http://www.sylvette-denefle.fr

Sylvette Denèfle, "De la léxicométrie en Anthropologie", L'anthropologue face à la langue, Journal des Anthropologues, n°57-58, 1994, p.62-74

Sylvette DENEFLE
LERSCO UNIVERSITÉ DE NANTES

De la lexicométrie en anthropologie

L'anthropologie, terme que l'on pourrait dans le contexte présent tout aussi bien considéré dans son acception étroite de synonymie avec ethnologie que dans son extension maximale à la manière kantienne, produit essentiellement des textes puisque, même dans le cas où elle appuie son analyse sur des objets, elle les transforme par l'interprétation discursive en signifiants scientifiques. Tous ces textes bien évidemment n'ont pas le même statut épistémologique puisque certains constituent les données ethnographiques, d'autres une analyse scientifique distanciée de ces données, d'autres encore appartiennent au métalangage épistémologique. Mais on peut dire qu'historiquement toute cette production discursive rencontre la difficulté inhérente aux sciences humaines de l'objectivation. L'analyse se fait sur des textes écrits ou transcrits et doit prendre en compte les subjectivités (chercheur, informateur, transcripteur, etc...) qui interviennent à des stades divers du travail de recherche : enquête, recueil de données, transcription des données, analyse, interprétation, etc...

Se posent donc, dans un propos très général, les problèmes des contextes d'enquête, des choix des interlocuteurs, de la

subjectivité du chercheur, de ses inscriptions théoriques, des traductions, etc...
Très souvent, ces questions ont fait l'objet de la réflexion épistémologique et toujours est apparue la nécessité du contrôle des subjectivités dans la mesure où on ne peut les éliminer. À ma connaissance, les techniques de ce contrôle restent elles-mêmes très liées au professionnalisme des chercheurs, c'est-à-dire à des critères encore bien subjectifs. C'est d'ailleurs, en fin de compte, ce butoir méthodologique qui distingue radicalement nos sciences des sciences dites exactes et pour longtemps encore vraisemblablement.
Cependant, si sur l'ensemble des pratiques anthropologiques cette situation semble devoir durer, il est peut-être certains aspects de la recherche qui peuvent bénéficier de techniques objectivantes.
Avec le développement des techniques de l'audiovisuel, par exemple, une nouvelle façon de recueillir les données est advenue qui, si elle n'est pas exempte de subjectivité bien sûr, n'en renouvelle cependant pas moins les problématiques.
De même, il me semble que, dans le champ de la parole transcrite ou de l'écrit de façon très générale, l'outil informatique peut apporter un renouvellement de perspectives conséquent. L'analyse informatisée des textes qui connaît un développement très important depuis une petite dizaine d'années peut grandement modifier la façon de travailler des chercheurs. Cette technique qui s'applique à tous les types de textes et intéresse, en ce sens, toutes les sciences humaines, me semble devoir servir tout particulièrement l'anthropologie.

C'est pourquoi je voudrais avancer quelques réflexions qui pourraient très probablement porter sur les différentes strates du discours anthropologique mais que je limiterai, pour l'exemple, à celui du matériau de base, le texte d'enquête. L'analyse statistique informatisée des textes permet, me semble-t-il, de traiter, à l'aide d'un outil relativement objectivant, des données d'enquête, de quelque

nature qu'elles soient, à partir du moment où elles sont transformées en textes, dans le sens le plus large qu'on puisse donner à ce mot : discours parlés, écrits, séquences gestuelles, bruits, mimiques etc... qui deviennent "textes" par la transcription qu'en fait le chercheur.
Je mesure parfaitement l'aspect paradoxal de l'introduction de méthodes considérées, à mon sens à tord, comme exclusivement quantitatives dans des réflexions sur l'intersubjectivité fine du rapport entre un chercheur et ses informateurs, cependant je souhaite rapporter mon expérience dans l'utilisation de cet outil pour montrer, du moins je l'espère, de quelle aide précieuse il peut être.

