0,01785714	0,14429854	0,6557971
S3-35	0,05357143	0,25918278	0,4729064
S3-36	0,125	0,22609107	0,51836735
S3-37	0,08928571	0,26091107	0,56405694
S3-38	0,08928571	0,2485357	0,44495413
S3-39	0,33928571	0,2901798	0,67866324
S3-40	0,05357143	0,2309639	0,58426966

BIBLIOGRAPHIE

Abric, J.C. (1994a). *L'organisation interne des représentations sociales: système central et système périphérique*. Dans Guimelli, C. (dir) (1994). *Structures et transformations des représentations sociales*. Lausanne, Delachaux et Nierstlé, p. 73-84.

Abric, J.C. (1994b). *Méthodologie de recueil des représentations sociales.* Dans Abric, J.C. (dir). *Pratiques sociales et représentations*. Paris, PUF, p. 59-82

Abric, J.C. (1994b). *Pratiques sociales et représentations.* Paris, PUF

Abric, J.C. (1994c). *Méthodologie de recueil des représentations sociales.* Dans Abric, J.C. (dir.) (1994b). *Pratiques sociales et représentations.* Paris, PUF, p. 59-82

Abric, J.C. (2003b). *La recherche de la zone muette des représentations sociales*. Dans Abric, J.C. (dir) (2003a). *Méthodes d'étude des représentations sociales*. Ramonville Saint-Agne, Eres, p. 59-80.

Abric, J.C. (2003c). *L'analyse structurale des représentations sociales*. Dans Moscovici, S. et F. Buschini (éd). *Les méthodes des sciences humaines*. Paris, PUF, p. 375-392.

Abric, J.C. (dir). (2003a). *Méthodes d'étude des représentations sociales*. Ramonville Saint-Agne, Eres.

Aïssani, Y., Bonardi, C. et B. Guelfucci (1990). « Représentation sociale et noyau central: problèmes de méthode », *Revue Internationale de Psychologie Sociale*, 3, 3. p. 335-356.

Andler, D. (dir) (2004). *Introduction aux sciences cognitives*. Paris, Gallimard.

Apostolidis, T. (2003). *Représentations sociales et triangulation : enjeux théorico-méthodologiques*. Dans Abric, J.C. (dir). *Méthodes d'étude des représentations sociales*. Ramonville Saint-Agne, Eres, p. 13-35.

Bardin, L. (1977). *L'analyse de contenu*, Paris, PUF.

Bataille, M. (2002). *Le noyau peut-il ne pas être central?*. Dans Garnier, C. et W. Doise (dir.). *Les représentations sociales. Balisage du domaine d'études.* Montréal, Éditions nouvelles, p. 25-34.

Baugnet, L. et A. Fouquet (2005). « L'Europe dans les médias : effets de contexte », *Connexions*, 84, 2, p. 87-109.

Benzécri Jean-Paul et coll. (1981). *Pratique de l'analyse des données, Linguistique et lexicologie*. Paris, Dunod.

Berger P., et L. Luckmann (2003). *La construction sociale de la réalité*. Paris, Armand Collin.

Berthelot, J.- M. (1990). *L'intelligence du social*. Paris, PUF.

Blackmore, S. J. (2000). *The Meme Machine*. Oxford, Oxford University Press.

Bless, H., K. Fiedler et F. Strack (2004). *Social Cognition: How Individuals Construct Social Reality*. London, Social Psychology Press.

Bloor, D. (1976). *Knowledge and Social Imagery*. Chicago, The Universiy of Chicago Press.

Borgatti, S.P., Everett, M.G. et L.C. Freeman (2002). *Ucinet for Windows: Software for Social Network Analysis. Harvard*. MA, Analytic Technologies.

Bouchard, G. et C. Taylor (2008). *Fonder l'avenir. Le temps de la conciliation. Rapport de la commission de consultation sur les pratiques d'accommodement reliées aux différences culturelle*. Québec, Bibliothèque et Archives nationales du Québec. [En ligne] http://www.accommodements.qc.ca/documentation/rapports/rapport-final-integral-fr.pdf.

Bourdieu, P. (1979). *La distinction. Critique sociale du jugement*. Paris, Minuit.

