Didier Demazière, Claire Brossaud, Patrick Trabal et Karl Van Metter (*dir.*), *Analyses textuelles en sociologie. Logiciels, méthodes, usages*, Rennes, Presses Universitaires de Rennes, 2006, p.156-173

# Le traitement des mots associés à des questions : l'apport du logiciel Trideux

Philippe Cibois
Laboratoire Printemps
philippe.cibois@printemps.uvsq.fr

## *Les concepts de base de la méthode*

Le but de cet article est de montrer comment Trideux rend possible l'exploitation de principes qui avaient été mis au point auparavant (Cibois 1989) à partir de Lebart et Salem (1988) mais qui n'avaient pas trouvé de développement logiciel fiable, ce qui n'est plus le cas désormais dans la version 4 de Trideux. En effet, les données textuelles sont très fréquentes dans les enquêtes de sciences sociales et il est rare que l'on n'ait pas de renseignements sur ceux qui ont produit ces textes. La logique qui est développée dans l'utilisation de Trideux pour les données textuelles est de croiser les renseignements dont on dispose sur les émetteurs avec les mots qu'ils utilisent. Le tableau ainsi constitué, que j'ai appelé "tableau lexical des questions" peut être soumis au traitements usuels : analyse des correspondances, recherche d'attractions entre mots et renseignements sur les émetteurs, recherche de similitudes entres mots.

Pour expliquer les concepts de base de la méthode je partirais d'un exemple fictif très simple. Soit une enquête où l'on pose trois questions fermées : le sexe, l'âge (trois tranches d'âge) et le groupe socio-professionnel (Supérieur, Intermédiaire, Inférieur). On demande ensuite, dans une question ouverte, à chaque personne interrogée qu'elle définisse elle-même sa classe sociale. On suppose les résultats suivants obtenus pour 4 individus.

**Données d'origine (fictives)**

| Num | Sex | Age | Gsp | Txt |
|---|---|---|---|---|
| n1 | 1 | 3 | 1 | Classe moyenne profession intellectuelle |
| n2 | 2 | 1 | 2 | Classe intellectuelle exploitée |
| n3 | 1 | 2 | 3 | Classe ouvrière exploitée |
| n4 | 2 | 1 | 1 | Profession intellectuelle supérieure |

Nb individus=4  Nb de mots=13  Nb de mots différents=7

Le tri à plat des variables nous donne la distribution des questions fermées

```
          Tri à plat des variables

Nb de questions = 3
Question SEX
Tot.    1    2
     Masc Fémi
   4    2    2
 100 50.0 50.0

Question AGE
Tot.     1         2    3
     Jeune AgeMédian  Agé
   4     2         1    1
 100  50.0      25.0 25.0

Question GSP Groupe Socio Professionnel
Tot.      1        2      3
     GSPsup GSPinter GSPinf
   4      2        1      1
 100   50.0     25.0   25.0
```

La manière la plus courante de construire un tableau pour traiter les mots émis est de n'utiliser qu'une seule des caractéristiques des individus, leur identité : c'est ce qu'on fait dans ce qu'on appelle le *Tableau lexical* où à l'intersection entre individu et un mot, on indique (en zéro/un) si l'individu à utilisé ou non le mot[1].

Voici le Tableau lexical associé aux données fictives :

```
          Tableau lexical

              Individus
           | 1 | 2 | 3 | 4 | Total
-----------|---|---|---|---|-------- -------
superieure | 0 | 0 | 0 | 1 | 1  (fréquence du mot)
ouvriere   | 0 | 0 | 1 | 0 | 1
moyenne    | 1 | 0 | 0 | 0 | 1
profession | 1 | 0 | 0 | 1 | 2
exploitee  | 0 | 1 | 1 | 0 | 2
intellectu | 1 | 1 | 0 | 1 | 3
classe     | 1 | 1 | 1 | 0 | 3
------------------------------ -------
Total      | 4 | 3 | 3 | 3 |13=nb total de mots émis
En col.  Nb de mots émis par individu
```

Quand on dispose de questions fermées et non de questions ouvertes, on construit habituellement un tableau dit de Burt qui croise chaque modalité de l'enquête avec toutes les autres modalités de l'enquête. Ce tableau de Burt est la

[1] On peut utiliser aussi, plutôt que la présence ou l'absence, le nombre de fois où l'on rencontre le mot.

juxtaposition de tous les tableaux deux à deux possibles, dupliqués en inversés par rapport à une diagonale où l'on trouve la distribution de chaque modalité.

Voici ce tableau de Burt pour les données fictives :

**Tableau de Burt habituel**

| Mod | Masc | | Age | | | GSP | | | Total |
|---|---|---|---|---|---|---|---|---|---|
| | Fémi | | − | = | + | Sup | = | inf | (eff modalité x NbQuestions) |
| SEX1 | 2 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 6 |
| SEX2 | 0 | 2 | 2 | 0 | 0 | 1 | 1 | 0 | 6 |
| AGE1 | 0 | 2 | 2 | 0 | 0 | 1 | 1 | 0 | 6 |
| AGE2 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 3 |
| AGE3 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 3 |
| GSP1 | 1 | 1 | 1 | 0 | 1 | 2 | 0 | 0 | 6 |
| GSP2 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 3 |
| GSP3 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 3 |
| | 6 | 6 | 6 | 3 | 3 | 6 | 3 | 3 | 36 (NbIndiv x NbQuestions²) |

Le concept fondamental utilisé dans Trideux consiste à conserver la structure du tableau de Burt pour les colonnes d'un nouveau tableau, et la structure du tableau lexical pour les lignes. A l'intersection d'une modalité et d'un mot, se trouve le nombre de fois où ce mot a été associé dans le corpus à cette modalité. Par exemple dans le *Tableau exical des questions* ci-dessous, le mot "classe" a été utilisé trois fois : deux fois par des personnes de sexe masculin et une fois par un individu de sexe féminin. De ce fait, pour une question donnée, on retrouve en ligne le nombre de fois où le mot a été émis dans le corpus, ce que l'on appelle sa "fréquence".

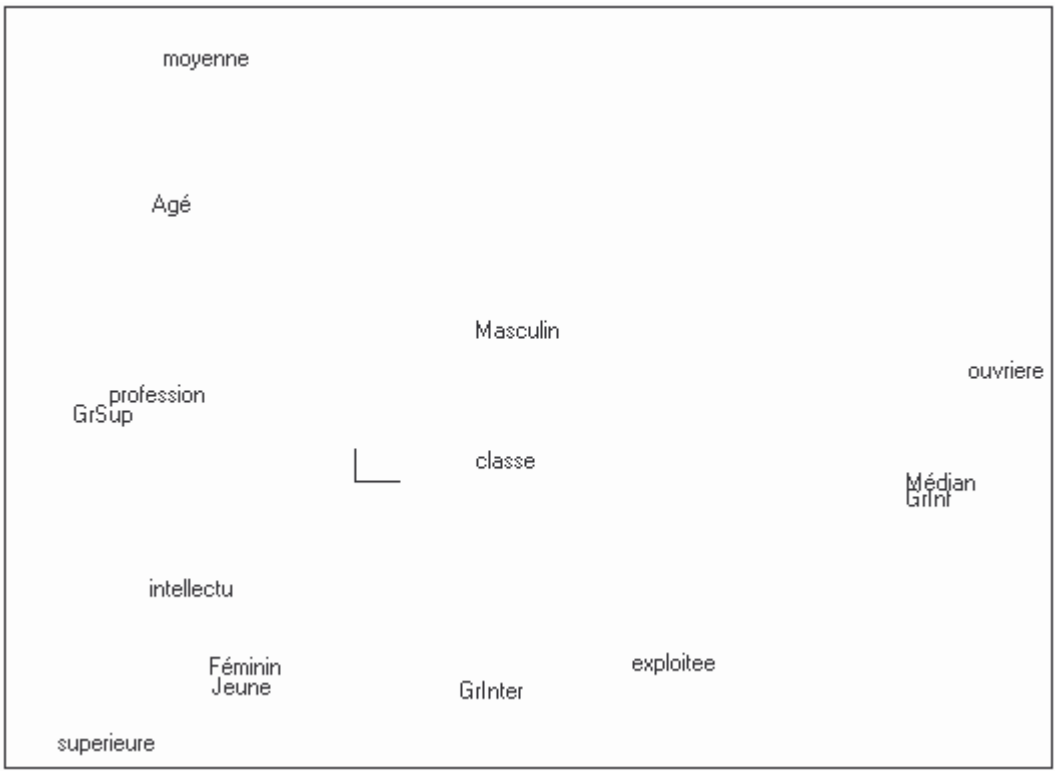
**Tableau lexical des questions**

| Fréq | Masc | | Age | | | GSP | | | Total | Mot |
|---|---|---|---|---|---|---|---|---|---|---|
| | Fémi | | − | = | + | Sup | = | inf | freq x nb questions | |
| 1 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 3 | superieure |
| 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 3 | ouvriere |
| 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 3 | moyenne |
| 2 | 1 | 1 | 1 | 0 | 1 | 2 | 0 | 0 | 6 | profession |
| 2 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 6 | exploitee |
| 3 | 1 | 2 | 2 | 0 | 1 | 2 | 1 | 0 | 9 | intellectu |
| 3 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 9 | classe |
| 13 | 7 | 6 | 6 | 3 | 4 | 7 | 3 | 3 | 39 (NbMots x NbQuestions) | |

