

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

Bernard Fallery, Florence Rodhain

## ► To cite this version:

Bernard Fallery, Florence Rodhain. Quatre approches pour l'analyse de données textuelles: lexicale, linguistique, cognitive, thématique.. XVI ème Conférence de l'Association Internationale de Management Stratégique AIMS, 2007, Montréal, Canada. AIMS, pp 1-16, 2007. <hal-00821448>

HAL Id: hal-00821448

<https://hal.archives-ouvertes.fr/hal-00821448>

Submitted on 10 May 2013

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

Bernard Fallery  
Florence Rodhain

MRM-Université Montpellier 2

## Résumé :

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. **De quoi** parle-t-on? C'est le domaine de l'approche lexicale. **Comment** en parle-t-on ? Il s'agit alors d'approche 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'approche thématique. Pour ces quatre approches (illustrées en exemple par quatre outils : Alceste, Tropes, Decision Explorer et NVivo) on discute des problèmes méthodologiques puis théoriques qu'elles posent. La conclusion envisage les facteurs de choix d'un type particulier d'analyse qualitative.

**Mots clés :** méthodologie qualitative, lexicale, linguistique, cartographie cognitive, analyse de contenu.

## Abstract :

Analysis of Textual Data (ADT) has many methods and many tools, which aim to discover the core information contained in a text. Four examples of analysis are presented in this paper. What do we speak? This is the domain of lexical analysis. How do we speak? This is the domain of linguistic analysis. How to structure a thought? It is the ambition of cognitive mapping. And finally how to interpret a content? It is possible thanks to thematic analysis. For these four approaches (illustrated with the help of four tools: Alceste, Tropes, Decision Explorer and NVivo) are discussed theoretical and methodological problems they pose. The article concludes with key factors that help choosing a particular type of qualitative analysis.

**Keywords:** qualitative methodology, vocabulary, language, cognitive mapping, content analysis.

## INTRODUCTION

Derrière le terme générique « d'analyse de contenu », de nombreux outils manuels ont d'abord été proposés : grille de concepts, comptage des propositions, des expressions... (voir l'ouvrage fondamental de Huberman et Miles en 2002 et celui de Bardin réédité en 2007). Les facilités offertes par l'informatique permettent d'apporter une aide nouvelle aux analystes. L'Analyse de Données Textuelles (A.D.T.) regroupe les méthodes qui visent à manipuler l'information contenue dans un texte (Bournois et al., 2002).

En s'intéressant plus particulièrement à la demande des chercheurs en gestion qui considèrent que les discours (les investigations de terrain) constituent une entrée privilégiée de leur objet d'étude (Tesch 1990), sont présentés ici quatre types d'analyses qui permettent à chaque fois de discuter les différentes approches de l'A.D.T.. **De quoi** parle-t-on? C'est le domaine de l'approche lexicale. **Comment** en parle-t-on ? Il s'agit alors d'approche 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.

Nous parlons bien de quatre approches (esprit général d'une recherche, une sorte de disposition générale qui en situe l'arrière-plan) et non de quatre méthodes particulières (une méthode est un assemblage de trois composantes : une démarche, mais aussi un ensemble de concepts et un ou plusieurs outils). Dans les quatre premières parties nous détaillerons donc, pour chacune des quatre approches, les démarches possibles, les outils possibles et les problèmes théoriques qu'elles posent. En conclusion nous discuterons alors des critères de choix d'un type d'analyse textuelle.

## 1. L'APPROCHE LEXICALE : POUR DÉCRIRE « DE QUOI » ON PARLE

L'analyse lexicale, ou lexicométrie, est fondée sur la statistique fréquentielle (la redondance des traces lexicales) et les proximités entre les mots employés. On présentera ici quelques outils classiques en France (1.1), puis une discussion sur les différentes démarches de lecture « assistée » d'un discours (1.2) et enfin une discussion sur le concept ici fondamental de « représentation » de la réalité par le langage (1.3)..

### 1.1. Les outils d'une analyse lexicale

Chaque logiciel présente bien sûr des caractéristiques particulières, mais on peut repérer différentes grandes fonctions :

- une première fonction consiste à fabriquer un **lexique des mots** du texte : les mots sont en quelque sorte « déshabillés » de leur mise en forme grammaticale, c'est la lemmatisation, une opération qui n'est d'ailleurs pas anodine (Lemaire, 2008). Des traitements interactifs peuvent alors être réalisés sur ce lexique (recherche des occurrences, richesse lexicale, indices de spécificité, regroupement de formes lexicales particulières... (Gavart-Perret et al, 1998).

- le logiciel fait ensuite un découpage automatique du texte en **unités** (une unité peut être une phrase, une proposition grammaticale, un paragraphe...) et construit une grande matrice « **Mots x Unités** », c'est-à-dire un tableau de présence/absence des mots dans chaque unité. A cette matrice on applique alors les méthodes de l'analyse de données multidimensionnelles fondées sur la distance au sens du Chi2 (analyse factorielle de correspondances, classification hiérarchique ...) pour mettre en évidence les classes de discours (chaque classe contient les mots proches les uns des autres dans le discours de la personne ou de l'auteur) et les oppositions (les mots s'excluant mutuellement). Les données du texte sont ici traitées de façon « automatique », c'est à dire sans *a priori* sur les classes à découvrir.

