Représentations des réseaux de mots associés

Christophe Lejeune

FNRS – Sociologie Générale – Université de Liège
Bd du Rectorat 7 B32 Bte 47 – Belgique
christophe.lejeune@ulg.ac.be

Abstract

The paper presents networks of associated words as it is used in qualitative data analysis software such as Candide, Réseau-Lu T-Lab on one side and Prospéro on the other side. The precise algorithm of both Candide and Prospéro is described.
Strength and limits of such a feature are compared with the gain it provides to social scientists.
The paper concludes on a procedure helping to judge the adequacy and accuracy of categories of analysis constructed by the researcher.

Résumé

Cet exposé présente les réseaux de mots associés utilisés dans les logiciels d'analyse des données qualitatives comme Candide, Réseau-Lu et T-Lab ou comme Prospéro. L'algorithme de Candide et Prospéro est précisément décrit.
Les contributions et les limites de ce type de fonctionnalités sont jaugées en fonction de les apports qu'elles fournissent aux chercheurs en sciences sociales.
L'exposé se termine sur une procédure permettant d'éprouver la pertinence et la validité des catégories d'analyse construites par le chercheur.

Mots-clés : Candide, Prospéro, CAQDAS, cooccurrence, cluster, carte de lien.

1. Délimitation du champ

Mon propos porte sur les outils logiciels qu'utilisent les sciences sociales lorsqu'elles s'adonnent à l'analyse qualitative[1]. Sont exclus de ce champ les outils quantitatifs provenant de l'analyse factorielle ou des statistiques (SAS, Spss, R). Restent donc les outils qualitatifs dont les plus connus sont employés par les chercheurs du monde entier et désignés par l'acronyme CAQDAS (pour Computer-Assisted Qualitative Data Analysis Software). Parmi ces outils figurent en bonne place Nud*Ist, AtlasTi, Kwalitan, The Ethnograph, Intext (à ne pas confondre avec Intex) et bien d'autres[2]. Je situe l'origine des outils d'analyse qualitative informatisée dans les années soixante avec The General Inquirer ; un logiciel initialement développé chez IBM, notamment par Philip Stone, et toujours en activité aujourd'hui.

S'agissant d'outiller l'analyse compréhensive de données textuelles, j'ai également restreint mon intérêt aux outils dédiés à la langue française. Cela écarte les outils cités ci-dessus. La

[1] Je remercie pour leur aide Jean-Pierre Charriau et Francis Chateauraynaud (Doxa) ainsi que le Fonds National de la Recherche Scientifique (FNRS) qui a rendu cette étude possible.

[2] Je renvoie le lecteur intéressé à la rubrique de l'Open Directory Project dédiée à ces outils que je tiens personnellement à jour : http://dmoz.org/Science/Social_Sciences/Methodology/Qualitative/Tools/ .

raison réside dans les différences linguistiques et culturelles transportées par des énoncés provenant d'Angleterre ou de Belgique.

Le champ des logiciels concernés compte donc des outils comme Alceste, Candide, Civilité, Hyperbase, Leximappe, Prospéro, Réseau-Lu, Sampler, Sato, Sphinx, T-Lab et Tropes[3]. Dans ces logiciels, plusieurs techniques sont présentes qui permettent de rendre compte des corpus analysés (Jenny, 1997). Je me propose de me pencher sur les réseaux de mots associés (RMA). Ce type de construction est très apprécié car il se double souvent d'une représentation en deux dimensions sous forme de graphes. À la différence d'une arborescence, ce graphe réticulaire est non dirigé, il provient de la génération d'agrégats de mots fortement liés (la métrique dépend de la récurrence de leur cooccurrence). Ces agrégats (ou clusters) permettent de dégager des pôles (représentés par les nœuds) correspondant aux associations les plus fréquentes, donc, selon les utilisateurs de ces cartes (Callon *et al.*, 1991b ; Courtial et Kerneur, 1995 ; Teil, 1991 ; Teil, 1994), aux différents grands thèmes du corpus ; toujours selon le même principe de calcul, ceux-ci apparaissent à leur tour liés (ou non) entre eux.

