Bulletin de Méthodologie Sociologique
http://bms.sagepub.com/

Alceste une méthodologie d'analyse des données textuelles et une application: Aurelia De Gerard De Nerval

Max Reinert

Bulletin de Méthodologie Sociologique 1990 26: 24
DOI: 10.1177/075910639002600103

The online version of this article can be found at:
http://bms.sagepub.com/content/26/1/24

Published by:
Association Internationale de Methodologie Sociologique

RC33 Logic and Methodology in Sociology

RC33

and

http://www.sagepublications.com

Additional services and information for *Bulletin de Méthodologie Sociologique* can be found at:

Email Alerts: http://bms.sagepub.com/cgi/alerts

Subscriptions: http://bms.sagepub.com/subscriptions

Reprints: http://www.sagepub.com/journalsReprints.nav

Permissions: http://www.sagepub.com/journalsPermissions.nav

Citations: http://bms.sagepub.com/content/26/1/24.refs.html

Downloaded from bms.sagepub.com at PORTLAND STATE UNIV on September 15, 2014

>> Version of Record - Mar 1, 1990

What is This?

Downloaded from bms.sagepub.com at PORTLAND STATE UNIV on September 15, 2014

Bulletin de méthodologie sociologique, n. 26, March 1990, pp. 24-54

ALCESTE
UNE METHODOLOGIE D'ANALYSE DES DONNEES TEXTUELLES ET UNE APPLICATION: *AURELIA* DE GERARD DE NERVAL

Max Reinert
(CNRS UA-259, Université de Toulouse-Le Mirail, 5 allées Antonio Machado, 31058 Toulouse cedex;
tél 61.44.48.37)

Abstract. **ALCESTE - A Methodology of Textual Data Analysis and an Application: *Aurélia* by Gérard de Nerval.** Beginning with a cross-tabulation with different all sentence fragments in rows and a selected vocabulary in columns for a specific corpus, the author presents: the methodology, including principle concepts and objectives of this form of analysis; the technique, the ALCESTE computer program of automatic classification based on resemblance or dissimilarity; and an application, the analysis of Gérard de Nerval's text Aurélia. The analysis distinguishes three types of fragments which are described and analyzed further. **ALCESTE. Textual Analysis. Gérard de Nerval. *Aurélia*. Hierarchical Descending Classification.**

Résumé. Basé sur un tableau à double entrée comprenant en ligne les différents énoncés et en colonne le vocabulaire retenu d'un corpus, l'auteur présente: la méthodologie, avec les pricipaux concepts utilisés et les objectifs de cette analyse; la technique, le programme d'ordinateur ALCESTE de classification automatique basée sur la ressemblance ou la dissemblance; et une application, l'analyse du text Aurélia de Gérard de Nerval. L'analys permet de distinguer trois classes d'énoncés qui sont décrites et interprétées. **ALCESTE. Analyse textuelle. Gérard de Nerval. *Aurélia*. Classification hierarchique descendante.**

"Tu m'attribues ce qui n'est pas dans le livre; ce que tu y as, crois-tu, deviné."
Aely d'E. Jabès

"Il y a , dans tout livre, une zone d'obscurité, une épaisseur d'ombre qu'on ne saurait évaluer et que le lecteur découvre peu à peu. Elle l'irrite mais il sent bien que là se tient le livre réel autour duquel s'organisent les pages qu'il lit."
Du désert au livre d'E. Jabès et M. Cohen

INTRODUCTION

L'analyse des données textuelles est actuellement une méthodologie en développement. Plusieurs logiciels sont présentés sur le marché ou existent à l'état de prototype: SPAD-T de L. Lebart et LEXICLOUD, d'A. Salem (Lebart & Salem 1988); LEXINET et LEXIMAPPE de G. Chartron (1988) et B. Michelet (1988). Nous-même avons mis au point, d'abord sur les ordinateurs du centre de calcul de Toulouse (CICT), et maintenant sur Macintosh (Mac II), un logiciel ALCESTE (Analyse Lexicale par Contexte d'un Ensemble de Segments de Texte, *BMS* n. 25, p. 58) qui a été présenté en plusieurs occasions (Reinert 1979, 1983, 1986 & 1987). Cette étude est davantage orientée vers une réflexion plus théorique sur les hypothèses

Downloaded from bms.sagepub.com at PORTLAND STATE UNIV on September 15, 2014

sous-jacentes à la méthodologie proposée ainsi que sur la signification des interprétations. C'est une manière de faire le point sur une approche que nous poursuivons depuis 1979 et sur les choix qu'elle présuppose.

Dans la première partie, nous présentons la méthodologie en fonction de cette approche. La seconde est consacrée à la description des techniques mises au point (logiciel ALCESTE) et à l'exposé des résultats "bruts" obtenus sur le texte *Aurélia* de Gérard de Nerval. Enfin la troisième partie permettra de préciser concrètement, dans cette application particulière, la manière dont nous nous servons de ces résultats et quelles significations nous leur donnons.

1. LA METHODOLOGIE ALCESTE

1.1. Lois de distribution du vocabulaire et contextes

En tant qu'étude des principales *lois de distribution* du vocabulaire dans un corpus, l'approche proposée entre dans le cadre général de l'analyse de données en linguistique qui, comme le fait remarquer J.P. Benzécri (1981, p. 4), prolonge l'approche distributionnelle de Bloomfield et Z.S. Harris.

En un certain sens seulement, car si l'*"On entend par distribution l'ensemble des environnements (des contextes) dans lequel on peut trouver une unité"* (Pottier 1973), pour L. Bloomfield, la notion d'environnement était associée à une procédure de segmentation des phrases en constituants immédiats qui a eu pour conséquence d'infléchir ce type d'approche vers une analyse mathématique de la structure syntaxique de la langue (Harris 1971).

Cela dit, une grande imprécision semble régner dans la définition de ce que peut être un contexte, ne fût-il que linguistique: *"Le terme de contexte ou celui d'environnement qu'emploie Z.S. Harris, offre la même ambiguïté que celui d'unité ou d'élément. En toute rigueur, le contexte d'un mot, c'est le texte moins le mot. Parler de contexte d'un élément suppose donc qu'on ait défini des unités d'ordre supérieur à l'élément qui seront à celui-ci ce que le texte est au mot."* (Benzécri 1973, pp. 5-6).

De son côté, F. Rastier (1987, p. 72) constate que *"si l'on a fréquemment recours à la notion de contexte linguistique, on omet presque aussi souvent de la définir"*. Il propose pour sa part *"trois paliers de description du contexte: le syntagme minimal (ou mot), l'énoncé, et son au-delà textuel."* (p. 73).

Dans l'approche que nous proposons, la définition de l'environnement est étendue à une unité de l'ordre de l'énoncé minimal, que nous appelons *"unité de contexte" (u.c.)*. En effet, si nous désirons définir ces u.c. dans un cadre purement linguistique, nous désirons aussi pouvoir les interpréter, dans un cadre plus cognitif, comme des *représentations élémentaires*. Au plan du fonctionnement du langage, ce choix se justifie, dans la mesure où, comme le souligne J.F. Le Ny (1989, p. 83), *"un contexte linguistique joue,*

Downloaded from bms.sagepub.com at PORTLAND STATE UNIV on September 15, 2014

lors de la compréhension d'une unité (par exemple un mot) à l'intérieur d'un énoncé, exactement le même rôle qu'une situation par rapport à une représentation d'objet". Cet auteur ajoute: *"C'est d'ailleurs pour cette raison que le mot "contexte" est si extensivement, et parfois si abusivement employé. "Contexte" signifie originellement, comme son étymologie l'indique, "contexte linguistique""*. La confusion entre contexte linguistique et contexte de situation, si elle peut être évitée, dans l'étude d'un corpus fini particulier, se justifie en partie, lorsque l'on doit appréhender, des corpus non finis comme par exemple, les productions langagières d'un groupe social donné à une époque donnée. Dans ce cas la notion de contexte interfère avec la notion d'usage et de représentation collective.

Cela nous conduit à distinguer deux aspects dans la notion de contexte: l'un en rapport avec celle d'environnement d'un mot dans un texte, et que nous avons cherché à opérationnaliser avec la notion d'unité de contexte; l'autre en rapport avec celle d'usage, qui dans un corpus particulier, pourrait être opérationnalisée par la recherche de régularités dans la distribution du vocabulaire dans ces *unités de contexte*. Nous appellerons *"contexte-type"* ce type de régularité. Si *l'unité de contexte* renvoie, au niveau cognitif, à une représentation, singulière, le *contexte-type* renvoie, en tant que classe de contextes locaux, à une représentation collective.

Si cette notion de *contexte type* reste ambiguë tant que n'est pas précisée la collectivité à laquelle elle se réfère (elle interfère avec la notion d'usage), nous la définissons ici dans un cadre strictement linguistique en nous référant à un corpus fini donné.

