Les problèmes posés par l'utilisation d'un logiciel d'analyse des données textuelles dans une perspective d'aide à la décision.

Daniel Guy

PARIS, LES 01, 02 ET 03 JUILLET 1996

DEUXIEME CONGRES D'ACTUALITE DE LA RECHERCHE EN EDUCATION ET FORMATION

ORGANISE A L'INITIATIVE DE L'AECSE PAR L'UNIVERSITE DE PARIS X-NANTERRE

DEPARTEMENT DES SCIENCES DE L'EDUCATION

Communication présentée le 02 juillet 1996 à Paris

Les problèmes posés par l'utilisation d'un logiciel d'analyse des données textuelles dans une perspective d'aide a la décision

Communication présentée le 02 juillet 1996 à Paris

Daniel GUY

Sciences de l'éducation, Université de Toulouse Le-Mirail

Dans le cadre de l'analyse des données textuelles, notre projet est d'examiner le problème de l'articulation entre les éléments objectifs déterminés par le calcul statistique et leur commentaire. En effet, dans une perspective d'aide à la décision, l'intervenant court le risque d'une lecture trop attentive à l'interprétation des données au détriment de l'analyse des éléments objectifs qui la fondent. Cette question est d'autant plus incontournable que la seule communication des déterminations objectives dont le prix serait le *sacrifice des significations*[1], rendrait inutile le recours à l'analyse des données textuelles. Nous appuierons notre réflexion sur : 1 / une enquête relative à une "demande sociale" concernant les préoccupations des jeunes âgés de 15 à 25 ans , 2 / la conception triadique du signe de Peirce, 3 / une méthodologie d'analyse des données textuelles particulière : ALCESTE. En conclusion, nous tracerons les contours d'une stratégie de communication des résultats.

1 - L'enquête relative à une demande sociale

Durant le premier semestre 1995, notre équipe a étudié la population âgée de 16 à 25 ans d'un département du Midi de la France afin de tracer les orientations possibles de la politique "jeunesse" du Conseil Général. Effectuée auprès d'un échantillon représentatif, construit selon la méthode des quotas (N = 600),

[1] GRANGER, G.G. (1968). Essai d'une philosophie du style. Paris : Armand Colin.

l'enquête associait un questionnaire fermé à un entretien enregistré au cours duquel les jeunes étaient invités à exprimer leurs préoccupations. Cette étude a fait l'objet d'un rapport qui reprenait l'ensemble des résultats de l'enquête.

Suite à la restitution de ce document aux membres du cabinet du Président du Conseil Général, et avant sa diffusion publique, un chargé de mission a rédigé et communiqué à l'ensemble des élus une synthèse des principaux résultats. En ce qui concerne l'analyse du contenu des entretiens, cette note ne retient, pour l'essentiel, que les interprétations développées par le chercheur.

Cette observation souligne le problème posé par l'articulation entre les éléments objectifs déterminés par le calcul statistique et leur commentaire dans un rapport d'étude. En effet, les lectures ultérieures du document original risquent d'être principalement orientées par la grille de lecture que constitue la note de synthèse, donc par les seuls éléments d'interprétation mis en place par le chercheur. C'est dire que dans un dispositif d'aide à la décision, le poids des registres d'interprétation est renforcé par les processus de diffusion et d'accès à l'information.

2 - La conception triadique du signe de Peirce

La conception triadique du signe de Peirce, et en particulier la notion d'*interprétant,* permet d'analyser les deux phases du processus d'interprétation des données textuelles dans un dispositif d'aide à la décision :

- Le cheminement méthodologique qui conduit le chercheur depuis la détermination statistique des éléments objectifs du contenu des données textuelles jusqu'à leur interprétation.
- Le passage de la mise en oeuvre par le chercheur de moyens sémiotiques pour communiquer les résultats de l'analyse de contenu à leur interprétation par la *praxis* des acteurs.

Mais, rappelons d'abord que le triangle sémiotique de Peirce est une relation triadique entre un objet, un signe et un *interprétant*. Gilles-Gaston Granger emprunte à Peirce l'exemple de l'homme ivre : "*on présente un homme ivre pour montrer, par contraste, l'excellence de la tempérance*." L'objet est ici la valeur de la sobriété, le signe est l'homme ivre et l'interprétant est constitué par l'ensemble des représentations de l'ivrognerie qui peuvent être enchaînées dans une suite indéfinie et qui renvoient toutes à l'objet. L'*interprétant* ne renvoie donc pas à un

individu singulier. Pour Marty[2], au contraire, c'est "*une norme sociale ou un habitus collectif déjà-là et la détermination ici et maintenant d'un esprit qui intériorise cette norme*". Dans cette conception, Ricoeur note l'intérêt du "rapport ouvert" de signe à interprétant, de telle manière qu'un autre *interprétant* peut toujours médiatiser le premier rapport, d'où le caractère indéfini de la série des *interprétants*. Dans le cas du dispositif d'aide à la décision que nous observons, l'origine de la chaîne interprétative est le rapport établi par le chercheur entre les données objectives de l'analyse de contenu et les interprétations qu'il a lui-même proposées.