#### Exemple 1. Une analyse lexicale avec « Alceste », pour établir une typologie de récits de vie.

Il s'agit ici d'un projet EQUAL pour étendre en Europe le statut professionnel spécifique de « femmes collaboratrices » (compagnes d'artisans ou d'agriculteurs travaillant à l'exploitation) (Fallery et Marti, 2007). L'objectif était 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 manuellement et l'outil Alceste (Reinert, 2008) a été ensuite utilisé.

Sur les 45.000 mots de départ, le logiciel en a retenu 3.600 pour construire automatiquement le lexique. Sur les 96 histoires de départ, le logiciel a fait un premier découpage en 1000 « unités de contextes élémentaires » (UCE), qu'il a ensuite regroupé en 600 « unités de contexte » (UC).

La matrice Mots/Unités a permis les différents traitements suivants :

- Cinq classes différentes ont été proposées par Classification Descendante Hiérarchique (CDH). En fonction de leurs mots caractéristiques le chercheur peut alors 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, salaire, 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 ont été données par l'Analyse Factorielle de Correspondances (AFC) dans une représentation graphique. Dans cet exemple, l'axe principal de l'AFC explique 34% de la dispersion totale, et il oppose la classe « la maison-famille » à la classe « les statuts ».
- Des tris croisés ont permis de croiser une variable signalétique de la personne qui parle avec le texte du discours, pour analyser « qui parle de quoi ? ». Dans cet exemple, les femmes d'artisans et travaillant à mi-temps parlaient beaucoup moins d'« Emploi du temps » que les autres.
- Une analyse par Classification Ascendante Hiérarchique (CAH) a permis, pour chaque classe, de définir différentes sous-classes. Ici par exemple, la classe 2 « Emploi du temps » a pu se décomposer en 7 sous-classes de discours sur le soir, les devoirs, les repas, les problèmes, les aides, l'école, les week-end.

En France le site de la revue électronique Lexicometrica ainsi que 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 l'analyste, 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 des extensions « qualitatives textuelles » à partir d'un logiciel de traitement d'enquête par questions codées. Pour Alceste on consultera l'étude de Aubert-Lotarski et al. (2002), la contribution de Peyrat-Guillard (2000) dans le domaine de la GRH, l'analyse de Chabchoub (2008) sur l'excès de confiance des dirigeants dans les cas d'acquisitions, l'analyse très détaillée de Kalampalikis (2003) sur les comportements alimentaires... Des grilles d'évaluation et de comparaison de ces outils d'analyse lexicale ont été proposées dans le contexte industriel d'EDF (Brugidou et al., 2000, Quatrain et al., 2004). On trouve dans l'ouvrage de Gauzente et Peyrat-Guillard (2007) une comparaison sur un même corpus de l'utilisation d'Alceste et de WordMapper.

## 1.2. Discussion sur la démarche dans une approche lexicale : la lecture « assistée »

Pour des corpus de grande taille l'approche lexicale présente évidemment l'avantage de réduire considérablement le volume d'information à analyser, mais le calcul des propriétés statistiques du texte offre surtout la possibilité de **lectures assistées** successives : découvertes de résultats statistiques surprenants... donc nouvelles interrogations... et donc retour au texte à partir de certaines entrées lexicales. On peut aussi faire des relectures du même corpus avec deux logiciels comme le propose Bonneau (2008) avec Alceste et Lexico. Des dictionnaires spécifiques (dans Alceste) ou des variables de codification (dans Sphinx-Lexica) permettant de dénombrer dans le texte des formes particulières, et donc de relire le texte avec des « quasi-variables » (variables qui pourraient presque alors s'apparenter aux items d'un questionnaire fermé). Quelle que soit la démarche de lecture assistée qui est adoptée, il faut garder à l'esprit deux problèmes ici très importants : le découpage du texte en unités et le traitement des ambiguïtés.

**Le découpage du texte en unités** (phrase, proposition grammaticale ou paragraphe) est certes réalisé par les logiciels, mais différentes simulations sont quelquefois possibles. Il faut prendre conscience que ces unités sont en fait des unités de sens et que c'est le logiciel qui fait ici ce travail de définition de sens, en choisissant le nombre de mots sur lesquels se base le calcul de découpage. Pour prendre un exemple un peu décalé, la phrase « je voyais tes yeux bleus dans la clarté du ciel déclinant » n'aurait évidemment pas le même sens que les deux phrases « je voyais tes yeux bleus » et « dans la clarté du ciel déclinant ».

