LES CAHIERS DE L'ANALYSE DES DONNÉES

M. REINERT

Un logiciel d'analyse lexicale

Les cahiers de l'analyse des données, tome 11, n° 4 (1986), p. 471-481

<http://www.numdam.org/item?id=CAD_1986__11_4_471_0>

Les Cahiers de l'Analyse des Données
Vol. XI - 1986 - N° 4 - pp. 471-484

UN LOGICIEL D'ANALYSE LEXICALE : [ALCESTE]

par M. Reinert

Après avoir explicité nos objectifs (1), nous présenterons quelques unes des procédures utilisées lors des principales étapes d'une analyse lexicale (2), et un bref exemple d'application (3) extrait d'une étude de Mme MAFFRE-SAVELLI (**) sur l'oeuvre de Lawrence Durell (préparation d'une thèse de doctorat à l'Université de Toulouse-le-Mirail).

1 *Introduction*

L'utilisation de l'analyse des données et plus spécifiquement de l'analyse factorielle des correspondances pour la descriptions des structures textuelles date du début même de cette discipline (BENZECRI 1962), cependant, ce n'est que bien plus tard, que les logiciels d'analyses des données spécifiques au traitement des données textuelles apparaissent (LEBART, SALEM et LAFON).

Les études effectuées ont plutôt été orientées vers une approche purement formelle des structures textuelles : il s'agit principalement de comparer les distributions de "mots" voire de "séquences de lettres", entre différents textes, ou d'étudier leurs cooccurrences dans de mêmes "contextes". Cela n'empêche pas des différences de points de vue dépendantes du type de corpus traité, du type de contenu à révéler, point de vue qui, en définitive, imprime sa marque sur la manière dont on définit les "contextes" et les "unités textuelles" à dénombrer : quelles formes graphiques doivent être retenues ? Doit-on ou non en rejeter certaines de l'analyse ? En fonction de leur rôle syntaxique ? en fonction de leur "longueur" ? (nombre de lettres) ? en fonction de leur fréquence . Cherche-t'on à comparer leur distribution sur plusieurs textes ou, au contraire, à décrire leur organisation dans un corpus particulier ? Et pour l'étude des distributions, quel type de découpage retenir pour opérationnaliser la notion de "contexte" ?

Nous avons été très liés dès l'origine aux travaux du laboratoire de statistique de PARIS VI (notamment par l'intermédiaire de M. ZLOTOWICZ et de son étude sur les peurs enfantines 1971). Nous avons aussi été influencés par la méthodologie utilisée par le psychologue H. TOME dans ses travaux sur "l'Identité" (1978), utilisant une technique d'enquête particulières par questions ouvertes. C'est dire que notre approche a été imprégnée aussi au départ par la méthodologie classique de l'analyse du contenu thématique alors en vigueur [1].

(*) *Ingénieur. Laboratoire CNRS UA 259. Université de Toulouse-le-Mirail.*

(**) que nous remercions pour son aimable accord.

Cette méthodologie est, sur le plan théorique, généralement rejetée aujourd'hui, même si, dans de nombreuses études, elle est toujours utilisée, faute d'outils mieux appropriés. Elle ne pose d'ailleurs pas seulement des problèmes théoriques, mais aussi des difficultés pratiques, les règles de décodage étant souvent difficiles à expliciter, ce qui ne facilite pas les prises de décision et occasionne une grande perte de temps tout en ne permettant pas une objectivation complète des procédures d'analyse.

La méthodologie que nous proposons porte la marque de cette double expérience (approche formelle, catégorisation conceptuelle) et si les difficultés rencontrées nous ont éloignés de l'analyse de contenu, pour nous rapprocher d'un type d'analyse plus lexical, nous en avons cependant conservé certains schèmes méthodologiques comme par exemple, la notion "d'unité de contexte". Aussi nous l'avons dénommée "Analyse Lexicale par Contexte" (ou A.L.C.).

