Comparaison d'un mode de sélection par le chercheur et d'un mode d'extraction automatisée de données textuelles

Nicole Landry[*]
Shenaz Bhanji-Pitman
Réjean Auger

Dans l'ensemble des procédures spécifiques liées à l'instrumentation en analyse de textes, il est proposé de centrer l'attention sur les tâches élémentaires d'extraction des unités textuelles, de repérage des mots signifiants, et de formation des classes. Pour ce faire, l'étude d'un corpus thématique constitué de deux textes, Patton (2002, chap. 9), Silverman (1997) est présentée. L'étude compare l'extraction et la classification par le chercheur de données textuelles et l'extraction, la classification automatisée de données textuelles à partir d'un traitement statistique effectué par le logiciel Alceste. Cette étude de cas unique est abordée sous l'angle de l'efficience et de la parcimonie dans l'extraction des données textuelles qui sont considérées comme d'importants critères liés au contrôle de la qualité des produits de recherche. Les résultats obtenus suggèrent que le chercheur serait plus parcimonieux que le logiciel Alceste dans l'extraction des unités contextuelles élémentaires; que le logiciel Alceste est susceptible de favoriser une collecte parcimonieuse des données, du moins dans la tâche de repérage et de classification des mots signifiants du corpus. De manière plus globale, Alceste pourrait permettre d'augmenter l'efficience dans la collecte et la classification de données textuelles en vue de leur catégorisation par le chercheur. Suite à l'étude comparative des deux modes de collecte des données, on se questionne au sujet de l'intégration de ceux-ci à l'intérieur d'une démarche générale d'analyse de contenu. Cette intégration serait-elle souhaitable, et si c'est le cas, dans quelles limites et selon quel processus?

* Pour la rédaction de cet article, l'auteure bénéficie d'une bourse de perfectionnement de courte durée des chargées et chargés de cours de l'Université du Québec à Montréal (SCCUQ).

RECHERCHES QUALITATIVES – Hors Série – numéro 2
Actes du colloque L'INSTRUMENTATION DANS LA COLLECTE DES DONNÉES, UQTR,26 novembre 2004

ISSN 1715-8702

INTRODUCTION

Depuis le début du XX^{e} siècle, l'usage et la variété des techniques d'analyse de contenu n'ont fait que s'accroître (Neuendorf, 2002; Robert et Bouillaguet, 1997). Un essor considérable s'est produit en ce qui concerne spécifiquement l'analyse de documents écrits à l'aide d'une variété de techniques et d'instruments, dont l'analyse de texte assisté par ordinateur (Krippendorff, 2004).

Dans un sens large, l'*analyse de contenu* réfère à un ensemble de techniques de recherche utilisées pour décrire et analyser de manière systématique différents types de contenu (Vogt, 1999). L'analyse de contenu peut ainsi être effectuée sur n'importe quel matériel symbolique, qu'il soit visuel, acoustique ou autre (Roberts, 1997; Vogt, 1999). De manière plus restreinte, l'*analyse de contenu textuel* ou *analyse de textes* s'applique seulement aux textes et aux transcriptions (Roberts, 1997). Par ailleurs, *l'analyse du discours* réfère à l'analyse des éléments discursifs des textes comme entre autres les pronoms personnels, les locutions adverbiales, les adverbes de temps et d'espace.

Dans la littérature spécialisée, l'analyse de contenu et l'analyse du discours ont souvent été mises en opposition. Selon Duchastel (2004), plusieurs disciplines ou champs disciplinaires ont contribué au développement de l'analyse de contenu soient notamment la sociologie, la psychosociologie, la communication, les sciences politiques, l'histoire, le droit, l'éducation. Par contre, les sciences du langage, les études littéraires, l'anthropologie ou l'ethnologie, s'intéressaient davantage à l'analyse du discours. Par ailleurs, l'analyse de texte assisté par l'ordinateur offre la possibilité d'analyser à la fois le contenu informationnel et le contenu discursif d'un ensemble de textes. Du point de vue de Duchastel (2004), l'analyse de texte assistée par l'ordinateur, dans la perspective des sciences cognitives, aurait une orientation psychologique davantage liée à l'analyse de contenu, et une orientation linguistique davantage liée à l'analyse du discours.

Il convient d'indiquer que la présente étude d'analyse de texte assistée par l'ordinateur relève de l'analyse du contenu informationnel d'un corpus de textes.

PROBLÉMATIQUE

Des études ont comparé l'analyse de contenu effectuée par l'ordinateur avec l'analyse de contenu réalisée par l'humain (Nacos et coll., 1991; Schnurr, Rosenberg, et Oxman, 1992, 1993; Zeldow et McAdams, 1993). De ces études, principalement celle de Nacos et coll. (1991), on peut conclure à l'intérêt d'utiliser l'ordinateur de manière complémentaire à l'analyse humaine de

textes plutôt que d'appréhender l'utilisation de l'ordinateur dans une optique de remplacement de l'un par l'autre (Krippendorff, 2004).

