Quelques problèmes méthodologiques posés par l'analyse de tableaux "Énonces x Vocabulaire"

Max Reinert

Université de Toulouse-Le Mirail

ABSTRACT:

Two types of units are counted in the statistical approach to texts presented here (computer method Alceste) : the "words" and the segments of texts known as "units of context". The author focusses his attention on the definition of the latter, on its links to the linguistic concept of utterance and on its implementation. A set of systematical tests are presented in order to evaluate the influence of the units' length (number of words) on the statistical outcome of the textual analysis. These tests were achieved by running Alceste on Gérard de Nerval's Aurélia - a text very familiar to the author, who analysed its structure several times before. The main result of the research is the display of strong stability in the textual structure identified by Alceste (where variations in units' lengths range from on to forty).

Dans l'approche statistique des textes utilisée (méthode "Alceste"), deux sortes d'unités sont comptabilisées : les "mots" et des segments de texte appelés "unités de contexte". Nous nous intéresserons plus particulièrement ici à la notion d'unité de contexte, à sa relation avec la notion d'énoncé, à son opérationalisation. Un ensemble d'essais systématiques permettra d'apprécier l'influence d'une variation de longueur de ces unités sur les résultats observés dans l'analyse du texte "Aurélia" de G. de Nerval.

Deux parties principales pour ce travail :

a) Une introduction plus théorique où sont évoqués les hypothèses de travail et les objectifs afin de préciser ce que nous attendons du découpage d'un texte en unités de contexte.

b) Une expérience à partir d'un corpus dont nous connaissons bien la structure pour l'avoir plusieurs fois analysée. Cette expérience est composée de deux fois deux séries de huit essais ainsi que leur comparaison : une des deux séries a été effectuée en prenant comme vocabulaire, les formes lemmatisées avec mise à l'écart des mots-outils et l'autre, les formes non lemmatisées sans reconnaissance des mots-outils.

1. Hypothèses et objectifs

La démarche adoptée pour l'analyse de corpus de textes est basée sur une hypothèse générale assez commune de topologie du sens. Brièvement, le point de vue que nous soutenons s'appuie sur les arguments suivants. Pour pouvoir énoncer, le sujet énonçant doit se représenter ce qu'il va dire dans un certain espace mental (qui lui sert de "référence"). Le choix de cet espace référentiel, de ce "lieu" - qui ne dépend pas forcément d'une opération consciente - implique le choix d'un type d'objet : il implique, par là même, un type de vocabulaire. *En conséquence, l'étude statistique de la distribution du vocabulaire dans les différents énoncés d'un corpus doit permettre une discrimination de ce vocabulaire révélatrice des différents choix référentiels effectués par l'énonciateur.*

D'un point de vue opérationnel, on étudie cette discrimination à partir d'un tableau de données binaires croisant, en lignes, les différents "énoncés" du corpus et, en colonnes, les différents "mots pleins" (noms, adjectifs, verbes, adverbes dans leur forme lemmatisée) avec à l'intersection d'une ligne et d'une colonne, la valeur "zéro" si le mot considéré est absent de l'énoncé et la valeur "un" s'il est présent.

Cette perspective implique des compétences particulières pour effectuer : a) le découpage du corpus en "énoncés" ; b) la reconnaissance du vocabulaire.
Nous insisterons surtout ici sur le point (a) : comment discriminer un "énoncé" ? Un énoncé est-il identifiable formellement hors de l'objet même du discours ? Un énoncé dans l'oeuvre de Proust peut-il être défini de la même manière qu'un énoncé dans une interview ? Qu'entendons-nous d'ailleurs par "énoncé" ?

