

# 17ème Université d'été en Analyse de Données Textuelles

28, 29 et 30 août 2006  
Carcassonne



## Compte-rendu

L'an deux mille six, le lundi vingt-huit août à neuf heures et quinze minutes la session 2006 s'est ouverte.

Choeb Zafar, président du comité d'organisation, a remercié les participants pour leur présence et l'intérêt qu'ils portent à l'Analyse de Données Textuelles et particulièrement à la méthodologie Alceste. Il a remercié également les intervenants pour leur participation et le partage de leurs expériences.

Il a présenté ensuite Max Reinert, ingénieur de recherche CNRS et responsable scientifique, Franck-Olivier Kakou-Marceau et Patrick Lescure, ingénieurs en informatique et spécialistes en ingénierie textuelle.

Après ce préambule, la parole a été donnée à Max Reinert, qui remercie à son tour les participants et présente le programme détaillé des trois journées de l'Université d'été, puis fait un tour de table afin d'évoquer les préoccupations de chaque participant, ainsi que la nature de leurs jeux de données. Ce tour de table s'achève aux alentours de dix heures quinze minutes.

Après 15 minutes de pause café, Max Reinert reprend son intervention en introduisant la méthodologie Alceste :

« Les méthodes statistiques d'analyse des données introduites par J.P. Benzécri dès les années 1970 ont montré leur efficacité dans le domaine des analyses textuelles. Du moins, c'est ce que les utilisateurs de ces méthodes pensent. Qu'est-ce qu'une telle méthode statistique peut représenter d'un discours ? Peut-on parler du contenu d'un discours sans analyser sa signification linguistique ?

Enfin, n'y a-t-il pas contradiction à proposer une méthode quasi-automatique d'analyse de discours comme ALCESTE alors même que le sens d'un discours ne peut se constituer en dehors de l'engagement d'un interprète (qu'il soit locuteur ou lecteur) ? Cette méthode est fondée sur les présupposés suivants :

a) Un texte n'a pas de contenu en soi, s'il n'est pas pris en charge par un locuteur ou un lecteur. En cela le texte n'est que la trace d'un discours possible. Tout discours n'est également que la reprise de discours antérieurs. De ce point de vue, la position du locuteur et du lecteur n'est pas si différente.

b) Par exemple, ce qu'on appelle le contenu d'un texte est d'abord vécu en situation, comme question. On cherche à viser, à travers le contenu, un objet, mais le contenu s'impose d'abord comme problème; il s'impose à l'interprète bien avant qu'il ne puisse se le représenter. Et cet écart n'est jamais complètement comblé par son discours. Le texte est donc davantage la trace de cet effort de représentation sans cesse renouvelé que la trace d'une représentation.

c) Le texte n'est en tout cas pas la trace d'une représentation achevée. Ce qu'il s'agissait de gloser dans l'avant texte n'était justement pas là. Son intérêt était de résister à la représentation. Dans le cas d'un entretien, par exemple, l'interviewé répond bien sûr aux questions de l'intervieweur, mais en fonction de ce qu'elles mobilisent chez lui comme contenu propre ; c'est-à-dire en tant qu'intérêt, en tant qu'énergie première, et d'abord en tant qu'affects. C'est en tournant autour de ce que le locuteur n'arrive pas à formuler, bien que cela l'affecte, que le discours se constitue.

En conséquence, ces hypothèses conduisent à choisir un corpus qui réunit des séquences d'activité discursive orientées autour d'une "même" problématique. C'est le cas d'une œuvre littéraire, d'entretiens libres ou semi-directifs, d'un recueil d'articles en tant qu'il est choisi en fonction d'un "même" thème.

La méthode d'analyse de discours proposée consiste à chercher dans le recueil des textes choisis pour l'analyse ce qui insiste comme écart, comme rupture. C'est généralement par là qu'une problématique finit par émerger à la conscience. Au plan statistique cela se perçoit par la différenciation de constellations particulières de mots que nous appelons les mondes lexicaux, au pluriel. Ce pluriel insiste justement sur l'aspect problématique d'un centre : le contenu s'exprime d'abord énergétiquement comme tension, comme expression d'un écart, d'un vide, et l'objectif d'une analyse est d'en rendre compte ; c'est à travers leurs oppositions que les mondes lexicaux finissent par constituer

pour le chercheur un espace de tension où va pouvoir se représenter sa problématique.

