LES CAHIERS DE L'ANALYSE DES DONNÉES

MAX REINERT

Une méthode de classification des énoncés d'un corpus présentée à l'aide d'une application

Les cahiers de l'analyse des données, tome 15, n° 1 (1990), p. 21-36

<http://www.numdam.org/item?id=CAD_1990__15_1_21_0>

Les Cahiers de l'Analyse des Données
Vol XV - 1990 - n°1 - pp. 21-36.

UNE MÉTHODE DE CLASSIFICATION DES ÉNONCÉS D'UN CORPUS PRÉSENTÉE À L'AIDE D'UNE APPLICATION

[CLASS. CORPUS]

Max REINERT

Les applications de l'analyse des données textuelles se multiplient. Plusieurs logiciels sont sur le marché ou existent à l'état de prototype (SPAD-T de L. Lebart; LEXICLOUD, d'A. Salem, LEXINET et LEXIMAPPE, de G. Chartron et B. Michelet [17,18]). Le présent article est consacré à notre logiciel ALCESTE [7,10], utilisé actuellement sur Macintosh II.

0 Introduction: vocabulaire et types de représentation

L'orientation générale est ici celle des recherches effectuées en Analyse des Données Textuelles [2, 5, 6] : il s'agit de décrire, de manière formelle, des lois de distribution du vocabulaire dans des textes. Cependant, notre objectif est d'étudier au travers de ces lois, des types de représentations. C'est pourquoi, la méthode que nous préconisons se distingue notamment par les deux points suivants: découpage du corpus en énoncés et choix du vocabulaire.

0.1 Découpage du corpus en énoncés

Nous attribuons à la notion d'énoncé le rôle central d'une représentation élémentaire : notons que la sémantique de l'énoncé se différencie nettement de la sémantique du mot en ce qu'elle contient la marque d'un sujet psychique. Par exemple, "Le ciel est bleu" ne peut être confondu avec "le ciel-bleu" puisque dans un cas, il y a affirmation et dans l'autre, non. Un *énoncé simple* pourrait être défini comme la plus petite partie d'un discours où un sujet exprime quelque-chose sur le monde.

En étudiant la ressemblance ou dissemblance du vocabulaire dans des énoncés différents, on peut espérer aboutir à un premier constat sur des types de ressemblance ou de dissemblance entre les formes de relation au monde. Ces formes n'étant visibles qu'au travers d'une représentation du monde, nous les appellerons des types de monde. Différencier des classes d'énoncés en fonction du vocabulaire, revient donc à différencier les types de mondes référentiels les

Les cahiers de l'analyse des données – 0339 3097/90/01 21 16/$ 3.60/

plus sollicités par un sujet psychique, lors de l'élaboration du corpus. Telle est, en tout cas, notre hypothèse.

Pratiquement, nous proposons d'étudier la distribution des mots dans les énoncés d'un corpus à partir d'un tableau à double entrée, avec en lignes, les énoncés du corpus et en colonnes, l'ensemble du vocabulaire reconnu. Mais faute de pouvoir définir opérationnellement un découpage rigoureux du texte en énoncés, nous adoptons un découpage quelque peu arbitraire en unités de contexte; unités dont la définition peut varier dans certaines limites, et que nous faisons donc varier. Ainsi, nous accédons à des résultats stables, c'est-à-dire ne dépendant pas de l'arbitrarité du découpage, mais uniquement de son ordre de grandeur, qui est celui de l'énoncé.

0.2 Choix du vocabulaire

Pour caractériser ces unités de contexte, nous cherchons à retenir le plus grand ensemble possible de mots, quitte à effectuer certaines transformations sur les formes brutes relevées, en supprimant par exemple les désinences de conjugaison, les marques de pluriel, certains suffixes, de manière à conserver la trace d'un contexte le plus large.

0.3 Aspects techniques et application

Nous présenterons les techniques utilisées à l'aide d'une application: l'analyse du texte *Aurélia* de Gérard de Nerval.

Pour nous, une analyse comporte schématiquement cinq étapes, dont chacune fait l'objet d'un des § qui suivent:

1) définition des unités de contexte;
2) recherche des formes réduites analysées;
3) définition des tableaux de données croisant formes réduites et unités de contexte;
4) recherche des classes d'unités de contexte caractéristiques;
5) description de ces classes pour aider à leur interprétation.

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, pp. 359-414).