Si l'on se réfère au dictionnaire (LEXIS), un énoncé est défini principalement selon deux types de critères : *"1) soit comme proposition, phrase dans laquelle une pensée est énoncée : «l'oiseau chante» par exemple est un énoncé élémentaire. 2) soit comme texte exact qui exprime un jugement, qui formule un problème : se reporter à l'énoncé de la loi."*

Dans les deux cas, on insiste plutôt sur l'aspect propositionnel ou prédicatif : on dit quelque chose à propos de quelque chose. On juge. Mais cet aspect reste flou dans le langage naturel notamment du fait de l'enchâssement des propositions. Il est alors nécessaire de saisir le propos principal et de délaisser momentanément les propos secondaires. Cette opération suppose des choix subjectifs dépendants d'un acte de lecture, des attentes (ou manipulations) qui peuvent être aussi bien imposées par le contexte (contrat de communication par exemple) que par des dispositions individuelles. Ainsi les contours d'un énoncé sont généralement flous du fait que les objets dont on parle sont plus ou moins ambigus (non complètement définis, reconstruits en partie), ou en concurrence les uns avec les autres.
On comprend donc la prudence de nombreux linguistes qui ne veulent voir dans l'énoncé que la trace textuelle ou phonologique d'un acte d'énonciation. Pour B.

Potier, par exemple, *"l'énoncé est une production linguistique «brute», antérieure à l'observation"*. Cela dit les frontières d'un acte ou d'une production sont tout aussi arbitraires (les causes d'un arrêt de production pouvant être de nature très diverse) et le problème reste entier.
Cette dernière définition a cependant le mérite de ne pas séparer énoncé et énonciation qui ne peuvent se définir l'un sans l'autre. Autant la notion d'énoncé présuppose l'unicité d'un objet du discours (ce que l'on veut dire), autant celle d'énonciation présuppose l'unicité d'un sujet (celui qui veut dire quelque chose). Mais cette unicité apparente est plus intentionnelle que réelle, que l'on se place au niveau de l'énonciateur - qui parle en définitive ? le locuteur de chair et d'os, celui qui affirme, celui qui se distancie ou nie, l'institution qui lui propose un rôle, l'interlocuteur auquel le locuteur peut chercher à s'identifier et qui peut parler à travers lui, etc… - ou que l'on se place au niveau du thème, celui-ci pouvant se démultiplier en une infinité de facettes selon les contextes dont on l'enrichit ou selon que l'on mette plutôt en valeur tel ou tel aspect de la thématisation.

Cela dit, sous cette multiplicité miroitante des éléments énoncés, l'énonciataire aussi bien que l'énonciateur doit faire le pari d'une cohérence globale qu'il cherche justement à apprécier dans ce qu'il appelle le sens de cet énoncé, que celui-ci soit attribué à la proposition, à la phrase, au paragraphe ou à une oeuvre entière. Il n'y a donc pas obligatoirement contradiction entre unicité et multiplicité. On ne trouve pas de contradiction dans le fait qu'un même objet puisse être ici et là quand on évoque des lois de symétries, entre la multiplicité des éléments d'un groupe ou d'un tableau cubiste et l'unité de leur représentation, entre les objets d'un monde à partir du moment où des opérations de contiguïté, de déplacement ou de transformation peuvent être mises en oeuvre.
Rechercher la cohérence sous-jacente à un énoncé c'est de même dévoiler l'unité logique des opérations effectuées sous la diversité des éléments réunis. C'est en cela qu'un énoncé renvoie à un monde où les objets interagissent en fonction de lois propres à ce monde. Cette cohérence n'est cependant pas externe à celui qui l'évoque. Elle est reconstruite en vue de certaines finalités. Elle exprime un point de vue, impliquant aussi bien un sujet qu'un lieu référentiel que l'on ne peut étudier indépendamment l'un de l'autre.

Au niveau méthodologique, l'enchâssement des propos dans un même énoncé implique qu'il n'y a pas de frontières sûres entre deux énoncés. Il n'y a que différents niveaux de cohérence, la cohérence globale d'une oeuvre pouvant résulter de l'agencement d'une multitude de cohérences locales. Dans un acte de lecture, nous utilisons les mots d'un contexte restreint pour reconstruire une référence possible, c'est à dire pour identifier un point de vue probable, hypothèse de lecteur sans cesse remise en cause au cours de la lecture. C'est cette interaction entre énoncés et points de vue qui justement nous intéresse et notamment les relations existantes entre cohérence globale et cohérences locales.

