

# **Système A.L.C.E.S.T.E. :** **une méthodologie d'analyse des données textuelles présentée à l'aide** **d'une application.**

Max REINERT

Laboratoire de psychologie associé au C.N.R.S n0 259  
Université de Toulouse-Le Mirail  
31054 Toulouse-cedex

**Résumé :** La source d'inspiration principale de la méthode proposée vient de l'analyse des données. Il ne s'agit pas tant de comparer des corpus entre eux que de s'intéresser à la "structure sémantique" d'un corpus donné unique. Cette notion de "structure sémantique", objet d'une analyse "Alceste" ne s'est pas donné de lui-même mais s'est construit progressivement en fonction d'une interrogation sur le sens des opérations effectuées. L'objectif de cet article est justement de présenter ces opérations en essayant d'explicitier les aspects théoriques afin de circonscrire au mieux cet "objet de l'analyse" et d'en discuter le statut et la position. Un exemple d'application permet d'en prendre une mesure concrète.

**mots-clés :** analyse des données, analyse de contenu, analyse du discours, sémantique, pragmatique.

## **1. Aspects théoriques :**

La méthodologie ALCESTE entre dans le cadre général des recherches en analyse de données linguistiques (Benzécri, 1981 ; Lebart, Salem, 1989) et consiste principalement en l'étude des lois de distribution du vocabulaire dans un corpus. De ce point de vue, comme le remarque d'ailleurs J.P. Benzécri, cette approche peut être considérée comme dérivée de l'approche distributionnelle de Z.S. Harris, au moins dans son objectif général puisqu'il s'agit *non pas de chercher le sens d'un texte mais de déterminer comment sont organisés les éléments qui le constituent.*

Selon Harris (cité par Benzécri, p5), *par distribution d'un élément, il faut entendre l'ensemble de ses environnements possibles.* Bien que les préoccupations d'Harris soient plus d'ordre logique et transformationnel que statistique, cette assertion, peut aussi être entendue pour l'approche statistique... à condition de pouvoir définir précisément ce que l'on entend par "élément" et par "environnement".

L'approche proposée se différencie principalement par la définition de ces éléments. Nous faisons de l'énoncé l'unité de base de nos investigations. On entend généralement par *énoncé*, la plus

petite partie d'un discours où un sujet psychique exprime quelque chose à quelqu'un sur le "monde" (ou un "monde") et le reconnaît comme tel. De ce point de vue, la sémantique de l'énoncé se différencie nettement de la sémantique du mot en ce qu'elle contient la marque d'un sujet psychique, d'un *vouloir dire* (Pottier 1987).

*Cependant la notion d'énoncé, en tant qu'unité, n'est pas clairement définie :*

a) l'énoncé peut être appréhendé comme acte de langage : il s'inscrit dans un lieu, un temps, une situation ... c'est l'objet de la pragmatique. En tant que stratégie discursive, en relation avec une situation d'interlocution, il relève de l'argumentation.

b) L'énoncé est aussi bien sûr, l'énoncé d'un propos. A ce titre, il renvoie à un contenu propositionnel relativement indépendant de la situation d'émission. Il joue le rôle d'une *représentation élémentaire*. Il se rapproche de la notion de proposition en logique à laquelle on peut attribuer une valeur de vérité, c'est à dire, le juger par rapport à sa conformité au monde (au moins, au monde du sujet). C'est alors l'objet du sémanticien.

Notre choix a été de nous intéresser plus expressément à l'organisation du second niveau, du niveau sémantique. *Il est donc assez naturel de délaisser dans cette approche, les marqueurs de la syntaxe, les mots-outils, les désinences de conjugaison, certains suffixes, pour ne retenir que les éléments signifiants des mots pleins, c'est à dire les lexèmes.*

Cela dit, cette différenciation entre niveaux ne peut être qu'approximative et ce choix n'implique pas non plus que nous ne nous désintéressions des autres. *Les analyses effectuées montrent d'ailleurs leur interdépendance* et il est habituel d'obtenir des classes d'énoncés construites à partir de la distribution des lexèmes très discriminantes des mots-outils. Mais cette discrimination est obtenue à partir d'un point de vue particulier, celui de la sémantique (on cherche en tout cas à ce qu'il le soit).