En conclusion, avec ALCESTE, il ne s'agit pas d'étudier une représentation en soi mais d'étudier la stabilisation dynamique d'un discours possible autour d'un même recueil de textes. Ce processus d'équilibration est préalable à la représentation. En explorant ce qui pose problème, ce qui semble s'opposer ou s'ignorer dans le corpus, on finit par prendre conscience des enjeux et des différentes facettes du contenu d'une problématique. En prendre conscience, c'est justement concevoir l'espace dans lequel il devient représentation.

En ce sens, la méthode ALCESTE offre une aide à la représentation en dégageant les différents mondes lexicaux associés à un corpus. »

En évoquant ensuite le prolongement de ses travaux et la méthodologie, il a expliqué la nouvelle approche des topiques « Imaginaire-Réel-Symbolique » à partir du corpus « Aurélia » de Gérard de Nerval.

La matinée prend fin par un débat questions-réponses.

A douze heures trente, un apéritif de bienvenue est offert par le Conseil général de l'Aude.

Après le déjeuner, à 14h00, Patrick Lescure commence son intervention en présentant le système Alceste à l'aide d'un exemple, le paramétrage des rapports d'analyse, les techniques d'aide à l'interprétation en utilisant les mappings, les réseaux et les dictionnaires thématiques.

Questions-réponses, pause café.

A 16h00, a débuté une séance consacrée à la préparation des jeux de données, notamment formatage au format Alceste, puis les participants ont "dégusté" les résultats de leurs premières analyses.

La séance de travaux pratiques s'achève à 19h00.

Mardi 29 août - 09h00

Max Reinert reprend les résultats du traitement du corpus « Aurélia », et explique

plus en détail la création des tableaux de données et la méthode de classification descendante hiérarchique, fondement de la méthodologie Alceste.

### Le corpus

On entendra par corpus un ensemble de textes réunis. On suppose que cet ensemble a été réuni en fonction d'un objectif particulier, autrement dit, qu'il constitue un objet pour l'analyste, par exemple, un ensemble d'entretiens ou de réponses à une question ouverte, une œuvre littéraire, un ensemble d'articles sur un thème donné, etc.

### Unité de Contexte Initiale (U.C.I.)

Les unités de contexte initiales que l'on note U.C.I. sont des divisions naturelles du corpus, par exemple les réponses à une question ouverte, chaque entretien d'une enquête, les différents chapitres d'un livre, etc. forment des U.C.I., elles sont introduites par une ligne contenant les variables signalétiques appelée ligne étoilée.

### Unité de Contexte Élémentaire (U.C.E.)

L'unité de contexte élémentaire notée U.C.E. est composée d'une ou de plusieurs lignes de texte consécutives. L'unité de contexte élémentaire est considérée comme l'unité statistique essentielle par Alceste.

### Unité de Contexte (U.C.)

Les unités de contexte sont à la base de la classification sous Alceste. L'objectif de l'analyse est leur classement en type de contexte, elles peuvent être définies a priori par l'utilisateur, ou calculées par Alceste.

Les unités de contexte sont calculées par concaténation des unités de contexte élémentaires (U.C.E.) de sorte que chaque unité de contexte (U.C.) contienne un nombre de mots analysés différents. On effectue ensuite une Classification Descendante Hiérarchique sur le tableau qui découle de cette concaténation.

## Classification simple

On effectue une seule classification sur les unités de contexte (U.C.), l'utilisateur peut alors définir a priori ses unités de contexte. En général une classification simple convient bien lorsque le corpus est de petite taille ou lorsque l'on traite des réponses à des questions ouvertes.

## Classification double

Comme l'indique son nom, on effectue deux classifications successives sur des unités de contexte de grandeur légèrement différente. La longueur de ces unités de contexte en nombre de mots est calculée par Alceste suivant la taille et la nature du corpus à traiter. Une classification double a pour avantage d'écarter tout risque dû au découpage et d'assurer la stabilité. Une telle classification convient bien dans le cas des corpus de grande taille.