Si l'analyse statistique ne permet pas d'expliciter dans le détail les opérations reliant les "objets" impliqués dans l'énoncé pour l'énonciateur, elle permet par contre d'expliciter les environnements probables dans lesquels des opérations particulières sont mises en oeuvre sur un grand ensemble d'énoncés et donc, d'effectuer une sorte de cartographie des points de vue de l'énonciateur à partir de leurs traces textuelles.

Au niveau opérationnel, le problème de la discrimination des différents énoncés d'une oeuvre reste entier. Notre approche jusqu'ici a été de découper le corpus en segments de texte assez courts, de grandeur comparable à la phrase ou au paragraphe, que nous appelons *"unités de contexte "*.

Nous présentons ici le résultat de quelques essais systématiques de variation de la longueur de ces unités de contexte sur les résultats de l'analyse du texte "Aurélia" de G. de Nerval, la "longueur" d'une unité de contexte étant appréciée grossièrement en nombre d'occurrences.

2. Expérimentation

2.1. *Première série d'essais*

Dans notre analyse de "Aurélia" de 1990, l'interprétation était basée sur une partition des unités de contexte en trois classes. Aussi pour les différents essais effectués ici, nous n'avons retenu que les trois premières classes de la classification descendante hiérarchique (C.D.H.).

Rappelons rapidement la procédure d'analyse : le corpus est découpé en segments de texte d'environ 200 caractères, si possible terminés par une ponctuation forte, sinon faible, et, dans le cas où il n'y a pas de ponctuation, par un séparateur de forme. Ces segments de texte sont appelés *unités de contexte élémentaire* ou *u.c.e.*. Ce calcul étant effectué, on cherche ensuite l'ensemble des formes utilisées. Cet ensemble est partagé en deux parties : les "mots-outils" (grossièrement, les articles, prépositions, conjonctions, pronoms, certains adverbes, les auxiliaires, certains verbes modaux) et les "mots pleins" (noms, verbes, adjectifs, adverbes). Les formes associées aux mots pleins sont réduites de leurs désinences grammaticales ou suffixes dans la mesure où cette réduction permet de regrouper sous une même racine plusieurs formes du corpus.

Une fois ces deux opérations effectuées, on considère le tableau de données croisant, en lignes, les u.c.e. et, en colonnes, les formes réduites. A partir de ce tableau, on construit de nouvelles unités de contexte en concaténant deux ou plusieurs u.c.e. successives jusqu'à ce que le nombre de formes analysées réunies dans cette nouvelle unité soit supérieur à un seuil fixé (la "longueur" de l'unité de contexte).

En outre, le corpus peut être prédécoupé en fonction d'unités de contexte naturelles que ce soit des réponses d'individus, les chapitres d'un livre ou ici les 19 paragraphes numérotés du texte : nous appelons "*unités de contexte initiales* "ou "*u.c.i.*" ces unités de base. L'unité de contexte considérée dans l'analyse est donc un segment de texte compris entre l'u.c.e. qui est la plus petite unité de contexte définissable sous "Alceste" et l'u.c.i. qui en est la plus grande.

Cela dit, quelle que soit l'unité de contexte utilisée, on ne considère pour l'analyse que la présence (1) ou l'absence (0) d'une forme dans cette unité.

2.1.1 *Analyse sur le vocabulaire lemmatisé.*

n° essai	longueur (nbre de formes)	nbre lignes par u.c.	nombre d'u.c.	nbre de formes	% zéros
1 (A)	u.c.e.	2,5	876	610	98,95
2 (B)	6	3,8	569	608	98,39
3 (C)	12	6,0	364	603	97,53
4 (D)	24	10,3	213	598	95,91
5 (E)	48	18,4	119	582	93,04
6 (F)	96	32,7	67	572	88,54
7 (G)	192	60,8	36	547	80,84
8 (H)	u.c.i.	115,3	19	508	68,65

Tableau 1. Caractéristiques des tableaux analysés .

