

QUELQUES ASPECTS DU CHOIX DES UNITES D'ANALYSE ET DE LEUR CONTRÔLE DANS LA METHODE "ALCESTE"

Max Reinert

C.N.R.S. - URA 1033
Université de Toulouse-Le Mirail
31058 Toulouse-cedex

The choice of units of analysis - contextual units or textual units - to perform statistical operations on texts is a delicate part of the practitioner's work. If that choice depends on the practitioner's objectives, it also implies an a-priori knowledge of the text which can rarely be acquired before analysis. It is therefore necessary to check the influence of the chosen units on the results after the analysis has been performed.

The autor presents here a few techniques used in the application software ALCESTE to make that final check..

The text chosen for illustrating these techniques is the philosophical book "Le Réel" by D. Parrochia.

KEY-WORDS : Statistical analysis of textual data ; dicourse analysis method ; Content analysis.

Au niveau méthodologique où nous nous situons ici, l'étude des cooccurrences, comme le remarque J.P. Benzécri (1981), suppose la définition des éléments à distribuer et des contextes distributeurs ou voisinages qui restent à décrire dans une topologie appropriée. Mais quelle distance utilisée quand on sait que les unités lexicales peuvent se rencontrer à différents niveaux d'organisations textuelles : syntaxe, rythme, consonance, catégories grammaticales, jeux de mots, associations, etc. ... une mesure précise de ces voisinages semble irréaliste. Elle ne peut être approchée indépendamment d'un type de préoccupation et, même alors, seulement de manière très approximative. C'est la raison pour laquelle, nous avons préféré approcher cette notion de contexte à l'aide d'un petit segment de texte Δx relativement arbitraire que nous appelons "unité de contexte".

En recourant à ce type de découpage, on se donne les moyens de rendre compte des changements distributionnels au cours du processus discursif d'un même énoncé naturel et d'en apprécier la composition, son homogénéité ou son hétérogénéité, sans présupposer pour autant l'existence d'énoncés minimaux. On doit cependant s'assurer de la stabilité des distributions dans un certain domaine de "grandeur" de ce segment qui reste à apprécier selon le type de texte à analyser.

Certes, il est assez naturel de partir de l'ordre de grandeur des phrases ou des paragraphes produits dans le discours étudié, c'est à dire, habituellement, quelques dizaines de mots. Il faut aussi tenir compte de la pertinence de cette unité relativement à la grandeur du corpus étudié. Ainsi dans l'étude d'un corpus de romans, l'unité d'analyse pourra être plus grande que pour l'analyse d'une oeuvre seule de ce même corpus. Cela suggère que la notion d'unité de contexte ne peut être définie uniquement en soi mais en rapport avec une échelle de mesure des énoncés naturels réunis dans le corpus.

Rappelons que l'objectif de ce découpage est la mise en évidence de "mondes lexicaux" (1993). Un monde lexical n' est pas non plus définissable en soi. Il est la trace lexicale d' un point de vue que l'on peut identifier, certes, en fonction d'un principe de cohérence, d'une logique propre, mais aussi en fonction d'une polémique, d'un parcours discursif construit par ancrage successif des différents points de vue.

Dans le cas où le corpus étudié est constitué des productions d'un seul auteur, ce "vouloir-dire" s'intègre naturellement à une argumentation plus ou moins maîtrisée. Nous aimerions montrer l'apport du découpage en unités de contexte pour une analyse de cette activité discursive.

L'autre problème que nous aimerions évoquer aussi est celui de la discrimination des unités à distribuer. La séparation des formes en fonction de la typographie comme unités d'analyse n'est pas satisfaisante de deux points de vue : celui, bien connu, de la lemmatisation (Bolasco) ; celui aussi des expressions figées : comment décider qu'un groupement de formes joue le rôle d'une unité ici ou de plusieurs là ? La recherche et l'utilisation des segments répétés ont mis en évidence ces problèmes (Salem). Quand doit-on considérer comme unité la forme simple ou l'expression ? Par exemple, quel statut donner à la terminologie technique, aux sigles, aux nombres écrits en lettres, etc....?

Bien sûr, l'approche statistique permet une certaine approximation et on peut espérer qu'usuellement ces divers choix n'influent pas trop sur la stabilité des résultats. Mais cette règle d'usage n'est pas obligatoirement valable pour tout corpus et nécessite au moins un contrôle a posteriori.