Analyse des données textuelles

Un mot pour rappeler en quoi consiste l'outil. Depuis une vingtaine d'années, des travaux ont été entrepris par des linguistes et des statisticiens sur l'analyse lexicale du discours à l'aide de la méthode d'analyse statistique des données mise au point en France dans les années 60 par J.P.Benzécri[1].En fait, l'analyse statistique multidimensionnelle apparaît dès le XIXème siècle mais son renouvellement fondamental sera lié au développement de l'informatique car elle nécessite le maniement d'un nombre énorme d'opérations. En effet, l'analyse multidimensionnelle traite de toutes les relations liant toutes les variables retenues comme significatives entre elles.
Et lorsque on en viendra, sous l'influence des linguistes d'ailleurs, à retenir comme variables significatives tous les mots d'un texte, débutera réellement ce qu'on appelle l'analyse des données textuelles.
Avec les possibilités de calcul des ordinateurs et les méthodes statistiques d'analyses multidimensionnelles, on se

[1]Le premier exposé de la méthode appelée "analyse des correspondances" fut donné en 1963 par J.P.Benzécri dans un cours au Collège de France.

trouve donc dans la possibilité d'analyser statistiquement tous les textes en eux-mêmes ou dans leur contexte d'enquête (indiqué par les variables retenues).
Il y aurait donc bien là un outil nouveau susceptible de modifier complètement notre rapport à l'analyse de textes, à la condition toutefois, et c'est là l'objet de bien des polémiques aujourd'hui, que cette analyse soit productrice d'autres choses que de comptages de mots et permette de percevoir du sens.

Si l'on veut bien dépasser le rejet viscéral de l'intervention du machinisme dans les choses de l'esprit, je proposerai d'examiner le résultat de ces pratiques[2].
Les différents logiciels permettent d'établir des dictionnaires sur le vocabulaire ou sa fréquence, permettent une lecture thématique du texte, en suivant un terme par exemple, ou une comparaison entre différentes parties du texte (déterminées par le chercheur) ou d'appréhender les contextes d'utilisation d'une forme et lorsqu'on a choisi des référents significatifs, il ouvre à une analyse statistique multidimensionnelle de ces éléments et permet d'établir des graphes factoriels qui peuvent faire ressortir des rapports non immédiatement perceptibles entre les notions. De façon plus intéressante encore, ils permettent de mettre en relation un texte ou un ensemble de textes avec toutes les variables que l'on souhaite et de déterminer qui dit quoi, quel est le discours type des femmes, des personnes âgées, des

[2] Les premiers logiciels d'analyse des données textuelles ou calquaient leur travail sur l'analyse des données numériques ou répondaient aux exigences des linguistes et proposaient l'établissement de dictionnaires de vocabulaire et une statistique relativement élémentaire de fréquence. A l'heure actuelle, les chercheurs en anthropologie peuvent disposer de logiciels beaucoup plus spécifiquement élaborés pour traiter les textes du type : relations d'observation, transcription d'entretiens, récits de vie, etc...
Ces outils permettent non seulement une analyse lexicométrique élémentaire mais encore des mises en évidence de champs sémantiques complexes. Pour ma part, j'ai utilisé trois de ces outils : les logiciels Hyperbase, Spad.t et Alceste, mais il en existe d'autres. Hyperbase est dû à M.Brunet (Université de Nice), Spad.t à M.Lebart (CISIA) et Alceste à M.Reinert (CNRS Toulouse).

cadres, etc... quels thèmes sont abordés par qui, quels champs sémantiques se dégagent du vocabulaire de tel ou tel individu, etc... Des méthodes de lexicométrie très fines permettent de faire apparaître les ensembles de mots répétés (ce qui permet entre autre de traiter la négation) et de les croiser également avec les variables jugées utiles, etc... Enfin, certains logiciels dont le type est "Alceste" se distinguent en abordant systématiquement les textes par la recherche de champs sémantiques qui s'obtiennent à partir de méthodes classificatoires spécifiques. Ils livrent ainsi des classes de mots et les variables qui y sont associées à la réflexion du chercheur. L'idée en est qu'à chaque discours appartient des champs sémantiques propres que le vocabulaire et la syntaxe reflètent. Pour illustrer cette hypothèse générale, citons Max Reinert : "Dans une petite expérience d'association d'idées auprès de ses deux nièces, A.Binet avait remarqué que si l'une cherchait sa source d'inspiration dans son imagination, l'autre, peut-être moins inspirée, se contentait de désigner les objets qui l'entouraient... Essayons d'imaginer l'évolution du vocabulaire des réponses de cette dernière si le lieu des séances avait varié. Un changement de pièce aurait induit un changement de vocabulaire et il est possible d'imaginer qu'une simple étude sur la distribution du vocabulaire durant les différentes séances aurait pu discriminer les différentes pièces du logement de Binet. En effet, à chacune d'elle serait associé le vocabulaire en caractérisant les objets...".