Exemple : Voici un exemple de regroupement d'unités de contexte élémentaires U.C.E. en unités de contexte U.C. pour une classification double, ce regroupement se fait bien sûr à l'intérieur de chaque U.C.I.

U.C.I.					
U.C.E. 1	U.C.E. 2	U.C.E. 3	U.C.E. 4	U.C.E. 5	U.C.E. 6
U.C. 1		U.C. 2			U.C. 3

On observe ainsi que dans ce cas le regroupement des U.C.E. aboutit à 3 U.C., les unités de contexte U.C.1, U.C.2, et U.C.3 ainsi constituées sont destinées à une Classification Descendante Hiérarchique.

Après la pause, à 11h00, Dominique Coquet Cardinale, Maître de conférence en linguistique anglaise à l'Université Paul Verlaine de Metz, a pris la parole pour son intervention : « Une méthodologie linguistique quantitative et psychanalytique comme outil de décryptage de la Trilogie new-yorkaise de Paul Auster » dont voici un résumé :

« Le voyage que le sujet austérien entreprend dans *The New York Trilogy* a toutes les apparences d'une enquête policière. Multipliées par 3 dans les volets de l'œuvre, les histoires des héros Quinn (dans *City of Glass*), Blue (dans *Ghosts*) et un mystérieux narrateur homodiégétique (dans *The Locked Room*) s'enchevêtrent, se répètent et se complètent. Elles proposent en fait au lecteur, en parallèle de ce qui n'est qu'une déconstruction du roman policier, une triple quête identitaire. Ce récit n'est autre que le voyage initiatique de celui qu'ils représentent : l'auteur implicite qui, par son acte d'écriture, tente de construire la dimension symbolique de l'auteur réel. Ainsi, le sujet parlant responsable de l'énonciation accomplit un voyage sémantique au travers de récits en abyme, et chaque volet amène un élément supplémentaire dans le monde propre que se compose l'auteur. Ce monde est labyrinthique, et il faut se situer au-delà de l'histoire de surface pour étudier l'écriture de Paul Auster. Or, pour parvenir à approcher cette part d'inconscient qui domine l'écrivain, et tenter de découvrir si son œuvre tripartite renferme des volets homogènes ou en évolution, dissociés ou interdépendants, l'analyse quantitative et statistique du corpus constitue un outil performant. L'analyse a mis en évidence l'importance de l'organisation du discours pour l'inscription du sujet dans la chaîne des signifiants. Grâce à l'analyse *Alceste*, on procède au décryptage des choix de contenus de pensée révélés par les lexèmes co-occurents. Dans un deuxième temps, grâce à l'outil *Hyperbase*, on étudie la hiérarchisation de ces contenus au niveau des opérations énonciatives effectuées. Le tout fournit une image globale de la structure Borroméenne de la trilogie new-yorkaise et fait ressortir un degré d'organisation du langage extrêmement élevé. »

Après la pause déjeuner, Patrick Lescure a repris la suite de son intervention sur les fonctionnalités du logiciel *Alceste* et leur rôle dans l'interprétation.

Il a présenté également les différentes méthodes de paramétrage sous *Alceste* et l'optimisation de ces derniers.

A 16h00, deuxième séance de travaux pratiques, durant laquelle les participants analysent et optimisent leur travail en modifiant les différents paramètres, afin de mieux interpréter les résultats. 😊

En fin d'après-midi, avant le repas de bienvenue, les participants ont été conviés par les organisateurs à une dégustation des vins de l'Aude.

20h00 Repas de bienvenue.



Mercredi 30 août - 9H00

Troisième et dernière journée, la session débute par l'intervention de Christian Roy, Maître de Conférence en Sociologie à l'Université de Toulouse le Mirail : « Analyse d'un dictionnaire pour enfants à l'aide du logiciel Alceste ».



« Conformément à la démarche maintenant classique cherchant à saisir, à travers la parole ou la langue écrite, des représentations, des référents, des intentions ou des buts visés, on a cherché ici, à partir du texte d'un dictionnaire pour enfants (Larousse Mini débutants, CP-CE1, 1990), à caractériser le monde de l'enfance – du moins celui que des rédacteurs adultes nous donnent à voir.