La forme initiale du corpus est assez libre : le texte est segmenté en grandes unités que nous appelons, les unités de contexte initiales (u.c.i.): dans l'exemple, ce sont 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, commençant par une étoile, qui identifient des caractéristiques *hors-corpus*, ici réduites à la composition du texte en deux parties, chacune 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. C'est un
souterrain vague qui s'éclaire peu à peu, et où se dégagent de l'ombre et
de la nuit les pales figures gravement immobiles qui habitent le séjour des
limbes. Puis le tableau se forme, une clarté nouvelle illumine et fait
jouer ces apparitions bizarres; le monde des esprits s'ouvre pour nous.
```

Le texte est ensuite reformaté et découpé en segments de quelques lignes, en respectant s'il se peut les 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 ces portes
    1 d'ivoire ou de corne qui nous separent du monde invisible.
    2 les premiers instants du sommeil sont l'image de la mort;
    3 un engourdissement nebuleux saisit notre pensee, et nous ne pouvons
    3 determiner l'instant precis ou le moi, sous une autre forme, continue
    3 l'oeuvre de l'existence.
    4 c-est un souterrain vague qui s'eclaire peu-a-peu,
    5 et ou se degagent de l'ombre et de la nuit les pales figures
    5 gravement immobiles qui habitent le sejour des limbes.
```

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. Un même mot peut prendre généralement plusieurs formes en fonction des marques de pluriel et des désinences de conjugaison.

Dans une première étape de calcul, on délimite les formes simples. 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 à un moindre nombre, afin d'enrichir le plus possible les liaisons statistiques impliquées par les cooccurrences de formes. Supposons qu'une u.c. contienne en moyenne 20 formes et que nous en analysions 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.

Deux méthodes de regroupement des formes simples sont utilisées: d'une part, reconnaître les formes de base directement à l'aide d'un dictionnaire propre: c'est ce qu'on fait, notamment pour les principaux verbes irréguliers; d'autre part regrouper les formes du corpus, dérivant d'une même racine: pour

être réduite à sa racine, la forme associée doit se composer de cette racine et d'une désinence reconnue (pour plus de précision, voir [10]).

clé	forme réduite	forme initiale	fréquence
0	agir.	agissait	3
0	agir.	agir	2
0	agir.	agissaient	1
0	agir.	agit	1
0	agit¨	agiter	1
0	agit¨	agiterent	1
0	agit¨	agitaient	1
0	agit¨	agitent	1
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.

3 Calcul des tableaux de données

Une fois effectué le découpage du corpus en u.c.e. et la reconnaissance des formes réduites, plusieurs tableaux de données, sont préparés. Ils croisent unités de contexte (10000 maximum) et formes réduites (1400 maximum) :

Les formes réduites retenues sont réparties en deux classes : les formes analysables qui seront utilisées pour définir les classes d'u.c. et les formes illustratives qui serviront uniquement à la description des classes obtenues. L'expérience nous a conduit à n'effectuer l'analyse que sur les mots pleins, c'est-à-dire, les noms, verbes, adjectifs et adverbes; et à considérer comme

formes illustratives, les mots outils : c'est-à-dire, les prépositions, pronoms, conjonctions, et auxiliaires être et avoir .

Nous considérons les mots "hors-corpus" comme des formes illustratives caractérisant toutes les u.c. contenues dans une même u.c.i. (lesquelles sont les différents chapitres d'AURELIA). Si ces mots permettent de caractériser les classes obtenues, ils peuvent aussi servir à définir des classes *a priori* (ce qui peut permettre, par exemple, de comparer le vocabulaire spécifique de chacune des deux parties par rapport à l'autre).

A la fin de cette étape on constitue trois tableaux de données:

un fichier numérique comprenant, par unité de contexte élémentaire, un enregistrement, dans lequel est transcrite la séquence des formes réduites retenues, en conservant leur ordre;

deux fichiers numériques comprenant un enregistrement par unité de contexte, avec une définition légèrement différente de ces unités et qui seront les tableaux utilisés pour définir les classes : dans chaque cas, l'unité de contexte analysée comprend un nombre entier d'u.c.e. mais sa "longueur" minima peut être fixée par l'utilisateur comme un nombre minimum de formes analysées par u.c.: dans l'exemple, un premier tableau a été constitué avec 10 formes analysées au moins par u.c. et le second tableau, avec 15 formes analysées au moins.

Dans l'exemple proposé, ces deux derniers tableaux ont les caractéristiques suivantes :

1er tableau (10 formes analysées au moins 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 au moins 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 différents (538 contre 416).

4 Recherche des classes caractéristiques

4.1 Méthode de classification utilisée

Nous avons mis au point [6, 8, 9] pour construire ces classes une méthode de classification descendante hiérarchique qui permet de traiter des tableaux logiques (codage '0' ou '1') de très grande dimension (4000 lignes par 1400 colonnes maximum) à condition que ceux-ci soient de faible effectif (45000 '1' au plus). Le procédé utilisé pour condenser les données est apparenté à celui conçu par L. Lebart: pour plus de détails on se reportera à [9 et 13].