On utilisera les résultats de l'analyse d'un livre de philosophie "le Réel" de Daniel Parrochia pour évoquer ces problèmes. Inutile de dire que nous resterons à un niveau de présentation purement méthodologique : rôle du découpage en unités de contexte ; statut des groupements de formes dans un même monde lexical.

1. Rôle du découpage en unités de contexte.

Le rôle du découpage en unités de contexte sera abordé de deux manières : en entrée, quelle longueur choisir pour ces unités ? en sortie, quel apport pour l'analyse des énoncés naturels ?

Cette partie comporte trois paragraphes : 1) le problème du choix de la longueur des unités de contexte ; 2) Une présentation succincte des mondes lexicaux de l'analyse ; 3) une présentation de la distribution des "énoncés naturels" de ce livre (chapitres et sous-chapitres) à l'aide d'une A.F.C. sur les classes.

1. 1 Problème de la longueur des unités de contexte.

Nous avons évoqué à Montpellier en 1993 des essais de stabilité en faisant varier la longueur des unités de contexte, de la longueur minimale des unités de contexte élémentaires (moins de 250 caractères de texte) jusqu'à la grandeur maximale des énoncés naturels (les 19 paragraphes d'Aurélia de G. de Nerval). Nous avons constaté une relative stabilité des résultats dans cet intervalle de grandeur. Nous aimerions montrer ici à la fois l'existence de stabilités locales et l'impossibilité d'attribuer un statut distributionnel clair à la notion d'énoncé naturel dans le cas traité, certains chapitres étant effectivement homogènes de ce point de vue et d'autres non, d'où l'utilité, selon nous, de ce prédécoupage du corpus en unités de contexte de longueur variable pour mettre en évidence ces phénomènes.

Dans les essais sur "Aurélia", les résultats sont apparus relativement stables avec, semblait-il, une meilleure discrimination quand l'unité de contexte était définie avec une longueur d'environ 16 à 20 formes analysées.

Dans l'analyse du livre "Le Réel", six essais avec double-classification (1989) utilisant des longueurs d'unité de contexte variables (en nombre de mots analysés) ont conduit aux résultats réunis dans le tableau 1.

Chaque ligne renvoie à une analyse avec double classification et comparaison des classes. La première colonne indique les longueurs utilisées pour définir les unités de contexte ; la seconde, le nombre de classes stables (indépendantes de ces variations) ; la troisième colonne, le pourcentage d'unités de contexte élémentaire qui n'ont pas changé de classes ; la quatrième,

le nombre de formes réduites totales discriminées par ces classes.

Croisement des classifications des U.C.	Nombre de classes stables obtenues	Pourcentage d'u.c.e. bien classées	Nombre de formes discriminées
de 08 et 10 mots	3	52.44 %	1424 / 1701
de 10 et 12 mots	4	64.96 %	1574 / 1701
de 12 et 14 mots	5	69.34 %	1636 / 1701
de 14 et 16 mots	6	62.64 %	1669 / 1701
de 16 et 18 mots	5	63.92 %	1637 / 1701
de 18 et 20 mots	5	60.33 %	1597 / 1701

Tableau 1 : Discrimination des classes en fonction de la longueur des unités de contexte.

Ce tableau 1 suggère que le classement des u.c. le plus discriminatif des formes se situe autour de la longueur 14 (maximum de classes stables ; plus fort pourcentage d'unités stables, meilleure discrimination des formes). Ce qui est proche de ce que l'on avait observé pour le texte "Aurélia" de G. de Nerval. Cela correspond approximativement à une dizaine de lignes de texte (car ne sont comptabilisées pour définir cette "longueur" que les formes réduites différentes retenues dans l'analyse, les mots outils et les formes rares étant donc exclus). Nous prendrons dorénavant, dans l'exposé de cet exemple, l'analyse 14-16 comme analyse de référence.

2.2 présentation succincte des "mondes lexicaux"

Le tableau 2 donne un aperçu rapide des profils des classes (permettant une approche des principaux mondes lexicaux). L'indication chiffrée qui suit les chapitres renvoie au nombre d'u.c.e. de la classe appartenant au chapitre. La ligne suivante du tableau distribue les principaux noms. Ensuite vient la liste des 8 mots outils les plus spécifiques puis la liste des mots "pleins" (qui sont rentrés dans le calcul de ces classes). Les listes sont ordonnées par spécificité décroissante. Les noms propres ont été analysés. Leur position dans les classes permet d'apprécier les courants scientifiques ou philosophiques sous-jacents.