Bouriche, B. (2003). *L'analyse de similitude*. Dans Abric, J.C. (dir) (2003a). *Méthodes d'étude des représentations sociales*. Ramonville Saint-Agne, Eres.

Bradley, P. S. et U. M. Fayyad (1998). *Refining Initial Points for K-Means Clustering*. Dans Shavlik, J. (éd). *Proceedings of 15th Internationnal Conference on Machine Learning (ICML98)*. San Francisco, Morgan Kaufmann, p. 91-99.

Bronner, G. (2003). *L'empire des croyances*. Paris, PUF.

Brunet, E. (1986). *Méthodes quantitatives et informatiques dans l'étude des textes*. Paris, Champion.

Buschini, F. et N. Kalampalikis (2002). *La synonymie, l'analogie et la taxinomie*. Dans Garnier, C. et W. Doise (dir). *Les représentations sociales. Balisage du domaine d'études*. Montréal, Éditions nouvelles, p. 187-206.

Carley, K. M. (1986). « An approach for relating social structure to cognitive structure ». *Journal of Mathematical Sociology*. 12, 2, p. 137-189.

Casson, R. W. (1983). « Schemata in Cognitive Anthropology ». *Annual Review of Anthropology*. 12, p. 429-462.

Cavnar, W. B. et J. M. Trenkle (1994). « N-Gram-Based Text Categorization ». *Paper Presented at the Symposium on Document Analysis and Information Retrieval* (Las Vegas).

Cerulo, K.A. (éd) (2002). *Culture in Mind. Toward a Sociology of Culture and Cognition*. New York, Routledge.

Chartier, J.-F., Meunier, J.-G., Jendoubi, M. et J. Danis (2008). *Le travail conceptuel collectif : une analyse assistée par ordinateur de la distribution du concept d'ACCOMMODEMENT RAISONNABLE dans les journaux québécois*. Dans Heiden, S. et B. Pincemin (dir). *Actes des 9e Journées internationales d'Analyse statistique des Données Textuelles,* Lyon, Presses Universitaires de Lyon, p. 297-307.

Choueka, Y. (1988). « Looking for Needles in a Haystack », *International conference on User-Oriented Context Based Text and Image Handling*. Cambridge, p. 609-623.

Cicourel, A. V. (1979). *La sociologie cognitive*. Paris, PUF.

Collins, R. (2004). *Interaction Ritual Chains*. Oxford, Princeton University Press.

Conein, B. (2005). *Les sens sociaux : trois essais de sociologie cognitive*. Paris, Economica.

D'Andrade, R. (1995). *The Development of Cognitive Anthropology*. New York, Cambridge University Press.

Damashek. M. (1989). « Gauging Similarity with N-grams: Language Independent Categorization of Text ». *Science*, 267, p. 843-848.

De Alba, M. (2004). « El méthodo ALCESTE y su applicacion al estudio de las representaciones sociales del espacio urbano : El Caso de la ciudad de México », *Papers on Social Rrepresentations / Textes sur les représentations sociales*, 13, p. 1-20.

De Rosa, A. S. (2003). *Le « réseau d'association »*. Dans Abric, J.C. (dir) (2003a). *Méthodes d'étude des représentations sociales*. Ramonville Saint-Agne, Eres, p. 81-118.

Demazière, D., Brossaud, C., Trabal, P. et Van Meter, K. (dir) (2006). *Analyses textuelles en sociologie - Logiciels, méthodes, usages*, Renne, PUR.

Desjardins, G. (2006). *Modélisation connexionniste du repérage de l'information*, Montréal, Thèse de Doctorat, Université du Québec à Montréal.

Diesner, J., et Carley, K.M. (2005). *Revealing Social Structure from Texts: Meta-Matrix Text Analysis as a novel method for Network Text Analysis*. Dans V.K. Narayanan et D.J. Armstrong (éd). *Causal Mapping for Information Systems and Technology Research: Approaches, Advances, and Illustrations*. Harrisburg, Idea Group Publishing, p. 81-108.

Doise, W. (1993). *Logiques sociales dans le raisonnement*. Lausanne, Delachaux et Nierstlé.

Doise, W. et A. Palmonari (1986). *L'étude des représentations sociales*. Paris, Delachaux et Niestlé.

