

MAX REINERT

Approche statistique et problème du sens dans une enquête ouverte

Journal de la société française de statistique, tome 142, n° 4 (2001),
p. 59-71

http://www.numdam.org/item?id=JSFS_2001__142_4_59_0

© Société française de statistique, 2001, tous droits réservés.

L'accès aux archives de la revue « Journal de la société française de statistique » (<http://publications-sfds.math.cnrs.fr/index.php/J-SFdS>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
<http://www.numdam.org/>

APPROCHE STATISTIQUE ET PROBLÈME DU SENS DANS UNE ENQUÊTE OUVERTE

Max REINERT *

RÉSUMÉ

Dans quelle mesure l'analyse statistique des réponses textuelles à une question ouverte s'apparente-t-elle à l'analyse de contenu (Bardin,1977), à l'analyse de discours (Harris,1966), à la sémiotique (Peirce,1978) ? L'auteur propose une réflexion en relation avec son expérience de la méthode ALCESTE. Par cette méthode, ce n'est pas le sens représenté par les locuteurs qui est visé, mais le sens comme activité problématique, répétitive. Une petite application à l'analyse d'un corpus d'évocations de projets d'avenir permettra d'illustrer le propos.

Mots clés : Statistique textuelle, Méthode ALCESTE, Analyse de discours, Analyse de contenu, sémiotique.

ABSTRACT

In what measure is statistical analysis of textual answers to open question surveys part of content analysis, discourse analysis, or semiotics? The author offers a reflection on this in relation to his experience with the ALCESTE method. By this method, it is not the meaning represented by the speakers which is targeted, but meaning as a 'problematic' activity - that is to say, an activity determined by a problem. The recurrence of the problem leads to a tendency to repeat formulations which seek to contain it. A short application to the analysis of a corpus of evocations of personal projects illustrates this.

Keywords : statistically processing textual data, the "ALCESTE" method, discourse analysis, content analysis, semiotic.

1. Analyse de discours et sémiotique peircienne

On cite souvent cette étude de J. Richard Zappella (Achard, 1993) pour montrer l'effet des formulations des questions sur les réponses. Cette auteur a systématiquement dépouillé les sondages publiés depuis les années cinquante et elle remarque qu'en février 1969 à la question « croyez-vous en Dieu » le nombre de oui est 81 % alors qu'en Août 1969, à la question « Est-ce que vous croyez en Dieu », le nombre de oui devient 66 %. Peut-être y a-t-il un effet d'échantillonnage, mais il est probable que cette chute de pourcentage soit attribuable à la formulation. Pierre Achard, qui cite cet exemple dans son

* C.N.R.S.-PRINTEMPS, Université de Versailles, Saint-Quentin en Yvelines, 45, bd Vauban, 78047 Guyancourt cedex. E-mail : max.reinert@printemps.uvsq.fr

livre «la sociologie du langage», écrit : «ces effets manifestent le caractère actif de l'énonciateur dans l'interaction, la forme «est-ce que» objectivant le contenu de la question et mettant le locuteur à distance de la réponse».

Il y a donc effet de sens par la simple manipulation de la position énonciative du questionné par le questionnement puisque le pourcentage a changé. Les deux questions n'ont pas la même signification alors même que nous leur attribuons le même contenu. Il s'agit donc d'éclaircir la notion de sens, dans son rapport avec le contenu, dans son rapport avec la signification ; enfin, dans son rapport avec la situation réelle.

Dans ces conditions, de quel sens parle-t-on quand on étudie les réponses d'une enquête ouverte en les soumettant aux techniques d'analyse des données ?

Revenons à l'origine théorique de ces pratiques. Pour présenter des applications de l'*Analyse des Correspondances*¹ aux traitements de «données linguistiques», Jean-Paul Benzécri (1981) se réfère aux travaux de Harris sur l'analyse distributionnelle. Il s'y réfère surtout pour donner un sens linguistique à la notion de distribution statistique. Harris a d'autre part beaucoup influencé l'analyse de discours à la française. (Voir notamment l'*Analyse Automatique du Discours* de Pêcheux, 1969).

Dans la version française² de son article de 1952 sur l'analyse du discours, Harris écrit : «*Cet article présente une méthode d'analyse de l'énoncé suivi (écrit ou oral) que nous appellerons discours. C'est une méthode formelle qui ne se fonde que sur l'occurrence des morphèmes en tant qu'éléments isolables ; elle ne dépend pas de la connaissance que le linguiste qui analyse peut avoir du sens spécifique de chaque morphème, et elle ne nous apprend rien de nouveau sur le sens particulier de chacun des morphèmes qui figurent dans le discours en question. Mais ceci ne signifie nullement que nous ne puissions pas découvrir autre chose que la manière dont la grammaire s'illustre dans ce discours. Car bien que nous usions de procédures formelles, proches de celles de la linguistique descriptive, nous pouvons obtenir sur le texte étudié des renseignements que cette dernière ne fournissait pas*». Harris ajoute plus loin : «*Nous avons soulevé deux questions : celle des rapports distributionnels entre phrases, et celle de la corrélation entre langue et situation sociale*». Ce dernier aspect a conduit à s'intéresser au rapport entre la forme distributionnelle des discours et leurs conditions de productions.