L'outil informatique permet donc, si on accepte l'hypothèse de la possibilité de percevoir le sens d'un discours à travers le vocabulaire utilisé et l'organisation du propos, de faire apparaître des champs sémantiques spécifiques au texte étudié. On en vient donc, d'une certaine façon, à une analyse de contenu sans qu'intervienne, à toutes les étapes, la subjectivité du chercheur.

Avant d'en revenir à cet aspect important et à la question de l'intérêt et de la fiabilité de telles méthodes, j'illustrerai

les explications qui précèdent par quelques résultats issus de ma recherche.

Quelques exemples

Les travaux qui m'occupent actuellement concernent les systèmes idéologiques, moraux et métaphysiques qui sous-tendent éventuellement les conceptions et les choix de vie de personnes qui, en France, se déclarent "sans-religion". Il me parait superflu d'insister sur la nécessité de mettre en oeuvre dans ce projet les méthodes qualitatives les plus diverses et de souligner le rendement assez limité du questionnaire ou du sondage.
Aussi, ai-je entrepris, dans l'Ouest de la France, plusieurs séries d'entretiens ainsi que des histoires de vie pour tenter de cerner d'éventuelles régularités dans les choix et les expressions axiologiques de personnes, sociologiquement différenciées, mais partageant toutes l'affirmation de leur non-inscription dans une Église. Plusieurs centaines de personnes, dont les propos ont été enregistrés, ont été interrogées et je me trouve, en conséquence, devant une production de textes tout à fait considérable à analyser. Or, il se trouve que l'outil informatique m'a apporté une façon nouvelle d'appréhender mes données que je voudrais évoquer sommairement en indiquant quelques aspects de la démarche et des ouvertures qu'elle donne.

Ainsi, par exemple, travaillant sur les valeurs, je me suis intéressée à l'utilisation des formes graphiques "bien" et "mal". La lecture déconstruite que permet Hyperbase m'a ainsi montré que le terme "bien" était utilisé un nombre très supérieur de fois à celui de son opposé. Mais il m'a permis de corriger et de prendre conscience par la mise en concordance de ces termes avec leur contexte que "bien" était non

seulement utilisé au sens axiologique mais aussi comme tic verbal par certains de mes informateurs. Charge à moi de prendre en compte ces distinctions ou de repenser les difficiles problèmes de transcription que tout anthropologue a expérimentés. On voit, sur cet exemple simple, l'exigence de transcription que supposent les méthodes évoquées, ce qui est à la fois une contrainte mais surtout, de mon point de vue, une incitation à plus de rigueur dans le recueil et la préparation des données.

Dans une autre utilisation, Spad.t m'a permis de déterminer, parmi les réponses à une même interrogation sur l'appartenance religieuse, que femmes et hommes se distinguaient et de mettre en évidence les caractéristiques biographiques des personnes interrogées par rapport à ces réponses types. Ainsi, j'ai pu montré que le type d'éducation religieuse reçue pendant l'enfance était un facteur déterminant des choix métaphysiques de l'adulte.

Sur des textes d'entretiens, le logiciel Alceste a fait ressortir des thèmes abordés dans l'enquête, ce qu'après tout une analyse classique permet de faire également, mais aussi des comportements très différenciés par rapport à la démonstration qui n'apparaissaient nullement à la lecture des entretiens et n'avaient pas été perçus comme déterminants. Les personnes "sans-religion" se distinguent selon leur appartenance sociale et surtout selon leur niveau d'instruction par des justifications de leur non-croyance qui font une place plus ou moins importante à la référence à la science. Et, comme il n'était pas si évident de le voir, il montre que plus le niveau d'instruction des informateurs est élevé et moins leur justification idéologique s'appuie sur la démonstration ou la référence scientifique.