La question est alors soulevée à savoir quelles tâches à l'intérieur d'une démarche d'analyse de textes seraient mieux servies par l'ordinateur, et quelles autres le seraient davantage par l'humain (Krippendorff, 2004). Les connaissances actuelles à cet effet sont limitées, notamment parce que l'on ne connaît pas suffisamment bien les opérations liées à l'analyse de documents écrits par le chercheur. Comme le soulignent Bertrand-Gastaldy, Daoust, Meunier, Pagola, et Paquin (1993) en se référant à David (1990), Engres-Niggemeyer (1990) et Farrow (1991): «l'étude cognitive des opérations d'analyse documentaire ne bénéficie pas d'une longue tradition». Néanmoins, il est généralement admis que les tâches qui nécessitent de plus grandes habiletés cognitives, celles liées notamment aux opérations de catégorisation et de plus haut niveau d'inférences cognitives, seraient avantageusement réservées à l'humain. Ces tâches viendraient compléter les opérations de tri et de classification exécutées efficacement par l'ordinateur. Une telle combinaison de modes de traitement des données serait d'autant plus valable dans la situation où le besoin du traitement d'une grande quantité de données textuelles est présent (Krippendorff, 2004).

Bref, dans une optique de contrôle de la qualité des produits de recherche (Auger, 2003, 2000; Lussier et Auger, 1997), un intérêt particulier doit être accordé aux procédures spécifiques liées à l'instrumentation en analyse combinée de textes. Plus précisément et à la lumière de la problématique exposée, il est proposé de centrer l'attention sur des procédures élémentaires, l'extraction des unités textuelles, le repérage des mots signifiants, et la formation des classes constituée de regroupements sémantiques.

À partir de ces objets, l'objectif de l'étude est d'apprécier la plus ou moins grande efficience et parcimonie de deux modes d'extraction et de classification de l'information, soit l'une effectuée par le chercheur à l'aide d'une base de données GRÉ et l'autre effectuée par extraction automatisée à l'aide du logiciel Alceste.

Pour ce faire, un corpus thématique est constitué de deux textes, à savoir le texte de Patton (2002, chap. 9) et le texte de Silverman (1997, dans Miller et Dingwall, chap. 1). L'étude compare (1) l'extraction et la classification par le chercheur de données textuelles (2) l'extraction et la classification automatisée de données textuelles à partir d'un traitement statistique effectué par le logiciel Alceste. Dans la présente étude, un *corpus* réfère à un ensemble de textes soumis à l'analyse systématique (Legendre, 1993). Le thème du corpus est le suivant : *La qualité et la crédibilité de la recherche.*

Brièvement, dans une approche traditionnelle d'analyse de textes, d'une part, le chercheur commente le corpus et en fait l'interprétation. D'autre part, dans une analyse assistée par l'ordinateur à l'aide du logiciel Alceste, un traitement automatisé des données textuelles est effectué qui comprend une

réorganisation du texte, un traitement statistique et une représentation graphique. L'analyse est assistée par des dictionnaires intégrés modifiables et par la présence de clés catégorielles paramétrables.

INSTRUMENTATION

Dans la présente étude, le traitement de données textuelles par le chercheur est facilité par l'utilisation d'une base de données électronique, à savoir la base de données GRÉ[1] de Auger et Landry (2003). La figure 1 suivante montre un exemple d'une fiche dans la base de données GRÉ.

