CAHIER DE RECHERCHE

SEPTEMBRE 1999 ■ N° 131

IMPACT DE LA LEMMATISATION SUR LA ROBUSTESSE DES TYPOLOGIES LEXICALES

Rôle des seuils de sélection des mots analysés

Anne-Delphine Brousseau

Département "Prospective de la Consommation"

CRÉDOC
L'ENTREPRISE DE RECHERCHE

Centre de Recherche pour l'Étude et l'Observation des Conditions de Vie

CRÉDOC

Impact de la lemmatisation sur la robustesse des typologies lexicales

Rôle des seuils de sélection des mots analysés

Anne-Delphine BROUSSEAU

Département Prospective de la Consommation

SEPTEMBRE 1999

142, rue du Chevaleret
75013 - PARIS

SOMMAIRE

INTRODUCTION

Depuis plusieurs années, le CRÉDOC effectue des travaux de recherche sur l'analyse lexicale[1] : son utilisation, ses extensions, ses performances méthodologiques et ses limites. Ces travaux se sont essentiellement inscrits dans le cadre des développements du logiciel *Alceste*, partant d'une collaboration soutenue avec son développeur, Max Reinert. Ils se poursuivent encore aujourd'hui dans un objectif de test, de compréhension et d'amélioration du logiciel.

Plusieurs types de travaux ont déjà été menés par le CRÉDOC : d'une part, des réflexions visant à améliorer, à enrichir les analyses via *Alceste*, mais aussi à tester ses limites et, d'autre part, des études comparatives des performances d'*Alceste* par rapport à d'autres outils de lexicométrie tels que les logiciels *SPAD.T*, *Tropes*, *Leximappe*, ou encore par rapport à la post-codification traditionnelle. Pour ces travaux exploratoires, divers matériaux ont été utilisés, aussi bien des réponses à des questions ouvertes, des corpus d'entretiens, des textes littéraires, des articles, des structures narratives, que des corpus de réponses en anglais et en français.

Les deux derniers travaux de recherche effectués dans ce domaine par le CRÉDOC ont porté sur des comparaisons d'analyses effectuées à partir de différents logiciels qui se sont souvent révélées être à l'avantage du logiciel *Alceste*. Il nous a donc paru intéressant cette année de changer d'orientation en approfondissant plus directement la connaissance des analyses d'*Alceste*, notamment en portant plus spécifiquement nos réflexions sur la robustesse de cette

[1] Cf. Bibliographie en fin de rapport.

méthodologie, question essentielle à l'utilisation et l'interprétation des résultats. Une étape critique est celle du choix des paramètres de transformation du corpus à étudier. Parmi ceux-ci, la lemmatisation et la fixation du seuil minimal de la fréquence des mots choisis pour l'analyse sont des originalités de l'approche d'*Alceste*. Aussi, allons-nous traiter dans ce travail de recherche la question de l'impact du degré de lemmatisation sur la stabilité des typologies lexicales.

Il est vrai que l'utilisation de la lemmatisation dans les analyses lexicométriques a été popularisée par le logiciel *Alceste*, mais un débat méthodologique récurrent en lexicométrie partage encore aujourd'hui les partisans de la lemmatisation et ceux qui préfèrent analyser directement les textes non lemmatisés, en distinguant les pluriels des singuliers, les formes conjuguées des verbes et le féminin du masculin.

L'avantage souvent reconnu de la lemmatisation est la plus grande stabilité des analyses et notamment l'évitement de corrélations entre formes graphiques difficilement interprétables. Par ailleurs, la lemmatisation augmente la fréquence des formes réellement analysées. De ce point de vue, elle combine son action à celle du choix des seuils minima des fréquences des mots analysés.

Dans ce rapport, nous allons analyser empiriquement l'impact du choix de ces paramètres à partir d'exemples concrets d'applications. Deux exemples spécifiques vont, en effet, nous aider tout au long de ce rapport à valider nos hypothèses : il s'agit de deux questions ouvertes extraites de l'enquête Consommation 1998 du CRÉDOC[1], la première concerne la perception du bonheur (*« Pour vous, qu'est-ce qu'être heureux ? »),* la seconde porte sur les préférences des Français dans divers domaines. Les réponses à ces questions constituent des corpus de mots que nous avons traités avec *Alceste* et qui ont l'avantage d'être très différents l'un de l'autre.

[1] Cf. Brousseau A.D., Volatier J.-L., (1999).- « Le consommateur français en 1998 – Une typologie des préférences », *CRÉDOC, Cahier de recherche* n°130, juin.

Ce rapport met en évidence deux étapes de recherche :

- La première étape s'interroge sur **le rôle de la lemmatisation** dans le traitement des corpus sous *Alceste*. Quelle option choisir pour analyser un ensemble de réponses individuelles à des questions ouvertes : la lemmatisation ou la non lemmatisation ? Pour tenter d'éclairer ce choix, nous avons comparé, pour un même corpus, les résultats des deux analyses, l'une sur le texte lemmatisé, l'autre à partir du corpus non lemmatisé.

- La seconde partie de ce rapport porte sur **l'impact du degré de sélection des mots sur la stabilité des typologies**. Cette phase part du postulat de base que le choix de la lemmatisation est celui à retenir ; une fois cette hypothèse établie, quatre degrés différents de sélection ont été testés. Nous en présentons ici les résultats, en précisant les similitudes, ainsi que les limites d'interprétation des classifications obtenues.

I. LEMMATISATION OU NON DANS LES ANALYSES LEXICALES

La première partie de notre recherche consiste à nous interroger sur le rôle de la lemmatisation dans le traitement des corpus sous *Alceste*. Le débat entre les partisans de la lemmatisation et ceux qui préfèrent les formes textuelles brutes ne date pas d'hier. Il semble donc bon de revenir un peu sur ce débat. Nous nous proposons ensuite de comparer, pour un même corpus, les résultats obtenus à partir de deux analyses, l'une sur le texte lemmatisé, l'autre à partir du corpus non lemmatisé, et de tester la stabilité des classifications obtenues.

Mais interrogeons-nous tout d'abord sur la signification de la « lemmatisation ».

I.1. UNE DEFINITION DE LA LEMMATISATION

La première étape qu'effectue *Alceste* dans le traitement du corpus d'une question ouverte est la reconnaissance du corpus et le calcul des dictionnaires. Dans un second temps, se présente la possibilité de lemmatiser les formes ou expressions employées par les enquêtés.

Lemmatiser le vocabulaire d'un texte revient à ramener les formes verbales à l'infinitif, les substantifs au singulier, les adjectifs au masculin singulier, c'est-à-dire à rassembler sous une même « forme réduite » des mots qui peuvent avoir une représentation différente dans le corpus et qui, avec une probabilité non négligeable, indiquent un même aspect de la « référence ».

Autrement dit, en privilégiant le point de vue lexicographique, la lemmatisation, qui est avant tout un traitement quantitatif sur un corpus, soumet les unités graphiques à un ensemble de règles d'identification qui permettent de regrouper sous une même unité les formes graphiques[1] qui correspondent aux différentes flexions d'un même lemme. Cette réduction a pour objectif d'améliorer l'analyse statistique et notamment le classement des *unités de contexte élémentaires* (uce).

[1] Les formes graphiques sont définies comme suites de caractères comprises entre deux caractères délimiteurs. Elles se distinguent donc des mots car un même mot peut prendre généralement plusieurs formes en fonction des marques du pluriel, du féminin et des désinences de conjugaison.

I.2. LE DEBAT SUR LA LEMMATISATION

Est-il judicieux de lemmatiser ou non un texte avant d'en faire une analyse lexicale ? À cette interrogation, deux écoles donnent leurs arguments.

Pour les partisans de la lemmatisation, cette technique a l'avantage de réduire la diversité du vocabulaire pour mieux mettre en évidence les proximités sémantiques : la différence entre l'emploi du nom ou de l'adjectif est pour eux, notamment dans les réponses aux questions ouvertes, négligeable au regard de la proximité des concepts sous-jacents. Le mérite de la lemmatisation réside dans la grille de lecture unifiée qu'elle permet d'obtenir. Les réponses, ou parties de réponses, renvoient à des représentations mentales.

La lemmatisation gomme ainsi les différences dues à une maîtrise du langage corrélée à l'appartenance sociale des individus. Ainsi, pour reprendre l'exemple donné par L. Lebart, quelle que soit la formulation de la réponse, « financières » et « finances » renvoient à la même idée de contrainte budgétaire[1].

Au contraire, les opposants à cette technique voient dans la richesse du vocabulaire employé et dans l'utilisation de formes grammaticales spécifiques une connotation sociale.

Le choix entre lemmatisation et non lemmatisation est, en réalité, motivé par le fait que la lexicométrie n'a pas la faculté de traiter la phrase dans sa composante syntaxique. Elle est réduite à ne prendre en compte que le vocabulaire.

Or, la taille du vocabulaire est souvent très importante. Il apparaît donc indispensable de la réduire pour faciliter les calculs matriciels et **rendre l'analyse plus stable**. Cette étape est possible avec la lemmatisation.

Mais d'un autre côté, sans lemmatisation, on prend en compte les mots d'un texte en tant que tels, bruts, sans aucune intervention. L'avantage principal de ce type d'analyses textuelles

[1] L. Lebart in *Analyse statistique des données textuelles*, Dunod, Paris, 1988.

réside dans la prise en compte d'*« empreintes textuelles » présentes dans les textes, liées à l'emploi de facto de termes donnés, qui révèlent de forts éléments d'homogénéité dans le langage concrètement employé.*

Depuis quelques années, les techniques de lemmatisation se sont améliorées : les erreurs de rapprochement de formes sont moins fréquentes aujourd'hui, même si des ajustements sont nécessaires a posteriori. Dans les premières versions des logiciels d'analyse textuelle, il n'y avait pas de lemmatiseur, les méthodes pour réduire la taille du vocabulaire étaient très archaïques et les résultats étaient affinés en manuel. Par ailleurs, certains travaux antérieurs[1] ont montré que le matériau textuel constitué par les réponses à une question ouverte est si redondant que des erreurs de lemmatisation n'ont finalement qu'une incidence minime. Cependant, le fait que la lemmatisation soit correcte accroît en tous cas la lisibilité des résultats.

Cette opération, pour être exacte, nécessiterait en fait une analyse syntaxique approfondie du texte, de la phrase pour « désambiguïser » les homographes et pouvoir rattacher le terme à sa forme lemmatisée. La méthode de lemmatisation *Alceste* et l'analyseur syntaxique ont apporté de remarquables améliorations dans ce domaine. Aujourd'hui, grâce à des analyseurs automatiques combinés avec des dictionnaires de grande dimension, la lemmatisation est considérée comme efficace à 95% pour rattacher une forme graphique à son lemme.

[1] Cf. Beaudouin V., Lahlou S., (1993).- « L'analyse lexicale : outil d'exploration des représentations », *CRÉDOC Cahier de recherche* n°48, septembre.

I.3. INCIDENCE DE LA LEMMATISATION SUR L'ANALYSE LEXICALE

➲ Les corpus étudiés

Deux corpus ont été analysés dans le cadre de ce travail d'observation des différences entre les deux méthodes (avec lemmatisation, sans lemmatisation). Ils sont tous deux extraits de l'enquête du CREDOC sur la Consommation des Français. Cette étude, réalisée en octobre 1998 auprès de 1006 individus français de 15 ans et plus, comprend notamment deux questions ouvertes :

- l'une sur la conception qu'ont les Français du **bonheur**, avec l'interrogation *« Si je vous dis être heureux, à quoi pensez-vous ? »*,
- l'autre sur les **préférences** des Français dans des domaines aussi variés que la culture, les loisirs, la consommation, la vie pratique, ... Pour cela, chaque enquêté a répondu aux questions suivantes :

Nous allons maintenant parler de vos préférences, de ce que vous aimez le plus.
(une seule réponse donnée spontanément)

Quel est votre plat préféré (en dehors des desserts) .....
Quel est votre dessert préféré .....
Citez une marque de vêtement que vous aimez .....
Quelle est votre voiture préférée (préciser le modèle) .....
Quel est votre sport préféré (à pratiquer ou à regarder) .....
En dehors du sport, quelle est votre activité de loisir préférée .....
Quelle est votre émission de télévision préférée .....
Quel est votre journal, revue ou magazine préféré .....
Quelle est votre couleur préférée .....
Quel est votre animal préféré (sauvage ou domestique) .....
Quelle est votre fleur ou arbre préféré .....
Quelle est la région de France que vous préférez (pour y vivre) .....
Quel est le pays du monde que vous préférez (pour y passer des vacances) .....
Pour vous, quel est le nombre idéal d'enfants dans une famille .....
Quel est le personnage historique que vous admirez le plus .....
Quel est votre chanteur ou musicien préféré .....
Pour un objet de décoration, quelle est votre matière préférée .....
Quel est le cadeau que vous aimeriez que votre meilleur ami vous fasse .....

Pour chacun de ces deux corpus, nous allons comparer les résultats d'un dépouillement lemmatisé[1] à ceux d'un dépouillement non lemmatisé dans lequel l'analyse lexicale se fait directement à partir des formes usuelles rencontrées dans les réponses des interviewés. Nous avons pour cela utilisé le logiciel de lexicométrie *Alceste*.

Ces deux matériaux ont des caractéristiques initiales différentes, notamment en ce qui concerne la taille du texte et le type de réponses faites par les interviewés. Leur analyse nous permettra de mettre en avant d'éventuelles différences en termes de résultats et de stabilité des dépouillements. Le tableau suivant donne ainsi pour chacun d'eux les principales propriétés observées initialement.

Caractéristiques initiales des deux corpus étudiés

Caractéristique	Corpus	
	« Perception du bonheur »	« Préférences des Français »
Nombre d'*uci* (unité de contexte initial = nombre d'interviews)	1 006	1 006
Nombre de segments de texte	1 198	1 086
Nombre de formes distinctes	2 276	3 821
Nombre de mots ou expressions citées (occurrences)	24 611	31 329
Nombre moyen de citations de chaque mot	11	8
Nombre d'hapax (mot n'apparaissant qu'une seule fois)	1 146	1 880

Le premier texte étudié est moins riche que le second : il y aurait donc une plus grande homogénéité des réponses à la question à laquelle il se rapporte. Le nombre d'hapax est également plus réduit. La confrontation des deux analyses permettra par conséquent de déceler si le fait de lemmatiser ou non ces corpus donne les mêmes effets sur deux textes de taille assez voisines mais de richesses de vocabulaire différentes.

[1] On trouvera le détail de la typologie effectuée à partir du dépouillement lemmatisé dans « Le consommateur français en 1998, une typologie des préférences », *CRÉDOC Cahier de recherche*, n°130 , juin 1999.

➲ Les résultats des analyses sur les deux corpus

La première étape de l'analyse lexicale par *Alceste* est la phase de traitement des formes du corpus ; celle-ci peut se faire avec différentes options, qui correspondent à des étapes plus ou moins avancées dans la recherche du vocabulaire et la lemmatisation. Quatre options s'offrent à nous avec le logiciel :

1. La première consiste à construire le dictionnaire des formes.
2. La seconde à reconnaître les formes usuelles (s'il s'agit d'un nom, d'un adjectif, d'un verbe, d'un auxiliaire, ...) car *Alceste* ne retient que certaines d'entre elles dans l'analyse.
3. La troisième étape touche à la lemmatisation elle-même, c'est-à-dire la réduction des formes usuelles.
4. Enfin, la dernière étape réduit les autres formes à leur « racine ». En effet, après la lemmatisation et la suppression des mots de fréquence inférieure à 4, on a stemmatisé le texte, en ramenant à une racine commune les différents dérivés. Cette stemmatisation ne concerne que les mots lexicaux ; les mots grammaticaux et les noms propres ne sont pas soumis à cette opération et seront considérés comme des variables illustratives.

Nous avons choisi de comparer les résultats obtenus à l'étape 1 (prise en compte des formes textuelles brutes) et l'étape 4 (lemmatisation + réduction aux racines).

A. La perception du bonheur

Le premier corpus étudié est assez riche pour mener une analyse lexicale intéressante : il comprend 2 276 formes distinctes (mots ou racines de mots), 24 611 mots ou expressions cités, soit un nombre moyen de citations de chaque mot utilisé égal à 11.

Qu'est-ce qui diffère entre le traitement par lemmatisation et le dépouillement classique, c'est-à-dire brut ? Tout d'abord, en termes de traitement du texte initial, le tableau suivant montre clairement des écarts significatifs sur le nombre de mots analysables, retenus pour les deux types de dépouillement.

La lemmatisation du corpus entraîne :

- une réduction, dans notre exemple, du nombre de mots analysés d'environ 25% : celui-ci passe en effet de 1 977 à 1 479 ;
- un accroissement du nombre d'occurrences analysables (de fréquence d'apparition supérieure à 3) : il passe de 8 112 à 8 795 (les mots rares, c'est-à-dire apparaissant moins de trois fois, sont éliminés).

« La perception du bonheur » - Comparaison avec ou sans lemmatisation

Caractéristique	Dépouillement	
	non lemmatisé	lemmatisé
Nombre de mots analysés	1 977	1 479
Nombre de mots supplémentaires de type 'r'	278	247
Nombre d'occurrences retenues	20 654	20 654
Fréquence moyenne par mot	8,2	11,1
Nombre d'occurrences analysables (fréquence > 3)	8 112	8 795

Autrement dit, la lemmatisation permet de conserver le noyau dur du vocabulaire, tout en « mettant aux oubliettes » les mots satellites qui n'apparaissent que peu fréquemment. Ce choix vise à enrichir le plus possible les liaisons statistiques entre les formes pour donner une meilleure cohérence à l'analyse. En effet, la présence de mots rares pourraient être défavorable à cette cohérence, car deux unités de contexte très différentes mais contenant un même mot rare risqueraient d'être rapprochées.

Cela a-t-il une incidence sur les résultats, notamment sur l'ordre hiérarchique des mots les plus cités, sur les regroupements des réponses et donc, au final, sur la classification obtenue et sa stabilité ?

➲ L'ordre hiérarchique des mots les plus fréquemment cités

L'ordre hiérarchique des mots les plus fréquemment cités n'est pas pour autant affecté par la lemmatisation : on retrouve en effet dans les deux cas (dépouillement lemmatisé ou non), les mêmes premiers mots, avec des fréquences de citation finalement assez proches (cf. tableau suivant).

Les quatre mots les plus souvent exprimés sont : *santé, famille, travail et argent* ; ils obtiennent un nombre quasi identique de suffrages et ce, dans les deux types de dépouillement. Seul *travail<* dans sa forme lemmatisée apparaît plus fréquemment que le mot *travail* dans sa forme brute (293 fois, contre 241 fois) et de fait s'intercale entre famille et argent quand il y a lemmatisation.

L'analyse comparative des mots suivants vient confirmer cette idée de non bouleversement total de l'ordre des termes privilégiés dans les discours[1] : certes, il existe quelques différences de numéro d'ordre ou de fréquence d'apparition, mais les écarts restent dans une fourchette acceptable, ne remettant pas en cause la qualité des résultats.

Il est vrai que cela dépend du corpus que l'on analyse ; or, ici, on a pu remarquer que la lemmatisation avait finalement réduit significativement le nombre de mots analysés et donc le vocabulaire. La stabilité de la hiérarchie des fréquences n'allait donc pas de soi.

Cet exemple montre aussi que le rang précis de la fréquence d'un mot ou d'une racine ne peut pas être interprété dans l'absolu car il dépend des options choisies. En revanche, l'évolution des fréquences des mots peut être interprétée si les paramètres de l'analyse lexicale sont conservés à l'identique.

[1] On trouvera en annexe le tableau donnant l'intégralité des mots cités selon leur fréquence pour chacun des types de dépouillement choisi.

« La perception du bonheur » - Les mots les plus fréquemment cités selon le dépouillement

Avec lemmatisation	
Fréquence	Mots analysés
365	santé+
312	famille+
293	travail<
292	argent
221	faire.
220	ne pas
193	vivre.
192	enf+ant
186	vie+
185	pouvoir+
181	bonne-santé+
173	heur+eux
116	loisir+
106	bonheur+
100	vacance+
93	voir.
93	mes enfants
92	aller.
86	monde+
82	temps
82	problem<
78	ami+
77	amour+
77	être-bien
67	chose+
65	souci+
60	bonne+
59	envi+e
58	financier+
58	dire+
58	maison+
58	je-suis
55	malade+
55	gens
54	aim+er
50	familia+l
50	pa+yer
48	être-heureux
45	besoin+
45	niveau+
42	chôm+
42	ma-famille
40	femme+
40	voyage+
38	petits-enfants

Sans lemmatisation	
Fréquence	Mots analysés
365	santé
310	famille
292	argent
241	travail
220	ne pas
210	faire
187	vivre
185	pouvoir
185	vie
172	enfants
105	loisirs
101	bonheur
98	vacances
93	heureux
93	mes enfants
92	être en bonne santé
85	monde
82	temps
77	être bien
75	amis
71	amour
71	voir
70	problèmes
65	soucis
58	choses
58	dire
58	je-suis
57	maison
55	gens
55	envie
52	malade
48	être-heureux
47	bonne
45	bonne-santé
44	travailler
44	en-bonne-santé
42	niveau
42	chômage
42	ma-famille
41	fait
39	femme
38	petits-enfants
36	couple
36	profiter
33	soleil

Ainsi, dans notre exemple, quelque soit le type de dépouillement choisi (lemmatisé ou non), les préférences qui obtiennent le plus de suffrages restent les mêmes et leurs rangs sont quasi identiques. On constate néanmoins que, quand les formes sont lemmatisées, leur importance et leurs fréquences augmentent, ce qui aura certainement un impact à l'étape suivante, à savoir dans la définition des classes d'unités de contexte par le biais de la classification descendante hiérarchique. On le verra plus loin, certaines classes se constituent plus facilement à partir de ces « lemmes majeurs ».

➲ Les typologies

Deux typologies ont ensuite été obtenues. Elles forment cinq classes homogènes[1], chacune d'elles mettant en avant une dimension spécifique du bonheur. La classification descendante hiérarchique utilisée par *Alceste* consiste à calculer les distances entre les différentes *unités de contexte élémentaires* (uce) définies précédemment puis, à partir de ces distances, à segmenter le corpus en groupes homogènes d'unités proches distinctes les unes des autres. La distance utilisée est une distance sur le lexique : deux unités de contexte sont d'autant plus proches qu'elles contiennent des mots identiques.

Mais à ce stade d'analyse, les résultats diffèrent selon que le dépouillement a été effectué avec ou sans lemmatisation. Les regroupements des uce sont distincts selon les modalités d'analyses choisies (cf. dendogramme ci-après) et on ne retrouve grosso modo que trois classes communes sur cinq (les trois premières définies dans le tableau ci-dessous).

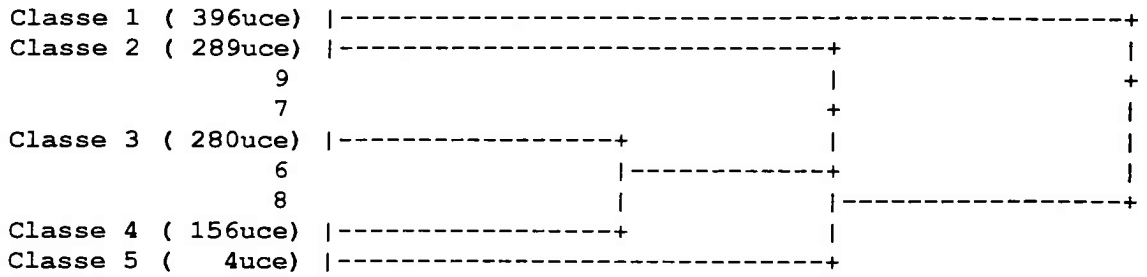
On trouvera un peu plus loin une sélection des mots fréquents utilisés pour chacune des classes dans les deux types de dépouillement.

[1] On trouvera l'analyse détaillée de la typologie (avec lemmatisation) dans « Le consommateur français en 1998 – une typologie des préférences », *CRÉDOC, Cahier de recherche* n°130, juin 1999.