Le même logiciel me permet de découper mes textes en sous-thèmes et de mettre ces textes spécifiques en relation avec les variables sociologiques, par exemple, que j'ai retenues. Il fait ressortir par ailleurs des formes syntaxiques que je choisis dans la continuité des textes, ce qui m'a permis de suivre les éléments de l'argumentation (saisie à travers des

mots retenus comme "donc" "parce que", "je pense que", "je crois que", etc...) de façon transversale et comparative entre les individus.
Comme il n'est guère facile de présenter en quelques lignes des résultats qui nécessitent la connaissance des logiciels et celle de l'objet de recherche, je prendrai un exemple, évidemment très partiel, qui devrait donner une indication des éléments informatiques à partir desquels on peut renouveler l'interprétation des textes.
Lors de l'analyse des entretiens de personnes "sans-religion", je me suis intéressée, en particulier, à la cohérence des opinions exprimées et à la construction des argumentaires justifiant leurs prises de position ou leurs choix.
L'ensemble des entretiens a été soumis au traitement informatique qui a fait ressortir des thèmes à l'intérieur des propos et je prendrai en exemple, pour cet exposé, celui des attitudes à l'égard des faits que l'on a coutume de désigner comme "paranormaux" ou "liés à l'irrationnel" ou "parasciences" etc...
La première analyse informatique de l'ensemble des textes concernés fait apparaître des champs de regroupement d'idées qui articulent l'attitude des personnes interrogées à propos du paranormal à leurs choix sur les questions métaphysiques à proprement parler et à leurs opinions sur la religion face à leurs choix moraux ou sociaux et à l'histoire de leur vie.
Pour analyser les attitudes à l'égard du paranormal plus précisément, j'ai constitué les éléments biographiques en variables nominales et confronté les propos concernant la religion et la métaphysique à ceux qui portait sur le paranormal. Ce deuxième traitement informatique fait cette fois apparaître quatre thèmes dont le caractère argumentatif devient plus précis : une attitude très positive à l'égard de la science, une exigence d'"expérimentation" dans le fondement de la croyance, une conception très ambivalente des faits paranormaux et un rejet de la référence à la religion institutionnelle.

Enfin, pour approfondir ces rapports ambivalents au "paranormal", j'ai réitéré le déplacement des propos, cette fois sur les questions religieuses et métaphysiques, en variables nominales. Le troisième traitement informatique, portant spécifiquement sur les attitudes à l'égard du paranormal, fait ressortir quatre conceptions très fortement déterminées : un rejet d'une lecture en termes de "surnaturel" des phénomènes de guérison par les médecines "traditionnelles" ou les miracles pour une lecture naturaliste et psychologisante, une attitude de doute favorable à l'égard des extraterrestres et du spiritisme, une dichotomie opposant en astrologie l'horoscope et les thèmes astraux et enfin un rapport exigeant à l'expérience.
Cette première analyse oppose assez nettement à l'intérieur d'une conception "scientiste" partagée par les gens lettrés les plus jeunes et les plus âgés et ceux qui sont issus d'une famille catholique à ceux dont la famille entretient des liens très lâches avec la religion. Elle oppose par ailleurs cet ensemble de gens favorables à la science à un autre groupe beaucoup plus féminin, d'âge et de niveau d'instruction moyens et d'origine et de position sociale modestes qui privilégie l'argumentation faisant référence à l'expérience vécue.

Pour concrétiser un peu plus les sources des résultats qui précède, on peut examiner les fichiers qui suivent. Ils sont issus du traitement des textes dont nous venons de parler par le logiciel Alceste. Les indications des deux premières lignes et de la dernière et des six premières colonnes sont des indices statistiques qui permettent de contrôler la qualité du traitement et sa fiabilité pour le projet de recherche. La dernière colonne qui est celle du vocabulaire permettra au lecteur, même non averti, de constater qu'on est en présence d'un champ sémantique qui traverse les textes analysés et que l'on reconnaîtra, me semble-t-il sans ambiguïté, comme l'astrologie. Les indications de la fin de colonne concernent les variables retenues. Ainsi, "ent_28b" affecté du chi2 "67,92" signifie que l'entretien 28 contribue très fortement