La conception triadique du signe de Peirce nous autorise à concevoir le processus d'interprétation d'une étude comme une relation triadique entre :

- le rapport d'étude conçu comme un système informationnel,
- les déterminations objectives des situations analysées,
- et les systèmes de significations qui donnent un sens aux résultats.

Les systèmes de significations jouant ici un rôle identique à celui des *interprétants* dans le triangle sémiotique de Peirce, alors que la phase objectivée des situations analysées renvoie à son objet.

Dans le rapport, l'auteur propose une structuration objective des situations qui est articulée, afin de produire des effets de sens, à un registre d'interprétation dont les formes sont diverses : titres, légendes, style, commentaires, métaphores, mise en page... A travers cette série de *contraintes interprétantes* (Marty), le chercheur tente d'orienter le processus d'interprétation dans des directions privilégiées. Mais, les acteurs interpréteront les données du rapport en fonction d'un ensemble de significations qui dépend aussi de leur position dans le système et de leur histoire personnelle. Au final, le terme provisoire du processus d'interprétation ne sera donc pas une interprétation commune à l'ensemble des lecteurs. Au contraire, les interprétations seront plurielles. Nous faisons cependant l'hypothèse, étayée par nos observations, qu'elles resteront inscrites dans les orientations tracées par le texte. Pour approfondir cette question, nous prendrons appui sur ALCESTE qui est la méthodologie particulière d'analyse des donnés textuelles que nous avons mise en oeuvre dans le cadre de notre enquête.

---

[2]MARTY, C., MARTY, R. (1992). Quatre-vingt-dix-neuf réponses sur la sémiotique. Montpellier : Centre Régional de Documentation Pédagogique, 100 p.

3 - ALCESTE : une méthodologie d'analyse des données textuelles

Pour présenter ALCESTE, nous avons choisi l'énoncé pédagogique de l'hypothèse générale[3]. *Le locuteur, au cours de son élocution, investit des mondes successifs divers et ces mondes, en imposant leurs objets, imposent du même coup leur type de vocabulaire. Par conséquent, l'étude statistique de la distribution de ce vocabulaire devrait permettre de retrouver la trace de ces pièces mentales que le locuteur a successivement habitées, trace perceptible en termes de mondes lexicaux, ces mondes lexicaux renvoyant à telle ou telle manière particulière de choisir à un moment de son discours un système de référence ou un autre.*

3.1 - Le principe de l'analyse[4]

Afin de dégager du corpus les différents mondes lexicaux habités par les locuteurs, la méthodologie et le logiciel ALCESTE produisent une **A**nalyse **L**exicale par **C**ontexte d'un **E**nsemble de **S**egments de **T**exte. Autrement dit, cette méthodologie vise *l'étude des principales lois statistiques de distribution du vocabulaire*. A cette fin, l'algorithme du logiciel commande l'analyse d'un tableau des données dont les lignes sont constituées par l'ensemble des énoncés élémentaires[5] qui constituent le corpus et les colonnes par l'ensemble des formes réduites, ou lexèmes, qui *fixent* le vocabulaire puisque la formation d'un mot varie en fonction des marques du pluriel ou de la conjugaison. Chaque cellule du tableau enregistre la présence (1) ou l'absence (0) d'une forme réduite F dans un énoncé E. Ce tableau est principalement composé de zéros (97%) si bien que la simple co-occurence de deux mots dans une même unité de contexte devient très significative ; ce qui explique le choix effectué par Reinert de privilégier l'analyse sémantique par rapport à l'analyse syntaxique. Une Classification Descendante Hiérarchique permet de classer les énoncés en fonction de la similitude de leur vocabulaire. Sans rentrer dans le détail des techniques

[3]REINERT, M. (1992). Alceste. Documentation. Toulouse.

[4] REINERT, M. (1990). ALCESTE : Une méthodologie d'analyse des données textuelles et une application : Aurélia de Gérard de Nerval. Bulletin de méthodologie sociologique, n° 26, Mars 1990, p. 24-54.

REINERT, M. (1987). Classification descendante hiérarchique et analyse lexicale par contexte - application au corpus des poésies d'Arthur Rimbaud. Bulletin de méthodologie sociologique, n° 13, janvier 1987, p. 53-90.

[5] Le découpage du corpus en unités de contexte opérationnalise la notion d'énoncé qui est définie comme la plus petite unité de texte susceptible de décrire, selon Reinert (1992) la représentation sous-jacente d'un sujet, et dont la marque est, au minimun, la relation qu'établit l'individu entre un prédicat et un sujet (grammatical). *Le corpus est considéré comme un ensemble de segments de texte. Ces segments de texte sont appelés unités de contexte* .

statistiques, notons que la distance mathématique utilisée pour calculer la C-H-D est le CHI 2.