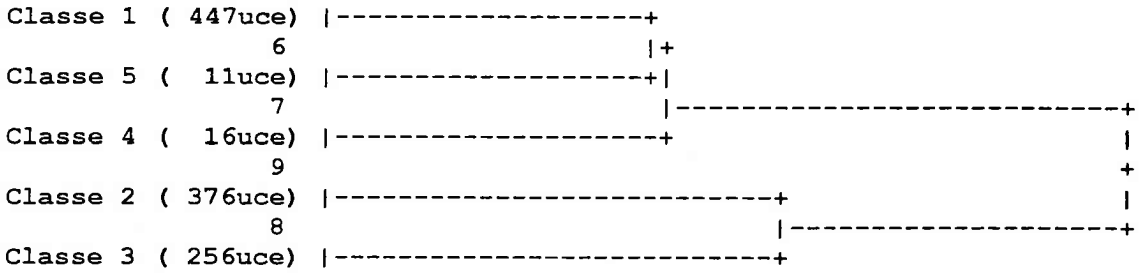
« La perception du bonheur »

Dendogramme des classes stables selon le type de dépouillement

Avec lemmatisation



Sans lemmatisation



La classification née du dépouillement sans lemmatisation donne, en effet, des groupes de taille très différente. Deux rassemblent en effet un très petit nombre d'unités (11 et 16 uce) et se forment essentiellement à partir de quelques mots, voire un ou deux. Deux autres classes regroupent à elles seules environ les trois quarts des uce. Autrement dit, la stabilité de l'analyse est ici moins nette, dans la mesure où des petites classes se forment autour de l'association très forte entre quelques mots.

Sans lemmatisation, la classe 4 (16 réponses) se forme autour des expressions « *sécurité financière* » et « *stabilité de l'emploi* », qui sont très peu fréquentes mais produisent une cooccurrence statistique très forte. Cette classe est centrée sur la notion de précarité, mais il y a évidemment bien plus de 16 réponses qui parlent du besoin de sortir de la précarité pour être heureux.

De même, la classe 5 se forme autour de l'expression « *fin de mois* ». Si on avait considéré cette expression comme une locution, cette classe ne serait probablement pas apparue.

La classification née du dépouillement avec lemmatisation n'échappe pas à ce problème, mais cela ne concerne qu'une seule des classes, les autres groupes étant plus consistants et homogènes.

Cette classe 5, qui apparaît notamment très réduite, est spécifique : elle regroupe les réponses qui mettent en avant le fait qu'une seule chose importe, la santé. Elle s'établit rapidement dans la hiérarchie, elle ne peut donc pas être « soustraite » ou regroupée, même avec une demande de classification en un nombre inférieur de classes (par exemple 4) qui aurait pu rendre l'analyse plus stable. En tout état de cause, la répartition des classes —à l'exception de la numéro 5— apparaît dans cette analyse plus cohérente et harmonieuse.

« La perception du bonheur » - Les classes obtenues

Avec lemmatisation		Sans lemmatisation	
Nombre d'uce classées	Nom de la classe	Nombre d'uce classées	Nom de la classe
396	Loisirs, vacances, consommer, profiter	447	Loisirs, vacances, consommer, profiter
280	Réussite globale de sa vie	376	Réussite globale de sa vie
289	Être entouré, penser aux autres	256	Être entouré, penser aux autres
156	Pas de soucis, pas de problèmes	16	Sécurité financière, trouver un travail
4	La santé, c'est le principal	11	Tranquillité - fin de mois

À travers cet exemple, il ressort que l'analyse lexicale avec lemmatisation semble mieux répondre à nos objectifs de mise en évidence de grands groupes de réponses. Même si cette méthode n'est pas parfaite, reconnaissons lui le mérite de mieux regrouper les réponses des individus en fonction de la signification de leur discours et d'éviter des corrélations entre formes graphiques difficilement interprétables. Cette méthode prend mieux en compte les distances entre unités de contexte, dans la mesure où ces distances sont basées sur le lexique fréquent —et donc sur un lexique plus restreint—, et surtout avec moins de corrélations artefactuelles que dans le cas d'un dépouillement non lemmatisé. Ainsi, les unités de contexte d'une même classe partagent-elles une idée sous-jacente commune plutôt qu'une expression particulière.

➲ **« La perception du bonheur » -**
Sélection de quelques formes réduites par classe

LOISIRS, VACANCES, CONSOMMER, PROFITER

Avec lemmatisation

Vocabulaire spécifique de la classe 1 :

loisir+(82), pouvoir+(93), temps(52), vacance+(87), faire.(109), maison+(36), soleil+(25), sport+(15), voiture+(16), gagn+er(22), temps-libre(21), belle+(13), achat+(12), boulot+(9), budget+(7), île+(6), liberté+(16), mer+(8), moyen+(17), musique+(8), plage+(6), plaisir+(22), voyage+(25), achet+er(22), mang+er(17), offrir.(10), partir.(12), permettre.(14), voyag+er(15), être-en-famille(10), vie-de-famille(17), loto(9), content+(8), régulier+(5), contrainte+(7), détente+(5), fille+(10), joie+(14), jour+(11), maladie+(14), mois(9), moment+(9), région+(6), dépendre.(5), pa+yer(25), retrouv+er(6), désert+ion(4), envi+e(27), montagn+e(5), prés+ent(6), cocotiers(5), stress+(9), cher+(5), simple+(6), actuellement(5), impôt+(6), retraite+(10), soir+(3), consacr+er(5), gér+er(5), profit+er(19), respect+er(3), tele(5), correct+(14), étranger+(4), france(7), avenir+(8), bateau+(4), dépense+(4), fin+(6), continu+er(4), promen+er(4), act+ion(8), superflu<(4);

Sans lemmatisation

Vocabulaire spécifique de la classe 1 :

loisirs(82), pouvoir(116), temps(51), vacances(87), faire(115), temps-libre(25), gagner(19), vivre(96), envie(34), correctement(14), gros(8), libre(19), avenir(10), loisir(10), maison(36), mer(8), moyens(15), soleil(24), sport(13), voiture(15), voyage(13), acheter(22), aller(22), garder(10), partir(12), profiter(23), voyager(14), vie-de-famille(18), tranquille(13), travailler(29), loto(9), achat(9), appartement(5), conditions(5), contraintes(7), île(6), impôts(7), liberté(16), maladie(12), maladies(6), musique(7), plaisir(17), région(6), retraite(11), consacrer(6), continuer(5), offrir(10), être-en-bonne-santé(48), ne-pas(92), cinéma(6), montagne(5), suffisamment(15), cocotiers(5), tele(6), chose(6), joie(13), terme(4), voyages(16), entretenir(4), occuper(10), retrouver(4), sortir(14), déserte(4), belle(11), financier(12), moral(3), nécessaire(8), filles(3), niveau(21), conjoint(5), permet(8), activités(5), décemment(5), maladie-grave(6);

RÉUSSITE GLOBALE DE SA VIE

Avec lemmatisation

Vocabulaire spécifique de la classe 3 :

amour+(47), famille+(113), santé+(117), couple+(22), dans son(15), être-bien(36), travail<(91), sentimenta+l(6), socia+l(6), amitié+(9), argent(88), emploi+(15), marche+(6), relation+(12), réussite+(9), situation+(10), épanou+ir(13), réuss+ir(12), vie-familiale(12), affect+ion(7), bonne+(27), professionn+el(16), revenu+(11), sta+ble(14), intéressant+(4), scolaire+(4), confort+(6), contribu+er(4), dépens+er(7), être-en-bonne-santé(34), être-heureux(20), ambian<(4), autonom<(4), enf+ant(58), harmoni+e(10), humain+(4), cadre+(3), carrière+(3), foyer+(10), projet+(5), vie+(53), aim+er(16), comprendre.(3), assur<(3), mari+age(4), tendre+(3), étude+(8);

Sans lemmatisation

Vocabulaire spécifique de la classe 2 :

famille(144), santé(161), amour(42), couple(28), dans-son(17), vie-familiale(16), bonne(30), familial(13), familiale(12), argent(114), souci(18), réussir(9), bonne-santé(26), être-bien(39), matériel(10), travail(99), financiers(12), principal(7), contacts(4), entente(13), minimum(10), projets(7), ambiance(5), enfants(71), essentiel(5), professionnelle(11), stable(10), être-riche(5), continue(7), familiaux(8), proches(8), foyer(12), mari(10), entoure(12), pas-de-problème(6), avoir-un-travail(14), plein(5), primordial(5), biens(4), marche(5), peur(4), satisfaction(4), situation(9), entendre(6), épanouie(5), réaliser(6), fait(20), harmonie(9), mariage(4), problèmes(28), riche(4);

ÊTRE ENTOURÉ, PENSER AUX AUTRES

Avec lemmatisation

Vocabulaire spécifique de la classe 2 :

heur+eux(65), petits-enfants(27), monde+(50), mes-enfants(49), ma-famille(27), chom+„(27), épou+x(13), guerre+(13), voir.(44), je-suis(29), age+(8), malheur+eux(12), gens(27), paix(18), bonheur+(39), chose+(28), esprit+(7), état+(7), forme+(5), misère+(12), parent+(12), savoir+(9), aller.(39), sentir.(13), autour-de-moi(9), en-bonne-santé(22), mon-mari(15), petit+(16), polit+3(6), possi+ble(11), rac+3(4), meilleur+(4), primordia+l(6), accord+(4), année+(4), difficulté+(6), dire+(22), personne+(7), porte+(6), terre+(4), aid+er(6), demand+er(6), exist+er(4), autour-de-soi(10), je-ne(9), import+ant(11), viol+ent(5), insécurité(4), grand+(10), naturel+(3), besoin+(18), sens(6), donn+er(5), évit+er(4), plaindre.(3), prendre.(11), cinéma<(4), médica<(3), vieill<(4), vill+„(4), ami+(27), nature+(7), société+(5), occup+er(8), rest+er(8), passe(6);

Sans lemmatisation

Vocabulaire spécifique de la classe 3 :

heureux(47), petits-enfants(27), voir(39), je-suis(37), mes-enfants(47), monde(39), ma-famille(27), gens(29), je-ne(14), chômage(24), malheureux(11), bonheur(40), savoir(10), contenter(9), donner(6), partager(6), penser(6), sentir(13), heureuse(14), choses(25), épouse(8), état(6), nature(9), solidarité(4), aider(6), vois(6), mon-mari(15), cote(7), possible(7), présent(6), aujourd(4), seule(5), ans(3), ménage(5), moments(4), partage(4), personnes(6), place(4), sens(6), fais(4), regarder(6), va(14), autour-de-moi(7), en-bonne-santé(19), important(6), voisin(4), bons(6), campagne(5), dire(19), esprit(5), fille(6), forme(3), instant(4), moment(4), parents(8), porte(5), écoute(3), envier(3), essayer(4), existe(3), prendre(9), autour-de-soi(9), cotes(3), petites(3), je-ne-sais-pas(3), misère(9), peau(7), rencontrer(5), jeunes(6), suffisant(4);

QUATRIÈME CLASSE

Avec lemmatisation : **Pas de soucis, pas de problèmes**

Vocabulaire spécifique de la classe 4 :

financier+(29), souci(23), souci+(28), bonne-santé(24), ne-pas(65), problem<(32), entente+(13), être-bien-dans-sa-p(19), familia+l(18), malade+(17), personnel+(5), quotidien+(6), contact+(4), équilibre+(7), finance+(2), question+(6), tête+(9), vue+(6), content+er(6), manqu+er(9), pos+er(5), priv+er(6), vivre.(41), confi+ant(3), vraiment(3), sécurité+(8), entendre.(9), sérénité(3), toit<(6), venir.(4), tranquil+e(9), négati+f(2), partie+(2), environnement(5), humeur+(2), mari+(5), priorité+(2), envi+er(2), gard+er(4), médecin<(2), voisin<(4), en-général(2);

Sans lemmatisation : **Sécurité financière, trouver un travail**

Vocabulaire spécifique de la classe 4 :

financièrement(3), emploi(8), finir(2), stabilité(3), sécurité(4), financière(2), boulot(2), salaire(2), compter(2), trouver(2), abri(1), désirs(1), épanoui(1), vient(1), indépendant(1), sérénité(1), dépenser(1), oblige(1), femme(2);

CINQUIÈME CLASSE

Avec lemmatisation : **La santé, c'est le principal**

port+er(4), être-riche(1), essenti+el(1), principa+l(1), pauvre+(1);

Sans lemmatisation : **Tranquillité - fin de mois**

découvert(2), fin(5), mois(6), tranquillité(6), fois(1), être-en-famille(2), resto(1), jours(2), achete(1), consommer(1), partir-en-vacances(2), calme(1), budget(1), faim(1), plage(1), questions(1), arriver(1), manger(2), payer(2), poser(1), petite(1), fête(1), entend(1), bonnes(1), relations(1);

B. Les préférences des Français

➲ Description du corpus

Le second corpus analysé est celui traitant des préférences des Français. Il est en réalité la reconstitution de toutes les réponses accolées les unes à la suite des autres pour chaque individu interrogé. Les réponses sont ici très variées et vont « dans tous les sens », elles découlent en effet de dix-huit questions distinctes (cf. l'encadré ci-dessus). Cette diversité et cette richesse des réponses donnent ici aussi un corpus intéressant à analyser.

« Les préférences des Français » - Exemples de réponses

les fruits de mer / la mousse au chocolat / Lee Cooper / Peugeot 406 / Basket-ball / la musique / Navarro / Sud-Ouest / bleu / chien / rosier / Aquitaine / Autriche / 2 enfants / Pasteur / Henri Salvador / bois / magnétoscope /
bœuf bourguignon / glace au chocolat / Caroll / Renault Mégane / tennis / marche/ l'Instit / l'Express / rouge / chien / roses / Bretagne / Égypte / 3 enfants / Victor Hugo / Phil Collins / bois / plantes /

Le corpus étudié est extrêmement riche, plus que le précédent : il comprend 3 821 formes distinctes (mots ou racines de mots), 31 329 mots ou expressions cités , soit un nombre moyen de citations de chaque mot utilisé égal à 8. Dans cet exemple également, le fait de lemmatiser le corpus (cf. tableau suivant) :

- réduit le nombre de mots analysés d'environ 10% : celui-ci passe en effet de 3 524 à 3 141;
- mais accroît légèrement le nombre d'occurrences analysables (de fréquence d'apparition supérieure à 3) : il passe de 17 275 à 17 769.

En fin de compte, les analyses ont porté respectivement sur 3 524 et 3 141 mots. Dans les méthodes d'analyse statistique, en fixant un seuil de fréquence en deçà duquel les mots ne sont pas analysés, on élimine une partie importante du vocabulaire, mais nettement moins que quand on procède à la lemmatisation.

Le nombre élevé de noms propres analysés et la faible fréquence des formes conjuguées des verbes expliquent ici l'impact relativement faible de la lemmatisation. Ainsi, la réduction du vocabulaire apparaît ici moins importante qu'elle ne l'a été dans notre premier exemple sur le bonheur.

Le corpus est donc plus riche, plus diversifié et, par conséquent, la fréquence moyenne des mots et de leurs différentes formes possibles est moindre.

« Les préférences des Français » - Comparaison avec ou sans lemmatisation

Caractéristique	Dépouillement	
	non lemmatisé	lemmatisé
Nombre de mots analysés	3 524	3 141
Nombre de mots supplémentaires de type 'r'	194	188
Nombre d'occurrences retenues	26 223	26 223
Moyenne par mot	6,1	7,0
Nombre d'occurrences analysables (fréquence > 3)	17 275	17 769

Cependant, la fréquence de citation de certains mots est particulièrement forte. Cela permet de dégager les préférences les plus marquées des Français : l'animal préféré est *le chien* pour 51% de Français, *le bois* est la matière de décoration préférée (50%), le nombre idéal est *deux enfants* pour 48 % et la couleur favorite est plus souvent *le bleu* (43%).

On constate que, quelque soit le type de dépouillement choisi (lemmatisé ou non), les préférences qui obtiennent le plus de suffrages restent les mêmes et leurs scores sont quasi identiques. La lemmatisation a donc une incidence très faible sur les mots qui sont fréquemment utilisés dans un corpus, mais son impact s'avère plus fort sur le vocabulaire peu ou moyennement courant.

« Les préférences des Français »

Les préférences qui obtiennent le plus de suffrages

Peu de différences selon que le corpus ait été ou non lemmatisé

Préférence pour ...		Part des enquêtés	
		Dépouillement lemmatisé	Dépouillement non lemmatisé
Animal	Le chien / le chat	51 % / 25 %	50 % / 25 %
Matière	Le bois	50 %	50 %
Nombre idéal d'enfants	Deux enfants	48 %	48 %
Fleur ou arbre	Les roses	45 %	41 %
Couleur	Le bleu	43 %	43 %
Marque de voiture	Renault / Peugeot	26 % / 14 %	16 % / 14 %
Pays (pour vacances)	La France	24 %	24 %
Sport	Le football (foot)	20 %	20 %
Région (où vivre)	Provence – Côte d'Azur	19 %	14 %
Dessert	Chocolat (mousse, gâteau,...)	19 %	19 %
Loisir	Marche, ballade / lecture	18 % / 14 %	10 % / 14 %
Personnage historique	Le Général de Gaulle / Napoléon	17 % / 13 %	17 % / 13 %
Chanteur, musicien	Michel Sardou	10 %	11 %

Source : CRÉDOC, Enquête Consommation 1998

Le dictionnaire des préférences donnant la fréquence de chacun des mots utilisés dans l'analyse montre bien les différences existant entre les deux dépouillements ; leur proximité dans cette étape de l'analyse reste néanmoins très nette.

Deux cas notamment sont très parlant : *le football* d'une part, *la rose* d'autre part. En effet, pour ces deux mots, le dépouillement lemmatisé donne une fréquence d'apparition assez élevée, de respectivement 198 et 451, tandis que l'analyse sans lemmatisation distingue *foot* de *football*, *rose* de *roses*, et donc restitue à chacun une fréquence plus faible (respectivement de 119, 79, 409 et 42). On ne peut pas dire que cette distinction apporte un sens particulier à l'analyse.

Si on compare les deux analyses, il y a certes des différences qui vont modifier la hiérarchie, mais surtout cela risque d'avoir un impact sur la formation des classes issues de la classification descendante d'*Alceste*. En effet, il n'est pas sûr que dans la classification, par la suite, les mots *foot* et *football* ou *rose* et *roses* par exemple se retrouvent ensemble dans les mêmes classes. En revanche, leur regroupement a davantage de chances de fournir une meilleure stabilité des résultats.

« Les préférences des Français »

Les mots les plus fréquemment cités selon le dépouillement

Avec lemmatisation	
Fréquence	Mots analysés
512	chien+
506	bois
485	deux-enfants
451	rose+
435	bleu+
417	trois-enfants
257	chat+
245	France
198	foot+
175	Gaull+e
173	rouge<
172	vert+
158	Renault
139	Peugeot
136	lecture+
132	Napoléon
106	Sardou
105	marche+
104	tarte+
92	fruit+
91	télé
90	noir+
89	chocolat<
89	Levis
88	glace+

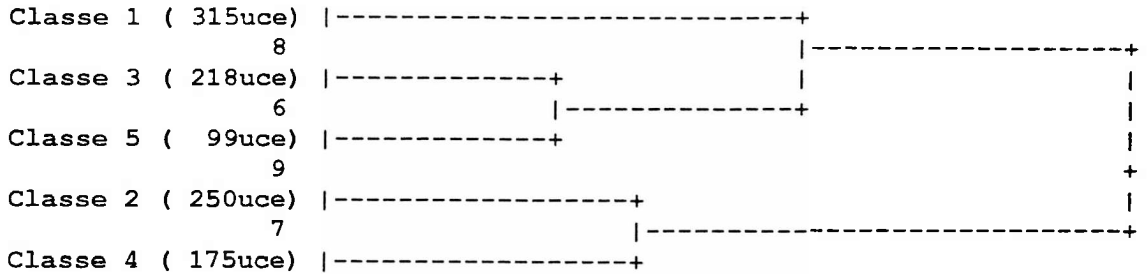
Sans lemmatisation	
Fréquence	Mots analysés
506	bois
503	chien
485	deux-enfants
432	bleu
417	trois-enfants
409	rose
250	chat
245	France
175	Gaulle
163	vert
162	rouge
158	Renault
139	Peugeot
136	lecture
132	Napoléon
119	foot
106	Sardou
105	marche
91	télé
89	Levis
87	Canada
87	tennis
85	chocolat
83	noir
83	Bretagne

Dans une deuxième étape, *Alceste* sélectionne les unités de contexte (uce) et prépare les calculs des données. Il effectue ensuite une classification descendante hiérarchique, en éliminant les mots de fréquence inférieure à 4 et supérieure à 3000, et constitue des classes (le dendogramme des classes stables est représenté ci-après).

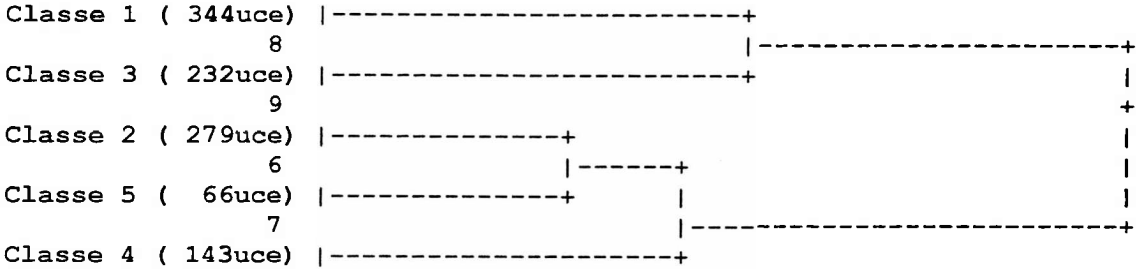
« Les préférences des Français »

Dendogramme des classes stables

Avec lemmatisation



Sans lemmatisation



L'analyse lexicale standard sur l'ensemble du corpus a permis de dégager cinq « portraits ». Chacun d'eux présente des préférences qui se recoupent.

« Les préférences des Français » - Les classes obtenues

Avec lemmatisation		Sans lemmatisation	
Nombre d'uce classées	**Nom de la classe**	**Nombre d'uce classées**	**Nom de la classe**
315	Les femmes actuelles	232	Les femmes actuelles
250	Les hédonistes modernes	279	Les hédonistes modernes
218	Les âgés traditionalistes	344	Les âgés traditionalistes
175	Les intellectuels bons vivants	143	Les intellectuels bons vivants
99	Les rustiques	66	*(« Les jeunes hommes modernes »)*

Les deux dépouillements, lemmatisé et sans lemmatisation, ont en commun un certain nombre de caractéristiques. Les deux classifications qui en découlent semblent assez cohérentes. Sur les cinq classes mises en évidence, quatre notamment se ressemblent fort[1].

[1] On trouvera un peu plus loin les tableaux détaillés des formes employées pour chacune des 5 classes.

Néanmoins, le découpage des classes n'est pas similaire d'une analyse à l'autre :

- D'une part, du point de vue des **unités prises en compte** : les ordres de grandeur du nombre d'uce retenues dans chaque classe varie assez sensiblement selon le type analyse.
- D'autre part, en terme de **découpage des classes** proprement dit. On ne retrouve pas en effet la même structure de l'arbre de classification. Dans l'analyse avec lemmatisation, la classe 5 (« les rustiques ») se disjoint de la classe 3, c'est-à-dire des « âgés traditionalistes » ; on y trouve toujours beaucoup de personnes âgées, mais surtout davantage d'ouvriers et de ruraux. Dans le dépouillement sans lemmatisation, la classe 5 provient d'une segmentation de la classe 2 (« les hédonistes modernes ») dans laquelle les hommes jeunes sont davantage sur-représentés.

On constate que la différence se fait essentiellement sur la dernière classe, c'est-à-dire celle qui est la plus petite en taille, qui rassemble le moins grand nombre de références. Est-il plus intéressant d'obtenir la classe « les rustiques » ou celle des « jeunes hommes modernes » ? Il est évidemment difficile de trancher scientifiquement.

On peut cependant avoir l'impression que la classe des « rustiques » se distingue mieux de celle des « âgés traditionalistes » par l'évocation de chanteurs populaires comme Johnny Hallyday, de marques de vêtements de sport et d'activités « terriennes » comme le jardinage. Les « traditionalistes » apprécient plutôt Michel Sardou, les marques de vêtements seniors et la télévision ou la pétanque.

En ce qui concerne les classes non lemmatisées « jeunes hommes modernes » et « hédonistes modernes », la distinction apparaît moins évidente. Les voitures puissantes, les émissions télévisées de dérision et les plats roboratifs sont fréquemment cités dans les deux groupes. Les voitures sont un peu plus sportives dans le groupe « jeunes hommes modernes » et les références musicales ou culturelles un peu plus intellectuelles dans la classe « hédonistes modernes ».

Aux cinq classes de réponses mises en évidence, s'ajoutent quelques réponses non classées par l'algorithme de constitution des classes. Environ 5% des individus n'ont pas été classés par l'analyseur, cela représente environ 10% des réponses. Le tableau suivant donne les parts respectives de ces classes dans l'ensemble des réponses.