**Le traitement des ambiguïtés** nécessite une très grande attention dans les démarches lexicales afin d'éviter les contresens. Le cas de l'affirmation et de la négation est ici un problème important : par défaut, les analyses lexicales ne se basent pas sur les « mots-outils » et notamment les marqueurs de modalisation (ne, pas...) pour établir la classification. 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 (un moyen de repérer et d'analyser des contraintes syntaxiques locales), définition d'automates pour identifier et étiqueter des concordances complexes (quasi-segments) (Bolasco, 2000, Silberztein, 2001). Enfin il faut 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 ...).

### **1.3 Discussion sur les présupposés théoriques d'une approche lexicale : le langage comme « représentation » de la réalité**

Les démarches de lecture assistée soulèvent en fait le problème du lien entre le fond et la forme, c'est à dire finalement celui des rapports entre la réalité et le langage. Dans une analyse lexicale on s'intéresse aux univers lexicaux (Reinert, 2008), on considère donc 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 du locuteur à la réalité (Kalampalikis, 2003), 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, le concept de « représentations » est proche des connaissances stabilisées (ce sont alors des paradigmes, des énoncés, des 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, des idées, des pensées...). Une analyse lexicale n'aborde pas cette question et elle considère finalement le langage comme une articulation de ces deux niveaux (le niveau représentations / connaissances collectives et le niveau représentations / idées individuelles) pour permettre de re-présenter « sans ambiguïté » une réalité **préexistante** : on peut parler d'une approche positiviste du rapport entre langage et réalité.

## **2. L'APPROCHE LINGUISTIQUE : POUR DÉCRIRE « COMMENT » ON EN PARLE**

Nous qualifions ici cette approche de « linguistique », dans la mesure où elle a l'ambition d'appréhender deux niveaux du discours : non seulement la catégorisation et l'agencement syntaxique (Qui dit quoi ?), mais aussi la correspondance sémantique et la modalité pragmatique (Comment ? Avec quels effets ?). On présentera ici quelques outils classiques en France (2.1), puis une discussion sur les démarches de traitement de la modalisation (la modification de la valeur d'un discours par des signes) (2.2) et enfin une discussion plus théorique sur le concept d'énonciation (la mise en discours) (2.3).

### **2.1. Les outils d'une analyse linguistique**

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 non seulement les aspects liés à la cohérence **référentielle** (« à quoi le texte se réfère-t-il ? » : c'est à dire des substantifs, des signes linguistiques qui renvoient à une réalité extra linguistique) mais aussi ceux relatifs au contexte d'**énonciation** (« comment est-ce dit ? » : c'est à dire 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).

### Exemple 2. Une analyse linguistique avec « Tropes », pour expliciter les convictions d'un entrepreneur

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 (ici le conseiller) 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 ses 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](http://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 Golfleur = Client), Tropes a permis de construire les « **Univers** » des discours en analysant les substantifs du lexique (noms communs et noms propres du texte). On a pu ainsi distinguer les Univers « Gens », « Client », « Besoin », « Activité », « Démarche », « Informations », « Expériences antérieures », « Relations », « Documentation », en plus des deux Univers « Connaissance » et « Compétence » (qui eux étaient déjà définis dans un dictionnaire préexistant dans Tropes appelé Concepts).

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 fine 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) a alors été couplée à un antécédent.

Le « vécu » de chacun de ces couplage a été 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 d'un locuteur a ainsi permis d'analyser la constitution **des actes ayant formé le projet** comme objet de pensée : constitution par familiarité avec un phénomène (« je peux le refaire »), constitution par schématisation du client-type (« je crois que »), etc.

Dans Tropes, après un découpage du texte en propositions grammaticales (sujet, verbe, prédicat) pour chacune desquelles est attribué un score (calculé en fonction de son poids relatif, de son ordre d'arrivée et de son rôle argumentatif), on ne propose aucun a priori pour définir automatiquement des classes de discours (appelées ici *Univers* de référence). Ceci est fait à partir des principaux substantifs du texte et grâce à des dictionnaires d'*équivalents sémantiques* (existants ou à construire par le chercheur). Pour chaque mot d'une proposition, les *Univers* représentent alors le contexte 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).

D'autres outils proposent au contraire de coder les unités du discours suivant des « genres » fondés sur une référence théorique préalable. L'outil MCA (Meaning Constitution Analysis [www.mcadev.com](http://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](http://www.qualisresearch.com)) le mode de fonctionnement relève aussi du codage des unités de texte par le chercheur puis d'un traitement quantitatif des codes résultants. Dans « Prospéro » ([www.prosperologie.org](http://www.prosperologie.org)), proposé en France par Chateauraynaud (2003), on se centre 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 l'analyste doit mettre à jour ses propres catégories en utilisant un double système de représentation (faits et interprétations).

## 2.2 Discussion sur la démarche dans une analyse linguistique : le traitement de la modalisation

Comprendre un texte devient ici identifier les intentions et donc le réseau de causalité sous-jacent, et les traces de ces intentions se voient lors de l'**articulation** de deux propositions (par exemple avec Tropes) ou de deux configurations (par exemple avec Prospéro). C'est ici la **modalisation** qui constitue la racine du lien entre langage et réalité. Modaliser un discours, c'est en modifier la valeur, de façon linguistique ou non (il y a aussi des 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.