Nous entendons par "unité de contexte" (u.c.), tout segment de texte pouvant servir de support à l'étude des représentations envisagées. Généralement la segmentation du corpus en u.c. suit la segmentation "naturelle" en unités de sens : proposition, phrases, paragraphes, réponses etc. . Opérationnellement ces unités ne sont pas toujours aisées à différencier avec précision et nous avons préféré parfois un découpage "automatique" du texte (par exemple, une ou plusieurs lignes consécutives du corpus retranscrit).

Il s'avère que lorsque le nombre d'u.c. est grand (c'est le cas habituel, les redondances entre u.c. atténuent les disparités du découpage pour distinguer ce qui est stable de ce qui ne l'est pas dans les résultats obtenus.

Les unités de contextes que nous définissons sont le plus souvent de petite taille (entre 1 à 10 lignes), ceci afin que les mots associés puissent être considérés comme provenant d'un même"énoncé".

a) *aspects méthodologiques*

Le but très général de l'analyse est de trouver des classes d'u.c. où sont utilisés spécifiquement certains types de vocables.

On fait l'hypothèse que ces ensembles de vocables renvoient à des représentations sous-jacentes qu'il est alors possible d'expliciter, notamment du fait des redondances et ceci, indépendamment d'une étude de la syntaxe (pour plus de précision se reporter à [2]).

On a appelé "contexte" l'ensemble des vocables associés à une telle représentation et "contexte statistique", l'ensemble effectivement observable des vocables significativement présents dans une classe caractéristique d'u.c...classe qu'il s'agit justement de définir de manière à ce qu'un contexte statistique renvoie en définitive à un contexte "réel" (même si la place de chaque vocable individuellement dans une classe d'u.c. empirique est peu sûre du fait des nombreux aléas, le hasard a peu de place quand se regroupent ensemble des types de mots caractéristiques : par exemple, dans l'analyse des poésies de Rimbaud [3], on trouve dans l'une des classes 15 mots désignant une couleur : si chacun d'entre eux est présent avec peu de certitude dans le "contexte statistique" considéré, le fait de les retrouver ensemble uniquement dans cette classe est en lui-même, si peu probable, qu'il devient l'expression d'une loi quasi certaine que nous associons à un trait caractéristique de la représentation sous-jacente.

b) *aspects logiciels*

Le logiciel ALCESTE présenté permet d'effectuer l'A.L.C. d'un corpus considéré comme un "Ensemble de Segments de TExtes" d'où son nom...).

La nature des données textuelles traitées par ALCESTE est assez diverse : questions ouvertes d'enquêtes, entretiens semi-directifs, corpus de résumés d'articles, corpus de discours politiques, roman, corpus de poésies ... Ce logiciel peut servir aussi à l'analyse de données de type présence (1)-absence (0) : il suffit de traduire sous forme de texte, les différentes modalités présentes chez un individu de la population étudiée. Cette retranscription est d'ailleurs économique si les modalités sont rares.

Le logiciel ALCESTE est opérationnal, sous système MULTICS (il comprend plus de 5000 instructions en FORTRAN 77 standard). Une version est implantée sur le DPS8 du Centre Inter-universitaire de Calcul de Toulouse. Cependant il serait assez aisé de l'implanter sous d'autres systèmes, d'autant que cette version d'ALCESTE provient d'une version antérieure tournant sur les IBM 3081 du C.N.U.S.C. (centre de calcul de Montpelleir). Elle permet d'analyser des corpus allant jusqu'à environ 10000 lignes de 74 caractères.

2 *Les principales étapes d'une analyse*

2.1 Retranscription du corpus : trois types d'informations peuvent être codées :

- des identificateurs numériques permettant de repérer les grandes unités de contextes (plusieurs niveaux d'unité peuvent être définis, chacun étant associé à une partition plus ou moins fine du corpus).

- des informations "hors corpus" permettant de préciser des caractéristiques qualitatives de ces unités (s'il s'agit de "réponses" à une question ouverte, les niveaux d'âge, de sexe, de milieu peuvent être ainsi repérés).

- le texte proprement dit, transcrit sans accent (ce choix permet d'éliminer un grand nombre d'irrégularités grammaticales).