« Les préférences des Français »

Classification des réponses par *Alceste* selon le mode de lemmatisation

Classe de réponses	Part avec lemmatisation		Part sans lemmatisation	
	Classées (en %)	Ensemble (en %)	Classées (en %)	Ensemble (en %)
Réponses non classées		5,2		6,4
Les femmes actuelles	24,0	22,7	22,2	20,8
Les hédonistes modernes	20,3	19,3	21,5	20,1
Les âgés traditionalistes	20,0	19,0	25,3	23,7
Les intellectuels bons vivants	19,8	18,8	18,8	17,6
Les rustiques / *« Les jeunes hommes modernes »*	15,0	15,0	12,2	11,4
Total	*100,0 %*	*100,0 %*	*100,0 %*	*100,0 %*

➲ **« Les préférences des Français » -**
Sélection de quelques formes réduites par classe

LES FEMMES ACTUELLES

Avec lemmatisation

Vocabulaire spécifique de la classe 1 :

bijou+ (55), femme+ (44), actu+el (41), parfum+ (30), lecture+ (67), Naf-Naf (19), natation (47), rose+ (165), crista+l (27), gymnast+3 (20), porcelaine+ (33), Fabian (10), Lara (9), mousse-au-chocolat (36), patinage-artistique (27), décorati+f (8), manuel+ (5), rôti+ (13), Canada (43), citron+ (15), fleur+ (31), lampe+ (6), langue+ (5), mark+ (7), Pascal (8), Burton (9), cabriolet (9), Faut-pas-rêver (19), Goldman (31), Louis-XIV (18), Mégane (25), Obispo (9), Sud-ouest (26), cote+ (32), Biarritz (5), hachis (7), parmentier (8), Girond+ (6), jaune+ (27), boeuf+ (13), chant+ (4), coutur+e (7), étain+ (17), île+ (18), mot+ (10), pyramide+ (9), tarte+ (43), travaux (6), verre+ (24), vêtement+ (14), prim+er (8), Marie (11), camaïeu (10), chantilly (10), Morgan (7), orchidée (10), paella (14), Pagny (16), pot-au-feu (12), saumon (11), top santé (7), avantag+e (5), azur< (20), religi< (5), Calédonie (5), instit (4), Kennedy (6), Manoukian (5), Modes-et-Travaux (5), montan+ (7), patchwork (3), Pimkie (3), spencer (6), Capita+l (31), clair+ (8).

Sans lemmatisation

Vocabulaire spécifique de la classe 3 :

femme (45), Naf-Naf (22), actuelle (43), bijou (37), lecture (61), porcelaine (32), natation (41), patinage-artistique (27), Fabian (10), Lara (9), décoration (7), gymnastique (17), parfum (20), pyramide (9), marie (12), cabriolet (9), Kiabi (10), Louis-XIV (17), Obispo (8), Claire (7), rôti (9), Canada (33), bijoux (6), cristal (19), étain (14), îles (8), marks (5), parents (6), variété (6), vêtement (12), fiat (7), pascal (7), mousse-au-chocolat (27), Calédonie (5), Modes-et-Travaux (6), spencer (6), unis (4), jaune (21), broderie (5), café (6), chant (3), chou (3), couture (6), fleur (7), fleurs (20), livres (4), plante (4), préférence (14), promenades (6), verre (19), préférée (5), Maurice (5), Aznavour (12), chantilly (8), Combien-ça-coûte (10), marche-à-pied (12), paella (12), Peugeot (43), Top Santé (6), Vendée (6), religieuse (4), Aveyron (4), crêpe (4), Seat (3), Gironde (4), nouvelle (7), noir (27), citron (9), famille (5), feux (4), gâteaux (6), tricot (7), violette (5), croises (6), Antilles (10), documentaires (8), magret-de-canard (6);

LES HÉDONISTES MODERNES

Avec lemmatisation

Vocabulaire spécifique de la classe 2 :

bois (164), pâte+ (39), Ferrari (20), moto+ (22), Levis (47), chêne+ (40), équipe+ (13), tigre+ (11), martin (8), carbonara (9), Collins (10), foot (48), Lacoste (17), Nulpartailleurs (15), Phil (10), Strauss (10), informat+3 (11), liégeois+ (8), austral+ (17), parisien+ (14), safrane+ (7), chaîne+ (7), dire+ (5), pin+ (9), sapin+ (14), ski+ (20), voiture+ (24), écout+er (6), bolognaise (7), Cooper (8), Einstein (8), Francis (9), Irlande (12), Lee (9), Léonard-de-Vinci (9), Marseille (6), pizza (7), Porsche (13), chass+e (11), Guignol< (6), a4 (5), Bob (5), Eddy-Mitchell (6), Floyd (4), Friends (5), Hendrix (5), Indonésie (5), Jimmy (6), judo (6), King (5), Lamborghini (4), Luther (5), pc (7), Pink (5), Strait+ (5), auto+ (8), box+e (8), Cana+l (5), parasol+ (4), pêche+ (18), science+ (6), séjour+ (4), surprise+ (7), voyage+ (25), fondre. (4), jou+er (5), sortir. (4), Audi (11), Bmw (17), Brassens (12), Cabrel (11), Mitterrand (10), profiteroles (9), Sud-est (7), Volkswagen (11), vtt (8), Beatle+ (4), Boss (5), Bouches-du-Rhône (5);

Sans lemmatisation

Vocabulaire spécifique de la classe 2 :

bois (189), Levis (51), pâtes (35), Phil (12), cd (9), Collins (12), Lacoste (18), pc (8), Capital (31), parisienne (8), Safrane (8), équipe (12), musique (20), ski (22), Martin (8), Audi (13), Brassens (15), carbonara (8), Celio (10), cerisier (10), Einstein (9), entrecôte (13), Envoyé-Spécial (28), Irlande (13), Laguna (13), Marche-du-siècle (17), Nulpartailleurs (14), cote (28), informatique (10), Beatles (5), Dauphine (6), footing (5), Friends (5), Georges-Brassens (4), Hendrix (5), Jimmy (6), King (5), Luther (5), Straits (4), Australie (15), bouteille (9), journal (15), lampe (5), ordinateur (7), pêche (19), région (12), sapin (11), séjour (4), voiture (20), bricolage (24), écouter (5), Passat (4), pierre (10), bolognaise (7), cassoulet (8), érable (6), forêt-noire (14), Goldman (27), Marseille (6), Mercedes (24), Mitterrand (10), Sud-est (8), azur (19), cinéma (30), libération (9), a4 (5), coccinelle (3), devred (5), jeans (4), Massif-Central (4), spaghetti (5), Volvo (5), yaourts (5), amitié (6), chaîne (5), Dire (4), espace (13);

LES ÂGÉS TRADITIONALISTES

Avec lemmatisation

Vocabulaire spécifique de la classe 3 :

fruit+ (49), marche+ (46), Citroën (26), Questions-pour-un-c (24), Rossi (14), Xm (10), France (72), Italie (25), légume+ (18), raisin+ (9), tableau+ (13), temps (7), gaul+er (59), Frédéric (14), François (16), patin+„ (15), frai+c„ (5), midi+ (24), Angleterre (5), émission+ (17), marne+ (9), tapisserie+ (5), variété+ (11), Claude (6), bouquet-de-fleurs (14), choucroute (16), Clio (19), pétanque (10), Sardou (36), montagn+e (14), Bx (4), chrysanthème (4), Damart (5), Foucault (4), Mgriffon (5), arc+ (6), belote+ (3), cuivre+ (13), grille+ (4), jour+ (11), livre+ (22), marque+ (17), pays (8), peinture+ (12), santé+ (6), sauce+ (11), soup+e (5), tissu+ (13), tour+ (3), connaître. (7), cors+er (11), marguerite (5), Aznavour (12), Figaro (10), Jeanne D'arc (6), agneau< (7), canevas (3), C&a (3), Fr3 (3), La-chance-aux-chans (5), magnétoscope (3), scrabble (4), Sergelama (4), Zx (3), rouge< (46), Loire (10), Centre+ (6), faïence+ (5), match+ (4), mer+ (6), poire+ (7), poisson+ (21), vélo+ (22), ferr+er (6), maxi (5), ancien< (4), petit+ (9).

Sans lemmatisation

Vocabulaire spécifique de la classe 1 :

Citroën (36), Républicain (25), Frédéric (20), François (23), pétanque (16), Questions-pour-un-c (32), Rossi (18), Sardou (60), bague (10), chien (200), cuivre (21), marche (55), Gaulle (86), Nord (13), peint (6), rose (158), midi (31), France (102), Italie (30), chiffres (6), flan (12), football (38), fruit (16), lettres (6), papier (10), télévision (12), temps (7), terre (25), variétés (10), jardinage (26), bouquet-de-fleurs (17), foot (52), journal-télévisé (10), Ouest-France (22), Xantia (12), Xm (10), Damart (7), Enrico (7), La-chance-aux-chans (7), Macias (7), allemand (4), dimanche (6), Allemagne (5), Charente-Maritime (8), belote (4), cartes (5), chats (6), cravate (4), éclair (5), fruits (35), légumes (18), marne (9), marque (19), pommes (18), raisin (6), soupe (6), tableau (12), tribune (6), voit (5), Charles (7), Claude (6), Duteil (8), endives (7), Fiesta (8), Ford (17), agneau (9), informations (11), patinage (13), sincère (4), C&a (4), lorrain (5), scrabble (5), yaourt-aux-fruits (4), pleine (4), Angleterre (4), Autriche (9).

LES INTELLECTUELS BONS VIVANTS

Avec lemmatisation

Vocabulaire spécifique de la classe 4 :

monde+ (25), St Laurent (10), baba (8), rhum (8), canard+ (18), St Honor+er (7), Marianne (9), Inde (4), art+ (8), jaguar+ (13), vin+ (13), voile+ (10), enchain+er (8), Yves (5), Beethoven (7), Jaurès (7), Thalassa (30), volley ball (7), Schubert (5), acier+ (4), actualité+ (5), ail (4), bouteille+ (9), cuisin+e (6), fer+ (6), forge+ (4), golf+ (15), hêtre+ (6), lion+ (6), maison+ (12), meuble+ (5), table+ (7), tennis (27), confire. (8), voyag+er (8), Henri (11), Jean (8), Michel (5), aime-pas (9), Bourgogne (6), Celio (9), fruits-de-mer (10), Gandhi (7), gâteau-au-chocolat (14), louis (7), Mercedes (19), Scenic (12), Libérat+ion (7), Chrysler (5), Diesel (5), nougat (4), Science-et-vie (5), ferre+ (3), naturel+ (3), particulier+ (4), plat+ (6), Amérique< (5), bouleau+ (5), cheva+l (20), débat+ (3), glace+ (24), lys (6), planche+ (4), plateau+ (3), bricol+er (18), blanquette (6), bleu-marine (5), Bordeaux (5), cd (5), Envoyé-special (17), express (8), magret-de-canard (6), Pays-basque (9), Seychelles (5), bonne+ (8), Chopin (3).

Sans lemmatisation

Vocabulaire spécifique de la classe 4 :

golf (20), jaguar (17), St Laurent (10), marques (10), St Honoré (7), Jacques (7), canard (16), meuble (5), brûlée (6), Thalassa (29), Schubert (5), Inde (4), arts (5), bouleau (7), hêtre (5), confit (8), Yves (5), Brel (21), Express (10), Faut-pas-rêver (15), Gandhi (8), man (6), marron (5), plat (5), saint (6), ail (4), cuisine (5), fer (5), feu (3), forge (4), lion (6), monde (16), Palette (3), planche (4), poisson (18), voile (7), voyage (16), enchaîné (6), regarde (10), voyager (7), Michel (5), Arte (5), ball (5), Bourgogne (6), foie-gras (7), new (6), Pays-basque (10), Télérama (7), Bach (4), Chrysler (4), mini (3), Nougaro (5), Vsd (4), rouges (4), vert (34), Portugal (5), argent (3), art (3), bouillon (4), champignons (4), cuite (6), maison (9), modèle (5), noix (4), peinture (9), tennis (21), voyages (5), Henri (9), Churchill (4), fruits-de-mer (8), gâteau-au-chocolat (11), Léonard-de-Vinci (5), Marianne (5), Mozart (6), saumon (6), Sud-ouest (13), volley (4), vtt (6), albisia (3), Indonésie (3), Queen (3), Bretagne (18), chevaux (4);

LA CINQUIÈME CLASSE

Avec lemmatisation : **« Les Rustiques »**

Vocabulaire spécifique de la classe 5 :

chiffre+ (6), lettre+ (6), Républicain+ (12), pomme+ (16), Enrico Macias (6), Mère Teresa (5), gratin+ (6), papier+ (6), terre+ (16), sincer+e (5), dahlia+ (5), allemand+ (3), peint+ (5), berger+ (4), champion+ (3), flan+ (8), pommier+ (6), jardin+er (13), Adidas (17), laser (3), R5 (3), Allemagne (3), Espagne (13), Portugal (5), bague+ (4), chien+ (65), course+ (9), paquet+ (3), poche+ (4), soie+ (6), télévision+ (5), tilleul+ (4), César (4), endives (4), fiesta (5), lorraine (5), Peugeot (27), son-amitié (8), tarte-aux-pommes (13), télé (17), montre+ (7), plast+3 (4), Astra (4), glaïeul (4), Gti (3), muguet (3), vidéo (4), yaourts (4), Nord+ (5), quotidien+ (3), Ardennes (3), breton+ (2), pied+ (7), singe+ (2), vanille+ (3), march+er (2), gratin-dauphinois (4), infos (4), frambois+ (3), lorrain (2), rosbif (3), Sénégal+ (2), Vaucluse (2), Weil (2), yaourt-aux-fruits (2), Auvergne (6), Montpellier (2), amitié+ (3), amour+ (3), eau+ (2), écho+ (2), prix (3), progrès (3), promen+eur (10), port+er (2), Hallyday (9), Ouest-France (8);

Sans lemmatisation : **« Les jeunes hommes modernes »**

Vocabulaire spécifique de la classe 5 :

moto (13), Cooper (8), Lee (9), turbo (7), automobile (7), chasse (8), Rolling (5), Stones (6), Ferrari (9), baba (5), rhum (5), hautes (4), Brésil (4), sport (10), Francis (6), daube (4), Floyd (3), Zone-interdite (4), Amérique (5), fraise (4), Cabrel (7), Porsche (7), judo (4), Pink (3), verte (3), Canal (3), chêne (14), classe (3), fromage (4), photo (3), république (2), rugby (8), super (3), vin (5), Balavoine (3), Alpes (8), Mexique (4), Nike (5), Sans-aucun-doute (5), Strauss (4), Guignols (3), mécanique (2), Reebok (2), Chine (3), loup (2), olivier (4), pin (3), restaurant (3), salade (4), sortir (2), crustacés (3), gratin-dauphinois (3), Var (5), Coluche (2), Diesel (2), hebdo (2), info (2), pistache (2), haute (6), Pyrénées (5), Sud (6), Bmw (6), gris (2), parisien (3), auto (3), série (3), tigre (3), tilleul (2), éclair-au-chocolat (2);

On a vu que le nombre d'uce retenu pour constituer chacune des classes varie assez sensiblement selon le mode de dépouillement. Néanmoins, si on replace ces résultats en nombre d'individus, on voit que la taille des classes change légèrement, mais la différence observée de quelques pour cent peut être considérée comme négligeable. Ce qui importe davantage, et qui est plus inquiétant, c'est que la structure globale de l'arbre de classification n'est pas identique dans les deux cas. Dès que l'on souhaite faire des découpages plus fins, on se heurte donc à des regroupements différents de vocabulaire selon que l'on effectue ou non la lemmatisation.

Autrement dit, les essais que nous avons effectués montrent que lemmatiser ou non change finalement assez peu la nature des classes obtenues, tout du moins celles qui regroupent un nombre suffisant d'uce. En revanche, la lemmatisation a l'avantage, par rapport à la « non lemmatisation », de réduire le nombre de « petites classes », c'est-à-dire celles formées à partir d'un très petit nombre d'unités de contexte et apparemment les moins stables. Elle limite en effet les corrélations entre formes graphiques difficilement interprétables. Le moins grand éparpillement du vocable (i.e. des mots exprimés sous des formes différentes mais ayant une même signification) induit alors une homogénéité des rapprochements par la mesure de distance sur le lexique et donc une plus grande robustesse des typologies.

Dans les deux exemples que nous avons traités, nous observons d'ailleurs la même incidence de la lemmatisation / non lemmatisation sur l'analyse lexicale : l'effet sur l'ordre hiérarchique des mots est relativement faible, mais on constate un effet plus important sur l'arbre de classification, c'est-à-dire sur le découpage même du corpus en classes et donc sur le rapprochement des réponses en fonction du vocabulaire.

La lemmatisation permet donc une classification à partir de « noyaux durs ». Elle rend plus stables les analyses car, tout en gardant la richesse du vocabulaire, elle le réduit en taille pour mieux associer les unités étudiées en fonction de leur sens.

Ces résultats nous mèneraient donc à opter plus facilement pour des analyses utilisant la lemmatisation. Néanmoins, on peut se demander si les caractéristiques des corpus pris en compte n'ont pas une incidence sur ces conclusions. Il pourrait alors être intéressant, dans le cadre d'un autre travail de recherche, de les valider sur des corpus de nature différente, par exemple des textes pour lesquels une lemmatisation entraînerait une plus grande réduction de la taille du vocabulaire qu'elle ne le fait sur les deux corpus que nous avons utilisés ici.

II. IMPACT DU SEUIL DE SÉLECTION DES MOTS SUR LA STABILITÉ DES TYPOLOGIES

L'un des avantages constaté de la lemmatisation est la plus grande stabilité des analyses et notamment l'évitement de corrélations entre formes graphiques difficilement interprétables. Étant donné que les analyses lexicales s'effectuent sur des tableaux de grande dimension très clairsemés, elles risquent d'être assez instables ; la lemmatisation a le mérite de réduire au maximum la taille du vocabulaire et donc de ces tableaux. Cependant, la lemmatisation est toujours combinée au choix d'un seuil minimal de fréquence pour sélectionner les mots actifs de l'analyse.

Nous allons donc tenter, dans cette seconde partie, de vérifier la robustesse des typologies lexicales établies avec le logiciel *Alceste*. Pour cela, nous sommes partis d'analyses effectuées avec lemmatisation et nous avons testé quatre degrés différents de lemmatisation, c'est-à-dire quatre seuils de fréquence des mots en deçà desquels le vocable n'est pas pris en compte dans l'analyse.

Notre objectif est alors de montrer (ou d'infirmer) que, quel que soit le seuil de fréquence des mots retenu, la classification obtenue reste homogène et cohérente.

Quatre hypothèses de seuils ont été retenues : le seuil minimum de 4 (qui correspond à la limite du plan d'analyse standard avec le logiciel *Alceste*) et les seuils de 6, 10 et 15. Autrement dit, les mots du corpus ne seront pris en compte dans l'analyse que s'ils apparaissent au moins respectivement 4 fois, 6, 10 ou 15 fois selon l'hypothèse. Les écarts entre les seuils sont volontairement assez espacés afin de tester la robustesse des classifications.

Les mêmes deux corpus que ceux analysés dans la première partie du rapport ont été retenus pour cette comparaison : « La perception du bonheur » et « Les préférences des Français ».

II.1. LE CORPUS « LA PERCEPTION DU BONHEUR »

Rappelons tout d'abord les caractéristiques propres à ce corpus. A l'état brut, ce corpus contient 24 611 mots qui se répartissent en 2 276 formes graphiques distinctes. C'est par la lemmatisation et la fixation d'un seuil de fréquence minimale de prise en compte d'une forme que nous avons réduit le vocabulaire analysé. Le nombre de formes distinctes variera ensuite selon la valeur du seuil retenu.

« La perception du bonheur » - Caractéristiques du corpus analysé

Nombre de formes distinctes (étendue du vocabulaire)	2 276
Nombre d'occurrences (étendue du texte)	24 611
Fréquence moyenne par forme	11
Nombre d'hapax (nombre de mots n'apparaissant qu'une seule fois)	1 146
Proportion d'hapax	50 % de l'ensemble des formes
Nombre d'uce enregistrées	1 169
Nombre d'uce classées	1 125, soit 96%

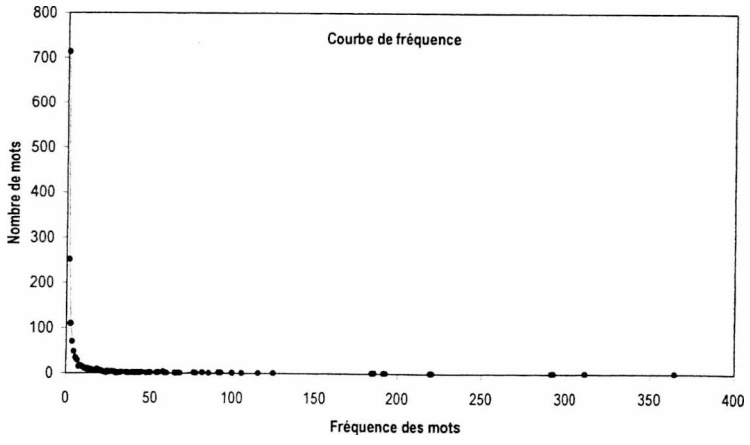
La distribution des fréquences de mots et la richesse du vocabulaire (qui en découle) sont depuis longtemps considérées comme des indicateurs pertinents pour la caractérisation des textes. La distribution de fréquence est un tableau qui associe à chaque classe de fréquence un effectif qui correspond au nombre de vocables qui ont cette fréquence. Elle fait abstraction du contenu lexical du texte, mais elle donne une image de sa structure lexicale. L'étude et la comparaison de gammes de fréquence permettent également de mieux mesurer sur quelle partie du vocabulaire porte l'analyse statistique textuelle. En effet, on est amené à fixer un seuil de fréquence en dessous duquel les mots ne sont pas analysés. L'incidence de ce seuil variant en fonction de la gamme de fréquence, il est nécessaire de bien connaître celle-ci, afin de faire des choix pertinents, ou au moins en connaissance de cause.

Connaître la répartition fréquentielle des mots dans un corpus permet donc de mieux maîtriser le fonctionnement de l'analyse lexicale.

Les basses fréquences représentent une proportion élevée du vocabulaire d'un corpus, or pour l'analyse statistique qui compare les profils lexicaux d'unités textuelles, elles sont peu signifiantes (c'est le cas des hapax : n'ayant qu'une seule occurrence, ils n'entrent activement dans aucun système de classification). Le cas des locutions apparaissant sous forme d'hapax est même plus gênant dans le cas où elles font apparaître des cooccurrences artefactuelles. C'est la raison pour laquelle les basses fréquences sont éliminées d'emblée, car le seuil minimal au-delà duquel on conserve les vocables est de 4 dans *Alceste*. On fait subir un appauvrissement au texte, mais dans le but d'obtenir des résultats plus stables.

➲ Distribution des fréquences

Quand la fréquence des mots croît, les effectifs correspondants décroissent selon la loi de Zipf. Sur le graphique suivant, les fréquences les plus faibles sont toutes représentées avec des effectifs élevés, alors que quand les effectifs deviennent plus faibles, certaines fréquences élevées ne sont pas représentées.



On a réalisé plusieurs tests en faisant varier le seuil de fréquence. Plus le seuil de fréquence est élevé, moins on conserve d'observations à analyser. Cela a-t-il une incidence sur les résultats de la classification issue de l'analyse lexicale ?

➲ La typologie obtenue avec un seuil de fréquence 4

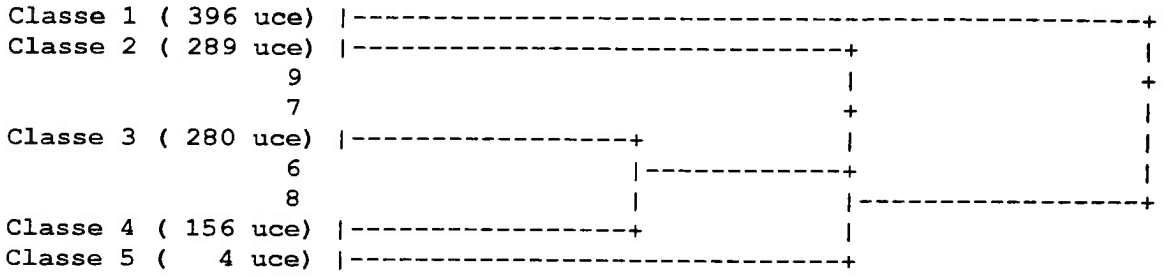
Afin d'obtenir une meilleure stabilité des résultats, on réduit de nouveau (après lemmatisation) la taille du vocabulaire en imposant un seuil de fréquence minimum aux formes réduites. Ce seuil est fixé à 4. Dans cette opération, on perd une partie importante du corpus puisque les mots rares apparaissant moins de 4 fois sont éliminés. Il s'agit d'un choix visant à enrichir le plus possible les liaisons statistiques impliquées par les cooccurrences des formes.