Doise, W., Clémence, A. et F. Lorenzi-Cioldi (1992). *Représentations sociales et analyses de données*. Grenoble. Paris, PUG.

Douglas, M. (2004). *Comment pensent les institutions*. Paris, La Découverte.

Durkheim, É. (1893). *De la division du travail social*. Édition électronique Classiques des sciences sociales, [En ligne] http://dx.doi.org/doi:10.1522/cla.due.del1.

Durkheim, É. (1898). « Les représentations individuelles et les représentations collectives ». Édition électronique Classiques des sciences sociales [en ligne] http://dx.doi.org/doi:10.1522/cla.due.rep1

Durkheim, É. (1912). *Les formes élémentaires de la vie religieuse*. Édition électronique Classiques des sciences sociales, Édition électronique Classiques des sciences sociales [En ligne] http://dx.doi.org/doi:10.1522/cla.due.for2.

Durkheim, É. (1914). « Le dualisme de la nature humaine et ses conditions sociales ». Édition électronique Classiques des sciences sociales [En ligne] http://dx.doi.org/doi:10.1522/cla.due.dua

Durkheim, É. (1955). *Pragmatisme et sociologie : Cours inédit prononcé à La Sorbonne en 1913-1914 et restitué par Armand Cuvillier d'après des notes d'étudiants*. Édition électronique Classiques des sciences sociales [En ligne] http://dx.doi.org/doi:10.1522/cla.due.pra

Durkheim, É. et M. Mauss (1903). « De quelques formes de classification : Contribution à l'étude des représentations collectives ». Édition électronique Classiques des sciences sociales [En ligne] http://dx.doi.org/doi:10.1522/cla.due.deq

Festinger, L. (1957). *A Theory of Cognitive Dissonance*. Evanston, Illinois.

Flament, C. (1981). « L'analyse de similitude: une technique pour les recherches sur les représentations sociales ». *Cahier de psychologie cognitive*, 4, p. 357-396.

Flament, C. (1986). *L'analyse de similitude: une technique pour les recherches sur les R.S.* Dans W. Doise et A. Palmonari (dir). *L'étude des représentations sociales*. Paris, Delachaux et Nestlé.

Flament, C. (1994a). *Aspects périphériques des représentations sociales*. Dans Guimelli, C. *Structures et transformations des représentations sociales*. Lauzanne, Delachaux et Niestlé, p. 85-118.

Flament, C. (1994b). *Structure, dynamique et transformation des représentations sociale*. Dans Abric, J.C. (1994 b). *Pratiques sociales et représentations*. Paris, PUF, p. 37-58.

Flament, C. (1996a). « Statistique classique et/ou logique de Boole dans l'analyse d'un questionnaire de représentation sociale : l'exemple du sport ». *Revue internationale de psychologie sociale*. 2, p. 109-121.

Flament, C. (1996b). « Quand les éléments centraux d'une représentation sont excentriques ». *Papers on social representations*. 5, 2, p. 145-150.

Flament, C. (1999). « Liberté d'opinion et limite normative dans une représentation sociale : le développement de l'intelligence ». *Swiss journal of psychology*. 58, 3, p. 201-206.

Flament, C. et M. L. Rouquette (2003). *L'anatomie des idées ordinaires: comment étudier les représentations sociales*. Paris, Armand Collins.

Fleck, L. (2005). *Genèse et développement d'un fait scientifique*. Paris, Belles lettres.

Forest D. (2006). *Application de techniques de forage de textes de nature prédictive et exploratoire à des fins de gestion et d'analyse thématique de documents textuels non structurés*. Montréal, Thèse de Doctorat, Université du Québec à Montréal.

Forest, D. et J.-G. Meunier (2000). *La classification mathématique des textes : un outil d'assistance à la lecture et à l'analyse de textes philosophiques*. Dans Rajman, M. et J.-C Chappelier (éd). *Actes des 5es Journées internationales d'Analyse statistique des Données Textuelles*, 9-11 mars 2000, EPFL, Lausanne, Suisse. Volume 1, p. 325-329.

Försterling, F. (2001). *Attribution : an Introduction to Theories, Researsh and Applications*. Philadelphia, Psychology Press.