Exemple de retranscription (tiré de l'étude de l'oeuvre de L. Durrell) :

```
020013  partie_1 *Justine *page_10

      la mer est de nouveau trop grosse aujourd'hui, et des bouffees de
      vent tiede viennent désorienter les sens. au coeur meme de l'hiver, on
      perçoit deja les premices du printemps ....

      ....

020023  partie_1 *Justine *page_20

      quelque chose me frappa dans son attitude l'air gauche d'un phoque
      savant qui s'efforce de mimer les emotions humaines et je compris
      pour la premiere fois qu'il aimait probablement Melissa autant que moi.
      j'eus pitie de sa laideur et ...

      ....
```

2.2 Le calcul du dictionnaire et du corpus numérique (CDT) : Le programme CDT permet de discriminer des chaînes de caractères comme "mot brut", de les recenser dans un dictionnaire (à créer ou déjà existant), éventuellement d'en éliminer certaines comme "non pertinentes", ou de les réduire. Il construit de plus, une forme numérique du corpus permettant sa manipulation plus aisée par les autres programmes de traitement, comprenant toutes les informations nécessaires à la reconstruction des différents types d'unités de contexte.

Pour chaque mot différencié dans le corpus, il s'agit de savoir s'il existe déjà dans le dictionnaire, et, s'il n'y existe pas, de l'y ajouter. La solution adoptée s'avère suffisamment rapide pour nos besoins :

Elle consiste à définir 28 "portes" d'accès au dictionnaire, associées à la première lettre du mot M considéré (les 26 lettres de l'alphabet (en minuscule), le singe " " associé aux mots "hors corpus", tout autre signe associé aux mots du corpus non analysés, considérés comme "éléments illustratifs"). Il est facile d'associer à ce mot M, le numéro d'ordre m de sa porte et de ne prospecter que les mots passés par cette porte (ceci évite d'avoir à ordonner le dictionnaire par ordre alphabétique).

- Suppression des mots usuels : L'élimination des mots très usuels, dès la constitution du dictionnaire a essentiellement pour but de diminuer le volume du fichier numérique associé au corpus retranscrit, ainsi que les temps de lecture de ce fichier. Ces mots sont en effet, des mots de liaison ayant peu d'intérêt pour la recherche des contextes, telle que nous l'envisageons (voir 2) : articles, prépositions, conjonctions principalement (aux homonymies près).

Le programme de réduction (REDUC : voir 2.3) permet également la suppression de certains mots, cette suppression n'est cependant pas irréversible contrairement à celle effectuée par CDT.

- Recherche des racines irrégulières : Cette recherche a surtout été axée sur la reconnaissance des mots irréguliers : cette approche complémentaire par rapport à celle effectuée par REDUC, permet de réduire les formes irrégulières les plus usuelles lorsqu'elles peuvent être facilement reconnues, et on diminue ainsi les risques de confusions ultérieures. Ceci dit, seuls les verbes dont la racine est très modifiée sont réduits dès cette étape : le programme REDUC tient compte en effet de certaines irrégularités locales de la racine (par exemple, le verbe manger, appuyer, etc. et peut permettre ce type de réduction.

Le principe de la reconnaissance est simple : on dispose de deux fichiers : un dictionnaire de désinences des conjugaisons, classées en plusieurs rubriques selon le type de conjugaison, et un dictionnaire des racines, chacune étant associée à un numéro permettant de repérer le type de conjugaison ou de terminaison permis.

Le dictionnaire des racines comprend actuellement plus de 400 racines et recouvre les racines irrégulières les plus courantes. Il est possible de le compléter au fur et à mesure des analyses, de l'élargir à des formes usuelles, non nécessairement associées à un verbe irrégulier, lorsque leur réduction élimine un risque de confusion ultérieure, l'objectif étant en définitive, le regroupement des formes brutes relevant d'un même "lexème", indépendamment de leur transformation due à leur rôle syntaxique.