Comme la fréquence d'un mot se retrouve dans chaque question, dans le tableau lexical des questions, l'effectif en ligne de chaque mot sera sa fréquence multipliée par le nombre de questions utilisées. Cette propriété est voisine de celle d'un tableau de Burt où l'effectif de l'enquête se retrouve dans chaque sous-tableau. Comme dans un tableau de Burt on utilise l'information contenue dans tous les tableaux juxtaposés qui croisent le vocabulaire et les questions fermées de l'enquête.

Sur un tel tableau on peut : faire une analyse des correspondances ; rechercher les attractions entre lignes et colonne ; les similitudes entre lignes.

1) On peut ainsi faire l'analyse des correspondances d'un tel tableau. On a les résultats suivants :




On voit la position centrale du mot *classe* qui est le mieux réparti de l'ensemble. Le premier facteur oppose la hiérarchie sociale et les mots associés (supérieur, profession, intellectuelle, moyenne pour le groupe supérieur et ouvrière et exploitée pour le groupe inférieur). Le deuxième facteur (vertical) oppose les points *féminin* et *jeune* (qui sont au même endroit), à l'âge le plus élevé.

2) On peut visualiser ces attractions entre mots et modalités en utilisant la technique du PEM (Pourcentage de l'écart maximum : Cibois 1993).

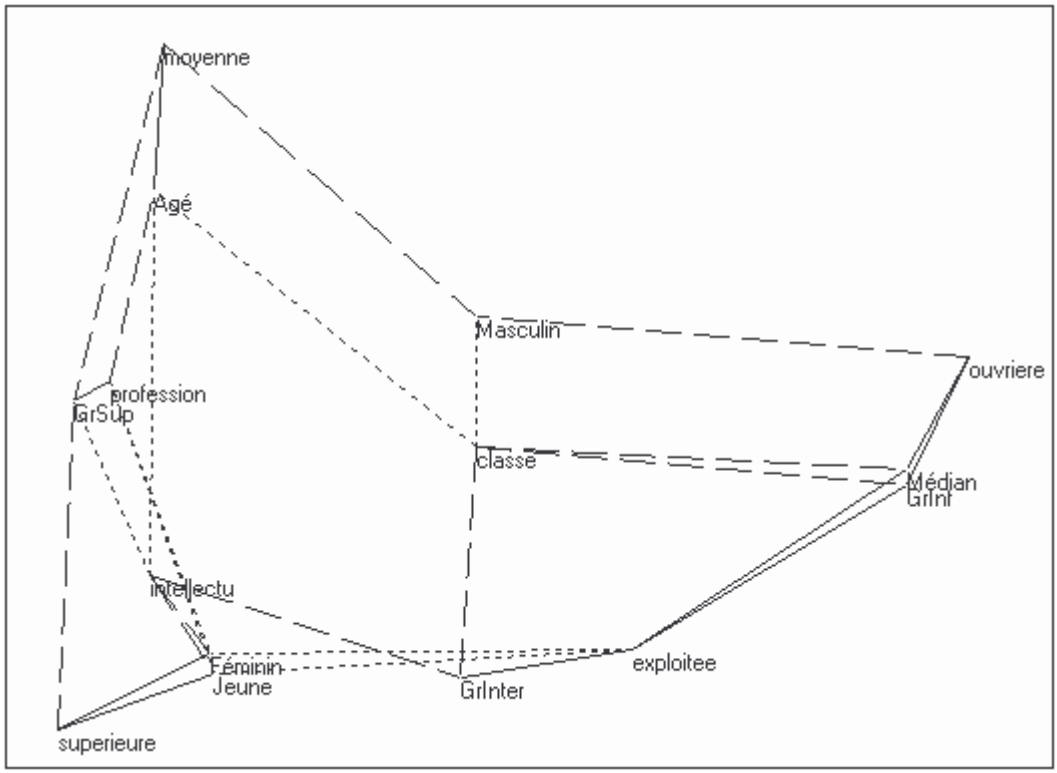
Dans le tableau de base, on isole par exemple une case, celle croisant le mot "profession" et le GSP1 (Groupe social supérieur)

| | SEX1 | SEX2 | AGE1 | AGE2 | AGE3 | GSP1 | GSP2 | GSP3 | TOT. |
|---|---|---|---|---|---|---|---|---|---|
| supe | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 3 |
| ouvr | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 3 |
| moye | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 3 |
| prof | 1 | 1 | 1 | 0 | 1 | 2 | 0 | 0 | 6 |
| expl | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 6 |
| inte | 1 | 2 | 2 | 0 | 1 | 2 | 1 | 0 | 9 |
| clas | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 9 |
| TOT. | 7 | 6 | 6 | 3 | 4 | 7 | 3 | 3 | 39 |

L'effectif observé pour cette case profession GSP1 est de 2
L'effectif théorique est de 6 x 7 / 39 = 1,08
L'écart à l'indépendance (observé – théorique) est de 2 – 1,08 = 0,92

Si la liaison était parfaite, le maximum dans la case serait de 6 (la plus petite des deux marges) et l'écart à l'indépendance serait de 6 – 1,08 = 4,92
Le rapport de l'écart à son maximum, mis en pourcentage donne le PEM
PEM = 0,92 / 4,92 x 100 =18,8%[2]

On calcule tous les PEM et on les visualise sur le fond de carte factoriel. On repère ainsi les attractions les plus fortes.




3) on peut enfin rechercher les similitudes entre lignes : à cette fin on calcule tous les PEM du tableau et on examine toutes les lignes prises deux à deux. Par exemple les lignes des mots "profession" et "intellectuelle" dans les données de base nous manifestent que ces mots sont souvent choisis dans le même contexte.

Dans le tableau ci-dessous, on a le calcul de tous les PEM pour toutes les cases du tableau. On isole la ligne du mot *profession* comme ligne 1 et *intellectuelle* comme ligne 2. Si ces mots avaient toujours le même PEM avec toutes les modalités de l'enquête, cela signifierait qu'ils sont toujours choisis ensemble. Quand les PEM ont de faibles différences, on se rapproche de cette situation de similitude.

[2] On pourra objecter avec raison qu'il faudrait mieux considérer le sous-tableau seul et prendre comme maximum non pas la marge mais la fréquence, ce qui serait plus réaliste et dans ce cas la liaison serait à 100%. Comme tous les PEM sont calculés de la même façon, c'est simplement l'échelle des PEM qui se déplace, ce qui est peu important pour l'interprétation.