Pour comparer les différents résultats obtenus, nous avons concaténé les tableaux résultats des huit essais, tableaux croisant vocabulaire par classes. Le tableau résultant ainsi construit contient autant de lignes que de mots retenus. Le nombre de colonnes est égal au nombre d'analyses multiplié par le nombre de classes demandées soit ici 8 X 3 = 24 colonnes.
Nous avons effectué sur ce tableau une classification ascendante hiérarchique avec la métrique du chi2 et avec pour indice l'inertie intraclasse. Avec cette procédure, deux classes de même profil sur le vocabulaire (donc s'interprétant de la même manière) seront agrégées en bas de l'arbre (voir figure 2).
La proximité des résultats des différentes analyses est très nette et montre que, du moins pour cette oeuvre, les mondes lexicaux interprétés sont plutôt dépendants de la structuration en paragraphes. On note cependant un léger effet séquentiel pour les noeuds 1 (A, B, C / D, E, F, G) et 2 (A, B / C, D, E / F, G, H), que nous ne chercherons pas à interpréter ici.

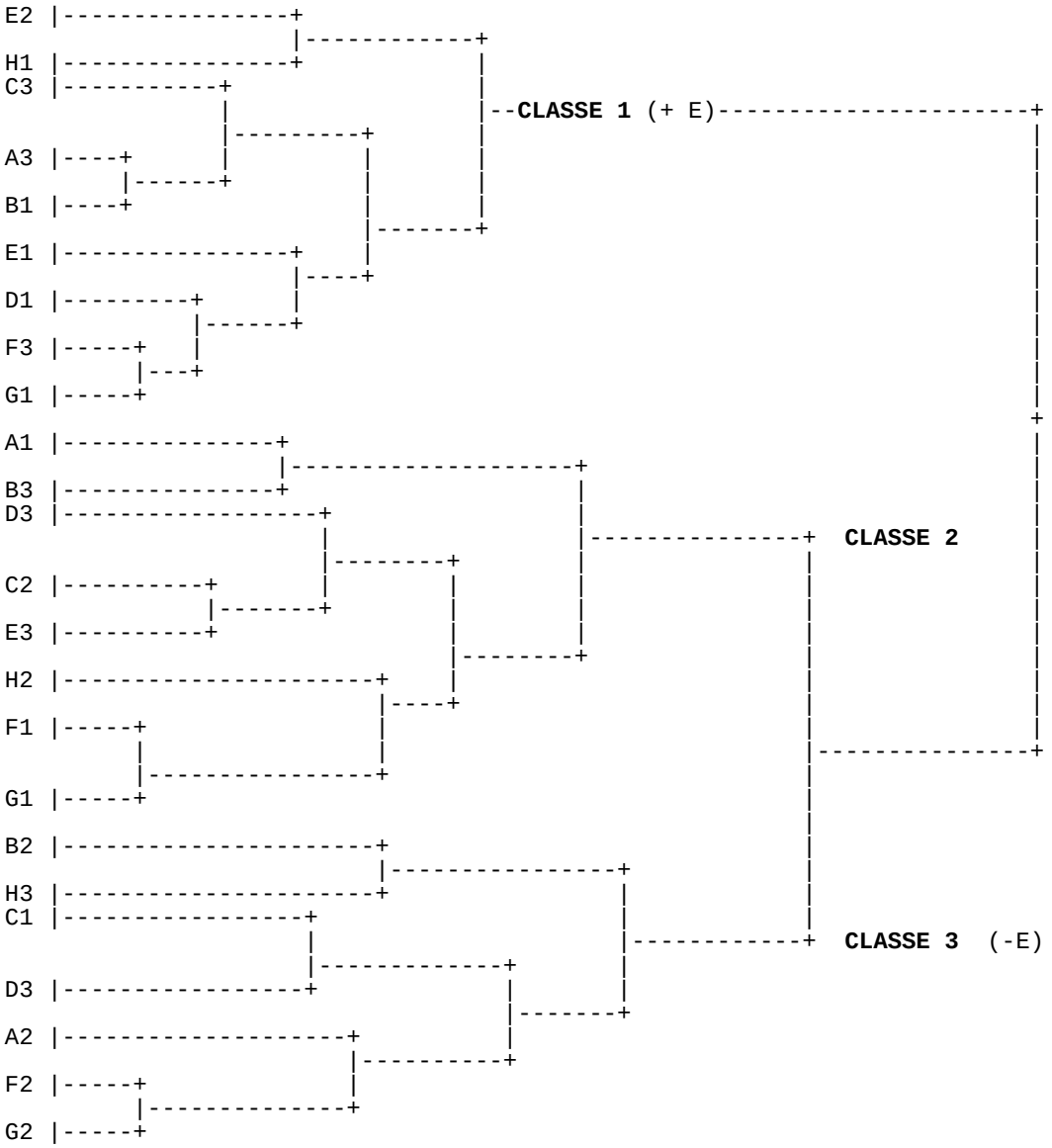

Figure 2. C.A.H. du tableau comprenant , en colonnes, 24 classes (3 classes par analyse et 8 analyses) et, en lignes, le vocabulaire lemmatisé.

Pour donner une représentation schématique de la structure du vocabulaire qui reste stable, on a reclassé les u.c.e. de la manière suivante : une u.c.e. est affectée à la classe n si elle appartient au moins à la moitié des classes regroupées (quatre sur les huit possibles) au noeud "classe n". Ce choix permet le reclassement de 98% des u.c.e. du corpus.

