Alceste

Un logiciel d'aide pour l'analyse de discours

Notice simplifiée

(de la version de base commune aux versions 4.x)

Max Reinert

max.reinert@printemps.uvsq.fr

Laboratoire PRINTEMPS
Université de Saint-Quentin-en-Yvelines
Centre National de la Recherche Scientifique

Sommaire général

Présentation, Structure & Caractéristiques générales du logiciel

Présentation

Il faut chercher l'origine statistique de la méthode Alceste dans le courant de l'Analyse des Données, animé dès la fin des années soixante par J.P. Benzécri, d'abord à l'Université de Rennes, puis à Paris VI. Ce courant a suscité de nombreuses approches informatisées pour l'analyse statistique des textes. Et le logiciel Alceste fut développé au même moment que d'autres logiciels marqués par leur contact avec ce courant Benzécriste, comme SPADT (Lebart, ENST & CISIA, Paris), LEXICO (Salem, Paris 3), HYPERBASE (Brunet, Université de Nice), pour les plus anciens[1].

La méthode ALCESTE[2] *est également la trace d'un parcours singulier avec ses rencontres et ses hasards. Si elle s'origine, par ses méthodes statistiques, dans les recherches sur l'analyse des données, elle s'est également différenciée aux contacts des méthodes et pratiques des chercheurs en psychologie sociale confrontés à des analyses de réponses à des questions ouvertes ou à des corpus d'entretiens. L' intérêt de l'auteur pour la psychanalyse et la sémiotique a également influencé certaines conceptions de base comme celle d'association ou de répétition.*

Cette activité s'inscrit aujourd'hui dans deux courants principaux de recherche : 1) « L'analyse de discours en sociologie », dans le cadre d'un groupe de recherche du laboratoire de sociologie PRINTEMPS et d'un séminaire de la revue "Langage & Société" ; 2) L'analyse des entretiens cliniques de recherche, en relation avec plusieurs laboratoires [3] *sensibilisés par l'analyse de discours en psychologie clinique.*

---

[1] TRIDEUX de Philippe Cibois (labo PRINTEMPS) date également de cette première époque. Il reprend certaines des techniques du courant Benzécriste sans pour cela en être issu directement. Citons également, aujourd'hui, WEBLEX de S. Heiden de l'E.N.S.&C.N.R.S. de Lyon ainsi que TALTAC de S. Bolasco de l'Universite « La Sapienza » de Rome.

[2] Le sigle *ALCESTE* vaut pour "Analyse des Lexèmes Cooccurrents dans un Ensemble de Segments de Texte"

[3] En contact avec M.C. Noël-Jorand, maître de conférence et chercheure dans une équipe du laboratoire de Biomathématique de la Faculté de Medecine de *La Timone* à Marseille, et également avec l'équipe de Recherche Clinique de l'Université de Toulouse-le Mirail dirigée par M. J. Sauret, professeur et psychanalyste. Ces différents contacts sont " institutionnalisés " à travers une action concertée incitative (ACI - Cognitique) sur *" l'analyse du discours de sujets en situations limites "* dirigée par M.C. Noël-Jorand.

La structuredu logiciel

Le logiciel est composé d'une *interface* et d'un ensemble de *modules de calcul.* L'exécution de ces modules est gérée par un *plan d'analyse* manipulable à partir de l'*interface.* Effectuer une analyse consiste à exécuter ce plan d'analyse sur le corpus préparé pour cela.

Une analyse comprend donc pour l'utilisateur deux moments :
a) Celui de la préparation de son corpus (à l'aide d'un éditeur de texte) ;
b) Celui de l'exécution du plan d'analyse adapté à la forme de son corpus.

Le plan d'analyse est divisé en quatre étapes, chaque étape étant elle-même composée de plusieurs opérations. Le nom d'une opération est désigné par une lettre suivi d'un numéro (A1, A1,…, D5), sigle qui identifie l'étape et le numéro d'ordre de son exécution dans l'étape.

Voici la liste des étapes et des opérations d'une analyse complète (version 4) :

ETAPE A : Segmentation, Lemmatisation & Numérisation du corpus

A1. *Préparation du texte et premier découpage*
A2. *Recherche du vocabulaire et « lemmatisation »*
A3. *Affectation des clés catégorielles aux formes réduites*

ETAPE B: Calcul des Tableaux de correspondances « U.C. x mots » et Classification Descendante Hierarchique

B1. *Définition et sélection des U.C.E.*
B2. *Calcul des tableaux DONN.n soumis à la CDH.*
B3. *Classification Descendante Hiérarchique.*

ETAPE C: Description des classes stabilisées

C1. *Définition des classes retenues.*
C2. *Profil des classes et reclassement contextuel.*
C3. *Analyse Factorielle des Correspondances*

ETAPE D: Calculs supplémentaires sur ces classes

D1. Sélection des U.C.E. significatives par classe.
D2. Recherche des "Segments Répétés Maximaux"
D3. Classification Ascendante des mots à clé valuée
D4. Calcul des concordances
D5. Extraction des sous-corpus associés aux classes

Caractéristiques générales du logiciel[1] et diffusion

Version 2. (1992) :

Corpus maximum traité…	environ 1 million de caractères.
Corpus minimum traité…	environ 70 000 caractères.
Nombre maximum d'unités de contexte élémentaires (U.C.E.)…	10 000
Nombre minimum d'unités de contexte élémentaires (U.C.E.)…	50
Longueur maximum d'une U.C.E. (en nombre de caractères)…	240
Nombre maximum d'unités de contexte initiales (U.C.I.)…	4 000
Nombre minimum d'unités de contexte initiales (U.C.I.)…	1
Nombre maximum de formes initiales…	10 000
Nombre maximum de formes réduites…	1 400
Nombre maximum de formes …	1 400
Nombre maximum de "uns" dans le tableau analysé…	50 000

Version 4. (1998) :

Corpus maximum traité…	environ 6 millions de caractères.
Corpus minimum traité…	environ 70 000 caractères.
Nombre maximum d'unités de contexte élémentaires (U.C.E.)…	10 000
Nombre minimum d'unités de contexte élémentaires (U.C.E.)…	50
Longueur maximum d'une U.C.E. (en nombre de mots)…	2 000
Nombre maximum d'unités de contexte initiales (U.C.I.)…	10 000
Nombre minimum d'unités de contexte initiales (U.C.I.)…	1
Nombre maximum de formes initiales…	90 000
Nombre maximum de formes réduites…	1 400
Nombre maximum de formes …	3 000
Nombre maximum de "uns" dans le tableau analysé…	600 000

Version 5. (2000) :

Corpus maximum traité…	environ 40 millions de caractères.
Corpus minimum traité…	environ 70 000 caractères.
Nombre maximum d'unités de contexte élémentaires (U.C.E.)…	40 000
Nombre minimum d'unités de contexte élémentaires (U.C.E.)…	50
Longueur maximum d'une U.C.E. (en nombre de mots)…	2 000
Nombre maximum d'unités de contexte initiales (U.C.I.)…	40 000
Nombre minimum d'unités de contexte initiales (U.C.I.)…	1
Nombre maximum de formes initiales…	90 000
Nombre maximum de formes réduites…	3 000
Nombre maximum de formes…	10 000
Nombre maximum de "uns" dans le tableau analysé…	1 500 000

[1] Avant 1990, le logiciel fonctionnait comme une bibliothèque de programmes sur les centres de calcul dédiés à la recherche (CIRCE, CNUSC, CICT). La version 1 correspond au premier essai de transfert de cette bibliothèque sur microordinateur. Elle ne fut guère opérationnelle. La version 3, au premier essai de passage sous Windows, qui ne fut pas non plus opérationnelle.

Chapitre I
Premier contact et Rapport d'Analyse[1]

Sommaire

[1] *Partie rédigée avec l'aide de Jean Reinert*

Introduction

Qu'est-ce qu'ALCESTE ?

Le logiciel ALCESTE est un outil d'aide à l'interprétation d'un corpus textuel : entretiens, réponses à une question ouverte, textes littéraires, c'est à dire, tout document écrit à l'aide de l'alphabet latin, des dix chiffres et des signes usuels de ponctuation pourvu qu'il présente une certaine homogénéité et un volume minimum.

Utilisé à l'origine dans des laboratoires de Sciences Humaines, il intéresse aussi à présent les entreprises et les services soucieux d'établir une communication avec un public. Il permet de dépasser les questionnaires à choix multiples des enquêtes habituelles pour l'analyse de questions ouvertes et d'entretiens. La méthode "Alceste", qui est purement formelle, se substitue avantageusement à l'analyse de contenu dans la première phase exploratoire d'une enquête.

Dans cette perspective, c'est la conception même d'une démarche de communication qui est renouvelée par ALCESTE.

A quoi sert ALCESTE ?

Le corpus étant supposé constitué en fonction d'un certain objet d'étude, ALCESTE va dégager les différentes fractures dans la distribution des mots qui pourront être prises par l'utilisateur comme autant de "faits bruts" et révéler l'aspect problématique, multipolaire, de cet objet d'étude. C'est à partir de cette prise de conscience qu'une démarche interprétative peut ensuite être tentée et ouvrir à une analyse de contenu.
Cela dit, le logiciel met en oeuvre des mécanismes d'analyse indépendants du contenu. L'objectif est d'obtenir un premier classement statistique des "unités de contexte" du corpus étudié en fonction de la distribution des mots dans ces "unités", ceci afin d'en dégager les mots les plus caractéristiques (approche des *"mondes lexicaux"* **:** se reporter à la bibliographie).

Dans un premier temps, l'intervention de l'utilisateur est limitée à des "formalités" purement utilitaires (cf. "préparation du corpus"), en sorte qu'aucun présupposé ne vienne influencer les résultats de l'analyse. Puis il ira chercher à l'intérieur des fichiers résultats une vision globale sur sa documentation. Ce niveau d'utilisation ne nécessite pas de connaissances statistiques particulières.

Dans un second temps, l'utilisateur pourra affiner l'analyse, vérifier ou essayer de nouvelles hypothèses interprétatives. La connaissance de certains outils statistiques (Chi2 d'association, Analyse Factorielle des Correspondances), ainsi qu'une familiarisation avec le logiciel, lui sera alors utile.

Il y a toutefois deux conditions pour obtenir un résultat signifiant : la première est que le corpus soit constitué par l'analyste relativement à son intérêt pour un certain objet. C'est le cas (en général!) des réponses à une question ouverte, de recueils d'articles sur un sujet, etc.... mais aussi de textes littéraires, de récits de vie, de récit de rêves, etc. A contrario on ne peut pas espérer une indication de contenu pour un patchwork de fragments disparates réunis par hasard, aussi intéressants soient-ils isolément...

La seconde est que le document soit suffisamment volumineux pour que l'élément statistique entre en ligne de compte. C'est du reste l'intérêt d'ALCESTE de donner très rapidement une vision globale sur une documentation volumineuse qui serait autrement très longue à dépouiller.

Comment se servir d'ALCESTE ?

C'est, bien sûr, le sujet du présent manuel. Nous allons développer ici, rubrique par rubrique, les étapes successives de son utilisation.

Pour ce qui est de l'installation du logiciel et les menus, il faut vous reporter à la notice spécifique qui dépend de la version du logiciel dont vous disposez.

(1) La préparation du corpus :

Il s'agit de l'étape de saisie de votre documentation et de sa mise en forme en sorte qu'elle ne présente pas d'ambiguïté pour ALCESTE. C'est aussi lors de cette étape que vous pourrez "marquer" les éléments d'information qu'il vous importe de distinguer dans l'analyse.

(2) L'analyse planifiée :

Bien qu'elle ne nécessite pas votre intervention et qu'un plan standard réponde à de nombreux usages, on donne dans cette rubrique un aperçu du déroulement de l'analyse au travers de ses différentes étapes ainsi qu'une présentation du plan d'analyse. Cette approche facilitera la lecture des fichiers résultats et ouvrira aux potentialités du logiciel.

(3) La lecture des fichiers résultats et le rapport d'analyse :

Au cours des différentes étapes de l'analyse, ALCESTE produit des fichiers qui n'ont pas tous le même intérêt pour l'utilisateur. Certains sont purement techniques, d'autres donnent des informations sur l'analyse elle-même et seront utiles pour un approfondissement de la démarche analytique. Dans cette rubrique, nous amènerons directement l'utilisateur néophyte aux fichiers qui permettent de construire une représentation synthétique du corpus traité. (Voir le rapport d'analyse).

(4) Quelques exemples de plans d'analyse :

On envisagera quelques possibilités d'intervention, d'une part sur le plan d'analyse, de l'autre sur les dictionnaires avec des exemples de paramétrage. Cela dit, le paramétrage est facilité par l'utilisation de l'interface qui est autodocumentée.

ALCESTE est un logiciel d'une grande transparence. A la façon d'une machine dont les rouages et les mécanismes sont apparents, il donne à voir sa complexité. Mais son utilisation première est simple et, en vous familiarisant avec sa conception, il peut devenir un outil d'investigation vous permettant de "coller" à une documentation touffue.

1.1 La préparation du corpus

Vérifiez tout d'abord que le document que vous voulez analyser dépasse 20 000 mots (environ 2 000 lignes de 70 caractères soit environ 140 000 caractères[1]) tout en n'excédant pas la capacité d'ALCESTE (environ 6 à 8 MO pour la version 4).

Saisie

Vous l'effectuez - par frappe kilométrique ou au scanner - sur un traitement de texte ou un éditeur quelconque, *pourvu qu'il ait une sauvegarde en mode Texte avec saut de ligne.* La présentation n'importe pas mais *vous devez conserver la ponctuation*, qui sera prise en compte pour le calcul des unités de contexte.
Faites l'enregistrement *dans un fichier unique* pour l'ensemble du corpus à traiter, et n'oubliez pas d'effectuer la sauvegarde en "Texte seul avec saut de ligne".

Par exemple sous Microsoft Word : il vous suffit de sauvegarder avec l'option "Texte seulement avec saut de ligne".

Ceci fait, vous allez devoir effectuer un petit travail de "toilettage" de votre document afin qu'il n'y ait pas d'interférence entre des éléments de présentation et des instructions adressées au logiciel.