Schématiquement, il s'agit d'une procédure itérative: La première classe analysée comprend toutes les unités retenues. Ensuite, à chaque pas, on cherche la partition en deux de la plus grande des classes restantes, maximisant un certain critère, ce qui conduit à effectuer une succession d'analyses.

La procédure s'arrête lorsque le nombre d'itérations demandé est épuisé.

La méthode de partition d'une classe en deux repose sur le critère suivant: considérons une partition candidate quelconque en deux classes et le tableau des marges associé; ce tableau comprend autant de colonnes que de formes analysées, avec uniquement deux lignes, une pour chaque classe de la partition candidate avec, par exemple, à l'intersection de la première ligne et de la j-ième colonne, le nombre k2j d'u.c. de la classe contenant la j-ième forme identifiée:

l'objectif est de rechercher, parmi toutes les partitions en deux classes, celle maximisant le chi2 de ce tableau (qui est donc le critère choisi).

L'algorithme utilisé ne permet pas d'affirmer que l'on obtient le chi2 maximum, même si l'on présume que le chi2 obtenu ne peut en être éloigné. On procède en trois étapes:

a) chercher le premier facteur de l'Analyse Factorielle des Correspondances du tableau considéré (voir notation dans [1]).

b) chercher l'hyperplan perpendiculaire au premier axe, maximisant l'inertie inter-classes des deux sous-nuages d'unités ainsi différenciés, cette inertie étant à un coefficient près égale au chi2 du tableau des marges.

c) améliorer la partition obtenue à l'aide d'un algorithme d'échange.

Notre démarche nous paraît justifiée sur la remarque suivante.

Considérons une partition des unités en deux classes quelconques; joignons le centre des deux classes par une droite: les valeurs de l'inertie inter-classes et de l'inertie extraite par la droite sont liées; notamment la première est forcément inférieure à la seconde. Aussi, est-il naturel de chercher la partition optimale à partir de la droite optimale, qui est justement le premier axe factoriel de l'A.F.C. du tableau.

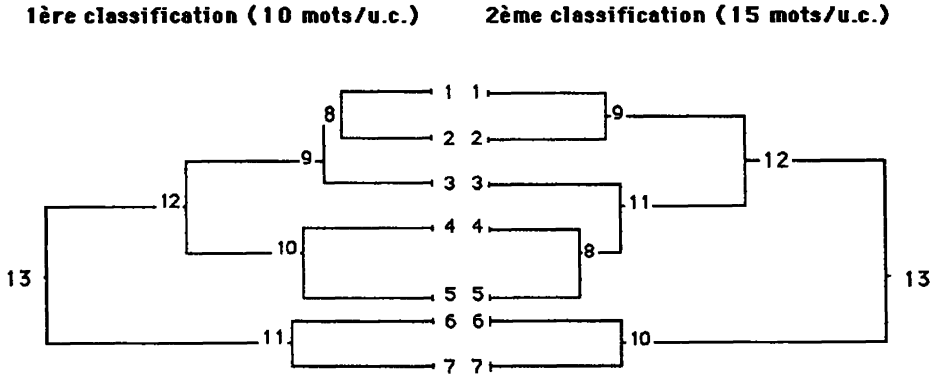
4.2 Choix des classes à considérer

La classification permet d'obtenir une hiérarchie de classes emboîtées les unes dans les autres. Quelles classes considérer pour l'interprétation ? Quelle confiance accorder à leur stabilité ?

La procédure que nous utilisons a un double objectif : d'une part, contrôler la stabilité des classes, en fonction d'une variation de la définition de l'unité de contexte, d'autre part, fournir à l'utilisateur, une méthode lui permettant de choisir dans la hiérarchie des partitions proposées, une partition acceptable. Pour cela, on compare les classes obtenues dans chacune des deux C.D.H, pour ne conserver que les classes relativement stables. Les u.c. analysées comprenant un nombre entier d'u.c.e., il est possible de comparer les classifications obtenues, chacune des deux classifications sur les u.c. pouvant être considérée comme une classification sur les u.c.e..

La technique pour comparer les classes est simple: on calcule les liens entre classes deux à deux (à l'aide d'un chi2), puis, on repère, parmi les couples ainsi définis, ceux ayant un lien maximum en ce sens que chacune des classes du couple, s'associe davantage à l'autre classe du couple qu'à toute autre classe de la hiérarchie.

Nous présentons les résultats obtenus dans le cas étudié.