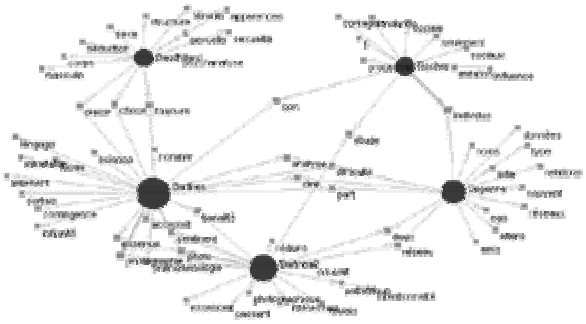

Fig. 1. Réseau-Lu

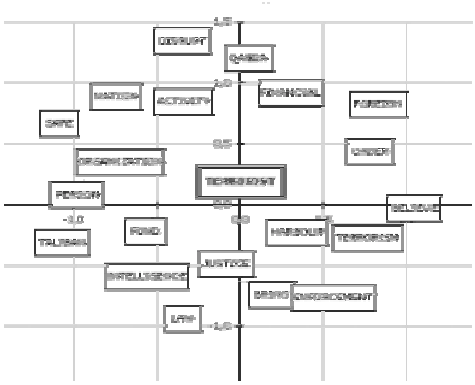

Fig. 2. T-lab

Les différents algorithmes qui permettent de construire ce type de réseau se basent sur la cooccurrence. Ce faisant, c'est le contenu des textes (et même plus précisément celui des énoncés) qui est analysé. J'écarte donc les réseaux de co-citations (Callon, 1993 ; Flake *et al.*, 2002) ou les réseaux de mots-clés pour conserver uniquement les techniques calculant un indice de cooccurrence déduit de données internes aux textes. Je rappelle que les techniques qui m'intéressent s'insèrent dans une démarche scientifique ; dès lors, l'exploitation de réseaux de mots associés dans des applications de représentation non-normée de l'information ne seront pas abordées frontalement, même si leur algorithme ne diffère pas radicalement de celui utilisé par les sociologues. Cette restriction écarte des outils comme ceux que mobilisent les outils de recherche d'information sur Internet, comme l'ancienne interface du moteur Altavista, appelée LiveTopics[4] (Bourdoncle, 1997), celle du méta-moteur Kartoo ou celle du nouveau moteur australien Mooter. Contrairement aux outils scientifiques, ces outils de recherche d'information produisent des graphes non-normés : leurs arcs sont dépourvus de métrique (ce qui signifie que leur longueur n'a pas de signification).

[3] Pour une description de ces logiciels, voir mon site Internet : http://www.smess.egss.ulg.ac.be/lejeune/logiciels/

[4] Après l'abandon par Altavista de la technologie Refine/Cow9/Live, son concepteur a lancé Exalead qui double la génération automatique de mots-clés d'une base de données conçue « à la main ».

Enfin, les réseaux que je vais évoquer ne doivent pas être confondus avec les ontologies (que l'on désigne parfois par le terme de réseau conceptuel) ou encore les réseaux sociaux (qui illustrent des liens entre personnes et ne sont que très rarement construits à partir de textes, il s'agit d'ailleurs le plus souvent d'une technique relevant de l'analyse quantitative des données qui, en en tant que telle, est extérieure à mon propos).

2. Des justifications sociologiques, philosophiques et linguistiques

Divers paradigmes sociologiques, philosophiques et linguistiques recourent à des méthodes prenant une forme réticulaire. Les justifications de l'usage des réseaux peuvent donc se rapporter à ces trois registres. La théorie de la traduction offre une justification **sociologique** à cette technologie : les acteurs visant à intéresser leurs interlocuteurs (comme les promoteurs industriels, les inventeurs ou les scientifiques) traduisent les objectifs des autres comme congruents avec les leurs (Callon *et al.*, 1991b ; Latour, 1995 ; Callon, 1993). Ce faisant, ils tentent de se constituer en point de passage obligé (Callon, 1986). Ces intéressements se concrétiseraient dans l'assemblage des thèmes (ceux des autres et les leurs propres) dans leurs productions textuelles. Par conséquent, l'analyse des réseaux de mots associés conviendrait parfaitement pour suivre les traductions des acteurs.

Au niveau **philosophique**, la référence se tourne vers le courant empiriste du XVIII^{e} siècle avec Étienne de Condillac et plus encore David Hume dont l'analyse de l'idée même de causalité comme résultant de l'habitude d'une conjonction d'événements (sans que la consécution n'existe ailleurs que dans l'esprit du sujet) sert de justification aux adeptes de ces méthodes[5].