Un rapide regard sur la sélection des noms propres, sur les intitulés des chapitres et sur la liste des vocables les plus spécifiques permet de construire une première schématisation des résultats :

classe 1 : *l'approche quantique du réel (chap. 4);*

classe 2 : *vers une approche du réel scientifique (chap. 2 et 3) ;*

classes 3 et 4 : *proches dans les deux analyses , elles renvoient au courant idéaliste de Platon à Kant et Hegel avec une opposition entre ces deux derniers (chap. 1 et 3);*

classe 5 : *la notion de réalité dans les sciences humaines (principalement psychanalyse et sociologie) (chap. 5 et 6) ;*

classe 6 : *elle fait davantage référence à des essais d'unification de cette notion de réalité à travers des courants épistémologiques actuels : systèmes ouverts, constructivisme, processus d'auto-organisation, notamment (chap. 6 et 4).*

Au niveau de la présentation des résultats, on fournit pour les formes réduites les plus spécifiques des classes, la liste des formes d'origine associées avec leur fréquence (en nombre d'occurrences dans la classe). Voici quelques formes réduites spécifiques de la première classe :

A9 onde+ : onde(35), ondes(11);
A9 vitesse+ : vitesse(41), vitesses(8);
A9 electron< : electron(8), electronique(1), electrons(19);
A8 rayonnement+ : rayonnement(15), rayonnements(1);
A8 corpuscul< : corpuscule(13), corpuscules(5);
A7 ega+1 : egal(4), egale(8), egales(1), egaux(1);
A7 accel+er : acceleration(3), accelere(10), acceleres(2);

A7 polaris+er : polarisation(9), polarisations(1), polarisee(2),
 polariseur(1), polariseurs(2);
 A7 photon : photon(12);
 A7 photons : photons(12);

Classe 1	Classe 2	Classe 3	Classe 4	Classe 5	Classe 6
Ch_4 (206)	Ch_2 (103) Ch_3 (71)	Ch_1 (107)	Ch_1 (67) Ch_3 (50)	Ch_5 (154) Ch_6 (57) Av propos Conclusion	Ch_6 (49) Ch_4 (35)
Bell Bohr De Broglie Planck	Descartes Galilée Newton Aristote	Hegel	Berkeley Kant Platon	Lacan Freud Marx	Bertalanffy D'Espagnat Popper Leibnitz
au-moment plus-d< alors a-cause lorsqu+ par pendant si	avec point chez loin avant devant vers assez	quant vouloir. deja autrement-dit comme en-effet en-quelque- sorte<	certes a-priori de-sorte< donc par-conseq< presqu+ moi soi	mal croire. dire. guere jamais juste ne pas	ailleurs ici sur plus au-point de-meme du-moins en-definitive
lumiere+ onde+ vitesse+ particule particules electron< rayonnement corpuscul< energ+3 ega+l acceler+er polaris+er photon photons atom+3 proba+ble elementaire mobile+ champ+ espace+ formule+ intervalle+ loi+ longueur+	siecle+ geometr+3 mathemat+3 revolution< qualitati+f cure+ debut+ etendue+ morceau+ science+ univers deduct+ion techn+3 euclidien+ fina+l inventi+f univers+el commune+ decouverte+ dieu+ epoque+ livre+ methode+ nature+	concret+ effecti+f essence+ moment+ unite+ depass+er determinat+ integr+er realitat log+3 concept+ section+ reflech+ir sais+ir approfondi< mediat+ion montre+ immediatete subjecti+f fondement+ histoire+ identite+ neant+ saisie+	intuiti+f categori< chose+ percevoir. sensib< esprit+ entendement intellig< percepti< idee+ objet+ transcend< pur+ ame+ espece+ existence+ idealisme parl+er pens+er arbitra< noumene distincti+f passi+f reduit+	homme+ vie+ signifi+er confusion+ effet+ manque+ ideolog< imaginaire+ inconscient+ socia+l mere+ besoin+ desir+ puissance+ troubl+er vivre. communic< polit+3 force+ interdit+ maniere+ mora+l parfait+ anima+l	remise+ theori< experimentat interact+ notion+ travaux phys+3 topologi+ attenti+f borne+ causa+l fondamental indirect+ rec+ent instrument+ langue+ tentative+ reconnaitre. rendre. accessi+ble independ+an epistemolog regularite+ multiple+

Tableau 2 : vocabulaire spécifique des classes stables de l'analyse 14-16.