Rappelons que l'école Harrissienne est issue d'une école influencée par le béhaviorisme avec les travaux de Bloomfield qui refusait d'analyser le sens à partir d'un point de vue mentaliste ce qu'impose la notion même de contenu, en tant qu'expérience subjective. Les analystes du discours se sont aussi fortement opposés à l'approche des contenus. L'intérêt pour le sens a été plutôt par eux inversé. Au lieu de chercher à interpréter le sens en aval de l'énoncé, il s'agit de remonter à sa source. L'intérêt sur les «conditions de productions» des énoncés conduit à un intérêt pour ce qui engendre l'énoncé,

1. Selon la terminologie de cet auteur (Benzécri, 1973).

2. Traduite par Dubois-Chalier dans le numéro 13 de la revue *Langage* consacré à l'analyse du discours (1969).

un intérêt pour la situation, le *sujet parlant*. On pourrait ajouter, un intérêt pour la référence dans le sens où l'emploie Benveniste dans cette citation : « *Si le sens de la phrase est l'idée qu'elle exprime, la référence de la phrase est l'état de choses qui la provoque* »³. Benveniste, qui a été très critique vis-à-vis de la théorie sémiotique de Peirce⁴, se trouve ici très Peircien. La référence est ce qui, d'une certaine manière, détermine le signe et son sens. Cela n'a plus grand-chose à voir avec l'idée d'un signe qui dénote arbitrairement son objet.

Peirce est un mathématicien et philosophe américain de la fin du XIX^e siècle, créateur avant Saussure de la sémiotique. Son point de départ n'est pas la linguistique mais la logique. Il cherche à définir le signe à partir de la notion d'inférence. Sa question : comment passe-t-on d'un signe à l'autre ? Une de ses définitions du sens est la suivante : « *Le sens d'un signe est le signe dans lequel il doit être traduit* »⁵.

Ce point de départ inférentiel de la sémiotique pourrait faire penser qu'elle ne constitue qu'une partie de la logique mais, à l'inverse, c'est à son dépassement qu'il aboutit d'où la création de la *sémiotique*. L'inférence purement logique, la déduction, n'est qu'un aspect de l'inférence. Un aspect trivial d'où le temps est exclu. Une déduction ne crée pas de nouveau signe, tout est déjà dans le premier signe. Par exemple déduire de « Socrate est un homme », « Socrate est mortel » sachant par ailleurs que « les hommes sont mortels » ne vient que d'une autre manière de lire le même schéma où l'ensemble des mortels contient l'ensemble des hommes dont un élément est Socrate.

Peirce distingue deux autres types d'inférences, l'induction et l'abduction. L'induction est liée à la capacité d'abstraction, de généralisation. Trouver la loi générale à partir de cas particuliers. Le temps n'y sert qu'à renforcer une conviction ou éventuellement à l'anéantir, mais tant qu'elle se vérifie l'induction n'implique ni histoire, ni parcours, ni sujet. Notre expérience quotidienne montre au contraire la capacité créative du temps. Par quel type d'inférence en arrive-t-on à poser une hypothèse nouvelle, par exemple ? Pour Peirce, il s'agit bien d'une inférence et il l'appelle « abduction » dont il cherche la véritable nature en analysant la notion d'habitude. Elle ressemble à l'induction par le fait qu'elle se constitue en fonction d'une série d'expériences antérieures, mais s'en sépare par le fait qu'elle ne présuppose pas une identité des cas dans la série qui mène à la loi. Dans ce qui advient, ici et maintenant, il y a seulement comme un air de famille, une ressemblance avec du déjà connu. C'est par cet intermédiaire que Peirce en vient à s'intéresser aux usages dans leurs aspects créatifs.

3. Problèmes de Linguistique générale, page 226 du tome 2 (Benveniste, 1966).

4. *Op.cit.*, page 45 : « *La difficulté qui empêche toute application particulière des concepts peirciens, hormis la tripartition bien connue (icône, indice, symbole), est qu'en définitive le signe est posé à la base de l'univers entier, et qu'il fonctionne à la fois comme principe et définition pour chaque élément et comme principe d'explication pour tout ensemble, abstrait et concret. L'homme entier est un signe, sa pensée est un signe, son émotion est un signe. Mais finalement ces signes étant tous signes les uns des autres, de quoi pourraient-ils être signe qui ne soit pas signe ?* ».