Les majuscules

Sous Alceste, le rôle des majuscules peut être paramétré. Dans l'utilisation standard, on utilise la règle de conversion suivante : toute majuscule suivie d'une minuscule est transformée en minuscule. Ainsi, la majuscule des mots en début de phrase est automatiquement transformée en minuscule. Par contre, les sigles ne le sont pas. *Un mot retranscrit complètement en majuscule reste inchangé*. Ces mots en majuscules sont placés dans une catégorie à part (marquée par la *clé catégorielle* M) qui n'est généralement pas analysée.

Le signe étoile (*)

Il va jouer un rôle particulier - à votre disposition - de marquage à l'intention d'ALCESTE. *Vous devez donc dans un premier temps le faire disparaître complètement du document,* qu'il figure en appel de note, dans le texte lui-même (La Marquise de ***) ou en signe d'introduction.

Le tiret haut (-) et le tiret bas (_)

Le tiret haut est réservé par ALCESTE pour identifier les locutions. Il ne le reconnaît pas comme signe de liaison dans le corpus : par exemple, *"y a-t-il"* sera reconnu par Alceste de la même manière que *"y a t il" ;* par contre, l'expression *"c'est à dire"* qui est retranscrite dans le dictionnaire des locutions sera reconnu par Alceste et apparaîtra sous la forme *"c'-est-a-dire"* dans les résultats. Cela dit, si vous désirez garder dans le texte même la forme composée d'un mot, vous remplacerez le tiret haut par le tiret bas : par exemple, *"monnaie_unique"*. Vous pouvez aussi introduire cette forme dans le dictionnaire des locutions (voir fichier ALC_LOC).

Si "savoir-faire" n'est pas dans le dictionnaire des locutions (ALC_LOC), on l'écrira "savoir_faire". Si vous voulez que "Général Boulanger" ou "Parti Radical" ou "Acte III, Scène 5" ou "Cat. soc. cult. 2" soient reconnus comme un seul mot, vous les écrirez alors

[1] Toutefois entre 70 000 et 140 000 caractères vous pouvez tenter une analyse, éventuellement en dupliquant le corpus, ce qui permet d'annalyser les mots présents au moins deux fois. Les résultats seront souvent instables… mais ils peuvent mettre en lumière tel ou tel aspect.

: "Général_Boulanger", "Parti_Radical", "Acte_III_scène_5", "cat_soc_cult_2". Cependant si un couple (comme Parti Radical) apparaît plusieurs fois, ALCESTE vous en indiquera tout de même la fréquence (cf. : le dictionnaire des segments répétés).

Le tiret haut (-) en première colonne

Le tiret haut est remplacé automatiquement par un espace sauf dans l'unique cas suivant : s'il sert à introduire un dialogue. Il doit alors être retranscrit en premier caractère de la ligne, et être suivi immédiatement d'un "mot étoilé" (par exemple, le nom de l'interlocuteur : voir les mots étoilés). Le texte du dialogue est retranscrit sur la ligne suivante.

Par exemple :

```
Don Dièguе : - Rodrigue, as-tu du coeur ?
```

devient :

```
-*DON_DIEGUE
Rodrigue, as-tu du coeur ?
```

Si le tiret haut n'introduit pas de dialogue, ALCESTE le supprimera. Notons au passage qu'ALCESTE remplace par un espace tout signe qu'il ne reconnaît pas, donc hors alphabet latin, chiffres, ponctuation (voir fichier ALC_COD).

L'apostrophe (')

Bien sûr, dans le cas général, ALCESTE la reconnaît et vous n'avez pas besoin de vous en préoccuper. Mais attention au rôle particulier qu'elle peut jouer dans certains textes en transcription phonétique : « Sur le boul'vard, déval' les loubards ». Il faudra écrire "boulevard" si on veut que ce mot soit reconnu comme tel (retranscription conseillée) ou "boul_vard" si on veut que cette forme soit reconnue sous cette forme.

Le même problème de transcription se présente **pour des textes en anglais**. Dans ce cas, il est nécessaire de procéder aux modifications suivantes (en respectant l'ordre des exécutions) :

1. Supprimer l'apostrophe quand elle est suivie d'un espace.
2. Changer les apostrophes restantes par le "tiret bas".
3. Supprimer la séquence de lettres "tiret bas" suivi de "s" quand elle est suivie d'un espace.

Par exemple : "Who's there?" devient "Who there?" ; "as by the same cov'nant" devient "as by the same cov_nant"; "Do you believe his tenders' as you call them?" devient "Do you believe his tenders as you call them?".

Il est possible ensuite d'utiliser le plan standard avec les dictionnaires anglais d'Alceste prévus pour cette transformation.

Mots étoilés et lignes étoilées :

Voici une rubrique essentielle parce qu'elle va vous permettre de "marquer" les mots qui vous sont indispensables en tant que repère ou comme information, mais que vous ne voulez pas faire intervenir dans l'analyse (en général simplement parce qu'ils ne figurent pas réellement dans le corpus étudié).
Généralement un corpus est composé de différents textes, chaque texte ayant sa spécificité de production : réponses à une question ouverte, chapitre d'un livre, etc.… Les lignes étoilées permettent de séparer et reconnaître ces énoncés naturels du corpus.
Ainsi, par exemple, dans une question ouverte, on voudra faire précéder chaque réponse par des informations concernant l'interlocuteur (âge, sexe, profession...), informations qu'il importe de retrouver dans les résultats, qui peuvent être objet de questionnement (cf.

"analyse par tris croisés"), mais qui ne sont pas à prendre en compte dans l'analyse elle-même.
Il suffira *d'écrire ces mots sur une ligne* (ou plusieurs) précédant le texte auxquels ils se rapportent et *de faire précéder chacun d'eux par* un espace *et* une étoile.

*On aura par exemple (remarquez au passage les quatre étoiles « **** » introduisant la ligne à partir du premier caractère de la ligne):*

```
**** *rep_3 *sex_masc *gr_soc_cult_2
J'ai profité de l'aide des pouvoirs_publics pour faire isoler ma
maison et c'est à ce moment-là que j'ai choisi le tout_électrique...
```

Ou encore :

```
**** *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
```

Sans doute vous interrogez-vous sur le signe « **** » en début de ligne. *Cette ligne étoilée* introduit pour ALCESTE une *unité de contexte initiale* (ou U.C.I.), notion sur laquelle il est nécessaire de s'attarder.

Les unités de contexte initiales (U.C.I.) :

Les U.C.I. sont les divisions naturelles du corpus (chapitres d'un livre, scènes d'une pièce de théâtre, réponses à une question ouverte etc...). Elles sont les premiers indices d'une structure qu'il convient de signaler à ALCESTE. Vous le ferez en les introduisant par des lignes étoilées. C'est donc l'utilisateur qui définit comme bon lui semble les U.C.I.. *Ce qu'il faut savoir :*

Une ligne étoilée s'ouvre nécessairement sur au moins un mot étoilé. A la place des quatre étoiles, il est possible d'utiliser un numéro d'identification de l'U.C.I.. Par ex :

```
00432 *rep_3 *sex_masc *gr_soc_cult_2

J'ai profité de l'aide des pouvoirs_publics pour faire isoler ma
maison et c'est à ce moment-là que j'ai choisi le tout_électrique...
```

Ce sont les deux seules façons d'introduire une nouvelle U.C.I. Notamment, dans le cas de dialogue, les mots étoilés (avec un tiret en premier caractère) ne constituent pas des séparateurs d'U.C.I..

Les unités de contexte élémentaires (U.C.E.)

Elles sont généralement définies par ALCESTE et vous n'avez pas besoin de vous en préoccuper dans cette phase de mise en forme du corpus. Mais il s'agit d'un concept de base d'ALCESTE qui intervient dans toutes les étapes de l'analyse : autant en dire tout de suite quelques mots.

L'U.C.E. répond à l'idée de " phrase " ou " d'énoncé " mais calibrée en fonction de la longueur (évaluée en nombre de mots) et de la ponctuation (dans l'ordre de priorité : **. ; ? ! : ,** et dominant tous les autres : **$**, ainsi que nous le voyons ci-dessous). C'est à partir de l'appartenance des mots du corpus à ces U.C.E. qu'ALCESTE va établir les matrices par lesquelles sera effectué le travail de classification.

Il y a des documents particuliers où ce découpage en U.C.E. est "naturel" : oeuvre poétique en vers, enchaînement de répliques courtes, chaînes codées de lettres et de chiffres comme on en définit dans les études comportementales. Vous signalerez à ALCESTE ces U.C.E. "naturelles" par le signe : £ suivi d'un retour à la ligne. Selon les

options choisies, le retour à la ligne seul peut être aussi considérée impérativement comme la fin d'une U.C.E..

Comme ci-dessous :

```
**** *ent_16_1

P: Un médecin va venir le soigner.£
D: Quoi ?£
P: Oui, on va venir le chercher. On va le soigner.£
D: Je ne suis pas malade !£
P: Il me décrit sa maladie et il me dit qu'il n'est pas malade.£
```

Notez un nouveau moyen de noter les interlocuteurs dans le cas de dialogues courts. L'usage du tiret-étoile en début de réplique est alors inutile. Et elle doit être terminée par le signe “ £ ” (ou “ $ ” si les réponses ne sont pas trop courtes).

Le nom du corpus

A présent votre corpus est prêt pour l'analyse par ALCESTE. Il ne vous reste plus qu'à le nommer pour l'introduire dans le dossier d'analyse du logiciel[1]. Choisissez un nom connexe (sans blanc) : "Aurélia" ou "Gérard_de_Nerval" mais pas "Gérard de Nerval".

En illustration, *voici un extrait, dans sa présentation pour ALCESTE, d'une enquête (sous forme de question ouverte) réalisée auprès de jeunes en situation scolaire.*

```
**** *sexe_m *assoc_oui

Moi je veux vivre loin de la ville dans une île déserte, avec de super
appareils de musique, et une image grand écran en direct du festival,
rien que musique et image, je veux pas m'inscrire dans la profession
après t'as envie d'une famille, d'une voiture, et puis tu arrêtes pas
d'avoir envie de ceci ou de cela.
Coté sentimental vraiment pas de projets, je veux vivre sur une île
déserte avec la mer en face et surtout pas de bateaux à l'horizon, au
cas où quelques jets de grenades et l'histoire est classée, loin du
trafic polluant des mécaniques et de la gente humaine

**** *sexe_f *assoc_oui

j'ai l'intention d'avoir au moins des enfants, mais en attendant, je
veux arriver à une profession par rapport au baccalauréat technique
que je veux passer, en premier, une bonne situation, et après fonder
une famille, ça il me faudra bien une dizaine d'années. ce qui est
difficile, c'est que les études c'est pas évident. Sinon je veux avant
de me marier vivre avec des copines et m'amuser, ça je le ferai à la
majorité, après je travaille dans mon métier, après je me marie les
gosses et après je suis grand-mère.
Je veux travailler dans le social, être assistante sociale, ou aide
ménagère, ou un boulot avec des gosses de toute manière
```

Chaque ligne étoilée introduit, à l’aide d’une liste de mots étoilés, le sexe et l'appartenance à une association sportive, culturelle ou autres[2].

---

[1] Dans la version de base, si le dossier de travail s'appelle "Dossier Aurelia", le nom du corpus devra s'appeler "Aurelia", et le nom du plan "P_Aurelia"

[2] Quand le corpus est constitué de réponses à une question ouverte, on peut également choisir d'introduire les U.C.I. par un nombre de quatre à huit chiffres au lieu des "****" afin d'identifier le numéro de l’interviewé : `0001 sexe_m assoc_oui.` On ne doit pas cependant introduire le signe « -* » dans le texte des réponses.

La frappe a négligé parfois les majuscules de début de ligne : cela n'importe pas pour l'analyse, les majuscules de début de mot étant retranscrit en minuscules par Alceste. Par contre, il est utile de bien retranscrire la ponctuation (même approximativement) celle-ci étant utilisée par le logiciel pour le découpage du texte en U.C.E..

1.2 L'analyse planifiée

Le corpus est à présent prêt pour l'analyse. La démarche à suivre pour effectuer une analyse standard à l'aide de la version de base consiste à créer un dossier d'analyse ou l'on place le fichier à analyser. Par exemple, le "dossier TOTO" va contenir le fichier texte à analyser appelé "TOTO" (sauvegarder en **"texte seulement avec saut de ligne"**).

Les différentes étapes de l'analyse.

Une analyse comprend 4 étapes au maximum :

L'étape A est une étape de mise en forme et de numérisation des textes. Elle reconnaît les U.C.I. que vous avez vous-même définies, ainsi que les mots étoilés. Différents dictionnaires permettent d'identifier les locutions, les mots outils, d'effectuer une lemmatisation des formes textuelles identifiées (c'est-à-dire, les mots sous leur forme d'entrée dans le dictionnaire). Elle établit un dictionnaire du vocabulaire de votre corpus, puis un dictionnaire des "formes réduites"...
par exemple elle va rassembler les formes "cache", cachées", "cachaient", sous le même radical "cach+er" dont la fréquence sera prise en compte ...

L'étape B est essentiellement une étape de calcul. Elle découpe le corpus en unités de contexte élémentaire (U.C.E.), regroupe ces U.C.E. dans des unités de contexte analysées plus larges de dimension variable, puis effectue leur classification en fonction de la distribution du vocabulaire[1]. Cette étape B est essentielle puisque c'est sur ces classes, caractérisées par leur vocabulaire dominant, que va s'appuyer ensuite votre démarche interprétative.
Dans l'option standard, deux classifications successives sont effectuées en faisant varier légèrement[2] la longueur de l'unité de contexte analysé afin de contrôler la stabilité des classes obtenues.

L'étape C donne une première description des classes obtenues. C'est elle qui fournit les principaux fichiers résultats. On y trouve les différentes classes retenues, leur dépendance mutuelle, le vocabulaire dominant de chacune d'elle, les mots étoilés et les mots outils caractéristiques. C'est sur ces éléments que vous baserez votre interprétation.

L'étape D effectue des calculs complémentaires sur chacune des classes. Par exemple, c'est à cette étape que les unités de contexte les plus représentatives de chaque classe sont calculées et extraites, que les segments répétés, les classifications ascendantes hiérarchiques sont calculés.

Le plan d'analyse

Une analyse se déroule donc en quatre étapes subdivisées chacune en plusieurs opérations. Le plan d'analyse consiste dans le paramétrage de ces opérations.