NB: Pour les deux analyses, on a demandé 7 classes terminales, numérotées de 1 à 7. D'où 6 classes non terminales, numérotées de 8 à 13 : la classe 8 de la première analyse, par exemple comprend toutes les unités classées dans les classes 1 et 2 plus, éventuellement quelques unités qui ont pu être éliminées lors de l'analyse de cette classe. La classe 13 comprend toutes les unités retenues du corpus : c'est la première classe analysée (on notera que la numérotation des classes est ascendante même si la méthode d'obtention de ces classes est descendante pour des raisons de commodité de calcul).

correspondances entre classes (en nombre d'u.c.e.):

RCDH10	<->	RCDH15	*	freq1	freq2	freq12	chi2	*
6	<->	7	*	247	241	142	261	*
7	<->	6	*	180	155	88	235	*
8	<->	9	*	276	358	179	200	*
10	<->	11	*	402	397	279	344	*
11	<->	10	*	431	397	309	435	*
12	<->	12	*	737	775	652	442	*

6 <-> 7, par exemple, signifie que la classe 6 de la première analyse (10 formes par u.c.) qui comprend 247 u.c.e., est en correspondance avec la classe 7 de la deuxième analyse, qui comprend 241 u.c.e., l'intersection entre ces deux classes comprenant 142 u.c.e. ce qui correspond à un lien égal à 261 (chi2 d'association à un degré de liberté).

L'objectif est de chercher, parmi ces couples en correspondance, ceux associés à une partition du plus grand nombre d'u.c.e.

Les classes d'une partition sont telles qu'une unité quelconque appartient à une et une seule des classes : le souci de choisir une partition pour la description des résultats est donc un souci d'exhaustivité: on désire que les traits observés soient observables sur le plus grand nombre d'unités.

Trois partitions peuvent être retenues ici: une partition en deux classes définissables à partir des couples 11<->10, 12<->12; une partition en trois classes, à partir des couples 8<->9, 10<->11, 11<->10 ; enfin, une partition en quatre classes à partir des couples 8<->9, 10<->11, 6<->7 et 7<->6.

On remarque que chacune de ces deux dernières partitions se déduit de la précédente par l'analyse d'une classe, la partition en deux classes correspondant à la structure la plus caractéristique.

Nous n'analyserons que la partition en trois classes, la troisième classe 11<->10 identifiant en définitive la structure la plus stable. L'intersection des classes de chaque couple retenu identifie les classes définitives (renumérotées de 1 à 3) qui serviront de base au calcul des profils.

5 Aides à l'interprétation des classes

Pour analyser la structure des classes d'u.c.e. extraites plusieurs procédures peuvent être utilisées. Nous en retiendrons deux, qui nous semblent les plus suggestives:

1) relevé du vocabulaire le plus spécifique de la classe retenue;

2) extraction des u.c.e. les plus représentatives de ce vocabulaire parmi les u.c. de la classe considérée.

5.1 Description du profil des classes

Pour chaque classe, on calcule la liste des mots les plus significativement présents. Cette procédure peut être utilisée pour des mots autres que ceux analysés.

NOTE: le coefficient d'association d'une forme à une classe est un chi2 à un degré de liberté, calculé sur le tableau de contingence croisant la présence ou l'absence du mot dans une u.c.e. et l'appartenance ou non de cette u.c.e. à la classe considérée.

Les mots relevés sont ceux ayant un chi2 d'association supérieur à 2.7; en gras, les mots associés à un chi2 supérieur à 10 pour les formes analysées et 4 pour les formes illustratives. Rappelons que les formes illustratives n'ont pas contribué au calcul des classes contrairement aux formes analysées.

1ère classe (387 u.c.e.)

formes analysées:

attendre. (5), **comprendre.** (25), connaitre. (4), couvrir. (7), **croire.** (23), errer. (9), faire. (5), **jeter.** (20), mettre. (7), ouvrir. (4), paraitre. (3), placer. (7), prier. (3), recevoir. (4), **sortir.** (11), souvenir. (3), venir. (6), voir. (6), vouloir. (13), ensuite (5), peut-etre (8), plusieurs (7), quelque (7), rien (8), **tout-a-coup** (14), tres (6) accompli+ (4), achet+ (9), ami+ (26), approch+ (4), arriv+ (15), aurelia (4), avou+ (7), ayant (6), campagne (7), certitude (3), **chant+** (11), **cherch+** (11), cite+ (9), continu+ (8), contree+ (7), conversation+ (3), **cri+** (9), demeur+ (4), dirige+ (8), d-abord (9), eglise (6), eloign+ (4), **entend+** (13), esper+ (3), **etoile+** (10), expliqu+ (6), exterieur+ (4), fatigu+ (4), fievre+ (3), galerie+ (6), georges (4), hasard (5), heure¨ (3), idee+ (6), impuiss+ (4), **inconnu+** (10), influ+ (6), inspir+ (9), larme+ (7), ligne+ (7), lit¨ (3), march+ (8), merveille+ (3), mot+ (2), mysterieus+ (4), **nuit** (21), oncle+ (4), or* (3), palais (4), parcour+ (3), pardon+ (7), **parl+** (10), **personnes** (13), priere+ (9), proie (9), racont+ (9), **rencontr+** (12), rentr+ (6), reveill+ (3), route (9), **rue+** (13), saint+ (2), salle (16), sens (5), seul+ (3), **sorte-de** (12), spectacle (4), suite (6), **temps** (11), termin+ (3), terrible+ (4), touch+ (6), transport+ (6), vague+ (5), **visit+** (10), voix (5)