Dans Tropes par exemple, deux notions sont utilisées pour analyser la modalisation : les connecteurs et les rafales. Les **connecteurs** (conjonctions de coordination et de subordination, verbes, adverbes) relient des parties de discours, ils permettent de qualifier l'action, c'est à dire : situer l'action, construire un raisonnement, énumérer des faits, argumenter... Les **rafales** regroupent 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) et un épisode correspond alors à une partie du texte où un certain nombre de rafales se sont formées ou terminées. Le style général du discours correspond à la répartition (en comparaison avec des « normes » de production langagière) des fréquences d'apparition des catégories de mots observées dans le texte : style Argumentatif, Narratif, Enonciatif ou Descriptif. On peut aussi repérer des mises en scène verbales : mises en scène Dynamique, Ancrée dans le réel, Prise en charge par le narrateur, Prise en charge à l'aide du « Je ».

On comprend bien qu'une analyse linguistique considère le discours comme le reflet d'un acte d'énonciation. Mais alors, au niveau de la démarche, n'y aurait-il pas une contradiction entre vouloir analyser un *acte* d'énonciation et traiter ensuite un corpus composé du discours de *plusieurs* personnes ? Le travail de Saint-Germes (2008) sur les pratiques d'évaluation de l'employabilité est un exemple d'analyse faite avec Tropes sur un corpus constitué de 19 entretiens avec des consultants RH.

## 2.3. Discussion sur les présupposés théoriques d'une approche linguistique : l'énonciation par un sujet ou l'énonciation suivant des règles sociales ?

Dans une approche linguistique l'**action** prend place dans les rapports du langage et de la réalité, car les paroles ne font pas seulement 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. Il ne s'agit plus de considérer le texte « en extension » (inspiration plutôt positiviste des analyses purement lexicales), mais il s'agit bien de vouloir le saisir « en intention » et de reconstruire les mondes possibles du locuteur en explorant les significations inscrites dans chaque fragment de texte (inspiration plutôt constructiviste, donc).

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)... Il n'est certes pas si simple d'identifier la théorie qui sous-tend tel ou tel logiciel, mais :

- Pour la **linguistique de l'énonciation**, un corpus doit être envisagé en tant qu'il a été produit par tel sujet, en se référant par exemple 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). 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 par exemple 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 alors 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, au-delà des différents outils et des différentes démarches, que la question de **l'interprétation de la modalisation** dans un texte renvoie à plusieurs théories des rapports du langage et de la réalité.

### **3. LA CARTOGRAPHIE COGNITIVE : POUR « STRUCTURER » UNE PENSÉE**

Une carte cognitive (un graphe des idées et des liens entre ces idées) est une représentation matérielle graphique des représentations mentales d'un ou plusieurs sujets à un moment donné. 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). On présentera ici quelques outils classiques (3.1), puis une discussion sur les différentes démarches de codage des idées (3.2) et enfin une discussion plus théorique sur la relation circulaire entre la carte et la pensée (3.3).

#### **Exemple 3. Une analyse cognitive avec l'outil « 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](http://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).

#### **3.1 Les différents 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, sémantiques, fortes/faibles... Dans les faits les outils disponibles sont plutôt spécifiques pour la représentation de certains types de liens.

Dans l'ensemble des outils de cartographie on peut distinguer :

- les outils utilisés dans l'analyse de **relations causales**, dont Decision Explorer est le représentant le plus connu. On peut y caractériser dans un premier temps les **propriétés structurelles** d'une carte (nombre total d'idées, nombre de relations, rapport idées/reliations, longueur des chaînes d'idées, nombre d'idées en entrée et en conclusion...). L'analyse automatique de « cluster » consiste à identifier dans la carte des groupes de concepts mutuellement exclusifs. 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;

- les outils 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.... Des outils comme MindManager ou TheBrain permettent de construire une carte conceptuelle, et prolongent cette carte par des liens avec des documents de travail, des systèmes de mémoire organisationnelle, des synchronisations avec Internet, etc.

Une fois construites de manière subjective mais rigoureuse, les cartes peuvent être analysées grâce aux outils logiciels, avec ici l'ambition d'une lecture plus « structurelle » que ne l'autoriserait une approche lexicale ou linguistique (bien que l'étude de S. Trébucq (2004) sur les discours de la finance d'entreprise associe à la fois Tropes, Lexter et la cartographie Decision Explorer). Un intérêt de la cartographie est de pouvoir donner un poids aux concepts en fonction d'un *indicateur*, et non pas en fonction de l'importance perçue attribuée par les *fréquences*. Des indicateurs de complexité (nombres de boucles par exemple) ou de complication (longueur des chaînes d'argumentation par exemple) permettent alors d'identifier les éléments autour desquels s'articulent les représentations des individus, leurs similarités et leurs divergences. On peut ainsi identifier les noyaux du réseau constitué par la carte, alors que les interviewés n'avaient pas toujours pleinement conscience du rôle que jouaient ces noyaux.