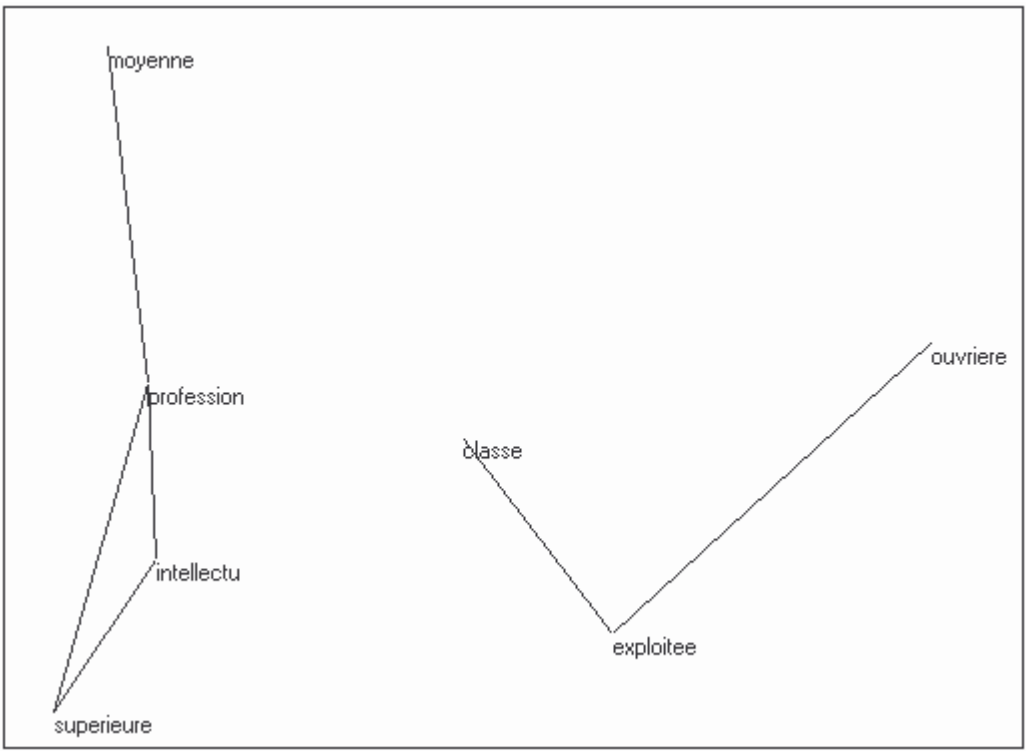
```
    Impression des PEM
           SEX1  SEX2  AGE1  AGE2  AGE3  GSP1  GSP2  GSP3
    supe-100.0  21.2  21.2-100.0-100.0  18.8-100.0-100.0
    ouvr  18.8-100.0-100.0  27.8-100.0-100.0-100.0  27.8
    moye  18.8-100.0-100.0-100.0  25.7  18.8-100.0-100.0
    expl  -7.1   1.5   1.5  21.2-100.0-100.0  21.2  21.2
    clas   7.1 -27.8 -27.8  13.3   2.5 -38.1  13.3  13.3
    ---------------------------------------------------
L1  prof  -7.1   1.5   1.5-100.0  11.4  18.8-100.0-100.0
L2  inte -38.1  13.3  13.3-100.0   2.5   7.1  13.3-100.0
    ---------------------------------------------------
L1-L2     31.0 -11.8 -11.8   0.0   8.9  11.7-113.3   0.0
```

On calcule donc les différences entre les PEM des deux lignes : comme ce qui importe est l'importance des écarts et non leur signe on calcule la somme des valeurs absolues des écarts égale à 188,5, on en prend la moyenne en divisant par le nombre colonnes (c'est-à-dire de modalités) soit 188,5 / 8 = 23,6

Si les deux profils étaient presque identiques, la moyenne serait très faible, on serait dans le cas de la ressemblance (alors qu'un PEM nul indique l'absence de liaison). Inversement si pour chaque différence on avait l'opposition la plus grande : +100 contre -100 soit une différence de 200, l'opposition serait la plus grande possible.

Pour retrouver la logique du PEM où l'attraction correspond aux fortes valeurs et zéro à l'indépendance, on centre la moyenne en calculant un indice des PEM en lignes (noté PEL) et égal à 100 diminué de la moyenne, ici PEL = 100 – 23,6 = 76,4

On peut ainsi visualiser les similitudes entres PEL : on a ici les PEL pour les 7 mots qui se répartissent, pour un seuil donné, en deux groupes intelligibles.



Nous disposons maintenant des trois outils pour examiner des données réelles.

*Le codage de la classe sociale*

D'une exploitation en cours de l'enquête 2003 de l'Insee "histoires de vie"[3] on isole les textes en clair en réponse à la question "si vous avez le sentiment d'appartenir à une classe sociale, comment la définissez-vous". On va constituer un tableau lexical des questions en croisant ce vocabulaire avec deux types de questions fermées :

- les questions définissant le répondant indépendamment de la classe sociale : sexe, âge, niveau d'étude.

- les questions qui correspondent à un codage fait par l'insee de la CS (code à deux chiffres mais aussi groupe à un chiffre), et codage fait par l'insee de la déclaration en clair :

1) réponse oui ou non à la question "avez-vous le sentiment d'appartenir à une classe sociale ?"

2) un premier codage par classe, groupe professionnel ou groupe social

| | |
|---|---|
| 00 | Ne sait pas |
| 10 | Moyenne |
| 20 | Ouvrière |
| 30 | Bourgeoisie |
| 40 | Défavorisée |
| 50 | Privilégiée |
| | **Groupe professionnel** |
| 61 | Travailleurs |
| 62 | Cadres |
| 63 | Agriculteurs |
| 64 | Indépendants |
| 65 | Employés |
| 68 | Fonctionnaires |
| | **Groupe social** |
| 71 | Prolétaires |
| 72 | Peuple |
| 73 | Citoyen |
| 74 | Intellectuels |
| 75 | Immigrés |
| 80 | Autre |

3) un deuxième codage par mode de déclaration de la classe sociale en fonction d'une échelle : sociale, de revenus ou sans précision.

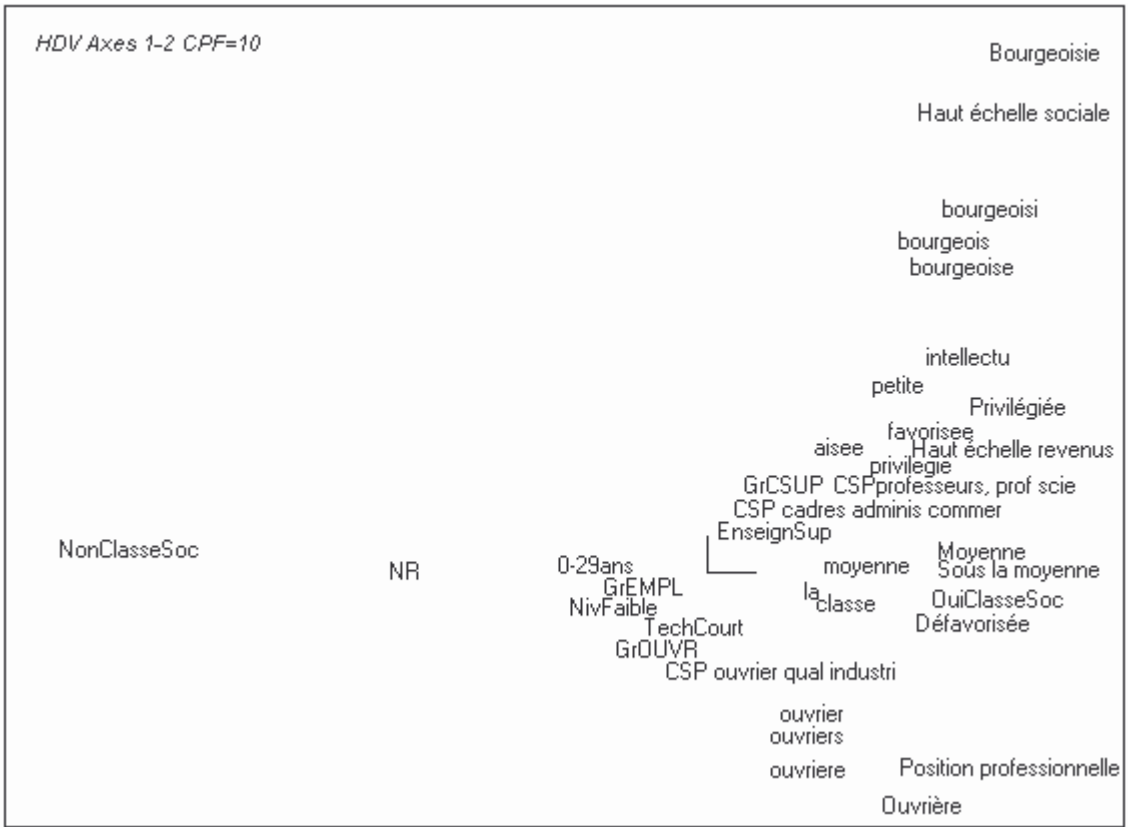
[3] En collaboration avec Jérôme Deauvieau, Laboratoire Printemps, CNRS-Université de Versailles-St-Quentin en Yvelines.

| | |
|---|---|
| 00 | Pas de classe déclarée |
| 11 | En dessous de la moyenne (échelle sans précision) |
| 12 | Moyenne (échelle sans précision) |
| 13 | Au-dessus de la moyenne (échelle sans précision) |
| 21 | Dans le bas de l'échelle sociale (échelle sociale) |
| 22 | Au milieu de l'échelle sociale (échelle sociale) |
| 23 | Dans le haut de l'échelle sociale (échelle sociale) |
| 30 | Sans précision du niveau (échelle de revenus) |
| 31 | Dans la bas de l'échelle de revenus (échelle de revenus) |
| 32 | Au milieu de l'échelle de revenus (échelle de revenus) |
| 33 | Dans la haut de l'échelle de revenus (échelle de revenus) |
| 40 | Position professionnelle |
| 50 | Situation par rapport à l'emploi |
| 60 | Valeurs |
| 70 | Autre mode de déclaration |

Pour 3988 répondants, on a 6633 mots soit des phrases de moins de deux mots par répondant : la saisie n'a recueilli que des phrases très stéréotypées comme le manifeste le début du corpus.