En ce qui concerne l'aspect méthodologique, il reste plusieurs questions à élucider. Comment opérationnaliser, à l'aide d'un algorithme, la notion d'unité de contexte? Comment définir ensuite des "similarités" entre *u.c.* pour dégager des *"contextes type"*? Le *contexte type* d'un mot en tant que trace dans un corpus donné d'un usage; c'est-à-dire, d'une forme commune de représentation, ne devrait-il pas être élargi à l'ensemble des mots susceptibles d'être activés par cette représentation? Dans ce cas, si un *"contexte type"* n'est plus caractéristique d'un mot, mais d'un ensemble de mots, pourquoi ne pas définir de même la notion d'*unité de contexte* indépendamment des unités lexicales qu'elle est supposée inclure?

La réponse à ces questions occupe toute cette étude. Auparavant, supposons le problème résolu, et remarquons qu'une fois effectuée la segmentation du corpus en un ensemble d'*unités de contexte*, un sens précis peut être accordé à l'étude des distributions puisque ces distributions sont représentables avec une bonne approximation par un tableau de données à double entrée comprenant, en ligne, les *unités de contexte* (*u.c.* pour simplifier) et, en colonne, le *vocabulaire* retenu avec, à l'intersection d'une ligne et d'une colonne: *"1"* si le mot est présent dans l'u.c.; *"0"*, sinon.

Nous nous éloignons ainsi du modèle classique de la notion de *distribution d'un mot*, puisque nous perdons la syntaxe des phrases, mais nous affirmons que d'un certain point de vue (celui de l'approche statistique de *contextes types* et non de l'approche sémantique de *contextes locaux*), cette perte est faible compte tenu de la dimension réduite de ces *unités de*

Downloaded from bms.sagepub.com at PORTLAND STATE UNIV on September 15, 2014

contexte: la simple apparition de mots dans une unité devient, en elle-même, très significative des liens pouvant unir ces mots. Par exemple, si nous considérons qu'une *u.c.* contient environ 20 mots, que le vocabulaire retenu comprend 800 mots, et si nous codons par la valeur "1" l'apparition d'un mot dans une *u.c.* et par la valeur "0", sa non-apparition, alors le tableau de données associé comprendra plus de 97 % de zéros: c'est tout dire sur la signification statistique de la simple présence d'un vocable dans une *u.c.*

En résumé, nous proposons de modéliser les *lois de distribution du vocabulaire* dans un corpus, à l'aide d'un tableau à double entrée croisant *unités de contexte* (u.c.) et *vocabulaire retenu*. D'où une première reformulation de nos questions: Quel corpus est susceptible d'un tel traitement? Quel découpage en *unités de contexte* choisir et comment? Quel vocabulaire retenir pour établir les liens entre u.c., et surtout, pour quel résultat?

1.2. La définition du corpus

Le choix d'étude d'un corpus présuppose ... que ce corpus constitue bien un *"objet d'étude"*; c'est-à-dire, l'analyste le perçoive comme une entité ou un *"objet"* dans l'univers référentiel qui l'intéresse. En définitive, même si ce n'est que de manière implicite, l'analyste fait des hypothèses sur les conditions d'existence de cet objet, sur ses lois de production, sur les paramètres qui le font reconnaître dans cet univers référentiel.

Par exemple, pour le sociologue qui étudie un corpus de réponses à une question ouverte, ce sont les caractéristiques sociales des individus qui pourront servir à "paramétrer" l'objet d'étude. Pour notre part, nous ne suggérons qu'une approche au niveau linguistique de type "structural" (si l'on élargit ce terme à une conception non seulement logique de la structure mais aussi statistique), dans la mesure où cette dernière consiste, comme le suggère R. Thom (1974, p. 21) à *"réduire l'arbitraire de la description du corpus en mettant en évidence ses régularités, ses symétries cachées"*. Autrement dit, on décrit des lois dans un cadre purement linguistique.

Mais cela ne veut pas dire, bien au contraire, que ces lois ne connotent pas des aspects extra-linguistiques sur lesquels il est possible d'inférer des hypothèses. Ces hypothèses sont licites, en ce sens que ce sont elles qui donnent un sens à l'objet d'étude, dans la mesure où les conditions de production du corpus peuvent être maîtrisées. On rentre d'ailleurs là, dans le cadre très général de toute expérimentation scientifique ("En quoi une variation des lois de production influe-t-elle sur la structure du corpus?").

Malgré cela, même l'approche structurale dénote un choix d'objet ou, plus exactement, un choix du type d'objet. Il est clair que ce que nous négligeons ou prenons en compte dans la modélisation de la structure linguistique du corpus influe sur le type d'objet que nous pourrons observer au travers de cette modélisation, ou du moins sur une qualité de sa représentation. C'est en cela que la signification des *"unités de contexte"*,

Downloaded from bms.sagepub.com at PORTLAND STATE UNIV on September 15, 2014

de leur relation avec l'objet d'étude, ainsi qu'avec le *vocabulaire retenu*, doit être précisée.

1.3. Unité de contexte, énoncé et proposition

Par ce découpage du corpus en u.c., on retrouve les caractéristiques habituelles d'un tableau de données avec en ligne, les *objets* et en colonne, les *attributs* de ces objets, si l'on identifie, d'une part, *unité de contexte* et *objet* et, d'autre part, *mot* et *attribut* . Qu'est ce qui peut justifier une telle individuation de ces unités? Qu'ont-elles en propre que ni le corpus, ni le mot n'ont? Pour répondre à ces questions, nous envisagerons d'abord en quoi la notion d'*énoncé* se distingue, sur le plan sémantique, de la notion de *mot*. Nous chercherons ensuite à montrer en quoi la définition des *unités de contexte* permet d'opérationnaliser la notion d'*énoncé*.

La définition la plus commune de l'*énoncé* (selon le *Lexis*) est: *"proposition, phrase, dans laquelle une pensée est énoncée"*, une pensée étant "*un acte particulier de l'esprit qui se porte sur un objet*". Sans revenir jusqu'à Aristote, la logique mathématique montre qu'à une proposition logique est associée une valeur de vérité que ne possèdent pas les termes de la proposition. Plus généralement, une proposition ou un énoncé dans le langage courant nous apprend "quelque chose" sur ce qui est, ou n'est pas pour quelqu'un. La notion d'énoncé renvoie à la notion d'idée en tant que cette "idée" est le reflet pour un individu donné d'une chose donnée, à un moment donné.

A propos de *"Qu'est-ce qu'une chose?"*, Heidegger écrit (1971, p. 55): *"est-ce un simple hasard si la détermination de l'essence de la chose, et celle de l'essence de la proposition et celle de l'essence de la vérité s'accomplissent simultanément, ou bien ces déterminations sont-elles toutes, l'une par rapport à l'autre, en état de connexion nécessaire ?"*, et plus loin: *"une chose est le support de propriétés, et la vérité qui lui correspond a son site dans l'énoncé, dans la proposition qui est jonction d'un sujet et d'un prédicat"*.

Si la notion de vérité caractérise la sémantique de l'énoncé par rapport à celle du mot dans l'appréhension rationnelle du monde par un sujet, cette notion ne semble être qu'une modalité parmi d'autres de l'attitude d'un sujet, dont par exemple, l'affirmation, la négation, la requête, le jugement. **En définitive, dans la sémantique de l'énoncé, et contrairement à la sémantique du mot, il y a la marque d'un sujet en tant qu'individu psychique.**

L'énoncé *"le ciel est bleu"*, ne peut être confondu avec *"le ciel bleu"*, en ce sens que dans un cas, il y a affirmation, dans l'autre non. Autrement dit, ce qui a valeur de vérité pour celui qui énonce, ne peut être réduit aux seules dimensions de ce qui est représenté.

En résumé, la trace du locuteur dans ces énoncés est le résultat d'une interférence entre deux entités: le monde et soi. Cette double dimension réalité/psyché est présente dans la *notion de vérité* mais la dépasse en exprimant une *attitude* générale d'un sujet face au monde. Elle se noue

Downloaded from bms.sagepub.com at PORTLAND STATE UNIV on September 15, 2014

dans l'*énoncé* par ce que nous appelons une idée ou une *représentation*. Dans cette perspective, une idée n'est pas simplement liée à la représentation d'un objet, mais elle est liée à la manière dont un sujet l'appréhende en fonction de sa propre identité, en fonction aussi de son intention. Le sens d'un énoncé est donc toujours double, puisqu'il se réfère et à un *"objet"* et à un *"sujet"*. Ce n'est que dans cette double référence qu'une *représentation* peut prendre corps. **C'est en cela que la trace linguistique de l'énoncé constitue la plus petite unité de texte susceptible de décrire, selon nous, la représentation sous-jacente d'un sujet.**

Cette manière de voir peut être rapprochée de la manière dont certains chercheurs en intelligence artificielle ont distingué dans la représentation d'un objet (par un robot par exemple), deux aspects duaux: l'un concernant la structure de cet objet, l'autre concernant sa fonction dans l'objectif à atteindre (Minski 1988, pp. 156-157). On retrouve bien là deux aspects fondamentaux de ce qui peut caractériser une représentation élémentaire: l'un descriptif et concernant le monde objectif, et l'autre relationnel et concernant le monde subjectif (limité pour un robot à l'objectif à atteindre).

