

FORMATION À LA RECHERCHE

REGARDS SUR DIVERSES APPROCHES DE TRAITEMENT DES DONNÉES TEXTUELLES : LES OUTILS, LEURS FONDEMENTS ET L'ÉPISTÉMOLOGIE DE LEURS USAGES

*François Larose, Université de Sherbrooke
Thierry Karsenti, Université du Québec à Hull
Vincent Grenon, Université de Sherbrooke*

INTRODUCTION

Il y a quelques mois, nous acceptons de produire un article introductoire portant sur différentes façons d'envisager le traitement des données textuelles. Plus que simplement objet de débat technique portant sur l'efficacité des méthodes, la question du traitement des données non quantitatives renvoie le plus souvent à l'épineuse question de la confrontation des méthodologies, et donc, des façons d'envisager les fondements épistémologiques de la recherche, notamment en sciences de l'éducation. Dans cet article, nous tenterons de réaliser le tour de force consistant à familiariser le lecteur avec certaines techniques tout en dédramatisant la question de la perspective dans laquelle se place le chercheur.

Cet article ne se situe donc pas sur le terrain des querelles mais plutôt sur celui de la complémentarité des méthodes et de la flexibilité des perspectives. Notons quand même que le chercheur se doit d'être prudent lorsqu'il parle de complémentarité des méthodes ou de méthodologie mixte. Selon Karsenti et Savoie-Zajc (2000), cette complémentarité doit être vue sous l'angle des choix des méthodes et des techniques de travail. Il est clair que les logiques épistémologiques sont, elles, différentes. « Un chercheur ne peut prétendre, à la fois, adopter une position neutre et objective dans sa recherche et, à la fois, subjective et immergée. [...] Un chercheur s'inscrit donc, par sa façon de poser son problème de recherche et sa question de recherche, dans une épistémologie particulière. Au plan du choix des techniques de travail, il veut toutefois peut-être élargir son éventail de moyens afin de produire des explications du phénomène étudié qui soient les plus riches possible. Sa recherche n'épousera donc pas une épistémologie positiviste et interprétative. C'est plutôt une vision pragmatique qui se développe, c'est-à-dire centrée sur une perspective intégratrice : la finalité de la production de résultats prévaut par rapport à la réflexion et au positionnement épistémologique » (p. 135).

DONNÉES TEXTUELLES, ORIGINE, NATURE ET FONDEMENTS

Qu'est-ce que le mot, le concept ou la phrase ? Pour un chercheur comme pour le commun des mortels, le mot ou la phrase sont des unités de sens ou des entités vectrices de sens. Bien entendu, ces unités ou ces vecteurs sont dépendants du contexte dans lequel ils se situent et, notamment, du contexte de communication ou d'interaction. À la différence du profane, le chercheur s'attache le plus souvent à la dimension conceptuelle que reflètent les mots ou les phrases émis par les tiers desquels il a obtenu quelques minutes ou quelques heures de production orale ou écrite. C'est ici que commence notre « calvaire » ! Qu'il s'agisse de mots ou de phrases, les énoncés se transforment subitement en données.

Qu'est-ce qu'une donnée textuelle, sinon une donnée qualitative, ou, si l'on préfère, une étiquette qui qualifie quelque chose, et si tel est le cas, qu'est-ce qui différencie cette donnée de toute autre donnée qualitative ? Rien du tout. Ainsi, si vous administrez un questionnaire à 100 personnes et que celui-ci inclut une rubrique d'identification qui, à son tour, offre au sujet la possibilité de se classer dans la catégorie « homme » ou « femme » de la variable « sexe », vous aurez créé deux types de données qualitatives (deux catégories) caractérisant une variable tout aussi qualitative. Que fait-on généralement de ces informations ? On leur appose une valeur numérique (un code) et on les additionne en attendant de s'en servir en tant que critère de distinction ou variable de contraste.

Les éléments que nous fournit le discours d'un ou de plusieurs sujets peuvent être considérés ou traités de deux façons. On peut en faire une lecture, une synthèse et utiliser le tout pour bâtir son propre discours. Sur ce terrain, on ne fait pas de la recherche mais bien de la production discursive, de l'énoncé d'opinion ou de la polémique. Idéal pour la philosophie, ce type de traitement de l'information ne se qualifie guère pour la recherche, même exploratoire, du moins *stricto sensu*. Inversement, on peut tenter de classer les unités de base du discours, de les regrouper selon un critère préalable correspondant à une conceptualisation et de les comparer, formellement ou informellement. Là, on entre sur le terrain de la méthode plus rigoureuse et systématique, et donc de la recherche. Pour ce faire, encore une fois formellement ou non, le chercheur devra considérer une unité (mot, concept ou segment d'énoncé) en tant que variable, ou encore, en tant qu'unité descriptive quantifiable. Pour illustrer notre propos, nous explorerons sommairement la logique de fonctionnement des logiciels de traitement de données textuelles.