Mais quels que puissent être les biais par là introduits, nous ne sommes pas moins en présence de la réalité d'un certain monde, disons donc plus ou moins enfantin.

L'ordre alphabétique du dictionnaire nous présente un tel monde dans le plus grand désordre, en le composant à partir d'une infinité de bribes : celles que constitue la multitude des définitions se succédant. Face à cet apparent chaos, et suite aux nombreux traitements déjà effectués avec Alceste, on a donc été conduit à suivre une méthode qui semble conduire, dans la plupart des cas, à des résultats assez détaillés, tout en facilitant une interprétation d'élaboration souvent délicate.

Nous avons donc vu qu'en sollicitant le plan d'analyse dit « expert », nous pouvions demander jusqu'à 15 classes (B31), tout en descendant jusqu'à un très faible poids de chacune (C11, 10 uce par classe). Ayant donc obtenu 15 classes – ce que rien ne garantissait –, on a cherché leur interprétation en associant profil des classes (par Khi2 décroissant), uce par classe (idem), et classification ascendante hiérarchique (CAH). Tous ces résultats – qui peuvent être diversement enrichis, selon les besoins – ne sont pas toujours d'une évidente compréhension, mais ils le deviennent bien davantage en observant les étapes mêmes par lesquelles Alceste y est parvenu. C'est ici le « dendrogramme » qu'il faut interroger, tel que reproduit dans le rapport d'analyse (étape C1).



Remarquant alors que la grande dispersion de nos classes, formant 15 rameaux, provient de 5 grandes branches, elles-mêmes issues de 2 « sous-troncs » (le corpus total formant le tronc de départ), nous passerons par ces étapes :

- 1) En demandant d'abord une analyse en 5 classes, et en autorisant toujours un effectif très faible – 10 uce – puisque notre très petite classe est dans le lot des 5 branches, quoique ne comptant que 14 uce ;
- 2) En demandant enfin une analyse en 2 classes, avec un effectif suffisamment important pour chacune, c'est-à-dire approchant les totaux prévisibles (ici 800 uce, demandées en C11 toujours).

Sans revenir sur le détail des résultats obtenus, on peut essayer de préciser vers quoi tend la méthode suivie : alors que la 1<sup>ère</sup> classification descendante fournit des renseignements très détaillés, mais dont on a du mal à saisir la cohérence, chacune des étapes suivantes – qui reviennent en quelque sorte à révéler l'ascendance de cette descendance – chacune de ces étapes met en évidence l'unité des sous-corpus à ces niveaux intermédiaires, avant leur prochaine dispersion dans des classes plus particulières. Ce sont encore ici les profils des classes, liste des uce et CAH qui aident principalement à comprendre ce qui fonde ces cohérences intermédiaires, compréhension qui aide alors à définir ce que l'analyse détaillée pouvait avoir d'éclaté, d'hétéroclite ou de disparate.

On se rappellera donc utilement la problématique classique de la compréhension et de l'extension, dont nous voyons ici comme une application : chaque sous-tronc « comprend » - c'est-à-dire à la fois contient et permet de concevoir – le lot des branches qui en sont issues (c'est son extension), de même que chacune de ces branches « comprend » à son tour le lot des rameaux qui en proviennent (qui en sont aussi l'extension).

C'est par ce va-et-vient descendant (extension)//ascendant (compréhension) que des compréhensions jusqu'alors incomplètes et approximatives peu à peu se corrigent, se précisent puis se complètent pour nous amener, ainsi qu'on l'a vu, à un monde globalement bipolaire :

- d'un côté le matériel, l'objectif, avec ses choses et ses lieux ;
- de l'autre l'idéal, le subjectif, avec ses cadres.

On conçoit bien alors que le pôle objectif comprenne une branche corporelle et sensorielle (des percepts), avec ses supports, et une branche classificatoire : celle de l'ordre des choses, et de l'ordre dans les choses (que parents et éducateurs mettent en effet souvent à l' « ordre du jour »).