2.3 *La réduction des "mots bruts" à leur racine (REDUC)*

2.3.1 *Généralités* : Compte tenu des objectifs généraux, il est assez naturel de chercher à extraire du lexique retenu les marqueurs associés à la syntaxe : désinences grammaticales, certains suffixes. De plus le parti pris de considérer les contextes les plus larges pour appréhender la structure sémantique conduit à essayer de perdre le moins d'information possible, en retenant le maximum de mots. Le regroupement de tous les mots qui peuvent l'être est un moyen d'approche pour cela, les mots peu fréquents se trouvant, autrement, éliminés.

L'objectif visé par ce programme est de donner un outil à l'utilisateur pour effectuer des regroupements à la fois plus aisément et

dans des conditions de plus grande objectivité. Pour cela il s'agit :

- de regrouper tous les mots ayant de "fortes chances" d'être liés sémantiquement, en les réduisant à leur racine commune ;

- d'éliminer tous les mots ayant de "fortes chances" de ne pouvoir être regroupés, et dont la fréquence est faible (lorsque le dictionnaire initial comprend quelque 8000 mots bruts, un premier nettoyage du lexique à étudier est fortement apprécié).

Ces deux objectifs sont partiellement contradictoires. Nous les avons concilié de la manière suivante : après examen de chaque mot du lexique, trois issues sont possibles :

- le mot est jugé réductible avec une "probabilité" d'erreur assez faible, et il est effectivement réduit (issue RE).

- le mot est jugé réductible mais la probabilité d'erreur est forte, et le mot n'est pas réduit (issue RP).

- le mot n'est pas jugé réductible (issue IR).

Ces objectifs étant précisés, il s'agit de choisir une stratégie pour l'analyse morphologique. Nous nous sommes limités à une approche très pragmatique : seuls sont réduits les groupes de mots ayant une même racine commune, et dont les terminaisons peuvent être des désinences grammaticales ou des suffixes reconnus à l'aide d'un dictionnaire des suffixes.

Certains cas d'irrégularité de la racine sont toutefois pris en compte : le cas important des verbes irréguliers est traité antérieurement par le programme CDT ; le cas d'irrégularité locale est traité au moment de la reconnaissance des racines et des suffixes, à l'aide de règles ("règles de liaison" et "règles de transformation" : voir [2]).

Une pondération est associée au groupe de mots pouvant être réduit à une même racine, en fonction du type des terminaisons reconnues (par exemple, si de nombreuses désinences de conjugaison sont observées, la pondération sera forte). Cette pondération permet d'effectuer la réduction la plus plausible, parmi celles possibles.

Le programme que nous proposons peut être encore amélioré. Au fur et à mesure des analyses, il est possible d'ailleurs d'affiner le dictionnaire des suffixes utilisé. Actuellement, nous avons traité environ une douzaine de corpus différents et le taux d'erreurs nous semble acceptable (environ 5%) pour une approche entièrement automatique des racines. Ceci dit, il est possible d'affiner le dictionnaire obtenu à l'aide d'un éditeur de texte : les conditions de travail de l'analyste sont, de toutes façons, grandement améliorées par cette première approche.

2.3.2 *Structure du dictionnaire en entrée (DICO) et en sortie (DICO1)*

En entrée, il comprend trois types de mots : les "mots bruts" susceptibles d'être analysés, les formes irrégulières reconnues par CDT et les mots "hors corpus" introduits par l'utilisateur. Il ne comprend pas les articles et prépositions les plus usuels, déjà éliminés. Ces trois types de mots sont repérés par les modalités d'une variable qualitative appelée "l'indicateur d'état". Seuls les mots bruts sont susceptibles d'être réduits par REDUC.

En sortie, un nouveau dictionnaire est créé, ainsi structuré :

col 35 à 64 le "mot brut" tel qu'il apparaît dans le corpus

col 15 à 33 : le mot réduit à sa racine, s'il a été effectivement réduit ; le terme brut sinon ;

col 9 : la modalité de l'indicateur d'état associé, éventuellement modifiée de la manière suivante, dans le cas où l'ancienne modalité était égale à "a" :

"a" : le mot est "analysable" : il est associé à une des issues RE, RP ou IR (dans ce dernier cas, il faut que sa fréquence soit supérieure à un seuil fixé).