**DU TRAITEMENT
DE TEXTE AU LOGICIEL
DE STATISTIQUE
TEXTUELLE :
LE MOT COMME
UNITÉ QUANTIFIABLE**

Les logiciels dits d'analyse qualitative, que les Anglo-Saxons considèrent comme des logiciels d'analyse de données qualitatives assistée par ordinateur (Barry, 1998), présentent tous certains éléments communs et partagent ces derniers avec tout logiciel permettant la saisie de données textuelles. Pour ces logiciels, un mot est une chaîne de caractères, donc un bloc graphique. Une phrase est donc une séquence de blocs graphiques séparés par un ou des espaces. En cela, tous ces programmes, incluant les traitements de texte¹, sont non seulement semblables, mais leurs procédures d'identification et de récupération des blocs graphiques sont identiques. Les « logiciels qualitatifs » offrent, bien entendu, des routines complémentaires au simple fait de pouvoir retrouver du texte, par exemple l'identification de liens de récurrence (nœuds) de concepts. En fait, leur principale caractéristique commune consiste à intégrer une procédure de marquage (*flagging* ou *tagging*) qui permet, à partir d'une console, d'accorder une valeur

¹ À ce propos, des chercheurs américains friands d'ethnométhodes, ont publié un article fort intéressant sur l'utilité du recours au traitement de textes en tant qu'outil de recherche « qualitative ». Voir sur ce point : Carney, Joiner et Tragou (1997).

hiérarchique à un mot (logiciels primaires comme Hyper-Qual, HyperRESEARCH, winMAX, code-A-Text ou Ethnograph) ou à une séquence de mots (logiciels plus complexes comme Atlas/ti ou NUD*IST).

Les logiciels de statistique textuelle, pour leur part (ex. : Lexico, SPAD-T, Sphinx², Alceste), bien qu'ils offrent aussi des routines de marquage du mot, par le truchement du mot-clé considéré en tant que variable, ont comme particularité que leur premier niveau d'opération n'est pas déterminé par le marquage. Ils considèrent simplement le texte en tant que série de blocs graphiques, pouvant être constitués d'unités (une chaîne de caractères) ou de séquences d'unités (plusieurs chaînes de caractères délimitées par les marqueurs lexicaux normaux du discours comme les points ou les virgules). L'utilisateur catégorise les éléments délimiteurs qui lui semblent organiser le discours. Certains logiciels offrent la possibilité, à partir d'une grammaire intégrée, de repérer les marqueurs syntaxiques du discours en tant que délimiteurs de zones ou variables.

À partir de la numérisation du discours, ces logiciels génèrent le plus souvent deux types de tableaux de fréquences croisées (tableaux à double entrée). Ceux-ci ont en commun d'utiliser un type de variable discriminante comme indice caractérisant les colonnes d'une part (par exemple, chaque sujet interviewé, une variable de contraste telle l'appartenance socioéconomique ou professionnelle, etc.) et, d'autre part, chaque mot, puis chaque regroupement de mots répétés au moins deux fois de façon identique dans le discours, en tant qu'indice caractérisant les lignes. Le premier tableau est donc un tableau de fréquence basé sur le mot et le second sur les segments de phrase répétés.

À partir de là, différents calculs, généralement non inférentiels et non paramétriques, peuvent être effectués. Trois types de calculs nous intéressent dans les exemples présentés ici. D'une part, ces logiciels évaluent les spécificités du discours. Par cela, nous entendons les éléments (mots ou segments répétés) qui sont surreprésentés ou sous-représentés dans le discours d'un sujet, d'une catégorie, etc. L'intérêt de cette opération réside en ce qu'elle ne se base pas sur un postulat de distribution normale des unités qui composent le discours mais bien plutôt sur une distribution probable ou bayésienne, qui tient compte de la structure réelle des corpus. Sur le plan pratique, cette procédure permet d'établir à coup sûr ce qui distingue le discours d'une entité. Par exemple, lors de l'analyse comparative du discours de divers auteurs (Conseil supérieur de l'éducation, ministère de l'Éducation, intervenants) au regard des concepts de pédagogie et de didactique, le logiciel Lexico indiquait ce qui suit (tableau 1) :

² Le Sphinx constitue un cas hybride en ce qu'il permet l'opération directe sur le texte, sans procédure statistique, ou la constitution de tableaux de fréquences complexes à partir desquels on peut procéder à des calculs de type multidimensionnel.