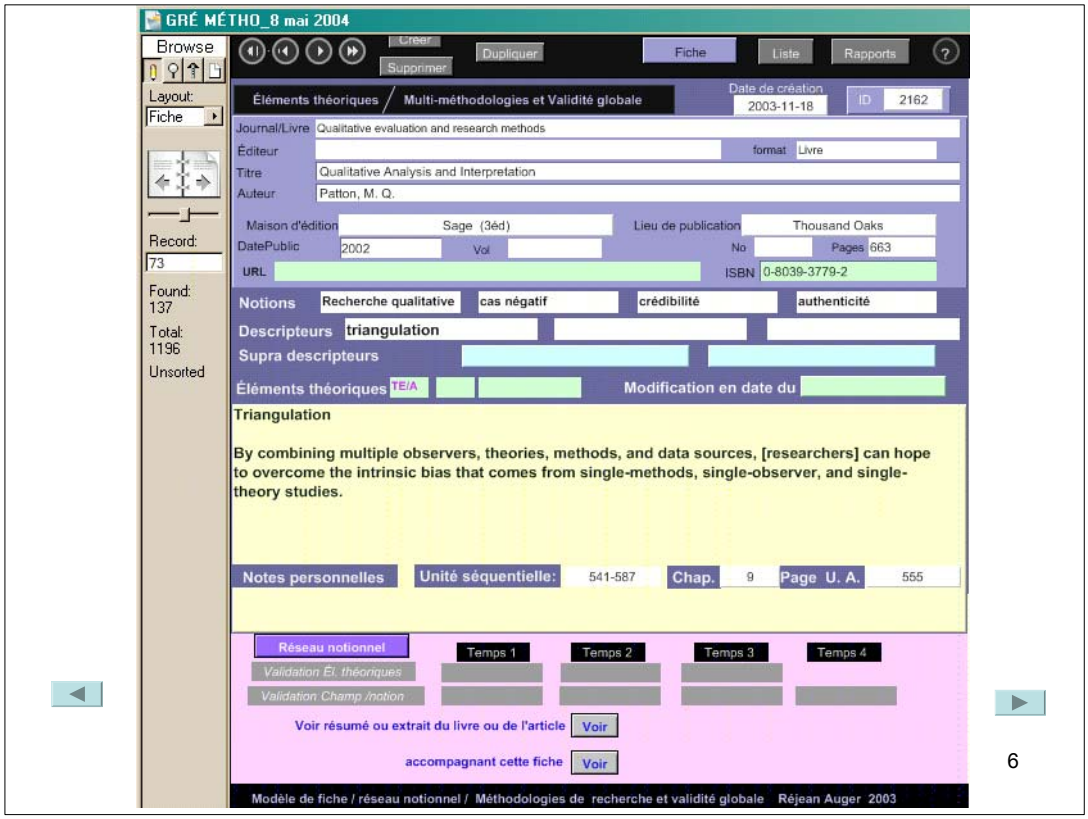

Figure 1. Exemple d'une fiche dans la base de données GRÉ

La fiche se divise en quatre champs distincts ayant chacun des fonctions spécifiques. Le champ supérieur de la fiche permet l'inscription de l'information bibliographique. Juste en dessous, un autre champ permet la catégorisation des UCE par éléments théoriques (TA, TF, TP ou TE); par notions, descripteurs et supra descripteurs. Le champ central de la fiche est

réservé à la saisie de l'unité contextuelle élémentaire (UCE) ou extrait de texte, ainsi qu'à l'inscription de commentaires. Enfin, le dernier champ permet le suivi de la procédure de validation.

Ainsi à l'aide de la base de données GRÉ et suite à la numérisation du texte, le chercheur sélectionne des extraits textuels par repérage visuel et découpage du texte en unités contextuelles élémentaires (UCE). Ensuite, ce même chercheur catégorise les extraits de texte en fonction d'éléments théoriques axiologique (TA), formel (TF), praxéologique (TP) ou explicatif (TP). Ces extraits sont également catégorisés en fonction de notions tenant compte de l'expression explicite dans le texte et de descripteurs de premier niveau d'inférence. Par la suite, le chercheur ajoute à la catégorisation un deuxième niveau d'inférence par l'entremise de supra descripteurs. Un processus de validation interne s'enclenche afin d'assurer des liens de cohérence à la catégorisation multidimensionnelle proposée. Enfin, il est toujours possible de visualiser l'ensemble de l'information disponible par des extraits de l'article ou du livre, des figures, des images, des notions ou par auteurs et dates ou par des éléments de synchronie, diachronie.

En ce qui concerne le traitement automatisé de données textuelles à l'aide du logiciel Alceste, l'objectif poursuivi est d'extraire du corpus les structures signifiantes les plus fortes. Pour ce faire, deux conditions doivent être satisfaites:

1. le corpus se présente comme un tout ayant une certaine cohérence, par exemple un corpus autour d'une thématique;
2. le document est suffisamment volumineux pour que l'élément statistique entre en ligne de compte.

Par ailleurs, il est essentiel de préparer le texte en vue du traitement automatisé. Cette préparation consiste notamment à l'inscription d'une ligne étoilée qui identifie les principales variables contextuelles: auteur, date, chapitre (**** *aut_Patton *dat_2002 ou **** *aut_Silverman *dat_1997). Le texte soumis à l'analyse est sauvegardé en format texte (*.txt).

Par la suite, quatre grandes étapes sont effectuées par le logiciel Alceste; il s'agit de:

a) la lecture des textes et du calcul des dictionnaires (dictionnaires de référence et ad hoc; attribution des clés catégorielles);
b) la définition des UCE et de leurs classifications (découpage en segments de textes; double classification; classification hiérarchique descendante) par une procédure statistique;
c) de la définition et description des classes, analyse factorielle des correspondances (représentation des classes; liste des mots pleins associés aux classes);
d) de calculs complémentaires permettant de nuancer certaines analyses.

DESCRIPTION DES ANALYSES

Précisons d'abord que le corpus se divise en deux unités de contexte initiales (UCI) à savoir l'ensemble des unités en provenance du texte de Patton (2002) et l'ensemble des unités en provenance du texte de Silverman (1997). Ces deux UCI ont en commun une même thématique soit « La qualité et la crédibilité de la recherche ».

Deux éléments de comparaison sont retenus: les extraits de textes faisant référence aux UCE et les mots, proprement dits, qui composent le corpus. La figure 2 présente dans la partie supérieure l'UCE extraite par le logiciel Alceste, et dans la partie inférieure l'UCE correspondante extraite par le chercheur. L'UCE présentée appartient à la classe 1 et est libellée [diversité des sources]. Les formes précédées d'un dièse sont les mots pleins distinctifs de la classe.