Vous n'avez pas a priori à vous en préoccuper. Si votre corpus est d'une nature textuelle ordinaire: entretiens, oeuvre littéraire, recueil d'articles, questions ouvertes, ***le plan standard*** *convient généralement à votre analyse dans une première approche.*

Mais, avec une certaine pratique d'ALCESTE, vous voudrez peut-être modifier les conditions de l'analyse pour avoir une plus grande maîtrise sur vos résultats, éventuellement en modifiant des dictionnaires.

---

[1] A la suite d'un calcul croisant les U.C.E. avec le vocabulaire, elle procède à une partition plus ou moins recouvrante de l'ensemble des U.C.E. en fonction de la fréquence des formes réduites.
[2] Cette modulation peut être aussi contrôlée par l'utilisateur grâce au paramétrage.

Au cas où votre corpus est d'une forme particulière : réponses courtes, oeuvre versifiée, textes en langue étrangère, transcriptions codées, vous adapterez le plan d'analyse à cette forme.

Vous pourrez le faire, au niveau de la retranscription du corpus, en intervenant sur les dictionnaires, ou en changeant le paramétrage du plan d'analyse.

Vous trouverez une information complète de la structure d'un plan dans l'annexe "Description du Plan d'Analyse" .

1.3 Les Fichiers Résultats (généralités)

Une fois l'analyse achevée (elle peut durer de quelques minutes à plusieurs heures selon l'importance de votre corpus et la rapidité de votre micro), vous éditez les fichiers résultats sur le traitement de texte ou l'éditeur de votre choix, mais avec *une police de caractères non proportionnelle (par exemple : courrier taille 10),* afin de respecter l'alignement des colonnes.

Leur volume va peut-être vous dérouter mais comme nous l'avons déjà signalé, certains sont purement techniques et d'autres ne vous intéresserons que si vous voulez modifier les conditions de l'analyse.

Tout d'abord, repérons-nous dans leur notation. Elle suit les quatre étapes de l'analyse, et en même temps l'ordre d'apparition des fichiers.

L'étape A produit les fichiers A1_..., A2_..., A3_... suivi du nom du fichier. Ainsi vous trouverez A2_DICO qui est la liste alphabétique du vocabulaire de votre corpus,
A3_DICB qui est le dictionnaire des formes réduites,
A3_DICB.tri qui est la liste des formes réduites les plus fréquentes.

L'étape B est, comme nous l'avons vu, surtout une étape de calcul. On pourra y consulter B3_arbre.1 et B3_arbre.2 qui sont les dendrogrammes des deux classifications descendantes hiérarchiques à l'issue desquelles est réalisée la partition du corpus en classes.

Ce sont les étapes C et D qui produisent les fichiers résultats proprement dits. Les classes stables y sont décrites dans le rapport d'analyse et dans les fichiers suivants :

C1.cpcdh.121, résultat de la comparaison des deux classifications.
C2_DICB.121, le dictionnaire des formes réduites affectées dans une classe.
D1_UCE.121, liste des U.C.E. avec leur appartenance aux classes.
D2_SR.121, "segments répétés" significatifs de chaque classe.

Concrètement tous les fichiers de la première étape sont situés dans le dossier d'analyse (par exemple le "dossier TOTO"). Ce dossier d'analyse comprend un sous-dossier intitulé, dans le cas standard, "&&_0", qui réunit tous les fichiers résultats des étapes B, C et D. On peut, en effet, modifier le plan et construire plusieurs sous-dossiers "&&_1", "&&_2", etc.…, avec dans chacun de ces sous-dossiers des analyses spécifiques, mais nous n'en sommes pas là…

*Les principaux résultats d'une analyse sont réunis dans le **Rapport d'analyse** qui est édité dans le dossier d'analyse (par exemple, dossier TOTO). Ce rapport d'analyse suffit bien souvent pour un premier dépouillement des résultats.*
*Après une analyse, nous vous conseillons donc de consulter directement ce **Rapport**. En voici une description précise.*

1.4 Le Rapport d'analyse… à l'aide d'un exemple

Le corpus d'essai "avenir" et le plan d'analyse

Prenons maintenant comme exemple l'analyse du corpus proposé pour l'essai, le corpus "avenir". Il s'agit d'un ensemble de réponses de jeunes adolescents de la banlieue de Toulouse à la question : "quels sont vos projets d'avenir dans le domaine professionnel, familial ou autre ?". En voici un extrait :

```
0011 *sexe_m *assoc_oui *sa_12
Je n'ai pas l'habitude de faire des projets, je vis au jour le jour.
les adolescents font des projets à partir du moment où il se rendent
compte qu'ils ne doivent compter que sur eux même et qu'ils doivent se
prendre en charge… de quel genre de projets s'agit-il ? projet pour
mon métier, une vie assez facile, sans trop d'argent ni trop peu, une
maison, une voiture, une moto, tout ça grâce à la police, car je veux
devenir policier…
```

Les lignes "étoilées" séparent les différentes réponses et contiennent des informations "exogènes" : sexe et appartenance à une association (*sa_12 définit la sous classe des sujets de sexe masculin participant à une association)…
Ce corpus a été placé dans le "dossier avenir" et l'on a exécuté le plan standard sans modification sauvegardé sous le nom : "P_avenir".

Après analyse, on ouvre le rapport d'analyse du "dossier avenir" à l'aide de Word par exemple. On sélectionne l'ensemble du texte pour le mettre sous la police de caractères "courrier 10 points". C'est le contenu de ce fichier qui est décrit présentement.

Le rapport d'analyse s'ouvre sur le nom de votre plan d'analyse et la liste des instructions contenues dans ce plan que nous n'expliciterons pas dans ce chapitre (voir en annexe, « description du plan d'analyse »).

```
 ---------------------------------
 * logiciel ALCESTE (version 4.5) *
 ---------------------------------

 Plan de l'analyse :P_avenir ; Date :  1/ 6/95; Heure : 11:22:12

 :Dossier avenir:&&_0:
```

Le logiciel contrôle l'existence du sous dossier &&_0. La mention &&_0 dans le rapport d'analyse permet d'identifier le sous-dossier comprenant les fichiers résultats obtenus avec ce plan d'analyse (P_avenir).

```
 &avenir
 ET 1 1 1 1
 A  1 1 1
 B  1 1 1
 C  1 1 1
 D  1 0 0 0 0
 A1   1   0   0
 A2   3   1
 A3   1   1   0
 B1   0   4   0   1   1   0   1   1   0
 B2   2   2   0   0   0   0   0   0
 B3  10   4   1   1   0   0   0   0   0   0
 C1   0   121
 C2   0   3
 C3   0   0   1   1   1   2
 D1   0   2   1   2
 D2   0
```

```
D3   5   a   2
D4   1  -2   1
D5   0   0   1
```

Il s'agit ici du plan standard, le plus utilisé. Ce plan s'adapte automatiquement à la grandeur du corpus analysé. Quand vous serez familiarisé avec ALCESTE, vous pourrez, au besoin, modifier le paramétrage et changer ainsi les conditions de l'analyse.

Résultats de l'étape A

Les sorties de la première étape A donnent des informations générales sur le corpus...

```
---------------------
A1: Lecture du corpus
---------------------

A12 : Traitement des fins de ligne du corpus :
N° marque de la fin de ligne :

Nombre de lignes étoilées    :          61
```

Le corpus est composé de 61 réponses (les U.C.I.). Chaque U.C.I. est découpée en petit segment de texte en fonction de la ponctuation si elle existe, avec la contrainte d'être inférieur à 250 caractères. Ces segments ponctués sont ensuite éventuellement réunis dans des segments plus longs (en restant cependant inférieurs à 250 caractères) en privilégiant les coupures associées à une ponctuation forte (les segments de texte calculés).

```
--------------------------
A2: Calcul du dictionnaire
--------------------------

Nombre de formes distinctes                :         827
Nombre d'occurrences                       :        4282
Fréquence moyenne par forme                :           5
Nombre de hapax                            :         437
Fréquence maximum d'une forme              :         221

 52.84% des formes de fréq. ≤     1 recouvrent 10.21% des occur.;
 76.66% des formes de fréq. ≤     3 recouvrent 20.74% des occur.;
 87.67% des formes de fréq. ≤     7 recouvrent 31.69% des occur.;
 92.38% des formes de fréq. ≤    13 recouvrent 40.71% des occur.;
 95.16% des formes de fréq. ≤    22 recouvrent 50.09% des occur.;
 97.34% des formes de fréq. ≤    36 recouvrent 62.26% des occur.;
 98.43% des formes de fréq. ≤    52 recouvrent 72.28% des occur.;
 99.15% des formes de fréq. ≤    73 recouvrent 81.08% des occur.;
 99.76% des formes de fréq. ≤   116 recouvrent 91.76% des occur.;
100.00% des formes de fréq. ≤   221 recouvrent100.00% des occur.;
```

Voir dans le glossaire, la terminologie suivante : occurrence, forme, forme réduite, hapax.
Relevons le nombre total d'occurrences du corpus : 4287 ; Le nombre de mots utilisé une fois (dit hapax) :437 ; La fréquence moyenne d'une forme : 5. Chaque forme différente apparaît ainsi, en moyenne, 5 fois dans ce corpus.

On remarquera que 50 % des occurrences recouvrent 95 % des formes les moins fréquentes… et donc 5 % des formes les plus fréquentes (généralement les articles, prépositions, conjonctions, etc…)

Durant l'opération A3, les mots sont catégorisés à l'aide de *clés catégorielles*[1] (lorsqu'ils sont reconnus par dictionnaire). L'utilisateur se sert des clés catégorielles pour choisir les

[1] Lettre ou chiffre permettant d'identifier une " catégorie " de mots a priori (voir A2_DICO, A3_DICB)

mots analysés : la catégorie de mots liée à une clé peut être mise dans l'analyse (code 1), rejetée de l'analyse (code 0) ou mise en élément supplémentaire (code 2).

Voici la liste des catégories de mots gérés par Alceste (version 4). Par exemple les noms, verbes, adjectifs et adverbes sont analysés si l'on utilise ce plan standard.

```
 ---------------------------------------------------
 A3 : Liste des clés et valeurs d'analyse (ALC_CLE) :
 ---------------------------------------------------

 A 1 Adjectifs et adverbes
 B 1 Adverbes en "ment"
 C 1 Couleurs
 D 1 mois/jour
 E 1 Epoques/ Mesures
 F 1 famille
 G 1 lieux, pays
 I 2 Interjections
 J 2 Nombres
 K 0 Nombres en chiffre
 M 2 Mots en majuscules
 N 1 Noms
 U 0 Mots non trouvés dans DICIN (si existe)
 V 1 Verbes
 W 1 Prénoms
 X 1 formes non reconnues et fréquentes
 Y 1 formes reconnues mais non codées
 0 2 Mots outils non classés et prépositions usuelles
 1 2 Verbes modaux(ou suceptibles de l'être)
 2 2 Marqueurs d'une modalisation
 3 2 Marqueurs d'une relation spatiale
 4 2 Marqueurs d'une relation temporelle
 5 2 Marqueurs d'une intensité
 6 2 Marqueurs d'une relation discursive
 7 2 Marqueurs de la personne (personnels possessifs)
 8 2 Démonstratifs, indéfinis et relatifs
 9 2 Auxiliaires être et avoir
   1 Formes non reconnues
```

Après reconnaissance des formes, on appelle parfois « mot » pour simplifier, la forme réduite :

```
A34 : Fréquence maximale d'un mot analysé   :        3000


 Nombre de mots analysés                     :         503
 Nombre de mots supplémentaires de type "r"  :         178
 Nombre de mots supplémentaires de type "s"  :           7
 Nombre d'occurrences retenues               :        3621
 Moyenne par mot                             :    4.464024
 Nombre d'occurrences analysables (freq > 3) :    937 soit   30.82  %
 Nombre d'occurrences supplémentaires        :        2103
```

Après réduction des pluriels, des désinences de conjugaison, après élimination des hapax, il reste donc 503 « mots » susceptibles d'être analysés, 178 « mots outils », 7 « mots étoilés » (ceux des lignes étoilées du corpus).

Résultats de l'étape B

Voici maintenant les valeurs des principaux paramètres de l'opération B1 : fréquences minimum maximum des formes retenues ; longueur des U.C.E. en nombre de mots (voir U.C.E. dans

glossaire). Le code de fin d'U.C.E. est en rapport avec la ponctuation : une valeur forte indique le ch d'une ponctuation forte comme fin d'U.C.E.. Le calcul de l'U.C.E. combine donc deux dimensions longueur en nombre de mots et la ponctuation. Selon les valeurs choisies, c'est l'une ou l'a dimension qui est dominante dans le calcul.

```
-----------------------------------------
B1: sélection des U.C.E. et calcul des données
-----------------------------------------


B11: Le nom du dossier des résultats est &&_0
B12: Fréquence minimum d'un "mot" analysé   :           4
B13: Fréquence maximum d'un  "mot" retenu   :        9999
B14: Fréquence minimum d'un  "mot étoilé"   :           1
B15: Code de fin d'U.C.E.                   :           1
B16: Nombre d'occurrences par U.C.E.        :          30
B17: Elimination des U.C.E. de longueur     ≤           0

Fréquence minimum finale d'une forme analysée           4
Fréquence minimum finale d'une forme type "s"           1

Nombre de mots analysés                     :          96
Nombre de mots sup type "r"                 :          84
Nombre total de mots                        :         180
Nombre de mots supplémentaires de type "s"  :           7
Nombre de lignes de B1_DICB                 :         187

Nombre d'occurrences analysées              :         937

Nombre d'u.c.i.                             :          61
Nombre moyen de "mots" analysés / u.c.e.     :    7.808333
Nombre d'u.c.e.                             :         120
Nombre d'u.c.e. selectionnées               :         120
Nombre de couples                           :        2398
```

L'opération B1 définit les lignes et les colonnes du tableau de données de base croisant les U.C.E. et le vocabulaire. En colonnes, ce tableau comprend ici 96 mots pleins (analysés), 84 mots outils (supplémentaires), et 7 "mots étoilés" (ceux des lignes étoilées). Le nombre d'U.C.I. est 61.
L'opération B1 calcule aussi la liste des *couples* d'occurrences composés par la succession de deux formes (voir glossaire), liste qui sera utilisée par l'opération D2 pour le calcul des *segments répétés*.