Tableau 1

Spécificités de la partie 20 (catégorie 20) - Lexico1					
n°	terme	F*	f**	spéc.	orig.
16	ou	53	5	+E04	0
84	classe	11	2	+E03	0
117	qui s'intéresse à une discipline	10	2	+E03	0
122	d'une discipline	10	2	+E03	0
46	pratiques	19	3	+E03	0
93	méthodes	10	2	+E03	0

*F = fréquence absolue de l'occurrence au travers du texte.
**f = fréquence spécifique à la partie ou à la catégorie.

	sp+	sp-		sp+	sp-
Formes	4	2	Segments	0	0

En clair, dans un fragment de document de 1992 (vingtième catégorie de la variable discriminante « source ») le MÉQ mentionnait 4 mots (formes) et 2 segments répétés de façon anormalement élevée par rapport à leur distribution dans l'ensemble des textes. Il s'agissait donc de spécificités (spéc.) et plus particulièrement de spécificités positives (sp+). Par contre, le document ne contenait pas de mots ou d'énoncés anormalement sous-distribués (sp-) par rapport à l'ensemble du corpus. Enfin, aucune forme ni aucun segment n'était surreprésenté (ou sous-représenté) au point qu'il méritait d'être reconnu comme énoncé original propre et exclusif au fragment documentaire (orig.).

Deuxièmement, les logiciels de statistique textuelle, en se basant sur l'association entre la structure de distribution du discours dans l'ensemble du corpus ou dans un sous-corpus représentant l'ensemble des énoncés d'une catégorie, permettent de repérer les éléments conceptuels qui sont réellement propres et communs à cette catégorie. Pour ce faire, le logiciel se base sur une métrique particulière dite de distance au khi carré³. Une façon simple de décrire la procédure consiste à considérer que la distance au khi carré permet de vérifier si des structures lexicales sont tellement et systématiquement récurrentes à l'intérieur d'un sous-corpus qu'elles peuvent caractériser le « bagage commun » des sujets qui y ont contribué. Pour illustrer ce qui précède, prenons les résultats d'analyse du discours (verbatim d'entrevues semi-structurées) de vingt enseignantes chevronnées du primaire. La question portait sur la définition de ce qui caractérise l'enseignante compétente (tableau 2).

³ Pour une entrée en matière, le lecteur se référera à Marchand (1998). Pour une introduction plus costaud à ce qui caractérise la statistique textuelle, il pourra consulter Lebart et Salem (1994).

Tableau 2

Recherche des structures d'énoncé caractéristiques (CRSH 1995-1998⁴)

Analyse lexicométrique du discours des enseignantes (SPAD-T)

Seuil de probabilité	Énoncé caractéristique
$p < 0,021$ (seuil critère : 0,05)	Une enseignante compétente, c'est quelqu'un qui a intégré plusieurs choses, dans le sens qu'elle est capable d'établir une relation avec chacun des enfants. C'est quelqu'un qui connaît bien son programme, qui a une pédagogie intéressante, accrochante pour les enfants. C'est une enseignante qui permet aux enfants d'apprendre, qui les soutient là-dedans.

Dans ce cas, le discours des sujets de la catégorie « enseignantes » fournissait suffisamment d'éléments communs pour que nous puissions reconnaître une structure qui le caractérise. Par contre, le fait que le logiciel ne puisse trouver suffisamment de similitude dans le discours d'une tierce catégorie évite au chercheur d'affirmer que « *les gens de cette catégorie pensent que...* ». Pour illustrer ce qui précède, dans la même recherche, lorsque nous avons procédé à l'analyse du discours du sous-échantillon « professeurs d'université » sur le même objet, si certaines structures discursives étaient relativement récurrentes, aucune n'atteignait le seuil critère statistique⁵ minimal permettant d'en affirmer la représentativité.

⁴ Recherche subventionnée CRSH (1995-1998) : *Compétences didactiques et formation didactique des enseignantes et des enseignants du primaire*. Chercheur principal : Yves Lenoir ; cochercheurs : Diane Biron, François Larose et Carlo Spallanzani.

⁵ Un seuil critère en statistique correspond à la probabilité maximale acceptable de risque lié au fait d'affirmer qu'il existe une relation entre deux variables, deux catégories ou deux entités alors que ces dernières seraient en réalité indépendantes l'une de l'autre. Par convention on considère qu'une probabilité de 5 % d'erreur de ce type correspond au seuil maximal acceptable (notation standard : $p < 0,05$).