5. Cité par Everaert-Desmedt (1990) des « *Collected Papers* » (4.132).

Par la notion d'habitude active, d'usage, le temps devient une composante à part entière, une composante créative de l'inférence, chaque parcours dépendant maintenant d'une histoire avec ses régularités et sa singularité. Mais une habitude, une pratique, n'est pas une représentation, ce qui ne signifie pas qu'on ne cherche pas à la représenter justement à travers de nouveaux signes d'où ce double mouvement entre ce que l'on cherche à voir et ce qu'on est déterminé à faire par l'habitude, détermination relative car il y a beaucoup de plasticité dans la notion de ressemblance. Dans la définition de Peirce : « *Le sens d'un signe est le signe dans lequel il doit être traduit* ». Le « il doit » dépasse donc la simple déduction logique, pour exprimer une détermination historique du sens. Du moins c'est là notre interprétation de Peirce.

Dans cette perspective dynamique, ce qu'on appelle signe n'est pas à définir en soi mais en rapport avec ce mouvement du sens. Un signe, c'est ce qui fait signe pour quelqu'un au cours de son activité, au cours d'une « sémiose ». Peirce distingue trois manières de « faire signe » : l'icône, l'indice et le symbole que nous développerons comme les trois *moments* caractéristiques d'une sémiose⁶ :

1) *Le moment iconique de la sémiose*, liée au temps de l'abduction. L'objet dynamique en situation, l'histoire du sujet, l'usage, le désir, font qu'une apparence s'impose pour quelqu'un, ici et maintenant. L'apparence est une forme d'abduction en ce sens qu'elle se donne comme conscience de... Il y a un aspect sensible, et immédiat de l'apparence (notion de « priméité » chez Peirce). Mais cet aspect premier pour la conscience s'inscrit dans la série des expériences passées, que la présence actuelle réordonne en la réactualisant.

Arrêtons-nous un instant au niveau d'un discours. Ce moment iconique de la saisie d'un sens se retrouve dans la perception d'un contenu ; contenu d'un énoncé, contenu d'un mot plein. Car, qu'est-ce qu'un contenu, sinon ce sentiment même de plénitude qui donne l'impression de voir la chose sous le mot. Impression purement imaginaire bien sûr. Mais qui anime réellement la trace d'une histoire dans un corps. Si le mot est symbole, sa plénitude est relative à sa capacité d'interférer avec la série des expériences singulières d'un sujet. Autrement dit, ce qu'on appelle généralement un mot plein symbolise cette capacité.

2) *Le moment indexical de la sémiose* : En captant un regard, l'icône de ce premier moment conduit à des effets réels, dans un corps réel, au niveau des désirs, des affects ou des actes, ici et maintenant. Le moment indexical est relatif à la prise de conscience de cette pression réelle de la référence sur ce qu'il advient comme conséquence de cette présence du signe. Je vois un mur : moment iconique de la sémiose. Je me cogne dedans : moment indexical de la sémiose. Je prends conscience d'une propriété du mur : moment symbolique et réflexif de la sémiose que l'usage social m'aide à représenter à travers ses symboles. Par exemple, le mot « mur », des expressions « se cogner dans des murs », etc.

6. Notre propos est d'exposer nos hypothèses le plus simplement et nous ne prétendons pas exposer la théorie peircienne. Cependant, il nous semble évident qu'elles en découlent très directement même si nous ne cherchons pas à le montrer ici.

L'indexicalité n'est perçue, bien sûr, qu'en situation. Au niveau d'un discours, cependant, certains mots la représentent, la symbolisent, et permettent de la montrer. Par exemple, les déictiques, shifters ou embrayeurs : je, tu, ici, là-bas, maintenant, voilà, dedans, etc...

Au moment iconique de la sémiologie, le signe est perçu uniquement en lui-même (notion de priméité chez Peirce). Au moment indexical, le signe semble se dédoubler, se révéler dans son rapport au référent. Il n'est pas simple apparence, mais affecte directement. Peirce relie ce moment à la notion d'effort, de résistance, de choc. Dire que cela affecte directement c'est prendre conscience d'une béance entre un intérieur affecté et un extérieur apparent. Ainsi la conscience d'un réel n'est sensible que par la conscience d'une coupure entre soi et soi ou entre soi et l'autre.