Enfin, le lien avec les **sciences de l'énonciation** réside dans la possibilité de déduire le sens du contexte, donc de rendre compte des phénomènes d'indexicalité. Les exemples les plus courants de cette détermination du sens par le contexte sont les anaphores et les déictiques, mais la pragmatique et l'ethnométhodologie ont montré que ce phénomène était bien plus large, au point que l'on puisse parler d'indexicalité généralisée ou absolue (Garfinkel et Sacks, 1970 ; Conein, 1993 ; Ducrot et Schaeffer, 1995 ; Ramognino, 1999). En outre, l'analyse d'un corpus de textes recourrant aux réseaux d'associations prend en considération le cotexte proche (celui de l'énoncé) ; cette proximité n'est cependant pas assimilable à l'entourage direct puisqu'elle tient compte des termes très éloignés au niveau syntagmatique mais néanmoins présents dans la même phrase, donc cooccurrents[6]. Par conséquent, la cooccurrence ne doit donc pas se confondre avec la distribution.

3. La génération des réseaux de mots associés

La réflexion critique sur la cooccurrence est le cœur de cette contribution. Il ne s'agit aucunement d'une nouveauté : elle est déjà présente à la fin des années cinquante (Osgood, 1959). J'évoque pour ma part l'utilisation de cette procédure au sein de deux types technologies littéraires : les cartes de liens (traitées dans ce troisième point) et Prospéro (qui est abordé dans le quatrième et dernier point).

Leximappe, Candide, Réseau-Lu et T-Lab mobilisent le calcul des réseaux de mots associés.

[5] Cette lecture de Hume est celle de Latour et Teil (1995).

[6] Je montre dans la troisième partie comment Prospéro introduit une « pondération de proximité » dans son calcul de la force d'association de deux termes.

Le simple comptage du nombre de fois que se présente la présence simultanée de deux termes (ou co-présence ou cooccurrence) donne une évaluation de la force du lien entre eux. Toutefois, ce type de comptage dépend trop de la fréquence. On utilise donc un indice statistique normédéfini par Bertrand Michelet et appelé *indice d'équivalence* (noté E_{ij}).

On calcule cet indice de la manière suivante :

$$E_{ij} = \frac{C_{ij}^{\ 2}}{C_i \bullet C_j}$$

E_{ij} *désigne l'indice d'équivalence des mots i et j.*
C_{ij} *désigne la cooccurrence des mots i et j.*
C_i *désigne l'occurrence du mot i. Ce nombre est nécessairement supérieur ou égal à* $C_{ij.}$

L'indice d'équivalence évalue la force du lien : lorsqu'il vaut 0, les mots ne sont jamais présents ensemble ; s'il vaut 1, ils apparaissent toujours ensemble, on dit qu'ils sont équivalents (c'est, par exemple, le cas des mots composés). Les liens les plus forts sont réunis en agrégats (Callon *et al.*, 1991a).

Développé à partir de Leximappe™ par Geneviève Teil dans l'environnement Hypercard™, Candide™ ne tient pas compte des mots vides (articles, conjonctions, négations) et opère une lemmatisation des mots pleins. Afin d'éviter les ambiguïtés anaphoriques, les pronoms sont remplacés par leur antécédent (Ramaux, 1993).

Chaque unité de contexte élémentaire (qui correspond, dans le cas de ce logiciel, à la phrase) est transformé en réseau élémentaire grâce aux principes de calculs définis par Michelet. La somme de tous ces réseaux donne une représentation du texte.

À chaque mot est donc lié une série d'autres mots. La liste de ceux-ci s'appelle profil d'association (PA). La longueur du PA, c'est-à-dire le nombre de ses composantes, informe l'utilisateur sur le nombre de mots auxquels un terme est connecté. La comparaison des profils d'association est possible grâce à l'*indice de similitude* (noté S_{ij}) de Bertrand Michelet.

$$S_{ij} = \frac{PA_i * PA_j}{\left\|PA_i\right\| \bullet \left\|PA_j\right\|}$$

*« * » est le signe du produit scalaire; « • » est le signe de la multiplication.*
Le quotient désigne le produit scalaire normé des deux vecteurs.

Lorsque l'indice de similitude vaut 0, donc que les PA des deux mots considérés sont totalement hétérogènes, les deux mots appartiennent à des contextes différents.

Si l'indice de similitude vaut 1, les deux mots sont tous deux reliés aux mêmes mots. Si l'on considère que le sens d'un mot dépend donc de son contexte, on peut dire que les deux mots sont identiques au niveau de leur PA ; selon cette conception (le sens est contextuel), on peut considérer que le PA d'un mot en donne la définition.