Fortin, C., et R. Rousseau (2005). *Psychologie cognitive. Une approche de traitement de l'information. 2e édition*. Québec, Presse de l'université du Québec.

Freeman, L.C. (1979). « Centrality in Social Network : Conceptual Clarification », *Social Networks*, 1, p. 215-239.

Gagnon, D. (2004). « NUMEXO et l'analyse par attracteurs et par classes des entrées de l'ECHO (Encyclopédie Culturelle hypermédia de l'Océanie) », *Lexicometrica, L'analyse de données textuelles : De l'enquête aux corpus littéraires*. Numéro spécial.

Garnier, C. et W. Doise (dir) (2002). *Les représentations sociales. Balisage du domaine d'études*. Montréal, Éditions nouvelles.

Gaymard, S. (2002). « Représentation sociale et algèbre de Boole: une étude des modèles normatifs dans une situation de biculturalisme ». *Revue internationale de psychologie sociale*. 15, 1, p. 163-184.

Goffman, E. (1974). *Les rites d'interactions*. Paris, Minuit.

Goffman, E. (1991). *Les cadres de l'expérience*. Paris, Les éditions de minuit.

Goldman, A. (1999). *Knowledge in a Social World*. Oxford, Clarendon Press.

Goody, J. (1979). *La raison graphique*. Paris, Minuit.

Guimelli, C. (1999). *La pensée sociale. Paris*, PUF.

Guimelli, C. (2003). *Le modèle des schemes cognitifs de base (SCB): méthodes et applications*. Dans Abric, J.C. (dir). (2003a). *Méthodes d'étude des représentations sociales*. Ramonville Saint-Agne, Eres, p. 119-143.

Guimelli, C. (dir). (1994). *Structures et transformations des représentations sociales*. Lausanne, Delachaux et Nierstlé.

Gurvitch, G. (1966). *Les cadres sociaux de la connaissance*. Paris, Tops/ H. Trinquier.

Halbwachs, M. (1994). *Les cadres sociaux de la mémoire*. Paris, Albin Michel.

Harman, D. (2005). *The History of IDF and its Influences on IR and Other Fields*. Dans Tait, J.I. (éd). *Charting a New Course : Natural Language Processing and Information Retrieval. Essays in Honour of Karen Spärck Jones*. Netherlands, Springer, p. 69-79.

Harris Z. S. (1968). *Mathematical Structures of Language*. New York, Wiley.

Harris Z. S. (1991). *A Theory of Language and Information: A Mathematical Approach*. Oxford, Clarendon Press.

Heider, F. (1958). *The Psychology of Interpersonal Relations*. New York, Wiley.

Hockey, S. (2001). *Electronic Texts in the Humanities: Principles and Practice*. Oxford, Oxford University Press.

Hotho, A., A. Nürnberger et G. Paaß (2005). « A Brief Survey of Text Mining ». *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology*, 201, 1, p. 19-62. [En ligne] http://www.kde.cs.uni-kassel.de/hotho/pub/2005/hotho05TextMining.pdf.

Hotho, A., Staab, S. et G. Stumme (2003). *Wordnet Improves Text Document Clustering*. Dans *Semantic Web Workshop of the 26th Annual International ACM SIGIR Conference*. Toronto, Canada.

Hruschka, E. R. et T. F. Covoes (2005). *Feature Selection for Cluster Analysis: an Approach Based on the Simplified Silhouette Criterion*. Dans *Proceeding of the 2005 Internationnal Conference on Computationnal Intelligence for Modelling, Control and Automation, and Internationnal Conference on Intelligent Agent, Web Technologies and Internet Commerce (CIMCA-IAWTIC'05)*. Computer society, p. 32-38.

Hutchins, E. (1995). *Cognition in the Wild*. Cambridge, MIT Press.

Jain, A. K., Murty, M. N. et P.J. Flynn (1999). « Data Clustering: a Review ». *ACM Computing Surveys*. 31, 3, p. 264-323.

Jodelet, D. (1989). *Représentations sociales : un domaine en expansion*. Dans Jodelet, D. (dir). *Les représentations sociales*. Paris, PUF, p. 47-78.

