Quatre approches pour l'analyse de données textuelles : lexicale, linguistique, cognitive, thématique

Bernard FALLERY, Professeur
Université Montpellier 2
CREGOR Montpellier-Management
Place Eugène Bataillon, 34000 Montpellier
Tel : (33) (0)4 67 14 42 21 ; Fax (33) (0)4 67 14 41 20
bernard.fallery@univ-montp2.fr

Florence RODHAIN, Maître de Conférences
Université Montpellier 2
CEROM Montpellier-Management
23000 Avenue des Moulins, 3185 Montpellier
Tel : (33) (0)4 67 10 28 20 ; Fax (33) (0)4 67 45 13 58
rodhain@polytech.univ-montp2.fr

Résumé
Cette contribution est d'ordre méthodologique.
L'Analyse de Données Textuelles (A.D.T.) regroupe aujourd'hui de nombreuses méthodes, et de nombreux outils, qui visent à découvrir l'information « essentielle » contenue dans un texte. En s'intéressant plus particulièrement à la demande des chercheurs en Management Stratégique, quatre exemples sont présentés, tous issus du même laboratoire de recherche, des différentes approches de l'A.D.T. **De quoi** parle-t-on? C'est le domaine de l'analyse lexicale. **Comment** en parle-t-on ? Il s'agit alors d'analyse linguistique. Comment **structurer** une pensée ? C'est l'ambition de la cartographie cognitive. Et enfin comment **interpréter** un contenu ? Il s'agit de l'assistance à l'analyse thématique. Pour ces quatre approches (illustrées par les quatre outils Alceste, Tropes, Decision Explorer, NVivo) on discute des problèmes méthodologiques et théoriques posés : discours et représentation, langage et énonciation, structuration et communication, interprétation et abstraction…

Mots clés : méthodologie, lexicale, linguistique, cartographie, qualitative.

INTRODUCTION

L'Analyse de Données Textuelles (A.D.T.) regroupe les méthodes qui visent à découvrir l'information « essentielle » contenue dans un texte, et le foisonnement de nouveaux outils auquel on peut assister aujourd'hui est à la conjonction de deux demandes différentes :

- d'une part une demande des entreprises, qui peuvent aujourd'hui collecter très facilement une grande quantité de textes avec Internet (articles, brevets, dépêches, rapports, études, mais aussi e-mails, messages de forums, enquêtes clients, fiches de centres d'appel, descriptifs de produits...). Il s'agit alors d'organiser automatiquement les contenus, d'extraire de l'information à partir d'un magma hétérogène de textes peu structurés. On constate alors une extension de la fouille de données textuelles *Textmining* ou de la cartographie *Web Positioning System*, pour la veille stratégique bien sûr, mais aussi pour l'indexation automatique de documents ou la capitalisation des connaissances (Wordmapper de GrimmerSoft, Zoom de Acetic, LexiQuest de SPSS, TextMiner de SAS …). Ces outils ont tendance aujourd'hui à compléter un « noyau dur » d'outils statistiques en ajoutant des environnements spécifiques : des lexiques et des ontologies de domaine, des serveurs d'annotations, le tout associé à des fonctions de robots aspirateurs et des outils de cartographie,

- et d'autre part une demande des chercheurs, qui ont besoin d'une alternative soit à de traditionnelles analyses de contenu jugées trop subjectives, soit à de simples analyses par mots-clés jugées trop pauvres (Bournois et al., 2002). On constate ici une extension des méthodologies qualitatives « assistées » par des outils quantitatifs (SpadT, Sphinx-Lexica, Alceste, Tropes, Decision Explorer, NVivo... parmi les plus cités en France) et les recherches en cours promettent encore de nombreux développements pour la formulation des requêtes « intelligentes » sur un corpus de données textuelles, que ce soit avec le Web sémantique (la spécification des ontologies en Informatique) ou avec le T.A.L. (la spécification des grammaires en Traitement Automatique des Langues).

En s'intéressant plus particulièrement à la demande des chercheurs en Management Stratégique qui considèrent que les discours (les investigations de terrain) constituent une entrée privilégiée de leur objet d'étude, seront d'abord présentés certains facteurs de choix d'un outil d'A.D.T. Quatre exemples, tous issus du même laboratoire de recherche, permettront ensuite de discuter les différentes approches de l'A.D.T. **De quoi** parle-t-on? C'est le domaine de l'analyse lexicale. **Comment** en parle-t-on ? Il s'agit alors d'analyse linguistique. Comment **représenter** une

pensée ? C'est l'ambition de la cartographie cognitive. Et enfin comment **interpréter** un contenu ? Il s'agit de l'assistance à l'analyse thématique. Pour chacune de ces quatre approches nous discuterons des problèmes méthodologiques et théoriques posés.

1. LES FACTEURS DE CHOIX D'UN TYPE D'ANALYSE DE DONNEES TEXTUELLES

Les chercheurs se situant dans le courant actuel de recherche en Stratégie qui valorise la dimension langagière et communicationnelle ont bien compris l'importance de se doter d'outils pour l'analyse des discours, outils par ailleurs devenus classiques en sciences humaines et sociales (en linguistique et en sociologie bien sûr, mais aussi en histoire, lettres, droit, médecine …). Analyser un discours relève toujours d'une créativité et d'un bricolage ; le profil de l'analyste reste donc une variable importante (discipline d'origine, référentiel théorique, compétences, entourage..). Au-delà de ce premier point, et pour une recherche en Stratégie, le choix d'un outil d'analyse devrait surtout dépendre de trois éléments : les choix méthodologiques, la constitution du corpus, et le moment de l'analyse statistique.

Tableau 1. Les facteurs de choix d'un type d'analyse de données textuelles

	Analyses Lexicales	**Analyses Linguistiques**	**Analyses Cognitives**	**Analyses Thématiques**
Cadre Méthodologique	- exploratoire - modèle	- exploratoire	- exploratoire	- exploratoire - modèle
Implication du chercheur	- Faible	- Forte - Faible	- Forte	- Forte
Axe temporel	- instantané - longitudinal	- instantané	- instantané	- instantané - longitudinal
Objet d'analyse	- un groupe	- un individu	- une situation	- un projet
Taille du corpus	- importante	- limitée	- limitée	- importante
Lisibilité Corpus	- forte	- forte	- faible	- faible
Homogénéité Corpus	- faible	- forte	- forte	- faible
Structuration langage	- faible	- faible		- forte
Moment de l'analyse statistique	- découverte ex-ante - contrôle ex-post	- ex-ante	- ex-post	- ex-post

1.1. LE CHOIX D'UN TYPE D'ANALYSE, EN FONCTION DES CHOIX METHODOLOGIQUES

Recherche exploratoire ou modèle d'hypothèses ? Coupe instantanée ou analyse longitudinale ? Implication du chercheur dans les récits (entretiens, observations…) ou analyse distanciée des pratiques (textes légitimés, discours officiels, enquêtes...) ? Les choix méthodologiques sont tous

acceptables, et ils n'ont qu'une seule exigence : être justifiés. Le choix de l'outil doit lui aussi être justifié par rapport à ces choix méthodologiques.

En prenant le cas des analyses lexicales, on pourrait dire qu'elles semblent adaptées pour une recherche exploratoire conduite sans a priori, puisqu'elles n'exigent au départ aucun présupposé concernant le contenu du texte. Mais le croisement possible de variables signalétiques (age, catégorie sociale..) avec des spécificités lexicales répond à l'idée de la détermination d'un contenu par son contexte, et l'analyse lexicale peut alors devenir aussi un outil pour une recherche fondée sur un corps d'hypothèses (on verra d'ailleurs que certains outils ont étés construits au départ sur cette idée).

1.2. LE CHOIX D'UN TYPE D'ANALYSE, EN FONCTION DU CORPUS

Pour l'adéquation entre un outil et un corpus, au moins deux questions méritent d'être débattues : la qualité et l'homogénéité de ce corpus.

La **qualité du corpus** est liée à sa taille et sa lisibilité. Pour **la taille** les avis sont partagés suivant que l'on envisage une analyse lexicale (plusieurs milliers de documents possibles) ou linguistique (cent pages est souvent proposé comme une limite). Il est clair que lorsque l'on a à traiter un grand nombre de données, un gain de temps significatif est obtenu soit par une première lecture lexicale, soit par une analyse thématique assistée par CAQDAS (Computer Aided Qualitative Data Analysis Systems). La **lisibilité** du corpus incite à la vigilance. Quand des ambiguïtés peuvent être liées à la non prise en compte du sens, il faudrait alors craindre une analyse lexicale : il est classique qu'un terme trop fréquent, ne présentant donc pas de cooccurrences particulières, ne soit pas distribué dans une classe particulière ; c'est alors la source d'une erreur d'interprétation qui amène l'analyste à considérer ce terme comme non signifiant pour les sujets, alors qu'il n'est que non spécifique. Quand l'analyse est rendue complexe par les stratégies discursives et les schèmes idéologiques des locuteurs, il faudrait alors craindre une analyse linguistique : on peut alors choisir une analyse thématique mais avec des allers-retours entre codage et décodage, et enrichir l'analyse avec des éléments illustratifs (sociaux, biographiques, thématiques, source du document, représentations du lecteur...) ou supplémentaires (rires, silences, points de suspension, questions du meneur de jeu...).