Lorsque S_{ij} vaut 1, on se reporte à l'indice d'équivalence. S'il vaut également 1, les mots sont toujours associés *et* ils apparaissent toujours dans le même contexte. On peut donc les considérer comme un mot composé et les remplacer par un seul mot. C'est également la configuration des termes qui forment ce que l'on pourrait appeler, à la suite de Pierre Bourdieu, un couple épistémologique, et sont, à ce titre, tout à la fois antinomiques et complices, indissociables (Bourdieu *et al.*, 1968).

Par contre si E_{ij} vaut 0 (*et* que S_{ij} vaut 1), on a affaire à des mots toujours associés au même terme, qui ont « la même définition, mais n'apparaissent jamais ensemble. C'est le cas des synonymes » (Ramaux, 1993).

Candide fixe la frontière entre les agrégats lorsque se manifeste une « différence trop forte – supérieure à un seuil donné – entre la force de la dernière association prise en compte dans l'agrégat et la plus forte association suivante » (Teil, 1994).

Le degré de cooccurrence de deux concepts est proportionnel à leur proximité, tant thématique que graphique. Donc, deux mots fortement associés, dont l'indice d'équivalence (E) est élevé, sont proches ; l'arc qui les relie doit donc être court. La longueur des arcs (ou tenseurs) est par conséquent inversement proportionnelle à la valeur de E (Teil, 1991 ; Noyer, 1995b ; Jenny, 1997).

Cet outil permet de résoudre des cas de polysémie : un mot polysémique présente plusieurs « sous-profils d'association »; il est fortement lié à différents agrégats pas – ou peu – liés entre eux. On peut donc le scinder en deux mots (par exemple « avocat juriste » et « avocat fruit »). On peut utiliser le même raisonnement pour identifier les mots outils (ou mots vides) : il s'agit de mots fortement associés à l'ensemble des agrégats.

La lemmatisation et la suppression des mots vides (comme les négations), en tant que réduction sémantique et syntaxique, font que la simple présence d'un terme prime sur ce qu'il en est dit (Osgood, 1959 ; Jenny, 1997).

4. Prospéro

Le Programme de Sociologie Pragmatique et Réflexive sur Ordinateur[7] désigne une famille regroupant Ariel, Caliban, Chéloné, Marlowe, Sycorax et Tirésias (Chateauraynaud, 2003). Suivant la terminologie de Prospéro, les formes nominales et verbales sont respectivement désignées par les vocables « entités » et « épreuves » ; cette dernière notion, issue des nouvelles sociologies (Corcuff, 1995), désigne toute possibilité d'un changement d'état (Latour, 1984 ; Chateauraynaud, 1991). Parmi les fonctionnalités disponibles, je vais me focaliser sur le réseau, opération principalement assurée par Caliban mais également mobilisée par Marlowe.

4.1. *Des réseaux de cooccurrences*

Sous Caliban, l'algorithme du réseau travaille sur les listes d'entités. Il mesure le degré d'association de couples d'entités au sein d'un corpus. Cette mesure d'association est calculée comme suit : lorsque deux entités apparaissent dans la même phrase (c'est-à-dire dans la même chaîne de caractères située entre deux points), il y a cooccurrence. Si, parmi les mots qui se trouvent entre les deux entités se trouve au plus une épreuve, la valence de l'association est fixée à deux. S'il y a plus d'une épreuve entre les deux entités, la valence est de un pour cette phrase[8]. L'association entre deux termes peut-être jaugée pour un texte (on parle alors d'un réseau local) ou tout le corpus (il s'agit du réseau global) : les scores de chaque phrase sont alors simplement additionnés pour donner la mesure qui est utilisée dans la fenêtre des réseaux. Les profils d'associations d'une entité pour différents textes (ie les différents réseaux locaux) peuvent être comparés au réseau global, ce qui permet d'évaluer les divergences et originalités des différents textes.

La fenêtre du réseau global (voir ci-dessous) indique, pour un terme, l'ensemble des entités qui lui sont associées (et la force de ces associations). Le réseau est affiché en double : la

[7] Le lecteur désireux d'en savoir plus peut consulter le site http://www.prosperologie.org/.

[8] Le recours à cette valence permet de tenir compte de la « proximité » sur l'axe syntagmatique sans restreindre le co-texte au simple entourage du terme.

colonne de gauche indique le réseau d'entités, mais agrège les scores des entités contenues dans des êtres fictifs ; alors que la colonne de droite n'indique que les scores agrégés sous les catégories d'entités (les notions d'« êtres fictifs » et de « catégories » sont définis dans les points suivants).