1.3. Distribution des énoncés naturels à travers l'A.F.C. du tableau des classes

L'analyse factorielle porte sur le tableau croisant, en lignes, les formes réduites retenues dans l'analyse et, en colonnes, les six classes stables avec, à l'intersection d'une ligne et d'une colonne, le nombre d'u.c.e. de la classe contenant la forme.

On ne présentera que le graphique du premier plan factoriel (tableau 3) relatif à la projection des classes (•1 à •6), des chapitres (Ch_1 à Ch_6, l'avant-propos et la conclusion) et des sous-chapitres. Les "chapitres" et "sous-chapitres" sont projetés en tant qu'éléments illustratifs.

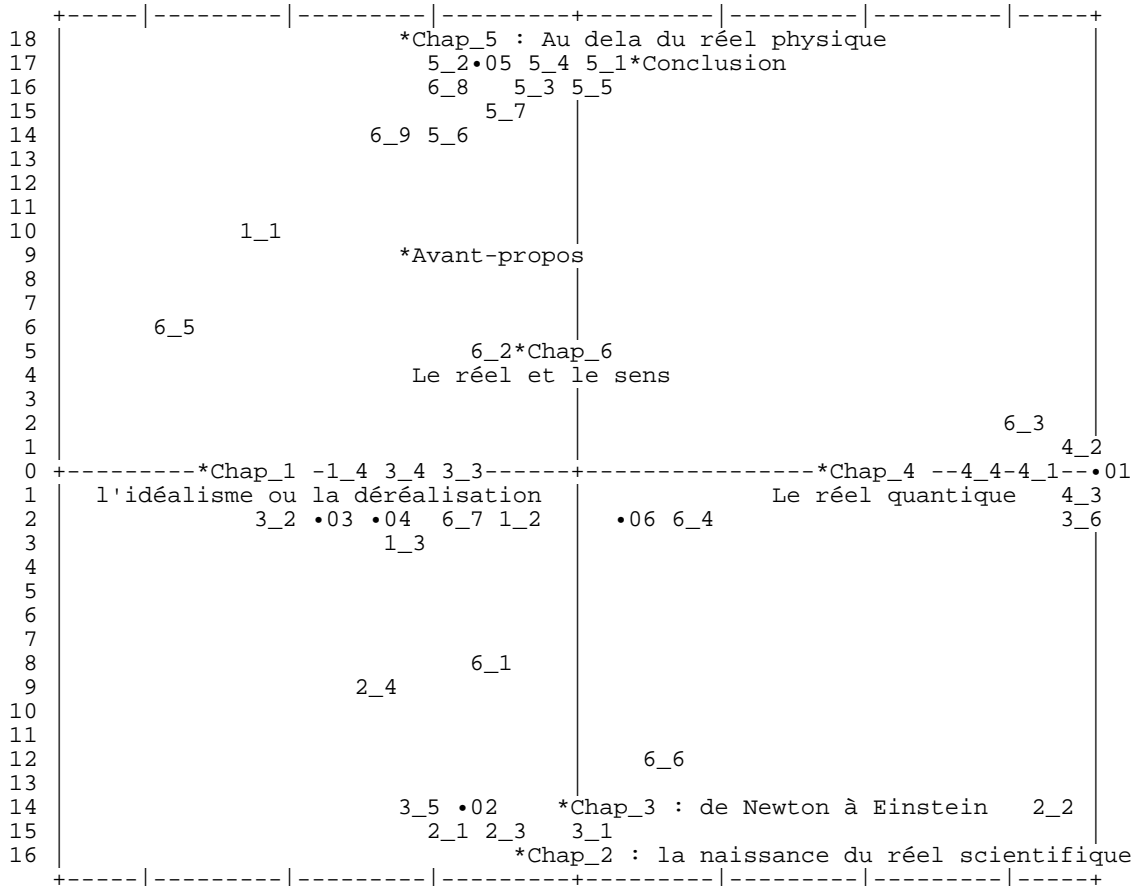


Tableau 3 : Projection des classes et des énoncés naturels sur le premier plan factoriel
 1e facteur (Axe horizontal) : V.P. =.3595 soit 29.41 % de l'inertie.
 2e facteur (Axe vertical) : V.P. =.2641 soit 21.60 % de l'inertie.