On recalcule ensuite le tableau croisant ces trois nouvelles classes avec le vocabulaire et on effectue l'A.F.C. de ce dernier. Seules les formes les plus spécifiques de chaque classe sont représentées (figure 3).

Figure 3. Premier plan factoriel de l'AFC du tableau croisant les trois classes avec le vocabulaire lemmatisé (représentation à partir des corrélations).

On retrouve très clairement les *mondes lexicaux* présentés dans l'étude de ce texte (1990). Le vocabulaire de la première classe pourrait être distribué dans différents champs lexicaux, en rapport avec le thème de la nature, les sensations (visuelles notamment), l'émergence des formes : *petit, profond, belle, grand, couleur, lumière, vaste, brille, peint, figure, longue, forme, trace, …*; *rayon, fleur, lune, soleil, montagne, air, arbre, ciel,….* Celui de la seconde, en fonction des déplacements, des sentiments, des relations, l'évocation d'êtres proches : *revenir, rester, conduire, aller, rencontrer, diriger, (se) rendre, attendre, passer, retrouver…*; *famille, ami, Aurelia…* ; *eprouver, mal, souvenir, larme, triste, malade, joie….* Et enfin, le vocabulaire spécifique de la troisième classe se distribuerait plutôt dans des catégories conceptuelles*: âme, religion, fatal(ité), mystique, esprit, vie, raison, science, destin, monde, chrétien, humain, dieu, éternel, harmonie, mystère, homme, etc….*