C'est en décrivant ces classes que l'analyste peut mettre à jour les mondes lexicaux qui structurent les discours des acteurs, et donc analyser les systèmes de référence qui les fondent. Les classes sont caractérisées par un sous-ensemble lexical spécifique qui définit leur *contexte statistique* que nous appellerons les objets concluants de l'analyse statistique, et qui constituent le matériau du commentaire interprétatif. Indiquons, à titre d'illustration, quelques uns des fichiers disponibles :

- Le dictionnaire des formes réduites qui permet de voir comment, à partir du vocabulaire, les formes réduites ont été déterminées.

- Les unités de contexte élémentaires rangées par classe.

- Le profil des classes qui regroupe l'ensemble des formes réduites liées significativement à une classe donnée.

- Le profil des classes par couples qui regroupe les couples de mots significativement présents dans chaque classe.

- La liste des segments répétés qui comprennent entre trois et cinq lexèmes successifs.

- L'analyse factorielle des correspondances, effectuée sur le tableau croisant les formes significatives et les classes d'appartenance.

... / ...

3.2 - De la procédure interprétative à la communication des résultats

Si le tableau des données est analysé par une classification hiérarchique descendante, l'interprétation s'effectue de manière ascendante à partir de l'interprétation séparée de chacune des classes terminales. A l'intérieur du lexique qui définit une classe, l'analyste recherche les proximités sémantiques. Les noyaux de sens ainsi dégagés jouent un rôle complémentaire. Chaque contexte peut être alors commenté et nommé. Un commentaire surplombant rendant compte de l'organisation globale des classes entres elles.

Nous illustrerons brièvement cette procédure en revenant à l'enquête auprès des jeunes dont les entretiens étaient engagés par la consigne suivante : "*Si vous avez quelque chose à ajouter, si vous voulez faire savoir quelque chose au Président*

du Conseil Général quelque chose qui vous tient à coeur, ne vous en privez pas... Il est possible que nous ayons oublié une importante question qui vous préoccupe..." Ne pouvant reproduire toutes les données dans le cadre de cette communication sans risquer de trop l'alourdir, nous avons sélectionné, pour chacune des quatre classes qui organisent la distribution du vocabulaire, les vingt-deux premiers termes classés du plus ou moins représentatif en fonction du CHI 2 d'association à la classe, et un ou deux énoncés significatifs du profil de chacune des classes.

Cl 1 : CHI2 moyen d'association des mots à la classe = 10,56 ; 78 formes : quartier, boulot, rue, vivre, français, drogue, sécurité, préoccuper, délinquance, chômage, racisme, scolaire, étude, travail, difficile, dur, défavorisé, SIDA, finir, soir, sortir, tranquille...

Je travaille dans un quartier défavorisé, c'est difficile. / Ce qui est important maintenant, c'est de finir les études et trouver un boulot, après le reste, c'est secondaire.

Cl. 2 : CHI2 moyen d'association des mots à la classe = 7,51 ; 78 formes : Aude, emploi, connaître, développer, région, chose, créer, informer, essayer, parler, discuter, département, donner, passer, spécial, salle, personnel, suffisant, continuer, échanger, information, bouger...

On parlerait de Carcassonne et de l'avenir de Carcassonne, de l'avenir de l'Aude du point de vue emploi.

Cl. 3 : CHI2 moyen d'association des mots à la classe = 12,67 ; 73 formes : politique, parole, prendre, intérêt, compte, désir, discussion, canton, lycée, homme, réunion, collège, réflexion, porte, jeune, libre, débat, participer, exprimer, comprendre, sujet, place...

Pendant une journée, participer à un débat, dans chaque canton, en prenant plusieurs jeunes de différents établissements.

Cl. 4 : CHI2 moyen d'association des mots à la classe = 13,56 ; 59 formes : question, genre, répondre, limité, venir, Balladur, besoin, vous, objectif, écouter, représenter, demander, servir, dire, compléter, expliquer, audois, chercher, rendre, facile, pouvoir, réfléchir...

A propos de toutes ces questions auxquelles nous avons répondu dans ce questionnaire, qu'il fasse ce qu'on demande. / On m'a demandé de venir, je suis venu, c'est pour vous rendre service.

Pour communiquer les résultats, nous avons choisi de nommer les classes en accompagnant la présentation des données les plus significatives d'un commentaire interprétatif. Ne sont reproduits ici que les titres des classes 1 et 2 sans leur commentaire :

- "**préoccupations proximales, sociales et sécuritaires**"
- "**préoccupations globales et économiques**".