Le calcul proprement dit des tableaux de données est effectué par l'opération B2. Trois tableaux sont calculés avec ce plan : B2_DONN.0, B2_DONN.1, et B2_DONN.2.
B2_DONN.0 est le tableau de base U.C.E. x formes avec les caractéristiques présentées ci-dessus. Il est calculé automatiquement pour tout plan d'analyse.
Quant aux deux autres tableaux, ils sont constitués spécifiquement pour la classification avec, en colonnes, les mots analysés et, en lignes, des unités de contexte de longueur variable. Cette stratégie un peu compliquée a été adoptée pour tester la stabilité des résultats en fonction d'une petite variation dans la définition des unités de contexte. En effet, si les résultats sont stables, les aspects arbitraires du choix des unités de contextes sont sans conséquence.
Dans l'analyse standard, les deux tableaux DONN.1 et DONN.2 sont calculés avec les caractéristiques suivantes :

```
-------------------
B2: Calcul de DONN1
-------------------

Nombre de formes par unité de contexte      :          10
Nombre d'unités de contexte                 :          96
```

Remarque : il s'agit du nombre de formes analysées différentes.

```
-------------------
B2: Calcul de DONN2
-------------------

Nombre de formes par unité de contexte      :          12
Nombre d'unités de contexte                 :          83
```

Dans le premier tableau, une unité de contexte est définie par concaténation des U.C.E. successives d'une même U.C.I. jusqu'à ce que le nombre de mots différents analysés dépassent 10 (12 pour le deuxième tableau).

Chaque tableau est ensuite soumis à la Classification Descendante Hiérarchique (voir glossaire) :

```
-----------------------------------------------------
B3: Classification descendante hiérarchique de DONN.1
-----------------------------------------------------

Elimination des mots de fréquence >  3000 et <   4
Nombre d'items analysables                  :          67
Nombre d'unités de contexte                 :          96
Nombre de uns                               :         807

-----------------------------------------------------
B3: Classification descendante hiérarchique de DONN.2
-----------------------------------------------------

Elimination des mots de fréquence >  3000 et <   4
Nombre d'items analysables                  :          67
Nombre d'unités de contexte                 :          83
Nombre de uns                               :         788
```

Les tableaux traités par la C.D.H. sont des tableaux logiques (valeur "zéro" pour l'absence d'un mot dans une unité de contexte et valeur "un" sinon). Les tableaux sont généralement très vides (jusqu'à 99% de "zéros"). Ils sont caractérisés par le nombre de "uns" analysés. Par exemple, le premier tableau comprend 67 x 96 = 6432 cases dont 807 contiennent la valeur "un", soit près de 87 % de "zéros".

Résultats de l'étape C

C1 : Comparaison des deux classifications.

Une fois les deux classifications effectuées, il s'agit de comparer les classes obtenues. Cette comparaison est simplifiée par le mode de calcul des U.C.. En effet, une unité de contexte de DONN.1 ou DONN.2 regroupe toujours un nombre entier d'U.C.E. si bien qu'une classe d'U.C. peut toujours être considérée comme une classe d'U.C.E.. Il suffit ensuite de comparer les classes d'U.C.E. entre elles :

```
----------------------------
C1: intersection des classes
----------------------------

Suffixe de l'analyse              :121
Date de l'analyse :25/ 8/96
Intersection des classes RCDH1 et RCDH2

Nombre minimum d'U.C.E. par classe   :    10

DONN.1 Nombre de mots par uc :          10
       Nombre d'uc           :          96

DONN.2 Nombre de mots par uc :          12
       Nombre d'uc           :          83

  78 u.c.e classées sur   120 soit  65.00 %

Nombre d'u.c.e. distribuées:          100

 tableau croisant les deux partitions :

  RCDH1 *   RCDH2

 classe  *    1    2    3

   poids *   23   44   33

  1   26 *   21    4    1
  2   57 *    2   40   15
  3   17 *    0    0   17
```

Dans notre exemple, sur les 120 U.C.E. définies par l'opération B1, 100 ont été classées simultanément dans les deux classifications, mais seulement 78 sont associées aux "mêmes classes". Le tableau ci-dessus permet de préciser le sens de cette expression : d'abord, le programme doit définir un niveau de partition stable parmi toutes les partitions possibles (ici, une partition en trois classes) ; puis, il mesure le degré de stabilité en construisant le tableau de cooccurrences entre la partition obtenue lors de la première analyse et la partition obtenue dans la seconde.
Les valeurs sur la diagonale indiquent le nombre d'U.C.E. restées stables dans les deux classements. Dans la suite des opérations, seules cette partie stable sera utilisée pour décrire les résultats. Elle représente ici 21 + 40 + 17 = 78 U.C.E. "bien classées" sur 120 soit 65 % des U.C.E. définies. On notera que 22 U.C.E. ont un classement différent et 20 U.C.E. ont été éliminées à une des étapes de calcul de l'une ou l'autre C.D.H. du fait de leur "poids" trop faible (i.e.: poids de l'U.C.E. = nombre de mots différents analysés présents dans l'U.C.E.). Suit le tableau des liens entre classe, exprimé à l'aide d'un χ^2 signé (voir glossaire).

```
 tableau des chi2 (signés) :

  RCDH1 *   RCDH2

 classe  *    1    2    3

   poids *   23   44   33

  1   26 *   66  -11  -13
  2   57 *  -28   36   -2
  3   17 *   -6  -16   41
```

Il est possible de consulter les arbres complets (dendrogrammes) des deux classifications dans les fichiers résultats correspondant (B3_arbre.1 et B3_arbre.2). On trouve cependant

dans le rapport d'analyse, les arbres reconstruits à partir de la partition stable mise en évidence précédemment. Les noeuds indiqués (18 et 19 ci-dessous) sont les noeuds des arbres d'origine... aux aléas des classes artefacts éliminées.

```
 Classification Descendante Hiérarchique…
  Dendrogramme des classes stables (à partir de B3_RCDH1) :

                 ----|----|----|----|----|----|----|----|----|----|

  Cl. 1 (  21uce) |-------------------------------------+
             18                                         |----------+
  Cl. 2 (  40uce) |-------------------------------------+          |
             19                                                    +
  Cl. 3 (  17uce) |-----------------------------------------------+

  Classification Descendante Hiérarchique…
  Dendrogramme des classes stables (à partir de B3_RCDH2) :

                 ----|----|----|----|----|----|----|----|----|----|

  Cl. 1 (  21uce) |-----------------------------------------------+
             13                                                    +
  Cl. 2 (  40uce) |---------------------------------+             |
             17                                    |-------------+
  Cl. 3 (  17uce) |---------------------------------+
```

Dans cette analyse, les deux arbres ne sont pas identiques bien que les classes terminales restent stables . Les classes 1 et 3 restent cependant toujours fortement opposées.

C2 : Profil des classes.

L'opération C2 calcule le profil des classes sur le vocabulaire et sélectionnent les mots les plus spécifiques de chacune d'elles :

```
  ----------------------
  C2: profil des classes
  ----------------------

  Chi2 minimum pour la sélection d'un mot     :        2.00

  Nombre de mots (formes réduites)            :         180
  Nombre de mots analysés                     :          96
  Nombre de mots "hors corpus"                :           7
  Nombre de classes                           :           3

            78 u.c.e. classées soit   65.00000    %

  Nombre de "uns" analysés                    :         554
  Nombre de "uns" suppl. ("r")                :         922

  Distribution des u.c.e. par classe...

   1ere classe :    21. u.c.e.  164. "uns" analysés ;  230. "uns" sup..
   2eme classe :    40. u.c.e.  276. "uns" analysés ;  525. "uns" sup..
   3eme classe :    17. u.c.e.  114. "uns" analysés ;  167. "uns" sup..

  --------------------------
  Classe n°  1 => Contexte A
  --------------------------
  Nombre d'u.c.e.                 :     21. soit : 26.92 %
  Nombre de "uns" (a+r)           :    394. soit : 26.69 %
  Nombre de mots analysés par uce :   7.81

  num   effectifs   pourc.    chi2 identification
```

```
  2     7.   7.  100.00   20.87      A belle+
  8     4.   6.   66.67    5.22      A plein+
 23     6.  10.   60.00    6.38      N devenir+
 27     6.   7.   85.71   13.51      N femme+
 32    11.  12.   91.67   30.21      N maison+
 36     7.   7.  100.00   20.87      N monde+
 42     2.   3.   66.67    2.50      N professeur+
 46     4.   6.   66.67    5.22      N sport+
 50    11.  13.   84.62   26.39      N voiture+
 58     5.   5.  100.00   14.50      V esper+er
 63     8.  16.   50.00    5.45      V mari+er
 75     2.   3.   66.67    2.50      V trouv+er
 78     3.   3.  100.00    8.47      V voyag+er
132 *  17.  48.   35.42    4.58 *    6 et
140 *   3.   3.  100.00    8.47 *    6 sans
150 *  10.  27.   37.04    2.15 *    7 me
167 *   8.  14.   57.14    7.92 *    8 tout
170 *  14.  41.   34.15    2.29 *    9 avoir
176 *   2.   3.   66.67    2.50 *    J cinq
183 *   5.   9.   55.56    4.24 *      *sa_11
187 *  13.  31.   41.94    5.89 *      *sexe_m

Nombre de mots sélectionnés    :   21
```

La classe 1 (21 U.C.E.) qui définit le "contexte A" contient 26.92 % des U.C.E. retenues dans l'analyse.
Son vocabulaire le plus spécifique est basé sur les racines (formes réduites) : "belle+", "maison+", "monde+", "voiture+".
Les mots outils dominants sont "sans" et "tout". Ces mots sont précédés d'une étoile pour indiquer qu'ils n'ont pas contribués au calcul de la classe (voir glossaire : élément illustratif)
Observons la ligne "*sexe_m". Elle indique que ce sont significativement des garçons qui contribuent aux U.C.E. de cette classe : 13 U.C.E. sur les 21 U.C.E. de la classe proviennent des réponses de garçons. Le nombre 31 renvoie au nombre d'U.C.E. classées dans l'une des trois classes relatives à une réponse de garçon. Autrement dit 41.94 % des U.C.E. "garçons" sont dans cette classe alors que cette dernière ne représente que 26.96% des U.C.E. classées. Cette différence est significative au sens du χ^2 (à un degré de liberté) égal ici à 5.89.

Une clé est attribuée à tout mot associé à une classe avec un χ^2 minimal. Cette clé est appelée “ clé contextuelle ”. Elle vaut “ A ” pour les mots spécifique de la classe 1 ; B, pour les mots spécifiques de la classe 2, etc.

```
-------------------------
Classe n°  2 => Contexte B
-------------------------
Nombre d'u.c.e.                :     40. soit : 51.28 %
Nombre de "uns" (a+r)          :    801. soit : 54.27 %
Nombre de mots analysés par uce :   6.90

num   effectifs   pourc.    chi2 identification

  3     5.   6.   83.33    2.67      A decide+
  4     3.   3.  100.00    2.96      A difficile+
 15    14.  19.   73.68    5.05      N an+
 33     3.   3.  100.00    2.96      N mari+
 34    14.  17.   82.35    8.40      N metier+
 41     4.   4.  100.00    4.01      N pouvoir+
 43    13.  18.   72.22    4.11      N projet+
 45     3.   3.  100.00    2.96      N societe+
 49    10.  13.   76.92    4.11      N vie+
 70    11.  13.   84.62    6.94      V realis+er
 73     5.   5.  100.00    5.08      V rest+er
```

```
  85     6.   6.  100.00    6.18      Y fait
  88     8.   8.  100.00    8.47      Y jeune+
  92     5.   6.   83.33    2.67      Y professionn+el
 101 *  13.  20.   65.00    2.03 *    1 falloir.
 102 *   8.  11.   72.73    2.36 *    1 pouvoir.
 119 *   4.   4.  100.00    4.01 *    5 beaucoup
 123 *   4.   4.  100.00    4.01 *    5 plus-d<
 125 *   6.   8.   75.00    2.01 *    6 aussi
 126 *  10.  14.   71.43    2.77 *    6 car
 139 *   6.   6.  100.00    6.18 *    6 pour-qu<
 145 *   4.   4.  100.00    4.01 *    7 ils
 147 *   3.   3.  100.00    2.96 *    7 leur
 148 *   6.   6.  100.00    6.18 *    7 leurs
 149 *  10.  14.   71.43    2.77 *    7 ma
 154 *   7.   7.  100.00    7.31 *    7 se
 156 *  13.  18.   72.22    4.11 *    8 ca
 163 *   7.   9.   77.78    2.86 *    8 on
 171 *   3.   3.  100.00    2.96 *    9 est
 172 *  13.  19.   68.42    2.95 *    9 etre
 174 *   6.   6.  100.00    6.18 *    9 sont
 178 *   4.   4.  100.00    4.01 *    J dix
 181 *  19.  29.   65.52    3.74 *      *assoc_non
 185 *  15.  20.   75.00    6.06 *      *sa_21
 186 *  19.  27.   70.37    6.02 *      *sa_22
 186 *  34.  47.   72.34   20.99 *      *sexe_f

Nombre de mots sélectionnés    :   35
```