2.1.2. *L'approche en fonction du choix du vocabulaire*

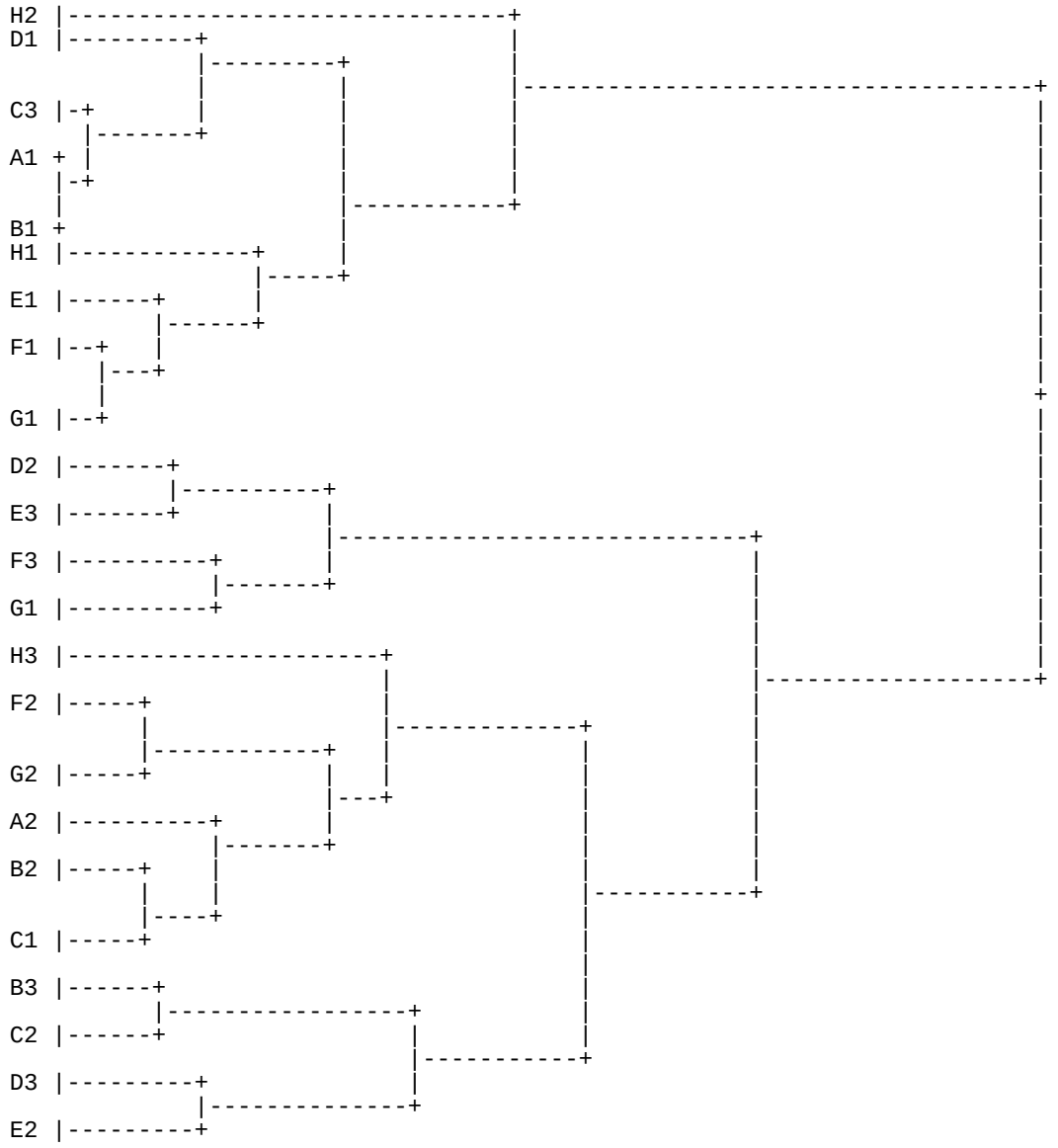

Figure 4. Classification des 24 classes obtenues lors des 8 analyses sans lemmatisation.

Pour étudier l'effet éventuel de la lemmatisation sur les résultats, nous avons procédé à la même série de huit analyses sur le corpus sans lemmatisation et sans reconnaissance des mots-outils (figure 4).

La première classe indique une bonne stabilité de la première dimension des différentes analyses ; elle regroupe en effet les huit classes des huit analyses. Il existe cependant une perturbation de la seconde dimension et la seconde branche de la seconde classe se segmente en trois parties distinctes, l'une regroupant six des huit classes et donc révélatrice d'un trait assez stable ; les deux autres, par contre, ne regroupent que quatre classes, chacune concernant des classes d'analyses contiguës (relativement à la longueur de l'u.c choisie : D, E, F, G d'une part et B, C, D, E, d'autre part).

Avant de s'interroger sur la signification de ces résultats, nous avons préféré effectuer une deuxième série de 8 essais en diminuant les contraintes de longueur : u.c.e., 10, 20, 30, 50, 70, 90, uci.

2.2. *Deuxième essai.*

Pour ce deuxième essai, nous avons appliqué le même protocole que précédemment : huit analyses avec lemmatisation et huit analyses sans. Les résultats de la série "avec lemmatisation" sont très proches de ceux déjà présentés. La classification des u.c.e. obtenue dans les deux séries est même très voisine puisqu'elle est commune à 84% (voir tableau 5).