Au plan technique, la notion d'énoncé est une notion peu opérationnalisable car elle fait référence à plusieurs niveaux d'analyse : le niveau syntaxique ne suffit pas à la déterminer même si l'on sent qu'un énoncé a une relation avec la notion de proposition, de phrase ou de paragraphe. En effet, un énoncé, aussi bien en tant qu'acte de langage, qu'en tant que propos d'un sujet sur le monde, fait référence à un sujet et donc fait référence à un élément psychique.

*Aussi, plutôt que de chercher à obtenir un découpage rigoureux du texte en énoncés (auquel nous ne croyons pas vraiment) nous lui avons substitué un découpage plus arbitraire en **unités de contexte**, dont la définition peut varier dans certaines limites, et que nous faisons varier. De cette manière, les résultats stables, c'est à dire indépendants de ces variations, ne devraient pas dépendre de l'arbitrarité du découpage, mais uniquement de son ordre de grandeur qui est l'ordre de grandeur d'un énoncé (qui, chez un locuteur moyen, est de l'ordre de quelques dizaines de mots).*

Opérationnellement, si la reconnaissance des mots-outils est assez aisée à l'aide d'un dictionnaire, la recherche des lexèmes pose un problème délicat. *De même que nous avons cherché une approximation statistique de la notion d'énoncé par le découpage en unités de contexte, nous chercherons à appréhender les lexèmes selon une heuristique statistique.* L'objectif n'est pas tant d'avoir une définition précise de la notion de lexème qui permettrait une analyse exacte de chaque forme, que d'avoir une approximation permettant une bonne modélisation *statistique* des données dans leur ensemble. En conséquence, l'idée est de procéder à l'analyse d'un mot si cette analyse est efficace au niveau de la structure des distributions. En clair, l'analyse d'un mot n'a de sens, dans cette approche, que dans la mesure où celle-ci conduit à le regrouper avec d'autres mots.

*Pratiquement, nous procédons en deux temps :*

a) à l'aide d'un "dictionnaire des racines", un algorithme reconnaît les mots-outils et les racines des principaux verbes irréguliers pour les réduire à leur forme infinitive ;

b) Les formes non reconnues lors de la première étape sont traitées à l'aide d'un algorithme particulier. Celui-ci ne réduit une forme que dans la mesure où, d'une part, d'autres formes commençant par la même racine existent dans le corpus traité et, d'autre part, dans la mesure où les terminaisons de ces formes sont reconnues comme des désinences ou des suffixes valides à l'aide d'un "dictionnaire des suffixes" (pour plus de détails sur l'algorithme, se reporter à [Reinert 1986]).

*Aussi plutôt que "lexème", nous utilisons le terme de "forme réduite" ou de "racine" pour désigner les produits par ces transformations.*

En définitive, la modélisation proposée est la suivante : on considère un tableau à double entrée comprenant en lignes, les *unités de contexte*, représentant les objets à décrire et, en colonnes, les *formes réduites* correspondant aux *attributs* de ces objets. A l'intersection d'une ligne et d'une colonne, la valeur "1" signifie la présence de la *forme* dans l'*unité* et la valeur "0", son absence.

C'est l'analyse statistique de ce tableau qui permet de distinguer des classes d'*unités de contexte* en fonction de la distribution différenciée du *vocabulaire* (i.e. l'ensemble des formes réduites associées aux noms, verbes, adjectifs et adverbes). D'un point de vue technique, la différenciation des classes est obtenue à l'aide d'un outil d'analyse de données purement descriptif que nous avons mis au point pour cela : la *classification descendante hiérarchique* (Reinert, 1983, 1986).

Les analyses effectuées sur des corpus très divers montrent, au moins expérimentalement, l'existence de telles classes d'énoncés spécifiques, autrement dit, montrent l'existence de lois d'association particulières entre les mots selon des types d'énoncés. Ces lois impliquent donc que le locuteur, à un moment donné de son énonciation, privilégie l'accès à certains mondes sémantiques plus qu'à d'autres, si l'on entend par là des ensembles de mots dont les associations

se révèlent suffisamment stables dans le temps, pour pouvoir structurer un discours ou un ensemble de discours.

*Notre hypothèse principale concernant ces mondes sémantiques est qu'ils renvoient à des manières particulières du locuteur de choisir à tel ou tel moment de son discours un système de référence ou un autre* (quelles que soient ces manières).