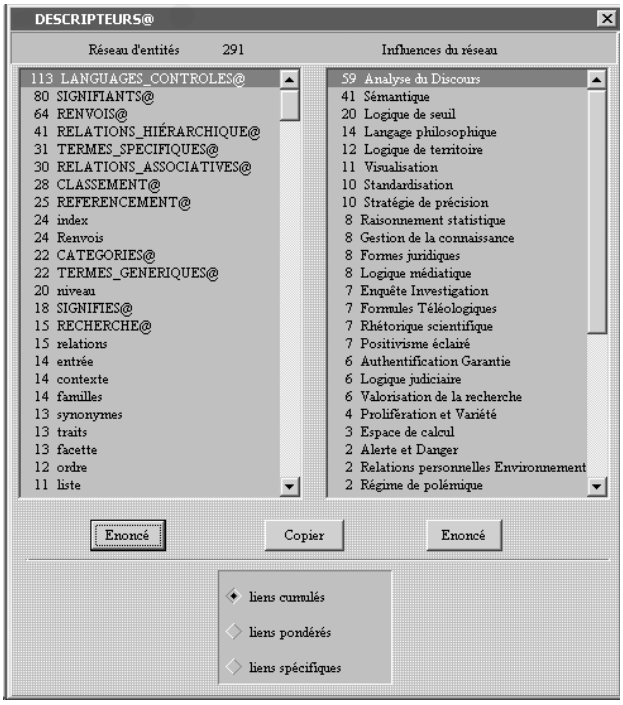

Fig. 3. Réseau global de l'entité DESCRIPTEURS@

Le réseau illustré indique l'univers conceptuel dans lequel est plongée une entité (Chateauraynaud 2003) ; il permet de suivre les traductions opérées par les différents acteurs (Callon, 1986). Ces traductions renseignent sur les connexions opérées par les acteurs entre différents thèmes ou préoccupations (Callon, 1993 ; Callon *et al.*, 1991b). Au bas de la fenêtre illustrée figurent deux autres options de calcul. Tout d'abord, l'algorithme dit des « liens pondérés » modifie la force des liens en fonction de la proportion de textes dans lesquels l'association est présente. Le calcul des « liens spécifiques » débarrasse quant à lui le réseau de la tête de liste d'entités associées ; celles-ci étant centrales dans le dossier, elles sont associées à tous les termes, ce qui est par conséquent peu parlant. Les liens spécifiques augmentent donc la probabilité de représenter des liens spécifiques les plus pertinents.

L'examen de la fenêtre du réseau global amène à trois remarques (que je commente dans les points suivants) : (1) aucune représentation graphique n'est ici mobilisée, contrairement à Candide, Leximappe, Réseau-Lu ou T-Lab ; (2) les réseaux sont parsemés de constructions de l'utilisateur (contrairement aux agrégats, qui sont générés par algorithme) ; (3) les deux colonnes représentent de manière différente le même réseau, en recourrant pour la première à des « ETRES-FICTIFS@ » et, pour la seconde, à des « Catégories ».

4.2. *Pas de graphe*

Contrairement au rendu des résultats exposés jusqu'ici, les concepteurs ont délibérément opté pour une présentation en liste. Ce choix a été motivé par les propriétés de ce format de présentation. Contrairement aux graphes réticulaires, en effet, la liste ne change pas de nature

lorsque son volume augmente. Ceci permet d'éviter les problèmes de lecture des réseaux touffus que le projet Textarc, par exemple, donne à voir (http://textarc.org/).

Fig. 4. Carte TextArc

Ce projet ne poursuit pas de retour critique sur la représentation des cartes de liens. Néanmoins, n'introduisant ni seuil ni pondération dans sa construction, il aboutit à des résultats qui ne gagnent ni en synthèse ni en clarté. Le travail avec des listes permet également de raisonner sur la queue de la série, donc de baser des interprétations sur les associations minoritaires, chose que l'accumulation quantitative des cartes cognitives interdit par définition (Chateauraynaud, 2003).

4.3. *Un masque de descripteurs surajoutés*

Outre l'absence de génération cartographique, les principe de construction des "agrégats" change également. Là où, précédemment, la composition des "clusters" était automatisée, elle est ici laissée à l'appréciation du chercheur. Cette différence influence donc les résultats et, par conséquent, la façon dont il faut les lire. Dans le premier cas, la cooccurrence amenait à agréger les liaisons les plus fortes représentant, selon les défenseurs de ce type d'outil (Callon, 1993), les thèmes, problèmes ou questions de préoccupation les mieux constitués. Dans le cas de Prospéro, l'accumulation des scores de plusieurs mots liés se fait selon d'autres principes que la connexion. Je vais les détailler et en dégager les conséquences.