Cette classe 2 définissant la clé contextuelle “ B ” contient 51.28 % des U.C.E. classées. On observera que le vocabulaire utilisé est plus socialisé (jeune+, société, métier). La présence de verbes modaux comme *"pouvoir", "falloir"* est caractéristique d'une attitude plus "active" voire "revendicative du sujet . Ce contexte est plus spécifiquement féminin si on le compare à celui de la classe 1 ou 3 par exemple.

```
--------------------------
Classe n°  3 => Contexte C
--------------------------
Nombre d'u.c.e.                :     17. soit : 21.79 %
Nombre de "uns" (a+r)          :    281. soit : 19.04 %
Nombre de mots analysés par uce :   6.71

num   effectifs   pourc.    chi2 identification

  9     3.   4.   75.00    7.00      A premier+
 14     5.   6.   83.33   14.44      N annee+
 18     3.   4.   75.00    7.00      N baccalaureat<
 21     4.   6.   66.67    7.68      N compte+
 25     7.  20.   35.00    2.75      N etude+
 26     4.  10.   40.00    2.23      N famille+
 56     5.  10.   50.00    5.35      V continu+er
 59     2.   3.   66.67    3.69      V essa+yer
 61     3.   5.   60.00    4.58      V fond+er
 65     4.   5.   80.00   10.62      V pass+er
 72     3.   5.   60.00    4.58      V rentr+er
 81     2.   3.   66.67    3.69      Y cote+
 86     9.   9.  100.00   36.51      Y format+ion
 87     3.   5.   60.00    4.58      Y independ+ant
 94     9.   9.  100.00   36.51      Y techn+16
112 *   5.   9.   55.56    6.80 *    4 apres
115 *   2.   3.   66.67    3.69 *    4 longtemps
129 *   3.   3.  100.00   11.20 *    6 encore
130 *   2.   3.   66.67    3.69 *    6 enfin
143 *   2.   3.   66.67    3.69 *    6 surtout
182 *  16.  49.   32.65    9.12 *      *assoc_oui
```

```
  184 *  12.  22.   54.55   19.28 *      *sa_12
  188 *  12.  31.   38.71    8.64 *      *sexe_m

 Nombre de formes sélectionnées :   19
```

Cette classe se passe de commentaire. Elle est plus spécifique des garçons participant à une association.

```
 Liste des valeurs de clé :
    0 si chi2 <  2.71
    1 si chi2 <  3.84
    2 si chi2 <  5.02
    3 si chi2 <  6.63
    4 si chi2 < 10.80
    5 si chi2 < 20.00
    6 si chi2 < 30.00
    7 si chi2 < 40.00
    8 si chi2 < 50.00
```

Les *clés catégorielles* (voir glossaire) affectées aux mots dès l'opération A2 sont distribuées dans les classes afin d'apprécier leurs liens avec chacune d'elles :

```
  Tableau croisant classes et clés :

        * Classes *     1     2     3

  Clés  * Poids   *   385   781   274

      A *      49 *    18    25     6
      B *       3 *     0     2     1
      J *      17 *     5    10     2
      N *     237 *    79   117    41
      V *     151 *    45    71    35
      W *       1 *     1     0     0
      Y *     110 *    21    58    31
      0 *      27 *     7    15     5
      1 *      80 *    20    47    13
      2 *      62 *    16    37     9
      3 *      45 *    11    27     7
      4 *      26 *     7     9    10
      5 *      34 *     8    21     5
      6 *     172 *    44    93    35
      7 *     196 *    48   115    33
      8 *     142 *    34    85    23
      9 *      88 *    21    49    18
```

Par exemple, la valeur "10" au croisement de la ligne "4" et de la colonne 3 signifie que 10 occurrences[1] de mots affectés de la clé "4" (les marqueurs du temps) sont présentes dans les unités de contextes de la classe 3. En regardant le tableau ci-dessous, dans la même case, on trouve la valeur du χ^2 d'association signé, qui indique la significativité (4 ici). On en déduit, dans cet exemple, une distribution légèrement plus spécifique de marqueurs du temps dans cette classe. Cet aspect ne peut être interprété seul mais doit être coordonné avec les autres spécificités de cette classe. Par exemple, ici, cela va dans le même sens d'une représentation des projets d'avenir plus structurés en continuité avec la position actuelle du sujet (contrairement à la classe 1), en relation notamment avec le cursus scolaire.

```
  tableau des chi2 (signés) :

        * Classes *     1     2     3
```

[1] Plus précisément, ce calcul ne tient pas compte de la répétion éventuelle d'une même forme réduite dans une même unité de contexte élémentaire.

Clés	Poids	385	781	274
A	49	2	0	-1
B	3	-1	0	0
J	17	0	0	0
N	237	6	-2	0
V	151	0	-3	1
W	1	2	-1	0
Y	110	-3	0	6
0	27	0	0	0
1	80	0	0	0
2	62	0	0	0
3	45	0	0	0
4	26	0	-4	6
5	34	0	0	0
6	172	0	0	0
7	196	0	1	0
8	142	0	2	0
9	88	0	0	0

A propos de la notion de Chi2 signé voir le glossaire. Le signe indique le sens de la significativité (« plus » pour la présence et « moins » pour l'absence).

C3 : l'analyse factorielle des correspondances

L'analyse factorielle des correspondances (opération C3) est effectuée sur le tableau de données présenté avec le dictionnaire des formes réduites (C2_DICB.121). Les graphiques des plans factoriels ne figurent que dans le rapport d'analyse. Les résultats numériques sont enregistrés dans le fichiers D1_AFC.121. Voici la suite du rapport d'analyse :

```
---------------------------------
C3: A.F.C. du tableau C2_DICB.suf
---------------------------------

A.F.C. de :Dossier avenir:&&_0:C2_DICB.121

Effectif minimum d'un mot        :           8

Nombre d'uce minimum par classe  :          10
Nombre de lignes analysees       :          36
Nombre total de lignes           :          98
Nombre de colonnes analysees     :           3
```

Sont donnés d'abord les caractéristiques du tableau analysé : 36 lignes analysés (les mots pleins de fréquence supérieure à 8) et 3 colonnes (les trois classes). Le tableau est présenté par ailleurs et contient, à l'intersection d'une ligne et d'une colonne, le nombre d'U.C.E. de la classe contenant la mot.

```
*********************************************
* Num.* Valeur Propre * Pourcentage *  Cumul  *
*********************************************
*   1 *    0.39650813 *    58.56433 *  58.564 *
*   2 *    0.28053904 *    41.43567 * 100.000 *
*********************************************

Seuls les mots a valeur de cle ≥ 0 sont representes

Nombre total de mots retenus  :          95
Nombre de mots pleins retenus :          33
Nombre total de points        :          98

Représentation séparée car plus de 60 points
```

On trouve ensuite le tableau des valeurs propres et le pourcentage d'inertie extrait par chaque facteur. Viennent enfin les graphiques tous relatifs au premier plan factoriel. On notera que la position des points sur ces graphiques est définie non pas par les coordonnées mais par les corrélations (ou cosinus).

Le premier graphique contient la projection des classes (•01, •02, etc…) et des mots étoilés (*sexe_f, *sexe_m, etc…).

Le second graphique contient les mots analysés. Un calcul automatique s'adaptant au nombre de points à représenter élimine de la représentation les mots dont la valeur de clé est inférieure à la valeur indiquée (valeur 0 ici).

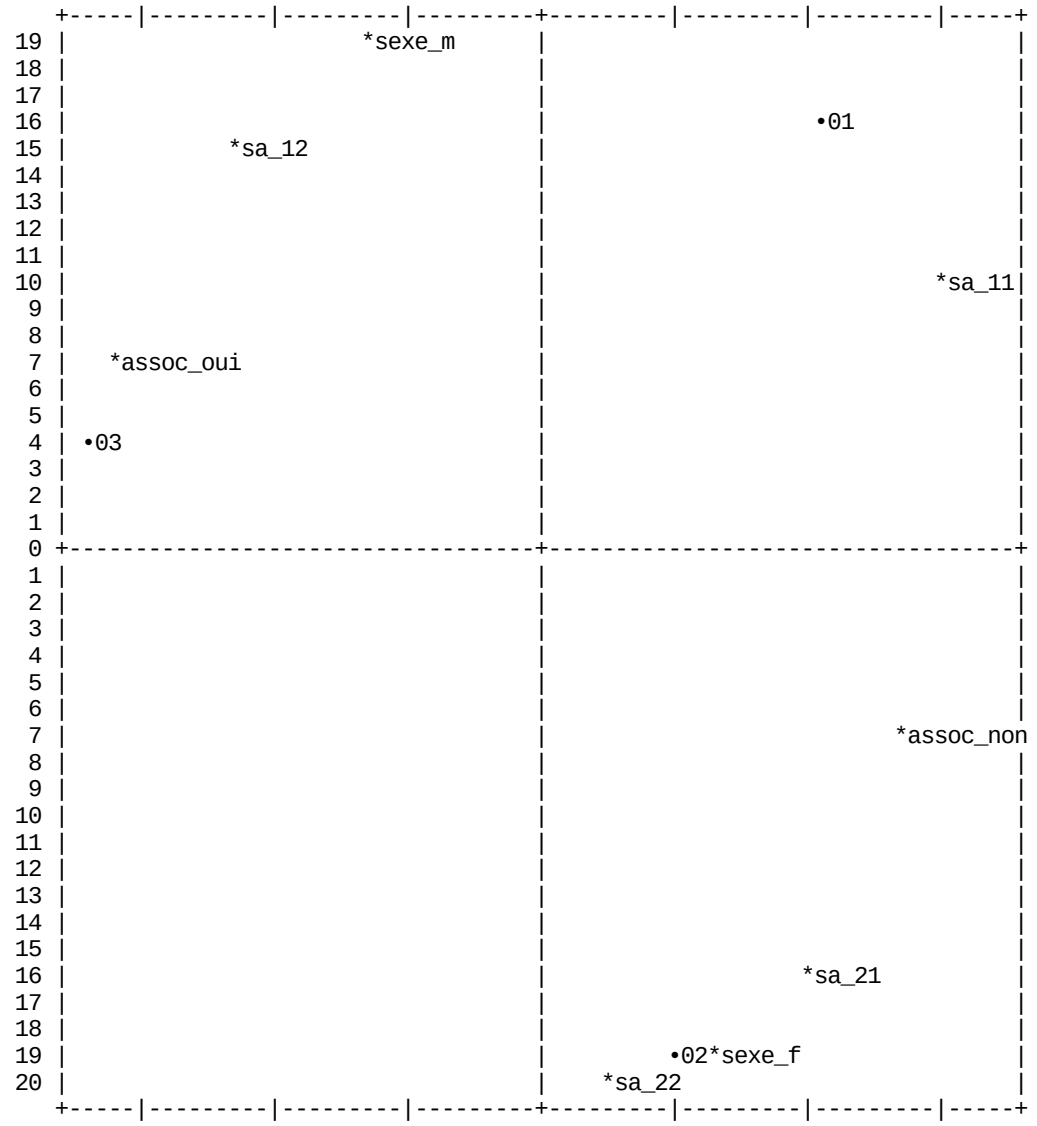
Le troisième graphique contient les mots en éléments supplémentaires (les mots outils dans l'option standard).

Ces trois graphiques sont superposables et sont relatifs au même premier plan factoriel.

Note : *l'aspect circulaire de la représentation vient du choix de la corrélation comme coordonnées des points du graphique et du nombre de colonnes analysées (l'espace vectoriel de référence est ici un espace à deux dimensions : il est égal au nombre de classes retenues pour le calcul des profils moins un ; pour vérifier, il suffit de lire le fichier C2_DICB.121 dans le cas d'une analyse standard. C'est d'ailleurs ce fichier qui est soumis à l'analyse factorielle des correspondances)*

Projection des colonnes et mots "*" sur le plan 1 2 (corrélations)

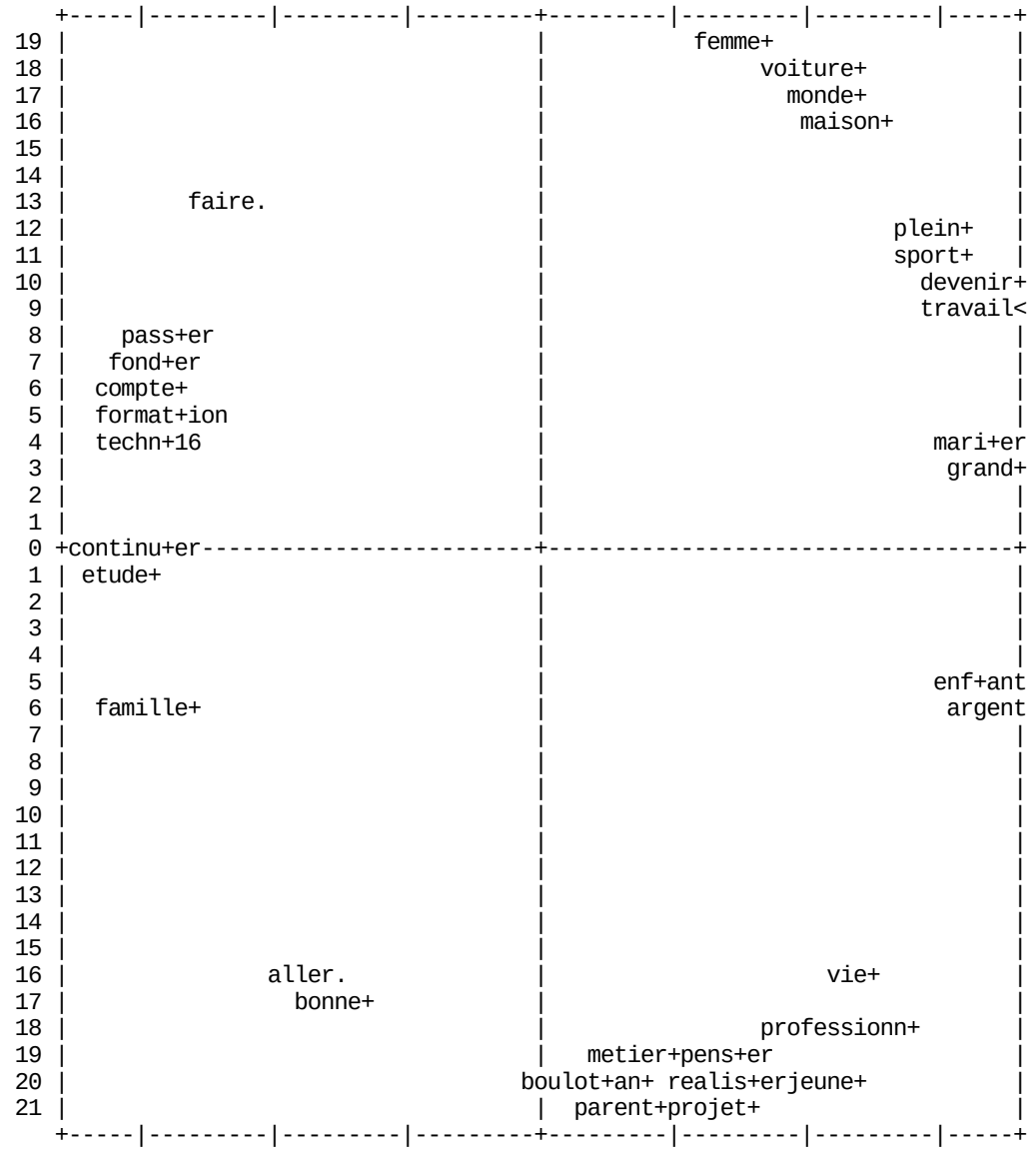
Axe horizontal : 1e facteur : V.P. =.3965 (58.56 % de l'inertie)
Axe vertical : 2e facteur : V.P. =.2805 (41.44 % de l'inertie)