Le pôle subjectif « comprend » bien, de son côté, les domaines classiques de la connaissance et du jugement (des concepts), ainsi qu'une importante branche des sentiments (des affects) ; la petite classe des souhaits et envies – Noël, anniversaire, disque...- étant probablement isolée du fait des goûts ici stéréotypés prêtés à la population visée. »

Ces 5 grandes branches comprennent donc – et par là aident enfin à mieux comprendre – la diversité des 15 rameaux qui en proviennent, et d'où nous étions partis. »

11h00 - Intervention de Laura Lima, qui mène une thèse en psychologie sociale dans l'équipe du Pr Moscovici: « L'Analyse Alceste et la dimension psychosociale : indicateurs lexicaux et indicateurs psychosociaux. Mis en rapport dans une étude de la RTT à EDF ».

« Le but de cette communication est de montrer comment les concepts sémiologiques rendus opératoires par des analyses textuelles pragmatiques (notamment à l'aide du programme Alceste) permettent d'accéder à la dimension sociale. Elle vise à expliquer comment, par l'intermédiaire de transpositions bien opérées entre indicateurs lexicaux et indicateurs psychosociaux, le chercheur peut saisir la dynamique des enjeux intergroupes à l'origine de la communication et de la production linguistique dont il est question.

Ces transpositions ont été décelées à l'issue d'une longue recherche empirique réalisée à EDF à l'occasion des accords signés en 1997 et 1999 ayant comme objet la *réduction du temps de travail*. Neuf études ont été menées, composées de trois procédures d'analyse lexicale, appliquées successivement à trois registres textuels différents. Les comparaisons et les mises en rapport de trois séries de trois résultats ont permis au chercheur de saisir huit clés de transposition, liant les indicateurs lexicaux (types de lexique, groupement lexicaux) aux indicateurs psychosociaux (opérations mentales qui participent aux processus de formation des représentations sociales et signes qui expriment les rapports intergroupes).

Les transpositions qui ont été établies sont d'ordre théorique et empirique. Trois correspondances ont été établies théoriquement ; d'une part, les types de lexique ou de groupements lexicaux : 1. termes référents, 2. objets référents et 3. fonds topiques, et, d'autre part, les opérations mentales : 1. objectivation, 2. ancrage et 3. thématization. Cinq clés de correspondance ont été découvertes empiriquement. Elles établissent un lien entre les indicateurs lexicaux (4. les noms, 5. les pronoms, 6. les énonciateurs, les groupes ou catégories de sujets qui ont produit le discours en question, 7. les adjectifs, 8. les verbes), d'une part; et, les indicateurs psychosociaux (4. groupes impliquées ou motivations, 5. positions intergroupes, 6. groupes ou catégories d'individus impliqués, 7.

valeurs hégémoniques, 8. Enjeux motivant les dynamiques intergroupes), d'autre part. La lecture conjointe de l'ensemble des huit indicateurs donne une idée complète du système de représentation concerné.

### Références :

Moscovici, S. (Dir.). (1984): *Psychologie sociale*. Paris : PUF.

Moscovici, S. (1994): Social representations and pragmatism communication. *Social science information*, vol 33, n°2, p. 163-177.

Moscovici, S., & Vignaux, G. (1994): Le concept de Thémata. In Ch. Guimelli (Dir.), *Structures et transformations des représentations sociales* (pp. 25-72): Neuchâtel : Delachaux et Niestlé.

Reinert, M. (1990): *Alceste*, une méthodologie d'analyse des données textuelles et une application: Aurélie de Gérard de Nerval. *Bulletin de Méthodologie Sociologique*, 26, 24-54.

Reinert, M. (1993): Les mondes lexicaux et leur logique à travers l'analyse statistique d'un corpus de récits de cauchemars. *Langage & Société*, 66, 5-39.

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, n° 90, 57-70.

Reinert, M. (2001b): Processus catégorique et co-construction des sujets et des mondes à travers l'analyse statistique de différents corpus (1-14) (Linguistique et Psychanalyse [ Colloque de Cerisy, septembre 1998], M. Arivé et Cl. Normand (éd.), 379-392 in Press).

Après la pause déjeuner, une dernière séance de travaux pratiques a eu lieu, permettant aux participants de peaufiner une nouvelle fois leurs analyses et leurs interprétations en toute autonomie.

A 17h30, l'école a été clôturée par Choeb Zafar, qui a remercié les participants pour leur présence et pour les riches débats qui ont eu lieu tout au long des trois journées.