à la constitution de cette classe sémantique qui porte sur l'astrologie. "sm_1" avec le chi2 "119,97" indique que les hommes contribuent beaucoup au discours sur l'astrologie, etc...
Attention toutefois, les exemples très rapides que je viens de donner ne doivent pas laisser penser qu'une interprétation mécanique peut être faite de ces fichiers. Il est bien sûr nécessaire de maîtriser la signification statistique des indices pour ne pas tomber dans le ridicule d'interprétations farfelues. La méthode est exigeante, c'est la condition de sa fiabilité.

```
 ------------------------------------------------
    4eme classe D ; Nombre d'u.c.e. :  299. soit : 28.50 %
 num   effectifs   pourc.    chi2 identification

   37     6.   6.  100.00   15.14      0 prevoir.
   50     6.   7.   85.71   11.32      0 valoir.
   60    14.  20.   70.00   17.23     X  aime+
   75    17.  18.   94.44   39.08     X  astrologi+
  155    22.  34.   64.71   22.60     X  destin+
  240    16.  21.   76.19   23.91     X  horoscope+
  328    28.  35.   80.00   47.12     X  ouais
  331    12.  24.   50.00    5.57     X  parce+
  350     6.  12.   50.00    2.75     X  planet+
  410    14.  17.   82.35   24.59     X  signe+
   20    10.  17.   58.82    7.80     X0 lire.
   73     7.   7.  100.00   17.68        ascendant+
   76     9.  10.   90.00   18.74        astr+
   89     5.   5.  100.00   12.60        belier
  104     5.   6.   83.33    8.90        caractere+
  122     3.   4.   75.00    4.26        ciel
  156     4.   5.   80.00    6.54        determin+
  190     5.   5.  100.00   12.60        exact+
  205     7.   7.  100.00   17.68        famille
  251     4.   5.   80.00    6.54        influenc+
  253     6.   9.   66.67    6.49        intere<
  258     5.   8.   62.50    4.57        journal+
  266     5.   7.   71.43    6.37        ligne+
  274     4.   7.   57.14    2.84        lune<
  278     5.   7.   71.43    6.37        main+
  308     8.   8.  100.00   20.22        naiss+
  313     5.   5.  100.00   12.60        nee
  338     3.   4.   75.00    4.26        pauvre+
  339     6.   6.  100.00   15.14        pays<
  345     4.   4.  100.00   10.07        pfouu+
  411     6.   6.  100.00   15.14        silence
  437     8.   8.  100.00   20.22        theme+
  454     3.   4.   75.00    4.26        vierge+
  493 *   6.   8.   75.00    8.55 *      ASTROLOGIE
  563 *   7.   8.   87.50   13.77 *      HOROSCOPES
  699 *  40. 112.   35.71    3.20 *    2 non
  741 *  30.  67.   44.78    9.30 *    5 plus
  769 *   4.   7.   57.14    2.84 *    6 d-ailleurs
  891 *  94. 183.   51.37   56.86 *      *age_1
  897 *   8.   8.  100.00   20.22 *      *choix_1
  900 * 235. 687.   34.21   31.78 *      *choix_nr
```

```
  905 * 101. 265.   38.11   16.07 *      *csp_6
  906 * 119. 273.   43.59   41.22 *      *csp_84
  908 * 106. 222.   47.75   51.17 *      *edrel_0
  909 *  36.  56.   64.29   37.17 *      *edrel_1
  915 *   8.   8.  100.00   20.22 *      *ent_14b
  919 *  46. 103.   44.66   14.63 *      *ent_21b
  924 *  16.  24.   66.67   17.55 *      *ent_28a
  925 *  63.  98.   64.29   67.92 *      *ent_28b
  927 *  27.  33.   81.82   47.52 *      *ent_36a
  942 *  28.  40.   70.00   35.14 *      *ent_9b
  945 * 134. 326.   41.10   36.85 *      *instr_3
  951 * 149. 347.   42.94   53.03 *      *orig_1
  954 *  85. 150.   56.67   68.12 *      *para_0
  955 *  27.  33.   81.82   47.52 *      *para_0_25
  956 *  28.  40.   70.00   35.14 *      *para_0_37
  966 *   8.   8.  100.00   20.22 *      *para_m0_28
  973 *  16.  24.   66.67   17.55 *      *para_m0_66
  977 *  14.  22.   63.64   13.61 *      *rej_2
  978 *  46. 103.   44.66   14.63 *      *rej_3
  980 * 120. 249.   48.19   62.11 *      *relfam_1
  984 * 215. 668.   32.19   12.24 *      *sex_2
  985 * 230. 526.   43.73  119.97 *      *sm_1
 -> Nombre de formes retenues :         131; Chi2 moy. :   14.57578
```