3) Si la sémiologie s'ouvre sur des apparences fluctuantes, se prolonge comme coupure, elle se réfléchit également en elle-même dans une représentation de ce qui se produit. Dire qu'il y a mouvement c'est déjà se représenter la coupure comme causalité entre deux moments. Ce ne sont pas les choses qui se représentent mais les actes qui nous ont amenés à les voir ou à les construire. Une recette de cuisine représente la préparation d'un plat. Le mot « concept » est parfois utilisé dans ce sens. Un concept de plat, de logement. Ce sens n'est pas opposé au sens classique du concept, comme prédicat, comme propriété, au contraire, il le restitue dans son dynamisme, dans son pragmatisme. Ce point est important car il permet de passer de la notion d'acte à la notion de structure⁷. En se représentant ses actes, on passe des actes particuliers aux concepts généraux, aux idées, aux lois (notion de tiercéité chez Peirce). Prendre conscience d'un sens, de ce point de vue, c'est aussi, par là même, sortir de l'expérience immédiate en se situant soi-même comme assujetti à des lois, à un cycle des répétitions, dans un ordre social ou mythique. On appellera ce moment, *le moment symbolique de la sémiologie*. Un signe se lit comme symbole s'il est général. Il ne peut se lire que médiatement à travers un usage des icônes. Peirce distingue ici la notion de type de la notion de réplique, que nous appelons plutôt « occurrence ». L'usage social fait que certaines icônes sont médiatement saisies comme des répliques de symboles. Il y a là un aspect récursif de la sémiologie à ce niveau car toute loi, en se représentant dans une icône, est susceptible d'être reconnue par quelqu'un comme expression singulière de son propre vécu de sujet, comme plénitude, comme nouvelle icône, comme métaphore de...

Mais revenons au discours. L'objet dynamique d'un discours est lui aussi, à la fois à l'origine du discours, qu'il provoque, et à l'horizon de tout nouvel énoncé, puisque, en situation, il nous apparaît comme ce qu'on cherche justement à dire comme sens. Mais le *sens* n'est pas dans le texte, le *sens* était dans le temps de cette circulation dynamique de la parole.

Si cependant la lisibilité d'un texte dépend de sa qualité intrinsèque, de cette qualité, on ne peut rien dire en dehors d'une lecture. Car l'apprécier dépend de la capacité du lecteur à y articuler sa propre temporalité, sa propre scansion,

7. Comment ne pas évoquer Piaget...

pour que chaque moment du texte soit saisi en plénitude avec sa propre expérience. Ce qui sera sa manière particulière de lui donner un sens.

2. La méthode « ALCESTE »

Mes propres travaux sur la méthode⁸ ALCESTE ont débuté avec ma thèse en 1979 dans cette école Benzécriste de l'analyse des données. La méthode *ALCESTE* repose sur un même paradigme que celui proposé par Harris. L'analyse distributionnelle d'un discours donne des informations sur les conditions qui l'ont produit. Elle se différencie de la perspective Harrissienne en se situant sur le versant sémiotique de notion de distribution et non pas linguistique, d'où mon évocation de Peirce. Dans la modélisation proposée, le discours est considéré essentiellement comme la trace d'une activité temporelle.

Comme nous l'avons suggéré, le sens se décline en trois modes : en fonction d'un contenu imaginaire de type iconique, en fonction d'une activité temporelle dépendante des aléas d'un réel, en fonction d'une représentation symbolique de cette activité, par la mise en évidence de lois, de concepts, de structures, etc. Selon notre point de vue, ces trois modes ne peuvent fonctionner séparément et nous avons tendance à ne retenir que le troisième mode pour étudier le sens. On cherche donc à traduire une représentation symbolique, ou que l'on prend comme telle, par une autre, en montrant en quoi elles sont logiquement équivalentes. Et l'on oublie l'origine dynamique du sens. Dans la perspective pragmatique, c'est parce que les activités se stabilisent qu'on peut les représenter. L'objectif d'une approche distributionnelle n'est plus de traduire une représentation dans une autre mais de modéliser directement l'activité pour voir en quoi elle se stabilise, en quoi elle détermine ainsi des représentations. Mais les activités elles-mêmes ne peuvent être saisies en chacun de leur moment qu'en rapport avec des contenus, avec des icônes. Tel est le cercle vicieux des sémoses. Nous retenons deux hypothèses :

- 1) Le texte est une trace d'une activité temporelle.
- 2) Les mots pleins sont une trace des contenus, au sens iconique du terme.

Le modèle que nous proposons est donc très simple : on cherche à étudier la distribution des mots pleins dans les différents moments d'un discours pour voir en quoi l'activité discursive se stabilise.

Cela dit, il serait contraire à notre conception du sens de penser qu'on peut modéliser la totalité du sens. Il y a un réel du sens qui ne peut se modéliser, réel qui ne peut se vivre que lors d'une lecture en fonction d'une scansion, en fonction de ruptures que le texte ne peut indiquer parfaitement. Le mot discours exprime cette discontinuité erratique d'un cours. Rien ne permet de fixer l'empan d'une interférence d'un segment de texte avec l'expérience d'un lecteur, autrement dit, sur ce qui fait contenu. Aussi, et nous insistons sur ce

8. On peut se reporter à (Reinert, 1983, 1986) pour des précisions techniques sur la classification descendante hiérarchique ; à (Reinert, 1990) pour une présentation synthétique de la méthodologie illustrée à l'aide d'une application ; à (Reinert, 1999, 2001) pour une discussion sur la méthodologie d'analyse de discours.

point, les limites d'un tel moment de saisi d'un contenu ne peuvent être que simulées arbitrairement par un algorithme.