"r" : le mot appartient à une catégorie reconnue par REDUC (prépositions, conjonctions, pronoms, certains adverbes), et l'utilisateur a demandé de la mettre en "élément illustratif".

"s" : le mot est "hors corpus".

"w" : le mot appartient à une catégorie reconnue, et l'utilisateur a demandé de l'éliminer.

"z" : le mot est associé à l'issue IR par REDUC avec un nombre d'occurrences inférieur à un seuil donné par l'utilisateur : il est considéré comme éliminé.

Remarque : ce dictionnaire est ensuite trié en fonction des modalités de l'indicateur d'état. Ce faisant, le programme ne fait que "suggérer" une nouvelle organisation du dictionnaire, mais celle-ci peut être remise en cause ou adaptée par l'utilisateur en fonction de ses hypothèses, soit en modifiant les indicateurs d'état, soit en modifiant les intitulés des "racines" donnés par REDUC (col 15 à 33).

2.3.3 *Les différentes étapes de REDUC* :

- 1-ère étape : Le programme cherche à reconnaître certains types de mots, à l'aide d'un dictionnaire propre. Ces mots sont classés en cinq rubriques :

1 : pronoms ou adjectifs personnels ou possessifs;

2 : auxiliaires être et avoir ;

3 : adverbes, prépositions, conjonctions d'usage courant (le même mot peut avoir plusieurs fonctions : avant, avant de , avant que etc.) ;

4 : les négations ;

5 : pronoms ou adjectifs relatifs, indéfinis ou démonstratifs.

Pour chacune de ces rubriques, l'utilisateur a la possibilité de les conserver, de les éliminer ou de les mettre en éléments supplémentaires (seul l'indicateur d'état est modifié). Quelle que soit l'issue choisie, les mots reconnus sont irréductibles pour REDUC.

- 2-ème étape : On réduit les "mots bruts" à leur "racine". Tout mot qui a été réduit est reconnaissable par le signe "+" positionné en fin de l'intitulé de cette racine (par exemple "chant+" sera associé à "chantons", "chantez"). A la fin de cette étape, chaque mot lié à l'indicateur d'état "a", est associé à l'une des trois modalités RE, RP, IR (réduction effectuée, possible ou irréductible).

- 3-ème étape : Le fichier DICO1 est créé :

Les mots RE sont remplacés par leur racine et associés à l'indicateur d'état "a".

Les mots RP sont conservés sous leur forme brute et associés à l'indicateur d'état "a".

Les mots IR, de fréquence "forte" (supérieure à un seuil donné) sont conservés sous forme brute et associés à l'indicateur d'état "a", les autres sont associés à l'indicateur d'état "z".

Remarque : Après vérification, les sous fichiers "w" et "z" de DICO1 peuvent être détruits définitivement.

2.3.4 Exemple de réduction (tiré de l'étude de l'oeuvre de Durell) :

Voici le début de la liste des mots réduits de DICO1, associés à la modalité "a" de l'indicateur d'état, ordonnés par ordre alphabétique, avec l'effectif de chaque mot brut associé.

Les racines terminées par le signe "+" sont réduites par REDUC, celles terminées par le ".", par CDT, à l'aide du dictionnaire des racines des verbes irréguliers (ou de certains verbes courants).