Le fichier qui précède n'est bien sûr pas le seul qui soit produit. Le même traitement nous donne aussi, toujours sur la classe concernant l'astrologie, les fichiers qui suivent (et bien d'autres). Ils indiquent respectivement les couples de mots contribuant le plus au champ sémantique en question, puis les segments répétés et enfin les phrases les plus représentatives du champ. On constatera qu'on est très éloigné, avec ce matériel, des dictionnaires et qu'on est bien engagé du côté de la sémantique.

```
4eme classe D ; Nombre d'u.c.e.:263. soit:27.45 %
 num   effectifs   pourc.    chi2 identification

   58    12.  14.   85.71   24.21        aime+-bien+
   64     4.   4.  100.00   10.61        astrologi+-mais
  259     4.   4.  100.00   10.61        croire.-plus
  429     4.   5.   80.00    6.97        c-est-vrai+
  454     6.  11.   54.55    4.10        c-est-vrai-que
  352     4.   4.  100.00   10.61        dans-famille
  625    12.  15.   80.00   21.13        je-aime+
  501     3.   4.   75.00    4.56        mais-astrologi+
  518     5.   7.   71.43    6.85        mais-pour
  532     3.   4.   75.00    4.56        meme<-signe+
  335    14.  34.   41.18    3.33        non-je
  332     4.   6.   66.67    4.66        non-mais
  219    13.  25.   52.00    7.77        pas-tout
  139     6.   6.  100.00   15.96        theme+-astr+
 1001 *  79. 162.   48.77   44.47 *      *age_1
 1007 *   7.   7.  100.00   18.63 *      *choix_1
 1010 * 205. 633.   32.39   22.79 *      *choix_nr
 1016 * 103. 247.   41.70   33.92 *      *csp_84
 1018 *  86. 192.   44.79   36.25 *      *edrel_0
 1019 *  33.  50.   66.00   39.36 *      *edrel_1
 1025 *   7.   7.  100.00   18.63 *      *ent_14b
 1031 *  14.  21.   66.67   16.58 *      *ent_24a
 1034 *  16.  23.   69.57   20.98 *      *ent_28a
 1035 *  50.  82.   60.98   50.60 *      *ent_28b
```

```
1037 *  20.  26.   76.92   32.84 *      *ent_36a
1052 *  26.  37.   70.27   35.43 *      *ent_9b
1055 * 115. 294.   39.12   28.97 *      *instr_3
1061 * 131. 317.   41.32   45.78 *      *orig_1
1064 *  72. 131.   54.96   57.66 *      *para_0
1065 *  20.  26.   76.92   32.84 *      *para_0_25
1066 *  26.  37.   70.27   35.43 *      *para_0_37
1070 *   9.  14.   64.29    9.68 *      *para_0_71
1076 *   7.   7.  100.00   18.63 *      *para_m0_28
1083 *  16.  23.   69.57   20.98 *      *para_m0_66
1087 *  14.  21.   66.67   16.58 *      *rej_2
1088 *  42.  97.   43.30   13.61 *      *rej_3
1090 *  98. 215.   45.58   45.74 *      *relfam_1
1095 * 201. 481.   41.79   99.67 *      *sm_1
->Nombre de formes retenues:65; Chi2 moy.:16.97640
```

****** classe numero 4 ******

```
  11  je aime+ bien+
   8  je y croire.
   7  ne savoir. pas
   7  je ne savoir.
   6  non je ne
   6  je ne savoir. pas
   5  je penser. que
   4  ne croire. pas
   4  ne penser. pas
   4  ne y croire.
   4  ne ai pas
   4  c-est peut-etre plus
   4  je savoir. pas
   4  je vouloir. dire.
   4  je ne croire. pas
   4  je ne penser. pas
   4  je ne y croire.
   4  je y croire. plus
   4  y croire. pas
   4  y croire. plus
   3  non je ne penser. pas
   3  c-est-vrai que c-est
   3  moi je aime+
   3  ca je y croire.
   3  ca je y croire. plus
   2  que c-est peut-etre plus
   2  non je ne y croire.
   2  non je y croire. pas
   2  quand je etais petit+
   2  c-est-vrai que c-est peut-etre
   2  c-est-vrai que c-est peut-etre plus
   2  mais c-est-vrai que c-est peut-etre
   2  par exemple+ je
   2  je y croire. pas
   2  je etais petit+
   2  moi je aime+ bien+
   2  moi je suis
   2  ca je aime+ bien+
```