Nombre de points recouverts 0 dont 0 superposes

Projection des mots analyses sur le plan 1 2 (correlations)

Axe horizontal : 1e facteur : V.P. =.3965 (58.56 % de l'inertie)
Axe vertical : 2e facteur : V.P. =.2805 (41.44 % de l'inertie)

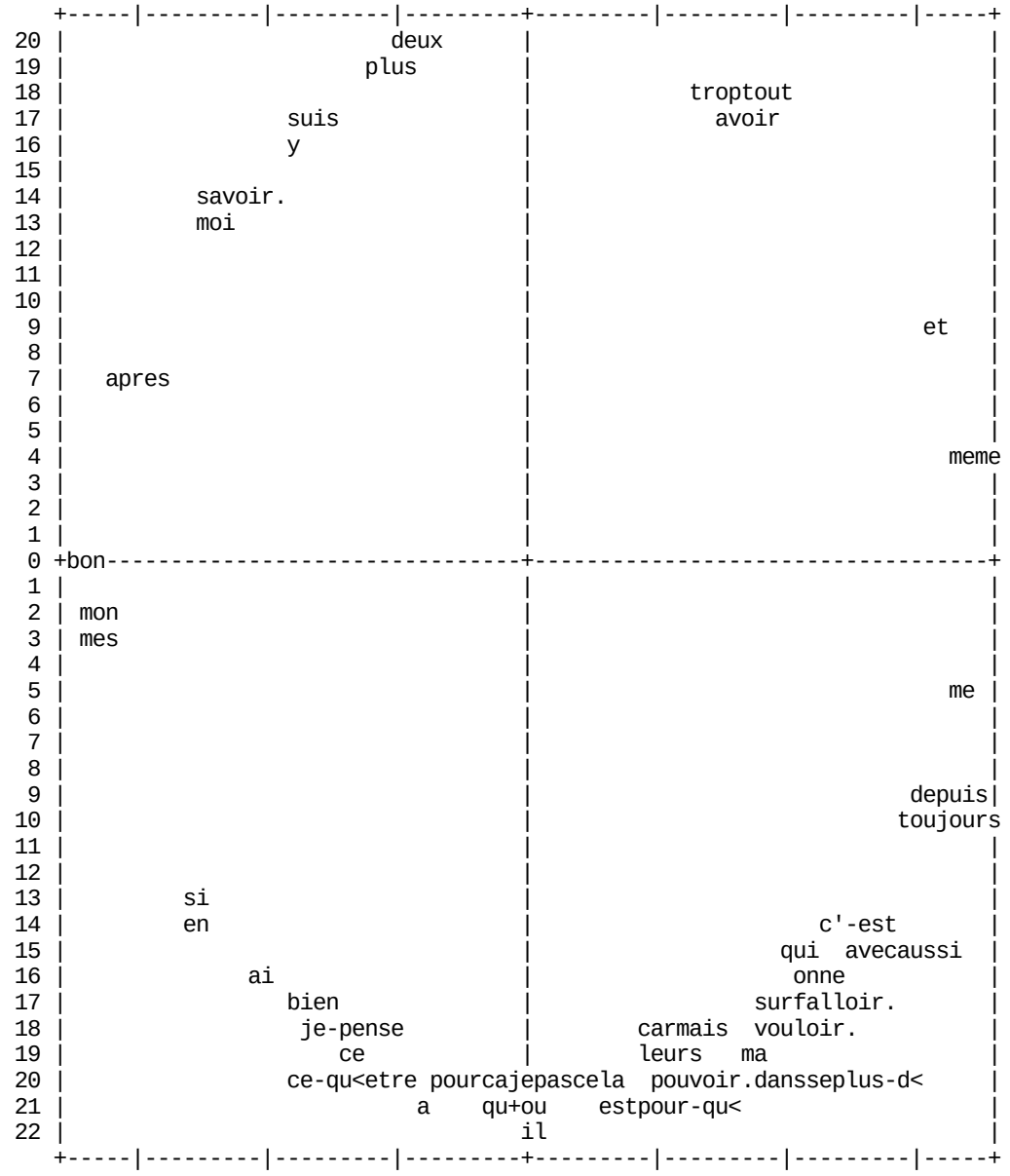


Nombre de points recouverts 0 dont 0 superposes

x y nom

```
 Projection des mots de type "r" sur le plan   1  2 (correlations)

 Axe horizontal : 1e facteur : V.P. =.3965 (   58.56 % de l'inertie)
 Axe vertical   : 2e facteur : V.P. =.2805 (   41.44 % de l'inertie)
```



Résultats de l'étape D

L'étape D prolonge l'étape C et fournit un certain nombre de calculs complémentaires à partir des classes. Ces calculs étant un peu plus longs ne sont généralement effectués qu'après s'être assuré de la bonne définition des classes.

D1 : Clés contextuelles et U.C.E. caractéristiques

Le rapport d'analyse présente un certain nombre de résultats "redondants", qui permet peu à peu de se faire une idée de la signification des classes obtenues. Le vocabulaire spécifique est aussi présenté en liste de la manière suivante. L'ordre des mots est celui de leur spécificité avec la classe (les *valeurs de clé : voir Glossaire*). Le nombre entre parenthèses correspond au nombre d'U.C.E. de la classe contenant le mot.

```
----------------------------------------
```

```
D1: Sélection de quelques mots par classe
-----------------------------------------

Valeur de clé minimum pour la sélection :   2

Vocabulaire spécifique de la classe  1 :
maison+(11),   belle+(7),   monde+(7),   voiture+(11),   femme+(6),   esper+er(5),
voyag+er(3), beau+(2), plein+(4), devenir+(6), permis(2), sport+(4), mari+er(8);

Vocabulaire spécifique de la classe  2 :
metier+(14), realis+er(11), jeune+(8), an+(14), rest+er(5), fait(6), pouvoir+(4),
projet+(13), vie+(10);

Vocabulaire spécifique de la classe  3 :
format+ion(9), techn+16(9), annee+(5), premier+(3), baccalaureat<(3), brevet+(2),
compte+(4), pass+er(4), continu+er(5), fond+er(3), rentr+er(3), independ+ant(3);

Mots outils spécifiques de la classe  1 :
et(17), sans(3), tout(8);

Mots outils spécifiques de la classe  2 :
beaucoup(4), plus-d<(4), pour-qu<(6), ils(4), leurs(6), se(7), ca(13), sont(6),
dix(4);

Mots outils spécifiques de la classe  3 :
apres(5), encore(3);

Mots étoilés spécifiques de la classe  1 :
*sa_11(5);

Mots étoilés spécifiques de la classe  2 :
*sa_21(15), *sa_22(19), *sexe_f(34);

Mots étoilés spécifiques de la classe  3 :
*assoc_oui(16), *sa_12(12), *sexe_m(12);

-------------------------------------------
D1: Sélection des mots et des uce par classe
-------------------------------------------
```

D1 : Distribution des formes d'origine par racine

Les formes réduites les plus spécifiques des classes sont présentées avec les distributions des formes d'origine dans la classe. Le nombre entre parenthèses indique le nombre d'occurrences de la forme dans les U.C.E. de la classe. Par exemple la forme réduite "maison+" est présentée dans le tableau ci dessus avec une fréquence de 11. Il s'agit en fait de 11 U.C.E. différentes de la classe 1. Dans la liste ci-dessous, cette forme réduite "maison+" se décompose en 11 occurrences sous la forme "maison" et 1 occurrence sous la forme "maisons", toujours dans la classe 1. Le nombre total d'occurrences de "maison+" étant 16 (voir B1_DICB). Le fait qu'il y ait 12 occurrences de "maison+" dans la classe 1 conduit donc à penser qu'une des 11 U.C.E. concernée contient deux fois "maison". Pour contrôler cela, il suffit de consulter le fichier D1_UCE.121 dont voici l'U.C.E. n° 90, qui a bien été classée dans la classe 1 avec un χ^2 *d'association* de 6 (voir glossaire) :

```
90    1    6  mais j' ai tout de-meme une idee de l' homme ideal #grand brun yeux bleus
      muscle, #beau quoi, je ne veux que deux #enfants, je veux une #grande #maison, pas
      ici et meme un autre #maison, je veux faire le tour du #monde, je veux une #grande
      #voiture,
```

Voici la distribution des formes d'origine dans chacune des classes. La clé "A" indique qu'il s'agit d'un mot spécifique de la classe 1; "B" pour la classe 2, etc.. Le coefficient 7 de A7 pour la forme réduite "maison+" est la valeur de clé qui indique un degré de significativité de cette spécificité (voir la liste de ces valeurs ci-dessus dans l'extrait du rapport concernant l'opération C2).

```
------------------------------
Formes associées au contexte A
------------------------------

A7 maison+ : maison(11), maisons(1);
A6 belle+ : belle(6), belles(1);
A6 monde+ : monde(7);
A6 voiture+ : voiture(8), voitures(3);
A5 femme+ : femme(5), femmes(4);
A5 esper+er : espere(6);
A4 voyag+er : voyager(3);
A3 beau+ : beau(1), beaux(1);
A3 plein+ : plein(5);
A3 devenir+ : devenir(11);
A3 permis : permis(2);
A3 sport+ : sport(4);
A3 mari+er : marier(8);

------------------------------
Formes associées au contexte B
------------------------------

B4 metier+ : metier(13), metiers(2);
B4 realis+er : realisables(3), realise(1), realisent(2), realiser(6), realiserai(1);
B4 jeune+ : jeune(1), jeunes(8);
B3 an+ : an(3), ans(14);
B3 rest+er : reste(1), rester(4);
B3 fait : fait(6);
B2 pouvoir+ : pouvoir(4);
B2 projet+ : projet(2), projets(13);
B2 vie+ : vie(11), vies(1);

------------------------------
Formes associées au contexte C
------------------------------

C7 format+ion : formation(10);
C7 techn+16 : technique(10);
C5 annee+ : annee(4), annees(1);
C4 premier+ : premier(2), premieres(1);
C4 baccalaureat< : baccalaureat(3);
C4 brevet+ : brevet(2);
C4 compte+ : compte(5);
C4 pass+er : passer(4);
C3 continu+er : continuer(6);
C2 fond+er : fonder(3);
C2 rentr+er : rentrer(4);
```

Dans cette dernière liste, la forme réduite «premier+ » apparaît 2 fois sous la forme d'origine « premier » et une fois sous la forme d'origine « premieres ».

D1: Tri des U.C.E. par classe

Le calcul des *Clés Contextuelles* (voir glossaire) permet d'ordonner les U.C.E. en fonction de la distribution des occurrences dans ces différentes clés et donc d'extraire une sélection des U.C.E. les plus représentatives. Le premier numéro indique le n° de l'U.C.I. ; Le second, le χ^2 d'association (ordonné ici de manière décroissante). Vingt U.C.E. par classe sont présentées dans le rapport d'analyse. Seules les cinq premières ont été retranscrites ici.

```
--------------------------
D1: Tri des uce par classe
--------------------------

Clé contextuelle sélectionnée : A

107   24   je veux #voyager, dans tout le #monde avoir #plein de #voitures et de
#femmes je veux etre riche et avoir #plein de #femmes.
```

6 18 mes projets seraient de #devenir veterinaire, d' avoir une #belle #maison a la montagne et une au bord de la mer, avoir une #femme un #enfant, une grosse #voiture de #sport et aussi une grosse moto.

7 14 aussi de #sport et aussi de competition. #devenir champion de tir a l' arc, tir a la carabine, #devenir champion du #monde de la chasse a courre, avec les meilleurs chiens, avoir un chenil.

96 13 dans un an je #pars de chez moi, je #passe mon #permis de #voiture et je #trouve un #travail je loue une #maison et je #passe une #belle vie et apres je me marie.

108 10 je voudrais etre #professeur de physique et pour cela il faut que je continue mes etudes, ensuite j' #espere me #marier, avoir une #voiture et une #maison.

Clé contextuelle sélectionnée : B

12 9 ce-que il faudrait ameliorer dans notre #societe pour-que les #jeunes puissent #realiser leurs #projets serait reduire le chomage, #donner plus-de #possibilite aux #jeunes dans les universite, initiation pour les #realiser, changer l' enseignement,

45 9 car la on ne peut pas imaginer tout ce-que on veut, mais pour cela il faut que je travaille beaucoup. dans notre #societe il faut que ca s' ameliore pour-que les #jeunes #realisent leurs #projets, il faut changer la pedagogie des professeurs.

60 9 certains de mes #projets sont #realisables maintenant, ne pas #rester toute ma #vie dans une cage a lapin, ne pas etre un mouton qui se fasse exploiter uniquement pour les autres etre libre meme si je ne gagne pas beaucoup d' argent,

87 9 j' ai ete a l' hopital et la j' ai compris que les #metiers dans le sanitaire c' etait important, essentiel pour la #societe, car ils savent les #vies et evitent les malheurs, c'-est pour cela que j' ai #decide d' etre infirmiere,

111 9 je-pense #pouvoir #realiser tout cela par-rapport a mon mariage.

Clé contextuelle sélectionnée : C

26 32 je voudrais #continuer mes #etudes, mais pas trop longtemps, je veux #passer mon #brevet, et ensuite #voir du #cote d' une #formation #technique, un truc de mecanicien, enfin des #etudes pas longues.