CLASSE SUPERIEURE
LES TRAVAILLEURS
PRIVILEGIEE
CLASSE PRIVILEGIEE
INTERNATIONALE
BOURGEOISIE
MICRO BOURGEOISIE
MOYENNE BOURGEOISIE
PETITE BOURGEOISIE
BOURGEOISIE
CLASSE AISEE

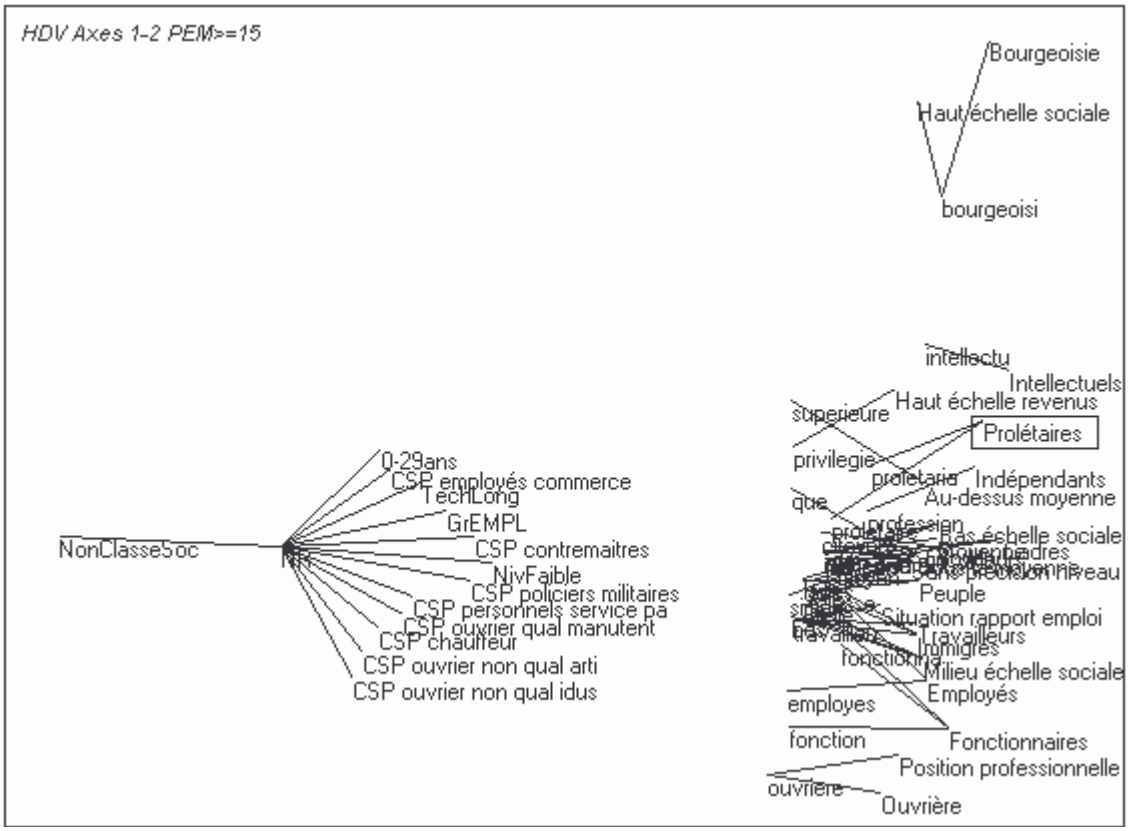
On va constituer un tableau lexical entre les 115 mots qui sont apparus dans le corpus au moins quatre fois et les 70 modalités de réponses aux questions fermées.



On a le premier plan factoriel ci-dessus où l'on a retenu que les points dont la contribution est supérieure à 10 pour mille (CPF=10). Le vocabulaire émis par les répondants est entièrement en minuscule et les questions fermées commencent toutes par une majuscule. On voit d'abord que le premier facteur oppose ceux qui n'ont pas de perception de classe sociale (par défaut, ils ont NR, non réponse comme mot émis) et les autres. Le deuxième facteur vertical est lié à une hiérarchie sociale, élevée en haut, inférieure en bas.

En haut par exemple on voit que les divers mots proches ont été codés *Bourgeoisie* comme classe et comme haut de l'échelle sociale, ce qui n'est pas surprenant. Inversement, en bas, divers mots proches ont été classés *Ouvrière* comme classe avec un mode de classement utilisant la position professionnelle. Ce plan est sans surprise et sans guère d'intérêt apparent.

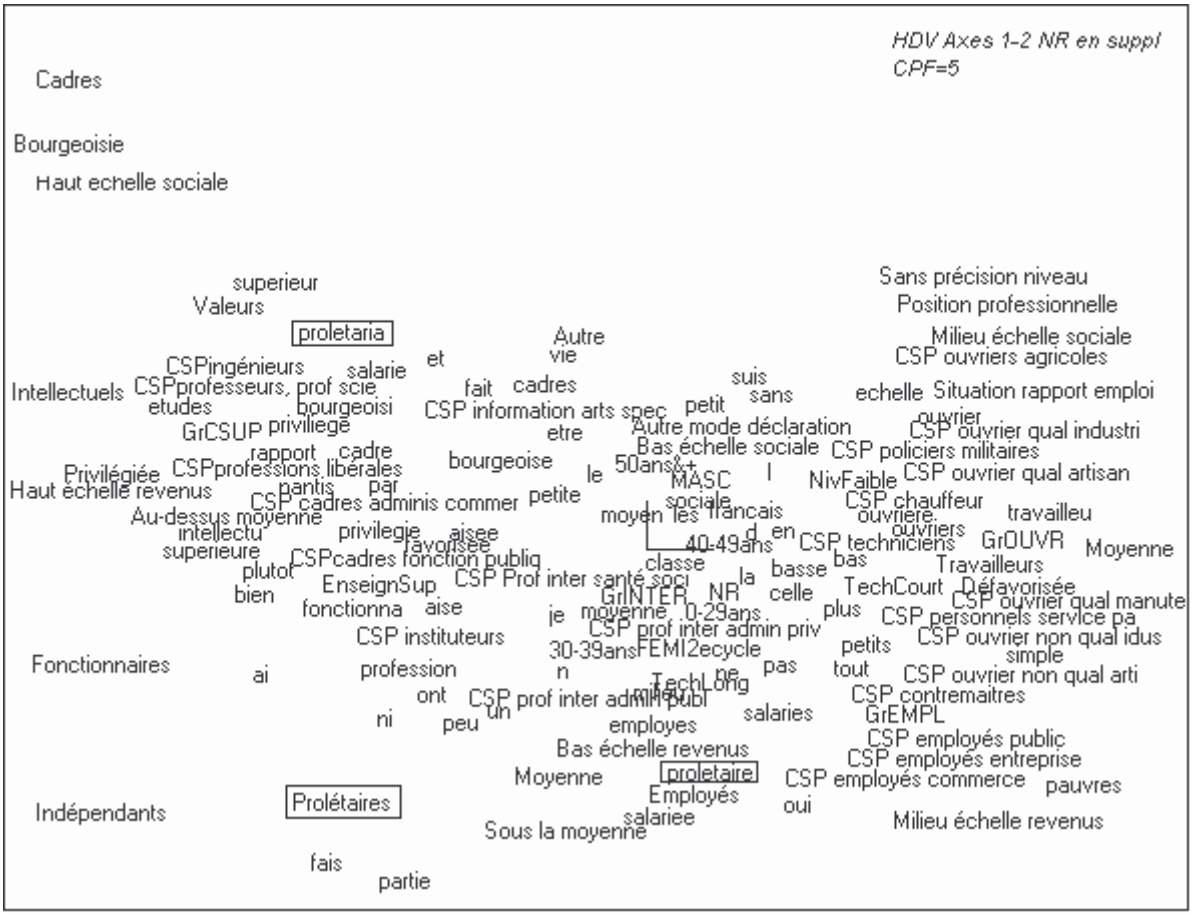
Cependant, si l'on projette sur ce plan le graphe des PEM associant pour un niveau de 15% du PEM les mots et les modalités qui sont en attraction, on a le graphique suivant :




On a désintriqué manuellement les liens avec l'absence de texte donné, qui correspond à une absence de sentiment de classe et l'on voit que ce sentiment est très lié à des niveaux bas (employés, ouvriers non qualifiés, chauffeurs, manutentionnaires, police armée) mais aussi techniciens et contremaitres, à des jeunes de moins de 30 ans, à un niveau faible d'étude. La conscience de classe est liée à l'âge et à la qualification. Ce phénomène étant enregistré, nous mettrons désormais la non-réponse en élément supplémentaire de façon à présenter dans un plan complet la partie droite actuellement sur le deuxième axe seul.