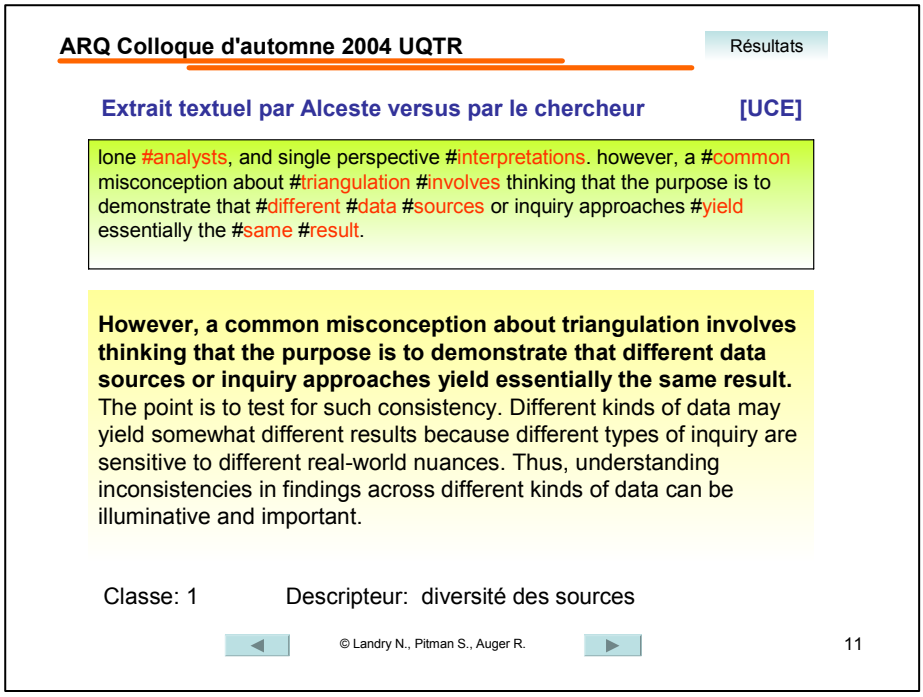

Figure 2. Exemple d'une unité de contexte élémentaire (UCE)

On note ici une convergence dans l'extraction de l'UCE entre le logiciel et le chercheur, à savoir : la présence du même contenu informationnel. La différence consiste en un extrait plus long de la part du chercheur. En effet, le logiciel Alceste procède à un découpage systématique des segments de texte qui sont approximativement de la même longueur; longueur déterminée à

l'étape du paramétrage du plan d'analyse. Le chercheur, quant à lui, prend habituellement toute la phrase et très souvent il retient un paragraphe entier. En conséquence, cette façon de faire entraîne une variation dans la longueur des unités de texte extraites par le chercheur, ce qui n'est pas le cas du logiciel Alceste. En dépit de cette différence, on observe que dans la plupart des cas, le contenu informationnel catégorisé par le chercheur et par Alceste s'avère être équivalent.

Le tableau 1 présente les statistiques descriptives de l'analyse textuelle par Alceste, tandis que le tableau 2 présente les statistiques descriptives de l'analyse textuelle par le chercheur. Le tableau 1 présente des éléments de comparaison liés aux UCE et aux mots pleins. En ce qui concerne les UCE, un total de 417 unités de textes extraites par le logiciel Alceste sur une possibilité de 729. En d'autres mots Alceste a retenu 57% des UCE du corpus, pour fin de classification. De plus, Alceste a distribué les 417 UCE retenues dans 4 classes, de la manière suivante: 17% des UCE dans la classe 1; 29% des UCE dans la classe 2; 15% des UCE dans la classe 3; 39% des UCE dans la classe 4. Ces valeurs correspondent au pourcentage relatif d'UCE.

Tableau 1
Statistiques descriptives de l'analyse textuelle par ALCESTE

Éléments de comparaison	**Classe 1**	**Classe 2**	**Classe 3**	**Classe 4**	**Total**	**Total absolu**
UCI	Patton (1-563)		Silverman (564-729)		2	2
UCE retenues	72	119	63	163	417 57 %	729
% relatif d'uce	17,28 %	28,54 %	15,10%	39,09 %		
% absolu d'uce	9,88 %	16,32%	8,64 %	22,36 %		
Mots pleins retenus	97	181	133	151	566 (66%)	860
Mots pleins significatifs	38	45	32	71	186 (33%)	
Mots pleins significatifs / Nombre de formes distinctes dans el corpus					186 4 %	4353
Temps d'exécution : 2 minutes 13 secondes pour 4 353 formes distinctes						

En ce qui concerne les mots retenus, Alceste a retenu 566 mots pleins sur une possibilité de 860, ce qui constitue 66% de l'ensemble des mots pleins du corpus. Les «mots pleins» sont en général des noms, des verbes, des adjectifs. Les mots pleins se distinguent des mots outils ou mots fonctions qui sont par exemple des pronoms, des adverbes ou des locutions adverbiales et qui sont davantage utilisés dans l'analyse du discours proprement dit. Enfin, sur les 566 mots pleins retenus Alceste a conservé 186 mots pleins significatifs, ceci

correspond à 33% du total des mots pleins retenus. Dans Alceste, pour qu'un mot plein soit significatif, il doit rencontrer deux conditions: se retrouver dans une classe stable et avoir un χ^2 plus grand que 0. Finalement, comme il y a 4 353 formes distinctes qui composent le corpus, ces 186 mots pleins significatifs comptent pour 4%. Ce pourcentage serait suffisant à mettre en évidence l'essentiel du contenu informationnel du corpus. Le temps d'exécution du traitement automatisé des données est de 2 minutes 13 secondes.