abandon+	abandon(2), abandonnait(1), abandonnames(1), abandonnant(1) abandonne(3), abandonner(1)
abatt+	abattaient(1), abattre(1)
aboi+	abois(1)
aboiement	aboiement(1), (réductible non réduit)
abord	abord(10 (irréductible)
absence	abscence(3)
absent+	absent(1), absente(1)
absolu+	absolu(1), absolue(1), absolument(7)
absurd+	absurde(5), absurdes(4), absurdites(1)
abuser.	abuse(3), abusive(1)
académi+	académie(1), académique(1)
accable+	accable(3), accablee(2)
accent+	accent(2), accents(2), accentue(1)
accent+	accentuait(1) ("accentue a été classé avec "accent" mais la réduction avec "accentuait" a été envisagée).
accept+	accepta(2), acceptables(1), acceptai(1), accepter(3), accepter(3)
acces+	acces(4)
accessoire+	accessoire(1), accessoires(1)
accompagn+	accompagnai (1), accompagne(2), accompagner(2), accompagnes(1)
accompli+	accompli(3) accomplir(2), accomplissement(1)
accord+	accord(6), accorde(3), accorder(1), accords(1)
accroch+	accrochait(1), accroche(1)
accroup+	accroupi(1), accroupirent(1), accroupis(1)
accueillir.	accueillons(1), accueillirent(1)
achete+	achete(1), acheter(1)
achev+	acheva(1), acheve(1), achevee(2), acheves(1)
acide	acide(1)
acier	acier(1) (REDUC a envisagé la possibilité d'un lien entre "acide" et "acier" ; ces mots seraient sinon "éliminés"-

	associés à l'indicateur d'état "z" - ils seront éliminés à l'étape suivante (programme (CDD))
acquerir.	acquiesca(2), acquisition(1), acquitte(1) (erreur due à une mauvaise définition de la racine dans le dictionnaire des racines, utilisé par CDT, erreur corrigible)
act+	acte(6), actes(2), actions(2)
acteur+	acteur(1), acteurs(1)
actif	actif(1)
action+	action(2)
activ+	activement(1), activités(4)
actuel+	actuel(1), actuellement(1)
admettre.	admettrez(2), admettre(2), admises(1)
administr+	administrateur(1), administre(1), administrer(1)
administra+	administrative(1)
admir+	admirable(2), admirait(2), admirateurs(2) admiration(4), admire(1), admiree(1), admiriez(1), admires(1)
adonnai+	adonnaient(1), adonnait(1)
adore+	adore(1), adoree(1)
adress+	adressa(1), adressait(1), adressant(3), adresse(3), adressees(1), adresser(1)

2.4 *Le calcul du tableau des données (prog CDD)* : Une fois la réduction du lexique obtenue, le programme CDD construit un tableau à double entrée, croisant les u.c. avec les mots réduits, dans lequel on note par "1", la présence du mot dans l'u.c. considérée et par "0", son absence. Ces tableaux peuvent atteindre de grandes dimensions (au maximum 5000 lignes et 1000 colonnes). Ils sont codés sous forme condensée (le nombre de "0" étant souvent grand - plus de 95% -, il est avantageux de ne coder que la position des "1"). Ce programme calcule d'autre part la liste des mots réduits avec pour chaque mot son effectif (nombre d'u.c. différentes dans lesquelles tel mot apparaît).

Remarque : à cette étape une redéfinition de l'u.c. est possible (choix d'un découpage automatique par exemple. Toutefois ce choix respecte les grandes u.c. définies par la numérotation initiale).

2.5 *Tris croisés ou classification* : Un programme de tris croisés permet de décrire le profil des classes d'u.c. associées à un même "mot hors corpus" à l'aide des "mots réduits" les plus cooccurrents. Ce type de tri est classique et très utile pratiquement.

Quant au programme de classification, il s'agit d'un programme que nous avons conçu spécialement pour le traitement de ce type de données. La procédure est descendante hiérarchique. Une première version a déjà été exposée dans ces cahiers [4]. Elle a été améliorée pour le traitement des grands tableaux [5]. Notons que des programmes de classification descendante existent aussi à l'étranger. Celui proposé par M^{r} HILL a été conçu dans un esprit assez proche du notre [5].