Le corpus initial contenait 2 276 formes différentes. Après lemmatisation et élimination des mots ayant une fréquence d'apparition inférieure au seuil fixé, le matériau définitif est constitué de 403 formes réduites apparaissant 4 fois ou plus. La lemmatisation et la réduction du vocabulaire ont donc permis de réduire la taille du lexique de 80%. Rappelons ici que la classification opérée par *Alceste* est une classification descendante hiérarchique gérée sur le tableau « hypercreux » croisant les uce et les formes analysées, soit ici un tableau de taille 1 169 * 403. Il y a en moyenne 7,5 formes analysées sur les 403 possibles par uce, ce qui signifie que 98 % des cases du tableau analysées sont des « zéros » et seulement 2 % des « uns » indiquant la présence d'une forme dans une uce.

Les mots, par la manière dont ils apparaissent ou non dans les uce permettent de classer les unités de contexte par maximisation de la variance interclasse. On a donc au départ un ensemble d'uce qui, grâce à une méthode de classification descendante, est segmenté en deux classes, de telle sorte que chaque classe soit aussi homogène que possible d'un point de vue lexical et se différencie fortement de l'autre. Le processus de classification est itératif et conduit à une série de cinq classes constituées. En fin de compte, chaque unité de contexte appartient à une classe, —sauf les unités qui ne sont pas classées parce qu'elles emploient un vocabulaire trop marginal ou qu'elles sont « à cheval » entre deux classes—.

Avec un seuil de fréquence de 4, nous avions obtenu une classification en cinq groupes homogènes, dont un est de taille très inférieure aux autres[1]. Chacune de ces classes rassemble néanmoins un nombre d'uce qui varie assez sensiblement.

L'arbre de classification est le suivant :



Loisirs, vacances, consommer, profiter	(27% des individus),	rassemble 396 uce	(Classe 1)
Être entouré, penser aux autres	(21% des individus),	289 uce	(Classe 2)
Réussite globale de sa vie	(29% des individus).	280 uce	(Classe 3)
Pas de soucis, pas de problèmes	(21% des individus),	156 uce	(Classe 4)
La santé, c'est le principal	(3% des individus),	4 uce	(Classe 5)

➲ Les différentes hypothèses de seuil retenues

Nous avons ensuite testé l'hypothèse de stabilité de cette classification en faisant varier le seuil de fréquence minimum d'apparition des mots.

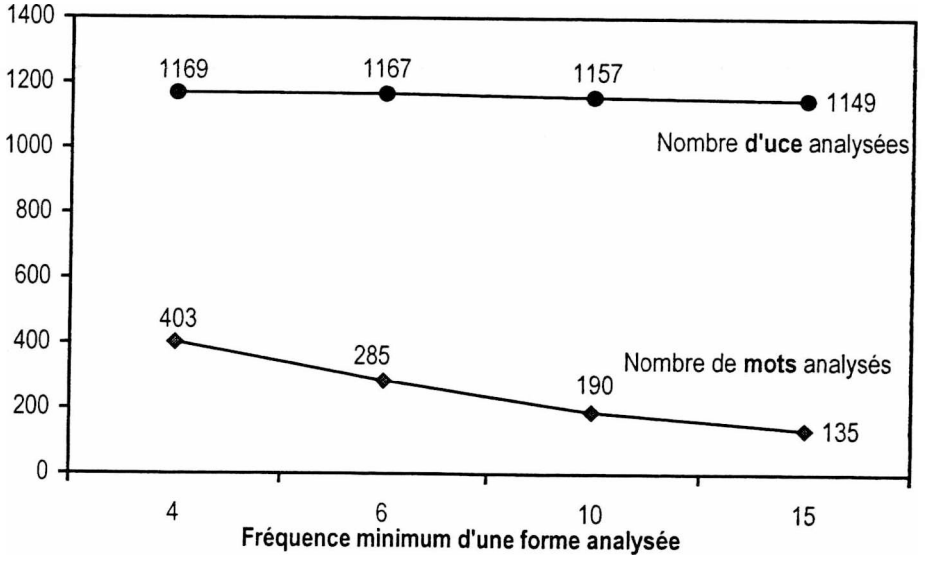
[1] On trouvera l'analyse détaillée dans « Le consommateur français en 1998 – une typologie des préférences », *CRÉDOC, Cahier de recherche* n°130, juin 1999.

« La perception du bonheur »

Caractéristiques suivant la variation du seuil de fréquence minimum d'une forme analysée

Caractéristique	Fréquence minimum finale d'une forme analysée			
	4	6	10	15
Nombre de mots analysés	403	285	190	135
Nombre de mots supplémentaires de type 'r'	143	123	104	88
Nombre total de mots	546	408	294	223
Nombre d'occurrences analysées	8 795	8 275	7 595	6 952
Nombre d'uce sélectionnées	1 169	1 167	1 157	1 149

Le tableau ci-dessus nous montre clairement les variations du nombre de mots et d'occurrences analysés en fonction de la fréquence minimale retenue pour conserver une forme. Il met bien en évidence, avec le graphe ci-dessous, la décroissance très nette du nombre de mots analysés en fonction du seuil fixé. Avec une hausse du seuil de fréquence des mots de l'ordre de 50%, on diminue d'environ un tiers le nombre de mots analysés. En revanche, le nombre d'uce analysées reste assez stable : il ne diminue que très légèrement.



Le nombre d'uce (unités de contexte élémentaire) constituant chacune des classes typologiques varie en fonction du seuil fixé. Néanmoins, la répartition des uce semble à peu près stable.

L'arbre de classification, en revanche, connaît des variations selon le seuil choisi, c'est-à-dire que le découpage en classes ne s'effectue pas dans le même ordre selon l'hypothèse. Par exemple, pour les seuils 4 et 6, la classe *« Loisirs, vacances, consommer, profiter »* se forme à la première étape de la classification, alors qu'avec les seuils 10 et 15 , la classe *« Loisirs, vacances, consommer, profiter »* reste d'abord assemblée à la classe *« Etre entouré, penser aux autres »*.

Dans les quatre typologies, on retrouve toujours une forte proximité entre les groupes *« Réussite globale de sa vie »* et *« Pas de souci, pas de problèmes »*.

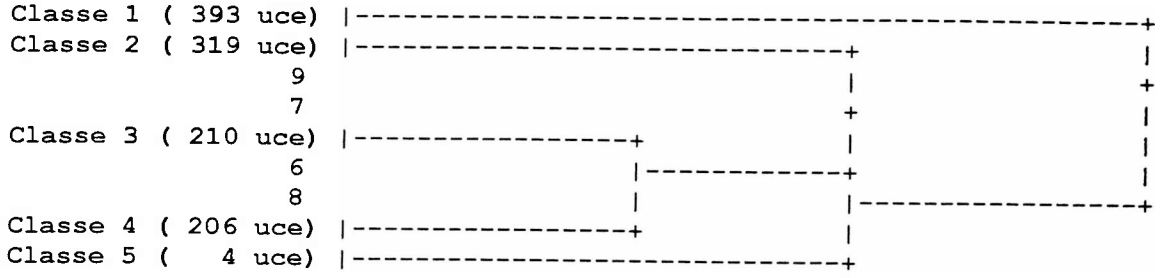
Cette comparaison montre que la classe dont la position est la moins stable est : *« Etre entouré, penser aux autres »* qui se trouve tantôt proche de la classe 1, tantôt proche des trois autres classes. Cette instabilité peut provenir du choix de la méthode de classification descendante qui privilégie la mise en évidence de dichotomies, alors que dans l'exemple étudié , trois grands types de réponses sont révélés par l'analyse.

Le premier bilan de cette comparaison est rassurant pour la nature des classes révélées, qui restent identiques dans les quatre essais, l'arbre de classification apparaît, en revanche, moins robuste et dépend du choix des paramètres initiaux de l'analyse.

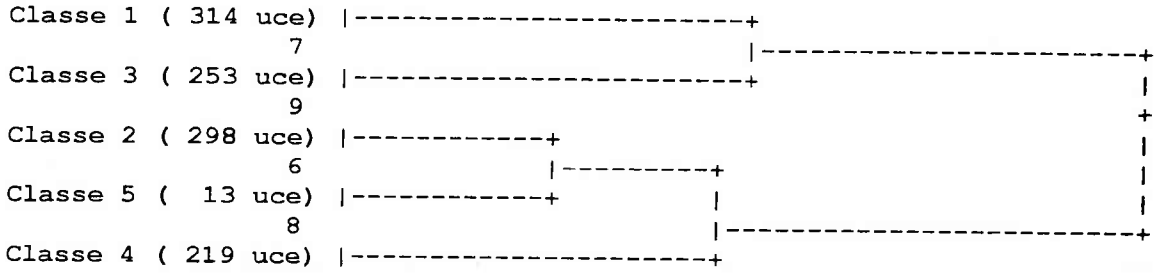
« La perception du bonheur »

Dendogrammes des classes stables selon le seuil de fréquence retenu

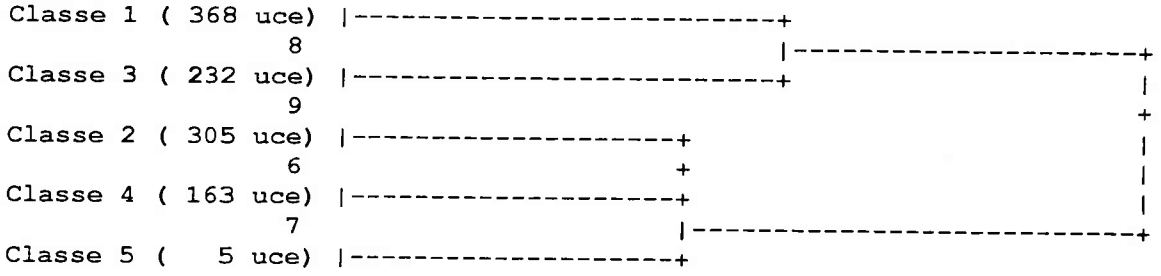
Seuil de fréquence≥ 6



Seuil de fréquence ≥10



Seuil de fréquence ≥15



« La perception du bonheur »

Nombre d'uce dans chacune des classes

Classe	Fréquence minimum finale d'une forme analysée							
	4		6		10		15	
Loisirs, vacances, consommer, profiter *(classe 1)*	396	*(34%)*	393	*(34%)*	314	*(27%)*	368	*(32%)*
Être entouré, penser aux autres *(classe 2)*	289	*(25%)*	319	*(27%)*	253 (=classe 3)	*(22%)*	232 (=classe 3)	*(20%)*
Réussite globale de sa vie *(classe 3)*	280	*(24%)*	210	*(18%)*	298 (=classe 2)	*(26%)*	305 (=classe 2)	*(26%)*
Pas de soucis, pas de problèmes *(classe 4)*	156	*(13%)*	206	*(18%)*	219	*(19%)*	163	*(14%)*
La santé, c'est le principal *(classe 5)*	4	*(-)*	4	*(-)*	13	*(1%)*	5	*(-)*
Non classées	44	*(4%)*	35	*(3%)*	60	*(5%)*	76	*(8%)*

Nota :

Le chiffre entre parenthèses correspond au pourcentage du nombre d'uce participant à la création de la classe par rapport au nombre total d'uce.

Les noms des classes en 1ère colonne correspondent à ceux donnés aux classes issues de la classification effectuée au seuil 4.

Quelles sont les répercussions sur la typologie ? Le vocabulaire spécifique de chaque classe mise en évidence en fonction du seuil de fréquence retenu est présenté dans les pages qui suivent.

On constate clairement une analogie très forte dans la définition des différentes classes. En effet, quel que soit le seuil de fréquence minimum des mots, les quatre grandes classes typologiques construites avec le plan d'analyse standard[1] se retrouvent complètement. Autrement dit, à partir d'un nombre restreint de mots analysés (mais à partir de mots dont la fréquence d'apparition dans le corpus est forte), on classifie aussi bien le discours des individus qu'en prenant pour base un corpus plus large (mais ayant un grand nombre de mots n'apparaissant que peu fréquemment). Cela revient à dire que **c'est sur les mots fréquents que se base l'analyse typologique effectuée à partir d'*Alceste* et que se forment les associations de mots et les noyaux durs**. Le choix de la méthode de classification descendante est sans doute pour beaucoup dans cette stabilité.

[1] On trouvera le détail de la typologie effectuée dans « Le consommateur français en 1998, une typologie des préférences », *CRÉDOC Cahier de recherche* n°130 , juin 1999.

Les mots s'associent à l'identique. La fréquence d'observation des mots, quant à elle, varie, mais reste du même ordre de grandeur quelle que soit la méthode d'analyse. On observe donc une très grande stabilité des résultats.

Pour la première classe décrite **« Loisirs, vacances, consommer, profiter »**, les quatre mêmes mots spécifiques arrivent en tête dans les différentes typologies. Il s'agit de *loisir+, pouvoir+, temps+* et *vacance+*; leurs fréquences de citations sont de plus tout à fait comparables.

En prenant en compte un plus grand nombre de mots caractéristiques et significatifs dans la constitution des classes, on retrouve bien une grande similitude entre dépouillements. Certes, le nombre de mots constitutifs des classes est largement inférieur quand le seuil est élevé, mais les caractérisants restent les mêmes.

Classe 1 : « Loisirs, vacances, consommer, profiter »
Citations de quelques mots spécifiques selon l'hypothèse de seuil retenue

Mot	Fréquence minimum finale d'une forme analysée			
	4	6	10	15
Loisir	82	82	78	89
Pouvoir	93	97	*89*	*109*
Temps	52	48	*40*	*50*
Vacance	87	87	67	73
Faire	109	111	81	110
Sport	15	15	13	14
Temps libre	21	21	21	23
Achat	12	21	23	23

Autrement dit, prendre un seuil plutôt bas permet d'affiner l'analyse, de souligner certaines nuances pouvant apparaître dans le groupe et, plus on prend un seuil élevé, plus la classe sera caractérisée par un petit nombre de mots, les plus significatifs mais aussi les plus basiques.

L'analyse de la classe 2 « **Être entouré, penser aux autres** » amène aux mêmes remarques. Ici, c'est *heur+eux, petits-enfants* et *monde+* qui prennent la tête des suffrages, et ce, quel que soit le seuil retenu.

Classe 2 : « Être entouré, penser aux autres »

Citations de quelques mots spécifiques selon l'hypothèse de seuil retenue

Mot	Fréquence minimum finale d'une forme analysée			
	4	6	10	15
Heureux	65	76	67	65
Petits-enfants	27	24	24	28
Monde	50	49	28	33
Mes enfants	49	51	44	57
Ma famille	27	24	22	23
Chômage	27	29	17	21
Je suis	29	34	39	39

Cette convergence traduit bien une réelle stabilité de la typologie obtenue, car même en ne prenant en compte dans l'analyse que 285 mots, on arrive à constituer des classes qui regroupent les mêmes unités de contexte et donc un vocabulaire spécifique identique.

La classe 3 « **Réussite globale de sa vie** » met en avant en première position des mots différents selon l'hypothèse de seuil fixé :

- Ainsi, avec un seuil de fréquence des mots de 4, le mot *santé+* prend la tête, avec 117 citations.
- Quand le seuil est de 6, c'est *famille+* et *santé+* qui ressortent en premier.
- Avec le seuil à 10, *santé+* est le plus significatif des vocables de cette classe (138 citations).
- Enfin, avec un seuil de 15, le mot *argent+* augmente nettement sa fréquence (111 citations).

Néanmoins, ces quatre mots, avec le mot *travail<*, sont tous spécifiques du vocabulaire de cette classe. Et globalement, les quatre classes ainsi constituées sont très proches en terme de vocabulaire caractéristique (cf. tableau ci-après).

Classe 3 : « Réussite globale de sa vie »

Citations de quelques mots spécifiques selon l'hypothèse de seuil retenue

Mot	Fréquence minimum finale d'une forme analysée			
	4	6	10	15
Famille	113	96	100	99
Santé	117	96	*138*	*136*
Travail	91	70	*90*	*90*
Amour	47	40	41	38
Argent	88	87	106	111
Couple	22	19	*	25
Stable	14	*	19	20
Réussite	9	9	*	*

Quelques nuances apparaissent néanmoins entre les quatre dépouillements :

- Avec les deux premiers seuils (4 et 6), les mots *famille* et *travail* semblent avoir un poids plus important. On y retrouve aussi la notion de *couple* et de *réussite*. On recherche donc ici avant tout la réussite familiale et professionnelle.
- Dans les typologies constituées avec les deux seuils les plus élevés, ce sont plutôt les mots *argent* et *santé* qui ont un rôle plus fort, de même que *stable*. Cela traduit un besoin de se retrouver soi-même, de dominer les éléments extérieurs, de maîtriser sa vie et ce, pas spécifiquement dans le cadre familial.

Ce phénomène est probablement dû au fait que le pôle famille-travail est défavorisé par l'élimination de mots peu fréquents, alors que les notions de « *santé* » et d'« *argent* » seraient plus isolées et ne souffriraient donc pas de ce phénomène.

Enfin, dans les « quatrièmes classes » qui sont constituées de vocabulaire connoté négativement **« Pas de soucis, pas de problèmes »,** on retrouve très fréquemment les mots ou expressions *ne pas*, *souci*, *problème*. La conception du bonheur des individus présents dans ces groupes s'organise autour du « non problème » : à partir du moment où l'on n'a pas de souci, on peut considérer que l'on est heureux.

Classe 4 : « Pas de soucis, pas de problèmes »

Citations de quelques mots spécifiques selon l'hypothèse de seuil retenue

Mot	Fréquence minimum finale d'une forme analysée			
	4	6	10	15
Ne pas	65	67	67	41
Problème	32	36	*	*
Financier	29	34	22	*
Souci	51	52	35	22

Plus on accroît le seuil de fréquence des mots, plus on a l'impression que la notion de souci diminue pour passer à une notion plus positive de *« être bien, se sentir bien, dans sa peau, dans sa tête, dans sa vie, ... »*. Le mot ***souci+*** baisse en effet en nombre de citations quand on augmente le seuil, alors que ***être bien***, par exemple, est cité plus de 40 fois dans les deux dernières typologies. On se situe donc davantage sur ce qu'on pourrait avoir ou être que sur ce qu'on ne voudrait pas avoir ou être.

Néanmoins, les mots ***souci***, ***ne pas***, ***financier***, ... ; ressortent toujours très clairement même quand les seuils sont élevés ; ce qui ne fait en quelque sorte que confirmer la stabilité de la typologie.

➲ **Sélection des formes réduites spécifiques en fonction du seuil de fréquence retenu**

Loisirs, vacances, consommer, profiter

Seuil de fréquence >4

loisir+(82), pouvoir+(93), temps(52), vacance+(87), faire.(109), maison+(36), soleil+(25), sport+(15), voiture+(16), gagn+er(22), temps-libre(21), belle+(13), achat+(12), boulot+(9), budget+(7), île+(6), liberté+(16), mer+(8), moyen+(17), musique+(8), plage+(6), plaisir+(22), voyage+(25), achet+er(22), mang+er(17), offrir.(10), partir.(12), permettre.(14), voyag+er(15), vie de famille(17), loto(9), détente+(5), fille+(10), joie+(14), dépendre.(5), pa+yer(25), retrouv+er(6), désert+ion(4), envi+e(27), montagn+e(5), prés+ent(6), cocotiers(5), stress+(9), cher+(5), impôt+(6), retraite+(10), soir+(3), consacr+er(5), gér+er(5), profit+er(19), respect+er(3), télé(5), correct+(14), étranger+(4), France(7), avenir+(8), bateau+(4), dépense+(4), fin+(6), continu+er(4), promen+er(4), act+ion(8), superflu<(4);

Seuil de fréquence >6

loisir+(82), pouvoir+(97), temps(48), vacance+(87), faire.(111), soleil+(27), maison+(37), plaisir+(22), sport+(15), voiture+(16), voyage+(29), gagn+er(22), voyag+er(16), temps-libre(21), belle+(14), content+(9), boulot+(9), budget+(7), fin+(8), ile+(6), mer+(8), mois(11), moment+(9), musique+(8), plage+(6), achet+er(21), offrir.(10), partir.(13), permettre.(14), vie-de-famille(18), loto(9), achat+(11), avenir+(9), contrainte+(7), détente+(5), impôt+(7), liberté+(15), maladie+(14), région+(6), salaire+(8), compt+er(9), pa+yer(27), prendre.(15), retrouv+er(6), être-en-famille(9), envi+e(27), cher+(5), simple+(6), actuellement(5), moyen+(14), consacr+er(5), gér+er(5), prés+ent(6), rêv+e(5), vieill<(5), beau+(6), étranger+(4), jour+(9), occup+er(10), sortir.(12), act+ion(8), déc+ent(5), stress+(8);

Seuil de fréquence >10

loisir+(78), pouvoir+(89), vacance+(67), temps(40), maison+(35), soleil+(24), achet+er(23), faire.(81), temps-libre(21), sport+(13), voyage+(24), gagn+er(19), offrir.(11), permettre.(15), vie-de-famille(17), belle+(12), correct+(16), achat+(12), boulot+(8), étude+(11), joie+(13), voiture+(12), pa+yer(24), vivre.(69), voyag+er(12), act+ion(10), loto(8), agréable+(9), liberté+(12), niveau+(20), partir.(10), profit+er(18), campagne+(6), épou+x(8), métier+(7), retraite+(8), sortir.(11), enf+ant(62), envi+e(22), plaire.(9), riche+(6).

Seuil de fréquence >15

loisir+(89), pouvoir+(109), temps(50), vacance+(73), faire.(110), vivre.(88), temps-libre(23), correct+(21), soleil+(22), sport+(14), voyage+(29), achet+er(23), envi+e(33), achat+(12), joie+(14), liberté+(16), voiture+(15), permettre.(15), profit+er(22), voyag+er(14), vie-de-famille(17), act+ion(11), libre+(16), étude+(11), maison+(30), maladie+(14), moyen+(16), plaisir+(18), retraite+(10), gagn+er(14), pa+yer(26), sortir.(14), agréable+(8), niveau+(21), être-en-bonne-santé(41), suffis+ant(16), jour+(9), partir.(9);

Être entouré, penser aux autres

Seuil de fréquence >4

heur+eux(65), petits-enfants(27), monde+(50), mes enfants(49), ma famille(27), chômage(27), épou+x(13), guerre+(13), voir.(44), je suis(29), age+(8), malheur+eux(12), gens(27), paix(18), bonheur+(39), esprit+(7), état+(7), misère+(12), parent+(12), autour de moi(9), en bonne santé(22), mon mari(15), dire+(22), personne+(7), porte+(6), aid+er(6), demand+er(6), exist+er(4), autour de soi(10), je ne(9), import+ant(11), viol+ent(5), insécurité(4), besoin+(18), sens(6), donn+er(5), prendre.(11), médica<(3), vieill<(4), ami+(27), société+(5), occup+er(8).

Seuil de fréquence >6

heur+eux(76), monde+(49), voir.(51), je-suis(34), mes enfants(51), chômage+(29), petits-enfants(24), age+(8), malheur+eux(13), bonheur+(46), épou+x(13), gens(29), guerre+(13), misère+(16), ma famille(24), besoin+(24), dire+(26), esprit+(8), état+(7), forme+(5), nature+(11), personne+(9), savoir+(9), donn+er(7), exist+er(5), rendre.(9), sentir.(13), autour-de-soi(14), en-bonne-santé(23), je ne(11), petit+(18), possi+ble(11), difficile+(8), aide+(5), chose+(27), difficulte+(6), ensemble+(9), logement+(10), parent+(10), société+(6), aid+er(6), aller.(37), mang+er(13), autour-de-moi(8), import+ant(12), polit+3(5), continu+(6), ennui+(5), fête+(7), essa+yer(5), regard+er(6), rest+er(9), jeune+(9), viol+ent(5), optimiste+(4), plein+(5), porte+(5), mon mari(13).