En ce qui concerne les UCE retenues par le chercheur, le tableau 2 montre que le chercheur a retenu 304 UCE sur une possibilité de 729. Ceci représente 42% des UCE potentielles. Le chercheur a ainsi retenu plus d'unités de textes que le logiciel, lequel a retenu 57% des UCE potentielles. Les 304 UCE retenues ont été distribuées dans 5 classes à raison de: 23% des UCE dans la classe 1; 29% des UCE dans la classe 2; 0.007% des UCE dans la classe 3; 34% des UCE dans la classe 4; 13% des UCE dans la classe 5. Ces valeurs correspondent à des pourcentages relatifs.

En ce qui concerne les mots retenus par le chercheur, on constate que sur les 4 353 formes distinctes qui composent le corpus, le chercheur a retenu 1091 mots pleins. Ceci constitue 25% des formes distinctes du corpus. Le temps d'exécution du traitement des données par le chercheur est de 135 heures.

Tableau 2
Statistiques descriptives de l'analyse textuelle par le chercheur

Éléments de comparaison	**Classe 1**	**Classe 2**	**Classe 3**	**Classe 4**	**Classe 5**	**Total**	**Total absolu**
UCI	Patton (1-255)		Silverman (256-304)			2	2
UCE retenues	70	89	2	104	40	304	729
						42 %	
% relatif d'uce	23,03 %	29,28 %	0,007%	34,21 %	13,16%		
% absolu d'uce	9,60 %	12,21%	0,0002 %	14,27 %	5,49%		
Mots pleins retenus	246	307	8	389	141	1091	4353
Mots pleins retenus / nombre de formes distinctes dans el corpus						(25 %)	

Temps d'exécution : 135 heures pour 4 353 formes distinctes

COMPARAISON DE L'EFFICIENCE ET DE LA PARCIMONIE DES MODES D'ANALYSE

Dans la présente étude, la comparaison du mode de sélection et de traitement par le chercheur et du mode d'extraction et de traitement automatisé des données textuelles relève de deux critères : l'efficience et la parcimonie. Dans des ouvrages de référence spécialisés (Keeves, 1997; De Lansheere, 1992)

l'efficience réfère à la relation entre deux facteurs s'exprimant sous la forme de ratio entre deux mesures. La parcimonie, quant à elle, réfère à la diminution des unités d'information tout en conservant toute l'information disponible par son caractère englobant et essentiel. Le recours au principe parcimonie vise ainsi à réduire la complexité du réel.

Brièvement et de manière opérationnelle, dans le cadre de la présente étude, *l'efficience absolue* réfère au rapport entre le nombre de formes distinctes du corpus intégral sur le nombre de classes et la durée du traitement en secondes. Quant à *l'efficience relative, elle* est définie en termes de ratio entre la mesure de l'efficience du logiciel Alceste et la mesure de l'efficience du chercheur. Comme le montre le tableau 3, la valeur d'efficience absolue pour le logiciel Alceste est de 8,18, tandis que la valeur d'efficience absolue pour le chercheur est de 0,11. La valeur d'efficience relative est donc de 74,36.

Tableau 3
Mesures d'efficience absolue et relative

Efficience (MOTS)	**Alceste**	**Chercheur**
absolue	(4 353 formes/4 classes*133 sec.)= **8,18**	(4 353 formes/ 5 classes*8100 sec.)=**0,11**
relative	**8,18/0,11=734,36**	

Ainsi, sur la base d'éléments de comparaison liés aux mots qui composent le corpus, on peut affirmer que le logiciel Alceste est relativement plus efficient que le chercheur à raison d'un ratio de 74,36 ou, en d'autres mots Alceste est 74 fois supérieur au chercheur dans la collecte et le traitement automatisé de données textuelles en ce qui concerne la sélection et la classification des mots.

Le tableau 4 suivant concerne les indices de parcimonie pour les UCE et les mots selon une mesure absolue. Au niveau des extraits de texte, la mesure absolue de l'indice de parcimonie équivaut au total des UCE retenues sur le total des UCE potentielles du corpus, multiplié par 100. Au niveau des mots du corpus, la mesure absolue de l'indice de parcimonie équivaut au total de mots pleins significatifs sur le grand total des formes distinctives du corpus, multiplié par 100. Pour ce qui est de l'interprétation de l'indice de parcimonie, plus la valeur est petite, plus grande est la parcimonie.