L'homogénéité du corpus devient de plus en plus problématique à mesure que l'on cherche à réaliser la fameuse « triangulation » des données, pourtant jugée si nécessaire à la validité logique des études qualitatives : analyser le discours d'une personne ou les discours de personnes différentes est un choix qui mérite réflexion. L'énonciation peut-elle être considérée comme liée à une certaine position socio-historique pour laquelle les énonciateurs seraient interchangeables ? Certains locuteurs parlent-ils au nom d'une structure (discours syndicaux et directoriaux…) ou s'expriment-ils en leur nom propre (entretiens, courriels …) ? Peut-on regrouper des communications orales retranscrites (répétitions nécessaires, dialogue orienté par un meneur de jeu, anonymat relatif, fonction émotive …) avec des écrits institutionnels (texte argumentatif, procédés rhétoriques, fonction conative..) ? Doit-on considérer les réponses à une question ouverte et les réponses données à l'issue de la relance, comme deux questions indépendantes ou comme une seule réponse ? Dans la construction progressive d'un échantillon pour des entretiens, la méthode classique des « choix raisonnés » assure-t-elle à la fois la similitude et la variété (représentation suffisante des statuts formels et informels, des rôles, des intérêts, des ressources, des relations d'alliances et d'oppositions…) ? Les réponses à ces questions devraient orienter le choix vers un type d'analyse, mais on verra que les outils proposés méritent une large discussion à la fois méthodologique et théorique.

1.3. LE CHOIX D'UN TYPE D'ANALYSE, EN FONCTION DU MOMENT DE L'ANALYSE STATISTIQUE

Face à des matériaux constitués en corpus, il est ensuite du ressort du chercheur de déterminer le niveau d'analyse auquel il prétend opérer et à quel type de questions il entend soumettre les textes : s'agit-il de *classer* des textes ou des fragments de textes, d'*extraire des informations* pour un commanditaire, d'effectuer *une synthèse*, d'effectuer l'*inventaire des thèmes* traités, d'enrichir un corpus de *commentaires*… ? Les différents outils n'opèrent pas tous sur les mêmes types d'objets et n'offrent pas tous les mêmes possibilités, et on peut finalement avoir une approche ex-ante ou ex-post.

Dans une approche statistique plutôt **ex-ante**, ce sont les traitements de données qui vont guider l'interprétation : on fait alors l'hypothèse que la structure formelle du discours implique des relations de sens établies par le sujet, et l'utilisation du logiciel constitue un outil méthodologique pour contrer les a priori du chercheur. On peut alors choisir une analyse lexicale : inventaire lexical du corpus traité, calcul des fréquences d'occurrence des mots, élaboration d'un réseau

graphique de différentes notions, élaboration de classes … le tout étant considéré comme un support pour une interprétation. Mais on peut aussi se situer dans une approche linguistique : étude des connecteurs dans le discours, progression thématique, analyse des marqueurs de forme dans le discours (forme assertive, interrogative, impérative...). Dans ces deux approches, lexicale ou linguistique, la question du sens est reportée en fin d'analyse au moment de la prise de connaissance des résultats (mais charge ensuite à l'utilisateur de dépasser les indices fournis en approfondissant « à la main » d'autres niveaux d'analyse).

Dans une approche statistique plutôt **ex-post** on se situera d'abord à un niveau extralinguistique, et l'outil ne sera utilisé dans ce premier temps que pour assister le codage du contenu dans une analyse thématique (affection à chaque fragment du discours de catégories thématiques que la lecture a permis de révéler, couplage avec des données sur le contexte de communication ou des données sociodémographiques sur l'émetteur…) ou pour assister le codage d'une carte cognitive (niveau d'abstraction des concepts, type de liens..). Ce n'est que dans un deuxième temps qu'on « contrôlera » l'analyse ou la carte par des traitements statistiques : ce type d'analyses statistiques ex-post permet notamment de contrôler les règles de la codification, mais surtout de découvrir des résultats contre intuitifs qui peuvent amener à modifier les catégories qui avaient été utilisées au départ.

2. L'ANALYSE LEXICALE : POUR DECRIRE « DE QUOI » ON PARLE

L'analyse lexicale est fondée sur les proximités entre les mots employés et la statistique fréquentielle : après une première étape de fabrication d'un *lexique* de mots puis de découpage du texte en *unités*, il est ensuite construit une matrice de présence/absence « *Mots* du lexique x *Unités* de texte ». A cette matrice on applique alors les méthodes de l'analyse de données multidimensionnelles fondées sur le Chi2 (analyse factorielle de correspondances, classification hiérarchique …) pour mettre en évidence les classes, les catégories ou les oppositions. L'objectivité proposée est liée au fait que les données sont traitées *sans a priori sur les catégories à découvrir.*

Encadré 1. Une analyse lexicale avec « Alceste » :

Pour établir une typologie des « récits de vie » des femmes collaboratrices

Il s'agit ici d'un projet EQUAL sur le statut professionnel des femmes en Europe (Fallery et Marti 2007). L'objectif était, pour des femmes collaboratrices d'artisans ou d'agriculteurs, de constituer un répertoire d'histoires que l'on puisse ensuite partager et commenter sur Internet. Le Corpus de départ était un ensemble de textes retranscrits après le recueil de dix premiers entretiens de type « récits de vie ». 96 histoires différentes racontées ont été d'abord repérées (unités initiales), et l'outil Alceste www.image.fr (Reinert, 1998) a été ensuite utilisé.

Sur environ 45.000 mots de départ, le logiciel en retient 3.600 pour construire le lexique. Sur les 96 histoires de départ, le logiciel fait un premier découpage en 1000 unités de contextes élémentaires, qu'il regroupe ensuite en 600 « Unités de Contexte » appelées UC. La matrice Lexique/Unités permet alors les différents traitements :

- Cinq classes différentes sont proposées par Classification Descendante Hiérarchique CDH, et en fonction de leurs mots caractéristiques le chercheur peut les « reconnaître » et les nommer. ***Classe 1 : la Maison-Famille*** *(essayer, chose, disponible, maison, organiser, important, séparer, fonction, projet, famille, semaine, bureau, ensemble, week-end, quotidien …)* ***Classe 2 : l'Emploi du temps*** *(matin, manger, soir, après-midi, devoir, scolaire, traite, partir, midi, école, linge, enfant, sœur, repas, quart…)* ***Classe 3 : le Commercial*** *(client, Internet, ordinateur, fournisseur, écouter, compte, appeler, structure, planning, lieu, message, rendu, expliquer, content, récent..)* ***Classe 4 : les Statuts*** *(exploiter, statut, installer, retraite, salarie, père, ville, conjoint jeune, reprendre, société, an, époux..)* ***Classe 5 : la Formation*** *(formation, comptable, gestion, technique, administration, acquérir, examen, coopérative, domaine, sein, prise, commission, entreprise, paie, établissement…)*

- Des relations d'opposition sont données par l'Analyse Factorielle de Correspondances AFC dans une représentation graphique. Dans cet exemple, l'axe 1 explique 34% de la dispersion totale, et il oppose « Maison-Famille » à « Statuts ».

- Des tris croisés permettent de croiser une variable signalétique avec le texte, pour analyser « qui parle de quoi ? ». Dans cet exemple, les femmes d'artisans et travaillant à mi-temps parlent beaucoup moins d'« Emploi du temps ».

- Une analyse par Classification Ascendante Hiérachique (CAH) permet, pour chaque classe, de définir différentes sous-classes. Ici par exemple, la Classe 2 « Emploi du temps » peut se décomposer en : le soir, les devoirs, les repas, les problèmes, les aides, l'école, les week-end.

2.1. PREMIER POINT DES ANALYSES LEXICALES : LE DECOUPAGE EN UNITES, PUIS LA CLASSIFICATION

Après une opération de lemmatisation (c'est-à-dire la fabrication d'une forme réduite du texte, standardisée par des dictionnaires) le premier découpage se fait dans Alceste en Unités

Elémentaires de Contexte (appelées UCE) qui sont automatiquement composées d'*une à trois lignes de texte* consécutives. Ces premières UCE sont ensuite regroupées en Unités de Contexte (appelées UC) qui contiennent un certain nombre de mots analysés différents (le logiciel calcule ce nombre suivant la taille et la nature du texte à analyser, mais on peut faire différentes simulations).