Au niveau de l'opérationnalisation de la notion d'énoncé, il est bien difficile de reconnaître dans un segment de texte particulier, où s'arrête et où commence un énoncé. Aucune procédure linguistique exacte ne permet d'affirmer qu'un segment de texte est un énoncé ou ne l'est pas, même si nous sentons que, puisque un énoncé comprend au minimum un sujet et un prédicat, il doit généralement être associé à des structures linguistiques permettant d'exprimer cette forme (la proposition).

Aussi notre stratégie pour définir les *unités de contexte* a été d'abord de définir des contraintes générales auxquelles doivent satisfaire ces unités, et ensuite, dans ce cadre, de procéder à un découpage arbitraire mais susceptible d'être varié, afin d'observer dans les résultats obtenus après telle ou telle variation ce qui est stable ou ne l'est pas. Pour être plus précis, nous avons retenu deux contraintes pour définir une *u.c.*: **la taille** et **la ponctuation**. Dans la mesure du possible, une fin d'*u.c.* correspondra à une fin de phrase, sa longueur étant au maximum de quelques lignes.

1.4. Représentation et contexte

Les liens que nous voulons établir entre *unités de contexte* doivent permettre d'appréhender, à la suite d'une analyse statistique, ce que nous avons appelé des *"contextes types"*, du moins les plus prégnants d'entre eux dans un corpus donné. Dans ce paragraphe, nous prendrons le problème de ce lien sous son aspect plus cognitif, l'énoncé étant la trace d'une représentation (mentale) singulière, et le contexte type étant celle d'une représentation collective.

Dans la mesure où une représentation collective exprime une certaine régularité de structure dans une classe de représentations singulières, l'hypothèse (peu ambitieuse), que nous chercherons ensuite à développer,

Downloaded from bms.sagepub.com at PORTLAND STATE UNIV on September 15, 2014

est que cette régularité est due aux contraintes de ce que nous appelons "un monde".

A propos de la notion de représentation

Pour P. Vergès (Grive, Vergès & Silem 1987), chacun *"entend représentation à sa façon"*. Sans prétendre que *"la chose est illégitime, elle complique cependant le travail et exige de ceux qui veulent faire usage de la notion qu'ils commencent par bien préciser en quel sens ils vont s'en servir"*. Nous nous conformerons à ce sage conseil, en partant de la définition du *Lexis*: *"image d'un objet donnée par les sens ou par la mémoire"*, sous-entendu *"image mentale"*.

Un cas bien étudié des psychologues est celui où l'image mentale est essentiellement spatiale. Dans certains tests de reconnaissance d'objet physique, dont on a plusieurs photos dans des perspectives différentes, *"tout se passe comme si nous possédions un réservoir tridimensionnel dans lequel nous conservions des modèles réduits des objets qui nous entourent: nous manipulons ces maquettes comme s'il s'agissait des vrais objets eux-mêmes"* (Kosslyn 1980). Ces expériences vont dans le sens d'hypothèses actuelles sur la manière dont fonctionne le cerveau. Pour J.C. Pérez (1988), *"la représentation des connaissances, donc la mémoire, serait peut-être de nature holographique"*.

D'autre part, la relation entre représentations imagées et représentations sémantiques a été étudiée par M. Denis (1982), dont l'hypothèse de travail est qu'*"il n'y a pas de séparation stricte entre l'approche conceptuelle du problème des représentations imagées et celle du problème des représentations sémantiques. Plus précisément, nous avons développé l'idée que les images constituent un mode d'actualisation des représentations sémantiques des énoncés ..."*.

Quoiqu'il en soit, nous voulons insister sur ce fait bien connu des cogniticiens qu'une représentation mentale n'a pas qu'un substrat linguistique, le problème étant dans la nature des liens entre ces représentations et la langue.

Pour R. Thom (1974, p. 120), *"toute théorie de production verbale (psycholinguistique) soulève nécessairement le problème philosophique de savoir s'il existe une pensée préverbale, dont le langage ne serait qu'une manifestation extérieure"*. Pour cet auteur, il ne fait d'ailleurs pas de doute sur le fait qu'*"il existe chez l'homme des modes de pensées non verbaux, qu'il partage avec les animaux. Parmi ces activités premières, la représentation sensorielle du monde qui nous entoure est fondamentale (...) Nous admettrons donc que l'hypothétique structure profonde des linguistes est constituée essentiellement de notre représentation sensorielle du monde extérieur. Au contraire, la structure de surface sera constituée des automatismes du langage proprement dits."*

Selon cette perspective, la notion de *"représentation"* renvoie à une forme préverbale en référence à la structure profonde, alors que la notion de

Downloaded from bms.sagepub.com at PORTLAND STATE UNIV on September 15, 2014

"contexte linguistique" renvoie à la trace linguistique que cette forme prend dans la langue, et, pour reprendre l'image très suggestive de R. Thom, selon *"un processus d'exfoliation permanent, à la manière de la peau, constituée de couches de cellules sécrétées par le derme profond et qui vont en se sclérosant vers l'extérieur, où elles se désagrègent."*

Représentation et type de monde

De même que nous avons distingué les *environnements locaux* des *types de contexte*, nous devrions distinguer une représentation psychique partielle, activée à un moment donné (par un énoncé, par exemple), d'un type de représentation autour duquel fluctuent plusieurs de ces représentations locales.

Ces représentations psychiques "locales" dépendent de la possibilité de focaliser à un moment donné son attention sur des événements particuliers, variés, et de les extraire momentanément d'un environnement aux contours plus flous, en vue de certaines actions. Elles supposent donc, par contraste, l'existence de cet environnement qui, bien que non perçu par la conscience, n'en est pas moins perçu de manière globale et imprécise, comme le fond d'un tableau duquel se détache un sujet particulier. Une part du mystère du sourire de la Joconde ne vient-il pas des lointains indistincts du fond du tableau?

Au niveau biologique, on retrouve cette dualité fond/motif à propos des fonctions de l'oeil. C'est le cas, par exemple, lorsque l'on distingue la vision, qui est une perception globalisante de l'environnement, du regard, qui permet de séparer un élément de cet environnement à l'aide d'un mouvement de poursuite occulaire.

Aussi nous semble-t-il légitime de distinguer les *représentations locales* venant dans le champ de la conscience à un moment donné, d'une *représentation globale* associée à tout un environnement, constituant le fond d'où elles émergent. Si les *représentations locales* sont, par nature, changeantes, instables, évanescentes, la *représentation globale* dont elles dépendent, paraît plus stable dans le temps bien qu'elle soit perçue de manière moins différenciée: elle constitue un cadre perceptivo-cognitif cohérent constituant ce que l'on peut appeler *"un monde"*, substrat d'où semble émerger et prendre sens chaque objet singulier.

C'est ce terme de *"monde"* que nous avons choisi pour désigner cette organisation psychique donnant une cohérence à la multiplicité de nos expériences sensorielles et motrices.

Au niveau d'un corpus de textes, nous pouvons de même distinguer les énoncés particuliers de la langue, associés à des mises en conscience locale de certains aspects du "monde", d'un ensemble d'énoncés connotant une même perception globale d'un *monde*.

Un *contexte type* servira donc à appréhender un type de monde, le terme *"monde"* convenant mieux à notre avis, que le terme "représentation", par la

Downloaded from bms.sagepub.com at PORTLAND STATE UNIV on September 15, 2014

non-différenciation qu'il suppose, traduisant bien l'aspect nocturne, inconscient de cette forme de représentation Comme le dit le poète E. Jabès: "*Le monde, dans l'homme, est une foule de mots réclamés*".

En conclusion, si ces idées sont celles qui nous ont guidé tout au long de notre démarche, elles ne prennent un sens véritablement précis que dans le cadre de l'analyse d'un corpus particulier. Notre hypothèse de travail est la suivante: les représentations locales, multiples, immédiates, associées aux énoncés d'un corpus, s'organisent en fonction de lois particulières identifiant des types de *"mondes"*. Leur trace dans la langue ne peut être révélée qu'au travers d'un grand nombre d'énoncés, semblables d'un certain point de vue. La représentation que nous pouvons avoir de ces *"mondes"*, au travers du corpus étudié, ne peut être, le plus souvent, qu'archaïque, du fait justement qu'elle n'est pas directement énoncée dans le corpus, et que nous la révélons à travers une analyse statistique assez grossière, par un classement.