Tableau croisant les deux partitions :

Essai 1 *	Essai 2		
classe *	1	2	3
poids *	321	405	108
1 324 *	304	12	8
2 394 *	6	367	21
3 116 *	11	26	79

750. u.c.e classees sur 892 soit 84.08 %

Tableau 5. Comparaison des deux séries d'essais "avec lemmatisation".

Quant à l'analyse effectuée à partir des formes non lemmatisées, la classification fait apparaître, cette fois, trois classes nettes indiquant une bonne stabilité des trois premières classes. Ces résultats sont donc plus faciles à comparer avec ceux obtenus dans l'essai 1. De même que pour celui-ci, on a réaffecté les u.c.e. appartenant au moins à 4 classes de chaque noeud conservé et effectué l'A.F.C. sur le tableau croisant le vocabulaire "non lemmatisé" avec ces trois nouvelles classes (figure 6).

```
Axe horizontal : 1e facteur : V.P. =.1115 (   61.34 % de l'inertie)
Axe vertical   : 2e facteur : V.P. =.0703 (   38.66 % de l'inertie)
```

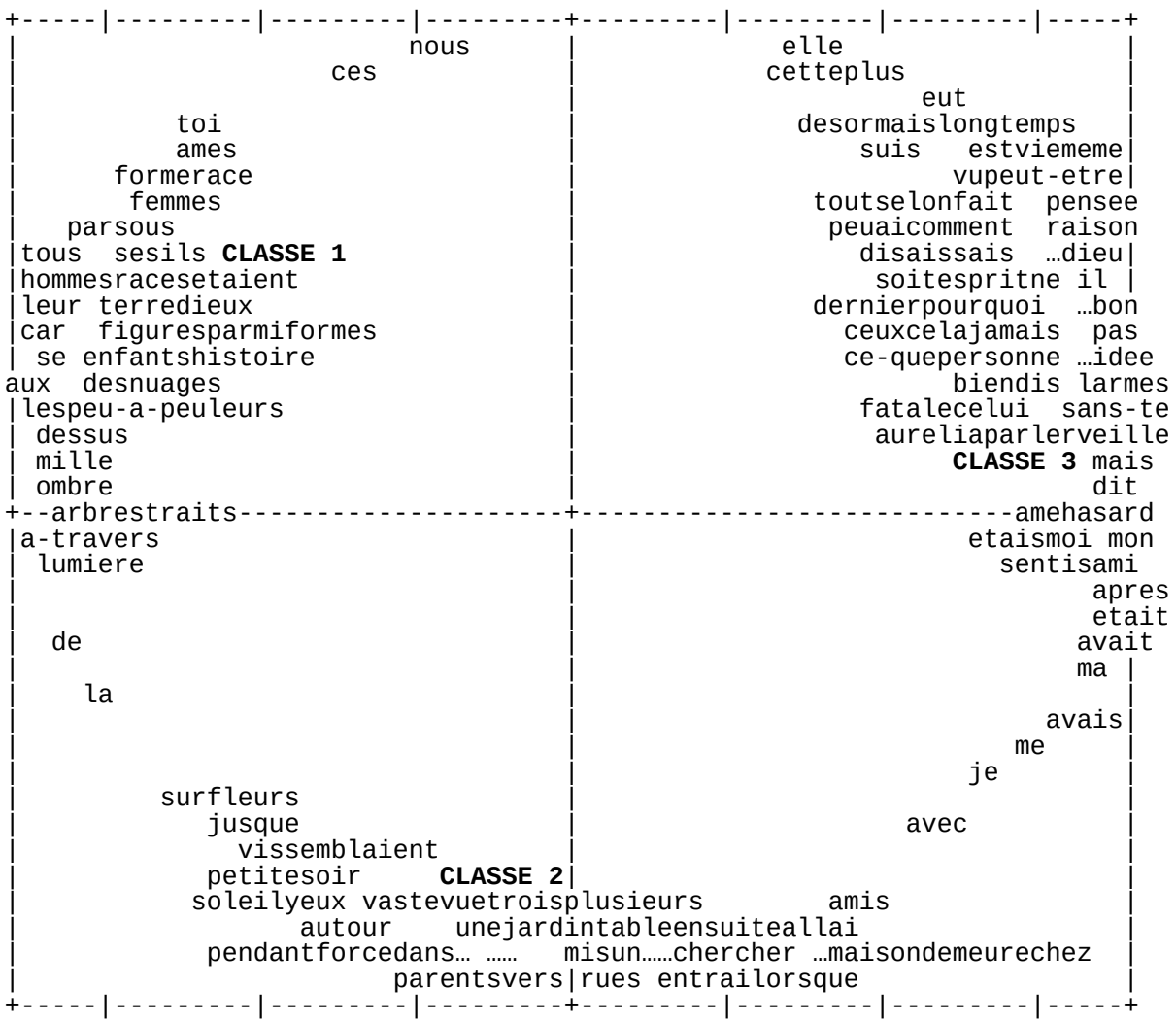

Figure 6. Premier plan factoriel de l'AFC du tableau des classes sur le vocabulaire non lemmatisé (représentation à partir des corrélations).

Les profils des classes obtenues sont sensiblement différents de ceux mis en évidence dans l'analyse précédente malgré quelques analogies. La classe 1 apparaît assez proche de la classe 1 de la précédente analyse. Par contre, les classes 2 et 3 sont peu reconnaissables. Cette inflexion est nette si l'on observe le classement des u.c.e. pour les formes réduites et non réduites de cet essai : seules 47,6 % des u.c.e. sont affectées aux "mêmes" classes (ce qui est toutefois révélateur d'un lien : à comparer avec les 33 % de l'hypothèse nulle).

Ainsi, le choix du vocabulaire semble influencé davantage les résultats que le choix de la longueur des unités de contexte. De plus, dans le cas d'une "non lemmatisation", les résultats obtenus sont plus sensibles à la variation de la longueur des u.c.. Un effet de codage n'est du reste pas à exclure, le poids des mots très fréquents se renforçant au fur et à mesure que l'unité de contexte diminue (du