Seuil de fréquence >10

heur+eux(67), je-suis(39), petits-enfants(24), prendre.(19), je ne(15), mes enfants(44), donn+er(9), voir.(42), difficile+(10), chose+(29), dire+(27), moment+(10), ma famille(22), cote+(10), possi+ble(12), beau+(7), seul+(11), fête+(8), guerre+(9), monde+(28), nature+(10), plaisir+(16), occup+er(10), rendre.(8), autour-de-soi(11), être-heureux(22), chômage+(17), petit+(15), gros+(6), France(7), jour+(8), sens(6), arriv+er(8), mang+er(12), regard+er(6), fait(17), bon+(6), content+(5), maladie+(10), import+ant(9), avenir+(6), manque+(5), personne+(5), savoir+(6), rencontr+er(5), autour-de-moi(6), mon mari(11);

Seuil de fréquence >15

heur+eux(65), petits-enfants(28), je-suis(39), mes enfants(57), voir.(41), je ne(14), ma famille(23), dire+(26), épou+x(11), monde+(33), nature+(10), en-bonne-santé(21), chômage+(21), bonheur+(34), chose+(23), autour-de-soi(11), possi+ble(9), occup+er(9), rest+er(9), petit+(12), grand+(9), prendre.(10), besoin+(15), trouv+er(7), import+ant(8);

Réussite globale de sa vie

Seuil de fréquence >4

famille+(113), santé+(117), travail<(91), amour+(47), couple+(22), être bien(36), sentimenta+l(6), socia+l(6), amitié+(9), argent(88), emploi+(15), marche+(6), relation+(12), réussite+(9), situation+(10), épanou+ir(13), réuss+ir(12), vie familiale(12), affect+ion(7), bonne+(27), professionn+el(16), revenu+(11), sta+ble(14), intéressant+(4), scolaire+(4), confort+(6), contribu+er(4), dépens+er(7), être en bonne santé(34), être heureux(20), ambian<(4), autonom<(4), enf+ant(58), harmoni+e(10), humain+(4), cadre+(3), carrière+(3), foyer+(10), projet+(5), vie+(53), aim+er(16), comprendre.(3), assur<(3), mari+(4), étude+(8).

Seuil de fréquence >6

amour+(40), argent(87), famille+(96), couple+(19), santé+(96), réussite+(9), socia+l(5), amitié+(9), relation+(11), situation+(8), dans-son(10), professionn+el(12), travail<(70), humain+(4), homme+(4), marche+(4), mari+(7), paix(12), vie-familiale(8), enf+ant(48), mari+_(4), emploi+(9), dépens+er(5), réuss+ir(8), être-en-bonne-santé(25), ami+(21), foyer+(8).

Seuil de fréquence >10

santé+(138), amour+(41), sta+ble(19), argent(106), emploi+(18), bonne+(35), entente+(14), famille+(100), sécurité+(16), familia+l(24), amitié+(10), bonheur+(41), foyer+(13), paix(16), relation+(12), situation+(12), vue+(10), entendre.(16), manqu+er(15), travail<(90), mari+(10), souci(15), problem<(31), malade+(22), phys+3(6)

Seuil de fréquence >15

argent(111), santé+(136), souci(24), sta+ble(20), familia+l(32), financier+(37), couple+(25), vie-familiale(15), bonne+(34), amour+(38), sécurité+(16), problem<(38), malade+(27), emploi+(16), entente+(14), famille+(99), situation+(11), mari+(10), manqu+er(14), avoir-un-travail(13), travail<(90), foyer+(11), entendre.(13), enf+ant(59)

Pas de soucis, pas de problèmes

Seuil de fréquence >4

ne pas(65), problem<(32), financier+(29), souci+(51), bonne santé(24), entente+(13), être-bien-dans-sa-peau(19), familia+l(18), malade+(17), personnel+(5), quotidien+(6), contact+(4), équilibre+(7), finance+(2), vue+(6), content+er(6), manqu+er(9), priv+er(6), vivre.(41), confi+ant(3), sécurité+(8), entendre.(9), sérénité(3), toit<(6), tranquil+e(9), négati+f(2), environnement(5), mari+(5), envi+er(2), gard+er(4).

Seuil de fréquence >6

familia+l(28), financier+(34), souci+(52), problem<(36), manqu+er(17), ne-pas(67), pos+er(7), bonne-santé(23), être-bien(30), confort+(8), entente+(12), équilibre+(11), tête+(13), bonne+(23), malade+(20), quotidien+(8), environnement(8), peau+(9), question+(8), entour+er(13), être-bien-dans-sa-peau(14), ses-enfants(14), harmoni+e(9), matéri+el(14), sta+ble(10), tranquil+e(14), personnel+(5), sain+(5), contact+(4), sécurité+(9), vue+(6), priv+er(6), réalis+er(6), vivre.(42), affect+ion(5), bien-être(7), épanou+ir(8), minimum(6), projet+(4), qualité+(3), vie+(39), consommat+ion(3), entourage<(6);

Seuil de fréquence >10

souci+(35), tête+(19), épanou+ir(17), être-bien(43), être-bien-dans-sa-peau(25), professionn+el(21), confort+(10), entour+er(19), dans-son(16), ne-pas(67), couple+(18), financier+(22), ami+(29), équilibre+(10), gens(22), projet+(6), content+er(8), réalis+er(7), bonne-santé(20), ses-enfants(14), vie-familiale(10), quotidien+(6), minimum(7), misère+(10), vie+(47), priv+er(6), rest+er(8), réuss+ir(9), sentir.(9), tranquil+e(13), voisin<(6), malheur+eux(7), toit<(8), matéri+el(10);

Seuil de fréquence >15

tête+(19), entour+er(20), sentir.(16), être-bien(41), être-bien-dans-sa-peau(23), professionn+el(18), dans-son(14), ami+(28), misère+(12), peau+(10), souci+(22), ses-enfants(13), environnement(7), gens(17), toit<(8), réuss+ir(9), malheur+eux(7), ensemble+(6), équilibre+(7), vie+(36), aim+er(12), épanou+ir(7), bonne-santé(13), ne-pas(41), entourage<(6), matéri+el(8);

En conclusion, l'analyse de ce premier corpus nous amène à plusieurs enseignements :

- Le nombre de mots retenu pour l'analyse diminue rapidement quand le seuil de sélection des mots s'accroît, mais cela n'empêche pas de constituer des classes stables.

- L'arbre de classification diffère selon le seuil de fréquence des mots retenu : la constitution des groupes ne s'opère pas dans le même ordre. Malgré tout, on constate que cela ne change pas le regroupement des mots en fonction de leur proximité et que, malgré des nuances d'ordre et de fréquence de citations, la classification obtenue est quasi identique selon l'hypothèse choisie.

On peut donc considérer que les typologies obtenues sont robustes et que l'impact du seuil de sélection sur celles-ci est faible. En effet, les classes sont constituées à partir des uce contenant les mots les plus fréquents, en quelque sorte un « noyau dur » de vocabulaire, un noyau robuste qui ne modifiera pas le fondement des corrélations.

II.2. LE CORPUS « LES PRÉFÉRENCES DES FRANÇAIS »

Afin de valider les résultats obtenus ci-dessus, nous allons maintenant procéder de la même façon que précédemment sur un autre corpus, celui des « préférences des Français ». Plus riche en vocabulaire, 3 821 mots et 31 329 occurrences, ce corpus possède un plus grand nombre d'hapax (le nombre de mots n'apparaissant qu'une seule fois s'élève à 1 880) et, au total, un nombre assez comparable au précédent d'unités analysées.

Caractéristiques du corpus analysé avant tout traitement

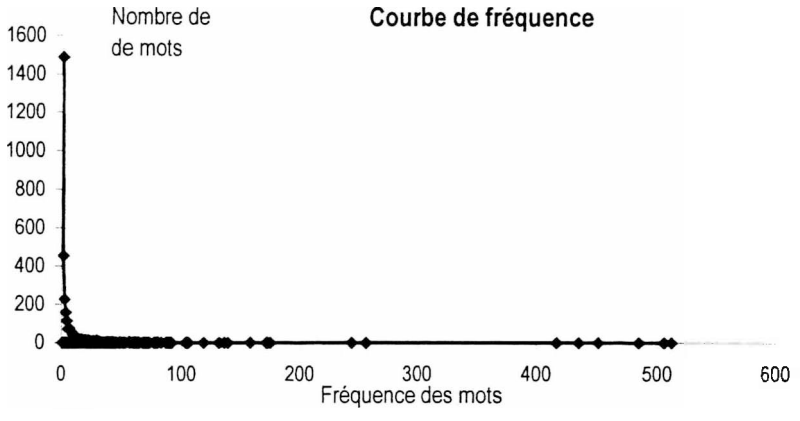
Nombre de formes distinctes (étendue du vocabulaire)	3 821
Nombre d'occurrences (étendue du texte)	31 329
Fréquence moyenne par forme	8
Nombre d'hapax (nombre de mots n'apparaissant qu'une seule fois)	1 880
Proportion d'hapax	49 % de l'ensemble des formes
Nombre d'uce enregistrées	1 074
Nombre d'uce classées	955, soit 89%

Quel est l'impact, sur l'analyse lexicale de ce corpus, de la variation du seuil de fréquence des mots ?

➲ Distribution des fréquences

Nous allons tout d'abord dessiner la distribution des fréquences de mots. Comme pour le corpus sur « La perception du bonheur », la gamme des fréquences est représentée par une courbe très ramassée, en forme de L. Elle met bien en évidence le fait que les formes les plus

fréquentes représentent une faible proportion du vocabulaire du corpus ; à l'inverse, les basses fréquences correspondent à des effectifs élevés. Néanmoins, en ne conservant que les fréquences supérieures à 50, on arrive à couvrir 50% des occurrences avec seulement 2% du vocabulaire.



Exemple de lecture : 1488 mots apparaissent une seule fois dans le corpus ; à l'inverse, un mot apparaît 512 fois.

Il n'est cependant pas nécessaire de réaliser une réduction aussi drastique du vocabulaire, qui aurait comme inconvénient d'éliminer trop de réponses de l'analyse.

Avec le logiciel *Alceste*, on fixe un seuil minimal au-delà duquel on conserve les vocables. On effectue donc par là une réduction importante du vocabulaire du texte. Dans les plans d'analyse standard, le seuil est fixé à 4 ; nous allons maintenant tester sur ce corpus la stabilité des résultats alors même que l'on accroît ce seuil et donc que l'on réduit de manière conséquente le pourcentage du vocabulaire analysé.

➲ Tests de différents seuils de fréquence

Nous avons testé plusieurs hypothèses de seuil minimum de fréquence : les quatre seuils de 4, 6, 10 et 15 ont été retenus pour l'analyse comparative.

« Les préférences des Français »

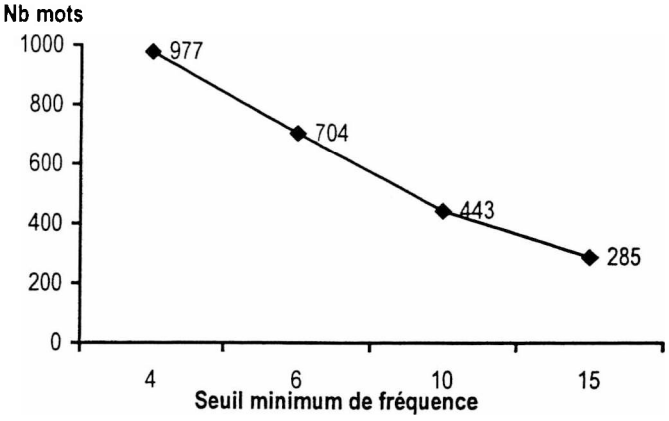
Caractéristiques suivant la variation du seuil de fréquence minimum d'une forme analysée

Caractéristique	Fréquence minimum finale d'une forme analysée			
	4	6	10	15
Nombre de mots analysés	977	704	443	285
Nombre de mots supplémentaires de type 'r'	84	72	52	37
Nombre total de mots	1 061	776	495	322
Nombre d'occurrences analysées	17 769	16 562	14 637	12 800
Nombre d'uce sélectionnées	1 076	1 074	1 069	1 064

Le *nombre de mots analysés* décroît très sensiblement en fonction du seuil fixé, passant de 977 (avec un seuil de 4) à 285 (avec un seuil de 15). Pour un seuil environ 50% plus élevé, on perd environ un tiers du nombre de mots.

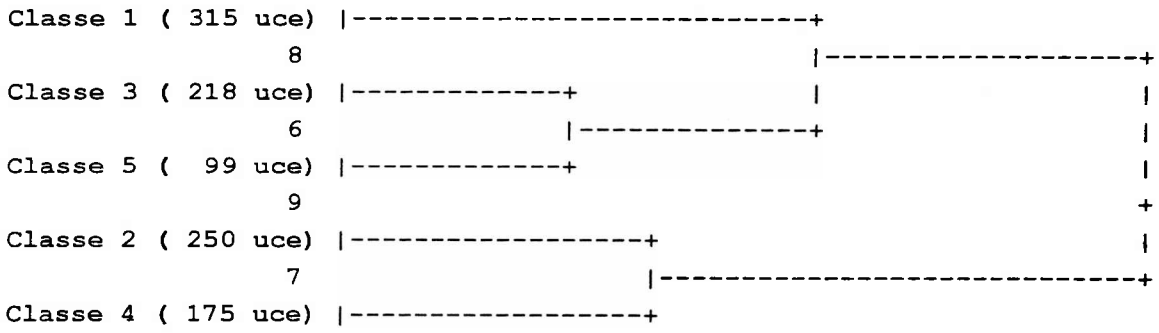
Le *nombre d'occurrences analysées* se réduit également quand le seuil minimal augmente ; mais il diminue dans des proportions moindres (on passe d'environ 17 800 à 12 800, baisse d'un peu moins de 30%, entre les seuils 4 et 15).

Le *nombre d'uce sélectionnées*, en revanche, reste quasiment identique quel que soit le seuil fixé (environ 1 070 uce retenues). Autrement dit, en accroissant le seuil de sélection dans la fourchette prévue, on réduit le vocabulaire véritablement analysé, mais on conserve pratiquement tout le texte, c'est-à-dire, dans notre exemple, les réponses de l'ensemble des individus interrogés.



➲ Les typologies obtenues

Initialement, l'analyse effectuée, qui correspond à une analyse standard, a pris en compte le seuil de 4. Les différentes classes retenues d'après cette analyse sont présentées et détaillées dans le Cahier de recherche sur la Consommation 1998. Elles sont au nombre de 5 et l'arbre de classification dont elles proviennent est le suivant :



Classe 1 : les femmes actuelles	(24% des individus). Cette classe rassemble	315 uce
Classe 2 : les hédonistes modernes	(20% des individus)	250 uce
Classe 3 : les âgés traditionalistes	(20% des individus)	218 uce
Classe 4 : les intellectuels bons vivants	(19% des individus)	175 uce
Classe 5 : les rustiques	(15% des individus)	99 uce

Pour chaque hypothèse de seuil de fréquence, nous avons également effectué des classifications descendantes hiérarchiques (en 5 classes afin de faciliter les comparaisons). Assez logiquement, avec un seuil de fréquence plus élevé, on diminue le nombre de mots analysés et les fréquences des mots peuvent varier assez sensiblement d'une classification à l'autre (cf. tableau suivant). Ceci a alors une répercussion directe sur le pourcentage d'unités classées et les rapprochements d'unités.

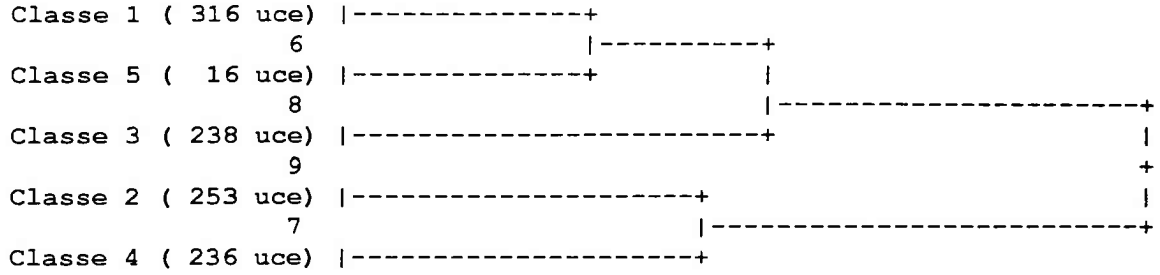
D'ailleurs, les quatre dendogrammes issus de cette opération sont différents. La formation des classes ne se fait pas dans le même ordre selon l'hypothèse du seuil retenue ; c'est le cas notamment pour la classe 5 qui n'est pas issue des mêmes groupes. Remarquons que c'est elle qui rassemble à chaque fois le plus petit nombre d'unités de contexte.

En effet, quand le seuil est de 4, le groupe ressemblant aux « Âgés traditionalistes » se partage en deux classes, la classe 5 « Les rustiques » et la classe 3, les purs « Âgés traditionalistes ». Pour le seuil à 6, la classe 5 provient d'une segmentation de la classe 2, « Les hédonistes modernes » en deux sous-groupes, qui sont, certes, de taille très différente. Il en est de même pour le seuil fixé à 15. En revanche, avec une limite de fréquence des mots à 10, on constate que c'est la classe 5 et celle des « Intellectuels bons vivants » qui étaient regroupées à l'étape précédente, c'est-à-dire lors de la typologie en quatre classes.

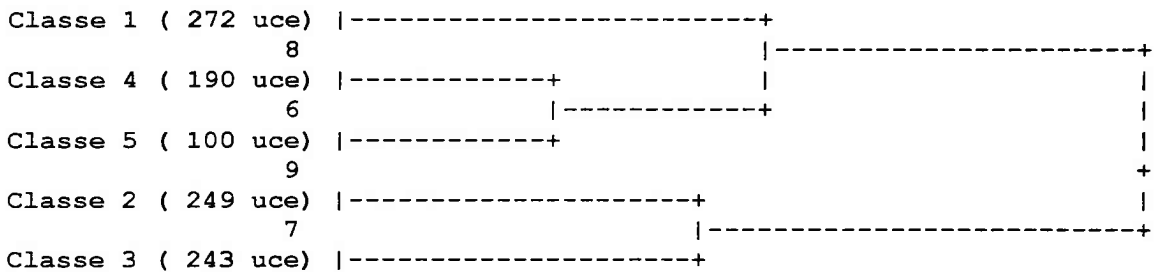
Autrement dit, les arbres de classification diffèrent d'un dépouillement à l'autre ; les différences se fondent essentiellement sur la classe la plus petite en taille, les groupes les plus importants en nombre d'uce restant stables. Cette petite classe 5 se détache une fois du groupe « Âgés traditionalistes », deux fois du groupe « Hédonistes modernes » et une fois du groupe « Intellectuels bons vivants ».

Dendogrammes des classes stables selon le seuil de fréquence retenu

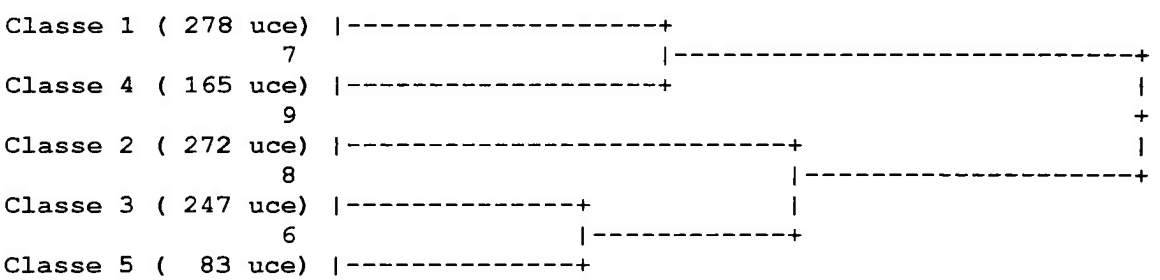
Seuil de fréquence ≥ 6



Seuil de fréquence ≥ 10



Seuil de fréquence ≥ 15



« Les préférences des Français »

Nombre d'uce dans chacune des classes

Classe	Fréquence minimum finale d'une forme analysée							
	4		6		10		15	
Les femmes actuelles	315	(Classe1)	236	(Cl.4)	243	(Cl.3)	165	(Cl.4)
Les hédonistes modernes	250	(Classe2)	**316**	**(Cl.1)**	272	(Cl.1)	**247**	**(Cl.3)**
Les âgés traditionalistes	**218**	**(Classe3)**	253	(Cl.2)	249	(Cl.2)	278	(Cl.1)
Les intellectuels bons vivants	175	(Classe4)	238	(Cl.3)	**190**	**(Cl.4)**	272	(Cl.2)
Classe 5	**99**	**(Classe5)**	**16**	**(Cl.5)**	**100**	**(Cl.5)**	**83**	**(Cl.5)**

Néanmoins, malgré les différences de regroupement des uce, les quatre premiers groupes mis en évidence dans l'enquête Consommation 1998 se retrouvent bien —en terme de citations de mots— dans les autres dépouillements[1]. Seule la cinquième classe n'apparaît pas à l'identique quand on accroît le seuil de fréquence. L'arbre de classification n'est donc pas stable, même si certaines régularités apparaissent. Les classes « Femmes actuelles » et « Âgés traditionalistes » sont toujours séparées des autres au premier niveau de segmentation. Il est vrai que la classe des personnes âgées est plutôt féminine. Parallèlement, les classes « Hédonistes modernes » et « Intellectuels bons vivants » sont toujours associées entre elles.

La classification

➲ Classe 1 : « Les femmes actuelles »

Cette classe n'apparaît pas en tête de chacune des analyses, mais elle se retrouve à chaque fois avec une grande similarité des mots employés. Certes, ces mots n'apparaissent pas avec la même fréquence, ni dans le même ordre ; néanmoins, on peut constater que les principaux vecteurs qui définissent ce groupe y sont bien présents.

Les mots *femme, bijou, actuel, parfum, lecture, patinage-artistique, cristal, porcelaine ...* sont bien spécifiques à ce groupe et, ce, dans chacune des hypothèses de seuil retenue. De plus, dans chaque analyse, on retrouve le profil type des « Femmes actuelles », c'est-à-dire des femmes, des employés, disposant de revenus moyens ou faibles. Elles sont également assez jeunes.

[1] On trouvera plus loin une sélection du vocabulaire spécifique à chacune des classes en fonction du seuil de fréquence retenu.

Classe 1 : « Les femmes actuelles »

Citations de quelques mots spécifiques selon l'hypothèse de seuil retenue

Mot	Fréquence minimum finale d'une forme analysée			
	4	6	10	15
Bijou	55	38	42	37
Femme	44	50	52	41
Actuel	41	47	49	37
Parfum	30	24	24	15
Lecture	67	58	64	42
Naf-Naf	19	13	19	17
Natation	47	33	40	24
Rose	165	119	123	
Cristal	27	21	19	18
Gymnastique	20			8
Porcelaine	33	30	25	25
Lara Fabian	10	8	9	
Mousse au chocolat	36	27	31	18
Patinage artistique	27	25	25	20
Décoratif+	8	8		
Fleur	31	31	32	
Étain	17	18	16	17
Ile+	18	20	18	
Louis XIV	18	16	17	

On constate assez logiquement que plus le seuil de fréquence s'accroît, plus le nombre de mots spécifiques à la classe diminue. Ainsi, par exemple, le mot *rose+* ne ressort pas comme spécifique de la classe dans la dernière analyse typologique au seuil de 15. Il en est de même pour *Lara Fabian*, *décoratif+* ou encore *fleur+*, *île+*, etc.

➲ **Classe 2 : « Les hédonistes modernes »**

Moto, foot, voiture, bois, plat de pâtes, l'équipe, informatique, Levis, Francis Cabrel, Lacoste, etc. sont les mots les plus caractéristiques de cette classe. On retrouve ces vocables dans les quatre typologies, même si ils ne sont pas hiérarchisés de façon identique. *Steak-frites* et *Etats-Unis* apparaissent également parmi leurs préférences.

Ce vocabulaire est spécifique d'une population masculine, plutôt jeune (moins de 35 ans), plus souvent étudiant ou ouvrier et ayant des revenus somme toute assez modestes.

Classe 2 : « Les hédonistes modernes »

Citations de quelques mots spécifiques selon l'hypothèse de seuil retenue

Mot	Fréquence minimum finale d'une forme analysée			
	4	6	10	15
Bois	164	184		154
Pâte	39	37	*34*	*23*
Ferrari	20		*16*	*15*
Moto	22	18	25	24
Levis	47	52		42
Chêne	40		31	36
Équipe	13	13	11	12
Tigre	11	9		
Foot	48	87	89	65
Lacoste	17	18	19	17
Nul part ailleurs	15	14		
Phil Collins	10	8		
Strauss	10	11	11	
Informatique	18	8	10	8
Australie	17	17	15	
Parisien	14	15	15	18
Voiture	24	35	32	32
Francis Cabrel	20	33	24	11
Pêche	18	25	26	29
Auto		14	14	
Chien	8		152	149
Rugby			22	23
Noir		45		31

➲ Classe 3 : « Les âgés traditionalistes »

Cette classe se caractérise par des préférences comme : *Citroën, marche, fruit, Tino Rossi, Frédéric François, l'émission 'Questions pour un champion', la France, de Gaulle*, etc.