Dans cette perspective, d'un point de vue opérationnel, il est logique, de ne pas tenir compte des éléments de la structure linguistique de surface la plus différenciée, éléments trop spécifiques de ce niveau. C'est la raison pour laquelle nous avons désiré, lors d'une première approche, négliger tous les traits syntaxiques. D'une part, il y a les "mots outils" qui, selon la dénomination usuelle, recouvrent les articles, prépositions, conjonctions et pronoms. Pour les pronoms personnels, notre attitude fut plus ambiguë. Dans un premier temps, nous les analysions, mais l'expérience a montré qu'il valait mieux ne pas les mettre directement dans l'analyse du fait de leur trop fort poids dans la définition des classes. On peut toutefois apprécier leur rôle indirectement (voir 2.3). D'autre part, il y a les désinences grammaticales telles que les pluriels, conjugaisons, certains suffixes. Il existe, en effet, entre ces désinences et les mots outils, des lois de transformation ou de substitution comme entre "de manière efficace", "efficacement", "avec efficacité".

En résumé, seule est considérée dans l'analyse des liens entre *unités de contexte*, la distribution des *formes* associables à des *"mots pleins"*; c'est-à-dire, des verbes, noms, adjectifs et adverbes, voire la distribution des *"formes réduites"*, une fois ôtées les désinences grammaticales ou certains suffixes.

1.5. Conclusion

L'orientation générale de cette recherche est celle des recherches effectuées en analyse des données textuelles par J.P. Benzécri, L. Lebart et A. Salem. Il s'agit de décrire, de manière purement formelle, des lois de distribution du vocabulaire dans des textes. Cependant, notre objectif est d'étudier au travers de ces lois de distribution, des types de représentations. Cela nous a conduit, sur le plan méthodologique, à différencier l'approche que nous préconisons notamment sur les deux points.

Le découpage en unités de contexte - Nous aurions aimé appréhender ces représentations à partir d'un découpage du corpus en "énoncés". Toutefois, la notion d'énoncé n'étant pas vraiment opérationnalisable.

Downloaded from bms.sagepub.com at PORTLAND STATE UNIV on September 15, 2014

Plutôt que de chercher à définir un découpage rigoureux du texte en énoncés, 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 (indépendant de ces variations) ne devraient dépendre que faiblement de l'arbitrarité du découpage.

Le 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 et certains suffixes, de manière à conserver la trace d'un contexte le plus large possible, en ne retenant toutefois que les mots pleins. Ces mots, porteurs d'un sens, sont les témoins d'un monde sémantique qu'il s'agit justement de préciser.

2. LES DIFFERENTES ETAPES TECHNIQUES D'UNE ANALYSE

Une analyse comporte schématiquement cinq étapes:

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

2.1. Définition des unités de contexte

Dans cette première étape, l'objectif est la recherche des *unités de contexte (u.c.)* et des *formes* répertoriées, afin de pouvoir étudier les lois de distribution du vocabulaire dans ces unités, à l'aide de tableaux de données à double entrée, avec en ligne l'ensemble des *u.c.* et en colonne les *formes* retenues, avec, à l'intersection d'une ligne et d'une colonne, la valeur "1" si la *forme* est présente dans l'*u.c.* et "0" sinon.

Le seul fichier nécessaire, en début d'analyse, est un fichier comprenant le texte à étudier. Le texte retenu est extrait de 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, les différents paragraphes d'*Aurélia*. Chaque paragraphe 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.

Downloaded from bms.sagepub.com at PORTLAND STATE UNIV on September 15, 2014

*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 ou 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, avec, si possible, le respect des coupures proposées par la ponctuation. Ces segments de texte constituent les *unités de contexte élémentaires (u.c.e.)*. Chaque segment est numéroté et se termine généralement sur une fin de phrase. Leur définition dépend d'un compromis entre la longueur 2, 3 lignes environ et le respect de la ponctuation. 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.

2.2. 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. Cela dit, pour fixer le vocabulaire, nous utiliserons le terme *forme* lorsque nous voudrons insister sur l'aspect formel des opérations effectuées. Par contre, nous utiliserons le terme *mot*, voire *vocable*, lorsque nous voudrons insister sur l'aspect sémantique. Bien entendu, même si les *formes* sont reconnues à l'aide d'opérations formelles, leur utilité est de pouvoir servir de support aux *sens*.

Dans cette 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. Il est possible, de réduire les racines irrégulières les plus courantes, de supprimer les pluriels, les désinences de conjugaison, certains suffixes.

Downloaded from bms.sagepub.com at PORTLAND STATE UNIV on September 15, 2014

L'objectif de cette réduction est de permettre d'enrichir le plus possible les liaisons statistiques impliquées par les cooccurrences des *formes*. Si une *u.c.* contient en moyenne 20 *formes* et que nous en analysions 600, le tableau de données qui aurait, en ligne, ces *u.c.* et, en colonne, 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. 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 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 (pour plus de détails se référer à Reinert, 1986). Cette racine servira de support à l'identification de sens et permet d'identifier une catégorie morpho-sémantique.

Le choix de cette technique nous a semblé plus souple que l'utilisation d'un dictionnaire. La réduction d'un mot n'est effectuée que dans la mesure où elle permet un regroupement et donc dépend de la distribution du vocabulaire dans un corpus donné. D'autre part, cette réduction est applicable à des mots de la langue parlée qui ne sont pas forcément répertoriés dans un dictionnaire, réduction utile lors de l'analyse d'entretiens ou de récits d'enfants, par exemple.

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

Downloaded from bms.sagepub.com at PORTLAND STATE UNIV on September 15, 2014

(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

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 contextes* (10,000 maximum) et *formes réduites* (1,400 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* (les noms, verbes, adjectifs et adverbes) et à considérer comme *formes illustratives*, les *mots outils* (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.* (les différents paragraphes d'*Aurélia*). 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 trois tableaux de données sont constitués. Un fichier numérique est constitué avec un enregistrement par *unité de contexte élémentaire* dans lequel est transcrite la séquence des *formes réduites* retenues, en conservant leur ordre. Deux fichiers numériques sont constitués avec 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" minimum peut être imposée par l'utilisateur en *nombre minimum de formes analysées par u.c.*. Dans l'exemple, un premier tableau a été constitué avec 10 *formes analysées* minimum par *u.c.* et le second tableau avec 15 *formes analysées* minimum.

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

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 *'zeros'*: **98.09 %**

Downloaded from bms.sagepub.com at PORTLAND STATE UNIV on September 15, 2014

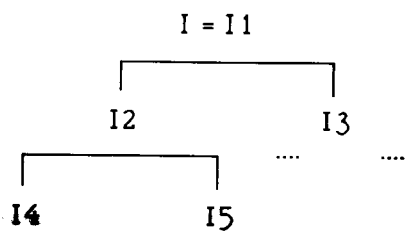
2ème tableau (15 *formes analysées* minimum par *unité de contexte*):
nombre d'*unités de contexte* analysées: **416**
nombre de *formes* analysées: **669**
nombre de *'uns'*: **6921**
pourcentage de *'zeros'*: **97.56 %**

(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

La méthode de classification utilisée

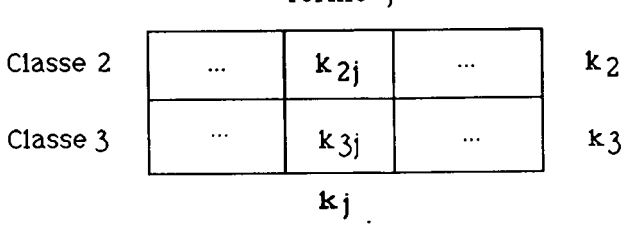
La méthode que nous avons mise au point pour construire ces classes est une méthode de classification descendante hiérarchique. Elle permet de traiter des tableaux logiques (codage "0" ou "1") de grande dimension (4.000 lignes par 1.400 colonnes maximum) mais de faible effectif (60.000 "1" maximum). La procédure proposée se situe au carrefour de plusieurs techniques: segmentation (Bertier & Bouroche 1975), classification hiérarchique (Benzécri 1973), dichotomie d'après une analyse factorielle (*ibid.*), nuées dynamiques (Diday *et al* 1982). Schématiquement, il s'agit d'une procédure itérative. La première classe analysée comprend toutes les *u.c.* 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 la succession d'analyses.



La procédure s'arrête lorsque le nombre d'itérations demandé est épuisé. La méthode de partitionnement d'une classe en deux considètr d'abord 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

Downloaded from bms.sagepub.com at PORTLAND STATE UNIV on September 15, 2014

jième colonne, le nombre k2j d'u.c. de la classe contenant la jième forme identifiée.