A partir du tableau binaire de présence/absence UC_i x Mots, la phase de Classification Descendante Hiérarchique CDH consiste à **extraire automatiquement des classes** d'énoncés en cherchant les partitions qui maximisent le Chi2 (une double classification est faite, sur des UC de grandeurs légèrement différentes, ce qui minimise le risque d'erreur dû au découpage).

Les résultats donnent alors, **pour chacune des classes trouvées,** les mots et les phrases les plus significatifs, les segments répétés, les concordances des mots les plus caractéristiques. Un dendrogramme restitue sous forme schématique les mesures de proximités et d'éloignements des classes.

2.2 . DEUXIEME POINT DES ANALYSES LEXICALES : LES ANALYSES COMPLEMENTAIRES

Les tableaux peuvent être soumis à l'Analyse Factorielle de Correspondances (AFC) pour donner une représentation graphique des relations d'opposition. Les tris croisés permettent de croiser une variable signalétique avec le texte, pour analyser « Qui parle de Quoi ? ». Une analyse par Classification Ascendante Hiérarchique (CAH) permet aussi, dans chaque classe, de définir les différentes sous-classes. Enfin en observant la fréquence de certains mots-outils (les adverbes ou les locutions adverbiales, exclus au départ dans le calcul des classes) dans leurs Unités de Contexte respectives, on peut s'intéresser à certaines formes de modalisation (Gavart-Perret et al, 1998).

Pour des corpus de grande taille l'approche lexicale présente l'avantage de réduire considérablement le volume d'information à lire et à analyser, mais le calcul des propriétés statistiques du texte (richesse lexicale, indices de spécificité, segments répétés, associations...) offre surtout la possibilité de différentes **lectures assistées** (découvertes de résultats statistiques surprenants… donc nouvelles interrogations… donc retour au texte à partir de certaines entrées lexicales). La richesse des calculs proposés par tel ou tel logiciel peut donc devenir un critère de choix. La possibilité de définir des **dictionnaires spécifiques** (on les appelle des scénarios dans Alceste) permet de dénombrer dans le texte des formes correspondant à un dictionnaire construit,

et donc de relire le texte avec des « quasi-variables » dont l'opérationnalisation peut alors presque s'apparenter aux échelles d'un questionnaire fermé.

2.3. LES DIFFERENTS LOGICIELS D'ANALYSE LEXICALE

En France le site des Journées de l'Analyse de Données Textuelles JADT constitue une excellente source d'information, en Allemagne le site de INTEXT recense de nombreux logiciels libres. L'article de Jenny (1997) reste une référence incontournable. Sphinx-Lexica (où des variables de codification et leur présentation à l'écran peuvent notamment être définies par le chercheur, Moscarola et al. 2001) et Spad-T (qui permet notamment la modification interactive du vocabulaire et la séparation en formes lexicales actives ou illustratives, Lebart et Salem, 1994) sont souvent cités en France comme extensions « qualitatives textuelles » à partir d'un logiciel classique de traitement d'enquête par questions codées. Pour Alceste on consultera l'étude de Aubert-Lotarski et al. (2002) et la contribution de Peyrat-Guillard (2000) dans le domaine de la GRH. Des grilles d'évaluation et de comparaison de ces outils d'analyse lexicale ont été proposées dans le contexte industriel d'EDF (Bruguidou et al. 2000, Quatrain et al. 2004).

2.4. DISCUSSION SUR LES ANALYSES LEXICALES : LE TRAITEMENT DES AMBIGUÏTES ET LE PRESUPPOSE D'UNE « REPRESENTATION » DE LA REALITE.

Quelle que soit leur efficacité, les analyses lexicales ne manquent pas de soulever des questions, aussi bien méthodologiques que théoriques.

D'un point de vue méthodologique **le traitement des ambiguïtés** nécessite une très grande attention afin d'éviter les contresens. Le cas de l'affirmation et de la négation est un problème important : par défaut, les analyses ne se basent pas sur les marqueurs de modalisation (ne, pas...) pour établir la classification, plusieurs tests supplémentaires sont donc nécessaires pour cerner le niveau d'expression de la négation qui a pu être pris en compte. D'une façon plus générale l'exclusion des mots-outils (à, afin, alors…) dans les analyses lexicales, relève bien du « paradigme des mots-clés » cher aux documentalistes pour lesquels la sélection des mots descripteurs « les plus pertinents » suffirait à résumer un texte. D'autres ambiguïtés doivent être levées par l'amélioration des dictionnaires : on devra par exemple lier ensemble des locutions composées qui présentent une unité de sens (*coûts_de_transactions*), ou à l'inverse séparer deux sens qui utilisent le même mot. De ce point de vue le travail réalisé dans la communauté INTEX

est un des plus aboutis : système intégré de dictionnaires de type Delaf, Delacf, Delafm.. (formes et polyformes, usages…), définition de graphes pour créer des grammaires locales personnalisées, définition d'automates pour identifier et étiqueter des concordances complexes (quasi-segments) (Bolasco 2000, Silberztein, 2001). Enfin on peut se demander si deux classes lexicales pourtant bien « différentes » relatent toujours des prises de position dissemblables : deux classes peuvent relever de modes d'expression hétérogènes au niveau de la forme et être pourtant très proches sur le fond, si elles concernent en fait les mêmes opinions mais exprimées par des synonymes, des paraphrases, des périphrases, des formulations incomplètes, des ellipses, des commentaires sur les mots utilisés…

D'un point de vue **théorique**, ce problème de fond et de forme révèle en fait une conception particulière des rapports entre la réalité et le langage. Dans une analyse lexicale, on considère que le langage **sert à représenter** « la » réalité, ou que **la parole reflète la pensée** : pensée et paroles ne font que rendre présent un Réel, qui était déjà là mais partiellement absent. On considère donc, dans une vision plutôt positiviste, que les « objets du monde » ont des propriétés essentielles en dehors de la manière dont ils sont décrits, et la vérité se définit ici comme une adéquation des énoncés à la réalité, le langage possédant alors *« un statut de désignation et de représentation* » (Quéré, 1990).

Mais ce concept de « représentation » est pourtant loin d'être clair au niveau théorique : s'agit-il d'un système d'interprétation de la réalité ? d'une image rapportée à autre chose ? ou encore d'un processus de communication avec soi-même ? … Dans une acception plutôt sociologique et objective, les « représentations » sont proches des connaissances stabilisées (ce sont alors des concepts, paradigmes, énoncés, visions du monde...), alors que dans une acception plutôt psychologique et cognitive les « représentations » sont plutôt qualifiées de modélisations contingentes pour traiter une situation (ce sont alors des mythes, idées, pensées...). Une analyse lexicale considère finalement le langage comme une articulation de ces deux niveaux (représentations/connaissances plutôt collectives et représentations/idées plutôt individuelles) pour permettre de re-présenter sans ambiguïté une réalité préexistante : on peut parler d'une approche positiviste du rapport entre langage et réalité.

3. L'ANALYSE LINGUISTIQUE : POUR DECRIRE « COMMENT » ON EN PARLE

Nous qualifions ici ces analyses de « linguistiques », dans la mesure où elles ont l'ambition

d'appréhender deux niveaux du discours, tout en gardant à distance la subjectivité du codeur : non seulement la catégorisation morphologique et l'agencement syntaxique (Qui dit quoi ? À qui ?), mais aussi la correspondance sémantique et la modalité pragmatique (Comment ? Avec quels effets ?).

L'analyse linguistique repose sur l'idée qu'il existe des connections entre système linguistique et système cognitif, et il s'agit alors de prendre en charge à la fois les aspects liés à la cohérence **référentielle** (ce à quoi le texte se réfère : des substantifs, signes linguistiques qui renvoient à une réalité extra linguistique) et aussi ceux relatifs au contexte d'**énonciation** (comment est-ce dit : des verbes, des adverbes, des conjonctions, des connecteurs.. qui servent à traduire la relation du locuteur à la situation, son point de vue et ses jugements).

Encadré 2. Une analyse linguistique avec « Tropes » :
Pour expliciter comment se construisent les convictions d'un créateur d'entreprise

Il s'agit ici de la thèse de Y. Andrieux (2005) sur l'élaboration des projets de création d'entreprise. Pour l'évaluation du projet, le créateur a en charge de faire partager par des tiers la vraisemblance de « l'ordre nouveau » qu'il propose. Pour asseoir la coordination des points de vue des différentes parties sur la viabilité du projet encore virtuel, l'objectif est ici de pouvoir expliciter la genèse des convictions du porteur, et pour cela de repérer les intentions en analysant la modalisation de son discours.