Tableau 3

Recherche des structures d'énoncé caractéristiques (CRSH 1995-1998)	
Analyse lexicométrique du discours des professeurs (SPAD-T)	
Seuil de probabilité	Énoncé caractéristique
<p>$p < 0,063$ (seuil critère : 0,05)</p>	<p>Dans un contexte primaire, c'est sa connaissance de la matière à enseigner qui est importante, sa capacité de comprendre la psychologie de l'enfant, son développement, son organisation mentale. L'enseignant compétent, c'est celui qui est capable de s'adapter à différents types de personnalité, à différents types d'élèves qui ont des expériences personnelles qui peuvent faire divergence. Alors tu n'as plus d'homogénéité dans un groupe-classe et un enseignant compétent, à mon avis, c'est celui qui est capable de jouer avec un registre incroyable de rapports de relation, d'établir des communications et de saisir le niveau de fonctionnement de son élève.</p>

Enfin, la majeure partie des logiciels lexicométriques offrent la possibilité de procéder à des analyses descriptives, à l'aide de modèles statistiques multidimensionnels puissants de type Analyse factorielle des correspondances (AFC) ou Analyse des correspondances multiples (ACM). Faute d'espace et puisque la statistique appliquée n'est pas l'objet de cet article, nous ne décrivons pas ces procédures. Il suffit de dire qu'elles permettent de synthétiser un volume d'informations plus qu'impressionnant à l'intérieur de plans factoriels (représentations graphiques dans un espace à deux dimensions) fort utiles lorsqu'il s'agit de déterminer la part de discours partagé par une population ou une population échantillon et, dans un même tableau, ce qui est distinct et spécifique à divers sous-groupes ou catégories d'un échantillon⁶.

**LOGICIELS D'ANALYSE
QUALITATIVE
OU LOGICIELS
LEXICOMÉTRIQUES :
PRINCIPALES
DIFFÉRENCES**

Si deux catégories de logiciels utilisent la même base d'information, soit la chaîne de caractères, donc le mot ou la chaîne de blocs de caractères (l'énoncé), qu'est-ce qui les distingue ? Pour répondre à cette question, il nous faut revenir aux fondements de ce qui définit « l'objectivité », certes toute relative, qui qualifie une recherche exploratoire ou descriptive. La majeure partie des logiciels dits d'analyse qualitative organisent l'information, comme nous l'avons dit, à partir de marqueurs. Une fois enregistrée la base de données textuelles, le chercheur repère des indices qu'il marque et qu'il associe à des métamarqueurs qui correspondent à des concepts eux-mêmes définis selon des critères qui lui sont

⁶ Le lecteur qui voudra prendre contact avec le type d'utilisation qui peut être faite de ces modèles statistiques en analyse du discours ou en analyse documentaire en éducation pourra consulter les sources suivantes : Larose, Jonnaert et Lenoir (1996) ; Larose, David, Dirand, Lenoir et Roy (1999).

propres. Au fil du déroulement du texte, l'utilisateur code des mots ou des séquences de mots qui, pour le programme, correspondront à des indices de classement fréquentiels. Autrement dit, le logiciel fait de la mathématique primaire, du décompte. Bien entendu, il le fait en créant un premier tableau de fréquences dont les catégories de classement correspondent aux indices choisis par le chercheur.

Le problème de base que pose la démarche « imposée » par la structure du logiciel est le suivant. Tout mot est contextualisé dans la phrase, et tout énoncé l'est à son tour dans la pragmatique propre à la situation de communication analysée. Les mots étant des entités polysémiques par essence, les énoncés le sont aussi. Seul le contexte de la communication fixe le sens de l'indice. Or si l'indice est déterminé par le chercheur, sa stabilité est potentiellement influencée par plusieurs facteurs tels l'habitude à la récurrence conceptuelle, la surcharge cognitive liée au volume du texte, etc. Comme les « nœuds » que repère un logiciel tel NUD*IST sont directement fonction de la structure du tableau de fréquence créé au moment du codage, ceux-ci, qui sont supposés déterminer les éléments structuraux stables, communs du texte, risquent fort d'avoir une valeur directement proportionnelle aux biais que le chercheur a induits au fil des glissements d'attribution des indices et, donc, de stabilité des variables.