****** CLASSE NUMERO : 4 ******

```
181 4161je crois a l'influence du theme astral sur la/ vie d'un/ homme.
199 4161je ne l'ai pas pratiquee, mais, / je crois/ surtout a l'influence du
199 4162theme astral sur lequel on vit.
587 4161(silence); comme je/ disais que je ne croyais pas que notre destin
587 4162etait ecrit, il/ ne peut pas etre ecrit dans notre main non plus!
```

```
 942 4161ouais, c-est-vrai que/ j'aime bien lire les horoscopes./ ET SI
1840 4161croire c-est un/ grand mot, / mais je m'y interesse, j'aime bien.
1537 4141c-est/ loin-d'etre/ evident. alors non, les horoscopes dans les
1537 4142journaux, non, je/ n'y crois/ pas, j'aime bien lire, ca m'amuse mais
1537 4143c-est tout.
1731 4141l'homme doit suivre son destin, les grandes/ lignes de sa/ vie sont
1731 4142deja tracees des sa naissance,
  15 4131/ bin c-est un petit peu pareil; mais je me suis pas trop/ interessee
  15 4132a/ cela;
1082 4131bien sur fonder/ une famille, avoir plein d'enfants parce-que j'adore
1082 4132les gosses/;
 182 4121par exemple, au niveau de mes enfants, ma premiere fille/ est vierge/
 182 4122ascendant belier, ma deuxieme est belier ascendant vierge.
 591 4111/ bin l'astrologie ca j'y croirais plus! (rire) pas a/ l'horoscope,
 591 4112mais a l'astrologie ouais;
 781 4111tout le/ monde a les memes chances d'aboutir, mais il y en a quand
 781 4112ils/ naissent dans une famille pauvre ou dans une famille riche,
1059 4111/ je sais qu'on en avait parle, parceque quand j'etais petite, / mes
1059 4112parents avaient voulu que je regarde une emission, enfin,
1077 4111/ ah, ouais, ouais, ouais; elle prend ca tres bien au/ contraire.
1088 4111pour la beaute des/ paysages aussi bien sur; et puis apprendre a
1088 4112connaitre tous/ les gens differents de nous aussi bien sur, je ne sais
```

Ces exemples bien trop partiels et plutôt délicats à livrer dans un texte général devraient permettre de saisir l'intérêt des méthodes informatiques pour le traitement des textes. Qu'on ne s'y trompe pas, si l'outil permet de traiter des textes de dimensions considérables, il n'est nullement mécanique et il me semblerait très inexact de croire qu'il est un gain de temps. Son intérêt réside dans une nouvelle façon de considérer des données qualitatives et dans la possibilité d'un traitement des textes qui soit relativement objectivant. Il y a dans ces méthodes une ouverture de recherche qui me semble fondamentale mais nous sommes fort éloignés encore de systèmes fermés et finis.

Pour conclure

Ouvrons un bref instant la boite de Pandore des points de vue sur les possibilités de faire, par le biais d'une machine, une lecture des données d'un discours.

Certains rejetteront, sans autre procès, l'idée que la subtilité du discours qui passe par les intonations, les hésitations, l'utilisation de vocabulaires euphémisants, voire d'expressions utilisées ironiquement et signifiant le contraire même de ce que le vocabulaire semble indiquer, sans parler même des synonymes ou homonymes si difficiles à traiter hors de tout contexte, puisse traverser l'utilisation simplificatrice des méthodes informatiques. D'autres mettront en avant la difficulté à trouver les mots porteurs de sens dans l'analyse de contenu et évoqueront les délicats problèmes de la transcription et de la lemmatisation des textes (que les outils ne font pas nécessairement d'ailleurs). D'autres encore objecteront qu'une machine ne produit que par l'activité de son utilisateur (point de vue irréfutable me semble-t-il) et que l'on trouvera à l'arrivée ce que l'on savait au départ. Ce à quoi les plus pratiques rajouteront qu'on perd beaucoup plus de temps à la manipulation informatique qu'à la lecture réfléchie de textes et que si l'outil ne doit servir qu'à confirmer ce que l'on savait initialement, ils ne voient pas très bien où est l'intérêt.

Il est impossible, comme chacun sait, de refermer la boite de Pandore. Aussi mon propos ne peut être convaincant pour qui a déjà étayé ses conceptions.

Toutefois, je voudrais insister sur les incidences, à mon sens positives, que génère l'analyse lexicométrique.