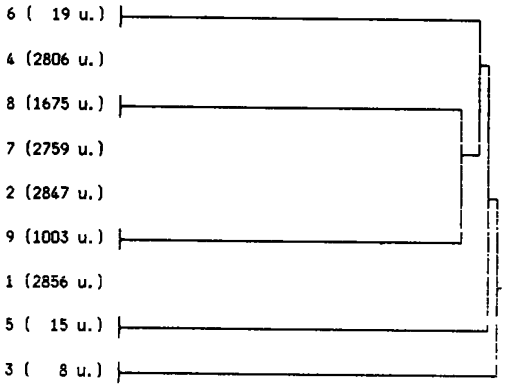
3 *Exemple d'application*

Cet exemple est extrait d'une analyse de l'oeuvre "Le Quatuor" de Laurence DURELL commanditée par Mme Maffre-Savelli pour son étude sur cet auteur. Cette oeuvre regroupe quatre romans : Justine, Balthazar, Mountolive et Cléa, soit environ 1000 pages. Le corpus retenu est constitué d'un échantillon de 100 pages (une sur dix).

Le tableau analysé a été construit automatiquement à l'aide du logiciel ALCESTE (CDT + REDUC + CDD), sans retouche des dictionnaires: il croise les 829 "racines" apparues plus de 3 fois (autres que les articles, prépositions, conjonctions, et pronoms - Les pronoms personnels ont été toutefois conservés en "éléments illustratifs" ou "supplémentaires" -) avec les 2856 u.c. retenues (nous avons choisi comme u.c., les lignes du corpus retranscrit : parmi les 3106 unités, seules sont considérées celles comprenant au moins deux vocables analysés). Le codage de ce tableau est de type logique : "1" si la racine est présente dans l'u.c., "0" sinon. Il comprend 10628 "1" soit 99.5% de "0".

a) *l'arbre des analyses successives* : La procédure de classification est descendante : la classe 1 comprend toutes les u.c..Elle est segmentée en deux : les classes 2 avec 2821 u.c. et 3 avec 13 u.c. ... etc. . (On remarquera que certaines unités peuvent se trouver éliminées en cours d'analyse ; par ex. : 1675 + 1003 (2759).

L'indice de la hiérarchie est simplement proportionnel au nombre d'unités par classe :



Des trois premières dichotomies ne se dégagent que des classes de faible effectif 'cl. 3 : 8 unités ; cl. 5 : 15 unités ; cl. 6 : 19 unités : ce "grapillage" de l'ensemble d'unités initiales est fréquent sur ce type de données du fait du grand nombre de zéros (il suffit de quelques lignes ne comprenant que des vocables rares). Ces classes ne révèlent que des propriétés très locales du corpus, propriétés qui, sur le plan de l'interprétation, sont souvent sans intérêt. Par contre, la quatrième analyse conduit à la définition de deux classes bien proportionnées (classes 8 et 9), susceptibles de traduire des structures sémantiques plus complexes (mettant en jeu, un grand nombre de vocables).

b) *description du profil des classes* : Chaque classe terminale ou non est décrite à l'aide de la liste des "racines" les plus significativement présentes dans cette classe, relativement à l'ensemble des autres (au sens du khi2). Cette procédure peut être utilisée pour des indicateurs logiques autres que ceux analysés dans la C.D.H. : Ils jouent alors le rôle "d'éléments illustratifs" ou "supplémentaires" (par exemple, les "mots hors corpus").

Habituellement, nous relevons tous les vocables associés à un khi2 supérieur à 2.7 : il y en a 260 liés à la classe 8 et 296 liés à la classe 9. En "éléments supplémentaires", nous avons placé les

adjectifs ou pronoms possessifs ou personnels, ainsi que les négations. Voici les listes des racines associées avec un khi2 > 6. et des éléments supplémentaires associés avec un khi2 > 2.7 :

classe numéro 8 (1675 u.c.) :

abord act+ aim+ ajout+ alexandr+ aller. amitie+ amour+ anglais+ Arnauti aucune besoin+ bien bless+ bonn+ cess+ chose+ cle+ compren+ connaitre. copte+ croire.demand+ desir+ devoir. dire. doute+ ecrire. envoy+ epoque eprouv+ errol etonn+ excus+ exempl+ fac+ facon+ faire. falloir. fin+ fois homm+ imagin+ jeune+ jour+ Justine livr+ maitre+ mieux monde mot mots Mountolive oui parc+ part passion+ pense+ person+ Pombal pouvoir. premier+ quart+ reelle+ rentr+ reponse+ rev+ savoir. sentiment+ serie+ seul+ simple+ sincer+ sort+ souvenir. surpris+ tard+ trouv+ verit+ vite voir. vouloir. vrai+ vraiment