### **3.2. Discussion sur la démarche dans une analyse linguistique : le codage des idées et les cartes collectives**

Pour la collecte des idées du texte, les démarches peuvent être très différentes :

- certaines 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) ou « Core » de Rodhain et Reix (1998);
- enfin certaines pourraient être qualifiées de mixtes : questions spontanées puis grilles d'exploration systématique (Cossette, 2003).

Pour le codage, ce sont les modalisations (connecteurs) qui permettent de repérer les liens : conjonctions de subordination du type « parce que », conjonctions de coordination du type « donc », « car »,... ou encore verbes qui sous-entendent l'existence de relations de causes à effets (« impliquer », « permettre », « entraîner », « causer », « empêcher »...). Mais, dans la définition d'une démarche, l'opération la plus délicate est la « fusion des concepts » : Quels sont les équivalents sémantiques pour une même idée? Certains préconisent de soumettre aux répondants l'ensemble de ces délicates opérations de fusion (Allard-Poési, 1997). Ces questions ont été débattues par Point et Voynnet-Fourboul (2006) en définissant le

codage à portée théorique qui permet de coupler une approche inductive et une logique déductive.

Il est aussi possible de construire des cartes collectives, et l'élaboration d'une carte peut alors 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) ou carte composite (qui résulte alors d'une communication, d'une véritable négociation de sens entre participants).

### **3.3 Discussion sur les présupposés théoriques d'une cartographie cognitive : la relation en boucle entre la carte et la pensée**

La considération de différentes représentations « intermédiaires » (représentations mentales, graphiques... représentations du sujet, du chercheur...) est bien soulignée dans la littérature (Verstraete, 1996). Mais 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 *pendant* la construction de la carte. 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 (on est alors dans un schéma « Pensée → Action »). Dans le domaine de la stratégie Laroche et Nioche (1994) critiquent alors les espoirs que certains chercheurs mettent dans les cartes cognitives, car le modèle de la dissonance cognitive montre une attitude souvent rationalisante des individus (on est alors dans un schéma « Action → Pensée », Weick, 1979). 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 unique 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 ?** 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 »*. Pour Weick (1979) : *« Comment puis-je savoir ce que je pense avant d'avoir entendu ce que je dis ? »*. Pour l'individu le discours qu'il tient peut devenir lui-même sujet à découverte. La carte devient source de questionnement sur la pensée qu'elle est censée modéliser... et ainsi de suite, jusqu'à ce qu'intervenant et individu 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).

### **4.L'APPROCHE THEMATIQUE : POUR « INTERPRETER » UN CONTENU**

Une analyse de contenu « classique » 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... Dans un deuxième temps intervient une analyse statistique assez simple sur les éléments de cette grille d'analyse : fréquence d'apparition, variation selon les locuteurs, selon les contextes, interdépendance entre les éléments du modèle...

Quelle place faut-il laisser à l'interprétation ? Le principe des outils d'analyse thématique (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 : des mots, des idées, des paragraphes, des documents, des images, des propositions...).

#### Exemple 4. Une analyse thématique avec l'outil « NVivo », pour analyser de multiples données sur les stratégies de Gestion de la Relation Client

Différentes stratégies de Gestion de la Relation Client (GRC) dans les filiales d'un groupe mondial ont été analysées par B. Bousquie (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 **dé-contextualisation**, chaque document a été numérisé (avec récupération en type texte des tableurs et diaporamas) et chaque document ou extrait de document a été classé, par le chercheur, suivant plusieurs « Nœuds » décrits par des « Attributs » : il s'agit soit de thème prédéfini pour un entretien, soit d'une idée émergeant à la lecture, soit d'un concept théorique issu de la littérature... Chaque Nœud peut être considéré comme une sorte de « post-it » décrivant des méta-données sur un document.

La collecte et l'analyse ne sont pas séparables : un des objectifs du travail étant d'analyser les capacités mobilisées dans un projet GRC, ont été par exemple qualifiés en détail tous les extraits de textes qui concernaient une « capacité organisationnelle » donnée (maturité, compétences, impact...).

Dans un deuxième temps, et c'est la phase de **re-contextualisation**, la manipulation des « Nœuds » du codage a permis, dans cet exemple, la gestion de l'arborescence des Nœuds, la fusion de Nœuds en une catégorie plus large avec héritage des Attributs. On peut considérer que dans cette phase on manipule les « post-it » au gré des relectures successives de l'ensemble des documents.

Dans cet exemple la recontextualisation a consisté à construire avec l'outil NVivo une matrice appelée ici un référentiel de capacités (« capacités fonctionnelles, capacités techniques... » croisées avec ... « maturité, Input nécessaires, Output possibles »...), cela équivaut, de façon toujours imagée, à croiser des « post-it » avec des autres « post-it ».