Les logiciels lexicométriques, pour leur part, ont pour base de référence la structure réelle du texte, indépendamment du traitement ultérieur qu'en fera le chercheur. Cela étant, ils éliminent d'emblée une source de biais interprétatif lors de l'attribution de sens aux données ou, si l'on préfère, lors de la détermination des concepts ou des structures conceptuelles qui caractérisent telle ou telle catégorie de « producteurs de discours ». Est-ce à dire que le second type de logiciel est plus « scientifique » que le premier ? Certes non. Un logiciel est un logiciel. Néanmoins, si la tâche d'un chercheur est de démontrer que des concepts existent, par delà les fréquences de mots utilisés, au sein d'une catégorie déterminée d'individus, et, si cette tâche implique de démontrer qu'une structure conceptuelle est stable et récurrente dans cette catégorie, alors la réduction des sources de biais lors de la qualification du discours est un objectif qu'aucun chercheur digne de ce nom ne peut se permettre de traiter à la légère. Les prémisses qui fondent la structure de certains logiciels le permettent plus facilement que d'autres.

**QUANTITATIF OU
QUALITATIF ?
UN FAUX PROBLÈME PAR
RAPPORT AU RESPECT
DES PRODUCTIONS DE
SES SOURCES**

En matière d'outillage de soutien informatique à la recherche, la qualification d'un logiciel en fonction de prémisses épistémologiques relève du non-sens. Les logiciels qualitatifs n'existent pas plus que les logiciels quantitatifs lorsqu'il s'agit de programmes destinés au traitement de données textuelles. Cela étant, que la posture épistémologique d'un chercheur l'amène à formuler des postulats *a priori* quant aux construits qui caractérisent une population échantillon ou qu'il préfère croire qu'il est utile d'engendrer un univers théorique par recherche à partir du discours de ses sujets, un élément demeure. L'éthique impose dans le cas d'une recherche descriptive, exploratoire, lorsque celle-ci utilise essentiellement les productions discursives des sujets formant un échantillon, on fasse un effort pour respecter les structures conceptuelles organisant le discours, en ce qu'elles ont d'individuel ou de partagé, de collectif. Le choix d'un logiciel n'y suffit pas. Encore faut-il se poser la question du rôle du discours des sujets dans la construction ou

dans la justification de postulats ou de schèmes structurant un processus de théorisation. Si ce rôle en est un de justification des postulats ou hypothèses, formelles ou non, du chercheur, il suffit alors de se livrer au jeu de piste qui permettra de trouver la bonne citation, le bon fragment illustrant notre propos. Si au contraire le rôle du discours est celui de toute base de données, en l'occurrence d'être une collection d'informations à organiser, certains éléments caractérisant le collectif, d'autres relevant des caractéristiques individuelles, alors l'approche se doit d'être distincte. Au-delà des méthodes qualitatives et quantitatives, Krathwohl (1998) souligne l'importance de combiner différentes méthodes afin de mieux « attaquer un problème de recherche » (p. 618). Il insiste également sur l'importance de la créativité du chercheur dans la combinaison de divers éléments méthodologiques, de façon cohérente et organisée, afin de mieux répondre à une question de recherche. Par delà la querelle des méthodes donc, et par delà l'énoncé des qualités intrinsèques qu'un logiciel ne peut avoir, se tient donc le critère de vérité : l'honnêteté intellectuelle du chercheur.

RÉFÉRENCES

- Barry, C.A. (1998). Choosing Qualitative Data Analysis Software : Atlas/ti and Nudist Compared. *Sociological Research Online*, 3(3). Document téléaccessible à l'URL : <http://www.socresonline.org.uk/socresonline/3/3/4.html>
- Carney, J.H., Joiner, J.F., et Tragou, H. (1997). Categorizing, Coding, and Manipulating Qualitative Data Using the WordPerfect®. *The Qualitative Report*, 3(1). Document téléaccessible à l'URL : <http://www.nova.edu/ssss/QR/QR3-1/carney.html>
- Karsenti, T., et Savoie-Zajc, L. (2000). *Introduction à la recherche en éducation*. Sherbrooke : Éditions du CRP.
- Krathwohl, D.R. (1998). *Methods of educational and social science research : An integrated approach* (2^e éd.). New York (NY) : Addison Wesley Longman.
- Larose, F., David, R., Dirand, J.M., Lenoir, Y., et Roy, G.-R. (1999). *Rapport de recherche portant sur le profil d'utilisation des TIC en pédagogie universitaire à Sherbrooke*. Sherbrooke : Université de Sherbrooke, Vice-rectorat à l'enseignement. Document téléaccessible à l'URL : <http://www.usherb.ca/PP/documents/tic99/>
- Larose, F., Jonnaert, P., et Lenoir, Y. (1996). Le construit de didactique : une étude lexicométrique illustrative d'un corpus de définitions. *Éduquer et former*, 8, 28-44.
- Lebart, L., et Salem, A. (1994). *Statistique textuelle*. Paris : Dunod.
- Marchand, P. (1998). *L'analyse du discours assistée par ordinateur*. Paris : Armand Colin.