L'analyse que nous proposons est donc bien une forme d'analyse du discours, puisqu'elle met en relief des lois de distribution du vocabulaire extérieures au domaine linguistique. La référence de J.P. Benzécry à Harris est, à ce propos, d'autant plus intéressante que cet auteur est reconnu comme un précurseur de l'analyse du discours ("Discourse Analysis" Language 28 n°1 1952). Pierre Achard, sociologue, note qu'"en proposant, en 1952, d'appliquer l'analyse distributionnelle au "discours", Harris se propose, d'une part de rajouter un étage au linguistique (dimension supra-phrastique), et d'autre part de prendre en compte une dimension de contraintes externe au langage". (Achard, 1987)

## **2. Aspects techniques et application :**

Une analyse comporte schématiquement trois étapes, chacune d'elles comprenant plusieurs opérations. Nous ne présenterons ici que les opérations principales d'une analyse à l'aide d'une application : l'analyse du texte *Aurélia* de Gérard de Nerval.

### **2.1. définition des unités de contexte :**

Le fichier initial nécessaire, en début d'analyse, est un fichier comprenant le texte à étudier. Dans l'exemple choisi, ce texte est celui retenu dans "La Pléïade" (édition Gallimard, 1974, p359-414). La forme initiale du corpus est assez libre : les unités naturelles du texte sont reconnues et distinguées sous le nom d'*unité de contexte initiale (u.c.i.)* : dans l'exemple, les différents chapitres d'AURELIA. Chaque chapitre est introduit à l'aide d'une ou plusieurs lignes spéciales, commençant par un numéro d'identification, et comprenant un nombre libre de mots "étoilés" identifiant des caractéristiques "*hors-corpus*", ici réduites à la composition du texte en deux parties, chacune étant segmentée en plusieurs chapitres :

```
1011 *Partie_1 *chapitre_1_1
```

```
Le rêve est une seconde vie. Je n'ai pu percer sans frémir ces portes d'ivoire ou de corne qui nous séparent du monde invisible. Les premiers instants du sommeil sont l'image de la mort ; un engourdissement nébuleux saisit notre pensée, et nous ne pouvons déterminer l'instant précis ou le moi, sous une autre forme, continue l'oeuvre de l'existence.
```

Le texte est ensuite reformaté et découpé en segments de quelques lignes, avec, si possible, le respect des coupures proposées par la ponctuation. Ces segments de texte constituent les *unités de contexte élémentaires* ou *u.c.e.*.

Les accents et les majuscules sont supprimés. Les locutions les plus usuelles sont reconnues et traitées ensuite comme des *formes simples* :

```
1011      *Partie_1 *Chapitre_1_1
1 le reve est une seconde vie. je n'ai pu percer sans fremir
1 ces portes d'ivoire ou de corne qui nous separent du monde
1 invisible.
2 les premiers instants du sommeil sont l'image de la mort;
3 un engourdissement nebuleux saisit notre pensee, et nous ne
3 pouvons determiner l'instant precis ou le moi, sous une autre
3 forme, continue l'oeuvre de l'existence.
```

## **2.2. formes répertoriées et calcul des dictionnaires :**

Une *forme simple* est un ensemble de lettres séparées par un délimiteur reconnu : espace, début de ligne, signe de ponctuation. Dans une première étape de calcul, les *formes simples* sont délimitées. Certaines sont reconnues, notamment celles associées aux principaux "*mots outils*" : *articles, prépositions, conjonctions, pronoms, auxiliaires être et avoir*.

Dans une seconde étape, ces formes simples sont réduites. Au niveau statistique, l'objectif de cette réduction est de permettre d'enrichir le plus possible les liaisons statistiques impliquées par les cooccurrences des *formes*, en négligeant des différences non fondamentales par rapport au point de vue choisi ici (organisation de la structure sémantique). Pour saisir concrètement le problème posé, il suffit de remarquer le fort pourcentage de "zéros" dans un tableau construit avec notre procédure : si une *u.c.* contient en moyenne 20 *formes* et que nous en analysons 600, le tableau de données qui aurait, en lignes, ces *u.c.* et en colonnes, ces *formes*, contiendrait au minimum 96 % de "zéros". Ce fait explique notre souci de perdre le moins d'information possible en regroupant les *formes* qui peuvent l'être.