On présentera ici quelques outils classiques (4.1), puis une discussion sur les différentes démarches de codage (4.2) et enfin une discussion plus théorique sur les concepts de thème et de représentation (4.3).

#### 4.1 Les outils d'une analyse thématique

Les outils d'analyse sont ici 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 vers des logiciels statistiques à travers la construction de « rapports ». 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 (Nvivo, Sato, HyperResearch, Max QDA et QDA Miner...), par exemple :

- NVivo propose d'abord une étape de codage entièrement libre et manuelle, qui permet de stocker les informations, de les qualifier et de les organiser : il s'agit de *dé-contextualiser*. Ainsi pour chaque **Document** de base (documents numérisés au format .rtf Rich Text Format, ce qui rend exploitables certaines données issues de diaporama ou de tableurs, et documents qui peuvent être annotés, liés entre eux, ou liés à un fichier extérieur) et pour chacun des

*Nodes* qui sont créés (un Nœud peut être considéré comme une sorte de *post-it*, 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). La deuxième étape dans NVivo consiste alors à *re-contextualiser*, à regrouper les Nœuds pour en faire un tout intelligible et porteur de sens. Les fonctionnalités offertes permettent de faire une **relecture assistée** du corpus (recherche des co-occurrences en croisant par exemple « hommes » x « en désaccord »), de créer des *matrices* (croisement de différents nœuds) ou des *modèles* (croisement de documents et/ou de nœuds).

- « Sato », développé au Québec [www.ling.uqam.ca/sato/index.html](http://www.ling.uqam.ca/sato/index.html), propose une véritable boîte 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 » (Armony et al., 1995). Dans l'outil Sato la catégorisation dite socio-sémantique vise à classer, de manière exhaustive et exclusive, les mots éventuellement ambigus (par exemple « Homme » en tant qu'être humain ou « homme » en tant que mâle...) en fonction d'un système de catégories thématiques (Gélinas-Chebat et al. 2008).

- « HyperResearch » [www.researchware.com](http://www.researchware.com) permet de gérer les différents *Cas* (les sujets abordés) et les différents *Codes* attribués (les catégories du codage). On peut coder différents types de données (du texte, mais aussi des parties d'images, des documents audio ou vidéo) et on peut ajouter des annotations. On peut gérer le regroupement automatique de toutes les données à partir de leurs codes, obtenir des rapports et des statistiques sur les codages et les exporter vers un tableur. Le test d'une proposition de recherche (*Hypothesis*) en possible en créant des règles conditionnelles sur les catégories.

#### **4.2 Discussion sur la démarche dans une analyse thématique : la qualité et le stabilité du codage**

Au niveau de la démarche, une catégorisation « en contexte » repose sur les qualités du codeur. 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. Le codage ne consiste pas simplement à appliquer un mot-étiquette à un extrait de matériau (Ayache et Dumez, 2011), car chaque unité de sens renvoie à plusieurs catégories ou noms. 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 (Point et Voynnet Fourboul, 2006). 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).

#### **4.3 Discussion sur les présupposés théoriques des analyses thématiques : les concepts de thème et de représentation**

D'un point de vue théorique, les analyses thématiques ont comme problème la définition du **concept de « Thème »**. Le thème, construction intellectuelle élaborée par le chercheur 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 du langage (le thème : ce dont on parle) et la fonction descriptive du langage (le rhème : ce qu'on en dit). Plus le thème est abstrait, plus est grande cette possibilité d'*écart* entre les mots du texte et le thème élaboré.

La fiabilité de cette interprétation est liée à la fois à la stabilité des « représentations » des énonciateurs et aussi à celle du chercheur. 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 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. Il y aurait donc des domaines où les énonciateurs et le chercheur 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 »).

## **5. DISCUSSION ET CONCLUSION : LES FACTEURS DE CHOIX D'UN TYPE D'ANALYSE DE DONNEES TEXTUELLES**

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, et pour une recherche en gestion, le choix d'un type d'analyse devrait surtout dépendre de trois éléments : le choix d'une démarche (exploratoire ou modèle), la constitution du corpus (5.1.), et le moment de l'analyse statistique (5.2).

### **5.1. 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, Reinert M., 2008) ou linguistique (cent pages est souvent proposé comme une limite, Escoubas-Benveniste M.P., 2010). 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 être très prudent dans 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 être très prudent dans une analyse linguistique. L'homogénéité du corpus peut devenir problématique : analyser le discours d'une personne ou les discours de personnes différentes est un choix qui mérite réflexion. 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, fonction émotive...) avec des écrits institutionnels (procédés rhétoriques, fonction conative...) ? Les réponses à ces questions devraient orienter le choix vers un type d'analyse, mais on a vu que les outils proposés méritent aussi une large discussion théorique.

### **5.2. Le choix d'un type d'analyse, en fonction du « moment » de l'analyse statistique**

Les différents outils n'opèrent pas tous sur les mêmes types d'objets et n'offrent pas tous les mêmes possibilités, et on peut avoir une approche statistique **ex-ante** ou **ex-post**.