Avec, par exemple :

$$k_{2j} = \sum_{i \in I_2} k_{ij} \; ; \; k_2 = \sum_{i \in I_1} k_{2j} \; ; \; k_j = k_{2j} + k_{3j} \; ;$$

le khi2 pouvant s'écrire sous la forme :

$$khi2 = k_2 . k_3 \sum_{j \in J} (k_{2j}/k_2 - k_{3j}/k_3)^2 / k_j$$

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 le chi2 obtenu ne peut en être éloigné. En effet, la technique consiste en: (a) rechercher le premier facteur de l'analyse factorielle des correspondances (AFC) du tableau considéré (espace R^J muni de la métrique du X^2 (voir notation dans Benzécri 1973); (b) rechercher 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. Cet algorithme considère une partition des unités en deux classes quelconque. 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 pour chercher la partition optimale, de partir de la droite optimale qui est justement le premier axe factoriel de l'AFC du tableau.

Le 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. Cette procédure se déroule selon les étapes suivantes: (a) on construit deux tableaux de données avec une définition légèrement

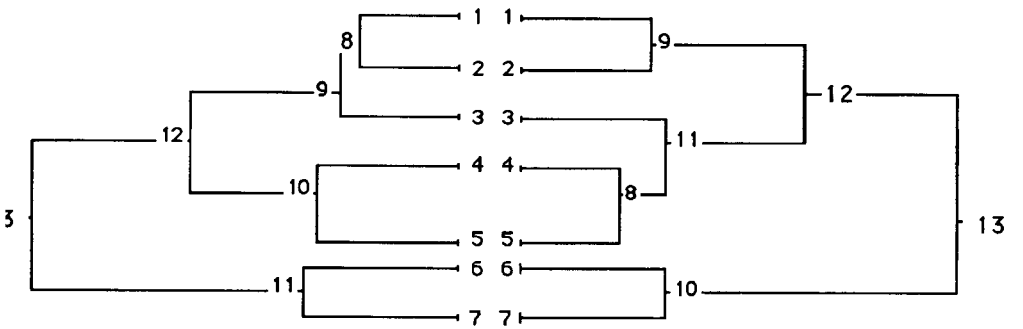
Downloaded from bms.sagepub.com at PORTLAND STATE UNIV on September 15, 2014

différentes des *u.c.*; (b) on effectue la classification des *u.c.* de chaque tableau; (c) on compare les classes obtenues 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 utilisée pour comparer les classes est simple. Elle consiste à calculer les liens entre classes deux à deux (à l'aide d'un chi2), puis, à repérer 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.

Dans le cas étudié, voici les résultats obtenus:

1ère classification (10 mots/u.c.) **2ème classification (15 mots/u.c.)**



Pour les deux analyses, 7 classes terminales ont été demandées, numérotées de 1 à 7, d'où l'existence de 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. La notation 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 *

Downloaded from bms.sagepub.com at PORTLAND STATE UNIV on September 15, 2014

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 à une degré de liberté - ce coefficient étant utilisé dans cette étude comme indice de lien et non pas comme test).

L'objectif est de rechercher, 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 sert ensuite de base au calcul des classes définitives (renumérotées de 1 à 3) à l'aide d'un algorithme du style "centres mobiles" (Diday *et al* 1982). Cet algorithme consiste schématiquement, à chaque pas, à calculer le centre de gravité des classes et à réaffecter les unités dans la classe dont le centre est le plus proche, ceci jusqu'à ce que toutes les unités soient bien classées. L'algorithme utilisé est une variante: après avoir rejeté les unités non caractéristiques (trop proche du centre de gravité de l'ensemble), on effectue ce calcul dans le même espace que précédemment (espace R^J muni de la métrique du X^2), jusqu'à un maximum local, le critère utilisé étant l'inertie inter-classe (il s'agit donc du même critère que celui utilisé lors de la classification de l'étape précédente). L'expérience montre que les classes intersections obtenues lors de l'étape précédente sont généralement très stables. Cet algorithme permet cependant d'affecter un certain nombre d'unités qui se trouveraient, sinon, rejetées de l'analyse.

2.5. Aides à l'interprétation des classes

Pour analyser la structure des classes d'u.c. extraites, plusieurs procédures peuvent être utilisées. Nous en retiendrons deux, qui nous semblent les plus suggestives: (1) le relevé du vocabulaire le plus spécifique de la classe retenue; (2) l'extraction des u.c. de la classe les plus représentatives de ce vocabulaire.

Downloaded from bms.sagepub.com at PORTLAND STATE UNIV on September 15, 2014

La 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 (résultats commentés dans la deuxième partie).

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), plus (2), plusieurs (7), quelque (7), rien (8), **tout-a-coup** (14), tres (6) accompli+ (4), achet+ (9), ami+ (26), approch+ (4), **arriv+** (15), art‹ (2), aurelia (4), avou+ (7), ayant (6), a-travers (2), 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), famille+ (2), fatigu+ (4), fievre+ (3), galerie+ (6), georges (4), hasard (5), heure‹ (3), idee+ (6), impression+ (2), 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), triste+ (2), vague+ (5), veille (2), vierge (2), **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), etais (2), soit (2)*
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)*

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), annee+ (2), apparence+ (3), arme+ (7), **bien** (27), bon (2), 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), ecriai*

Downloaded from bms.sagepub.com at PORTLAND STATE UNIV on September 15, 2014

*(2), effet (2), eloim (9), envahi+ (7), **epreuve+** (11), **eprouv+** (14), esprit+ (6), eternel+ (3), etrang+ (8), etres (2), 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), memoire (2), 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), sience‹ (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), notre (2), nous (3), ton (2), **tu** (5), cela (3), ces (3), **telle** (4), dont (3), **ai** (10), **as** (6), **avons** (9), **est** (14), furent (2), **fut** (10), **serait** (4), **sommes** (6)*
****1_1** (28/41), ***2_1** (46/72), *2_9 (16/34)*

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), habit+ (2), 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)*

Downloaded from bms.sagepub.com at PORTLAND STATE UNIV on September 15, 2014

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. 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.
987 13 on crie, on chante, on rit aux eclats; c-est gai ou triste a entendre, selon les heures et selon les impressions.
1129 13 tout-a-coup, o merveille! je me mis a songer a cette auguste soeur de l'empereur de russie, dont j'ai vu le palais imperial a weimar.
101 11 mon ami m'avait quitte, voyant ses efforts inutiles, et me croyant sans-doute en proie a quelque idee fixe que la marche calmerait.
658 11 je me trouvais dans une salle inconnue et je causais avec quelqu un du monde exterieur, l'ami dont je viens de parler, peut-etre.
83 10 " et pendant qu'il m'accompagnait, je me mis a chercher dans le ciel une etoile, que je croyais connaitre,
85 10 l'ayant trouvee, je continuai ma marche en suivant les rues dans la direction desquelles elle etait visible,

******** 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;

Downloaded from bms.sagepub.com at PORTLAND STATE UNIV on September 15, 2014

563 17 cependant pouvons nous rejeter de notre esprit ce que tant de generations intelligentes y ont verse de bon ou de funeste? l'ignorance ne s'apprend pas.
319 15 oh! ne fuis pas m'ecriai je; car la nature meurt avec toi!
549 15 le systeme fatal qui s'etait cree dans mon esprit n'admettait pas cette royaute solitaire;
1014 15 bien des lettres manquent, bien d'autres sont dechirees ou raturees; voici ce que je retrouve:
112 14 si je ne pensais que la mission d'un ecrivain est d'analyser sincerement ce qu'il eprouve dans les graves circonstances de la vie,
193 14 ne te hate pas, dit il, de te rejouir, car tu appartiens encore au monde d'en-haut et tu as a supporter de rudes annees d'epreuves.
209 14 nous sommes sept, dis je a mon oncle./ c-est en effet, dit il, le nombre typique de chaque famille humaine, et, par extension, sept fois sept, et davantage;
276 14 l'un d'eux me dit en pleurant: "n'est ce pas que c-est vrai qu'il-y-a un dieu? oui! " lui dis je avec enthousiasme.
426 14 des circonstances fatales preparerent, longtemps apres, une rechute qui renoua la serie interrompue de ces etranges reveries.
613 14 j'ignore meme si le sentiment qui en resulte n'est pas conforme a l'idee chretienne;
637 14 je n'osai pas dire aux gardiens le nom d'une morte sur laquelle je n'avais religieusement aucun droit;

******** 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;
786 20 portant sur l'epaule gauche un enfant vetu d'une robe couleur d'hyacinthe.
161 19 je portais les yeux sur une toile qui representais une femme en costume ancien a l'allemande, penchee sur le bord du fleuve,
309 19 et parmi des touffes d'herbes parasites s'epanouissaient quelques fleurs de jardin revenues a l'etat sauvage.
312 19 j'apercus devant moi un entassement de rochers couverts de lierre d'ou jaillissait une source d'eau vive,
360 19 la pale lumiere des astres eclairait seule les perspectives bleuatres de cet etrange horizon;
900 19 etaient tracees des figures dont l'une representait la forme de la lune avec des yeux et une bouche traces geometriquement; sur cette figure on avait peint une sorte-de masque;
1021 19 un paysage eclaire par la lune m'apparaissait au-travers des treillages de la porte, et il me semblait reconnaitre la figure des troncs d'arbres et des rochers.
215 16 au sol des figures qu'il trace, au soleil des couleurs qu'il produit.