Tous ces mots reviennent en nombre important dans chacune des typologies, constituant un noyau dur identique, malgré l'ordre hiérarchique différent. Cela crée de fait des classes qui se ressemblent fortement et qui expriment les mêmes préférences.

Les variations observées d'un dépouillement à l'autre n'entraînent pas néanmoins de modification profonde de la signification de la classe.

Classe 3 : « Les âgés traditionalistes »

Citations de quelques mots spécifiques selon l'hypothèse de seuil retenue

Mot	Fréquence minimum finale d'une forme analysée			
	4	6	10	15
Fruit	49	44	36	44
Marche	46	47	50	55
Citroën	26	34	34	29
Questions pour un champion	24	25	20	21
Tino Rossi	14	17	17	17
Xm	10	9	10	
France	72	72	76	93
Italie	25			28
Légume	18	18	14	19
Raisin	9	8		
Tableau	13	14	12	
De Gaulle	59	62	64	
Frédéric François	16	14	24	24
Midi	24	30	29	27
Pomme		24	28	28
Jardiner		23	26	22
Pétanque	10	12	11	11
Républicain		16	16	13
Bouquet de fleurs	14	12	12	16
Michel Sardou	36	45	40	49

➲ Classe 4 : « Les intellectuels bons vivants »

On retrouve également dans cette classe des mots identiques quel que soit le seuil. C'est le cas notamment de *Monde, Canard, Jaguar*, ... Ces mots sont caractéristiques d'une certaine aisance de vie et du souhait de bien vivre.

Certains liens apparaissent néanmoins uniquement quand les seuils de fréquence sont élevés. Des mots comme *musique, Brel, randonnée, l'émission 'Faut pas rêver'*, par exemple, ne ressortent que dans les trois classifications au seuil élevé (6 à 15).

Autrement dit, prendre en compte davantage de mots dans l'analyse typologique permet d'une certaine façon d'affiner la description des classes, mais cela ne change pas pour autant la base de constitution des groupes. On obtient peut-être plus de précision et de nuances pour comprendre les corrélations.

Classe 4 : « Les intellectuels bons vivants »

Citations de quelques mots spécifiques selon l'hypothèse de seuil retenue

Mot	Fréquence minimum finale d'une forme analysée			
	4	6	10	15
Monde	25	30	22	39
Yves St Laurent	10	11	*10*	*14*
Baba au rhum	8			
Canard	18	2	19	23
St Honoré	7			
Marianne	9	9	8	
Inde	4			
Art	8	9	6	
Jaguar	13	19	13	18
Vin	13	17	11	13
Voile	10	9		12
Voyage	8	9		36
Golf	15	25	21	19
Bouteille	9	14		13
Musique		24	20	23
Thalassa	30	38	31	36

Mot	Fréquence minimum finale d'une forme analysée			
	4	6	10	15
Gâteau au chocolat	14	16	15	18
Ile flottante		20	19	
Bricoler	18	27	28	26
Beethoven	7	7	9	
Brel		25	27	27
Randonnée			16	18
Faut pas rêver		15	18	20

➲ **Classe 5 : la dernière classe constituée**

Cette cinquième classe apparaît différente d'un dépouillement à l'autre. Certes, on retrouve certains mots communs comme *tarte aux pommes, Ouest-France, télévision, Johnny Hallyday, vert*, mais finalement assez peu.

D'ailleurs, la classe obtenue avec le seuil de fréquence 10 est 'unique' d'un point de vue du vocabulaire : quasiment aucun mot n'est commun avec les groupes des autres dépouillements. En revanche, elle a de fortes similarités avec la classe des « intellectuels bons vivants ».

Autrement dit, cette dernière classe, la plus petite en nombre d'uce la constituant, ne définit pas les mêmes profils lexicaux, ni même d'ailleurs socio-démographiques. On ne peut donc pas parler de stabilité à son égard. Cette instabilité est importante car il s'agit de la seule classe à tendance ouvrière ressortant de l'analyse. Plutôt jeune, masculine et provinciale, ses contours sont assez fluctuants. Cette difficulté à se constituer tranche par rapport à l'autre classe de préférences à forte typicité sociale, celle des « intellectuels bons vivants ».

Classe 5 : Caractéristiques socio-démographiques caractéristiques selon l'hypothèse de seuil retenue

Caractéristiques	Fréquence minimum finale d'une forme analysée			
	4	6	10	15
Sexe		Homme	Homme	Homme
Profession	Ouvrier	Ouvrier	*Étudiant*	Ouvrier
	Retraité		Profession intermédiaire	
Âge	25-34 ans	25-44 ans	18-34 ans	25-44 ans
	65 ans et plus			
Région de résidence	Région Est	Bassin Parisien		Région Ouest
	Région Centre Est	Région Méditerranée		Région Sud-Ouest
Taille d'agglomération de résidence	Milieu rural	Milieu rural		
Revenus mensuels du foyer	< 8 000 F	8 à 24 000 F	12 à 24 000 F	12 à 24 000 F

Les formes réduites spécifiques aux différentes classes en fonction du seuil de fréquence retenu

Classe 1 : Les femmes actuelles

Seuil de fréquence > 4 (Vocabulaire spécifique de la classe 1)

bijou+(55), femme+(44), actu+el(41), parfum+(30), lecture+(67), Naf-Naf(19), natation(47), rose+(165), crista+l(27), gymnast+3(20), porcelaine+(33), Lara Fabian(10), mousse-au-chocolat(36), patinage-artistique(27), décorati+f(8), manuel+(5), rôti+(13), Canada(43), citron+(15), fleur+(31), lampe+(6), langue+(5), MarkSpencer+(7), Pascal(8), Burton(9), cabriolet(9), Faut-pas-rêver(19), Goldman(31), Louis-XIV(18), Mégane(25), Obispo(9), Sud-Ouest(26), côte+(32), Biarritz(5), hachis parmentier(8), Girond+(6), jaune+(27), boeuf+(13), chant+(4), coutur+e(7), étain+(17), île+(18), mot+(10), pyramide+(9), tarte+(43), travaux(6), verre+(24), vêtement+(14), camaïeu(10), chantilly(10), Morgan(7), orchidée(10), paella(14), Pagny(16), pot-au-feu(12), saumon(11), Top Santé(7), Avantag+e(5), azur<(20), religi<(5), Calédonie(5), instit(4), Kennedy(6), Manoukian(5), Modes-et-Travaux(5), patchwork(3), Pimkie(3), Capita+l(31).

Seuil de fréquence > 6 (Vocabulaire spécifique de la classe 4)

femme+(50), actu+el(47), lecture+(58), bijou+(38), parfum+(24), porcelaine+(30), patinage-artistique(25), décorati+f(8), étain+(18), fleur+(31), île+(20), poulet+(14), chantilly(11), rose+(119), rôti+(11), Suisse+(8), blan+c„(23), jaune+(23), crista+l(21), Mark+(6), veau+(8), Aznavour(14), caramel(7), Chanel(7), Clio(20), Lara Fabian(8), Louis-XIV(16), marche-à-pied(13), mousse-au-chocolat(27), Naf-Naf(13), natation(33), pot-au-feu(12), Avantag+e(5), religi<(5), Devernois(6), Girond+(5), grand+(7), Italie(21), amour+(6), broderie+(6), coutur+e(6), dans+e(18), gâteau+(14), objet+(5), plante+(5), prix(6), pyramide+(7), reportage+(13), tomate+(5), tricot+(8), Caroll(5), marguerite(6), marie(9), cabriolet(7), céramique(7), Kiabi(8), Mégane(17), Obispo(6), Renault(50), Top Santé(6), Twingo(8), Calédonie(4), crudités(5), Julien Clerc(4), jt(4), Modes-et-Travaux(5), Rodier(5), Spencer(5), Télé-Star(5), Trenet(4), Turquie(5), dauphin+(5), mousse+(5), noix(5).

Seuil de fréquence >10 (Vocabulaire spécifique de la classe 3)

femme+(52), lecture+(64), Naf-Naf(19), actu+el(49), bijou+(42), parfum+(24), cabriolet(11), natation(40), patinage-artistique(25), fleur+(32), Clio(26), Lara Fabian(9), Figaro(15), clair+(9), orange+(11), rose+(123), coutur+e(8), crista+l(19), étain+(16), forêt+(8), île+(18), porcelaine+(25), riz(8), marie(12), Kiabi(9), Louis-XIV(17), mousse-au-chocolat(31), Mozart(11), azur<(19), côte+(28), beau+(9), rôti+(9), Suisse+(6), blan+c„(20), jaune+(21), amour+(6), citron+(10), gâteau+(13), préférence+(15), tricot+(8), veau+(7), vêtement+(11), Aznavour(12), Burton(6), caramel(7), céramique(7), Chanel(6), chantilly(9), Ford(14), Goldman(24), marche-à-pied(12), Obispo(6), Pagny(13), Peugeot(46), pot-au-feu(10), Canada(29), chou+(7), poisson+(23), reportage+(12), sauce+(11), Alpes(15), lasagnes(10), tulipe(8), modèle+(6), Pasteur+(9), pyramide+(6), Top Santé(5).

Seuil de fréquence =15 (Vocabulaire spécifique de la classe 4)

bijou+(37), femme+(41), Naf-Naf(17), actu+el(37), étain+(17), chantilly(12), porcelaine+(25), patinage-artistique(20), crista+l(18), lecture+(42), Goldman(24), citron+(9), gâteau+(12), lys(8), parfum+(15), tricot+(7), marie(9), Aznavour(11), Clio(16), Grèce(12), natation(24), orchidée(8), Peugeot(36), pot-au-feu(10), azur<(15), Pasteur+(9), aller.(8), regard+er(13), mousse-au-chocolat(18), crèm+e(17), patin+(10), orange+(7), Pyrénées(9), tarte+(24), mille-feuille(7), dans+e(12), gymnast+3(8), documentaires(6), Géo(6).

Classe 2 : Les hédonistes modernes

Seuil de fréquence > 4 (Vocabulaire spécifique de la classe 2)

bois(164), pate+(39), Ferrari(20), moto+(22), Levis(47), chêne+(40), Équipe+(13), tigre+(11), martin(8), carbonara(9), foot(48), Lacoste(17), nul part ailleurs (15), Phil Collins (10), Strauss(10), informat+3(11), liégeois+(8), austral+(17), Parisien+(14), Safrane+(7), chaîne+(7), dire Strait+(5), pin+(9), sapin+(14), ski+(20), voiture+(24), écout+er(6), bolognaise(7), Lee Cooper(8), Einstein(8), Francis Cabrel (20), Irlande(12), Lee Cooper(9), Léonard-de-Vinci(9), Marseille(6), pizza(7), Porsche(13), chass+e(11), guignol<(6), Eddy-Mitchell(6), Floyd(4), Friends(5), Jimmy Hendrix(5), Indonésie(5), Jimmy Hendrix(6), judo(6), King(5), Lamborghini(4), Luther(5), pc(7), Pink(5), Dire Strait+(5), auto+(8), box+e(8), Cana+l(5), parasol+(4), pêche+(18), science+(6), séjour+(4), surprise+(7), voyage+(25), fondre.(4), jou+er(5), sortir.(4), Audi(11), BMW(17), Brassens(12), Mitterrand(10), profiteroles(9), sud-est(7), Volkswagen(11), vtt(8), Beatle+(4), Boss(5), Bouches-du-Rhône(5).

Seuil de fréquence > 6 (Vocabulaire spécifique de la classe 1)

Francis Cabrel(33), voiture+(35), Levis(52), noir+(45), auto+(14), bois(184), pate+(37), pied+(20), Adidas(35), Capita+l(34), Équipe+(13), moto+(18), pêche+(25), balad+er(15), charlotte(11), Atlantique(8), Audi(15), Charlemagne(16), foot+(87), Goldman(32), Lacoste(18), Laguna(14), magazine(18), Mitterrand(14), Nul part ailleurs (14), Strauss(11), vtt(10), chocolat<(38), cinéma<(36), A4(6), Beatle+(6), Dauphiné(5), liégeois+(7), U2(6), austral+(17), Parisien+(15), États-Unis(19), course+(18), famille+(7), fraise+(15), loup+(5), ordinateur+(8), sapin+(13), vacance+(7), BMW(19), Combien-ça-coûte(13), Céline Dion(9), fruits-de-mer(14), La-voix-du-Nord(10), mille-feuille(11), panthère(7), Paris-Brest(9), Phil Collins(8), Var(14), complic+e(5), coupe+(18), Bob(5), Eddy-Mitchell(6), Jimmy Hendrix(5), Kennedy(6), pc(6), pivoine(5), ouest+(6), Canada(36), box+e(8), pin+(7), région+(13), sortie+(8), surprise+(8), tigre+(9), carbonara(6), érable(6).

Seuil de fréquence > 10 (Vocabulaire spécifique de la classe 1)

moto+(25), sport+(23), chass+e(17), auto+(14), voiture+(32), foot+(89), ordinateur+(10), pate+(34), pêche+(26), rugby(22), Adidas(34), Bigdil(18), Brassens(16), Ferrari(16), Francis Cabrel (24), Johnny Hallyday(25), Lacoste(19), Laguna(15), Strauss(11), Parisien+(15), Safrane+(8), chien+(152), fraise+(17), pied+(17), Atlantique(8), BMW(19), Lee Cooper(10), Céline Dion(9), mille-feuille(12), Mitterrand(12), Porsche(12), turbo(7), automobil<(9), enf+ant(13), austral+(15), amitié+(7), chêne+(31), course+(15), Équipe+(11), région+(13), vélo+(28), balad+er(12), Jules César(8), Alpes-Maritimes(9), couscous(21), Einstein(7), infos(7), La-voix-du-Nord(9), Mégane(18), Paris-Brest(8), Savoie(11), steak-frites(16), Sud-Est(7), Tahiti(14), télé(35), coupe+(15), Guadeloupe(13), Nike(8), Toyota(6), Var(11), Afrique(6), États-Unis(16), olivier+(8), charlotte(8), documentaires(8), entrecôte(9).

Seuil de fréquence > 15 (Vocabulaire spécifique de la classe 3)

moto+(24), foot(65), pêche+(29), Parisien+(18), bois(154), voiture+(32), Adidas(35), chêne+(36), chien+(149), rugby(23), BMW(22), Ferrari(15), Lacoste(17), Levis(42), enf+ant(14), Capita+l(30), Équipe+(12), glace+(35), journa+l(17), région+(13), Bigdil(15), Francis Cabrel(11), Laguna(13), magazine(14), Porsche(13), Rhône-Alpes(9), Savoie(12), noir+(31), Charente<(12), États-Unis(17), série+(9), Alpes-Maritimes(9), couscous(19), forêt-noire(13), La-voix-du-Nord(9), Lorraine(7), Tahiti(12), cinéma<(26), coupe+(14), bleu+(117), équitation+(10), fraise+(12), pate+(23), sapin+(10), informat+3(8), Auvergne(12), steak-frites(14), Var(10).

Classe 3 : Les âgés traditionalistes

Seuil de fréquence > 4 (Vocabulaire spécifique de la classe 3)
fruit+(49), marche+(46), Citroën(26), Questions-pour-un-champion(24), Tino Rossi(14), Xm(10), France(72), Italie(25), légume+(18), raisin+(9), tableau+(13), temps(7), de Gaulle(59), Frédéric François(16), patin+„(15), frai+c„(5), midi+(24), Angleterre(5), émission+(17), Marne+(9), tapisserie+(5), variété+(11), Claude François (6), bouquet-de-fleurs(14), choucroute(16), Clio(19), pétanque(10), Sardou(36), montagn+e(14), Bx(4), chrysanthème(4), Damart(5), Foucault(4), Mgriffon(5), arc+(6), belote+(3), cuivre+(13), grille+(4), jour+(11), livre+(22), marque+(17), pays(8), peinture+(12), santé+(6), sauce+(11), soup+e(5), tissu+(13), tour+(3), connaître.(7), cors+er(11), marguerite(5), Aznavour(12), Figaro(10), Jeanne d'Arc(6), agneau<(7), canevas(3), C&A(3), Fr3(3), La-chance-aux-chanson(5), magnétoscope(3), scrabble(4), Serge Lama(4), Zx(3), rouge<(46), Loire(10), centre+(6), faïence+(5), match+(4), mer+(6), poire+(7), poisson+(21), vélo+(22), ferr+er(6), ancien<(4).

Seuil de fréquence > 6 (Vocabulaire spécifique de la classe 2)
Citroën(34), Tino Rossi(17), fruit+(44), marche+(47), peint+(8), midi+(30), pomme+(24), jardin+er(23) Claude François(17), journal-télévisé(10), pétanque(12), Questions-pour-un-champion(25), Sardou(45), Xantia(12), Damart(7), Enrico(7), Macias(7), Nord+(12), Républicain+(16), Allemagne(5), Autriche(9), bague+(7), chien+(149), chiffre+(6), flan+(11), légume+(18), lettre+(6), mot+(11), nature+(8), papier+(8), pommier+(8), raisin+(8), santé+(8), tableau+(14), télévision+(10), temps(6), terre+(23), variété+(11), connaître.(8), crois+er(8), de Gaulle(62), Frédéric François(14), choucroute(17), Duteil(8), Fiesta(8), Xm(9), informat+ion(12), patin+„(13), Bx(5), La-chance-aux-chanson(6), frai+c„(5), quotidien+(5), Mère Teresa+(6), France(72), Loire(12), arc+(7), cuivre+(14), cyclis+me(10), jour+(12), marque+(20), peinture+(13), tapisserie+(5), toile+(5), vie+(10), ferr+er(7), bouquet-de-fleurs(12), César(6), Derrick(6), Ford(14), Jeanne d'Arc(6), Paca(10), sincèr+e(5), Astra(5), dahlia+(5), gris+(5).

Seuil de fréquence > 10 (Vocabulaire spécifique de la classe 2)
Citroën(34), Frédéric François(24), marche+(50), pomme+(28), Tino Rossi(17), jardin+er(26), midi+(29), bague+(9), flan+(15), nature+(9), télévision+(11), terre+(24), ferr+er(10), de Gaulle(64), journal-télévisé(10), Xantia(12), Xm(10), Républicain+(16), Autriche(9), France(76), arc+(8), centre+(9), fruit+(36), Marne+(9), mot+(10), papier+(8), poulet+(12), raisin+(8), santé+(8), tableau+(12), Fiesta(8), Jeanne d'Arc(8), pétanque(11), Questions-pour-un-champion(20), Sardou(40), Mère Teresa(7), informat+ion(12), patin+„(14), Nord+(10), cyclis+me(10), jour+(12), légume+(14), prix(6), variété+(10), crois+er(7), bouquet-de-fleurs(12), Duteil(7), maxi(6), Paca(10), violet+(7), Charente<(10), Italie(20), cuivre+(13), oeillet+(6), soie+(8), Lorraine(7), tarte-aux-pommes(20), Loire(11), tissu+(13), connaître.(6), choucroute(14), Vercingétorix(5), agneau<(6), plast+3(5).

Seuil de fréquence > 15 (Vocabulaire spécifique de la classe 1)
Frédéric François(24), marche+(55), France(93), fleur+(36), pomme+(28), terre+(30), Citroën(29), Tino Rossi(17), fruit+(44), Figaro(16), Ford(21), Sardou(49), Xantia(13), midi+(27), Italie(28), Loire(15), émission+(22), légume+(19), livre+(28), préférence+(16), tableau+(15), variété+(13), aim+er(12), de Gaulle(64), jardin+er(22), bouquet-de-fleurs(16), pétanque(11), petit+(14), cuivre+(15), flan+(11), île+(18), mot+(9), pays(10), peinture+(14), Réunion+(10), soie+(10), tissu+(16), choucroute(17), Paca(10), Questions-pour-un-champion(21), Nord+(9), Républicain+(13), rose+(130), vie+(10), Napoléon(44), informat+ion(10), Alpes(16).

Classe 4 : Les intellectuels bons vivants

Seuil de fréquence > 4 (Vocabulaire spécifique de la classe 4)

Monde+(25), St Laurent(10), baba rhum(8), Canard+(18), St Honor+er(7), Marianne(9), Inde(4), art+(8), jaguar+(13), vin+(13), voile+(10), enchaîn+er(8), Beethoven(7), Jaurès(7), Thalassa(30), volley ball(7), Schubert(5), acier+(4), actualité+(5), ail(4), bouteille+(9), cuisin+e(6), fer+(6), forge+(4), golf+(15), hêtre+(6), lion+(6), maison+(12), meuble+(5), table+(7), tennis(27), confire.(8), voyag+er(8), Henri(11), Jean(8), Michel(5), Bourgogne(6), Celio(9), fruits-de-mer(10), Gandhi(7), gâteau-au-chocolat(14), louis(7), Mercedes(19), Scenic(12), libérat+ion(7), Chrysler(5), Diesel(5), nougat(4), Science-et-vie(5), ferre+(3), naturel+(3), particulier+(4), plat+(6), Amérique<(5), bouleau+(5), cheva+l(20), débat+(3), glace+(24), lys(6), planche+(4), plateau+(3), bricol+er(18), blanquette(6), bleu-marine(5), bordeaux(5), cd(5), Envoyé-Spécial(17), Express(8), magret-de-canard(6), Pays-basque(9), Seychelles(5), bonne+(8), Chopin(3);

Seuil de fréquence > 6 (Vocabulaire spécifique de la classe 3)

Canard+(25), golf+(25), jaguar+(19), Monde+(30), bouteille+(14), musique+(24), vin+(17), St Laurent(11), Celio(13), Thalassa(38), art+(9), tennis(36), confire.(11), enchaîn+er(10), Yves(6), Gandhi(11), Marianne(9), Pays-basque(14), volley ball (8), baba rhum(8), Dordogne(11), chêne+(30), cuisin+e(7), fer+(6), hêtre+(7), lion+(8), livre+(27), restaurant+(8), table+(8), voile+(9), bricol+er(27), écout+er(7), honor+er(6), Henri(14), Jacques(6), Jean(9), Beethoven(7), Brel(25), Churchill(7), Envoyé-Spécial(26), île-flottante(20), Sud-Ouest(22), basqu<(8), bonne+(10), Chrysler(6), dernier+(4), argent(4), cheva+l(24), fromage+(9), maison+(13), meuble+(5), ski+(17), super(5), verre+(18), voyage+(24), bro+yer(4), brûl+er(5), voyag+er(9), Michel(5), bolognaise(6), Bourgogne(6), cd(6), express(9), Faut-pas-rêver(15), gâteau-au-chocolat(16), Léonard-de-Vinci(7), Mercedes(20), Pavarotti(6), Scenic(12), Télérama(7), class+3(4), Boss(5), Charlie-Hebdo(4)

Seuil de fréquence > 10 (Vocabulaire spécifique de la classe 4)

golf+(21), Canard+(19), musique+(20), bricol+er(28), St Laurent(10), Beethoven(9), Brel(27), randonnée(16), lion+(8), Monde+(22), confire.(9), enchaîn+er(8), bleu-marine(8), Express(11), Faut-pas-rêver(18), Gandhi(9), Marianne(8), Scenic(15), Sud-Ouest(21), Thalassa(31), vert+(48), boeuf+(10), bois(112), jaguar+(13), livre+(23), vin+(11), Aquitaine(11), Churchill(6), gâteau-au-chocolat(15), île-flottante(19), Marche-du-siècle(13), Renault-Espace(7), Tatin(7), athlèt<(7), bonne+(9), art+(6), cheva+l(20), lapin+(9), maison+(11), ski+(15), table+(6), cors+er(10), faire.(8), bolognaise(5), Bourgogne(5), crustacés(5), magnolia(6), Pavarotti(5), vtt(7), Capita+l(20), voir.(7), Opel(11), pizza(5), Provence(21), saumon(7), Télérama(6), chat+(58), gigot+(6).

Seuil de fréquence > 15 (Vocabulaire spécifique de la classe 2)

Monde+(39), Express(17), Canard+(23), ski+(26), voyage+(36), bouteille+(13), cheva+l(33), fromage+(12), jaguar+(18), musique+(23), tennis(41), voile+(12), cors+er(17), Envoyé-Spécial(31), randonnée(18), Scenic(18), St Laurent(14), Sud-Ouest(29), saint+(10), golf+(19), maison+(17), vin+(13), voyag+er(11), Brel(27), Faut-pas-rêver(20), foie-gras(10), Mégane(22), Pays-basque(15), Thalassa(36), Twingo(10), bonne+(11), côte+(30), libérat+ion(10), austral+(14), jaune+(24), oeuf+(12), olivier+(9), bricol+er(26), Henri(13), camaïeu(10), gâteau-au-chocolat(18), Dordogne(9), Espace+(13), faire.(10), aime-pas(9), Aquitaine(10), Marche-du-siècle(13), profiteroles(9), Paris(8), Jean(7), Audi(10), fruits-de-mer(11).