Tableau 4
Indices de parcimonie

Parcimonie mesure absolue	Alceste		Chercheur	
UCE	**(417 / 729) *100 =**	**57 %**	**(304 / 729) * 100 =**	**42%**
MOTS	**(186/ 4353) *100 =**	**4 %**	**(1091/ 4353) *100 =**	**25%**

Basé sur les mots, on peut dire que le logiciel Alceste s'est avéré plus parcimonieux que le chercheur. Très précisément le logiciel Alceste présente un indice de parcimonie de 4% tandis que le chercheur présente un indice de 25%. Ensuite, en se basant sur les UCE, on peut dire que le logiciel Alceste s'est avéré moins parcimonieux que le chercheur. En effet, le logiciel Alceste présente un indice de parcimonie de 57% tandis que le chercheur présente un indice de 42%.

Cette comparaison ne serait complète sans un regard attentif sur la classification et la catégorisation des UCE. Rappelons d'abord que le chercheur a réparti les UCE retenues dans plus de classes que le logiciel Alceste. Alceste a distribué 417 UCE dans 4 classes, tandis que les 304 UCE retenues par le chercheur ont été distribuées dans 5 classes. Le tableau 5 présente la distribution des UCE par Alceste et par le chercheur en fonction de chacune des classes formées.

Tableau 5
Nombre d'UCE réparti par classes selon Alceste et le chercheur

UCE retenues	Classe 1	Classe 2	Classe 3	Classe 4	Classe 5
Alceste (417 uce)	72	119	63	163	……
Chercheur (304 uce)	70	89	2	104	41

Au tableau 5, l'on observe que le nombre d'UCE réparti dans la classe 3 se distingue nettement par une disproportion très marquée. À la lecture des résultats obtenus, on peut se questionner sur la pertinence de la cinquième classe formée par le chercheur ou par l'absence d'une cinquième classe par le logiciel. Pour mieux comprendre le cas de la classe 5, le tableau 6 suivant présente l'occurrence de la notion «crédibilité» en fonction de chacune des classes formées par le chercheur.

Tableau 6
Le cas de la classe 5 selon la notion de « crédibilité »

Cas de la notion de « crédibilité» selon chacune des classes identifiées par le chercheur		Classes et descripteurs				
		Diversité des sources (Classe 1)	Rigueur méthodologique (Classe 2)	L'impact du chercheur (Classe 3)	Types d'approche (Classe 4)	Crédibilité du chercheur (Classe 5)
		Fréquence	Fréquence	Fréquence	Fréquence	Fréquence
Crédibilité	N1	0	2	1	2	2
	N2	0	0	0	1	0
	N3	2	1	0	2	2
	N4	4	1	0	1	0
Total		6	4	1	6	4

D'abord, en colonne sont listées les 5 classes identifiées par le chercheur et nommées de la manière suivante: Diversité des sources (classe 1); Rigueur méthodologique (classe 2); L'impact du chercheur (classe 3); Types d'approche (Classe 4); Crédibilité du chercheur (classe 5). Ensuite, en rangée, N1, N2, N3 et N4 représentent les quatre emplacements dans la base de données GRÉ où le chercheur aurait pu inscrire la notion « crédibilité ». Ainsi, l'on retrace le nombre de fois que le chercheur a repéré la notion «crédibilité» dans les unités de texte, à raison de 6 fois dans la classe 1, 4 fois dans la classe 2, 1 fois dans la classe 3, 6 fois dans la classe 4, et 4 fois dans la classe 5.

Un regard attentif sur la classe 5, permet de constater que la notion « crédibilité » se retrouve dans chacune des classes et non pas exclusivement dans la classe 5; ce qui va à l'encontre de tout système de classification où chacune des catégories doit être mutuellement exclusive (Bailey, 1994). Par conséquence, la cinquième classe apparaît non nécessaire et redondante, si l'on considère la proximité sémantique entre les classes 3 et 5. La pertinence de la classe 5 est d'autant plus difficile à soutenir si l'on considère le fait que «la qualité et crédibilité de la recherche» est le thème général du corpus et que la création d'une classe spécifique n'ajoute rien à la compréhension du contenu informationnel de l'ensemble du corpus.

DISCUSSION ET CONCLUSION

En somme, cette étude de nature comparative montre que :

- Le logiciel Alceste est un instrument susceptible d'augmenter l'efficience dans le traitement de données textuelles (collecte et classification) lorsque le corpus est suffisamment volumineux et lié à une thématique précise.

- Le logiciel Alceste est un instrument susceptible de favoriser une collecte parcimonieuse de données textuelles, très précisément en ce qui concerne le repérage et la classification des mots pleins du corpus.
- Le chercheur pourrait s'avérer plus parcimonieux que le logiciel Alceste lorsqu'il s'agit d'extraire des UCE ou unités de texte en autant que ces UCE soient congruentes ou en lien de cohérence avec la classification proposée.

Les résultats et conclusions sont limités par l'étude du cas unique (N =1) qu'est le chercheur. Aucune généralisation n'est permise quant à l'utilisation systématique du logiciel pour fins d'extraction et de classification d'unités de texte.