Le Corpus a été constitué de 21 discours de créateurs (de 12 pages en moyenne) sur leurs convictions quant à la viabilité de leur projet. On a utilisé l'outil « Tropes » (www.acetic.fr) fondé sur l'Analyse Cognitivo-Discursive (ACD, Ghiglione et al, 1998).

Dans un premier temps, grâce à des scénarios préexistants (dictionnaires d'*équivalents sémantiques)* et à des scénarios spécifiques (dictionnaires construits pour chaque entretien, par exemple Golfeur = Client), Tropes a permis de repérer les « **Univers »** des discours en analysant les substantifs (noms communs et noms propres du texte). On a pu ainsi distinguer « Gens », « Client », « Besoin », « Activité », « Démarche », « Informations », « Expériences antérieures », « Relations », « Documentation », en plus des deux Univers « Connaissance » et « Compétence » qui étaient déjà définis dans le scénario préexistant Concept.

Dans un deuxième temps on voulait comprendre le poids des « antécédents » (expériences, relations, documentation/observation) dans la genèse des « convictions » du porteur de projet, en étudiant **la modalisation** des discours (par analyse des verbes du texte). La façon dont un antécédent a été vécu a été décrite grâce aux différents verbes employés : 48 verbes d'action mentale (croire, penser, voir, sentir, ressentir …), 8 verbes dialogiques (dire, demander, montrer…), 11 verbes de volition (aimer, plaire, vouloir…) et 23 verbes d'autres actions humaines (créer, développer..). Chacune des milliers de propositions grammaticales numérotée et identifiée (c'est à dire chaque « conviction » se rapportant à l'un des Univers : besoin de la clientèle, compétences nécessaires, viabilité de l'activité…) a alors été couplée à un antécédent, le vécu de chaque couplage étant décrit en termes d'actes, grâce aux différentes catégories de verbes illustrant l'action : acte

d'interaction, acte d'observation, acte de perception….). Le récit a ainsi permis d'analyser la constitution des actes ayant formé le projet comme objet de pensée : par familiarité avec un phénomène (« je peux le refaire »), par schématisation du client type (« je crois que »), etc.

3.1. PREMIER POINT DES ANALYSES LINGUISTIQUES : LE DECOUPAGE PAR PROPOSITIONS, PUIS LA DEFINITION DES UNIVERS

Tropes prend non pas la phrase mais la **proposition grammaticale** (sujet, verbe, prédicat) comme unité de découpage : unité pertinente dans les théories cognitives et en même temps unité de découpage appropriée à un texte. A chaque proposition peut être attribué un score calculé en fonction de son poids relatif, de son ordre d'arrivée et de son rôle argumentatif, ce qui permet de repérer des propositions remarquables (thèmes, personnages, événements…) hors de toute interprétation préalable.

La relation entre l'activité cognitive et ses traces dans le discours se justifie ici par la notion de « **micro univers »** : « *Un sujet traite une information en mettant en scène un ensemble structuré et plus ou moins cohérent de micro univers, chacun étant peuplé a minima d'un actant qui fait l'action et de l'acte que le verbe accomplit »* (Ghiglione et al, 1998). Pour chaque mot d'une proposition, les **Univers** représentent le contexte, ils sont construits en regroupant les principaux **substantifs** du texte (noms communs et noms propres) grâce à des scénarios existants (dictionnaires d'*équivalents sémantiques)* et/ou à construire par le chercheur. Les relations entre univers peuvent alors indiquer quels sont les univers fréquemment rencontrés côte à côte à l'intérieur d'une même proposition, et on peut distinguer les univers qui sont généralement placés en position d'actant avant le verbe (*effectue l'action)* ou en position d'*acté* après le verbe (*subit l'action*). Dans l'ensemble d'un texte on peut repérer la répartition chronologique d'un univers (il peut apparaître beaucoup plus au début ou à la fin du texte).

3.2. DEUXIEME POINT DES ANALYSES LINGUISTIQUES : LE REPERAGE DES INTENTIONS PAR LA MODALISATION ET LES ENCHAINEMENTS

Comprendre un texte devient ici identifier les intentions, et les traces de l'intention se voient lors de **l'articulation de deux propositions** et le réseau de causalité sous-jacent. Dans la pratique deux notions sont alors utilisées : les connecteurs et les rafales.

Les **connecteurs** et **joncteurs** (conjonctions de coordination et de subordination, verbes, adverbes) relient des parties de discours par des notions de condition, cause, but, disjonction, opposition, comparaison, temps, lieu ou de manière. Ils permettent de situer l'action, de construire un raisonnement, d'énumérer des faits ou des caractéristiques, d'argumenter...

Une **rafale** regroupe des occurrences de mots (contenus dans un univers) ayant une probabilité à se répéter de manière importante dans une partie limitée du texte (au début, au milieu ou à la fin). Un **épisode** correspond alors à une partie du texte où un certain nombre de rafales se sont formées et terminées : ruptures thématiques (fin d'une série de rafales), passages où un nouvel épisode est développé (nouvelle série de rafales)… Le **style général** du discours correspond à la répartition des fréquences d'apparition des catégories de mots observées dans le texte, en comparaison avec des « normes » de production langagière : style Argumentatif, Narratif, Enonciatif ou Descriptif. Quand aux **mises en scène** verbales possibles elles sont les suivantes : mises en scène Dynamique, Ancrée dans le réel, Prise en charge par le narrateur, Prise en charge à l'aide du « Je ».

3.3. LES DIFFERENTS LOGICIELS D'ANALYSE LINGUISTIQUE

Alors que Tropes ne propose aucun a priori sur les Univers de référence, d'autres outils proposent au contraire de coder les fragments du discours suivant des « genres » fondés sur une référence théorique. L'outil MCA (Meaning Constitution Analysis www.mcadev.com), proposé en Suède par R. Sages, propose par exemple six dimensions fixes, inspirées de l'approche phénoménologique, où chaque *Unit* est à coder selon différentes *Views* : le type de croyance affichée (opinion générale, probabilité, hésitation…), la fonction (perceptive, imaginative, conative), le temps (passé, présent…), l'évaluation portée (positive, négative, neutre), la volonté (engagement, aspiration, absence) l'implication du sujet (je, nous, aucune) (Moscarola, 2001). De la même manière dans « The Ethnograph » (www.qualisresearch.com) le mode de fonctionnement relève aussi du codage de segments de texte par le chercheur puis d'un traitement quantitatif des codes résultants. « Prospéro » (www.prosperologie.org), proposé en France par Chateauraynaud (2003), est lui centré sur les *configurations* (dans lesquelles on définit des acteurs, des événements, des dispositifs, des arguments) et sur les *transformations* subies par ces configurations (basculements, mis en rapport avec le passé…), et le lecteur doit mettre à jour ses

propres catégories d'analyse en utilisant un double système de représentation (faits et interprétations).

3.4. DISCUSSION SUR LES ANALYSES LINGUISTIQUES : LE TRAITEMENT DES MODALISATIONS ET LE PRESUPPOSE DE L'ENONCIATION

Dans une approche linguistique il ne s'agit plus de considérer le texte « en extension » (inspiration plutôt positiviste), mais il s'agit bien de vouloir le saisir « en intention » et de reconstruire les mondes possibles du locuteur (inspiration plutôt constructiviste) en explorant les significations inscrites dans chaque fragment de texte. L'action prend place dans les rapports du langage et de la réalité, car les paroles ne font pas que véhiculer des informations ou fournir une « représentation » d'un objet indépendant : elles sont aussi, dans leur énonciation même, plus ou moins *performatives* et doivent être analysées en tant qu'actes, événements, pratiques sociales à part entière. Description **et** justification sont considérées comme relevant d'une même activité.

D'un point de vue **méthodologique**, c'est ici le **traitement des modalisations** qui constitue la racine du lien entre langage et réalité. Modaliser un discours, c'est en modifier la valeur, de façon linguistique ou non (signes non verbaux) : *« La modalité n'est jamais que le supplément de langue, ce par quoi, telle une supplique, j'essaye de fléchir son pouvoir implacable de constatation.* », R. Barthes (in Ghiglione & al 1998). La modalisation caractérise l'insertion du discours dans des contextes sociaux, elle traduit donc l'activité cognitive. Mais alors, n'y a-t-il pas une contradiction à considérer le discours comme le reflet d'un **acte** d'énonciation et à traiter un corpus composé du discours de plusieurs personnes ? Et si le message est considéré comme la trace d'une intentionnalité, les univers de référence ne sont-ils pas à définir en référence à un contexte social ou historique, à un ensemble des connaissances conditionnant une pratique ?