Downloaded from bms.sagepub.com at PORTLAND STATE UNIV on September 15, 2014

317 16 tandis-que sa figure et ses bras imprimaient leurs contours aux nuages pourpres du ciel.
1116 16 la fleur soufree, la fleur eclatante du soleil!
229 15 ca et la, des terrasses revetues de treillages, des jardinets menages sur quelques espaces aplatis, des toits, des pavillons legerement construits,
316 15 de telle sorte-que peu-a-peu le jardin prenait sa forme, et les parterres et les arbres devenaient les rosaces et les festons de ses vetements;
484 15 s'eclaircissait peu-a-peu par l'epanouissement du feu central, dont la blancheur se fondait avec les teintes cerises qui coloraient les flancs de l'orbe interieur.
498 15 j'entrai dans un atelier ou je vis des ouvriers qui modelaient en glaise un animal enorme de la forme d'un lama,
654 15 il disparut, baignant de feux rougeatres la cime des bois qui bordaient de hautes collines.
926 15 je me promenai le soir plein de serenite aux rayons de la lune, et, en levant les yeux vers les arbres,

2.6. Comment interpréter les classes?

La notion de "champ contextuel"

Dans la mesure où le vocabulaire spécifique d'une classe caractérise un *type de contexte*, nous proposons de l'appeler *"champ contextuel"*. Cette notion, bien que différente de la notion de *champ lexical*, recouvrant *"tous les mots associés à un même secteur de réalité"* (Pottier 1973), a en commun avec cette notion de caractériser un *espace sémantique particulier* à l'aide d'une classe de mots. La notion de *"champ lexical"* renvoie en définitive à une représentation générale susceptible d'obtenir l'adhésion de tout lecteur. Le *champ contextuel*, au contraire, ne dépend que de la manière spécifique dont un corpus particulier a été constitué, de ses lois de production. Sa définition dépend des lois de distribution du vocabulaire, dans un corpus donné et ne dépend pas du sens. Il s'agit donc d'une notion purement linguistique. Cela dit, il existe des liens entre champs contextuels et champs lexicaux que nous allons essayer de présenter.

L'analyse d'un "champ contextuel"

Nous nous sommes servis de la distinction faite par F. de Saussure (1972), entre les *rapports syntagmatiques* et les *rapports associatifs* (nous préférons ce terme à "paradigmatique", dont le sens est plus formel). Dans la chaîne du discours, les éléments en rapport syntagmatique se coordonnent dans un même énoncé, alors que les éléments en rapport associatif sont susceptibles de se substituer les uns aux autres. Cette distinction n'est pas spécifique de la langue et peut être utilisée pour appréhender n'importe quelle représentation. F. de Saussure utilise d'ailleurs l'image suggestive suivante pour faire comprendre son propos: *"A ce double point de vue, une unité linguistique est comparable à une partie déterminée d'un édifice, une colonne par exemple; celle-ci se trouve, d'une part, dans un certain rapport avec l'architrave qu'elle supporte; cet*

Downloaded from bms.sagepub.com at PORTLAND STATE UNIV on September 15, 2014

agencement de deux unités également présentes dans l'espace fait penser au rapport syntagmatique; d'autre part, si cette colonne est d'ordre dorique, elle évoque la comparaison mentale avec les autres ordres (ionique, corinthien, etc.), qui sont des éléments non présents dans l'espace: le rapport est associatif."

Etant donné que nous cherchons à analyser, non pas un énoncé, mais une classe d'énoncés, redondante d'un certain point de vue, nous pouvons distinguer dans le vocabulaire spécifique de cette classe, les mots plutôt en rapport associatifs (ils jouent un rôle analogue dans des énoncés différents), des mots plutôt en rapport syntagmatique (ils jouent des rôles complémentaires dans un même énoncé). La procédure utilisée pour l'approche d'un champ contextuel consiste, dans un premier temps, à regrouper le vocabulaire dans des classes associatives. Cette procédure rappelle celle utilisée par J. Dubois (Mounin 1975, p. 69), dans son étude sur *le vocabulaire politique et social en France de 1869 à 1872*, quand il réunit, par exemple dans une même classe, les termes *"ouvriers, travailleurs, salariés, pauvres, déshérités"*.

Si cette procédure s'apparente nettement avec la manière dont on construit un *champ lexical*, elle en diffère par le fait essentiel que ces associations n'ont pas de sens "absolu" mais relativement à un *champ contextuel* particulier (dépendant d'une classe d'énoncés particulière à l'intérieur d'un corpus précis) dont le vocabulaire est fixé préalablement à l'aide d'une analyse statistique. Dans un second temps, on peut chercher à intégrer les différentes classes associatives d'un même champ, dans une représentation plus globale.

L'expérience montre que cette représentation est à mettre davantage en rapport avec l'espace sémantique référentiel du locuteur, que nous appellerons un *"type de monde"*, qu'avec le "contenu" proprement dit des énoncés de ce locuteur. Elle ne permet pas d'interpréter ce qui est dit, mais de savoir dans quel cadre cela est dit.

3. APPICATION: *AURELIA* DE GERARD DE NERVAL

Aurélia est généralement présentée comme le testament spirituel de Gérard de Nerval. Il s'agit d'une sorte d'autobiographie rêvée avec, pèle-mêle, ses expériences, ses visions et ses quêtes spirituelles. L'auteur y décrit deux phases de sa vie intérieure, de sa "*maladie*" *"passée toute entière dans les mystères de son esprit"*. La première partie est surtout consacrée à l'élucidation de l'expérience de sa première crise qui aboutit à son internement en 1841, et où il fut très impressionné par la force des images oniriques qui s'imposèrent à lui, et de laquelle il sortit avec la certitude d'une existence dans la vie éternelle. Dans la seconde partie le récit reprend, dix ans plus tard. Interrogations sur la religion, perte de Dieu, culpabilité d'une faute jamais exprimée, et nostalgie de ce monde de rêves, à peine entrevu, où il a cru un instant retrouver les êtres chers maintenant disparus, vagabondages incessants dans Paris et ses alentours, pour diminuer la tension qui le mine et supporter les nuits d'insomnie, telle semble être l'atmosphère du début de cette seconde partie. Lors d'un internement, il reprend espoir en s'occupant d'un malade anorexique et

Downloaded from bms.sagepub.com at PORTLAND STATE UNIV on September 15, 2014

muet auquel il réapprend la parole et le goût de vivre (il lui chante durant des heures des comptines enfantines). Sa guérison fut de courte durée puisque, quelque temps après, dans la nuit du 25 janvier du froid hiver 1855, Gérard devait se pendre impasse de la Vieille-Lanterne alors qu'il errait dans Paris à peine nourri et à peine couvert (voir Sébillote 1948).

Chacune des deux parties de *Aurélia* est segmentée en une dizaine de "chapitres". La deuxième partie n'a pas été revue par l'auteur et la composition de la fin de la seconde partie est moins claire. Il est probable que l'oeuvre soit en partie inachevée. Nous avons distingué dans chaque chapitre les différents paragraphes qui concordent avec la composition de l'édition de la Pléïade. Les paragraphes constituent ici ce que nous avons appelé les *unités de contexte initiales* (u.c.i.).

Voici un mnémonyque des différents chapitres de *Aurélia*, qui servira, de plus, à introduire les principaux résultats de l'analyse précédente. Chaque chapitre peut être considéré comme un ensemble d'*u.c.e.* (grossièrement, les différentes phrases). Ces *u.c.e.* peuvent provenir éventuellement, de façon privilégiée, d'une des trois classes définies précédemment. Lorsque c'est le cas, une lettre est placée en regard du chapitre concerné qui identifie la classe associée correspondante (A pour la première classe, B pour la seconde, et C pour la troisième). Le chiffre associé à la lettre indique la force du lien. Il devient nettement significatif au dessus de 4.