Par ailleurs, à la lumière des résultats obtenus, on peut se demander si l'intégration à l'intérieur d'une démarche générale d'analyse de contenu, de ces deux modes de collectes et de traitement de données textuelles, serait souhaitable ? Et, si c'est le cas, dans quelles limites et selon quel processus ? La question est importante puisqu'elle concerne la qualité du processus de recherche dans un contexte d'analyse de documents écrits, un type d'analyse qui a pris énormément d'ampleur en recherche en sciences humaines ou sociales, notamment en éducation. Il est important d'insister sur le fait, également, que les extraits textuels constituent les données empiriques de base sur lesquelles reposent les autres étapes dans la compréhension de l'objet d'étude.

Ainsi en plus d'une recherche d'efficience et de parcimonie dans l'extraction des données textuelles, il serait nécessaire que les extraits textuels retenus soient représentatifs du texte original, et qu'ils soient adéquats en fonction du contenu informationnel livré par l'auteur du texte. Lorsque ces conditions seront atteintes, on sera en mesure d'assurer l'inférence de connaissances [2] valides et réplicables en analyse de textes.

Par ailleurs, il reste à étendre la présente étude de cas à un nombre substantiel de sujets. Nous souhaitons bien évidemment obtenir une cohorte d'une cinquantaine de chercheurs afin d'établir avec plus de stabilité les observations faites au cours de cette étude. Nous souhaitons également prendre en compte d'autres variables liées au contrôle de la qualité du processus d'extraction, de codification, de classification des unités d'information en analyse de textes comme la saturation, l'exhaustivité de l'information, et l'exclusivité catégorielle.

NOTES

[1] CD-Rom «Gestion des recensions d'écrits (GRÉ)» par Auger et Landry (2003, version bêta).

[2] En ce qui concerne l'inférence de connaissances en analyse de contenu on peut lire Bardin (1996, 1977), Ghiglione, Beauvois, Chabrol et Trognon (1980), Mucchielli (1974), ou Krippendorff (2004, 1980) pour ne nommer qu'eux.

RÉFÉRENCES

Auger, R. (2003). *Clarification conceptuelle et proposition d'opérationnalisation des critères de scientificité de la recherche en éducation; le cas de la saturation et de la complétude*. ARQ. Université du Québec à Trois-Rivières.

Auger, R. (2000). *Qualité du processus de recherche dans une approche multi-méthodologique portant sur la construction des représentations culturelles.* Communication présentée à la 5e Conférence internationale sur les représentations sociales " Constructions nouvelles ", Montréal.

Auger, R., & Landry, N. (2003). *Base de données électronique. Gestion des recensions d'écrits GRE. Version de rodage* [CD-ROM]. Montréal: LABFORM.

Auger, R., & Landry N. (2002). Recension des écrits _ Base de données (version Bêta). Dans *Le développement des représentations culturelles et leur impact sur l'apprentissage des langues vivantes.* Création d'une application autonome de base de données FILEMAKER PRO.

Bailey, K. (1994). Typologies and taxonomies: an introduction to classification *techniques*. Thousand Oaks, Sage.

Bertrand-Gastaldy, S., Daoust, F., Meunier, J.-G., Pagola, G., & Paquin, L.-C., (1993). *Les traitements statistico-linguistiques et l'enquête cognitive comme moyens de reconstituer l'expertise des spécialistes en analyse documentaire: le cas de la jurisprudence* (Cahiers de recherche 2). Montréal : Université du Québec à Montréal, Centre de recherche en Cognition et Information ATO.CI.

David, C. (1990). *Élaboration d'une méthodologie d'analyse des processus cognitifs dans l'indexation documentaire.* Mémoire de maîtrise inédit, Université de Montréal.

De Lansheere, G. (1992). *Dictionnaire de l'évaluation et de la recherche en éducation*. (2e éd.). Paris : PUF.

Duchastel, D. (2004). Introduction à l'ATO et aux débats qu'elle suscite. Formation en analyse de texte assistée par ordinateur- École d'été, organisée par la Chaire de recherche du Canada en Mondialisation, Citoyenneté et Démocratie. UQAM.

Endres-Niggemeyer, B. (2-4 Octobre 1990). A procedural model of abstracting, and some ideas for its implementation. Dans H. C. & W. Nedobity (Dir.), *TKE'90: Terminology and Knowledge Engineering; Proceedings of the Second International Congress on Terminology and Knowledge Engineering* (p. 230-243). University of Trier (FRG), Frankfurt: Indeks Verlag.

Farrow, J. (1991, Juin). A cognitive process model of document indexing. *Journal of Documentation, 47*(2), 149-166.

Keeves, J. (1997). *Educational research, methodology, and measurement: an international handbook* (Ed. John P. Keeves) (2e éd.). Cambridge: Pergamon.

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology* (2e éd.). Philadelphia: Sage.

Legendre, R. (1993). *Dictionnaire actuel de l'éducation* (2e éd.). Montréal: Guérin.

Lussier, D., & Auger, R. (1997). *L'apport de la validité globale dans le contrôle de la qualité des opérations de recherche.* Communication présentée à la Colloque du CRÉAL, Ottawa.