Jodelet, D. (dir). (1989). *Les représentations sociales*. Paris, PUF.

Kalampalikis, N. (2003). *L'apport de la méthode Alceste dans l'analyse des représentations sociales*. Dans Abric, J.C. (dir) (2003a). *Méthodes d'étude des représentations sociales*. Ramonville Saint-Agne, Eres, p. 147-163

Kalampalikis, N. et S. Moscovici (2005). « Une approche pragmatique de l'analyse Alceste ». *Les Cahiers Internationaux de Psychologie Sociale*. 66, p. 15-24.

Kaufman, L et P.J. Rousseeuw (1990). *Finding Groups in Data : an Introduction to Cluster Analysis*. New York, Wiley.

Kohonen T. (2001). *Self-Organizing Maps*. Springer.

Kuhn, T. (1970). *La structure des révolutions scientifiques*. Paris, Flammarion.

Kunda, Z. (1999). *Social Cognition: Making Sense of People*. Cambridge, MIT Press.

Lahlou, L. (1996). « A Method to Extract Social Representations from Linguistic Corpora ». *Japanese Journal of Experimental Social Psychology*. 36, p. 278–291.

Lahlou, S. (1998). *Penser manger*. Paris, PUF.

Lahlou, S. (2003). *L'exploration des représentations sociales à partir des dictionnaires*. Dans Abric, J.C. (dir). *Méthodes d'étude des représentations sociales*. Ramonville Saint-Agne, Eres, p. 37-58.

Lalhou, S. (1995a). *Vers une théorie de l'interprétation en analyse des données textuelles*. Dans Bolasco, S., Lebart, L. et A. Salem (dir). *JADT 1995. 3rd International Conference on Statistical Analysis of Textual Data*. Roma, CISU, p. 221-228.

Lalhou, S. (1995b). *Penser manger. Les représentations sociales de l'alimentation*. Paris, EHESS, Thèse de doctorat. [En ligne] http://hal.archives-ouvertes.fr/docs/00/16/72/57/PDF/THEDE100a.pdf.

Landauer, T. K., Foltz, P. W. et D. Laham (1998). « Introduction to Latent Semantic Analysis ». *Discourse Processes*. 25, p. 259-284. [En ligne] http://lsa.colorado.edu/.

Lave, J. (1988). *Cognition in Practice. Mind, Mathematics and Culture in Everyday Life*. Cambridge, Cambridge University Press.

Lazega, E. (1998). *Réseaux sociaux et structures relationnelles.* Paris, PUF.

Lebart S. et S. A. Salem (1994). *Statistique textuelle*. Paris, Dunod.

MacQueen, J.B. (1967). *Some Methods for Classification and Snalysis of Multivariate Observations*. Dans *Proceeding Berkeley Symposium on Mathematics, Statistics and Probability*. University of California Press, p. 281-297.

Mannheim, K. (1985). *Ideology and Utopia : an Introduction to the Sociology of Knowledge*. Harvest book.

Manning C. et H. Schutze (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MIT Press.

Masson, E. et S. Moscovici (1997). *Les mutations dans la pratique alimentaire : processus symboliques et représentations sociales*. Paris, Rapport de recherche, Laboratoire de psychologie sociale, EHESS.

Mauss, M. (2002). *Essai sur le don*. Édition électronique Classiques des sciences sociales, [en ligne] http://dx.doi.org/doi:10.1522/cla.mam.ess3.

McCarthy, W. (2004). *Humanities Computing*. Palgrave MacMillan, Blackwell Publishers.

Mead, G.H. (2006). *L'esprit, le soi et la société*. Paris, PUF.

Memmi, D. (2000). « Le modèle vectoriel pour le traitement de documents », *Cahier Leibniz*. Grenoble, 2000, 14, p. 1-30.

Merton, R. K. (1947). *La sociologie de la connaissance*. Dans Gurvitch, G., (dir). *La sociologie au XXe siècle*. Paris, PUF.

Meunier, J.G. (2006). « Le concept: de la singularité à la synthèse ». *Cahiers du LANCI.* 2, octobre, p. 1-35.