Ces questions nous ramènent au débat théorique sur la langue (outil de communication) et la parole (assimilée à un acte), ou au débat sur l'énoncé (le contenu) et l'énonciation (la mise en discours)… et il n'est pas si simple d'identifier la théorie qui sous-tend tel ou tel logiciel. Pour la **linguistique de l'énonciation**, un corpus doit être envisagé en tant qu'il a été produit par tel sujet, en se référant à Benveniste : *« la subjectivité … n'est que l'émergence dans l'être d'une propriété fondamentale du langage. Est Ego qui dit Ego »,* (cité par Andrieux 2005). C'est la subjectivité qui trouve son fondement dans le langage. La subjectivité ne précède pas la possibilité de son expression, c'est au contraire le matériel linguistique qui permet l'expression

de la subjectivité, qui permet au sujet de se situer dans et par le langage. L'acte d'énonciation révèle le sujet qui le pose, avant même de dire quelque chose sur le monde. A l'inverse, ce qu'on appelle l'école française de **l'analyse du discours** (Maingueneau, 1998) insiste sur les formations discursives en se référant à Michel Foucault : *« les discours religieux, judiciaires, thérapeutiques, et pour une part aussi politiques, ne sont guère dissociables de cette mise en œuvre d'un rituel qui détermine pour les sujets parlant à la fois des propriétés singulières et des rôles convenus »* (cité par Jenny 1997). Le discours est ici envisagé comme un ensemble de règles socio-historiques, déterminées dans le temps et l'espace, et qui définissent les conditions d'exercice de la fonction énonciative : le discours médical, le journal télévisé ou le cours magistral ne sont pas dissociables du personnage statutairement défini qui a le droit de les articuler. L'accent est mis ici sur les « stratégies discursives », que l'on peut alors considérer soit comme des conventions langagières plus ou moins consensuelles, soit comme des pratiques antagonistes de domination/résistance.

On voit qu'au-delà des aspects techniques et méthodologiques des logiciels d'analyse sémantique, la question **de l'interprétation de la modalisation** dans un texte renvoie à plusieurs théories des rapports du langage et de la réalité.

4. LA CARTOGRAPHIE COGNITIVE : POUR « STRUCTURER » UNE PENSEE

Une carte cognitive (un graphe des idées et des liens entre ces idées), représentation matérielle graphique des représentations mentales d'un ou plusieurs sujets à un moment donné, est généralement obtenue à partir d'une représentation discursive exprimée dans un texte ou un entretien.

Encadré 3. Une analyse cognitive avec « Decision Explorer» :
Pour structurer les différentes argumentations dans l'emploi des seniors

Il s'agit ici d'un projet EQUAL portant sur la gestion des seniors (Pijoan 2005). L'objectif de l'étude était de comprendre pourquoi peu d'organisations mettent en place des pratiques favorisant le maintien en emploi des seniors. Le Corpus était l'ensemble des textes retranscrits après entretiens auprès de directeurs de maisons de retraite sur le thème des employés seniors : recrutements, conditions de travail … L'outil Decision Explorer (www.banxia.com) a été utilisé, et des cartes cognitives ont pu être construites pour onze directeurs interviewés.

Dans un premier temps, un total de 172 idées différentes ont été repérées sur les onze cartes individuelles qui ont été construites, et 149 idées ont finalement pu être classées dans six catégories : les caractéristiques des seniors et des jeunes, les modalités du travail, le problème de l'âge, et les trois politiques de

GRH (politique en général, politiques centrées sur les seniors, politiques centrées sur les jeunes). Chaque carte contient une cinquantaine de concepts inter-reliés.

Dans un deuxième temps, comme l'objectif de l'étude était ainsi de comprendre pourquoi peu d'organisations mettent en place des pratiques favorisant le maintien en emploi des seniors, on a étudié les chaînes d'argumentation qui apparaissent sur les cartes. On a ainsi pu **classer les chaînes d'argumentation** concernant les stratégies de régulation et celles concernant les leviers d'actions possibles : les argumentations des directeurs apparaissent différentes suivant le type de situations rencontrées (situations harmonieuses ou situations conflictuelles) et suivant les visions du problème de l'employabilité (visions centrées sur les avantages/inconvénients des jeunes ou visions centrées sur les avantages/inconvénients des seniors).

4.1 . PREMIER POINT DES CARTOGRAPHIES COGNITIVES : LA COLLECTE ET LE CODAGE MANUEL DES IDEES

Pour la collecte certaines approches sont très structurées pour assurer la fidélité (« Self-Q » de Bougon, 1986), d'autres sont délibérément ouvertes pour assurer la validité (« Soda » de Eden et al. (1992), « Core » de Rodhain et Reix (1998), enfin certaines pourraient être qualifiées de mixtes (questions spontanées puis grilles d'exploration systématique, de Cossette (2003)). On peut travailler à partir de documents écrits, mais dès qu'il s'agit d'entretiens retranscrits, la place du chercheur est toujours considérée comme cruciale : *« Une carte cognitive est une représentation graphique de la représentation mentale* ***que le chercheur se fait*** *d'un ensemble de représentations discursives énoncées par un sujet à partir de ses propres représentations cognitives, à propos d'un objet particulier. »* (Cossette et Audet 1994).

Pour le codage, ce sont les modalisations (connecteurs et joncteurs) qui permettent de repérer les liens, et pour améliorer la fiabilité certains préconisent de soumettre aux répondants les délicates opérations de « fusion des concepts » (Allard-Poési, 1997). Il est alors possible de construire des cartes collectives, et l'élaboration d'une carte peut faciliter la transmission d'idées entre plusieurs individus : carte moyenne (un lien est retenu en fonction du score obtenu à un vote), carte assemblée (réunion de sous-cartes, après exclusion des concepts non communs) et souvent carte composite (qui résulte alors d'une communication, d'une véritable négociation de sens entre participants). Ceci ne doit pas cacher les difficultés du codage (différences entre données de faits et variables d'action, différences de niveau d'abstraction, équivalents sémantiques…) et le retour aux sujets apparaît alors comme un gage de validité.

4.2. DEUXIEME POINT DES CARTOGRAPHIES COGNITIVES : LA STRUCTURATION DES REPRESENTATIONS

Une fois construites de manière subjective mais rigoureuse, les cartes cognitives peuvent être analysées, avec ici l'ambition d'une lecture plus « structurelle » que ne l'autoriserait une approche lexicale ou linguistique. L'intérêt est de donner un poids aux concepts en fonction d'un indicateur, et non pas en fonction de l'importance perçue attribuée par les fréquences. Ces indicateurs de complexité et de complication permettent alors d'identifier les éléments autour desquels s'articulent les représentations des individus, leurs similarités et leurs divergences.

On peut d'abord caractériser les **propriétés structurelles** d'une carte, qui révèlent l'organisation des connaissances d'un sujet, sans considération quant à leur contenu : nombre total d'idées, nombre d'idées isolées, nombre de relations, rapport idées/relations, nombre de boucles, longueur des chaînes d'idées, nombre d'idées en entrée et en conclusion sur une chaîne d'argumentation…

L'analyse automatique de « cluster » consiste à identifier dans la carte des groupes de concepts mutuellement exclusifs, groupes d'idées faiblement dépendant les uns des autres.

La mesure de **l'importance d'un concept** peut ensuite être appréhendée par le nombre de facteurs auxquels il est relié directement ou indirectement : dans « Decision Explorer » on parle de « domaine » si on ne prend en compte que les concepts qui lui sont directement reliés, et on parle de « centralité » si on prend en considération la longueur moyenne de tous les sentiers reliant ce concept à d'autres. Bien que les cartes cognitives, dans la plupart des cas, ne prennent pas en compte la force des liens qui unissent les concepts, ces analyses permettent quand même d'identifier les noyaux du réseau constitué par la carte, sans que les interviewés aient toujours pleinement conscience de leur rôle.

4.3. LES DIFFERENTS LOGICIELS DE CARTOGRAPHIE COGNITIVE

Seule une carte capable de représenter l'ensemble des liens, quelle que soit leur nature, pourrait légitimement se voir attribuer le qualificatif de « cognitive » : relations causales, conatives, temporelles, composites, fortes/faibles… Ceci semble peu réalisable et dans la pratique on peut séparer les outils utilisés dans l'analyse de **relations causales** (bien que Decision Explorer propose différentes catégories de relations) et ceux utilisés dans les **associations sémantiques**. C'est dans cette catégorie que l'offre est aujourd'hui abondante dans une perspective soit de

veille sur Internet, soit de Gestion de Connaissances : WebRain, Internet Cartographer, MindManager, OpenMind, Inspiration, Freemind sous licence GNU.