Par exemple 5_2 identifie la position du sous-chapitre 2 du chapitre 5. Les intitulés des chapitres sont retranscrits pour mettre en évidence la composition argumentative de l'ouvrage. On retrouve d'une certaine façon la table des matières, chaque chapitre s'associant au suivant. L'avant-propos et la conclusion se rejoignent.

Seul le dernier chapitre, le chapitre 6, a une position centrale. Si l'on regarde la dispersion des sous-chapitres, ce chapitre 6 éclate complètement. Cette composition dénote une volonté chez l'auteur de reprendre dans ce dernier texte l'ensemble des lieux de discours (repérés par les principaux mondes lexicaux) évoqués à travers l'ouvrage sur cette notion de réel afin d'en tenter une première synthèse. Le chapitre 3 joue un rôle assez analogue pour la première partie de l'ouvrage. Les autres chapitres apparaissent au contraire très centrés.

On en déduit que la notion de "chapitre" ne fonctionne pas toujours de la même manière relativement à la discrimination de champs lexicaux propres. L'analyse suggère de différencier ici deux types de chapitres : des chapitres "référentiels" bien ancrés dans un type de discours et les

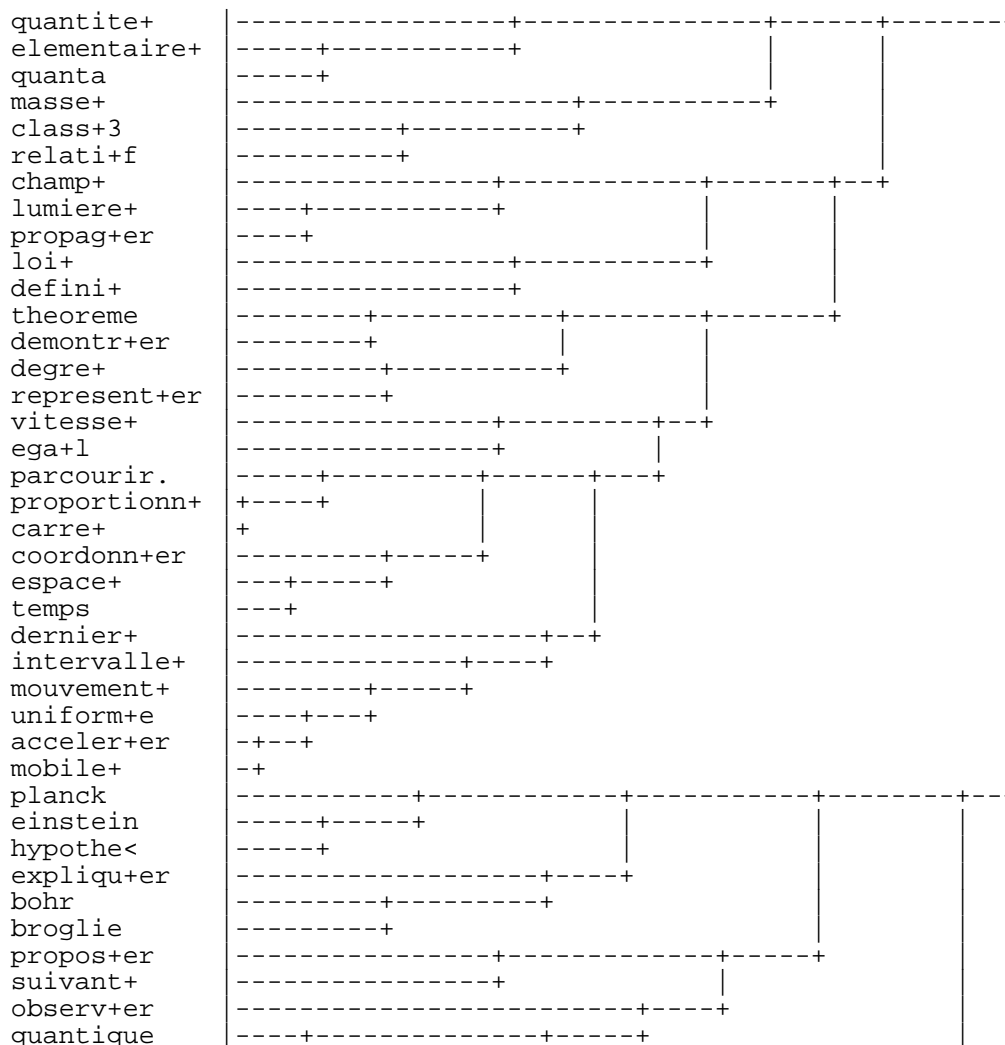
chapitres "rhétoriques" dont l'objectif serait d'avantage de discuter des passages d'un lieu de discours à l'autre. On remarquera que les chapitres 3 et 6 sont respectivement situés à la fin de la première partie et à la fin de la seconde partie de l'ouvrage.