Sur ce deuxième axe, un point intéressant a été montré par le seuil du PEM, c'est le codage "Prolétaires" (encadré sur le graphique) qu'il est surprenant de rencontrer plutôt dans le haut de la hiérarchie sociale. Nous allons reprendre ce

problème dans le nouveau plan où l'effet vu au premier facteur un est éliminé du fait d'avoir mis les non-réponses en éléments supplémentaires.



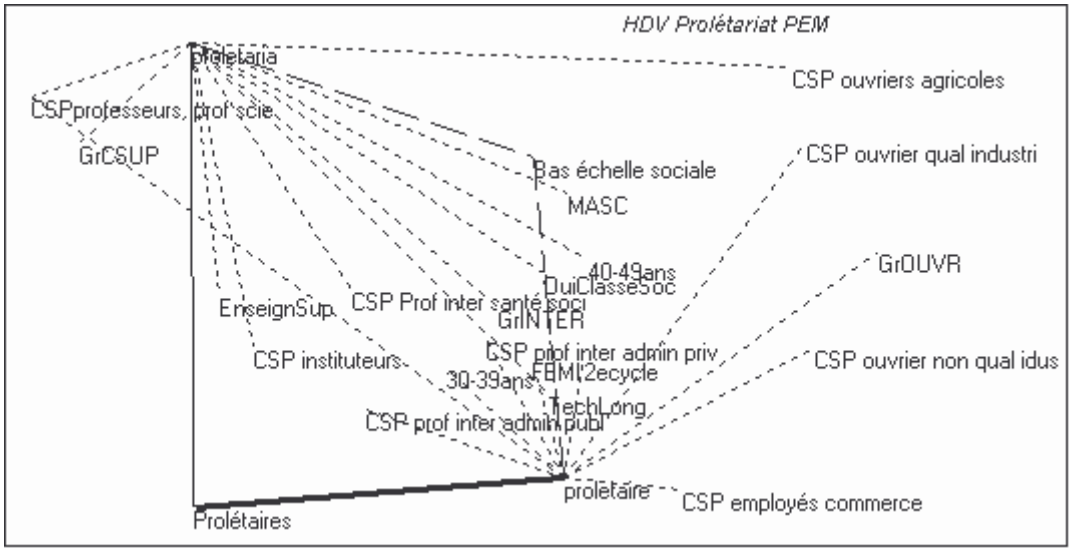

Le groupe CS supérieures est en haut à gauche, le groupe intermédiaire au centre, le groupe employés en bas à droite et le groupe ouvrier au milieu à droite, ce qui est sans surprise ainsi que le vocabulaire employé par chaque groupe. Le mot *classe* est au centre, non spécifique, les termes *aisée, privilégiée, intellectuelle, supérieur* à gauche.

Le problème du *prolétariat* déjà évoqué ressurgit : on trouve ce mot en haut à gauche (sous *supérieur*) mais on trouve le terme prolétaire près des employés et le codage *Prolétaires* en bas (encadrés). Les "prolétaires" ne sont pas là où on les attendrait, c'est-à-dire à droite sur le graphique. Pour éclairer le problème, on constuit les graphes de PEM sur le fond de carte précédent en sélectionnant les deux mots *prolétariat* et *proléraire*.

Le graphique ci-dessous nous apporte la solution de l'énigme. Le codage *Prolétariat* (avec une majuscule, en bas à gauche) a été fait pour les deux mots *prolétariat* et *prolétaire*, de même que le codage "Bas de l'échelle sociale" et l'on comprend ce geste du codeur pour qui la notion de prolétaire ne pose pas de problème, mais il faut examiner séparément les deux mots :

1) *prolétariat* (en haut) est associé aux professeurs, aux professions scientifiques, à des instituteurs, à des professions intermédiaires de la santé ou du social, à un diplôme supérieur, à une tranche d'âge autour de 45 ans , à des personnes de sexe masculin. On retrouve là le concept de prolétariat intellectuel qui est ressenti par des individus à fort niveau intellectuel qui ont le sentiment d'un

déclassement par rapport à leurs attentes, à une frustration. Il s'agit de l'intériorisation d'un thème ancien. Roger Chartier l'a déjà repéré au 18e siècle[4], le thème est un lieu commun depuis le 19e chez Taine[5] ou Barrès[6] et il a connu une reviviscence en 1968 avec le thème du sociologue chômeur radical[7].




2) *prolétaire* en bas est plus lié à des individus de sexe féminin, plus jeunes, de diplômes moins élevés, de professions intermédiaires de l'administration ou employés. C'est la même attitude à un niveau inférieur de qualification. C'est plutôt ce mot qui attire les catégories ouvrières.

En conclusion on voit comment le fait de ne pas prendre en compte les caractéristiques associées à un mot peut entrainer des erreurs de codage. Utiliser le terme de "prolétaires" sans autres indications pour désigner des intellectuels frustrés d'une part et des professions intermédiaires d'autre part, plus que le monde ouvrier, est une erreur qui peut être évitée par une démarche qui prend en compte les informations du tableau lexical des questions par le biais d'un plan factoriel ou d'un graphe de PEM.

*Les abonnés d'une maison de la culture*

Ce deuxième exemple est une nouvelle exploitation en cours d'un dossier déjà étudié : il s'agit d'une trentaine d'entretiens faits auprès d'abonnés de la Maison de la culture d'Amiens qui avaient été sélectionnés sur le fait d'avoir vu pendant la saison précédente des pièces d'auteurs contemporains. On a donc 29 entretiens d'une heure environ qui se présentent de la manière suivante comme un dialogue avec l'enquêteur :

---

[4] Roger Chartier, Espace social et imaginaire social : les intellectuels frustrés au XVIIe siècle, *Annales ESC*, mars-avril 1982, p.389

[5] Taine, *Etienne Mayran*, 1910

[6] Maurice Barrès, *Les déracinés*, 1898.

[7] et un avatar récent : Didier Lapeyronnie, L'académisme radical ou le monologue sociologique. Avec qui parlent les sociologues, *Revue française de sociologie*, 45-4, 2004, 621-651 et le débat qui en a suivi dans le numéro 1 de *Socio-logos*, http://socio-logos.revues.org/

E : Je vais tout d'abord vous demander de vous présenter.

Mme A : Je suis X ; Je vis chez mon compagnon au …, je ne travaille plus mais mon compagnon travaille encore, mes journées sont très chargées malgré tout, j'étais infirmière et j'ai pris ma retraite à 55 ans puisque dans ce métier on y a droit. Depuis que j'ai arrêté de travailler je suis encore plus occupée…

E : Occupée en partie par la pratique du théâtre justement ; je crois savoir que vous avez vus beaucoup de spectacles… J'aimerai connaître vos motivations, ce qui vous incite à une telle pratique.

Mme A : Nous n'allons pas qu'à la Maison de la Culture, il y a aussi la Comédie de Picardie ; nous étions abonnés aux deux ; ce qu'il y a de bien à la Maison de la Culture, c'est la variété des spectacles, on aime bien l'Orchestre de Picardie, les musiques du monde…

Le texte de Mme A fait environ 3000 mots et comme caractéristiques la concernant, nous savons seulement qu'elle de sexe féminin, qu'elle est dans la tranche d'âge supérieure à 60 ans et qu'elle n'est pas enseignante (car on a retenu cette opposition qui était pertinente du fait du nombre important de personnes enseignantes dans l'échantillon). On possède aussi le nom de l'enquêteur (ils étaient trois) mais c'est n'est pas pertinent pour l'analyse proprement dite.