Cinquième Classe

Seuil de fréquence > 4 (Vocabulaire spécifique de la classe 5)

chiffre+(6), lettre+(6), Républicain+(12), pomme+(16), Enrico(6), Macias(6), Mère Teresa+(5), gratin+(6), papier+(6), terre+(16), sincèr+e(5), dahlia+(5), allemand+(3), peint+(5), berger+(4), champion+(3), flan+(8), pommier+(6), jardin+er(13), Adidas(17), laser(3), r5(3), Allemagne(3), Espagne(13), Portugal(5), bague+(4), chien+(65), course+(9), paquet+(3), poche+(4), soie+(6), télé+(22), tilleul+(4), César(4), endives(4), Fiesta(5), Lorraine(5), Peugeot(27), son-amitié(8), tarte-aux-pommes(13), montre+(7), plast+3(4), Astra(4), glaïeul(4), Gti(3), muguet(3), vidéo(4), yaourts(4), Nord+(5), quotidien+(3), Ardennes(3), breton+(2), pied+(7), singe+(2), vanille+(3), march+er(2), gratin-dauphinois(4), infos(4), frambois+(3), lorrain(2), rosbif(3), Sénégal+(2), Vaucluse(2), Weil(2), yaourt-aux-fruits(2), Auvergne(6), Montpellier(2), amitié+(3), amour+(3), eau+(2), écho+(2), promen+eur(10), port+er(2), Johnny Hallyday(9), Ouest-France(8).

Seuil de fréquence > 6 (Vocabulaire spécifique de la classe 5)

couleur+(2), cuite+(4), olivier+(4), Lee Cooper(9), boite+(2), chass+e(4), formule+(3), bateau+(2), théâtre+(2), cal+er(2), prim+er(2), Ferrari(3), Porsche(3), haut+(4), Marne+(2), sport+(3), pierre(2), saumon(2), automobil<(2), Cana+l(1), lapin+(2), match+(1), vase+(1), aller.(2), bala+yer(1), Maurice(1), Savoie(2), enf+ant(2), daube(1), Kenya(1), M6(1), Picasso(1), Stones(1), vert+(6), poche+(1), jou+er(1), occup+er(1), pass+er(1), Napoléon(5), Sud-Ouest-France(1), Tahiti(2), télé(4), éclair-au-chocolat(1), judo(1), faïence+(1), marbre+(1), crustacés(1), Sud-Est(1), Toyota(1).

Seuil de fréquence > 10 (Vocabulaire spécifique de la classe 5)

Phil Collins(8), Envoyé-Spécial(19), salade+(8), sortie+(6), tigre+(7), fi+er(6), voyag+er(8), Ça-se-discute(7), Levis(22), paella(9), profiteroles(8), Libérat+ion(7), marron+(5), nouvel+(6), noir+(19), Maroc(6), bouteille+(6), fromage+(6), science+(5), soir+(5), tennis(17), voile+(6), voyage+(14), Celio(6), Léonard-de-Vinci(5), Mercedes(13), basqu<(5), crêpe+(5), belle+(5), Paris(5), anima+l(4), équitation+(6), iris(4), pays(5), pin+(4), Audi(6), en-général(4), Hugo(4), magret-de-canard(4), Pays-basque(6), polo(4), cinéma<(15), informat+3(5), montre+(6), ami+(6), poire+(4), vie+(5), blanquette(4), cerisier(4), Charlemagne(6), foie-gras(4), Amérique<(3), Arte(3), bordeaux(3), camaïeu(4), cd(3), Drôme(3), Nul part ailleurs(5);

Seuil de fréquence > 15 (Vocabulaire spécifique de la classe 5)

pied+(12), Brassens(12), auto+(8), Ouest-France(13), frite+(11), Celio(6), Normandie(10), chass+e(7), montre+(8), course+(8), gigot+(5), salade+(7), balad+er(6), entrecôte(6), Irlande(6), Louis-XIV(8), Mitterrand(6), tarte-aux-pommes(11), vert+(23), film+(8), football(13), jour+(6), sport+(7), vélo+(12), vêtement+(6), Antilles(6), Johnny Hallyday(9), Martinique(5), Opel(7), Provence(12), Bretagne(12), verre+(8), cerisier(4), île-flottante(8), sans-aucun-doute(4), télé(13), pierre(4)

Ainsi, le corpus de préférences et ses analyses typologiques mettent en avant plusieurs éléments, qui viennent confirmer ceux obtenus lors des analyses sur le corpus « La perception du bonheur » :

- Le seuil de sélection des mots a une incidence directe sur le nombre de mots analysés statistiquement : quand celui-ci s'accroît (c'est-à-dire plus le seuil de fréquence s'accroît), le vocabulaire se réduit très rapidement, sans pour autant conduire à la perte d'uce et de réponses.
- Le seuil de sélection des mots a un impact assez fort sur l'arbre de classification. La constitution des classes typologiques ne se fait pas de la même manière selon le seuil pris en compte pour réduire le vocabulaire. De fait, les unités rassemblées dans chaque groupe ne sont pas identiques. Cela engendre des différences en terme de fréquences des mots employés.
- Prendre en compte davantage de mots dans l'analyse typologique permet d'une certaine façon d'affiner la description des classes, mais cela ne change pas pour autant la base de constitution des groupes. On obtient peut-être plus de précision et de nuances pour comprendre les corrélations
- Au total, néanmoins, on retrouve quasiment les mêmes groupes typologiques, à l'exception de la dernière classe constituée, la moins homogène, la moins typée et la moins facile à interpréter, et très logiquement la moins stable.

Les résultats apportent également un enseignement nouveau :

- Le seuil de sélection des mots a surtout un fort impact sur les classes typologiques de petite taille, c'est-à-dire celles regroupant un nombre faible d'unités de contexte élémentaires.

On peut donc en conclure qu'il existe réellement des incidences du degré de lemmatisation sur la robustesse des analyses et des typologies lexicales ; mais les effets restent minimes à l'exception des classes de petite taille. Il conviendra donc d'analyser avec précaution ces « petits groupes ».

Autrement dit, sur des corpus comme ceux analysés dans ce travail, les typologies peuvent être considérées comme stables, si on se limite à n'étudier que les classes regroupant un grand nombre d'uce (plusieurs centaines).

CONCLUSION

La première conclusion générale que l'on peut retirer des diverses comparaisons méthodologiques réalisées dans ce rapport est la grande stabilité de la méthode de classification descendante utilisée par *Alceste*. Que l'on utilise ou non la lemmatisation, quelle que soit l'étendue du vocabulaire sélectionné pour les analyses, les grandes classes typologiques obtenues sont toujours les mêmes et sont toujours de taille comparable.

La deuxième conclusion est l'incontestable contribution de la lemmatisation à la stabilité des résultats obtenus. Si l'utilisateur perfectionniste préfère éviter cette étape, il n'obtiendra pas de résultats radicalement différents, mais il s'expose davantage à l'irruption malencontreuse de classes de réponses artefactuelles et improductives.

L'intérêt de la lemmatisation est clair lorsque cette opération conduit à une réduction de 10 à 30 % du vocabulaire. Par indisponibilité de corpus adéquat, nous n'avons pas pu tester les avantages et inconvénients de lemmatisations plus radicales.

La troisième conclusion réside dans l'intérêt de conserver comme actif dans l'analyse un vocabulaire assez étendu, puisque cette ouverture n'hypothèque manifestement pas la stabilité des résultats typologiques obtenus. Ce résultat remarquable est dû principalement à la robustesse de la méthode de classification descendante hiérarchique appliquée aux tableaux hypercreux étudiés en analyse lexicométrique.

On peut même affirmer que parmi les originalités du logiciel *Alceste*, c'est davantage cette méthode de classification descendante que la présence de la lemmatisation qui conduit à une telle stabilité des résultats. En effet, les résultats des analyses sans lemmatisation effectués dans ce rapport restent interprétables et proches de ceux obtenus avec lemmatisation. Dans un rapport précédent, la comparaison entre les résultats obtenus par *Alceste* et ceux obtenus par *Leximappe* —qui utilise une méthode de classification proche des nuées dynamiques— avait démontré toute la robustesse de la classification descendante.

Enfin, ces résultats rassurants quant à la stabilité de la méthodologie *Alceste* ne doivent pas nous encourager à abandonner notre esprit critique vis-à-vis de cet outil, mais plutôt à continuer à en tester les nouvelles fonctionnalités. Le point faible, souligné à de multiples reprises dans ce rapport, est la génération assez fréquente de petites classes de quelques dizaines d'unités de contexte instables et difficilement interprétables. Il nous reste à mieux comprendre l'origine de la forte variance inter-classe qui initie la création de ces dernières et les moyens de mieux contrôler cette création, par exemple par croisement de plusieurs typologies successives avec des seuils de sélection différents entre elles.

BIBLIOGRAPHIE

ANASTEX S.J., (1993) - Actes des Secondes Journées Internationales d'Analyse Statistique de Données Textuelles, JADT 1993.

BABAYOU P. (1997).- « Traitement des questions ouvertes : comparaison d'une postcodification et de méthodes lexicométrique et d'analyse du discours », *CRÉDOC, Cahier de recherche* n°101, septembre.

BAUER D., MARESCA B., (1992).- « Lignes de vie : méthodologie de recueil et de traitement des données bibliographiques », *CRÉDOC, Cahier de recherche* n°37.

BEAUDOUIN V. (1995).- « Analyse textuelle et structures narratives de récits », *CRÉDOC, Cahier de Recherche* n°82.

BEAUDOUIN V., BROCHET F., (1996).- « Analyse lexicale de corpus en anglais », *CRÉDOC, Cahier de Recherche* n°95, septembre.

BEAUDOUIN V., HÉBEL P. (1994).- « Avancées en analyse lexicale », *CRÉDOC, Cahier de Recherche* n°61, mai.

BEAUDOUIN V., LAHLOU S., (1993).- « L'analyse lexicale : outil d'exploration des représentations », *CRÉDOC, Cahier de Recherche* n°48.

BROUSSEAU A.-D., VOLATIER J.-L., (1999).- « Le consommateur français en 1998 – Une typologie des préférences », *CRÉDOC, Cahier de recherche* n°130, janvier, 170 p.

COLLERIE DE BORELY A., (avec la participation de SIGOGNEAU A., de l'OST), (1998).- « Étude de réseaux de mots - Confrontations des résultats issus de deux méthodes d'analyse textuelle *Alceste* et *Leximappe* », *CRÉDOC, Cahier de recherche* n°115, juillet, 111 p.

INED, (1995).- « L'analyse textuelle dans les enquêtes », Compte-Rendu du Séminaire de méthodes d'enquêtes à Paris le 17 janvier 1995 (24 février).

JENNY J., (1997).- « Méthodes et pratiques formalisées d'analyse de contenu et de discours dans la recherche sociologique française contemporaine. Etat des lieux et essai de classification », *Bulletin de Méthodologie Scientifique* n°54.

LAFON P. (1984).- *Dépouillements et statistiques en lexicométrie*, Éd. Slatkine – Champion.

LEBART L., SALEM A. (1988).- *Analyse statistique des données textuelles*, Éd. Dunod.

LION S. (1991).- « Construction d'un corpus et perte d'information en analyse lexicale », *CRÉDOC, Cahier de recherche* n°13, avril.

REINERT M, (1993).- « Les "mondes lexicaux" et leur logique », *Langage et société*, Maison des Sciences de l'Homme, n°66, pp. 5-39.

REINERT M., (1983).- « Une méthode de classification descendante hiérarchique : application à l'analyse lexicale par contexte », *Les cahiers de l'analyse des données*, Vol VIII, n°2.

REINERT M., (1990).- « Une méthode de classification des énoncés d'un corpus présentée à l'aide d'une application », *Les cahiers de l'analyse des données*, Vol XV, n°1, pp. 21-36.

YVON F. (1990).- « L'analyse lexicale appliquée à des données d'enquête : état de lieux », *CRÉDOC, Cahier de recherche* n°5, décembre.

ANNEXES

CORPUS « LA PERCEPTION DU BONHEUR »

« La perception du bonheur » - Fréquence des formes réduites

Avec lemmatisation	
Fréquence	Forme réduite
365	santé+
312	famille+
293	travail<
292	argent
221	faire.
220	ne-pas
193	vivre.
192	enf+ant
186	vie+
185	pouvoir+
125	heur+eux
116	loisir+
106	bonheur+
100	vacance+
93	voir.
93	mes-enfants
92	aller.
92	être-en-bonne-santé
86	monde+
82	temps
82	problem<
78	ami+
77	amour+
77	être-bien
68	quand-on
67	chose+
65	souci+
60	bonne+
59	envi+e
58	financier+
58	dire+
58	maison+
58	je-suis
55	malade+
55	gens
54	aim+er
50	familia+l
50	pa+yer

Sans lemmatisation	
Fréquence	Forme réduite
365	santé
310	famille
292	argent
241	travail
220	ne-pas
210	faire
187	vivre
185	pouvoir
185	vie
172	enfants
105	loisirs
101	bonheur
98	vacances
93	heureux
93	mes-enfants
92	être-en-bonne-santé
85	monde
82	temps
77	être-bien
75	amis
71	amour
71	voir
70	problèmes
68	quand-on
65	soucis
58	choses
58	dire
58	je-suis
57	maison
55	gens
55	envie
52	malade
48	être-heureux
47	bonne
45	bonne-santé
44	travailler
44	en-bonne-santé
42	niveau

« La perception du bonheur » - Fréquence des formes réduites

Avec lemmatisation	
Fréquence	**Forme réduite**
48	être-heureux
45	besoin+
45	niveau+
45	bonne-santé
44	en-bonne-santé
42	chom+„
42	ma-famille
41	fait
40	femme+
40	voyage+
38	petits-enfants
37	profit+er
36	couple+
36	être-bien-dans-sa-p
33	plaisir+
33	soleil+
33	achet+er
32	petit+
32	suffis+ant
31	tranquil+e
30	matéri+el
29	paix
29	souci
29	entour+er
29	mon-mari
28	libre+
28	entendre.
28	gagn+er
28	ses-enfants
27	emploi+
27	manqu+er
27	prendre.
27	professionn+el
26	moyen+
26	mang+er
26	avoir-un-travail
25	correct+
25	liberté+
25	tête+
25	temps-libre
25	vie-de-famille
24	maladie+
23	grand+
23	réuss+ir
23	sortir.
22	foyer+

Sans lemmatisation	
Fréquence	**Forme réduite**
42	chômage
42	ma-famille
41	fait
39	femme
38	petits-enfants
36	couple
36	profiter
36	va
36	être-bien-dans-sa-p
33	soleil
30	heureuse
30	aime
29	paix
29	souci
29	acheter
29	mon-mari
28	plaisir
28	aller
28	ses-enfants
26	libre
26	avoir-un-travail
25	besoin
25	liberté
25	tête
25	voyages
25	manger
25	manquer
25	temps-libre
25	vie-de-famille
24	emploi
24	gagner
23	entoure
23	prendre
22	suffisamment
22	autour-de-soi
22	dans-son
21	financiers
21	foyer
21	misère
21	sécurité
21	payer
21	sortir
20	besoins
20	entente
20	joie
20	moyens

« La perception du bonheur » - Fréquence des formes réduites

Avec lemmatisation	
Fréquence	**Forme réduite**
22	joie+
22	autour-de-soi
22	dans-son
21	misère+
21	relation+
21	sécurité+
21	permettre.
21	sentir.
21	sta+ble
20	seul+
20	entente+
20	épanou+ir
20	plaire.
20	import+ant
20	revenu+
19	belle+
19	équilibre+
19	situation+
19	toit<
19	voiture+
19	voyag+er
19	harmoni+e
19	possi+ble
19	je-ne
18	bien-être
18	étude+
18	logement+
18	rest+er
18	vie-familiale
17	malheur+eux
17	jour+
17	parent+
17	occup+er
17	trouv+er
17	jeune+
16	agréable+
16	ensemble+
16	mari+
16	minimum
16	peau+
16	arriv+er
16	entourage<
15	achat+
15	environnement
15	épou+x
15	fille+

Sans lemmatisation	
Fréquence	**Forme réduite**
20	enfant
19	financier
19	relations
19	situation
19	voiture
19	aille
19	aimer
19	je-ne
18	familial
18	bien-être
18	études
18	logement
18	maladie
18	toit
18	sentir
18	voyager
18	vie-familiale
17	belle
17	familiale
16	ensemble
16	mari
16	minimum
16	parents
16	peau
16	partir-en-vacances
16	entourage
16	harmonie
16	tranquille
15	correctement
15	malheureux
15	environnement
15	nature
15	voyage
15	petit
15	professionnelle
15	stable
14	amitié
14	avenir
14	équilibre
14	savoir
14	sport
14	occuper
14	partir
14	plait
14	revenus
13	agréable

« La perception du bonheur » - Fréquence des formes réduites

Avec lemmatisation	
Fréquence	Forme réduite
15	nature+
15	retraite+
15	sport+
15	partir.
15	act+ion
14	proche+
14	amitié+
14	avenir+
14	guerre+
14	savoir+
14	rendre.
14	cote+
13	norma+l
13	quotidien+
13	métier+
13	mois
13	réussite+
13	vue+
13	compt+er
13	priv+er
13	réalis+er
13	stress+
13	autour-de-moi
12	difficile+
12	nécessaire+
12	France
12	fête+
12	personne+
12	question+
12	gard+er
12	offrir.
12	passe
12	être-en-famille
11	beau+
11	bon+
11	campagne+
11	confort+
11	moment+
11	salaire+
11	content+er
11	dépens+er
11	oblig+er
11	rencontr+er
11	phys+3
11	voisin<
10	content+

Sans lemmatisation	
Fréquence	Forme réduite
13	épouse
13	guerre
13	jours
13	mois
13	retraite
13	compter
13	priver
13	bonnes
13	autour-de-moi
12	familiaux
12	proches
12	seul
12	France
12	achat
12	métier
12	personnes
12	réussite
12	vue
12	garder
12	offrir
12	réussir
12	jeunes
12	matériel
12	passe
12	possible
12	problème
12	être-en-famille
11	bons
11	nécessaire
11	normalement
11	campagne
11	confort
11	fille
11	loisir
11	permet
11	réaliser
11	rencontrer
11	voit
10	continue
10	principal
10	boulot
10	fête
10	manque
10	sens
10	contenter
10	dépenser

« La perception du bonheur » - Fréquence des formes réduites

Avec lemmatisation	
Fréquence	Forme réduite
10	continu+
10	gros+
10	principa+l
10	boulot+
10	manque+
10	projet+
10	sens
10	aid+er
10	désir+er
10	donn+er
10	regard+er
10	venir.
10	riche+
10	loto
9	age+
9	mora+l
9	plein+
9	primordia+l
9	contrainte+
9	difficulté+
9	ennui+
9	esprit+
9	état+
9	fin+
9	musique+
9	porte+
9	société+
9	connaître.
9	demand+er
9	pens+er
9	pauvre+
8	calm+
8	personnel+
8	sain+
8	simple+
8	condition+
8	île+
8	impôt+
8	mer+
8	amélior+er
8	partag+er
8	affect+ion
8	différ+ent
8	éducat+ion
8	prés+ent
8	maladie-grave

Sans lemmatisation	
Fréquence	Forme réduite
10	entend
10	entendre
10	oblige
10	trouver
10	cote
10	loto
10	stress
9	age
9	beau
9	difficile
9	grande
9	chose
9	esprit
9	état
9	musique
9	porte
9	salaire
9	société
9	aider
9	regarder
9	reste
9	important
9	tranquillité
8	financière
8	grand
8	gros
8	seule
8	contraintes
8	difficultés
8	mer
8	projets
8	partager
8	pas-de-problème
8	rester
8	maladie-grave
7	calme
7	financièrement
7	plein
7	primordial
7	forcement
7	budget
7	ennuis
7	île
7	impôts
7	instant
7	marche

« La perception du bonheur » - Fréquence des formes réduites

Avec lemmatisation	
Fréquence	**Forme réduite**
7	sentimenta+l
7	socia+l
7	forcement
7	an+
7	budget+
7	désir+
7	heure+
7	homme+
7	instant+
7	limite+
7	marche+
7	ménage+
7	qualité+
7	région+
7	conjoindre.
7	consomm+er
7	essa+yer
7	fin+ir
7	habit+er
7	nourr+ir
7	pos+er
7	retrouv+er
7	satisfaire.
7	chanc+e
7	conforta+ble
7	déc+ent
7	rêv+e
7	viol+ent
7	handicap+
6	cher+
6	convenable+
6	découvert+
6	étranger+
6	grave+
6	humain+
6	optimiste+
6	actuellement
6	aide+
6	contact+
6	détente+
6	faim+
6	forme+
6	pays
6	place+
6	plage+
6	promen+eur

Sans lemmatisation	
Fréquence	**Forme réduite**
7	ménage
7	qualité
7	questions
7	arriver
7	conjoint
7	consommer
7	désire
7	donner
7	épanouie
7	nourrir
7	réussissent
7	satisfaire
7	vois
7	chance
7	matériels
7	présent
7	professionnel
7	suffisant
7	travaille
6	découvert
6	optimiste
6	quotidien
6	actuellement
6	conditions
6	détente
6	faim
6	fin
6	forme
6	heure
6	maladies
6	moment
6	pays
6	place
6	plage
6	région
6	sérénité
6	consacrer
6	essayer
6	penser
6	rendre
6	activité
6	activités
6	cinéma
6	consommation
6	décemment
6	éducation

« La perception du bonheur » - Fréquence des formes réduites

Avec lemmatisation	
Fréquence	Forme réduite
6	sérénité
6	truc+
6	consacr+er
6	écout+er
6	évit+er
6	exist+er
6	gér+er
6	port+er
6	cinéma<
6	consommat+ion
6	essenti+el
6	mari+„
6	polit+3
6	vieill<
6	vill+„
6	être-à-deux
6	télé
5	aise+
5	intéressant+
5	meilleur+
5	partie+
5	prive+
5	régulier+
5	scolaire+
5	vraiment
5	abord+
5	accord+
5	année+
5	appartement+
5	bateau+
5	biens
5	copain+
5	dépense+
5	généra+l
5	partage+
5	rapport+
5	satisfaction+
5	soir+
5	solidarité+
5	terre+
5	amus+er
5	chang+er
5	communiqu+er
5	continu+er
5	contribu+er
5	dépendre.