formes illustratives:

en (5), **chez** (9), **dans** (11), **pour** (4), **vers** (7), **je** (35), **me** (19), **mes** (12), **mon** (11), avait (7), avoir (3)

*1_3 (24/54), *1_9 (24/49), *2_2 (28/62), *2_3 (2/24),*2_4 (48/112), *2_5 (40/77), *Partie_2 (222/611)

On notera que le vocabulaire spécifique de la classe comprend:

de nombreux verbes d'action, des déplacements: errer, faire, jeter, mettre, recevoir, sortir, venir, vouloir, approch+, arriv+, cherch+, dirige+, eloign+, march+, parcour+;

des personnages, ou des vocables exprimant une relation sociale: ami, aurelia, georges, oncle, personnes; et aussi: achet+, avou+, conversation+, mot+, parl+, racont+, visit+;

des lieux: campagn+, contree+, eglise, exterieur+, galerie+, palais, route, rue+, salle+;

des mots évoquant une tension, une surexcitation: chant+, cri+, fievre+, impuiss+, larme+, terrible+;

une forte présence des indicateurs de la première personne (d'autant plus intéressante que ces indicateurs ont été considérés comme des formes illustratives): je, me, mes, mon.

2éme classe (386 u.c.e.)

formes analysées :

amener. (4), appartenir. (4), appel¨ (4), craindre. (7), creer. (4), devoir. (5), **dire.** (23), ecrire. (7), **falloir.** (11), mourir. (4), pens¨ (3), pouvoir. (8), **savoir.** (13), sentir. (6), souffrir. (6), user. (3), vivre. (3), encore (9), ici (14), **jamais** (23), longtemps (3), maintenant (8), **ne** (55), ni (7), partout (7), **pas** (49), souvent (2), tant (9), toutefois (8)
aim+ (10), amer+ (9), apparence+ (3), arme+ (7), **bien** (27), celeste+ (4), certain (3), chretien+ (2), circonstance+ (7), coincid+ (3), compt+ (4), conserv+ (12), consult+ (3), **c-est** (26), demand+ (4), details (3), **dieu+** (31), divers+ (2), domin+ (3), dout+ (7), eloim (9), envahi+ (7), **epreuve+** (11), **eprouv+** (14), esprit+ (6), eternel+ (3), etrang+ (8), even+ (3), **fatal+** (12), fils (4), **fois** (17), formul+ (3), frere+ (5), froid+ (3), gard+ (6), generation+ (3), heureu¨ (7), humain+ (4), **ignor+** (10), il-y-a (7), impos+ (4), indiqu+ (3), juge¨ (4), juste+ (3), **lettre+** (15), lien+ (4), livr+ (6), lutt+ (4), **malheur+** (11), manqu+ (2), meilleur+ (3), mere (8), milieu (5), mort¨ (5), moyen+ (4), mystique+ (4), naturelle+ (9), natur+ (4), **neant** (11), nom (2), papier+ (9), peine¨ (3), peniblement (3), poete+ (3), possible (7), prepar+ (3), primitiv+ (3), race+ (3), raison+ (14), rapport+ (9), regret+ (6), **religi+** (12), **resolu¨** (13), result+ (7), retourner (7), retraite+ (4), **retrouv+** (10), science¨ (3), sens+ (2), **sentiment+** (10), sept+ (7), serie+ (5), simple+ (4), singulier+ (3), somme (3), songe+ (3), subir (7), supreme+ (4), talisman+ (3), terme+ (7), tout (5), tradition+ (7), tresor+ (3), vaincu+ (7), verit+ (4), vie (22), vision¨ (4)

formes illustratives:

ce (13), **si** (10), **suis** (12), t (3), **ainsi** (10), **car** (6), **contre** (4), **pourquoi** (6), **pourtant** (4), **que** (17), lui (3), **nos** (4), nous (3), **tu** (5), cela (3), ces (3), **telle** (4), dont (3), **ai** (10), **as** (6), **avons** (9), **est** (14), **fut** (10), **serait** (4), **sommes** (6)
*1_1 (28/41), *2_1 (46/72), *2_9 (16/34)