Ces limites une fois fixées, un « quelque chose » du contenu peut être relevé de manière relativement objective, comme trace de l'acte de saisi : la cooccurrence des mots pleins. Concrètement, on découpe donc le corpus en fragments⁹ de texte d'à peu près même longueur, en tenant compte toutefois de la ponctuation si elle existe. Cela, pour modéliser la scansion possible d'un locuteur ou d'un lecteur quelconque. On vérifie cette indifférenciation relative en faisant varier la longueur, en effectuant plusieurs analyses et en ne retenant que les résultats stables. On ne résout pas tout ainsi, mais cela permet de réfléchir à la notion d'amplitude de la scansion, d'en explorer les limites. Pour l'analyse de textes littéraires les résultats sont peu sensibles aux grandes variations¹⁰. Pour certains discours cliniques ce n'est pas le cas. Il y a là un champ de recherche à explorer.

Le discours est donc modélisé par un tableau binaire à double entrée. En lignes, les différents fragments de texte, appelés « Unités de contexte ». En colonnes, les mots pleins, dans leur acception courante : plutôt des noms, verbes, adjectifs et adverbes.

Dans la complexité des sémoses ayant produit un discours, lignes et colonnes du tableau ne modélisent qu'une toute petite part des traces relatives aux deux premiers moments des sémoses, saisi d'un contenu et indexicalité de chaque moment. Le tableau de données représente donc symboliquement la dispersion même de ces micro-moments. Il est une trace symbolique de l'activité discursive. Par l'analyse statistique du tableau, on cherche à rendre compte de la stabilisation de cette activité relativement aux contenus. Cette stabilisation éventuelle peut se représenter, se signifier, se calculer, par exemple à l'aide d'axes factoriels ou de classes. Dans ALCESTE, nous avons privilégié la représentation en classes, mais la procédure de classement est descendante hiérarchique¹¹. Cette méthode privilégie le concept d'opposition entre classes à celui de proximité entre deux unités. C'est la raison pour laquelle elle est descendante.

Cette technique a été élaborée pour traiter des tableaux binaires très clairsemés mais de grandes dimensions. Dans sa version 5, le logiciel ALCESTE permet d'analyser des corpus composés jusqu'à 40 000 fragments et de les classer en tenant compte de la distribution simultanée de 3 000 mots

9. Ces « fragments » sont appelés « unités de contexte » (UC). Leur ensemble constitue une partition du corpus. Étant donné qu'il y a plusieurs analyses portant sur des longueurs variables de ces UC, nous distinguons d'autre part les « unités de contexte élémentaires » (UCE) qui servent de « briques de base » pour composer les différentes unités de contexte qui seront analysées et classées : toute UC est donc constituée par la concaténation d'une ou plusieurs UCE successives dans le texte (en fonction du paramétrage du plan).

10. Voir, par exemple, notre communication aux JADT de 1993 (Reinert, 1993).

11. Au premier pas, ces fragments sont distribués en deux classes les plus différenciées possibles pour ce qui concerne leur vocabulaire spécifique. Au second pas, on reprend les fragments de texte de la plus grande des deux classes et l'on réapplique le même algorithme. Au troisième pas, les fragments du corpus sont maintenant distribués en trois classes. On choisit de nouveau la plus grande et ainsi de suite jusqu'à obtenir le nombre désiré de classes (Reinert, 1983, 1986).

pleins. Le nombre de « un » du tableau est limité à un maximum de 1 500 000. Cela permet de traiter des corpus allant jusqu'à 60 millions de caractères.

3. Présentation d'un exemple

Nous proposons d'illustrer ces conceptions par la présentation d'une brève application. Il s'agit de l'analyse d'un petit corpus recueilli dans les années 83 par un étudiant en maîtrise de psychologie sociale de l'université de Toulouse-le Mirail. Le corpus est composé de 61 réponses d'adolescents de la banlieue du Mirail à la question : « Quels sont vos projets d'avenir dans le domaine professionnel, familial ou autre ? ». Il sert de corpus d'essai pour le logiciel (*cf.* Annexe, encadré 1).

Le logiciel est programmé pour que les paramètres d'une analyse s'adaptent automatiquement à la grandeur du corpus. C'est ce que nous appelons « l'analyse standard ». Ce sont les résultats de cette analyse que nous proposons ici. Le corpus des 61 réponses a été segmenté par le logiciel en 120 fragments ou UCE compatibles bien sûr avec les réponses. L'ensemble des « mots pleins » différents d'au moins quatre occurrences (après lemmatisation) ont été retenus pour l'étude des distributions (soit 96).

Dans l'analyse standard, le découpage des UCE est relativement arbitraire, même si l'algorithme tient compte de la ponctuation lorsqu'elle existe. Aussi, dans la procédure standard, on procède à un double découpage pour tester la stabilité des résultats. Ainsi, le nombre de classes stables obtenues varie selon les corpus. 3 classes terminales émergent ici du calcul regroupant 65 % des UCE (*cf.* Annexe, encadré 2).