C5		***Partie_1***
B9	1_01	l'expérience du rêve; Aurélia et son amie; voyages.
A2	1_02	retour a paris; rêve_1: l'ange de la mélancolie; l'heure fatale.
A2	1_03	délire; l'étoile; le double; l'hospitalisation.
C3	1_04	*MS: rêve_2: la maison des aieux.*
C6	1_05	*MS: rêve_2: le peuple élu.*
C9	1_06	*MS: rêve_3: la dame-jardin.*
C2	1_07	*MS: jardin aux sycomores; rêve_4: germes de la création.*
C4	1_08	*MS: rêve_4: le combat des esprits.*
A5	1_09	rêve_5: le double menacant.
	1_10	rêve_5: mariage d'Aurélia et du double.

A3		***Partie_2***
B9	2_01	monde réel et monde des esprits; pensées religieuses.
A3	2_02	désespoir; souvenirs; rêve_6: Aurélia entrevue.
A2	2_03	remords et reflexions; rêve_7: reproche de la vieille nourrice.
A4	2_04	dépression et vagabondage; hospitalisation.
A9	2_05	vagabondage et délire; hospitalisation.
B2	2_06	MS: la marche des astres; magnétisme et monde mystique.
	2_07	MS: le hameau; la chambre; rêve_8: le pardon de la déesse.
C9	2_08	*MS: mémorables (rêves et chants d'extase).*
B2	2_09	MS: guérison; le sens des rêves; conclusion.

(MS signifie "Maison de Santé", ce qui permet d'identifier les périodes d'internement.)

La clé A identifie plutôt les chapitres où Gérard de Nerval évoque ses vagabondages dans Paris, plus particulièrement à l'approche des crises, en proie à la surexcitation ou à la dépression. La clé B est plutôt associée aux

Downloaded from bms.sagepub.com at PORTLAND STATE UNIV on September 15, 2014

chapitres où il disserte sur la religion, le monde, le sens des rêves avec, notamment, les chapitres introducteurs de chacune des deux parties. Enfin, la clé C distingue l'évocation des "grands" rêves, ceux où l'auteur a cru un moment trouver la révélation d'un au-delà paradisiaque et l'assurance d'une vie éternelle, notamment ceux de la première partie (chapitres 4, 5, 6, 7, 8). Cette dernière clé sépare approximativent les périodes d'internement. Rappelons qu'il s'agit de la structure la plus caractéristique du corpus. Approfondissons l'interprétation de chacune de ces classes séparément à partir des résultats présentés en 2.5.

3.1 Analyse de la classe 1 (clé A)

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, famille, 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+, frapp+, terrible+.***
- une forte présence des indicateurs de la première personne (d'autant plus intéressante que ces indicateurs ont été considérées comme des *formes illustratives*): ***je, me, mes, mon.***

Il est d'ailleurs intéressant de noter qu'au cours des différentes analyses effectuées, il se trouve que de nombreux "mots outils" sont assez bien discriminés par les classes obtenues à partir des "mots pleins" (notamment, les pronoms personnels). 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 langagière, il existe néanmoins une présyntaxe, déjà opérante dans les structures plus profondes de la langue.

Le vocabulaire spécifique, et plus encore les phrases sélectionnées associées (voir 2.5), évoquent les vagabondages de l'auteur, dans Paris, la nuit, emporté par une force extatique lui donnant un sentiment de toute puissance, d'invulnérabilité, souvent avant le déclenchement de crises qui l'amenaient, quelques temps plus tard, à l'internement. Il s'agit généralement d'épisodes de son histoire, aussi nous placerons ce contexte sous le signe du monde réel (qui constitue en quelque sorte le décor, l'arrière fond du drame).

Downloaded from bms.sagepub.com at PORTLAND STATE UNIV on September 15, 2014

3.2. Analyse de la classe 2 (clé B)

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+, memoire, mystique+, race+, raison, religi+, science-, sept, serie+, simple+, verit+, vie.***

A 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 ce contexte: ***craindre-, devoir-, falloir-, aim+, bien, bon, epreuve+, fatal+, juge-, juste+, malheur-, supreme+.***

Comme en contrepoint des valeurs affectives positives ou négatives, les personnages ne paraissent exister que par leur rapport "fusionnel" au narrateur: fusion empathique - ***fils, frere+, mere+***; renforcée aussi par l'utilisation plus fréquente de la première personne du pluriel - ***nos, notre, nous;*** ou au contraire, sous l'identité d'étrangers menaçants - ***gard+, race+***; plus facilement perceptible, peut-être, au travers d'autre termes comme - ***mourir-, souffrir, arm+, envahi+ , lutt+;*** comme si le désir de fusion empathique et la peur de l'envahissement se nourrissaient mutuellement.

Dans la thématique de Gérard de Nerval, le combat des races et la lutte avec le double sont souvent évoqués en contrepoint de son désir de retrouver ses ancêtres morts ou l'amour d'Aurélia.

Notons la présence de nombreuses négations: ***jamais, ne, ni, pas.*** Elles ont doublement l'occasion de se révéler ici en tant que liées à l'expression de valeurs de vérité ou de mensonge, ou en tant qu'expression d'un sentiment de culpabilité, d'interdit, d'impuissance ou de doute. La relation entre ces deux aspects serait sans doute intéressante à approfondir sur le plan philosophique. Quels liens existent-il entre la notion de vérité et le sentiment de culpabilité? L'une, expression d'une loi, ne serait-elle pas l'émanation rationnelle et sublimée de l'autre, expression d'un devoir?

Quoiqu'il en soit, la tonalité sémantique de cette classe, saisie au travers des différents résultats paraît claire. C'est le monde mystique et rationnel de Gérard de Nerval qui est ici circonscrit, univers où le narrateur semble chercher une cohérence rationnelle à l'incohérence de ses désirs et de ses peurs, univers que l'on placera sous le signe du monde symbolique.

3.3 Analyse de la classe 3 (clé C)

Evocation de couleurs et de sensations lumineuses: ***blanc+, bleu+, brill+, color+, couleur+, eclair+, eclat+, lumiere+, lumineus+, pal+, rayon+, rose, roug+, vert+.***

Evocation d'éléments de la nature, d'un monde aérien: ***aile+, arbre+, bois+, ciel+, colline+, eau, feu, feuill+, fleur+, fleuve, herb+, horizon+, jardin+, lune+, matin+,***

Downloaded from bms.sagepub.com at PORTLAND STATE UNIV on September 15, 2014

mont+, montagne+, nuages, paterre+, plante+, rocher+, rustique, sauvage+, serpent+, soleil, source, terrasse, terre+, treill+, vaste, verdure.

Mots exprimant l'indistinct, l'émergence de formes, la création: ***combinaison, creation+, distingu+, ebauch+, elanc+, epanoui, figur+, form+, germe, model+, nouvelle, represent+, ressembl+, trac+, trait+, troble+, vari+.***

Mots évoquant des êtres sans nom propre, plus éthérés que réels: ***cheveu+, enfant+, femme+, fill+, figur+, front, habit+, peupl+, pied+, revet+, robe, vetement, yeux*** .

Ce contexte évoque irrésistiblement un monde paradisiaque, sans tension. Des termes comme (***jouer. lier. plaire., attir+, epanou+, harmoni+, immense+, immortel+, infini+, sein, vaste+***) renforcent ce sentiment. De plus, l'utilisation préférentielle de la troisème personne (***leurs, se***) éloigne de nous ce monde idyllique que l'on peut voir, mais avec lequel on ne peut communiquer. L'existence de quelques termes plus négatifs ne doit pas cependant être oubliée (***descend+, menac+, monstre, profond+, serpent+, soldat+, souterrain+***), autant de termes fortement chargés de valeurs symboliques, comme si l'on ne pouvait évoquer les images fascinantes des sources de vie sans évoquer les profondeurs d'une nuit primordiale.

Un des charmes de la prose de de Nerval n'est-il pas dans l'art de traduire, par le langage, ce miroitement des formes incertaines, naissantes, souvent associées aux images d'une végétation qui s'épanouit? Les phrases extraites, représentatives de ce *champ contextuel*, en donnent, selon nous, un aperçu significatif (voir 2.5).

3.4 Commentaires interprétatifs

En conclusion, au vu des résultats, trois types de "monde" semblent se dessiner dans cette oeuvre:

- 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, prend le nom d'Aurélia.

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

- enfin le monde symbolique, à la fois mystique et rationnel, celui à qui de Nerval confie ses doutes et ses interrogations sur la vie et la religion, sur le sens des rêves, de ses rêves et de cet espoir fou, une nuit pressenti, qu'il pourra retrouver, à la fin des épreuves de cette vie, tous les êtres chers qui l'ont définitivement quitté.