La fenêtre comporte les constituants du réseau d'uns seule entité (ce dernier apparaît dans la barre de titre plus foncée en haut de la ***Fig. 3***). Outre les mots associés à ce terme (qui apparaissent en minuscules dans la première colonne) figurent des constructions de l'analyste qui

cumulent l'indice d'association de plusieurs entités. Dans la première colonne, ces constructions sont des êtres fictifs (leur nom apparaît en majuscules suivi d'un arobase @) ; dans la seconde, il s'agit de catégories d'entités (dont le nom est capitalisé).

4.3.1. *Êtres fictifs*

Les êtres fictifs réunissent la plage de variation des désignations d'un même actant sous une même fiction. Ces constructions – réalisées par le chercheur – consistent donc en une inscription des équivalences : l'opération consiste à collecter l'ensemble des désignations d'un même référent. La nature des objets étudiés en nouvelle sociologie (les controverses, les enrôlements, les disputes, etc.) rend ce travail d'autant plus nécessaire qu'ils impliquent plusieurs acteurs (dont les points de vue sont concurrents). Les variations de désignations se multiplient donc d'autant.

Comme le précise Sacks, cette multiplication des dénominations d'un même référent par les acteurs témoigne non seulement de leur intérêt pour celui-ci, mais également d'une « administration » du domaine qu'il recouvre. Cette gestion terminologique met en scène l'exercice d'une autorité : tenter d'imposer un nom revient à exercer un « contrôle social » ; le contester est une forme de « rébellion » contre celui-ci (Sacks, 1992a ; Silverman, 1998).

La plage des variations terminologiques qui composent un être fictif procède de la co-référence ou de la synonymie. La synonymie est entendue dans le sens suivant : suivant un point de vue de sociologie pragmatique, sont considérés synonymes deux termes que les acteurs mobilisent pour désigner le même phénomène. La construction des êtres fictifs n'est cependant pas indépendante du contexte : selon les affaires, les acteurs distinguent des désignations qui, dans d'autres circonstances, sont assimilées. Tantôt policiers et gendarmes seront substituables en tant que représentants des forces de l'ordre, tantôt ils se distribuent selon deux groupes aux intérêts divergeant (ce qui fut en partie le cas, par exemple, lors de la réforme belge des polices). Les êtres fictifs sont donc construits en fonction de l'étude sociologique.

4.3.2. *Catégories*

Dans la seconde colonne, on trouve des catégories. Contrairement à la colonne des êtres-fictifs, aucun terme n'apparaît seul. C'est pourtant exactement le même réseau que celui de la première colonne qui est représenté ici. Mais figurent uniquement les scores des termes codés sous une des catégories du cadre d'analyse du sociologue. Celles-ci ne sont plus construites comme une variation de la désignation. Les catégories recouvrent plutôt des univers de cohérence qui prennent sens dans un corpus déterminé et en fonction de l'analyse. Elles permettent d'asseoir l'interprétation sociologique et sont par conséquent modifiées, fusionnées ou disséquées en fonction du sujet traité, du contexte ou de ce que l'on veut montrer. Une catégorie correspond à une logique sociale, un registre d'action, un champ social (Bourdieu, 1980), une cité ou un monde commun (Boltanski et Thévenot, 1987).

4.4. *L'évolution du réseau*

Jusqu'ici, j'ai détaillé le calcul du réseau de manière statique : ce dernier agrège les scores des associations de tous les textes traités et synthétise donc ces associations sans tenir compte de la nature diachronique du corpus (la temporalité est donc écrasée). Or, les corpus étudiés en nouvelle sociologie sont souvent des affaires impliquant non seulement de multiples acteurs mais de nombreux événements. Les associations dont le réseau rend compte sont donc mouvantes au cours du déroulement du dossier. Pour rendre compte de cette dynamique temporelle, l'évolution et les reconfigurations successives du réseau doivent être examinées. On

peut ainsi identifier des périodes sur base de la stabilité des associations (ou, pour l'exprimer autrement, des dates de rupture, caractérisées par une modification des têtes de réseau).

4.5. *Le retour réflexif assisté par ordinateur*

Enfin, le réseau permet de tester la pertinence du codage des catégories. Cette pertinence du cadre d'analyse est tributaire de son adéquation aux données dont il entend rendre compte. Aussi le réseau fournit-il une ressource simple pour éprouver la pertinence du codage de l'utilisateur (c'est-à-dire du sociologue). La façon de procéder est la suivante. L'algorithme sélectionne parmi les termes appartenant à une catégorie celui qui établit le meilleur score (la plus grande occurrence). Le réseau de mots associés à cette forme nominale est examiné. Si, dans la tête de celui-ci, la catégorie sous laquelle il est codé est représentée, alors ce recouvrement est considéré comme un bon indice de la pertinence de la catégorie en question pour représenter les enjeux du corpus en présence.