L'étude de S. Trébucq (2004) sur les discours de la finance d'entreprise associe Tropes, Lexter et la cartographie Decision Explorer.

4.4. DISCUSSION SUR LES CARTOGRAPHIES COGNITIVES : LA RELATION CIRCULAIRE ENTRE LA CARTE ET LA PENSEE

La considération de différentes représentations « intermédiaires » (représentations mentales, discursives, graphiques… représentations du sujet, du chercheur..) est bien soulignée dans la littérature (Verstraete, 1996). Mais les relations entre ces représentations relèvent souvent d'une causalité linéaire et non pas d'un processus circulaire. Or la production de discours et de graphiques (la représentation) n'est pas sans produire d'effet sur la pensée (le représenté), ce processus conduisant alors à re-construire la représentation mentale. Deux questions théoriques sont alors posées : celle des rapports entre la pensée et l'action, celle des rapports entre la pensée et le langage.

La pensée est-elle première et l'action seconde ? L'élaboration d'une carte cognitive peut certes permettre de clarifier une idée confuse (structuration), d'envisager des voies d'actions possibles (aide à la décision), de faire prendre conscience à certains que ce qui est évident pour eux ne l'est pas pour les autres (communication), de passer du tacite à l'explicite (formalisation)… Mais en général les approches de la cartographie cognitive considèrent implicitement l'existence d'un **représenté statique** et posent comme hypothèse que la représentation **décrit et prévoit le comportement** d'un individu sincère qui agit en fonction des théories qu'il a adoptées (Pensée → Action). Laroche et Nioche (1994) critiquent alors les espoirs que certains chercheurs en stratégie mettent dans les cartes cognitives, à savoir qu'elles permettent de déceler ce qui initie le changement stratégique et de saisir la stratégie en tant qu'ensemble d'actions coordonnées. Cela revient à établir un lien de causalité du type « Problème → Réflexion → Action » : l'action stratégique suivrait la réflexion, que la carte permettrait de mettre en lumière. Or le modèle de la dissonance cognitive montre une attitude souvent rationalisante des individus et montre des théories reconstruites après l'action afin de retrouver la consistance et l'équilibre (Action → Pensée). On doit donc au moins considérer que le lien qui unit Action et Pensée est complexe et bouclé, il ne peut se réduire à un sens de la relation.

Les pensées d'un sujet, reflétées dans son discours, sont-elles antérieures à la demande du chercheur ? **La pensée est-elle première et le langage second ?** Merleau-Ponty (1945) répond clairement par la négative, il n'y a pas de pensée hors des mots, la vie intérieure est un langage intérieur : « *une pensée qui se contenterait d'exister pour soi, hors des gênes de la parole et de la communication, aussitôt apparue tomberait à l'inconscience, ce qui revient à dire qu'elle n'existerait pas même pour soi* ». Selon Pichot (1991) il s'agit d'une quasi-assimilation : « *la conscience des abstractions et concepts est exclusivement linguistique, le langage est donc l'expression consciente de la pensée, laquelle est alors conçue comme une activité psychique (voire nerveuse) discursive calquée sur l'activité linguistique qui est sa forme consciente*». Le discours met en forme les représentations mentales, il les influence : au fur et à mesure que l'individu s'entend parler, il modifie sensiblement ou insensiblement ses représentations mentales. « *Comment puis-je savoir ce que je pense avant d'avoir entendu ce que je dis ?* (Weick 1979) : pour l'individu le discours qu'il tient peut devenir lui-même sujet à découverte.

S'il n'y a pas indépendance entre la pensée et le langage, les représentations discursives influent alors sur la représentation mentale **durant le processus** de construction de la carte cognitive, comme elles ont influencé le processus de représentation mentale des concepts. Que se passe-t-il lorsque l'individu se trouve face à la carte tracée par le chercheur ? Il serait surprenant que la carte ne soit pas source de questionnement sur la pensée qu'elle est censée modéliser… et ainsi de suite jusqu'à ce qu'intervenant et individu, fatigués par ce jeu, admettent que la représentation graphique représente de manière satisfaisante une pensée que l'un et l'autre vont supposer stable. Le discours est de toute façon partial, puisqu'il a été aménagé de manière à ce qu'il soit reçu par le chercheur, et que la neutralité dans la réception du discours n'existe pas, « *on ne peut pas ne pas communiquer* » disent Watzlawick et al. (1972). Il n'existe pas de non-comportement, tout comportement a valeur de message.

5. L'ANALYSE THEMATIQUE : POUR « INTERPRETER » UN CONTENU

Quelle place faut-il laisser à l'**interprétation** ? Les outils lexicaux, linguistiques et cartographiques proposent tous une certaine objectivation, en standardisant la définition des catégories ou la structure des liens. A l'inverse le principe des CAQDAS (Computer Aided Qualitative Data Analysis Systems) est ici celui d'une analyse « *top-down* » qui laisse le codage des catégories au soin de l'analyste, mais en proposant de l'assister dans la gestion de ce codage

(gestion des liens entre les verbatim et les catégories en construction, annotations à volonté en ajoutant des propriétés aux segments textuels...). On prend donc ici en compte les processus interprétatifs dans la construction de la donnée, mais avec la possibilité d'augmenter la validité des analyses de contenu « classiques » qui ne proposaient qu'une approche méthodique fondée sur l'explicitation des règles de lecture, d'interprétation et de codage. Ces outils ont l'avantage de permettre de manipuler des unités non-linguistiques, ou du moins des unités qui sont hétérogènes: ce ne sont plus ni des lemmes ni des phrases, mais plutôt des notions (des mots, des idées, des paragraphes, des documents, des images, des propositions…).

Une analyse de contenu consiste à lire un corpus, fragment par fragment, pour en définir le contenu en le codant selon des catégories qui peuvent être construites et améliorées au cours de la lecture (c'est une approche constructiviste, avec le risque de changer la question de recherche en cours de travail). Dans un premier temps les significations des textes sont catégorisées selon le modèle qui guide le chercheur, c'est **la fameuse « grille d'analyse »** : matrices par phases ou par thèmes, évolution de ces matrices, cartes cognitives…. Dans un deuxième temps intervient l'analyse statistique sur les éléments de la grille d'analyse : fréquence d'apparition, variation selon les locuteurs, selon les contextes, interdépendance entre les éléments du modèle…

« NVivo » http://www.qsrinternational.com, « HyperResearch » www.researchware.com sont des logiciels pour gérer les liens entre des verbatim et des catégories en construction. Ils permettent au chercheur de manipuler des masses importantes de documents hétérogènes de façon itérative (allers-retours entre codage et décodage) pour étudier dynamiquement la complexité d'un corpus. Ils n'ont pas été conçus comme des outils d'analyse statistique, mais ils permettent l'exportation à travers la construction de « rapports ».

Encadré 4. Une analyse thématique avec « NVivo » :
Pour analyser de multiples données sur les stratégies de Gestion de la Relation Client

Il s'agit ici de la thèse de B. Bousquié sur les stratégies de Gestion de la Relation Client (Bousquié 2006). Le travail de terrain est une étude de cas en recherche participative sur plus d'une année, qui bénéficie donc d'un volume très important de données : plusieurs vagues d'entretiens directifs approfondis, des entretiens individuels semi directifs avec cinq nationalités, des entretiens semi directifs en groupe de travail, des notes de réunions, et très nombreux documents secondaires (au départ 60 Go de fichiers divers disponibles) : gestion de projet, suivi de projet, communication autour du projet… Dans un premier temps, et c'est la phase de **décontextualisation,** chaque document a été numérisé (avec récupération en type texte des tableurs et diaporamas) et chaque document ou extrait de document a été classé suivant plusieurs « Nœuds » décrits par leurs attributs : thème prédéfini pour un entretien, idée

émergeant à la lecture, concept théorique issu de la littérature, chapitre de thèse… La collecte et l'analyse ne sont pas séparables : un des objectifs du travail étant d'analyser les capacités mobilisées dans un projet CRM, on a par exemple qualifié en détail tous les extraits de textes qui concernaient une « capacité organisationnelle » donnée (maturité, compétences, impact…)
Dans un deuxième temps la manipulation du codage permet la gestion de l'arborescence des Nœuds, ou la fusion de Noeuds en une catégorie plus large avec héritage des attributs. La recontextualisation a par exemple consisté ici à construire « automatiquement » une matrice, un référentiel de capacités (capacités fonctionnelles, capacités techniques… X… maturité, Input nécessaires, Output possibles…).