Rappelons que deux méthodes de regroupement des *formes simples* sont utilisées : l'une consiste à reconnaître ces *formes* directement à l'aide d'un dictionnaire propre : c'est le cas notamment des principaux verbes irréguliers. L'autre méthode consiste, comme on l'a vu, à regrouper les *formes* du corpus, associables à une même racine : pour être réduite à sa racine, la *forme* associée doit se composer de cette racine et d'une désinence reconnue.

clé	forme réduite	forme initiale	fréquence
0	agir.	agissait	3
0	aller.	vais	1
0	aller.	allai	15
0	aller.	aller	7
0	aller.	vas	1
0	aller.	allait	6
0	aller.	allais	4
1	souvent	souvent	8
1	surtout	surtout	3
1	tant	tant	5
1	tard	tard	10
1	toujours	toujours	16
1	toutefois	toutefois	7
1	tout-a-coup	tout-a-coup	10
1	tres	tres	6
1	trop	trop	12
	abandon+	abandon	1
	abandon+	abandonne	1
	abandon+	abandonnee	1
	accompagn+	accompagna	2
	accompagn+	accompagnaient	1
	accompagn+	accompagnait	3
	accompagn+	accompagnees	2
	accompagn+	accompagnent	1

**note :** la clé permet d'organiser le dictionnaire en fonction de certaines catégories de mots reconnues a priori. Les *formes réduites* terminées par "." ou associées à une clé ont été reconnues à l'aide d'un dictionnaire; les *formes réduites* terminées par "+" ont été réduites uniquement par reconnaissance des désinences et déduction des racines.

### 2.3 calcul des tableaux de données :

Le découpage du corpus en *u.c.e.*, la reconnaissance et la réduction des *formes ayant été effectués*, un premier tableau de données à double entrée est calculé comprenant, en lignes, les *unités de contexte élémentaires* (10 000 maximum) et, en colonnes, les *formes réduites* (1 400 maximum) :

Ces *formes réduites* sont réparties en deux classes : les *formes analysables* qui seront utilisées pour définir les classes caractéristiques d'*u.c.* et les *formes illustratives*, utilisées uniquement pour la description des classes obtenues. Comme nous l'avons déjà dit, les *formes réduites* analysées proviennent des *mots pleins*, noms, verbes, adjectifs et adverbes. Les *mots outils*, prépositions, pronoms, conjonctions, auxiliaires *être* et *avoir*, sont considérés comme des *formes illustratives*

Ces formes illustratives comprennent aussi les *mots "hors-corpus"* qui se distinguent par le fait qu'ils sont "transportables" : ils caractérisent toutes les *u.c.e.* contenues dans une même *u.c.i.* (les différents chapitres d'AURELIA). Ils permettent de définir des classes d'*u.c.* a priori (qui peuvent être aussi directement décrites).

A partir du tableau "*u.c.e. x formes réduites*", deux autres tableaux de données sont ensuite calculés avec une définition légèrement différente des lignes : dans chaque cas, l'*unité de contexte* associée à la ligne considérée regroupe un nombre entier d'*u.c.e.*, la "longueur" minimum d'une *u.c.* étant définie par un *nombre minimum de formes analysées* (choisi par l'utilisateur : dans l'exemple, 10 formes pour le premier tableau, et 15, pour le deuxième). Ces deux tableaux sont soumis ensuite à la classification et comparés pour obtenir les classes stables :

Dans l'exemple proposé, en voici les principales caractéristiques :

**1er tableau (10 formes analysées minimum par unité de contexte) :**

nombre d'*unités de contexte* analysées : **538**

nombre de *formes* analysées : **672**

nombre de '*uns*' : **7019**

pourcentage de '*zéros*' : **98.09 %**

**2ème tableau (15 formes analysées minimum par unité de contexte) :**

nombre d'*unités de contexte* analysées : **416**

nombre de *formes* analysées : **669**

nombre de '*uns*' : **6921**

pourcentage de '*zéros*' : **97.56 %**

**note :** les petites variations dans le nombre de *formes* et le nombre des "uns" s'expliquent par le fait qu'une même forme apparue plusieurs fois dans une même *u.c.* n'est comptabilisée qu'une fois (tableau logique présence/absence); d'autre part, les *formes* n'apparaissant pas dans plus de 3 *u.c. différentes* sont éliminées. Par contre, le nombre d'*u.c.* analysées dans chaque tableau est très différent (538 contre 416).

## **2.4. recherche des classes caractéristiques :**

### **a) la méthode de classification utilisée :**

La méthode mise au point [1983,1987] pour construire ces classes est une méthode de classification descendante hiérarchique. Elle permet de traiter des tableaux logiques (codage "0" ou "1") de grande dimension (4 000 lignes par 1 400 colonnes maximum) mais de faible effectif (60 000 "1" maximum)