Une première exploitation a été faite de ce corpus par imprégnation et analyse approfondie qui a donné lieu à plusieurs publications (Cibois 2003a et 2003b). Le but de la présente exploitation en cours est de reprendre et de modifier éventuellement la typologie qui avait été découverte précédemment.

Pour faire l'analyse de ce corpus dans Trideux, on constitue le tableau lexical des questions en prenant en compte quatre questions : les trois indiquées plus haut (sexe, âge, profession), mais aussi la question à 29 modalités correspondant aux 29 individus enquêtés. On a éliminé tous les paragraphes émis par les enquêteurs et l'on met en tête de chaque paragraphe les caractéristiques de l'individu qui les a émis. Pour le début de l'entretien de Mme A, on a ceci :

011232Je suis X ; Je vis chez mon compagnon au ..., je ne travaille plus mais mon compagnon travaille encore, mes journées sont très chargées malgré tout, j'étais infirmière et j'ai pris ma retraite à 55 ans puisque dans ce métier on y a droit. Depuis que j'ai arrêté de travailler je suis encore plus occupée...

011232Nous n'allons pas qu'à la MCA, il y a aussi la ComPic ; nous étions abonnés aux deux ; ce qu'il y a de bien à la MCA, c'est la variété des spectacles, on aime bien OrchestrePic, les musiques du monde...

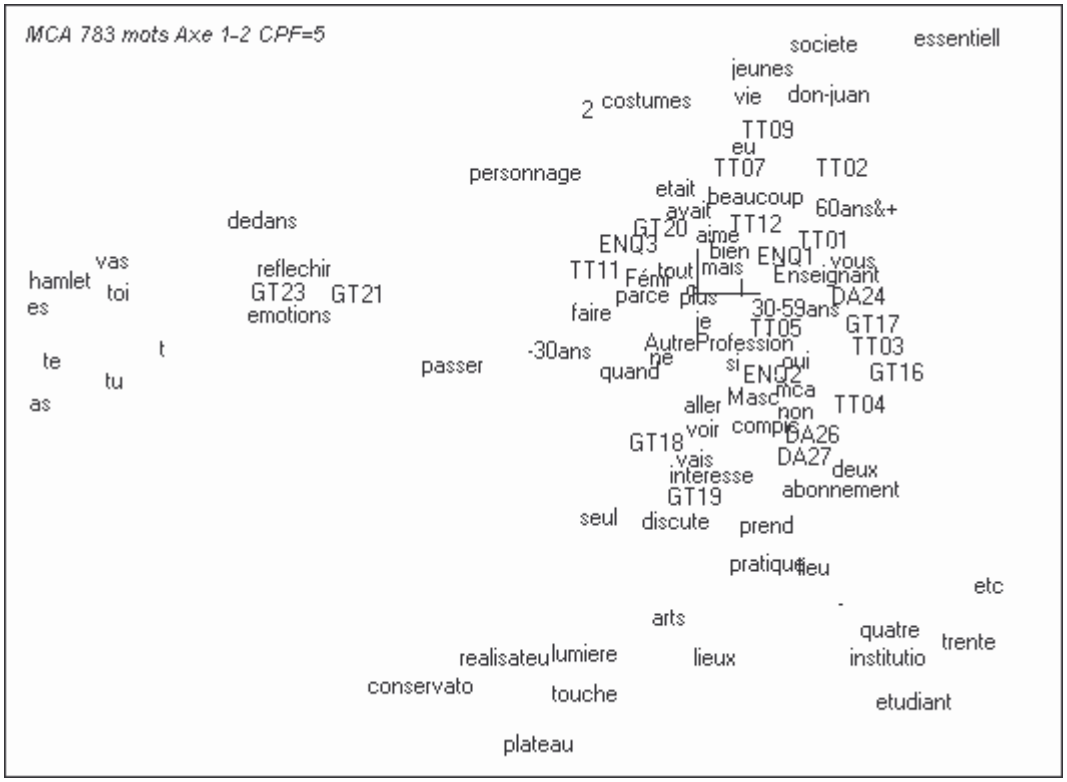
Chaque paragraphe est transformé en une ligne de longueur variable avec une partie fixe en tête, le numéro d'individu (01), le numéro de l'enquêteur (1), le sexe (2), la tranche d'âge (3) et le type de métier (2). Les noms propres qu'on retrouvera ailleurs sont standardisés (Maison de la Culture en MCA par exemple).

Une première étape va supprimer toute la ponctuation et les accents afin d'arriver à définir un mot comme une chaine de caractères entre espaces. Tout le corpus est découpé en mots : le millier de paragraphes des 29 enquêtés génère près de 100.000 mots. A chaque mot sont associées les caractéristiques de l'émetteur, ce qui permet ensuite, par croisement entre les mots et les caractéristiques, la construction du tableau lexical des questions.

La décroissance des fréquences des mots est la suivante : fréquence 1 = nombre de mots différents, fréquence 2 = mots apparus au moins deux fois, etc.

| Fréquence | Effectif |
|---|---|
| 1 | 5511 |
| 2 | 2864 |
| 3 | 2029 |
| 4 | 1594 |
| 5 | 1330 |
| 6 | 1132 |
| 10 | 783 |
| 25 | 360 |
| 50 | 213 |

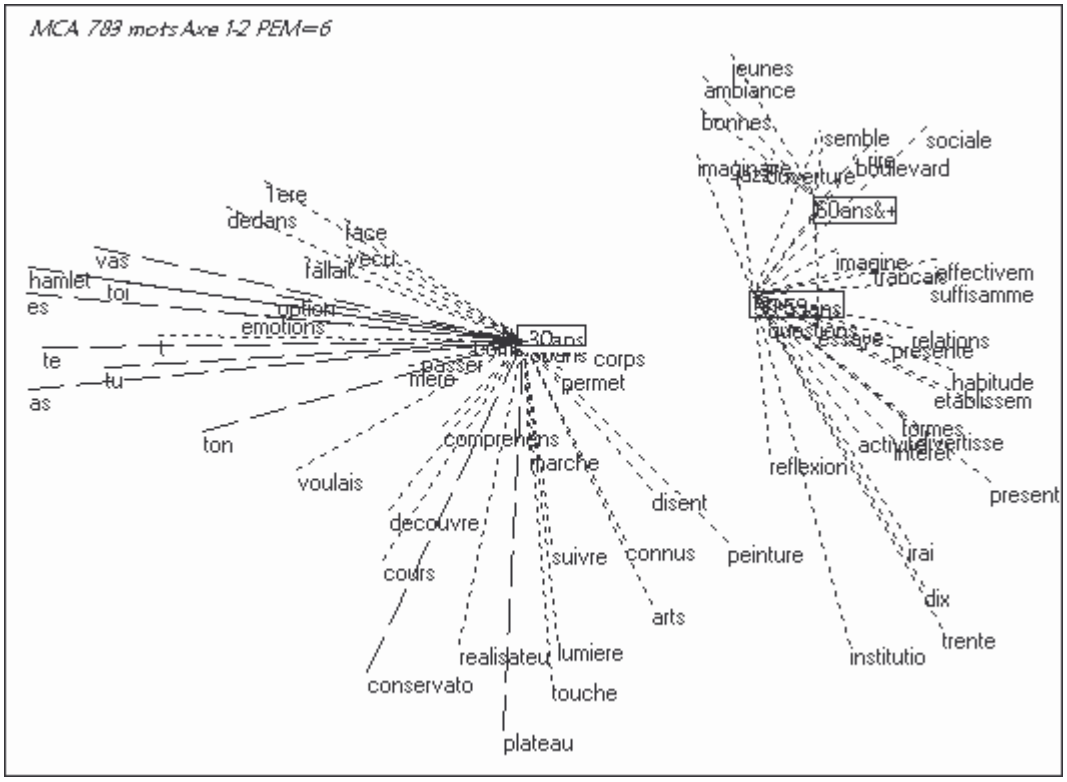
En prenant le vocabulaire des mots de fréquence 10 et plus, on a le premier plan factoriel suivant (CPF=5):




On voit que le premier facteur oppose des formes verbales du tutoiement à gauche aux autres mots. Les individus sont repérés par des numéros précédés d'initiales utilisant la typologie manuelle des analyses précédentes : GT correspond aux Gens du théâtre, spectateurs jeunes, en formation ; TT correspond à Théâtre total, personnes pour qui le spectacle contemporain permet un renouvellement des manières de voir la société ; DA correspond à Divertissement assumé, individus qui pensent que le théâtre doit être d'abord une distraction, un divertissement auquel on doit trouver un plaisir.