5.1. PREMIER POINT DES ANALYSES THEMATIQUES : LA DECONTEXTUALISATION PAR LE CODAGE DES THEMES

NVivo utilise tout type de documents enregistrés au format .rtf (Rich Text Format), ce qui rend quand même exploitables certaines données issues de diaporama ou de tableurs. La décontextualisation consiste à sortir de son contexte un extrait du texte, afin de le rendre sémantiquement indépendant : cette étape de codage, entièrement libre et le plus souvent manuelle, permet de stocker les informations, de les qualifier et de les organiser. Pour **chaque Document** de base (documents numérisées qui peuvent être annotés, liés entre eux, ou liés à un fichier extérieur) et pour **chacun des Nodes** qui sont créés (un Nœud est comme un répertoire qui permet de coder chaque **extrait** de documents), on est amené à décrire ainsi des Attributs (avec un type et une valeur, qui peuvent d'ailleurs être importés d'un tableur) et des Sets (ensembles de Documents similaires ou de Nœuds similaires).

5.2. DEUXIEME POINT DES ANALYSES THEMATIQUES : LA RECONTEXTUALISATION PAR LES MATRICES ET MODELES

Recontextualiser consiste dans NVivo à regrouper les Noeuds pour en faire un tout intelligible et porteur de sens. La première fonctionnalité offerte permet de faire une **relecture assistée** du corpus : recherche textuelle sur un mot ou une expression (avec création possible d'un nouveau Nœud pour chaque recherche), recherche des co-occurrences en croisant un Attribut et un Noeud (ex : « hommes » x « en désaccord »), ou recherche matricielle (ex : Attributs x Valeurs x Nœuds) avec intersection, union, négation, différence, matrice d'intersection, matrice de différence.

La deuxième fonctionnalité consiste à créer des matrices (croisement de différents nœuds) et à créer des modèles (croisement de documents et/ou de nœuds). Une **matrice** est constituée d'un nœud-parent A (contenant plusieurs nœuds-enfants A1, A2, A3) que l'on peut croiser avec un autre nœud-parent B (contenant plusieurs nœuds-enfants B1, B2, B3). Un **modèle** est un schéma des relations, qui fait apparaître tous les éléments liés ensemble et qu'on peut alors étendre (différents types de flèches sont possibles). Une organisation en hypertexte de ces modèles permet de définir différentes couches à mesure que la compréhension progresse.

5.3. LES DIFFERENTS LOGICIELS D'ANALYSE THEMATIQUE

Bien que le point commun des outils d'analyses thématiques soit de proposer une assistance libre et en partie manuelle du codage des thèmes, on peut trouver dans cette catégorie des outils fort différents depuis « NVivo » qui ne propose donc ni dictionnaire ni analyse statistique … jusqu'à « Sato », développé au Québec www.ling.uqam.ca/sato/index.html, qui propose une véritable boite à outils collaborative d'indexation semi-automatique, allant de la désambiguïsation manuelle jusqu'à la création de lexiques spécifiques : les utilisateurs non satisfaits des analyseurs lexicaux « prêts-à-porter » peuvent alors mettre au point leur analyseur « sur mesure », puisqu'il semble en effet peu satisfaisant d'utiliser des méthodes et des dictionnaires uniformes pour des types de discours aussi différents que le management, la littérature, la chimie… (Armony et al. 1995). Dans « Sato » la catégorisation dite socio-sémantique vise à classer, de manière exhaustive et exclusive, les mots à valence référentielle (noms et adjectifs) en fonction d'un système de catégories thématiques. L'originalité réside ici dans le fait que l'outil permet d'ajouter tous types de propriétés aux mots ou segments textuels (propriétés syntaxiques, sémantiques, thématiques, contextuelles, etc…) et d'obtenir des « indices de thématisation » : ceci est le résultat soit d'une opération automatique (association des modalités d'une variable thématique à un ensemble de formes lexicales et de champs sémantiques larges ou restreints qui sont repérés automatiquement), soit d'une opération manuelle effectuée au cas par cas dans le texte (segmentation, puis nomination des diverses subdivisions) ou dans le lexique (catégorisation sémantique du vocabulaire).

5.4 DISCUSSION SUR LES ANALYSES THEMATIQUES : LE TRAVAIL DU CODEUR ET LE PRESUPPOSE CONSTRUCTIVISTE

D'un point de vue **méthodologique**, une catégorisation « en contexte » repose sur les **qualités du codeur**. Chaque occurrence est soumise à une décision : établir d'abord la pertinence de retenir le terme (a-t-il une signification « forte » et « précise », par rapport à la grille ?) et, le cas échéant, lui affecter un « marqueur » informatique. Les codeurs sont ainsi appelés à choisir parmi les différentes appartenances socio-sémantiques possibles d'un mot, celle qui est la plus proche de la signification en contexte de ce mot. Cela présuppose une connaissance des implications théoriques du système de catégories, mais une dynamique d'aller-retour fait en sorte qu'il soit possible de détecter des régularités dans les décisions qui n'étaient pas prévues et de détecter des inconsistances dans l'application de la grille. On peut donc dire qu'il s'agit d'un double processus d'apprentissage (sur la base de l'accumulation de décisions correctes) et de correction d'erreurs (sur la base de l'identification des décisions incorrectes).

D'un point de vue **théorique**, les analyses thématiques ont précisément comme problème la définition du **concept de « Thème »**. **Le thème**, construction intellectuelle élaborée par le lecteur à partir d'éléments textuels récurrents, est une *abstraction*. Il est donc tout à fait possible que le thème construit ne corresponde à aucune expression précise du texte, autrement dit que le thème ne soit pas *inscrit* dans le texte (le thème du « conflit de rôle » peut être prépondérant dans un texte, sans que les mots « conflit » ou « rôle » y apparaissent jamais). On ne peut ignorer la distinction fondamentale entre la fonction référentielle (le thème : ce dont on parle) et la fonction descriptive (le rhème : ce qu'on en dit) du langage. Plus le thème est abstrait, plus est grande cette possibilité d'*écart* entre les mots du texte et le thème élaboré. Un thème étant une construction, on peut alors considérer deux attitudes : soit préférer, comme avec « NVivo », **partir de lectures humaines** du texte (il s'agit donc de superposer aux données textuelles brutes un premier système de repères) puis réaliser ensuite des recherches lexicométriques ou hyper-textuelles, soit comme c'est possible avec « Sato », obtenir des **défrichements logiciels préalables** (richesse, originalité lexicale ou syntaxique…) que les interprétations humaines du thème et du contexte viendront ensuite compléter.

Une « bonne » interprétation des thèmes devrait pouvoir expliquer une pratique sans en réduire la richesse (c'est-à-dire la diversité avec laquelle elle peut donner lieu à des réalisations concrètes, dont l'échantillon d'observation peut rendre compte). La fiabilité de cette interprétation est liée à

la fois à la stabilité des « représentations » des énonciateurs et à celle du lecteur. Mais nous avons vu que ce concept de « représentation » est loin d'être clair au niveau théorique : « *Ce n'est pas un hasard si ce concept de représentation apparaît inopérant à des neuro-biologistes, délicat à utiliser à des psychologues, utilisable pour des ergonomes et des gestionnaires, et imprécis aux informaticiens de l'intelligence artificielle* » (Teulier-Bourgine, 1997). On peut au moins dire avec J.C. Abric (2001) que *« la représentation est un système de pré-décodage de la réalité, car elle détermine un ensemble d'anticipations et d'entente »* Dans la pratique ce **système de pré-codage de la réalité** est évidemment plus ou moins stable, et il se révèle donc dans un langage plus ou moins partagé. Indépendamment d'une démarche exploratoire ou confirmatoire (car le choix d'une de ces démarches ne dépend pas de l'état du langage plus ou moins partagé, mais de l'état des connaissances sur un sujet particulier), il y aurait donc des domaines où les énonciateurs et le lecteur peuvent disposer d'un **langage commun partagé** et structuré (système plutôt clos, qui autorise un pré-décodage manuel de la réalité et permet une analyse thématique « avec a priori ») et d'autres domaines où le langage est **en construction** (système plutôt ouvert, où les analyses lexicales et linguistiques du texte permettent dans un premier temps de travailler « sans a priori »).

CONCLUSION

Au terme de cette présentation, on peut faire deux constatations et une proposition.