On voit que de nombreux mots du vocabulaire spécifique de la classe 2 renvoient à des concepts plus abstraits et par là-même nous évoquent les

passages où l'auteur disserte sur le sens de la vie, de la religion, de Dieu, sur un système du monde: celeste, chretien+, dieu, esprit+, eternel+, evenement, generation, humain+, livr+, mystique+, race+, raison, religi+, science-, sept, serie+, simple+, verit+, vie.

À cette réflexion philosophique, se mêlent des valeurs morales du bien et du mal, des valeurs affectives, le sentiment du devoir, d'une culpabilité, ou d'une fatalité malheureuse. De nombreux termes peuvent être mis en rapport avec ces valeurs spécifiques de la classe2: craindre-, devoir-, falloir-, aim+, bien, epreuve+, fatal+, juge-, juste+, malheur-, supreme+.

3ème classe (368 u.c.e.)

formes analysées:

agit¨ (8), aperc¨ (3), **apparaitre.** (16), entrer. (4), jouer. (5), lier. (3), monter. (3), plaire. (3), **porter.** (17), tenir. (3), assez (3), devant (3), non (4)
action+ (3), aile+ (5), **angl+** (12), apport+ (3), arbre+ (6), astre+ (6), attir+ (3), **a-mesure-que** (10), bizarre+ (6), **blanc+** (17), **bleu+** (12), bois+ (4), bord+ (5), **bras** (10), brill+ (4), chambre+ (3), **charge+** (12), cheveu¨ (5), ciel+ (3), colline+ (7), **color+** (14), combinaison+ (3), corps (3), correspond+ (3), cote+ (8), **couleur+** (12), couvert¨ (5), creation+ (4), creuse+ (3), decoup+ (3), demi (3), descend+ (8), distingu+ (9), **divin+** (15), eau (6), ebauch+ (8), **eclair+** (15), eclat+ (4), **elanc+** (14), enfant+ (5), **entour+** (18), epanou+ (8), escalier+ (7), etend+ (9), femme+ (3), ferm+ (3), feu (3), feuill+ (6), **figur+** (27), fill¨ (6), fleur+ (6), fleuve+ (5), **form+** (15), front (3), germe+ (5), gliss+ (3), gout+ (3), harmoni+ (4), haut+ (4), herb+ (8), horizon+ (3), immense+ (3), immortel+ (3), infini+ (4), **jardin+** (18), **jeune+** (17), **longue+** (17), lumiere+ (8), lumineus+ (8), lune+ (6), main+ (8), **maison** (18), **matin+** (12), memes (4), menac+ (4), menage+ (3), model+ (3), monstre+ (8), montagne+ (3), **mont+** (10), mouvement+ (3), mur+ (8), nouvelle+ (2), nuages (8), oppose+ (3), orage+ (3), ouvrier+ (8), pal+ (3), pareil+ (3), parterre+ (3), particulier+ (8), **pays¨** (16), penetr+ (3), perspective+ (3), **petit+** (25), **peupl+** (10), peu-a-peu (6), pied+ (6), **plante+** (11), plein+ (5), present+ (3), profond+ (3), promen+ (4), rayon+ (16), recit+ (3), represent+ (5), ressembl+ (8), revet+ (5), robe (8), rocher+ (8), rose (4), roug+ (3), rustique (8), sauvage+ (8), scene+ (7), **sein** (14), serpent+ (8), situ+ (8), soci+ (3), soldat+ (3), soleil (2), source+ (8), souterrain+ (3), supporte+ (3), tableau+ (7), taille+ (7), **teint+** (15), terrasse+ (5), terre¨ (5), tete+ (8), toile+ (8), **touff+** (10), tourn+ (3), trac+ (3), trait+ (8), travail+ (3), treill+ (4), trouble+ (3), vari+ (3), vaste+ (5), verdure+ (3), vert+ (3), vetement+ (9), vetu+ (9), **vis** (34), **yeux** (14)

formes illustratives:

comme (2), **jusqu¨** (5), **sous** (4), **sur** (12), **leurs** (6), **se** (23), tous (3)
*1_4 (28/66), *1_5 (31/66), *1_6 (28/41), *1_8 (27/61), *2_8 (41/70), *Partie_1 (190/530)

On notera l'évocation fréquente de couleurs et de sensations lumineuses: blanc+, bleu+, brill+, color+, couleur+, eclair+, eclat+, lumiere+, lumineus+, pal+, rayon+, rose, roug+, vert+.