On ne s'étonnera pas de retrouver au niveau de l'interprétation, les différents moments d'une sémiologie déjà présentés. Ainsi, les classes obtenues peuvent être interprétées de trois points de vue :

- a) Comme contenu, c'est-à-dire, en elles-mêmes en observant notamment la liste des mots pleins ou des fragments qui leur sont spécifiques ;
- b) Comme activité, car leurs oppositions les unes aux autres, expriment un certain dynamisme du parcours discursif. Ce dynamisme peut traduire des différences de postures au niveau de l'énonciation ce que peuvent souligner les distributions de mots outils ;
- c) Comme représentation puisque ces classes font système et reflètent une certaine stabilisation de l'activité discursive des interviewés. Les caractéristiques de la population interviewée sont susceptibles d'apporter un éclairage pour interpréter cette détermination du sens dans la représentation. Ici, n'ont été relevés que le sexe, en tant qu'il peut différencier des formes de projet et l'appartenance à une association en tant qu'elle indique un degré d'intégration sociale. Trois moments donc que nous allons survoler rapidement :

Commençons par les mots outils et les mots étoilés afin de fixer les différentes orientations de l'activité et leurs déterminants (*cf.* Annexe, encadré 3) :

APPROCHE STATISTIQUE ET PROBLÈME DU SENS

• 1^{re} classe : Peu de mots outils. Des adverbes en relation avec des quantités. Cette classe est plus spécifique des garçons n'appartenant pas à des associations. On notera la présence de l'auxiliaire « avoir ».

• 2^e classe : Beaucoup de mots outils. Indication d'une troisième personne : on, leur, leurs, ils. Des verbes modaux : pouvoir, falloir qui évoquent un style plus prescriptif. Des joncteurs argumentatifs : pour que, car. Une forte présence de l'auxiliaire être. Cette classe est plus spécifique des filles.

• 3^e classe : Peu de mots outils mais forte présence de marqueurs temporels : encore, après, enfin, longtemps... Cette classe est plus spécifique des garçons appartenant à des associations.

Nous avons dénommé les classes en fonction de l'orientation des postures : *avoir, être et devenir*.

Au niveau des contenus observons la distribution des mots pleins (cf. Annexe, encadré 3) et les segments de texte caractéristiques des classes (cf. Annexe, encadré 4) :

1) La première classe évoque un imaginaire d'affiches publicitaires : *maison, voiture, voyage, femme*;

2) Dans la seconde classe, des mots de la question sont repris et argumentés : *projet, vie, professionnel*. Des difficultés à résoudre : *difficulté, décide, pouvoir*. Des catégories sociales : *jeunes, société*.

Les UCE associées tendent à montrer une difficulté à concevoir des projets d'où des revendications sociales exprimant une certaine révolte.

3) La troisième classe évoque la représentation d'un parcours scolaire et professionnel : *formation, technique, année, passer, baccalauréat, continuer, étude*.

Ce descriptif des contenus (tel que l'analyste les perçoit) doit maintenant être prolongé par une réflexion sur ce qu'ils peuvent recouvrir, dynamiquement. Revenons *au problème de la formulation de la question* et observons chaque classe, non plus comme contenu, mais comme la trace d'une activité répétitive des interviewés évoquant une « posture » vis-à-vis de la question. On demande à des adolescents de milieu défavorisé leurs projets d'avenir. Il y a là quelque chose de l'ordre de la provocation. Involontaire sans doute, mais qui entre dans ce qui sera produit comme sens.

Les élèves ayant des projets n'entendent pas la question comme embarrassante. Ils sont plutôt ouverts à une proposition qu'ils se sont déjà formulée et ils présentent leurs projets avec complaisance. Ce sont plutôt des garçons déjà intégrés à des réseaux sociaux et, sans doute, mieux intégrés dans le système scolaire.

Les adolescents en difficulté scolaire préféreraient sans doute éluder la question. Semble se distinguer sur ce trop petit échantillon une différence de comportement liée au sexe. Les filles de cette population expriment une certaine révolte, un agacement, mais elles restent dans la suggestion de la question qu'il ne fallait pas poser, même si elles la reprennent sous un mode plus revendicatif.

Elles restent dans la problématique d'un projet réaliste qui ne peut arriver à se formuler.

Par contre, les garçons, sans projet «réaliste», cherchent à échapper à la consigne par un moyen ludique. D'où la dérive «imaginaire» du contenu des réponses, dérive influencée, il est vrai, par les médias, les feuillets et la publicité.