Downloaded from bms.sagepub.com at PORTLAND STATE UNIV on September 15, 2014

4. CONCLUSION

Notre objectif était de présenter une méthodologie mettant en valeur des caractéristiques formelles des textes, caractéristiques qui peuvent, cependant, induire chez un lecteur intéressé de nouvelles pistes de réflexion sur les mondes sous-jacents structurant une oeuvre. Ce n'est qu'en tant que lecteur que nous nous sommes permis de présenter nos "interprétations" de ces mondes dans le cadre d'une étude particulière.

Il est vrai que lorsqu'on aborde le "contenu" d'un corpus, on ne peut espérer aboutir à autre chose qu'au résultat 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 toute une série d'indices épars dans le texte, due aussi au renforcement statistique de ces indices au cours de la lecture.

Cependant les traits formels sur lesquels reposent les interprétations sont indéniables et peuvent être observés par tout lecteur. En cela, 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.

Le concept de sens est sans doute le concept le plus difficile à définir (il y a comme une *Boucle Etrange*, Hofstadter 1985) entre référé et signifié). Que peut-on donc en dire qui ne soit pas lié à une intuition introspective première, intuition induisant l'utilisation de nombreuses métaphores? Il serait sans doute intéressant de les étudier. Nous en distinguons deux types. D'abord, il y a celles basées sur un modèle spatial. On parlera, par exemple, d'*espace sémantique*, de *proximité sémantique*, etc. Ensuite, il y a celles basées sur un modèle logique ou propositionnel. On parlera, par exemple, de *syntaxe* du sens. Nous voulons suggérer par là, l'existence implicite derrière les formulations conceptuelles de modèles (mathématiques) particuliers. Le modèle spatial semble plus opérant lorsqu'il s'agit d'analyser la *structure profonde* de la langue. C'est en tout cas ce que semble montrer notre étude, puisque c'est le modèle implicite choisi ici (les outils mathématiques utilisés le présupposent). Le second, par contre, serait plus opérant pour étudier la structure de surface et le processus même de l'énonciation (aspect temporel). Nous rejoignons implicitement R. Thom, dans ce passage déjà cité, lorsqu'il admet que *"l'hypothétique structure profonde des linguistes est constituée essentiellement de notre représentation sensorielle du monde extérieur. Au contraire, la structure de surface sera constituée des automatismes du langage proprement dits."*

Avant de terminer, nous aimerions rendre hommage à Bachelard, dont la démarche controversée, exposée dans ses ouvrages sur la poétique des éléments (Bachelard 1942), est, en partie, à l'origine de cette étude. N'a-t-il pas cherché, à travers une collection d'extraits cueillis dans les oeuvres de différents poètes, à appréhender la *substance* active dans l'élaboration des images? Et, après tout, une matière n'est-elle pas, par essence, la définition

Downloaded from bms.sagepub.com at PORTLAND STATE UNIV on September 15, 2014

même d'un monde obscur (on croit chercher un contenu et l'on trouve un cadre, on croit chercher un lieu, et l'on trouve un horizon, on croit chercher une matière, et l'on trouve un espace)? Certes, l'objectif de l'auteur était très différent du nôtre puisqu'il s'agissait d'un essai poétique guidé par une rêverie sur la fascination exercée par les matières élémentaires - l'eau, la terre, l'air et le feu - plus ou moins influencée par les profondeurs de la psychologie junguienne. Malgré cela, il y a dans la démarche de Bachelard, un souci de rigueur, dans la manière de découper les textes, d'associer les fragments extraits pour appuyer son interprétation, que nous avons cherché à systématiser. Certes, à la différence de cet auteur, nous ne faisons aucune hypothèse *a priori* sur la nature particulière des lois organisant ces fragments, sinon qu'elles peuvent être révélées au travers une analyse statistique. Cependant, nous soupçonnons leur existence.

En définitive, dans les deux cas, il s'agit de rassembler des extraits de texte autour de lois et, dans les deux cas, cela amène le lecteur à s'interroger sur de nouvelles conspirations du sens en liaison avec les associations ainsi révélées, sans toutefois l'obliger à l'univocité d'une interprétation.

Cela dit, cette méthodologie a été appliquée à des corpus de textes divers, comme des ensembles d'entretiens, de réponses à une question ouverte, ou même à des ensembles d'articles, de rapports, etc. Dans ce cas, les "mondes sous-jacents" n'ont rien de mystérieux, mais permettent surtout un premier classement plus rationnel des documents, avant toute analyse de contenu.

Les techniques utilisées devraient pouvoir être couplées avec des techniques d'analyse du discours, ne serait-ce que pour décrire certaines caractéristiques syntaxiques ou plus simplement séquentielles (lois d'ordre entre les mots) des classes d'*énoncés* extraites. C'est dans cette voie que nous poursuivons actuellement cette recherche.

Post-scriptum

Au niveau de la terminologie, le petit tableau de correspondances suivant nous a servi à préciser les concepts linguistiques utilisés:

monde	linguistique	cognitif	"réel"
élément	signifiant forme	signifié concept	référé objet
cadre	contexte linguistique	représentation image	référent subsctrat

Downloaded from bms.sagepub.com at PORTLAND STATE UNIV on September 15, 2014

BIBLIOGRAPHIE

Bachelard, G., *L'eau et les rêves, essai sur l'imagination de la matière* (Librairie José Corte, Paris, 1942)

Benzécri, J.P., *L'Analyse des Données* (DUNOD, Paris, 1973)

Benzécri, J.P., *Pratique de l'Analyse des Données: linguistique et lexicologie* (Dunod, Paris, 1981)

Bertier, P., Bouroche, J.M., *Analyse des données multidimensionnelles* (PUF, Paris, 1975)

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

DE Nerval, *Oeuvres tome 1* (Bibliothèque de la Pléiade, 1974)

De Saussure, F., *Cours de linguistique générale* (Payot, Paris, 1972)

Denis, M., "Images et représentations sémantiques", *Bulletin de Psychologie*, XXXV, n° 356, 1982, 545-552

Diday, E., Lemaire, J., Pouget, J., Testu, F., *Eléments d'analyse de données* (Dunod, Paris, 1982)

Grive, J.B., Vegès, P., Silem, A., *Salariés face aux nouvelles technologies: vers une approche socio-logique des représentations sociales* (CNRS, Paris, 1987)

Harris, Z.S., *Structures mathématiques du langage* (Dunod, Paris, 1971)

Heidegger, M., *Qu'est-ce qu'une chose?*, traduction française par J. Reboul et J. Tamiviaux (Gallimard, Paris, 1971)

Hill, M.O., *TWISPAN, A FORTRAN Programm for Arranging Multivariate Data in an Ordered Two-Way Table by Classification of the Individuals and Attributes* (Cornell University, Ithaca, New York, 1979)

Hofstadter? D., *Gödel Escher Bach* (InterEditions, Paris, 1985)

Kosslyn, S M., "Les images mentales", *La Recherche*, n° 108, 1980, 156-163

Le Ny, J.F., *Science cognitive et compréhension du langage* (PUF, Paris, 1989)

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

Lebart, L., Salem, A., *Analyse statistique des données textuelles* (Dunod, Paris, 1988)

Downloaded from bms.sagepub.com at PORTLAND STATE UNIV on September 15, 2014

Michelet, B., *L'analyse des associations*, Thèse de nouveau doctorat en traitement de l'information (Université de Paris VII, 1988)

Minski, M., *La société de l'esprit* (InterEditions, Paris, 1988)

Mounin, G., *Clefs pour la sémantique* (Seghers, Paris, 1975)

Pérez, J.Cl., *L'intelligence artificielle* (Masson, Paris, 1988)

Pottier B. et al., *Le langage* (CEPL, Paris, 1973)

Rastier, F., *Sémantique interprétative* (PUF, Paris, 1987)

Reinert, M., "Classification descendante hiérarchique: un algorithme pour le traitement des tableaux logiques de grandes dimensions", in *Data Analysis and Informatics* (North Holland, Amsterdam, 1986), pp. 23-28.

Reinert, M., "Un logiciel d'analyse des données textuelles: ALCESTE", Communication aux Cinquièmes Journées Internationnales "Analyse de données et informatique", INRIA (1987)

Reinert, M., "Un logiciel d'analyse lexicale: ALCESTE", *Cahiers de l'Analyse des Données*, 4, 1986, pp. 471-484

Reinert, M., "Une méthode de classification descendante hiérarchique", *Cahiers de l'Analyse des Données*, 3, 1983, pp. 187-198

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

Salem, A., *Pratique des segments répétés* (Publication de l'INALF, collection "St-Cloud", Klincksieck, Paris, 1987)

Sébillote, L.H., *Le secret de G. de Nerval* (Librairie José Corti, Paris, 1948)

Thom, R., *Modèles mathématiques de la morphogénèse* (Presses de la Cité, collection 10-18, 1974)

Virbel, J., *Encyclopaedia Universalis*, tome 9, 1980, p. 1054

==========