De même, l'évocation d'éléments de la nature, d'un monde aérien: aile+, arbre+, bois+, ciel+, colline+, eau, feu, feuill+, fleur+, fleuve, herb+, horizon+, jardin+, lune+, matin+, mont+, montagne+, nuages, paterre+, plante+, rocher+, rustique, sauvage+, serpent+, soleil, source, terrasse, terre+, treill+, vaste, verdure.

Certains mots expriment l'indistinct, l'émergence de formes, la création: combinaison, creation+, distingu+, ebauch+, elanc+, epanoui, figur+, form+, germe, model+, nouvelle, represent+, ressembl+, trac+, trait+, trouble+, vari+.

Enfin, d'autres mots évoquent des êtres sans nom propre, plus éthérés que réels: cheveu+, enfant+, femme+, fill+, figur+, front, peupl+, pied+, revet+, robe, vetement, yeux.

5.2 Les u.c.e. les plus représentatives

L'extraction des u.c.e. les plus représentatives de chaque classe permet d'appréhender le sens des classes à l'aide de phrases réelles extraites du corpus. Nous tenterons au §5.3 d'en extraire une interprétation selon , trois types de "monde" qui semblent se dessiner dans l'œuvre analysée.

NOTE: chaque u.c.e. est précédée de son numéro d'ordre dans le corpus et du chi2 d'association à la classe (1dl). Le choix est effectué par ordre décroissant du chi2.

```
****** CLASSE NUMERO :  1 ******

103 24 je chantais en marchant un hymne mysterieux dont je
croyais me souvenir comme l'ayant entendu dans quelque autre
existence,

870 18 je continuai ma route" et j'arrivai aux galeries du
palais Royal.

876 18 de la, je sortis des galeries et je me dirigeai vers la
rue saint-Honore.

524 14 je me mis a parler avec violence, expliquant mes griefs
et invoquant le secours de ceux qui me connaissaient.

847 14 j'allai ensuite visiter les galeries d'osteologie.
```

888 14 des medecins vinrent alors, et je continuai mes discours sur l'impuissance de leur art.

1015 14 une nuit, je parlais et chantais dans une sorte-d'extase.

470 13 les personnes les plus cheres qui venaient me voir et me consoler me paraissaient en proie a l'incertitude,

798 13 on termina ensuite la priere, et le pretre fit un discours qui me semblait faire allusion a moi seul.

****** CLASSE NUMERO : 2 ******

475 31 "eh bien, me dis je, luttons contre l'esprit fatal, luttons contre le dieu lui meme avec les armes de la tradition et de la science.

969 19 "cette pensee me rassura, mais ne m'ota pas la crainte d'etre a jamais classe parmi les malheureux.

395 17 quand au peuple, a tout jamais engrene dans les divisions des castes, il ne pouvait compter ni sur la vie, ni sur la liberte.

406 17 ici ma memoire se trouble, et je ne sais quel fut le resultat de cette lutte supreme.

541 17 c-est un de ces rapports etranges dont je ne me rends pas compte moi meme et qu'il est plus aise d'indiquer que de definir;

563 17 "cependant pouvons nous rejeter de notre esprit ce que tant de generations intelligentes y ont verse de bon ou de funeste? l'ignorence ne s'apprend pas.

****** CLASSE NUMERO : 3 ******

358 24 les figures arides des rochers s'elancaient comme des squelettes de cette ebauche de creation, et de hideux reptiles serpentaient,

942 24 des combinaisons de cailloux, des figures d'angles, de fentes ou d'ouvertures, des decoupures de feuilles, des couleurs, des odeurs et des sons,

117 23 d'immenses cercles se tracaient dans l'infini, comme les orbes que forme l'eau troublee par la chute d'un corps;

336 23 la maison ou je me trouvais, situee sur une hauteur, avait un vaste jardin plante d'arbres precieux.

367 23 les variations se succedaient a l'infini, la planete s'eclairait peu-a-peu, des formes divines se dessinaient sur la verdure et sur la profondeur des bocages,

258 20 la se promenaient et jouaient des jeunes filles et des enfants

302 20 je me vis dans un petit parc ou se prolongeaient des treilles en berceaux chargees de lourdes grappes de raisins blancs et noirs;

5.3 Essai d'interprétation de l'exemple traité

Au vu des résultats, trois types de "monde" semblent se dessiner dans cette œuvre :

Le monde réel : Paris et ses environs ; les amis, les parents, les inconnus ; les rues où Gérard de Nerval erre, des nuits durant, en proie à l'ivresse ou à la dépression.