- d'une part <u>les textes constituent bien des données</u>. On perçoit aujourd'hui l'intérêt de ces données pour éviter certains biais introduits par des techniques plus classiques comme le questionnaire, qui impose des rubriques préétablies et influence les réponses des sujets. Mais ceci impose alors des processus d'objectivation des unités textuelles (processus de réduction et de formalisation), et la statistique permet justement de tirer parti de la redondance de la langue pour réduire considérablement l'effort de lecture. L'analyse de données textuelles ne prétend pas se substituer à l'interprétation du sens des textes, il s'agit d'extraire des contenus ou une structure pour répondre à des questions précises, il s'agit aussi de construire des procédures exposant le regard du lecteur à des niveaux opaques de l'action stratégique d'un sujet. L'intérêt des classes d'énoncés qui rendent compte de l'organisation formelle du corpus réside finalement dans les possibilités d'interprétation sémantique qu'elles offrent : *« la linguistique nous propose des*

visions schématiques de la langue permettant de disposer des repères et d'aller, un peu plus sécurisé, explorer les plis et replis de nos textes » (Chateauraynaud 2003)

- car d'autre part les textes sont aussi le fruit d'une intention de la part des acteurs et l'objet d'une interprétation de la part de l'analyste. Comment faire cette interprétation ? Quel sens est-il possible de donner à ces classes ? On pourrait comparer les classes obtenues aux résultats d'un électrocardiogramme, et l'interprétation des courbes ou le choix d'une intervention revient toujours au chirurgien… Il n'est pas possible d'interpréter les classes en se souciant uniquement des significations apparentes auxquelles renvoient les mots qui lui sont spécifiques. Il importe de replacer chaque terme dans son contexte, et les données textuelles n'ont pas de sens a priori : la recherche du sens doit être menée parallèlement à celle des mesures et des structures. Il s'agit finalement de « confronter » la lecture du texte et les idées sur le texte (Desmarais et Moscarola, 2002).

Que l'on souhaite confronter un texte à un modèle de référence ou qu'on s'engage dans un processus exploratoire, la rigueur scientifique exige l'explicitation des méthodes et une certaine formalisation. Les outils qui existent aujourd'hui offrent déjà une liberté méthodologique, sans s'enfermer dans une technique imposée par un logiciel. Alors plutôt que d'opposer une approche algorithmique à une approche heuristique (analyse de contenu considérée comme subjective, analyse linguistique considérée comme objective, analyse de la constitution du sens considérée comme projective), on peut appeler à leur usage complémentaire dans une démarche algorithmique ET heuristique, composée des nécessaires cycles itératifs grille/texte, codage/décodage, extraction/validation…

REFERENCES

Abric J.C., 2001, *Pratiques sociales et représentations*, Paris, PUF, 2001

Allard-Poesi F., 1997, Nature et processus d'émergence des représentations collectives dans les groupes de travail restreints, Thèse de doctorat, Université Paris-Dauphine.

Andrieux Y., 2005, Contribution à la réflexion sur l'évaluation des projets de création d'entreprise : une approche centrée sur l'élaboration du projet. Thèse, décembre 2005, Université Montpellier 2

Aubert-Lotarski A., Capdevielle-Mougnibas V., 2002, Dialogue méthodologique autour l'utilisation du logiciel Alceste : lisibilité du corpus et interprétation des résultats. 6èmes journées JADT

Armony V., Duchastel J., 1995, La catégorisation socio-sémantique, 3èmes Journées JADT.

Bolasco, S. 2000, Taltac: un environnement pour l'exploitation de ressources statistiques et linguistiques dans l'analyse textuelle. Un exemple d'application au discours politique. 5èmes Journées JADT.

Bougon G.M., 1986, *Using the self-Q interview process*, Manual, Pennsylvania State University, Fithh Edition, June 1986.

Bournois F., Point S., Voynnet-Fourboul C., 2002, L'analyse de données qualitatives assistée par ordinateur : une évaluation, *Revue française de Gestion*, 137, janv-mars 2002.

Bousquié B. 2006, Gérer la relation client : les spécificités du contexte, 15 ème Conférence Internationale de Management Stratégique, AIMS, Annecy / Genève 13-16 Juin 2006

Brugidou M., Escoffier C., Folch H.,Lahlou S., Le Roux D., Morin-Andreani P., Piat G. , 2000, Les facteurs de choix et d'utilisation de logiciels d'Analyse de Données Textuelles, 5èmes Journées JADT

Chateauraynaud F., 2003, *Prospéro, une technologie littéraire pour les sciences humaines*, Paris, CNRS Editions, 2003.

Cossette P. et Audet M., 1994, « Qu'est-ce qu'une carte cognitive ? », *Cartes cognitives et organisations*, sous la direction de P.Cossette, Editions Eska, 1994.

Cossette P., 2003, Méthode systématique d'aide à la formulation de la vision stratégique : illustration auprès d'un propriétaire dirigeant, Revue de l'entrepreneuriat, vol 2, n1, pp 1-18

Desmarais C., Moscarola J., 2002, Analyse de contenu et analyse lexicale, le cas d'une étude en management public. Communication IREGE.

Eden C., Ackermann F. et Cropper S., The analysis of cause maps, *Journal of Management Studies*, vol.29, n°3, pp.309-324, may 1992.

Fallery B., Marti C., 2007, Storytelling on the Internet to develop weak-link networks. 9th International Conference on Enterprise Information Systems, EICIS 2007, Madère, Portugal.

Gavart-Perret M.L. Moscarola J., 1998, Enoncé ou énonciation ? Deux objets différents de l'analyse lexicale en marketing, *Recherche et application en marketing*, 1998, vol. 13, n°2.

Ghiglione R., Landre A., Bromberg M., Molette P., 1998, *L'analyse automatique des contenus*, Paris, Dunod , 1998.

JADT, Journées de l'Analyse de Données Textuelles, toutes les communications : http://www.cavi.univ-paris3.fr/lexicometrica/jadt/

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 (BMS) N°* 54

Laroche H. et Nioche J-P., « L'approche cognitive de la stratégie d'entreprise », *Revue Française de Gestion*, pp.64-78, juin-juillet-août 1994.

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

Maingueneau D., 1998. Les tendances françaises en analyse du discours, conférence à l'Université d'Osaka, compte-rendu sur Internet http://www2005.lang.osaka-u.ac.jp/~benoit/fle/conferences/maingueneau.html

Merleau-Ponty M., 1945, *Phénoménologie de la perception*, Gallimard, 1945.

Moscarola J., Papatsiba V., Baulac Y., 2001, Exploration sans a priori ou recherche orientée par un modèle : Contributions et limites de l'analyse lexicale pour l'étude de corpus documentaires. 5èmes journées JADT.

Moscarola J., 2001, Contributions des méthodes de l'analyse qualitative à la recherche en psychologie interculturelle : Sphinx et MCA, 8ème Congrès International de l'ARIC, Genève 2001.

Peyrat-Guillard D., 2000, Une application de la statistique textuelle à la gestion des ressources humaines : appréhender le concept d'implication au travail de façon alternative, 5èmes journées JADT.

Pichot A., 1991, *Petite phénoménologie de la connaissance*, Aubier, 1991.

Pijoan N. Expliciter les représentations des seniors chez des directeurs : une analyse a partir de cartes causales idiosyncrasiques, Journée de recherche AGRH, IAE Poitiers, Mai 2005.

Quatrain Y., Nugier S., Peradotto A., Garrouste D., 2004, Evaluation d'outils de TextMining : démarche et résultats, 7èmes Journées JADT.

Quéré L., 1990, Agir dans l'espace public. L'intentionnalité des actions comme phénomène social, in *Les formes de l'action*, Paris, Éd. de l'EHESS, p. 85-112

Reinert M., 1998, Quel objet pour une analyse statistique du discours ? Quelques réflexions à propos de la réponse Alceste. 4èmes Journées JADT

Rodhain F., Reix R., 1998, CORE : proposition d'une méthode pour l'élaboration des portefeuilles de projets SI, *Revue Systèmes d'Information et Management*, v.3, n°3, pp.49-83.

Simon H., 1981, *Sciences des Systèmes, Sciences de l'Artificiel*, traduction Dunod, Paris, 1996

Teulier-Bourgine R., 1997, Les représentations : médiations de l'action stratégique, in Avenier M.J, *La stratégie chemin faisant*, Paris, Economica, 1997

Trebucq S. 2004, Finance organisationnelle : un essai de représentation, 7èmes Journées JADT.

Silberztein M., 2001, *Manuel INTEX*, en français, disponible sur le site www.intex.de

Verstraete T., « La cartographie cognitive : outil pour une démarche d'essence heuristique d'identification des Facteurs Clés de Succès », Communication à la 5e Conférence Internationale de Management Stratégique. AIMS, Lille, mai 1996.

Watzlawick P., Helmick Beavin J. et Don D.Jackson, *Une logique de la communication*, Editions du Seuil, 1972.

Weick K.E., *The social psychology of organizing*, Mc Graw Hill Inc., (première édition : 1969), 1979.