30 28 l' #annee prochaine, je voudrais bien #rentrer en seconde, et puis #continuer mes #etudes, #rentrer a la faculte, ou bien dans une #formation plus #technique, enfin le #minimum c'-est d' avoir mon #baccalaureat, c'-est mon #premier projet,

46 23 moi je #compte surtout avoir mon #independance, pour ca il-y-a pas de mystere, il faut que je travaille par-rapport aux #etudes, #passer mon #brevet, ensuite aller jusqu' a la terminale et preparer une #formation #technique, je-pense a l' informatique,

23 20 l' #annee prochaine j' aimerai #rentrer au lycee pour #continuer mes #etudes, #faire de l' electronique, apres si je #reussi ces #premieres #etudes je #compte #continuer encore deux ans dans l' informatique, etre technicien superieur.

Chapitre II
Les différents dictionnaires intégrés
(dans le dossier “ ALC ”)

Sommaire

Chapitre III
Les fichiers résultats

Sommaire

5	C2_PROFp.121	*Profil des classes d'U.C.E. obtenues*
6	C2_TUCE.121	*Tableau des "U.C.E. x Classes"*
7	C2_TUCI.121	*Tableau des "U.C.I. x Classes"*
8	D1_UCE.121	*Liste des U.C.E. reconstruites du corpus*
9	D1_UCE_A.121	*Liste des U.C.E. spécifiques du contexte A*
10	D2_SR	*Liste des "Segments Répétés"*
11	D2_SR.121	*Liste des "Segments Répétés" par classe*
12	D3_CAHa.121	*C.A.H. des formes spécifiques par classe*
13	D4_CONC.121	*Concordancier des formes spécifiques*
14	D5_UCE.121	*Corpus des U.C.E. accentuées.*

Chapitre IV
Le Glossaire

la terminologie utilisée, les informations techniques et quelques références bibliographiques

Ce glossaire regroupe un certain nombre de termes dont beaucoup sont spécifiques à la méthodologie "Alceste". Il arrive que des termes "communs" soient utilisés dans un sens spécifique ici. Tous ces termes sont décrits dans les paragraphes suivants organisés de manière à en faciliter une lecture linéaire et à se familiariser avec les principales techniques proposées.

Glossaire des termes avec le numéro du paragraphe :

Index des termes du glossaire avec le numéro du paragraphe

1. Le corpus, les Unités de Contexte Initiales, les lignes étoilées

On entendra par "corpus", tout ensemble de textes réunis par un analyste en vue d'une recherche particulière. Les différents " textes " ou " énoncés naturels " composant le corpus seront appelés " *Unités de Contexte Initiales* " ou U.C.I.. Pour faire une analyse, il faut au moins une U.C.I… Chaque U.C.I. doit être introduite par une *ligne étoilée.* Elle peut être décrite à l'aide de variables particulières les *variables exogènes* (ou *variables étoilées*). Une ligne étoilée est une ligne qui commence par une séquence de 4 étoiles et qui contient au moins un *mot étoilé* ou *mot hors corpus.* Exemple :

```
**** *Partie_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;
```

2. Les Segments de Texte Calibrés et les Unités de Contexte Elémentaires

• On appelle *segment de texte calibré (s.t.c.),* un segment de texte de longueur inférieur à 240 caractères et se terminant, si possible, par une ponctuation. L'utilisateur peut aussi définir des fins de segments de texte à l'aide des signes "£" et "$" (voir *ponctuation*).
• L'*unité de contexte élémentaire* est composée de un ou plusieurs *segments de texte calibrés* consécutifs. *L'U.C.E. est considérée comme l'unité statistique de base* par le logiciel. L'objectif principal de la méthodologie est justement d'obtenir un classement de ces U.C.E. en fonction de la distribution du vocabulaire.
• L'U.C.E. est définie dans l'opération B1. Deux paramètres permettent d'orienter sa construction : sa longueur en nombre de mots analysés et le type de ponctuation devant la terminer

Note. L 'U.C.E. est la plus petite unité statistique définissable sous Alceste et l'U.C.I. la plus grande. Le segment de texte calibré ne sert qu'à la définition des U.C.E.. Elle. est l'unité statistique utilisée pour les calculs des étapes C et D.

3. Le découpage du corpus en Unités de Contexte (U.C.)

On considère le corpus traité comme un ensemble de segments de texte non recouvrants et de petite dimension (de l'ordre de la phrase, de quelques lignes). Ces segments sont appelés *Unités de Contexte.* Toute *Unité de Contexte* est composée par un nombre entier d'*Unités de Contexte Elémentaires* (ou U.C.E.).

Dans une *classification double*, l'*unité de contexte analysée* est un segment de texte intermédiaire entre l'U.C.E. et l'U.C.I.. *Elle est définie par concaténation des U.C.E. successives d'une même U.C.I.* jusqu'à ce que le nombre de "mots" analysés de cette nouvelle unité soit supérieur au seuil fixé dans le plan d'analyse. Par exemple, une U.C.I. composée de 6 U.C.E. :

U.C.E 1	U.C.E 2	U.C.E 3	U.C.E 4	U.C.E 5	U.C.E 6

pourrait prendre l'aspect suivant après regroupement :

U.C 1 : uce1 + uce2	U.C 2 = uce3 + uce4 + uce5	U.C 3= uce6

4. Mots hors corpus ou mots étoilés

Lors de la retranscription du corpus, l'utilisateur peut introduire des informations exogènes caractérisant les unités de contexte initiales (par exemple : l'âge, le milieu, le sexe, s'il s'agit de réponses à une question ouverte). Ces informations sont introduites à l'aide de mots commençant par le symbole "*":

Exemple : `*age_1 *agriculteur *Le_Bateau_ivre *Jean-Paul`

Ces *mots étoilés* doivent être impérativement placés au début de l'U.C.I., sur une ligne d'au plus 256 caractères terminée par un "saut de ligne" (Return).
Leur propriété spécifique est d'être "transportable" à toutes les U.C.E. composant l'U.C.I. qu'ils identifient. Ces mots permettent de définir des classes d'U.C.E. a priori pour les tris croisés.

5. Les variables exogènes

On appelle variable exogène un ensemble de mots hors corpus associés aux modalités d'une même variable : par exemple, l'ensemble {*ag_1, *ag_2, *ag_3} pourra être considéré comme variable exogène, si chaque U.C.I. ne contient pas plus d'un élément de cet ensemble. Une variable exogène définit donc une partition a priori des U.C.I. (et donc des U.C.E.). Elle est identifiée par sa racine (par exemple : *ag_).

6. Les formes et leur réduction.

• Une *forme* d'origine est un ensemble de lettres séparé par un caractère délimiteur reconnu : Le retour à la ligne, l'espace ou l'un des signes suivants : *$ / . ? ! ; : , "* (voir Chapitre 1)
• Les *formes* d'origine renvoient globalement aux différentes formes prises par les mots d'un texte, aux aléas statistiques et orthographiques près (voir fichier A2_DICO)
• Une *forme réduite* est une forme transformée à l'aide d'un module d'ALCESTE ou par l'utilisateur à l'aide d'un éditeur de texte pour approcher le signifiant d'un mot (voir fichier B1_DICB & C2_DICB.121). Dans les listes de résultats, toute forme des dictionnaires de type DICB est appelée « forme réduite » voire simplement « mot ».
Les types de réduction sopérés dans Alceste :
1. Par reconnaissance de la racine et de la désinence (principalement pour les verbes irréguliers). *Les formes ainsi réduites se terminent par le symbole "."*
2. Par reconnaissance uniquement de la racine. *Les formes réduites ainsi traitées se terminent par le symbole "<".*
3. Par reconnaissance de la désinence seulement et dans le cas où plusieurs formes du corpus commencent par la même racine. Les *formes réduites* se terminent par le symbole "+".
• *Lemmatisation :* l'opération qui consiste à remplacer une forme textuelle par son lemme (la forme standardisée pour une entrée de dictionnaire de langue).
• *Comptage des formes, Occurrence, Hapax :* On appelle *occurrence* d'une forme sa position "matérielle" dans le texte par opposition à sa définition comme *" type "*. Pour une même *forme,* il y a généralement plusieurs *occurrences* dans un texte. Lorsqu'il n'y en a qu'une, cette forme est appelée « Hapax ».
Pour chaque *forme,* on peut calculer le nombre d'occurrences où elle apparaît. Ce calcul peut être effectué soit pour les *formes d'origine,* soit pour les *formes réduites.*

Note. Dans Alceste, le comptage des *formes d'origine* est en *nombre d'occurrences* alors que le comptage des *formes réduites* est en *nombre d'U.C.E.* contenant au moins une occurrence de la *forme réduite.*

7. Les couples de formes directement successives

Il s'agit des couples de formes directement successives dans le corpus : par exemple, dans "le chat est noir", les couples de formes directement successives sont (le, chat), (chat, est), (est, noir)).
Au sens "ALCESTE", on entendra par "couple de formes directement successives", une suite de deux occurrences successives d'une même U.C.E. après réduction et élimination des formes rares ou rejetées. Ces couples peuvent être analysés comme des formes simples. Ils entrent dans le calcul des segments répétés.

Note : dans le cas de données séquentielles, une analyse sur les couples plutôt que sur les items d'origine permet l'analyse du graphe de transitions en sous-graphes spécifiques de certaines classes de séquences.

8. Les segments composés de couples répétés.

Un segment répété est un segment de texte composé de plusieurs formes successives apparaissant au moins deux fois dans le corpus.
La procédure de calcul que nous utilisons consiste à rechercher d'abord les couples répétés.
A partir des couples répétés, on calcule les segments composés de couples répétés directement successifs. Tout segment de texte maximal composé de couples répétés est susceptible de recouvrir ce que nous appelons un *segment répété maximal*. Le logiciel calcule la fréquence de ces segments en tant qu'ils sont maximaux.

9. Clé catégorielle, Clé contextuelle, et Valeur de clé.

Deux sortes de clés sont considérées : les *clés catégorielles* et les *clés contextuelles* selon qu'elles sont affectées a priori par le logiciel ou l'utilisateur, ou bien, selon qu'elles sont affectées en fonction des résultats des classifications.

• La clé catégorielle identifie à l'aide d'une lettre minuscule ou majuscule ou encore d'un chiffre des catégories de mots reconnus a priori… On trouvera la liste des catégories dans le fichier ALC_CLE (dans le dossier "ALC"). Ce fichier est modifiable par l'utilisateur. Elle est imprimée dans le rapport d'analyse.

• La clé contextuelle est définie par une lettre identifiant la classe de la C.D.H. où la forme est plus particulièrement présente… par exemple, la lettre A indiquera un lien du mot ainsi marqué avec la classe 1, etc…
Cette clé contextuelle est suivie d'une valeur comprise entre 0 et 9 indiquant la force du lien du mot avec la classe (en fonction d'un χ^2 *d'association*)

10. Les éléments supplémentaires ou illustratifs

Les mots en éléments supplémentaires sont des mots n'entrant pas dans les calculs effectués pour obtenir la classification des U.C. (opération B3) mais ils apparaissent dans le descriptif du profil de ces classes et plus généralement dans tous les calculs des opérations des étapes C et D.

On distingue deux sortes d'éléments supplémentaires :

• les mots hors corpus (ou " mots étoilés " ou mot de type " s "), définis lors de la retranscription du corpus pour décrire certaines caractéristiques des U.C.I..

• les mots du corpus (ou mot de type " r "), que l'on ne désire pas analyser, mais dont on veut conserver la trace dans les résultats généralement repérés par leur clé d'origine : les mots outils, par exemple.

La gestion des mots supplémentaires est effectuée à l'aide d'un code (9ème caractère des dictionnaires) appelé l'indicateur d'analyse.
Dans les Profils des classes, les mots supplémentaires apparaissent en fin de liste avec la marque "*".

La gestion générale des mots analysés, rejetés et illustratifs est effectuée sous Alceste à l'aide d'un système de *clés catégorielles.*

11. Les indicateurs d'analyse

Un indicateur d'analyse permet de savoir si une forme est analysable, illustrative ou rejetée. Cet indicateur est codé en colonne 9 (i.e. : colonne 9 = 9e caractère de la ligne) des dictionnaires DICB. Voici la liste des modalités de cet indicateur dans l'ordre de leur apparition dans DICB :

" " : *forme analysable* reconnue (marquée d'une *clé* catégorielle).
"a" : *forme analysable* non reconnue.
"r" : *forme illustrative* (par exemple certaines catégories de mots faisant partie du corpus mais non conservés dans les analyses de données).
"s" : forme caractérisant toutes les *U.C.E.* de l'*U.C.I.* qui la contient (par exemple les *mots hors corpus*).

Les *indicateurs d'analyse* sont modifiables par l'utilisateur dans A3_DICB (avec un éditeur de texte). Ils sont gérés de deux manières complémentaires :
a) Les *clés catégorielles* (voir les codes *des clés* dans le fichier ALC_CLE).
b) La fréquence de la forme réduite. Les formes de trop faible fréquence sont rejetés de l'analyse. Notons qu'un calcul automatique des seuils peut être effectué par l'opération B1 (cas standard).
Les choix effectues sont retranscrits dans le rapport d'analyse (opération B1).

12. Chi2 d'association, Profil d'une classe, Clé contextuelle, Spécificité, Tris croisés.

Soit n, le nombre d'U.C.E. retenues dans l'analyse. Notons :
n1, le nombre d'U.C.E. de la classe considérée (d'une C.D.H. par exemple);
n2, le nombre d'U.C.E. où le mot est présent ;
n12, le nombre d'U.C.E. de la classe où le mot est présent ;

		Forme réduite choisie		
		Présent	**Absent**	
Classe X sélectionnée	**Présent**	$\mathbf{n_{12}}$	•	$\mathbf{n_1}$
Autres classes	**Absent**	•	•	•
		$\mathbf{n_2}$		**n**

On compare alors $\mathbf{(n_{12})}$ à $\mathbf{(n_1.n_2/n)}$ à l'aide d'un χ^2 calculé à partir d'un tableau à 4 cases.
Ce χ^2 est ensuite affecté du signe de la différence $\mathbf{n_{12}.(n_1.n_2/n)}$ pour identifier le sens de la corrélation. Il est appelé alors *« Chi2 d'association »*.
Ce calcul est utilisé dans plusieurs procédures ; notamment pour rechercher le vocabulaire spécifique de chaque classe *(Profil des classes)* et pour marquer chaque forme à l'aide d'une *Clé contextuelle.*

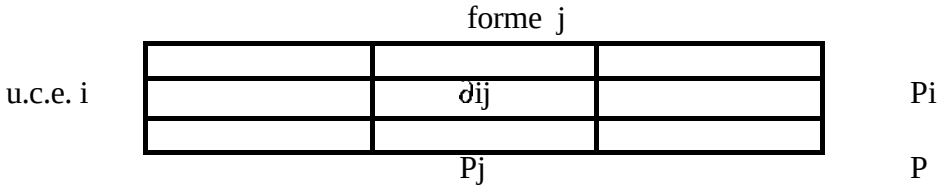
Note sur la notion de *Spécificité.* Lorsque les classes sont définies a priori, soit à l'aide d'une *variable exogène*, soit à l'aide d'un mot du texte, le calcul des profils des classes est appelé en lexicométrie *«calcul des spécificités »*. On peut également utiliser ce terme dans Alceste en retenant toutefois que le calcul des spécificités est effectué à partir des chi2 d'association et non à partir de la loi hypergéométrique (Lebart & Salem, Lafon). On désigne également ces calculs sous le nom de *Tris Croisés* car ces calculs sont effectués à partir de tableaux croisant les modalités de la variable exogène choisie avec l'ensemble du vocabulaire.