Dans une approche statistique plutôt ex-ante, ce sont les traitements de données qui vont guider l'interprétation 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 statistiques. Dans une approche statistique plutôt ex-post il s'agit à l'inverse de contrôler l'analyse ou la carte par des traitements statistiques mais surtout de découvrir des résultats contre intuitifs qui peuvent amener à modifier les catégories qui avaient été utilisées au départ.

Regrouper ces remarques dans un tableau sera certes jugé difficilement admissible par nombre de spécialistes, tellement les types de logiciels offrent des fonctionnalités qui parfois se recouvrent. Nous le tenterons quand même, ne serait-ce que pour lancer la discussion.

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

	<b>Analyses lexicales</b>	<b>Analyses linguistiques</b>	<b>Analyses cognitives</b>	<b>Analyses thématiques</b>
<b>Démarche</b>	exploratoire ou modèle	exploratoire	exploratoire	exploratoire ou modèle
<b>Fondements théoriques</b>	positivisme	constructivisme	constructivisme	interprétativisme
<b>Implication du chercheur</b>	faible	Forte ou faible	forte	forte
<b>Axe temporel</b>	instantané ou longitudinal	instantané	instantané	instantané ou longitudinal
<b>Objet d'analyse</b>	un groupe	un individu	une situation	un projet
<b>Taille du corpus</b>	Importante	limitée	limitée	importante
<b>Lisibilité du corpus</b>	Forte	forte	faible	faible
<b>Écueil possible</b>	Ambiguïté sur le sens	Stratégies discursives	Ambiguïtés sur les concepts	Interprétation des thèmes
<b>Homogénéité corpus</b>	Faible	forte	forte	faible
<b>Structuration du langage</b>	Faible	faible	faible ou forte	forte
<b>Moment de l'analyse statistique</b>	découverte ex-ante contrôle ex-post	ex-ante	ex-post	ex-post

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

- d'une part les textes constituent bien des données, comme l'a montré l'ouvrage fondamental de Silverman (2006). 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. 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 : « *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)

- 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 aux classes obtenues ? On pourrait les comparer aux résultats d'un électrocardiogramme : l'interprétation des courbes ou le choix d'une intervention revient

toujours au chirurgien... 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 (nous n'avons pas évoqué ici les approches ethno-méthodologiques voir archéologiques pourtant utilisées en gestion, car nous voulions nous limiter à l'analyse de données textuelles). Dans ce cadre, 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 qui serait algorithmique (analyse lexicale considérée comme objective) à une approche qui serait heuristique (analyse thématique considérée comme subjective ou analyse cartographique considérée comme projective), on peut appeler à un usage complémentaire de ces approches dans une démarche à la fois algorithmique et heuristique, composée des nécessaires cycles itératifs entre grilles et textes, entre codage et décodage, entre extraction et validation.

## REFERENCES BIBLIOGRAPHIQUES

- Abric J.C. (2001), *Pratiques sociales et représentations*, Paris, PUF.
- 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*, Thèse, Université Montpellier 2.
- Aubert-Lotarski A. et Capdevielle-Mougnibas V. (2002), « Dialogue méthodologique autour l'utilisation du logiciel Alceste ». 6èmes journées JADT.
- Armony V. et Duchastel J. (1995), « La catégorisation socio-sémantique », 3èmes JADT.
- Ayache M. et Dumez H. (2011), « Le codage dans la recherche qualitative, une nouvelle perspective ? », *AEGIS*, Vol. 7, n° 2, pages 33-46
- Bardin L. (2007), *L'analyse de contenu*, Paris, PUF.
- Bolasco S. (2000), « Taltac: un environnement pour l'exploitation de ressources statistiques et linguistiques dans l'analyse textuelle », *5èmes Journées JADT*.
- Bonneau J. (2008), « Outils d'aide à l'exploitation d'entretiens semi-directifs : étude de l'interaction entre intervieweur et interviewés », *9èmes Journées JADT*.
- Bougon G.M. (1986), *Using the self-Q interview process*, Pennsylvania State University.
- Bournois F., Point S. et Voynnet-Fourboul C. (2002), « L'analyse de données qualitatives assistée par ordinateur : une évaluation », *Revue française de Gestion*, 137.
- Bousquié B. (2006), « Gérer la relation client : les spécificités du contexte », *15ème Conférence AIMS*, Annecy, Genève, 13-16 Juin.
- 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*, Paris, CNRS Editions.
- Chabchoub A. (2008), « Une application de la statistique textuelle à l'excès de confiance des dirigeants : cas des acquisitions », *9èmes Journées JADT*.
- Cossette P. et Audet M. (1994), « Qu'est-ce qu'une carte cognitive ? », in *Cartes cognitives et organisations*, sous la direction de P. Cossette, Editions Eska.
- 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, p 1-18.
- Desmarais C. et Moscarola J. (2008), « Analyse de contenu et analyse lexicale, le cas d'une étude en management public », *9èmes Journées JADT*.
- Eden C., Ackermann F. et Cropper S. (1992), « The analysis of cause maps », *Journal of Management Studies*, vol.29, n°3, pp.309-324, may.