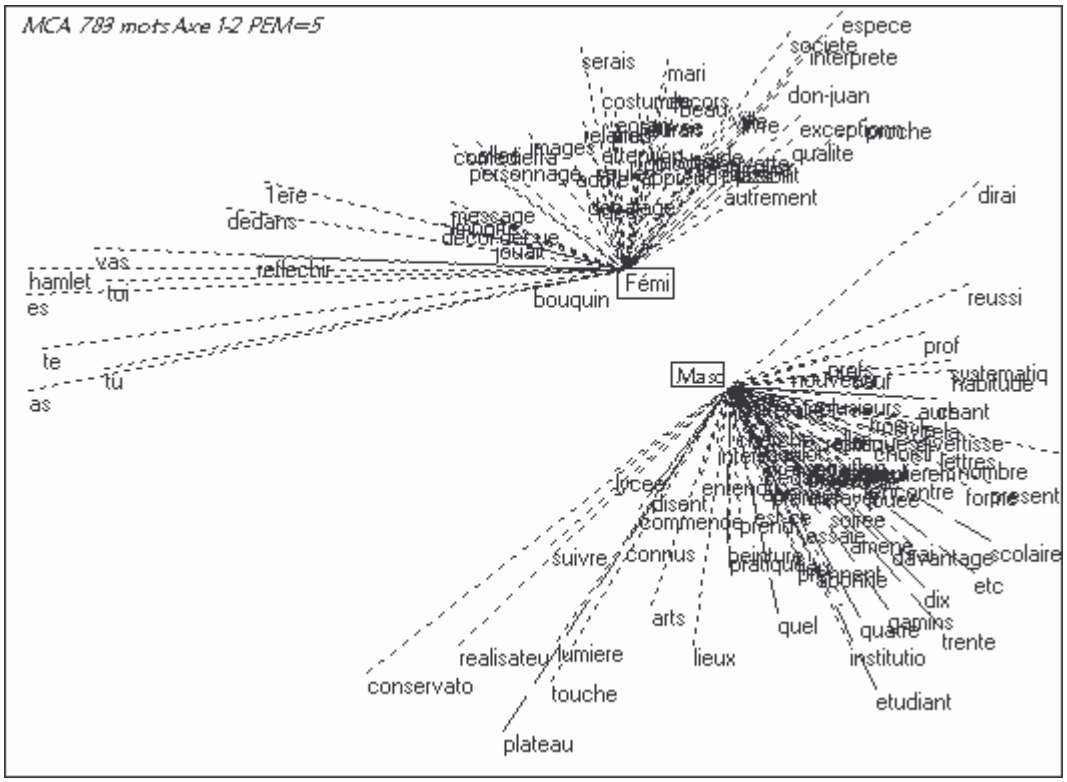
En prenant le graphe des PEM associé à la tranche d'âge inférieure (-30ans), on vérifie facilement que la partie inférieure du plan factoriel correspond bien aux

gens du théâtre en cours de formation, étudiants qui ont pratiqué le tutoiement avec l'enquêteur, lui-même de même âge.




On voit aussi les liens avec les mots techniques (plateau, lumière, réalisateur), les allusions à la formation (cours, conservatoire), qui sont caractéristiques des gens du théâtre en formation. A droite, les liens entre 30-59 ans et plus de 60 ans sont mélangés et correspondent à plusieurs types de répondants.

Si l'on fait la même opération avec l'opposition des sexes, le vocabulaire subit une nouvelle division qui se superpose à la première.




Le vocabulaire féminin reprend le tutoiement (et d'autres mots), le vocabulaire masculin reprend le vocabulaire technique des gens du théâtre (et également d'autres mots). Une expérience analogue de nouveau découpage pourrait être faite avec l'opposition *enseignants* contre *autres professions*.

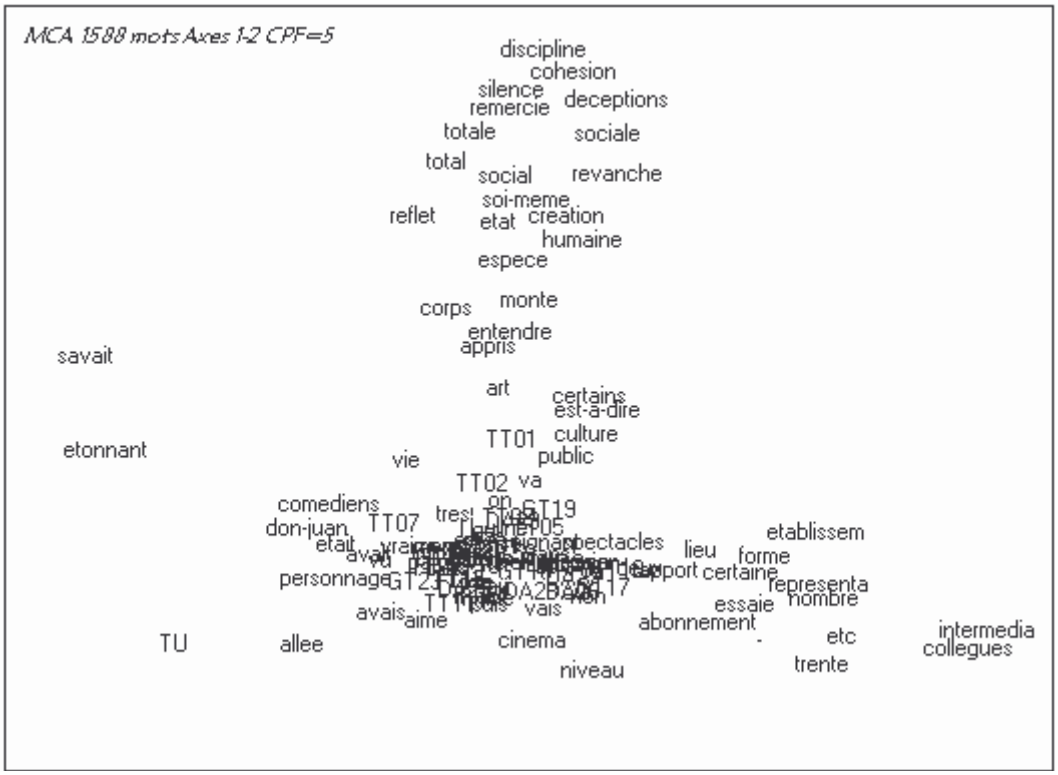
Cette première exploitation nous montre que les caractéristiques dont on dispose ne sont pas pour autant les plus pertinentes puisqu'elles mettent en avant des caractéristiques comme le style de relation à l'enquêteur en les mélangeant à d'autres critères comme le sexe ou la profession. Pour avoir une typologie des répondants eux-mêmes, on va mettre ces trois caractéristiques (sexe, âge, profession) en éléments supplémentaires et ne conserver en variables actives que les numéros des individus. Par contre on augmente le niveau des mots en prenant les mots à partir de la fréquence 4 (1594 mots) en mettant ceux de fréquence 4 et 5 en éléments supplémentaires.

De même, on va mettre en supplémentaires les mots relatifs au tutoiement : cette opération peut être faite directement en cliquant sur le graphique les mots suivants (où les indicateurs de lignes correspondent à la fréquence du mot:

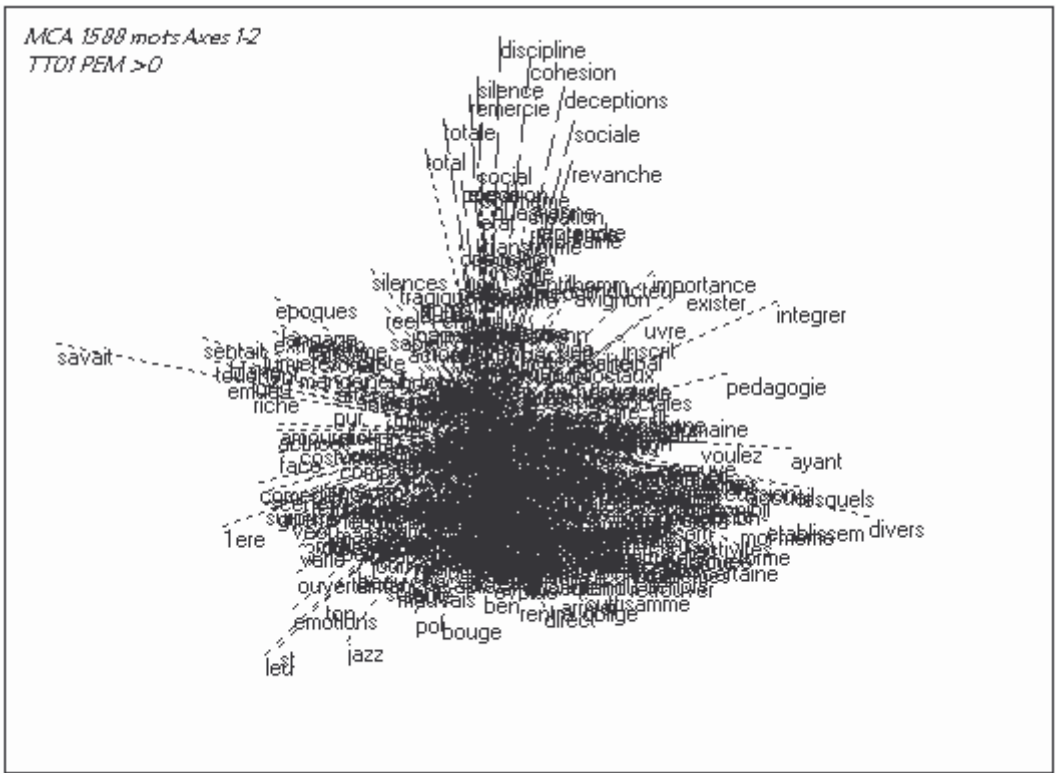
```
La ligne   21 fait l'objet d'un regroupement : t
La ligne   25 fait l'objet d'un regroupement : toi
La ligne   29 fait l'objet d'un regroupement : es
La ligne   31 fait l'objet d'un regroupement : vas
La ligne   42 fait l'objet d'un regroupement : te
La ligne   61 fait l'objet d'un regroupement : as
La ligne  281 fait l'objet d'un regroupement : tu
```

Le résultat final, somme des lignes correspondantes à pour fréquence la somme des fréquences soit 90, il est nommé "TU" en majuscule et mis en élément supplémentaire.

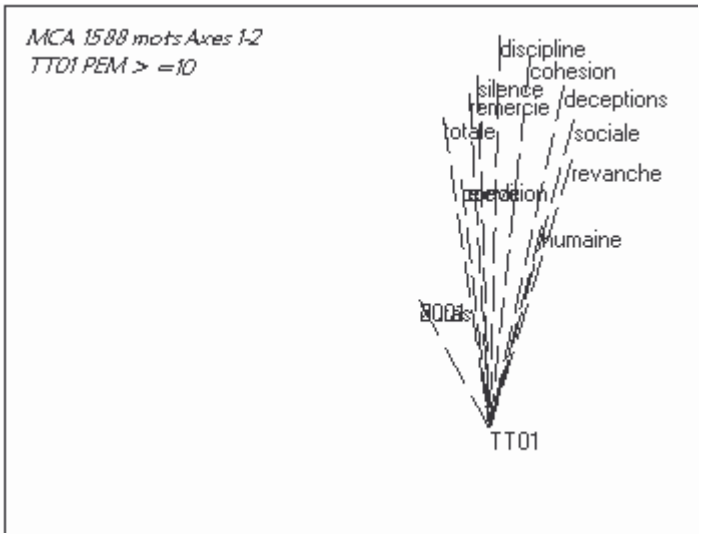
Voici à titre d'exemple des possibilités de traitement, le premier plan factoriel résultant (CPF=5 pour actives et supplémentaires).




On voit que le tutoiement ("TU"), du fait de l'équivalence distributionnelle dans une analyse des correspondances, reste comme élément supplémentaire dans le premier facteur. Ce qui apparait le mieux dans ce plan factoriel, est le regroupement des points en haut du 2$^{e}$ facteur : il s'agit d'un vocabulaire spécifique au spectateur de type "théâtre total" (TT01 et TT02). C'est bien un vocabulaire spécifique et non tout le vocabulaire d'un émetteur. Pour le vérifier, prenons par exemple l'ensemble du vocabulaire de TT01 qui est visualisé en prenant tous les mots associés (PEM>0).

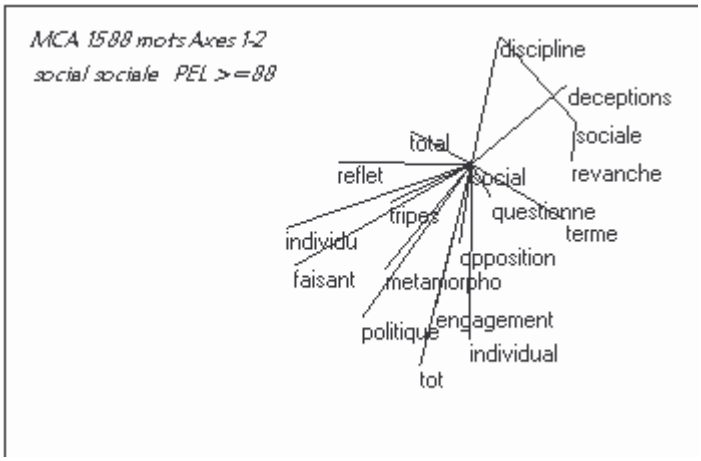


On voit que toutes les régions du graphe comportent des mots utilisés par TT01, par contre si l'on fixe le niveau de PEM à 10, on voit bien que son vocabulaire spécifique correspond bien au côté positif de l'axe 2.




On voit ainsi comment le graphe du PEM permet d'isoler le vocabulaire d'un émetteur qui ici fait allusion à l'aspect "social" du théâtre. Pour approfondir ce thème, nous allons maintenant utiliser, le troisième instrument annoncé, l'indice de proximité entre lignes du tableau.

A cette fin, nous prendrons le pourcentage de l'écart entre lignes (PEL>=88) pour les deux mots proches "social" et "sociale".




On voit que la discipline et la revanche sont "sociale", alors que "social" est associé à l'engagement et au politique, forme spécifique du "théâtre total".

*Conclusion*

Les deux exemples développés ici on pour objet de montrer les possibilités de Trideux en termes d'exploitation d'un tableau lexical des questions par le biais de plans factoriels et de graphes de PEM ou de PEL. On ajoutera que les agrégations de lignes du tableau peuvent être réalisées de deux façons complémentaires :

- soit directement sur le plan factoriel, en sélectionnant les points que l'on veut agréger. Mais une seule agrégation peut être faite ; il faut ensuite mettre cet agrégat en élément supplémentaire et relancer une analyse sur ce nouveau fichier (le

précédent étant toujours conservé dans Trideux afin de pouvoir toujours revenir en arrière).

- soit sur une liste des intitulés de ligne qui peut être triée soit en ordre de fréquence (pour mettre systématiquement en supplémentaires un niveau de fréquence), soit par ordre alphabétique (pour pouvoir retrouver un terme), soit par ordre d'un des facteurs calculés. On peut faire plusieurs agrégations successives en fonction de plusieurs tris. Ce n'est qu'à la fin qu'un nouveau jeu de données sera créé.

Le nombre de points élevés n'est pas un obstacle dans un graphique de Trideux car les points, positionnés à leur place exacte, peuvent être facilement déplacés pour désintriquer les intitulés, mais leur nombre peut facilement être réduit en ne faisant apparaitre que ceux dont la contribution est égale à un niveau donné, qui peut être différent pour les éléments actifs et les éléments supplémentaires. On trouvera les données techniques dans l'aide du logiciel qui est librement chargeable au site indiqué.

## Références bibliographiques


Philippe Cibois, 1989 "Eclairer le vocabulaire des questions ouvertes par les questions fermées : le tableau lexical des questions", *Bulletin de Méthodologie Sociologique* 26, mars 1989 pp. 12-23

Philippe Cibois, 1993 "Le PEM, pourcentage de l'écart maximum : un indice de liaison entre modalités d'un tableau de contingence", *Bulletin de méthodologie sociologique*, n°40, p.43-63.

Philippe Cibois, 2003a "Les abonnés du théâtre : un public hétérogène", *in* Olivier Donnat (Dir.), *Regards croisés sur les pratiques culturelles*, La documentation française, p.171-187.

Philippe Cibois, 2003b "Comprendre les publics du théâtre : l'exemple des abonnés d'une scène nationale", in Olivier Donnat, Paul Tolila (Dir.), *Le(s) public(s) de la culture*, Presses de Sciences-Po, vol. 2, p.169-174.

Ludovic Lebart, André Salem, 1988 *Analyse statistique des données textuelles*, Dunod.


## Site pour chargement de Trideux

http://perso.wanadoo.fr/cibois/SitePhCibois.htm (ou http://www.printemps.uvsq.fr/ rubrique logiciels )