fait du codage binaire utilisé). Nous laissons pour l'instant ces faits à l'appréciation du lecteur.

3. Conclusion provisoire

Il est, bien sûr, délicat de tirer des conclusions trop affirmatives sur des résultats obtenus à partir d'un seul corpus. Par contre, il est utile d'essayer de formuler un certain nombre d'observations afin justement de voir, dans d'autres cas, si elles se répètent ou non. Notamment :

a) *L'arbitrarité du découpage en unités de contexte semble avoir peu d'influence sur les résultats, notamment dans le cas où le vocabulaire retenu est lemmatisé ;*
b) *Dans ce protocole d'analyse, le rôle de la lemmatisation n'est pas négligeable et l'on ne trouve pas exactement les mêmes résultats avec ou sans lemmatisation.*

L'interprétation de cette variation est cependant délicate du fait de la taille réduite du corpus étudié.
En ce qui concerne l'approche méthodologique "Alceste", c'est surtout le point (a) qui nous intéresse et c'est sans doute le résultat le plus surprenant et le plus sûr de cette expérience. Cela dit, le découpage en unités de contexte n'a qu'une influence limitée sur les résultats obtenus dans la mesure où ce découpage est compatible avec les grandes unités naturelles du corpus considéré : chapitres ou paragraphes, qui constituent en quelques sortes, les énoncés de base.

Ce phénomène suggère plusieurs orientations de travail. Entre autres :
a) lorsque ce découpage n'est pas connu ou bien lorsqu'il est problématique, l'utilisation de plusieurs segmentations arbitraires du texte peut permettre d'apprécier ce que l'on obtiendrait si un tel découpage naturel existait. Il peut permettre dans ce cas d'en concevoir, après analyse, des frontières probables ; Cette procédure peut donc apporter une aide pour définir, a posteriori, ce que l'on peut appeler "énoncé" dans une oeuvre donné ;
b) Dans le cas où ce découpage semble connu a priori, par construction même de l'oeuvre, cette procédure peut permettre d'en tester la pertinence en mettant en évidence, par exemple, l'homogénéité des parties ou des chapitres analysés (voir notamment les études de A.M. Pezous ou de C. Roy dans ce volume).