- Escoubas-Benveniste M.P. (2010), « Le Monde et le grenelle de l'environnement pistes pour l'analyse sémantique assistée par ordinateur d'un corpus de presse », *11èmes journées JADT*.
- Fallery B. et Marti C. (2007), « Storytelling on the Internet to develop weak-link networks. 9<sup>th</sup> International Conference on Enterprise Information Systems », *EICIS 2007*, Portugal.
- Gauzente C. et Peyrat-Guillard D. (2007), *Analyse statistique de données textuelles en sciences de gestion : Concepts, méthodes et applications*, Paris, EMS.
- Gavart-Perret M.L. et Moscarola J. (1998), « Énoncé ou énonciation ? Deux objets différents de l'analyse lexicale en marketing », *Recherche et application en marketing*, vol. 13, n°2.
- Gélinas-Chebat C., Chebat JC et Therrien K. (2008), « Contribution de SATO à des analyses statistiques SPSS pour parler d'orientation dans les centres commerciaux », *9èmes JADT*.
- Ghiglione R., Landre A., Bromberg M. et Molette P. (1998), *L'analyse automatique des contenus*, Paris, Dunod.
- Huberman A. et Miles M. (2002), *The qualitative Researcher's Companion: Classic and Cotemporary Readings*, London, Sage.
- JADT, Journées de l'Analyse de Données Textuelles : toutes les communications, ainsi que la revue électronique Lexicometrica, disponible sur <http://lexicometrica.univ-paris3.fr/>
- 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*.
- Kalampalikis N. (2003), « L'apport de la méthode Alceste dans l'analyse des représentations sociales », in J-C Abric *Méthodes d'étude des représentations sociales*, Érès p. 147-163.
- Laroche H. et Nioche J-P. (1994), « L'approche cognitive de la stratégie d'entreprise », *Revue Française de Gestion*, pp.64-78, juin-juillet-août.
- Lebart L. et Salem A. (1994), *Statistique textuelle*. Paris, Dunod.
- Lemaire B. (2008), « Limites de la lemmatisation pour l'extraction de significations », *9è JADT*.
- Maingueneau D. (1998), « Les tendances françaises en analyse du discours », *conférence Osaka*, <http://www2005.lang.osaka-u.ac.jp/~benoit/fle/conferences/maingueneau.html>
- Moscarola J., Papatsiba V. et Baulac Y. (2001), « Exploration sans a priori ou recherche orientée par un modèle », *5èmes journées JADT*.
- Moscarola J. (2001), « Contributions des méthodes de l'analyse qualitative à la recherche en psychologie interculturelle : Sphinx et MCA », *8ème Congrès de l'ARIC*, Genève.
- Peyrat-Guillard D. (2000), « Une application de la statistique textuelle à la gestion des ressources humaines », *5èmes journées JADT*.
- Pichot A. (1991), *Petite phénoménologie de la connaissance*, Aubier.
- Pijoan N (2005), « Expliciter les représentations des seniors chez des directeurs : une analyse a partir de cartes causales idiosyncrasiques », *Journée de recherche AGRH*, IAE Poitiers.
- Point S., Voynnet-Fourboul C. (2006), « Le codage à visée théorique », *Recherche et Applications en Marketing*, Dec, 21, 4.
- Quatrain Y., Nugier S., Peradotto A. et 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. (2008), « Mondes lexicaux stabilisés et analyse statistique de discours », *9è JADT*.
- Rodhain F. et Reix R. (1998), « CORE : proposition d'une méthode pour l'élaboration des portefeuilles de projets SI », *Systèmes d'Information et Management*, v.3, n°3, pp.49-83.
- R. Tesch (1990), *Qualitative research: Analysis types and software tools*, Falmer, London.
- Teulier-Bourguine R. (1997), « Les représentations : médiations de l'action stratégique », in Avenier M.J, *La stratégie chemin faisant*, Paris, Economica.
- Trebucq S. (2004), « Finance organisationnelle : un essai de représentation », *7èmes JADT*.



Saint-Germes (2008), « Les pratiques d'évaluation de l'employabilité : une mise en tension de facteurs individuels, organisationnels et territoriaux », *Congrès de l'AGRH*, Reims.

Silberztein M. (2001), *Manuel INTEX*, en français, disponible sur le site [www.intex.de](http://www.intex.de)

Silverman D. (2006), *Interpreting Qualitative Data*, London, Sage.

Verstraete T. (1996), « La cartographie cognitive : outil pour une démarche d'essence heuristique d'identification des Facteurs Clés de Succès », *Conférence AIMS*, Lille

Watzlawick P., Helmick B. et D.Jackson (1972), *Une logique de la communication*, Seuil.

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