*elle *jamais *me *moi *ne *pas *rien *sans *ta *te *tu *vos *vous

classe 9 (1003 u.c.) :

accompagn+ agiter. air allum+ ame ans approch+ argent+ aveugle+ baign+ basse baudet belle+ blanc+ bleu+ boire. boite+ bord+ bouche+ bras+ brill+ bris+ bruit+ brusqu+ cadet+ cafe+ canot caress+ cercle+ chal+ chambre+ chandel+ chemin+ cheve+ ciel+ cigare+ clair+ claqu+ contempl+ cord+ cote+ coul+ coup+ coutume+ cri+ demeur+ depos+ descend+ desert+ doigt+ domestique+ dore+ douce+ droit+ eau eaux ebranl+ ecart+ echapp+ eclair+ ecout+ el embrass+ emporte+ empreint+ enfonc+ entour+ envi+ epaule+ escalier+ espece+ etend+ fenetre+ fixe+ fleur+ fleuve+ flott+ fond+ fort+ fraiche+ frisson+ fum+ genou+ grand+ grecq+ grinc+ groupe guerre hat+ horrible+ ile jet+ lac lampe+ large+ larmes leger+ lente+ levre+ long+ lueur lugubre+ luisant+ lumiere+ main+ mange+ manuscrit+ marin+ mele+ mettre. milieu moment+ mont+ mur murmur+ murs musique narouz neige+ nez noir+ nuages nuit+ obscur+ odeur oeil ongles oreille+ ouvrir. palmiers paraitre. peau petit+ plant+ plonge+ poche+ poitrine port+ prouv+ rat+ reconnaiss+ reflet+ regard+ relev+ rempl+ repos+ repr+ rose+ roug+ roul+ route rue sable+ sal+ sang sembl+ serr+ sombre+ sortir. souffl+ souri+ splend+ sueur tapi+ tendre. tenebres terr+ tete+ tiede+ tir+ tomb+ tour+ trembl+ vetements vid+ vieill+ visage+ vol+ yeux

*leurs *ses *son

La structure observée se rapproche de manière étonnante de celle observées sur les poésies de Rimbaud (voir [3]) :

Les indicateurs de la première et deuxième personne, les négations et l'indicateur de la troisième personne "elle" dans une classe, avec notamment des schèmes d'action (plutôt des verbes) ou exprimant une relation (act+, aller., compren+, connaite., croire., demand+, desir+, devoir., dire., ecrire., envoy+, eprouv+, excus+, faire., falloir., pense+, pouvoir., savoir., vouloir) ou une relation sentimentale ou passionnelle (aim+, amitie+, amour+, passion+, sentiment+) évocation du monde social, humain (forte présence de prénoms),

s'opposant aux autres indicateurs de la 3-ème personne, avec les évocations d'un monde naturel (air, ciel, eau, eaux, fleur+, fleuve+, flott+, lac, marin, nuages, nuit+, palmiers, sable+ souffl+) du corps (bouche, bras, cheve+, epaule+, genou+, levre+, nez, oeil, ongles, oreille+, peau, poitrine, tete+, visage+, yeux) des sensations (plutôt des adjectifs : belle+, douce+, fort+, fraiche+, large+, leger+, lente+, long+, petit+) et notamment ... les couleurs (blanc+, bleu+, dore+, noir+, rose+, roug+), la lumière et l'obscurité (lueur+, luisant+, lumiere+, reflet+, obscru+, sombr+, tenebres).

Cette ressemblance est extrêmement curieuse ... nous nous garderons d'en tirer pour l'instant des conclusions, sinon que la structure observée a peu de chance d'être l'oeuvre du hasard.