13. Coefficient d'appartenance d'une U.C.E. à une classe

Le coefficient d'appartenance d'une U.C.E. à sa classe n'est généralement calculé que sur les U.C.E. classées (voir opération C2). Le calcul est semblable à celui du χ^2 d'association, l'unité statistique n'étant plus l'U.C.E. mais le *trait unaire* d'une forme retenue dans l'analyse.

14. Tableau de données, Poids, Marge, Trait unaire

La structure du corpus qui est retenue est celle modélisée par un tableau logique à double entrée ayant, en lignes, les U.C.E. et, en colonnes, les formes réduites analysées. A l'intersection de la ligne i et de la colonne j, la valeur ∂ij est égale à 1 si la forme j appartient à l'U.C.E. i et est égale à 0 sinon.



Ce tableau modélise le corpus pour Alceste. Il est le tableau de base pour tous les calculs. Plusieurs autres tableaux sont cependant considérés :

Les tableaux des unités de contexte de longueur minimum k : Ils sont construits à partir du tableau présenté en concaténant les U.C.E. successives d'une même U.C.I. jusqu'à ce que le nombre de " un " de la ligne associée soit ≥ k (de codage logique).

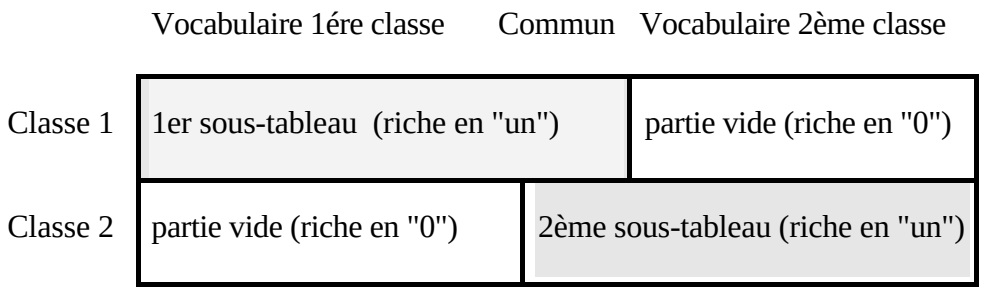
Note sur les autres tableaux statistiques de Alceste. Les autres tableaux sont calculés à partir du tableau logique de base. On note des tableaux de fréquences en nombre d'u.c.e. comme C2_DICB.xxx ; ou des tableaux qui comptabilisent les *traits unaires*, si on appelle « *trait unaire* », la marque d'une simple « présence » du mot dans le contexte considéré. Ce trait se différencie de la simple occurrence, par le fait qu'un même mot ne compte que pour une unité, quelle que soit sa fréquence dans l'unité de contexte considérée .

15. La Classification Descendante Hiérarchique (C.D.H.)

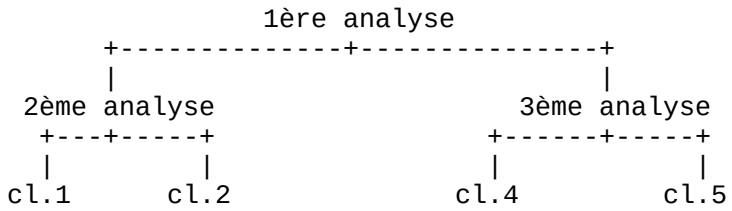
Il s'agit d'une technique descriptive d'Analyse des Données applicable à des tableaux présence/absence (croisant ici le vocabulaire et les unités de contexte : la valeur 1 signifie la présence du mot dans l'unité ; la valeur zéro son absence).

La technique est itérative : Au premier pas, le tableau d'origine est décomposé comme ci-dessous. Une fois obtenu, chaque sous-tableau est analysé selon la même procédure. La succession des analyses définit ainsi un arbre comme le suggère le schéma ci-dessous :

Analyse d'un sous-tableau



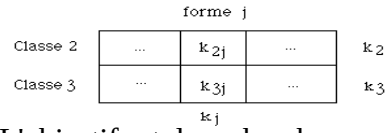
Succession des analyses :



Dans l'exemple ci-dessus, trois itérations conduisent à la définition de quatre classes terminales. A chaque itération, l'analyse porte sur le plus grand des sous-tableaux restant à traiter (en nombre de lignes).

Précision sur la méthode de classification utilisée :

Il s'agit d'une méthode de classification descendante hiérarchique. Elle a été mise au point pour traiter des tableaux logiques (codage "0" ou "1") de grandes dimensions (10 000 lignes par 1 500 colonnes maximum dans la version 4.0) mais de faible effectif.
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. La procédure s'arrête lorsque le nombre d'itérations demandé est épuisé (15 maximum).
La méthode de partitionnement d'une classe en deux repose sur le critère suivant : Considérons une partition candidate quelconque en deux classes et le tableau des marges associé ; ce tableau comprend autant de colonnes que de *formes analysées*, avec uniquement deux lignes, une pour chaque classe de la partition candidate avec, par exemple, à l'intersection de la première ligne et de la jième colonne, le nombre k2j d'*U.C.* de la classe contenant la jième *forme* identifiée :



L'objectif est de rechercher, parmi toutes les partitions en deux classes, celle maximisant le χ^2 de ce tableau (qui est donc le critère choisi).

16. Problème de la stabilité des classes obtenues par la C.D.H..

Dans la procédure standard d'analyse, une première appréciation de la stabilité des résultats est basée sur le calcul suivant. On effectue deux classifications sur des unités de contexte de grandeur légèrement différentes (voir le Rapport d'Analyse et le Paramétrage d'un plan d'analyse).
Chaque U.C. de l'une ou l'autre analyse est composée d'un nombre entier d'U.C.E. On peut donc considérer les deux classifications comme des classifications sur les U.C.E..

La comparaison entre les deux hiérarchies s'effectue ainsi :
• On construit un tableau de cooccurrences entre les classes de la première classification et les classes de la seconde classification qu'elles soient terminales ou non. L'unité distribuée est l'U.C.E.. Autrement dit, les U.C. utilisées pour les deux classifications sont remplacées par les U.C.E. qui les composent. Ainsi chacune des deux classifications peut être considérée comme une classification sur les U.C.E.
• On calcule le χ^2 d'association entre chaque couple de classes en correspondance.
• Puis, on retient l'ensemble des couples dont le χ^2 d'association est maximum en ligne et en colonne.

Parmi l'ensemble des couples retenus, on choisit ensuite ceux associés, au moins pour l'une des classifications, à une même partition. Les classes retenues sont les classes-

intersections entre les 2 classes des couples distingués. Elles recouvrent généralement plus de 50 % des U.C.E. du corpus. C'est sur cet ensemble plus stable que le profil des classes est calculé (voir opération C1 dans le Rapport d'Analyse).

Note : Si le pourcentage des U.C.E. classées est inférieur à 50 %, il vaut mieux recommencer l'analyse en modifiant la longueur des U.C. (voir les paramètres de l'opération B2)

17. L'analyse factorielle des correspondances (A.F.C.)

J.P. Benzécri, le créateur de cette méthode d'analyse des données, écrivait, il y a plus de 20 ans : *" C'est principalement en vue de l'étude des langues que nous nous sommes engagés dans l'analyse factorielle des correspondances"* [A.D/tome 2/p 327] ou encore :*"L'analyse des correspondances a été initialement proposée comme une méthode inductive d'analyse des données linguistiques"* [HPAD p102]

Cette méthode est à la base de la classification descendante hiérarchique. Elle est utilisée également pour décrire à l'aide de quelques facteurs la structure de tableau de cooccurrences. Les tableaux soumis à l'A.F.C. dans Alceste croisent le vocabulaire retenu avec des "classes" d'U.C.E. (définies dans une opération précédente : classes d'une c.d.h. ou classes dérivées d'une variable exogène). L'objectif est de donner une représentation spatiale simplifiée des relations entre classes.

18. La classification ascendante hiérarchique (c.a.h.)

Cette technique est décrite dans les ouvrages de Benzécri (1972). Elle est utilisée dans la méthodologie Alceste qu'en deux occasions :
a) en complément pour une aide à la représentation des liens entre classes (opération C2 dans le cas de tris croisés) ;
b) pour présenter des relations locales entre formes d'un même contexte (opération D3).
Pour le détail de cette procédure nous renvoyons à la bibliographie.
Le critère utilisé dans Alceste est le rapport entre variance intra-classe et variance interclasse, la variance étant calculée avec la métrique du $\chi 2$.

Quelques références bibliographiques

Beaudouin, Valérie, Lahlou Saadi, *L'Analyse lexicale : outil d'exploration des représentations,* Cahier de Recherche n° 48, CREDOC, Paris, 1993.

Benzécri, Jean-Paul, *L'Analyse des Données* (tome 1 et 2), DUNOD, Paris, 1973.

Benzécri, Jean-Paul, *Pratique de l'Analyse des Données : linguistique et lexicologie* , DUNOD, Paris, 1981

Brunet Etienne, Voyage autour des mots, *Dictionnairique et Lexicolographie,* vol 2, 167-184, Didier Erudition, Nancy, 1992.

Cibois, Philippe, *L'analyse factorielle, P.U.F.,* 1983

Guiraud, Pierre, *Problèmes et méthodes de la statistique linguistique*, P.U.F., Paris, 1960.

Lafon, Pierre, Salem, André , L'inventaire des segments répétés d'un texte, *Mots*, 1983, 6,161-177.

Lahlou Saadi, L'analyse lexicale : une technique nouvelle illustrée par un exemple qui ouvre l'appétit, *Variances,* 1994, n° 3, 13-24

Le Roux, Dominique, Blot, I. *L'utilisation du logiciel Alceste au département GRETS,* HN-52/92/067, EDF-DER, Clamart, 1993.

Lebart, Ludovic & Fénelon, Jean-Pierre, *Statistique et informatique appliquée*, DUNOD, Paris, 1971.

Lebart, Ludovic & Salem, André, *Statistique textuelle* DUNOD, Paris, 1994.

Looze (de) M.-A., Roy A., Coronni R., Reinert M., Jouve O., Two measures for identifying the perception of risk associated with the introduction of transgenic plants, *Scientometrics,* , Elsevier Science, 1999, vol 44, n° 3, 401-426.

Marchand Pascal, *L'analyse du discours assistée par ordinateur,* Armand Colin, Paris, 1998.

Muller Charles, *Principes et méthodes de statistique lexicale,* Hachette, Paris, 1977.

Noël-Jorand M-C,Dassa D., Giudicelli S. "A new approach to discourse analysis in psychiatry, applied to a schizophrenic patient's speech", *Schizophrenia Research*, Elsevier Science, 1997, 25, 183-198.

Reinert Max

Une méthode de classification descendante hiérarchique : Application à l'analyse lexicale par contexte. *Cahiers de l'Analyse des Données,* 1983, 3,187-198.

- Un logiciel d'analyse lexicale (ALCESTE). *Cahiers Analyse des Données*, 1987, 4, 471-484.

- Une méthode d'analyse des données textuelles et une application : Aurelia de G. de Nerval. *Bulletin de Méthodogie Sociologique,* IRESCO, Paris,1990, 26, 24-54.

- Les mondes lexicaux et leur logique à travers l'analyse statistique d'un corpus de récits de cauchemars, *Langage et Société*, 1993, 66, 5-39

- Quelques aspects du choix des unités d'analyse et de leur contrôle dans la méthode “Alceste”, in Bolasco, Lebart, Salem (Eds), *JADT 1995* (*Analisi Statistica dei Dati Testali),* CISU, Roma, 1993, p. 27-34

- Quelques problèmes méthodologiques posés par l'analyse de tableaux "Enoncés x Vocabulaire", in Bécue, Lebart, Rajadell (Eds), *JADT 1993 (Journées Internationales d'Analyse des Données Textuelles)*, Montpellier, Telecom Paris 93 S 003, 1993, p 539-549

- Les "Mondes lexicaux" des six numéros de la revue "Le Surréalisme au Service de la Révolution" , *Mélusine N° XVI*, Editions L'Age d'Homme, Lausanne, 1997, p 270-302.

- Quelques interrogations à propos de l'"objet" d'une analyse de discours de type statistique et de la réponse "Alceste", *Langage & Société,* décembre 1999, n° 90, p. 57-79

- “ La tresse du sens et la méthode Alceste ”, *5èmes Journées Internationales d'Analyse Statistique des Données Textuelles*, Lausanne, 9-11 mars 2000.

- “ Alceste, une méthode statistique et sémiotique d’analyse de discours ; Application aux “ Rêveries du promeneur solitaire ””, *La Revue Française de Psychiatrie et de Psychologie Médicale,* 2001,V, n°49, p 32-36

- “ Approche statistique et problème du sens dans une enquête ouverte ”, *Journal de la Société Française de Statistique*, 2001, tome 42, n°4, p 59-71

Salem André, *Pratique des segments répétés* , klincksieck, Paris, 1987.

Wald Paul, Classes d'énoncés, dimensions modales et catégories sociales dans ALCESTE, *Utinam,* 1999-1/2, p. 303-24

Yvon François, *L'analyse lexicale appliquée a des données d'enquète : état des lieux,* Cahier de Recherche du CREDOC, 1990, n°5.