Meunier, J.-G. et D. Forest (2009). *L'analyse conceptuelle assistée par ordinateur: premières expériences.* Dans Le Priol, F., Djioua, B. et J.-P. Desclés (dir). *L'annotation.* Paris, Hermes.

Meunier, J.-G., Forest, D. et I. Biskri (2005). *Classification and Categorization in Computer Assisted Reading and Analysis of Texts.* Dans Cohen, H. et C. Lefebvre (dir). *Handbook of Categorization in Cognitive Science.* Amsterdam, Elsevier, p. 955-978.

Meunier, J.-G., Forest, D. et I. Biskri (2006). « A Model for Computer Analysis and Reading of Text (CARAT): The SATIM Approach ». *TEXT Technology*. 1, p. 127-155.

Mitkov, R. (dir) (2003). *The Oxford Handbook of Computational Linguistics*. Oxford, Oxford University Press.

Moliner, P. (1994). *Les méthodes de repérage et d'identification du noyau des représentations sociales*. Dans Guimelli, C. (dir). *Structures et transformations des représentations sociales*. Lausanne, Delachaux et Nierstlé, p. 199-232.

Moliner, P. (dir) (2001). *La dynamique des représentations sociales*. Grenoble, PUG.

Moliner, P. et A. Martos (2005). « Une redéfinition des fonctions du noyau des représentations sociales ». *Journal international sur les représentations sociales*. 2, 1, p. 89-96.

Moliner, P., Rateau, P. et V. Cohen-Scali (2002). *Les représentations sociales: Pratique des études de terrain*. Rennes, PUR.

Monge, P. et N. Contractor (2003). *Theories of Communication Networks*. Oxford University Press.

Morin, E. (1991). *La méthode tome 4: Les idées*. Paris, Seuil.

Moscovici, S. (1976). *La psychanalyse, son image, son public*. Paris, PUF.

Moscovici, S. (1989). *Des représentations collectives aux représentations sociales: éléments pour une histoire*. Dans Jodelet, D. (dir). *Les représentations sociales*. Paris, PUF, p. 78-103.

Negura, L. (2006). « L'analyse de contenu dans l'étude des représentations sociales », *SociologieS*. [En ligne] URL : http://sociologies.revues.org/document993.html.

Nevin, B. (1993). « A Minimalist Program for Linguistics: The Work of Zellig Harris on Meaning and Information ». *Historiographia Linguistica*, XX, 2, p. 355–398.

Pincemin, B., Issac, F., Chanove, M. et M. Mathieu-Colas (2006). *Concordanciers : thème et variations*, Dans J.-M. Viprey (éd). *Actes des 8e Journées internationales d'Analyse statistique des Données Textuelles*. Besançon, Presses universitaires de Franche-Comté, p. 773-784.

Popper, K. (1998). *La connaissance objective*. Paris, Flammarion.

Popping, R. (2000). *Computer-Assisted Text Analysis*. London, Sage Publications.

Porter, M.F. (1980). « An Algorithm For Suffix Stripping ». *Program*. 14, p. 130-137.

Priss, U. (2006). *Formal Concept Analysis In Information Science*. Dans *Proceedings of the American Society for Information Science and Technology*. 40, 1, p. 521-543.

Quéré, L. (1999). « La situation toujours négligée? ». *Réseaux*. 85.

Reinert M. (1993). « Les mondes lexicaux et leur logique ». *Langage et société*, 66, p. 5-39.

Reinert M. (2008). *Mondes lexicaux stabilisés et analyse statistique de discours*. Dans Heiden, S. et B. Pincemin (dir). *Actes des 9e Journées internationales d'Analyse statistique des Données Textuelles*. Lyon, Presses Universitaires de Lyon, p. 579-590.

Reinert, M. (1997). « Les mondes lexicaux et leur logique à travers l'analyse statistique de divers corpus », *Lexicometria*. [En ligne] www.cavi.univ-paris3.fr/lexicometrica/article/numero0/MRMondLex.html.

Rice, E. (1980). «On Cultural Schemata ». *American Ethnologist*. 7, 1, p. 152-171.

Rockwell, G. (2003). « What Is Text Analysis, Really? ». *Literary and Linguistic Computing*, 18, 2, p. 209–219.

Rogers, E. M. (2003). *Diffusion of Innovations*. New York, Free Press.