2 L'étude des quasi - figements à l'aide de la classification ascendante appliquée aux formes les plus spécifiques de chaque contexte.

La classification descendante hiérarchique des unités de contexte permet de séparer globalement des classes d'unités de profils contrastés. Ce mode de calcul est utile pour discriminer globalement des mondes lexicaux et les liens entre formes spécifiques ne peuvent être appréciés aussi que globalement.

L'objectif du calcul proposé est de fournir à l'analyste une visualisation plus fine des relations de cooccurrences dans une même classe d'unités afin de mettre en lumière des propriétés locales entre voisinages de mots.

Plus précisément, le tableau analysé est un tableau de cooccurrences calculé sur les formes spécifiques de la classe d'u.c.e. considérée. A l'intersection d'une ligne et d'une colonne la valeur est égale au nombre d'uce de la classe contenant simultanément les deux formes. La diagonale comprend le nombre d'uce contenant la forme sélectionnée. Le mode de classification est une classification ascendante hiérarchique. La métrique utilisée est la métrique du chi2 et l'indice retenu pour la présentation du dendrogramme est l'inertie intra-classe.



984 il est nécessaire de supposer que certaines #quantités physiques, regardées jusqu' à-présent comme #continues, sont composées de #quanta #élémentaires.

On remarquera dans cet exemple le segment répété *quanta élémentaires* ; l'association avec quantité est une association conceptuelle. Une telle approche nous semble pouvoir constituer une aide heuristique pour contrôler certains figements liés notamment à l'utilisation d'un vocabulaire technique.

Conclusion

Beaucoup d'autres aspects auraient pu être évoqués, notamment celui de la définition de la fenêtre de mots à analyser (choix fréquentiels, linguistiques sémantiques ? stabilité des résultats en fonction de ces choix ?). On peut s'attendre à des développements sur ces sujets du fait de l'apparition d'analyseurs syntaxiques performants. La possibilité d'obtenir un marquage fin des catégories de mots en tenant compte de caractéristiques syntaxiques pourra permettre d'effectuer des statistiques sur la distribution de ces catégories dans les énoncés et permettre une explicitation plus précise des choix des mots à analyser (voir notamment les travaux de D. Bourrigault, de P. Constant, de R. Ghiglione). L'analyse syntaxique peut aussi enrichir et diversifier les recherches de concordances, de segments contraints, de termes candidats en terminologie. Un grand nombre de logiciels apparaît actuellement qui montre la productivité du domaine. Reste à construire des ponts entre ces approches et éventuellement de mettre en commun un certain nombre d'outils et de dictionnaires pour permettre à chacun d'avoir des conditions de recherche les plus appropriées.

RÉFÉRENCES

- Bécue, M., Peiro, R. (1993) Les quasi-segments pour une classification automatique de réponses ouvertes, In Anastex (ed.), *JADT 1993*, TELECOM Paris 93 S 003, p 411-423
- Bolasco, S. (1990). Sur différentes stratégies dans une analyse des formes textuelles. In Bécue et al. (ed.), *JADT 1990*, Barcelona, U.P.C., p 69-88
- Constant, P. (1991) *Analyse syntaxique par couche*, Thèse de doctorat de l'E.N.S.T., TELECOM Paris 91 E 007
- Ghiglione, R., Kekenbosch, Ch., Landré, A. (1995) *L'analyse cognitivo-discursive*, document de travail, PARIS VIII.
- Lebart, L., Salem, A (1994) *Satistique textuelle*, Dunod.
- Parrochia, D. (1991) *Le Réel*, Bordas.
- Reinert, M. (1993) Quelques problèmes méthodologiques posés par l'analyse de tableaux "Enoncés x Vocabulaire", In Anastex (ed.), *JADT 1993*, p 539-549
- Salem, A. (1987) *Pratique des segments répétés*, collection "ST-CLOUD", Klincksieck.