Bibliographie

- ACHARD P. (1993), *La sociologie du Langage*, P.U.F., Paris.
- BARDIN L. (1977), *L'analyse de contenu*, P.U.F., Paris.
- BENVENISTE E. (1966), *Problèmes de linguistique générale*, 2, Gallimard, Paris.
- BENZÉCRI J.-P. (1973), *L'Analyse des Données*, 2, *L'analyse des correspondances*, Dunod, Paris.
- BENZÉCRI J.-P. (1981), *Pratique de l'Analyse des Données : linguistique et lexicologie*, Dunod, Paris.
- EVERAERT-DESMEDT N. (1990), *Le processus interprétatif*, Mardaga, Liège.
- HARRIS Z.S. (1969), Analyse du discours, *Langage*, n° 13, pp. 8-45.
- LEBART L., SALEM A. (1994), *Statistique textuelle*, Dunod, Paris.
- PÊCHEUX M. (1969), *Analyse automatique du discours*, Dunod, Paris.
- PEIRCE Ch. S. (1978), *Écrits sur le signe*. Éditions du Seuil, Paris.
- REINERT M. (1983), Une méthode de classification descendante hiérarchique, *Cahiers de l'Analyse des Données*, 3, pp. 187-198.
- REINERT M. (1986), Classification descendante hiérarchique : un algorithme pour le traitement des tableaux logiques de grandes dimensions, In *Data analysis and informatics*, edited by E. Diday *et al.*, pp. 23-28. Elsevier Science Publishers, Amsterdam.
- REINERT M. (1993), Quelques problèmes méthodologiques posés par l'analyse de tableaux Énoncés x Vocabulaire, In *JADT 1993*, edited by S.J. Anastex, pp. 539- 549, E.N.S.T. - TELECOM (Paris 93 S 003), Paris.
- REINERT M. (1999), Quelques interrogations à propos de l'«objet» d'une analyse de discours de type statistique et de la réponse ALCESTE, *Langage & Société*, décembre 1999, 90, pp. 57-70.
- REINERT M. (2001), Alceste, une méthode statistique et sémiotique d'analyse de discours; application aux Réveries du promeneur solitaire, *La Revue Française de Psychiatrie et de Psychologie Médicale*, octobre 2001, V, 49, pp. 32-36.

Annexe : les résultats de l'analyse ALCESTE

ENCADRÉ 1. – Extrait du corpus soumis à l'analyse.

**** *sexe_m *assoc_oui *sa_12¹²
 Je n'ai pas l'habitude de faire des projets, je vis au jour le jour.
 Les adolescents font des projets à partir du moment où ils se rendent
 compte qu'ils ne doivent compter que sur eux-mêmes et qu'ils doivent
 se prendre en charge... De quel genre de projets s'agit-il ?...

ENCADRÉ 2. – Du corpus au dendrogramme de la classification descendante hiérarchique.

Premiers résultats :

Nombre de mots pleins analysés: 96
 Nombre de mots outils : 84
 Nombre de mots étoilés : 7
 Nombre d'occurrences analysées: 937
 Nombre de réponses (u.c.i) : 61
 Nombre de segments (u.c.e.)¹³ : 120

DONN.1 Nombre de mots par u.c.: 10 Nombre d'u.c. : 96
 DONN.2 Nombre de mots par u.c.: 12 Nombre d'u.c. : 83

Recherche des classes stables : 78 uce bien classées sur 120 (65.00 %)
 Nombre d'u.c.e. distribuées simultanément dans les deux CDH : 100

Tableau croisant les deux partitions retenues :

	RCDH1 *	RCDH2		
classe *		1	2	3
poids *		23	44	33
1	26 *	21	4	1
2	57 *	2	40	15
3	17 *	0	0	17

Classification Descendante Hiérarchique...
 Dendrogramme des classes stables (analyse de DONN1) :

```

    ----|----|----|----|----|----|----|----|----|
    Cl. 1 ( 21uce) |-----+
                   18 |-----+
    Cl. 2 ( 40uce) |-----+
                   19 |-----+
    Cl. 3 ( 17uce) |-----+
```

12. Les lignes étoilées séparent les réponses ou unités de contexte initiales (UCI) et les mots étoilés de la ligne caractériseront l'ensemble des UCE construites ultérieurement et composant cette UCI.

*sa_12, par exemple, indique que la réponse vient d'un garçon appartenant à une association.

13. À partir d'une première segmentation du corpus en « unités de contexte élémentaire » (UCE), on définit des « unités de contexte » (UC) de longueur variable. Par exemple, dans

APPROCHE STATISTIQUE ET PROBLÈME DU SENS

ENCADRÉ 3. – Distribution des mots outils, des mots étoilés et des mots pleins dans les classes.