Ces expressions tentent de résumer, de condenser le contexte sémantique que traduit le profil des classes, mais leur choix relève de l'arbitraire du chercheur. Aucune règle statistique ne permet de les déterminer. Or, nous avons souligné que dans les dispositifs d'aide à la décision, les modes d'accès et de diffusion de l'information renforcent l'impact de ces éléments qui jouent, au sens fort, le rôle de *contraintes interprétantes*. Souvent pressés de consommer les résultats des études, les acteurs socio-économiques ne s'attardent guère, à l'inverse des acteurs de la communauté scientifique, sur l'analyse critique de la procédure interprétative. Comment, dans ces conditions, ne pas enfermer l'interprétation des acteurs dans la seule direction tracée par l'intervenant ? Comment ne pas réduire la richesse du profil d'une classe à l'expression qui la désigne ? D'autant plus que de synthèses en résumés, peuvent apparaître comme déterminations objectives, les éléments d'interprétation mis en place par le chercheur. C'est-à-dire que l'interprétation des acteurs socio-économiques risque d'être fondée sur un rapport biaisé entre ce qu'ils pensent être une donnée objective de la situation problématique et leurs propres systèmes de significations liés au contexte de terrain. A titre d'illustration, nous reproduisons (en respectant la mise en forme de l'auteur) le texte de la note de synthèse qui reprend le commentaire des classes 1 et 2 :

Ces jeunes erratiques qui se sont le plus librement exprimés dans les entretiens permettent de faire émerger plusieurs préoccupations et modalités d'engagement civique.

** Des préoccupations proximales, sociales et sécuritaires.*

Dans un environnement difficile pour les jeunes, émerge un souci de sécurité, de tranquillité : prévention, santé, emploi.

** Des préoccupations globales et économiques.*

Avec un engagement soumis à une condition : un nécessaire effort d'information et de communication, entre les jeunes eux-mêmes mais aussi avec le Conseil Général.

En conclusion

Le problème est difficile. Cependant, nous pensons qu'une stratégie de communication peut contribuer au dessin d'une solution. En effet, au cours de la procédure interprétative, chacun des éléments objectifs déterminés par le calcul statistique est utilisé comme *interprétant* des autres objets concluants de l'analyse des données. Ainsi, les couples de mots, ou les énoncés représentatifs du profil d'une classe, jouent le rôle de système donateur de sens pour le profil du vocabulaire caractéristique de la même classe. De même, dans notre exemple, le fait que les deux mots, les plus significativement associés à la classe 1, soient "quartier" et "boulot" prend tout son sens lorsqu'on observe que les deux mots les plus significativement associés à la classe 2 sont "Aude" et "emploi". C'est pourquoi,

notre première proposition est de considérer qu'interpréter des données, c'est constituer un ensemble organique de significations où : 1 / chaque élément est déterminé - déterminant par (de) chacun des autres éléments de l'ensemble, 2 / chaque élément est déterminé - déterminant par (de) chacune des intersignifications des éléments entre eux. 3 / chaque élément est déterminé - déterminant par (de) l'ensemble des intersignifications entre le tout de la signification et chacune de ses parties. Cet ensemble est structuré, organisé, par une grille d'analyse dont le rôle "*est de mettre en relation un système d'hypothèses articulé à une problématique avec une structure objective du corpus*" (Reinert, 1987). Dans ce contexte, une stratégie possible de communication des résultats est de proposer comme *contraintes interprétantes*, chaque fois que cela est possible, des éléments déterminés objectivement par le calcul statistique afin de relativiser le poids des reformulations et des commentaires rédigés par le chercheur. Enoncés significatifs, segments répétés, profil caractéristique du vocabulaire peuvent jouer ici le rôle des tableaux, figures et autres courbes, légendes et titres dans la communication des données métriques. Notre volonté est de construire un espace de libre mouvement au processus d'interprétation des acteurs socio-économiques. En conséquence, nous orientons nos travaux sur la recherche des conditions qui permettent de confronter les acteurs socio-économiques aux ensembles organiques de significations. Dans cette perpective, mise en page et ergonomie de la lecture peuvent jouer un rôle efficace d'accompagnement du lecteur.

Notes bibliographiques

GRANGER, G.G. (1968). Essai d'une philosophie du style. Paris : Armand Colin.

REINERT, M. (1987). Classification descendante hiérarchique et analyse lexicale par contexte - application au corpus des poésies d'Arthur Rimbaud. Bulletin de méthodologie sociologique, n° 13, janvier 1987, p. 53-90.

REINERT, M. (1990). ALCESTE : Une méthodologie d'analyse des données textuelles et une application : Aurélia de Gérard de Nerval. Bulletin de méthodologie sociologique, n° 26, Mars 1990, p. 24-54.

REINERT, M. (1992).ALCESTE. Documentation. Toulouse.