Le monde symbolique, à la fois mystique et rationnel, celui auquel Gérard de Nerval confie ses doutes et ses interrogations sur la vie et la religion.

Et enfin, le monde imaginaire, celui des rêves, lié à l'évocation de la nature et des "forces végétantes" (pour reprendre un terme de Bachelard), monde des sensations (visuelles surtout), lieu d'un désir premier qui, chez Gérard de Nerval, prend le nom d'Aurélia.

6 Conclusion méthodologique

Notre démarche ressemble davantage à la démarche d'un cartographe, qu'à celle d'un chercheur d'or. Il s'agit d'abord d'explorer un monde inconnu dans ses principaux reliefs; avant de tenter de s'y frayer un chemin, en fonction de ses intérêts, en fonction aussi des aléas du terrain, pour trouver l'or du sens convoité.

Lorsque l'on aborde le "contenu" d'un corpus, on ne peut espérer aboutir à autre chose qu'aux résultats d'une interférence entre deux représentations, celle de l'auteur, celle du lecteur, interférence due à la plus ou moins grande sensibilité du lecteur à réagir à tout une série d'indices épars dans le texte, dûe aussi au renforcement statistique de ces indices au cours de la lecture.

La méthode proposée permet une première approche objective de ces répétitions pour appréhender les modes de représentation, que nous appelons des "mondes", ceux qui sont le plus souvent présents chez l'auteur (qu'il en ait ou non conscience).

Il est intéressant de noter que les classes obtenues à partir des mots pleins, lors de l'analyse de différents corpus, discriminent généralement de nombreux mots outils (notamment, les pronoms personnels; mais il est assez bien connu que ces mots outils carctérisent différents genres, où se rencontrent des dialogues). Ce phénomène semblerait infirmer l'hypothèse que ces mots ne joueraient qu'un rôle syntaxique. Nous croyons plutôt que la syntaxe même d'une phrase n'est pas indépendante du choix des mots pleins qui la constituent. Cette expérience nous incite à penser que si la syntaxe elle-même n'est élaborée qu'à un stade terminal de la mise en forme d'une production de langage, il existe néanmoins une présyntaxe, déjà opérante dans les structures plus profondes de la langue.

Les techniques utilisées par nous sont, certes, encore très archaïques et doivent pouvoir être fortement améliorées, notamment en développant des outils d'analyse plus fins, ne serait-ce que pour décrire certaines caractéristiques syntaxiques ou plus simplement séquentielles (lois d'ordre entre les mots) des classes d'unités de contexte extraites. C'est dans cette voie que nous pensons poursuivre ultérieurement cette recherche.

Références bibliographiques

[1] Benzécri, J.-P. et collaborateurs, *l'Analyse des Données*; DUNOD, Paris, (1973).

[2] Benzécri, J.-P. et collaborateurs, *Pratique de l'Analyse des Données en Linguistique et Lexicologie*; DUNOD, Paris, (1981).

[5] Lebart, L., Salem,A., *Analyse statistique des données textuelles*; DUNOD, Paris, (1988).

[6] Reinert, M., *Analyse de deux corpus verbaux et Présentation d'un programme de classifiation descendante hiérarchique*, Thèse de 3ème cycle, (Université Pierre et Marie Curie, Paris VI, 1979).

[7] Reinert, M., *Un logiciel d'analyse des données textuelles : ALCESTE.* Communication aux Cinquièmes Journées Internationnales "ANALYSE DE DONNEES ET INFORMATIQUE", organisé par l'INRIA,(1987).

[8] Reinert, M., Une méthode de classification descendante hiérarchique, *Cahiers de l'Analyse des Données*, Vol VIII, n° 3, pp. 187-198, (1983).

[9] Reinert, M., Classification descendante hiérarchique : un algorithme pour le traitement des tableaux logiques de grandes dimensions. in *DATA ANALYSIS AND INFORMATICS* ; NORTH-HOLLAND, Amsterdam, pp. 23-28 (1986).

[10] Reinert, M., Un logiciel d'analyse lexicale: [ALCESTE], *Cahiers de l'Analyse des Données*, Vol XI, n° 4, pp. 471-484, (1986).

[13] Lebart, L., Exploratory Analysis of Large Sparse Matrices with Application to Textual Data; *COMPSTAT*, Physica Verlag, pp. 67-76, (1982).

[14] G. DE NERVAL, *Œuvres*, tome 1. Bibliothèque de la Pléiade, (1974).

[15] Chartron, G., *Analyse des corpus de données textuelles, sondage d'un flux d'informations*, thèse de doctorat en traitement de l'information de l'Université de Paris VII, (1988).