Classe 1 : 21 uce " AVOIR "	Classe 2 : 40 uce " ETRE "	Classe 3 : 17 uce " DEVENIR "
8.47 sans 7.92 tout 4.58 et 2.50 cinq 2.15 me 2.29 avoir	7.31 se 6.18 sont 6.18 pour-qu< 6.18 leurs 4.11 ça 4.01 plus-d< 4.01 ils 4.01 dix 4.01 beaucoup 2.96 leur 2.96 est 2.95 etre 2.86 on 2.77 ma 2.77 car 2.36 pouvoir. 2.03 falloir. 2.01 aussi	11.20 encore 6.80 apres 3.69 surtout 3.69 enfin 3.69 longtemps
5.89 *sexe_m 4.24 *sa_11	20.99 *sexe_f 6.06 *sa_21 6.02 *sa_22 3.74 *assoc_non	19.28 *sa_12 9.12 *assoc_oui 8.64 *sexe_m
30.21 maison+ 26.39 voiture+ 20.87 belle+ 20.87 monde+ 14.50 esper+er 13.51 femme+ 8.47 voyag+er 6.38 devenir+ 5.45 mari+er 5.22 plein+ 5.22 sport+ 2.50 professeur+ 2.50 trouv+er	8.47 jeune+ 8.40 metier+ 6.94 realis+er 6.18 fait 5.08 rest+er 5.05 an+ 4.11 vie+ 4.11 projet+ 4.01 pouvoir+ 2.96 difficile+ 2.96 societe+ 2.96 mari+ 2.67 decide+ 2.67 professionn+el	36.51 techn+16 36.51 format+ion 14.44 annee+ 10.62 pass+er 7.68 compte+ 7.00 baccalaureat< 7.00 premier+ 5.35 continu+er 4.58 independ+ant 4.58 rentr+er 4.58 fond+er 3.69 cote+ 3.69 essa+yer 2.75 etude+ 2.23 famille+

le tableau DONN1, les UC sont définies par des concaténations d'UCE de manière à ce que chaque UC comporte, au moins, 10 mots pleins analysés différents.

APPROCHE STATISTIQUE ET PROBLÈME DU SENS

ENCADRÉ 4. – Sélection des segments (UCE) les plus spécifiques des classes.

<i>Unités de contexte sélectionnées dans la classe 1</i>	
107	24 je veux #voyager, dans tout le #monde avoir #plein de #voitures et de #femmes je veux etre riche et avoir #plein de #femmes.
6	18 mes projets seraient de #devenir veterinaire, d' avoir une #belle #maison a la montagne et une au bord de la mer, avoir une #femme un #enfant, une grosse #voiture de #sport et aussi une grosse moto.
7	14 aussi de #sport et aussi de competition. #devenir champion de tir a l' arc, tir a la carabine, #devenir champion du #monde de la chasse a courre, avec les meilleurs chiens, avoir un chenil.
96	13 dans un an je #pars de chez moi, je #passe mon #permis de #voiture et je #trouve un #travail je loue une #maison et je #passe une #belle vie et apres je me marie.
<i>Unités de contexte sélectionnées dans la classe 2</i>	
12	9 ce-que il faudrait ameliorer dans notre #societe pour-que les #jeunes puissent #realiser leurs #projets serait reduire le chomage, #donner plus-de #possibilite aux #jeunes dans les universite, initiation pour les #realiser, changer l' enseignement,
45	9 car la on ne peut pas imaginer tout ce-que on veut, mais pour cela il faut que je travaille beaucoup. dans notre #societe il faut que ca s' ameliore pour-que les #jeunes #realisent leurs #projets, il faut changer la pedagogie des professeurs.
60	9 certains de mes #projets sont #realisables maintenant, ne pas #rester toute ma #vie dans une cage a lapin, ne pas etre un mouton qui se fasse exploiter uniquement pour les autres etre libre meme si je ne gagne pas beaucoup d' argent,
87	9 j' ai ete a l' hopital et la j' ai compris que les #metiers dans le sanitaire c' etait important, essentiel pour la #societe, car ils savent les #vies et evitent les malheurs, c'-est pour cela que j' ai #decide d' etre infirmiere,
<i>Unités de contexte sélectionnées dans la classe 3</i>	
26	32 je voudrais #continuer mes #etudes, mais pas trop longtemps, je veux #passer mon #brevet, et ensuite #voir du #cote d' une #formation #technique, un truc de mecanicien, enfin des #etudes pas longues.
30	28 l' #annee prochaine, je voudrais bien #rentrer en seconde, et puis #continuer mes #etudes, #rentrer a la faculte, ou bien dans une #formation plus #technique, enfin le #minimum c'-est d' avoir mon #baccalaureat, c'-est mon #premier projet,
46	23 moi je #compte surtout avoir mon #independance, pour ca il-y-a pas de mystere, il faut que je travaille par-rapport aux #etudes, #passer mon #brevet, ensuite aller jusqu' a la terminale et preparer une #formation #technique, je-pense a l' informatique,
23	20 l' #annee prochaine j' aimerai #rentrer au lycee pour #continuer mes #etudes, #faire de l' electronique, apres si je #reussi ces #premieres #etudes je #compte #continuer encore deux ans dans l' informatique, etre technicien superieur.