Nacos, B., Shapiro, R., Young, J., Fan, D., Kjellstrand, T., & McCaa, C. (1991). Content analysis of news reports: Comparing human coding and a computer-assisted method. *Communication, 12*, 111-128.

Neuendorf, K. (2002). *The content analysis guidebook*. Thousand Oaks: Sage.

Patton, M. (2002). Enhancing the quality and credibility of qualitative analysis in q*ualitative research and evaluation methods* (3e éd.). Thousand Oaks: Sage.

Robert, A., & Bouillaguet, A. (1997). *L'analyse de contenu*. Paris: Presses universitaires de France.

Roberts, C. (Dir.). (1997). *Text analysis for the social sciences: methods for drawing inferences from texts and transcripts.* Mahwah : Lawrence Erlbaum.

Schnurr, P., Rosenberg, S., & Oxman, T. (1992). Comparison of TAT and free speech techniques for eliciting source materials in computerized content analysis. *Journal of personality assessment, 58*, 311-325.

Schnurr, P., Rosenberg, S., & Oxman, T. (1993). Issues in the comparison of techniques for eliciting source material in computerized content analysis. *Journal of personality assessment, 61*, 337-342.

Silverman, D. (1997). The logics of qualitative research. Dans G.Miller & R. Dingwall (Éds.) *Context & method in qualitative research*. (pp. Début-fin). Thousand Oaks: Sage.

Vogt, P. (1999). *Dictionary of statistics & methodology. A nontechnical guide for the social sciences* (2e éd.). Thousand Oaks: Sage.

Zeldow, P., & McAdams, D. (1993). On the comparison of TAT and free

speech techniques in personality assessment. *Journal of personality assessment, 60*, 181-185.

Nicole Landry (M.Sc.). Doctorante en éducation (UQAM). Assistante de recherche et coordonnatrice des activités du LABFORM au Département d'éducation et pédagogie (DÉP) de l'UQAM. Objet de la thèse: «Classification des capacités perceptives dans une visée d'apprentissage et de développement cognitif en éducation préprimaire» avec développement méthodologique pour fins de modélisation d'éléments théoriques fondamentaux. Intérêts de recherche actuels: aspect méthodologie portant sur l'analyse de textes, à savoir l'utilisation combinée de deux modes d'analyses de contenu, par le chercheur utilisant une base de donnés électronique (GRÉ) et, à l'aide du logiciel d'analyses statistiques Alceste. Membre active de l'Association des Méthodologies d'Évaluation en Éducation (ADMEE) et responsable technique du E-Bulletin. Maîtrise ès sciences du Département de kinanthropologie de l'Université du Québec à Montréal, concentration neurocinétique. Chargée de cours au Département de kinanthropologie, UQAM, depuis neuf ans. Enseignante en éducation physique, en éducation préscolaire, éducatrice à la petite enfance et au Centre jeunesse de Montréal (1983-1997).

Shehnaz Bhanji-Pitman (M.Ed.). Doctorante en éducation (UQAM). Objet de la thèse: « Perceptions et démarches pédagogiques d'enseignants de français langue seconde au regard de la diversité culturelle des apprenants immigrants adultes ». Maîtrise en éducation du Département de la didactique des langues de l'Université McGilll, spécialisation en difficultés d'apprentissage des adultes sinophones en français langue seconde. Professeure de français langue seconde au MRCI depuis 1995 et au projet MRCI-UQAM depuis 2000. Conseillère pédagogique au MRCI (1991-995). Professeure de français langue seconde à l'Université Mémorial (1983-1991). Enseignante d'allemand et d'anglais au lycée polyvalent des îles Saint Pierre et Miquelon (1980-1983).

Réjean Auger (Ph.D.). Professeur titulaire agrégé au Département d'Éducation et Pédagogie, regroupement Mesure et Évaluation, Université du Québec à Montréal, depuis 1990. Directeur fondateur du Laboratoire de méthodologie de la recherche, d'analyse de données et de Formation en mesure et évaluation en éducation (LABFORM) depuis 1994. Membre d'un comité d'évaluation pour le CRSH (2004). Expert conseil en méthodologie de la recherche en éducation pour le Centre Européen des Langues Vivantes (GRAZ, Autriche) 2000-2003. En 1990, professeur invité à l'Université de Montréal. Diplômé de l'Université du Québec à Montréal, Ph.D. en éducation, mai 1989. Conseiller expert auprès de la DDÉ (Direction du Développement en Évaluation, MEQ) de 1986-2000.

Conseiller occasionnel auprès de diverses commissions scolaires (1983 – 1990). Quatre ans au ministère de l'Éducation (1983-86, à la DDÉ). Concepteur de la démarche de construction d'instruments de mesure à l'intérieur du projet BIM de l'actuelle GRICS. Concepteur du concept « Définition du domaine » pour les examens de la sanction des études, MEQ. Enseignant à l'ordre du primaire en éducation physique pendant 10 ans et trois ans au secondaire en mathématiques (1970-1983).