5. Conclusions

Du côté de l'automatisation, j'apprécie le large éventail de résultats déductibles des réseaux de mots associés, qu'il s'agisse de ceux de Leximappe, de Candide, de Réseau-Lu, de T-Lab ou de Prospéro. Ce dernier logiciel est moins attractif que les autres : du coté de l'apparence, les listes ont un côté spartiate un peu rebutant. Ceci n'est pas qu'un jugement esthétique : la présentation des résultats en liste semble intéresser bien moins que les graphes (qu'il s'agisse des comités de revues scientifiques ou de l'assistance des colloques).

Néanmoins, la voie choisie par les concepteurs de ce dernier logiciel m'a séduit. Contrairement aux agrégats générés automatiquement, des constructions comme les êtres-fictifs et les catégories permettent au chercheur d'assurer la solution de continuité entre méthode et orientation théorique. Par conséquent, les outils offerts par Prospéro n'enferment pas leurs utilisateurs dans un paradigme scientifique déterminé (Kuhn, 1983) : ceux-ci peuvent tout aussi bien être psychanalystes, bourdieusiens ou ethnométhodologues. Prospéro consiste donc en un bon outil résistant aux boîtes noires. Mais cette liberté a comme pendant une responsabilité : chaque chercheur est responsable de ses catégories d'analyse, donc de son interprétation. Cet outil rend à l'activité scientifique sa dimension spécifique telle qu'elle a été analysée par les philosophes des sciences : la prise de risque (Stengers, 1992).

Références

Akrich M. (1991). L'analyse socio-technique. In Vinck D. (sous la coordination de), *Gestion de la recherche. Nouveaux problèmes, nouveaux outils*. Armand Colin.

Boltanski L. et Thévenot L. (1987). *Les économies de la grandeur*. PUF.

Boltanski L. et Chiapello E. (1999). *Le nouvel esprit du capitalisme*. Gallimard.

Bourdieu P. (1980). *Le sens pratique*. Minuit.

Bourdieu P., Chamboredon J.-C. et Passeron J.-C. (1968). *Le métier de sociologue*. Mouton/Bordas.

Bourdoncle F. (1997). LiveTopics : Recherche Visuelle d'Information sur l'Internet. *La documentation française, Dossiers de l'audiovisuel*, vol. (74) : 36-38.

Callon M. (1986). Éléments pour une sociologie de la traduction. La domestication des coquilles Saint-Jacques et des marins-pêcheurs dans la baie de Saint-Brieuc. *L'Année sociologique*, vol. (36) : 169-208.

Callon M. (1993). *La scientométrie*. PUF.

Callon M., Courtial J.-P. et Turner W. (1991a). La méthode Leximappe : un outil pour l'analyse stratégique du développement scientifique et technique. In Vinck D. (sous la coordination de), *Gestion de la recherche. Nouveaux problèmes, nouveaux outils*. Armand Colin.

Callon M., Courtial J.-P., Turner W. et Bauin S. (1991b). From translations to problematic networks : An introduction to co-word analysis. *Information sur les sciences sociales*, 2 (vol. 22) : 191-235.

Chateauraynaud F. (1991). *La Faute professionnelle. Une sociologie des conflits de responsabilité*. Métailié.

Chateauraynaud F. et Torny D. (1999). *Les sombres précurseurs. Une sociologie pragmatique de l'alerte et du risque*. EHESS.

Chateauraynaud F. (2003). *Prospéro : Une technologie littéraire pour les sciences humaines*. CNRS.

Corcuff P. (1995). *Les nouvelles sociologies. Constructions de la réalité sociale*. Nathan.

Courtial J.-P. et Kerneur L. (1995). Contribution de l'analyse des mots associés au suivi du développement d'un champ scientifique. In Cocaud M. (textes réunis par), *Histoire et Informatique. Base de données, recherche documentaire multimédia*. PUR.

Conein B. (1993). L'ethnométhodologie comme sociologie descriptive : réflexivité et sui-référentialité des catégories sociales. *Cahiers de recherche ethnométhodologique*, vol. (1) : 73-88.

Deloison J.-P. (1994). *Acquisition automatique de connaissances linguistiques pour l'indexation automatique. Rapport de DEA « Contrôle des Systèmes »*. Université de Technologie de Compiègne.