Roth, C. (2005). *Co-evolution in Epistemic Networks: Reconstructing Social Complex Systems*. Palaiseau, Ecole Polytechnique, Sciences Humaines et Sociales.

Rouquette, M.L. (1994). *Une classe de modèles pour l'analyse des relations entre cognèmes*. Dans Guimelli, C. (dir). (1994). *Structures et transformations des représentations sociales*. Lausanne, Delachaux et Nierstlé, p. 153-170.

Rouquette, M.L. (1998). *La communication sociale*. Paris, Dunod.

Sales-Wuillemin, E. (2005). *Psychologie sociale expérimentale de l'usage du langage. Représentation sociales, catégorisation et attitudes : perspectives nouvelles*. Paris, L'Harmattan.

Salton G. et M. McGill (1983). *Introduction to Modern Information Retrieval*, New-York, McGraw-Hill.

Salton, G. (1989). *Automatic Text Processing*. Addison-Wesley, Reading.

Scheler, M. (1993). *Problème de sociologie de la connaissance*. Paris, PUF.

Sebastiani, F. (2002). « Machine Learning In Automated Text Categorization ». *ACM Computing Surveys*. 34, 1, p. 1-47.

Shore, B. (1996). *Culture in Mind. Cognition, Culture and the Problem of Meaning*. New York, Oxford University Press.

Simmel, G. (1908). *Le croisement des cercles sociaux*. Dans *Sociologie: étude sur les formes de la socialisation*. Paris, PUF.

Sperber, D. (1989). *L'étude anthropologique des représentations: problèmes et perspectives*. Dans Jodelet, D. (dir) (1989). *Les représentations sociales*. Paris, PUF, p. 133-148.

Sperber, D. (1996). *La contagion des idées*, Paris, Odile Jacob.

Sperber, D. (2000). *Outils conceptuels pour une science naturelle de la société et de la culture*. Dans Livet, P. et R. Ogien (dir). *L'Enquête ontologique. Du mode d'existence des objets sociaux*, Paris, EHESS, Raisons pratiques, p. 209-230.

Sperber, D. et D. Wilson (1989). *La pertinence. Communication et cognition*. Paris, Minuit.

Stark, W. (1958). *The Sociology of Knowledge*. London, Routledge and Kegan Paul.

Strauss, C. et N. Quinn (1998). *A Cognitive Theory of Cultural Meaning*, New-York, Cambridge University Press.

Wasserman, S. et Faust, K. (1994). *Social Network Analysis – Methods and Applications*. Cambridge, Cambridge University Press.

Watzlawick, P., Helminck Beavin, J., et Don D. Jackson (1972). *Une logique de la communication*. Paris, Seuil, Essais.

Weiss, S. M., Indurkhya, N., Zhang, T. et Damereau, F. J. (2005). *Text Mining. Predictive Methods for Analyzing Unstructured Information*. New York, Springer-Verlag.

Wille, R. (1992). «Concept Lattices and Conceptual Knowledge Systems». *Computers Mathematics and Applications*. 23, p. 493.

Willett, P. (2006). « The Porter Stemming Algorithme : then and now », *Program : Electronic Library and Information Systems*. 40, 3, p. 219-223.

Wu, X., Kumar, V., Quinlan, JR., Ghosh, J., Yang, Q., Motoda, H., McLachlan, GJ., Ng, A., Liu, B., Yu, PS., Zhou, Z-H., Steinbach, M., Hand, DJ. et D. Steinberg (2008). « Top 10 Algorithms in Data Mining ». *Knowledge and Information Systems*, 14, 1, p. 1–37.

Zerubavel, E. (1997). *Social Mindscapes : An Invitation to Cognitive Sociology*. London, Harvard University Press.

Zhao, Y. et G. Karypis (2002). *Evaluation of Hierarchical Clustering Algorithms for Document Datasets*. Dans *Proceedings of the 11th Conference of Information and Knowledge Management (CIKM)*. ACM Press, p. 515-524.

Zhao, Y. et G. Karypis (2005). « Hierarchical Clustering Algorithms for Document Datasets ». *Data Mining and Knowledge Discovery*. 10, p. 141–168.