Les capacités de la machine à gérer des variables en nombres aussi grands qu'on le souhaite, suppose que l'on mène une réflexion importante sur le choix des contextes d'enquête puis sur les variables qui peuvent être porteuses d'interprétations intéressantes. Chaque chercheur procède bien évidemment de cette façon mais je veux dire que la rigidité de l'outil implique un soin très grand porté à la préparation des protocoles d'enquête, ce qui ne peut qu'ajouter à la qualité de la recherche. L'outil peut, donc, modifier dans le sens de la précision et de la rigueur le recueil des données lui-même.

Puis, l'utilisation d'un traitement informatique renouvelle la question de la transcription des textes oraux. Aucun signe ne

doit être laissé au hasard. La précision et surtout l'homogénéité après des choix explicites des procédés de retranscription s'imposent. La qualité des données ne saurait, encore une fois, en pâtir en tout état de cause.
Les exigences et les contraintes de l'outil poussent à la rigueur dans l'établissement des données certes, mais ne serait-ce pas qu'un surcroît de travail si les résultats de ces efforts restent médiocres ou insignifiants ?
Il me semble, sur ce point essentiel, que ceux qui ont pratiqué de telles méthodes[3] me suivront pour dire qu'elles ouvrent véritablement des pistes nouvelles pour l'interprétation des textes, en faisant apparaître, par le fait du très grand nombre de traitements que permet la machine, des thématiques non immédiatement perceptibles, des rapprochements de termes non perçus à la lecture continue, des croisements de formes graphiques et de variables bien difficiles à gérer manuellement dans le cas d'entretiens et puis surtout et bien trivialement qu'elles permettent de travailler sur des textes de dimensions considérables puisque la machine ne se lasse pas et n'opère pas de glissement sémantique à la lecture de données aussi vastes soient elles.
La question reste donc de savoir si la fiabilité du procédé électronique compense la pauvreté de ses fonctions systématisées. À cette question encore, seule l'expérimentation peut permettre d'apporter des éléments de réponses et il ne me paraît pas utile de dévider trop de généralités sur ces problèmes. Je souhaite seulement les évoquer et revenir à l'objectif qui était le mien en écrivant ces lignes.
L'anthropologie produit des textes, et souvent des textes oraux, très nombreux et de très grandes dimensions.
L'évolution scientifique et technique amène un outillage qui renouvelle potentiellement le traitement de ces données.

[3] Voir à ce sujet, les travaux des Journées Internationales d'Analyse des Données Textuelles de 1991 et 1993 qui se sont tenues à Barcelone puis à Montpellier et qui ont présenté des exemples nombreux, dans toutes les disciplines, d'utilisation de ces méthodes.

Est-il sage de l'ignorer voire de le rejeter avant de l'essayer parce que nos taches habituelles sont trop lourdes pour nous permettre de nouveaux apprentissages, parce qu'il n'est pas parfaitement adéquate à notre attente ou parce qu'on a choisi une conception de sa discipline de sciences humaines qui fait plus de place à l'humain qu'à la science?

Je ne détiens pas la réponse définitive à ces questions et la diversité des méthodes ne peut, à mon sens, qu'enrichir la recherche, mais comme je plaide pour le développement de méthodes naissantes, je souhaiterais que ce papier permette à ceux qui sont convaincus de l'intérêt de l'analyse des données textuelles de joindre leurs efforts aux miens pour des échanges méthodologiques qu'on peut espérer féconds, que ceux qui sont intéressés suivent la pente de leur curiosité et que ceux qui y voient la perte de toute sensibilité n'oublient pas que les machines reflètent les exigences de leurs utilisateurs.

L'analyse statistique des données textuelles n'est pas un procédé magique qui permet de générer des interprétations de discours en appuyant sur un bouton. Elle ne permet même pas de limiter le temps de travail. Simplement, elle ouvre à des explorations fouillées et renouvelées des données complexes que véhiculent les discours humains.

Bibliographie

Achard, P. "Analyse de discours et sociologie du langage" *Langage et Société* n°37 p.5-60 1986

Benzecri, J.P. *L'Analyse des Données* Dunod Paris 1980

Lebart, L., Salem, A. *Analyse statistique des données textuelles* Dunod Paris 1981

Muller, C. *Principes et Méthodes de Statistique Lexicale* Hachette Paris 1977