Ducrot O. et Schaeffer J.-M. (1995). *Nouveau dictionnaire encyclopédique des sciences du langage*. Seuil.

Flake G., Lawrence S., Giles L. et Coetzee F. (2002). Self-Organization and Identification of Web Communities. *IEEE Computer*, vol. (35/3) : 66-71.

Garfinkel H. et Sacks H. (1970). On Formal Structures of Practical Actions. In McKinney J. et Tiryakian E. (Eds), *Theorical Sociology. Perspectives and Developments*. Appleton-Century-Crofts.

Jenny J. (1996). Analyse de contenu et de discours dans la recherche sociologique française : pratiques micro-informatiques actuelles et potentielles. *Current Sociology*, vol. (44/3) : 279-290.

Jenny J. (1997). Méthodes et pratiques formalisées d'analyse de contenu et de discours dans la recherche sociologique française contemporaine. État des lieux et essai de classification. *Bulletin de Méthodologie Sociologique*, vol. (54) : 64-112.

Kuhn T. (1983). *La Structure des Révolutions Scientifiques*. Flammarion.

Latour Br. (1984). *Les microbes. Guerre et paix suivi de Irréductions*. Métailié/Pandore.

Latour Br. (1995). *La Science en action*. Gallimard.

Lejeune Chr. (2001). Du mode de définition de deux programmes de recherche en sociologie et en ethnométhodologie. *Carnets de bord*, vol. (2) : 56-66.

Lejeune Chr. (2002). Indexation et organisation de la connaissance. La régulation des décisions sur un forum de discussion. *Les cahiers du numérique*, vol (3/2) : 129-144.

Malingre M.-L. (1995). Une application de Candide, logiciel d'analyse textuelle pour une histoire de la traduction littéraire. In Cocaud M. (textes réunis par), *Histoire et Informatique. Base de données, recherche documentaire multimédia*. PUR.

Noyer J.-M. (1995b). Utilisation d'un outil Infométrique, « Candide » dans le contexte d'une réflexion stratégique. Les réseaux de simulation distribuée de l'armée américaine : émergence et description de l'émergence. In Noyer J-M. (sous la direction de), *Les sciences de l'information. Bibliométrie, scientométrie, infométrie*. PUR.

Osgood C. (1959). The representational model and relevant research methods. In De Sola Pool I. (Ed.), *Trends in Content Analysis*. University of Illinois Press.

Ramaux N. (1993). *Acquisition automatique et polysémique en Langage naturel. Rapport de D.E.A. « Contrôle des systèmes »*. Université de Compiègne.

Ramognino N. (1999). Linguistique et sociologie, un point de vue méthodologique. *Sociologie et sociétés*, vol. (XXXI/1) : 35-50.

Sacks H. (1992a). *Lectures on Conversation. Volume 1*. Blackwell.

Silverman D. (1998). *Harvey Sacks. Social Science & Conversation Analysis*. Polity Press.

Stengers I. (1992). *La volonté de faire science*. Les Empêcheurs de penser en rond.

Stone P., Dunphy D., Smith M. et Ogilvie D. (1966). *The General Inquirer: A Computer Approach to Content Analysis.* The MIT Press.

Teil G. (1991). Candide™ : un outil de veille technologique basé sur l'analyse des réseaux. In Vinck D.(sous la coordination de), *Gestion de la recherche. Nouveaux problèmes, nouveaux outils*. Armand Colin.

Teil G. (1994). L'analyse des goûts alimentaires par les réseaux. Une cartographie des critères du goût des consommateurs. In Courtial J.-P. (sous la direction de), *Science cognitive et sociologie des sciences*. PUF.

Teil G. et Latour Br. (1995). The Hume machine. Can association networks do more than formal rules ? *Stanford Humanities Review*, vol. (4, issue 2).

Thomas A. et Shearer J. (2000). *Internet Searching and Indexing. The Subject Approach*. The Haworth Press.

Weitzman E. et Miles M. (1995). *A Software Source Book. Computer Programs for Qualitative Data Analysis*. Sage Publications.

Yates S. (1996). Oral and Written Linguistic Aspects of Computer Conferencing. A Corpus Based Study. In Herring S. (Ed.), *Computer-Mediated Communication. Linguistic, Social and Cross-Cultural Perspectives*. John Benjamins.

Yates S. (2001). Researching Internet Interaction: Sociolinguistics and Corpus Analysis. In Wetherell M. Taylor S. et Yates S., *Discourse as Data. A Guide for Analysis*. The Open University.