Sans lemmatisation	
Fréquence	Forme réduite
6	petite
6	petits
6	physique
6	revenu
6	stabilité
6	violence
6	être-à-deux
6	télé
5	content
5	correct
5	étranger
5	partie
5	vraiment
5	abord
5	accord
5	aide
5	amoureux
5	appartement
5	bateau
5	biens
5	bonheurs
5	copains
5	désirs
5	équilibrée
5	général
5	limite
5	moments
5	moyen
5	partage
5	peur
5	plaisirs
5	question
5	satisfaction
5	solidarité
5	terre
5	continuer
5	contribue
5	éviter
5	existe
5	fais
5	finir
5	fonder
5	poser
5	promener
5	subvenir
5	vient

« La perception du bonheur » - Fréquence des formes réduites

Avec lemmatisation	
Fréquence	Forme réduite
5	fond+er
5	pass+er
5	promen+er
5	subvenir.
5	marie
5	aliment<
5	ambian<
5	autonom<
5	confi+ant
5	indépend+ant
5	médica<
5	montagn+e
5	souhait<
5	superflu<
5	cocotiers
5	être-riche
5	insécurité
4	attenti+f
4	chaud+
4	large+
4	menta+l
4	naturel+
4	négati+f
4	parfait+
4	plan+
4	premier+
4	restricti+f
4	serein+
4	unie+
4	pleinement
4	abri+
4	bout+
4	cadre+
4	carrière+
4	cas
4	compte+
4	conflit+
4	coup+
4	culture+
4	domaine+
4	entreprise+
4	façon+
4	finance+
4	fois
4	gêne+
4	humeur+

Sans lemmatisation	
Fréquence	Forme réduite
5	marie
5	ambiance
5	confiance
5	essentiel
5	mariage
5	matérielles
5	montagne
5	pauvreté
5	petites
5	rêves
5	riche
5	voisin
5	cocotiers
5	être-riche
5	insécurité
4	chaud
4	grands
4	graves
4	large
4	moral
4	plan
4	privée
4	quotidienne
4	sain
4	unie
4	pleinement
4	abri
4	années
4	ans
4	cadre
4	carrière
4	cas
4	compte
4	contacts
4	coup
4	culture
4	filles
4	fois
4	gêne
4	homme
4	humeur
4	jour
4	lave
4	paysage
4	propriétaire
4	rapports

« La perception du bonheur » - Fréquence des formes réduites

Avec lemmatisation	
Fréquence	Forme réduite
4	lave+
4	passion+
4	paysage+
4	peine+
4	priorité+
4	propriétaire+
4	sortie+
4	tas
4	terme+
4	accord+er
4	assouv+ir
4	balad+er
4	comprendre.
4	conserv+er
4	cré+er
4	enrich+ir
4	entretenir.
4	envi+er
4	espér+er
4	grand+ir
4	habill+er
4	intéress+er
4	men+er
4	peindre.
4	plaindre.
4	respect+er
4	réun+ir
4	suivre.
4	assur<
4	désert+ion
4	médecin<
4	rac+3
4	tendre+
4	aujourd
4	avoir-un-métier
4	en-général
4	Internet
4	je-ne-sais-pas
4	peut-être-heureux
4	resto
4	tout-ce-qui-est
3	amica+l
3	certain+
3	dur+
3	futur+
3	immédiat+

Sans lemmatisation	
Fréquence	Forme réduite
4	tas
4	terme
4	trucs
4	achète
4	améliorer
4	assouvir
4	changer
4	connaître
4	conserver
4	demander
4	écoute
4	entretenir
4	envier
4	épanoui
4	gérer
4	habiller
4	habiter
4	intéresse
4	passer
4	plaise
4	rend
4	respecter
4	retrouver
4	trouvent
4	vit
4	confortablement
4	cotes
4	déserte
4	importante
4	indépendant
4	matérielle
4	superflu
4	tranquillement
4	aujourd
4	avoir-un-métier
4	en-général
4	Internet
4	je-ne-sais-pas
4	peut-être-heureux
4	resto
4	tout-ce-qui-est
3	agréables
3	aise
3	attention
3	certaine
3	cher

« La perception du bonheur » - Fréquence des formes réduites

Avec lemmatisation	
Fréquence	Forme réduite
3	matin+
3	obligatoire+
3	perdu+
3	positi+f
3	propre+
3	publi+c„
3	relationnel+
3	rempli+
3	satisfait+
3	uni+
3	constamment
3	noir+
3	frère+
3	bruit+
3	cancer+
3	chaleur
3	changement+
3	chat+
3	citoyen+
3	coin+
3	collègue+
3	compagne+
3	corps
3	course+
3	crainte+
3	crédit+
3	dette+
3	devoir+
3	dieu+
3	égalité+
3	égoïs+me
3	espérance+
3	facilite+
3	fils
3	foret+
3	idée+
3	impression+
3	lecture+
3	livre+
3	lune+
3	maximum
3	milieu+
3	minute+
3	mot+
3	naissance+
3	nourriture+

Sans lemmatisation	
Fréquence	Forme réduite
3	contente
3	convenable
3	convenablement
3	familiales
3	financières
3	futur
3	humaine
3	immédiat
3	intéressant
3	malades
3	matin
3	meilleur
3	obligatoirement
3	parfaite
3	personnel
3	personnelle
3	premier
3	quotidiens
3	régulièrement
3	relationnel
3	remplie
3	saine
3	scolaire
3	sentimental
3	simple
3	simples
3	constamment
3	achats
3	an
3	bout
3	cancer
3	chaleur
3	changement
3	citoyen
3	collègues
3	compagne
3	corps
3	courses
3	crainte
3	dépense
3	devoir
3	dieu
3	domaines
3	égalité
3	égoïste
3	emplois

« La perception du bonheur » - Fréquence des formes réduites

Avec lemmatisation	
Fréquence	Forme réduite
3	ordre+
3	part+
3	prix
3	rencontre+
3	ressource+
3	restaurant+
3	réunion+
3	rire+
3	soeur+
3	spectacle+
3	télévision+
3	tension+
3	touche+
3	train+
3	vaisselle+
3	vécu+
3	veu+f
3	affront+er
3	amen+er
3	apprendre.
3	arrang+er
3	arrêt+er
3	augment+er
3	cherch+er
3	découl+er
3	découvrir.
3	évolu+er
3	gât+er
3	impos+er
3	jou+er
3	lev+er
3	maintenir.
3	ouvrir.
3	perdre.
3	recevoir.
3	répondre.
3	revenir.
3	servir.
3	soign+er
3	souffrir.
3	soutenir.
3	valoir.
3	actu+el
3	communic<
3	délinqu+ant
3	individu<

Sans lemmatisation	
Fréquence	Forme réduite
3	entreprise
3	espérance
3	facilite
3	façon
3	fils
3	finances
3	fins
3	foret
3	hommes
3	impression
3	lecture
3	lune
3	maximum
3	milieu
3	minute
3	niveaux
3	nourriture
3	ordre
3	part
3	passions
3	peine
3	priorité
3	prix
3	produits
3	promenade
3	promenades
3	ressources
3	restaurant
3	rire
3	soirs
3	télévision
3	touche
3	train
3	vaisselle
3	vécu
3	veuve
3	affronter
3	aillent
3	aimerais
3	amélioration
3	apprendre
3	arrivent
3	comprendre
3	connaît
3	créer
3	demande

« La perception du bonheur » - Fréquence des formes réduites

Avec lemmatisation	
Fréquence	Forme réduite
3	intellectu+el
3	mutu+el
3	philo<
3	sorte+
3	spiritu+el
3	vis
3	deux-enfants
3	être-bien-être
3	hui
3	parce
3	petit-fils
3	se-faire-plaisir
3	tolérance

Sans lemmatisation	
Fréquence	Forme réduite
3	dépend
3	donne
3	entende
3	entourée
3	entourent
3	épanouissement
3	espère
3	gâter
3	grandir
3	impose
3	jouer
3	lever
3	occupe
3	ouvrir
3	paye
3	peindre
3	pense
3	permettrait
3	permettre
3	portant
3	porter
3	rendra
3	répondre
3	revenir
3	soigner
3	souffrir
3	trouve
3	viennent
3	affection
3	affective
3	alimentation
3	autonome
3	communication
3	confortable
3	harmonieuse
3	importance
3	jeunesse
3	médecin
3	physiques
3	politique
3	politiques
3	possibilité
3	richesse
3	tendresse
3	ville
3	vis

« La perception du bonheur » - Fréquence des formes réduites

Avec lemmatisation	
Fréquence	Forme réduite

Sans lemmatisation	
Fréquence	Forme réduite
3	voisins
3	deux-enfants
3	être-bien-être
3	hui
3	parce
3	petit-fils
3	se-faire-plaisir
3	tolérance

CORPUS « LES PRÉFÉRENCES DES FRANÇAIS »

« Les préférences des Français » - Fréquence des formes réduites

Avec lemmatisation	
Fréquence	Forme réduite
512	chien+
506	bois
485	deux enfants
451	rose+
435	bleu+
417	trois enfants
257	chat+
245	France
175	gaul+er
173	rouge<
172	vert+
158	Renault
139	Peugeot
136	lecture+
132	Napoléon
119	foot
106	Sardou
105	marche+
104	tarte+
92	fruit+
91	télé
90	noir+
89	chocolat<
89	Levis
88	glace+
87	Canada
87	tennis
83	Bretagne
80	chêne+
80	natation
80	Provence
79	football
79	Thalassa
78	bijou+
77	cinéma<
77	quatre enfants
73	Capita+l
72	vélo+
71	cheva+l

Sans lemmatisation	
Fréquence	Forme réduite
506	bois
503	chien
485	deux enfants
432	bleu
417	trois enfants
409	rose
250	chat
245	France
175	Gaulle
163	vert
162	rouge
158	Renault
139	Peugeot
136	lecture
132	Napoléon
119	foot
106	Sardou
105	marche
91	télé
89	Levis
87	Canada
87	tennis
85	chocolat
83	noir
83	Bretagne
83	tarte
80	chêne
80	natation
80	Provence
79	football
79	Thalassa
77	cinéma
73	Capital
72	vélo
70	bijou
70	fruits
69	mousse-au-chocolat
64	poisson
64	cote

« Les préférences des Français » - Fréquence des formes réduites

Avec lemmatisation	
Fréquence	**Forme réduite**
70	voyage+
69	mousse-au-chocolat
68	poisson+
65	cote+
64	fleur+
64	pâte+
64	Adidas
63	crèm+e
62	livre+
62	Goldman
61	bricol+er
61	Brel
60	Envoyé-spécial
59	porcelaine+
57	midi+
57	femme+
57	promen+eur
57	Sud-Ouest
56	jaune+
56	Espagne
56	Italie
56	voiture+
56	tarte-aux-pommes
53	monde+
53	Mercedes
52	haut+
51	pêche+
51	actu+el
49	marque+
49	verre+
49	Clio
49	couscous
49	Hallyday
48	blan+c„
46	terre+
46	île-flottante
46	Questions-pour-un-c
45	pomme+
44	dans+e
44	jardin+er
44	Citroën
43	Mégane
43	patinage-artistique
42	crista+l
42	préférence+
42	ski+

Sans lemmatisation	
Fréquence	**Forme réduite**
64	Adidas
62	Goldman
61	cheval
61	glace
61	crème
61	Brel
60	Envoyé-spécial
59	porcelaine
59	bricolage
58	pâtes
57	midi
57	Sud-Ouest
56	Espagne
56	Italie
56	femme
56	livre
56	voyage
56	tarte-aux-pommes
55	jaune
53	monde
53	Mercedes
51	fleurs
51	pêche
51	actuelle
49	verre
49	Clio
49	couscous
49	Hallyday
47	promenade
47	voiture
46	île-flottante
46	Questions-pour-un-c
44	danse
44	terre
44	Citroën
43	haute
43	jardinage
43	Mégane
43	patinage-artistique
42	roses
42	cristal
42	ski
42	azur
41	parfum
41	préférence
40	États-Unis

« Les préférences des Français » - Fréquence des formes réduites

Avec lemmatisation	
Fréquence	**Forme réduite**
42	regard+er
42	azur<
41	émission+
41	film+
41	musique+
41	parfum+
40	États-Unis
40	frite+
39	rugby
39	alpes
39	BMW
39	Ouest-France
38	île+
38	moto+
38	sud+
38	gâteau-au-chocolat
37	golf+
37	choucroute
37	steak-frites
36	canard+
35	course+
34	Faut-pas-rêver
33	journa+l
33	tissu+
33	coupé+
32	légume+
32	Louis-XIV
31	auvergne
31	cuivre+
31	sport+
31	Ford
31	Grèce
31	magazine
31	Opel
30	austral+
30	étain+
30	gâteau+
30	montagn+e
29	Pyrénées
29	espace+
29	fraise+
29	maison+
29	peinture+
29	reportage+
29	Bigdil
29	forêt-noire

Sans lemmatisation	
Fréquence	**Forme réduite**
40	musique
39	rugby
39	Alpes
39	Bmw
39	Ouest-France
38	blanc
38	frites
38	moto
38	Sud
38	gâteau-au-chocolat
37	golf
37	choucroute
37	steak-frites
35	canard
35	marque
35	pommes
34	Faut-pas-rêver
32	Louis-XIV
31	Auvergne
31	films
31	légumes
31	sport
31	tissu
31	coupé
31	Ford
31	Grèce
31	magazine
31	Opel
30	Australie
30	cuivre
30	étain
30	journal
30	montagne
29	Pyrénées
29	espace
29	maison
29	regarde
29	Bigdil
29	forêt-noire
29	Guadeloupe
29	Marche-du-siècle
29	Normandie
29	Pagny
29	Scenic
28	républicain
28	peinture

« Les préférences des Français » - Fréquence des formes réduites

Avec lemmatisation	
Fréquence	**Forme réduite**
29	Guadeloupe
29	Marche-du-siecle
29	Normandie
29	Pagny
29	Scenic
28	républicain+
28	pied+
28	charlemagne
28	François
28	Lacoste
27	parisien+
27	Loire
27	gymnast+3
27	idée+
27	rosier+
27	viande+
27	cors+er
27	Aznavour
27	Tahiti
26	Charente<
26	ami+
26	jaguar+
26	jour+
26	sauce+
26	Henri
26	patinage+
26	bouquet-de-fleurs
26	marche-à-pied
26	randonnée
25	région+
25	vêtement+
25	fruits-de-mer
25	paella
25	son-amitié
25	var
24	oeuf+
24	montre+
24	Antilles
24	Brassens
23	bourguignon+
23	pasteur+
23	poulet+
23	sapin+
23	balad+er
23	chass+e
23	Ferrari

Sans lemmatisation	
Fréquence	**Forme réduite**
28	Charlemagne
28	François
28	Lacoste
27	Loire
27	glaces
27	gymnastique
27	île
27	Aznavour
27	Tahiti
26	jaguar
26	reportages
26	Henri
26	bouquet-de-fleurs
26	marche-à-pied
26	randonnée
25	sauce
25	fruits-de-mer
25	paella
25	son-amitié
25	Var
24	idée
24	région
24	viande
24	montre
24	Antilles
24	Brassens
23	jours
23	Pasteur
23	pied
23	poulet
23	Ferrari
23	figaro
23	Nulpartailleurs
23	pays-basque
23	Volkswagen
22	bourguignon
22	boeuf
22	émission
22	équitation
22	fruit
22	Frédéric
22	Audi
22	Cabrel
22	lasagnes
22	Naf-Naf
22	Porsche

« Les préférences des Français » - Fréquence des formes réduites

Avec lemmatisation	
Fréquence	Forme réduite
23	figaro
23	Nulpartailleurs
23	pays-basque
23	Volkswagen
23	sept
22	boeuf+
22	équitation+
22	tableau+
22	vin+
22	aller.
22	Frédéric
22	alsac<
22	Audi
22	Cabrel
22	lasagnes
22	Naf-Naf
22	Porsche
22	Savoie
22	cinq
21	lapin+
21	salade+
21	informat+ion
21	petit+
21	Combien-ça-coûte
21	Irlande
21	Laguna
21	Martinique
21	pot-au-feu
20	Nord+
20	citron+
20	équipe+
20	flan+
20	vie+
20	aim+er
20	faire.
20	préfer+er
20	enf+ant
20	Aquitaine
20	entrecôte
20	je-ne
20	St
20	yaourt
19	orange+
19	cuir
19	cyclis+me
19	soie+

Sans lemmatisation	
Fréquence	Forme réduite
22	Savoie
21	course
21	lapin
21	oeufs
21	rosier
21	sapin
21	tartes
21	Alsace
21	patinage
21	Combien-ça-coûte
21	Irlande
21	Laguna
21	Martinique
21	pot-au-feu
20	Nord
20	citron
20	Équipe
20	fraises
20	vêtement
20	vie
20	vin
20	aller
20	Aquitaine
20	entrecôte
20	je-ne
20	St
20	yaourt
19	amis
19	cuir
19	émissions
19	soie
19	tableau
19	aime
19	Corse
19	intéresse
19	voyager
19	pierre
19	chasse
19	informations
19	Express
19	Géo
19	mille-feuille
19	Mitterrand
19	Paca
19	Rossi
18	orange

« Les préférences des Français » - Fréquence des formes réduites

Avec lemmatisation	
Fréquence	Forme réduite
19	variété+
19	intéress+er
19	voyag+er
19	pierre
19	express
19	Géo
19	mille-feuille
19	Mitterrand
19	Paca
19	Rossi
18	Dordogne
18	Maroc
18	pays
18	Réunion+
18	série+
18	voir.
18	marie
18	bonne+
18	aime-pas
18	Mozart
18	profiteroles
17	beau+
17	nouvel+
17	rôti+
17	Paris
17	lilas
17	olivier+
17	charlotte
17	Alpes-Maritimes
17	camaïeu
17	documentaires
17	orchidée
17	pétanque
17	Sans-aucun-doute
17	saumon
17	steak
17	tulipe
17	Xantia
16	auto+
16	bouteille+
16	formule+
16	fromage+
16	gigot+
16	jeu+
16	mot+
16	informat+3

Sans lemmatisation	
Fréquence	Forme réduite
18	Dordogne
18	Maroc
18	cyclisme
18	flan
18	gâteau
18	pays
18	Réunion
18	marie
18	aime-pas
18	Mozart
18	profiteroles
17	beau
17	parisien
17	Paris
17	lilas
17	olivier
17	charlotte
17	bonne
17	enfants
17	Alpes-Maritimes
17	camaïeu
17	documentaires
17	orchidée
17	pétanque
17	Sans-aucun-doute
17	saumon
17	steak
17	tulipe
17	Xantia
16	rôti
16	formule
16	gigot
16	mots
16	salade
16	série
16	faire
16	informatique
16	Celio
16	cerisier
16	chantilly
16	La-Voix-du-Nord
16	Nike
16	Nouvel-Observateur
16	Twingo
15	belle
15	nouvelle

« Les préférences des Français » - Fréquence des formes réduites

Avec lemmatisation	
Fréquence	**Forme réduite**
16	Ça-se-discute
16	Celio
16	cerisier
16	chantilly
16	La-Voix-du-nord
16	Nike
16	Nouvel-Observateur
16	Twingo
15	belle+
15	saint+
15	lys
15	poire+
15	tigre+
15	tricot+
15	voile+
15	jean
15	Libérat+ion
15	foie-gras
15	Lorraine
15	Rhône-Alpes
15	Strauss
14	violet+
14	Autriche
14	box+e
14	chou+
14	cuite+
14	marne+
14	modèle+
14	riz
14	télévision+
14	Jules
14	automobil<
14	spaghetti+
14	blanquette
14	céramique
14	Dion
14	Kiabi
14	Télérama
14	vtt
14	Égypte
13	clair+
13	farci+
13	grand+
13	Portugal
13	centre+
13	foret+

Sans lemmatisation	
Fréquence	**Forme réduite**
15	auto
15	lys
15	tigre
15	voile
15	jean
15	Libération
15	foie-gras
15	Lorraine
15	Rhône-Alpes
15	Strauss
14	violet
14	Autriche
14	Charente
14	bouteille
14	boxe
14	courses
14	cuite
14	fromage
14	marne
14	marques
14	modèle
14	riz
14	télévision
14	tricot
14	voyages
14	balade
14	Jules
14	blanquette
14	céramique
14	Dion
14	Kiabi
14	Télérama
14	vtt
13	Saint
13	Portugal
13	centre
13	fleur
13	foret
13	histoire
13	iris
13	mer
13	surprise
13	confit
13	regarder
13	basket
13	cabriolet

« Les préférences des Français » - Fréquence des formes réduites

Avec lemmatisation	
Fréquence	**Forme réduite**
13	histoire+
13	iris
13	mer+
13	pyramide+
13	sortie+
13	surprise+
13	confire.
13	crêpe+
13	basket
13	cabriolet
13	cassoulet
13	Collins
13	Francis
13	Gandhi
13	Languedoc-Roussillon
13	Lee
13	Léonard-de-Vinci
13	Louis
13	magret-de-canard
13	Navarro
13	Paris-Brest
13	Paris-Match
12	marron+
12	plat+
12	anima+l
12	disque+
12	pin+
12	restaurant+
12	santé+
12	connaître.
12	crois+er
12	ferr+er
12	promen+er
12	Laurent
12	agneau<
12	athlet<
12	basqu<
12	Bordeaux
12	Einstein
12	en-général
12	gratin-dauphinois
12	Hugo
12	infos
12	journal-télévisé
12	Mexique
12	new

Sans lemmatisation	
Fréquence	**Forme réduite**
13	cassoulet
13	Collins
13	Francis
13	Gandhi
13	Languedoc-Roussillon
13	Lee
13	Léonard-de-Vinci
13	Louis
13	magret-de-canard
13	Navarro
13	Paris-Brest
13	Paris-Match
12	gâteaux
12	pyramide
12	restaurant
12	santé
12	variétés
12	croises
12	Ferrat
12	préféré
12	promener
12	Laurent
12	agneau
12	athlétisme
12	petit
12	Bordeaux
12	Einstein
12	en-général
12	gratin-dauphinois
12	Hugo
12	infos
12	journal-télévisé
12	Mexique
12	New
12	Phil
12	polo
12	Sud-Est
12	Tatin
12	Thaïlande
12	Xm
11	Safrane
11	Afrique
11	Amérique
11	Chine
11	Tunisie
11	amitié

« Les préférences des Français » - Fréquence des formes réduites

Avec lemmatisation	
Fréquence	Forme réduite
12	Phil
12	polo
12	Sud-Est
12	tatin
12	Thaïlande
12	Xm
11	Safrane+
11	Afrique
11	Amérique<
11	Chine
11	Tunisie
11	amitié+
11	arc+
11	art+
11	bateau+
11	bouleau+
11	café+
11	feu+
11	lion+
11	marbre+
11	nature+
11	oeillet+
11	ordinateur+
11	papier+
11	pommier+
11	table+
11	veau+
11	enchaîn+er
11	fi+er
11	prim+er
11	neig+e
11	bleu-marine
11	break
11	caramel
11	Citroen-Xantia
11	ClaudeFrancois
11	Cooper
11	Drome
11	Duteil
11	Fiesta
11	magnolia
11	Marianne
11	pizza
11	Polynésie
11	Renault-Espace
11	Toyota

Sans lemmatisation	
Fréquence	Forme réduite
11	arc
11	bouleau
11	café
11	lion
11	marbre
11	ordinateur
11	papier
11	pommier
11	table
11	veau
11	enchaîne
11	Fiat
11	Prima
11	automobile
11	neige
11	bleu-marine
11	break
11	caramel
11	Citroen-Xantia
11	Claudefrancois
11	Cooper
11	Drome
11	Duteil
11	Fiesta
11	magnolia
11	Marianne
11	pizza
11	Polynésie
11	Renault-Espace
11	Toyota
11	Turbo
10	parisienne
10	Charente-Maritime
10	animaux
10	bague
10	bateau
10	chevaux
10	choux
10	couture
10	disque
10	faïence
10	film
10	lire
10	nature
10	pin
10	pomme

« Les préférences des Français » - Fréquence des formes réduites

Avec lemmatisation	
Fréquence	Forme réduite
11	Turbo
10	Suisse+
10	amour+
10	bague+
10	coutur+e
10	dessert+
10	faïence+
10	gratin+
10	lire+
10	progrès
10	raisin+
10	science+
10	soir+
10	tilleul+
10	cal+er
10	fait
10	plast+3
10	Ardèche
10	Arte
10	Atlantique
10	Beethoven
10	bolognaise
10	Bourgogne
10	Burton
10	cd
10	César
10	Chanel
10	Churchill
10	Colomb
10	crustacés
10	Fabian
10	jeanned
10	Laure
10	maxi
10	Obispo
10	panthère
10	Pavarotti
10	Teresa
10	Topsanté
10	Vendée
10	Vercingétorix
9	decorati+f
9	Ouest+
9	policier+
9	année+
9	bouillon+

Sans lemmatisation	
Fréquence	Forme réduite
10	prix
10	progrès
10	promenades
10	soir
10	tilleul
10	fait
10	Ardèche
10	Arte
10	Atlantique
10	Beethoven
10	bolognaise
10	Bourgogne
10	Burton
10	cd
10	César
10	Chanel
10	Churchill
10	Colomb
10	crustacés
10	Fabian
10	jeanned
10	Laure
10	maxi
10	Obispo
10	panthère
10	Pavarotti
10	Teresa
10	Topsanté
10	Vendée
10	Vercingétorix
9	claire
9	Ouest
9	rouges
9	bouillon
9	champignons
9	chiens
9	culture
9	famille
9	fraise
9	gratin
9	jeux
9	lait
9	Landes
9	noix
9	oeillet
9	purée

« Les préférences des Français » - Fréquence des formes réduites

Avec lemmatisation	
Fréquence	Forme réduite
9	broderie+
9	champignon+
9	cuisin+e
9	culture+
9	dauphin+
9	éclair+
9	famille+
9	lait
9	lande+
9	mousse+
9	noix
9	parent+
9	poche+
9	purée+
9	saule+
9	temps
9	théâtre+
9	tomate+
9	vacance+
9	violette+
9	import+er
9	occup+er
9	pass+er
9	venir.
9	Charles
9	Marguerite
9	Martin
9	Pascal
9	ancien<
9	mimosa+
9	montan+
9	ball
9	Caraïbes
9	carbonara
9	Derrick
9	endives
9	érable
9	Jaurès
9	Lara
9	man
9	Marseille
9	Michael
9	Morgan
9	Périgord
9	Seychelles
9	Sud-Ouest-France

Sans lemmatisation	
Fréquence	Forme réduite
9	saule
9	temps
9	théâtre
9	vacances
9	violette
9	voitures
9	balades
9	Calais
9	connaît
9	importe
9	occuper
9	Charles
9	marguerite
9	Martin
9	Pascal
9	basque
9	plastique
9	ball
9	Caraïbes
9	carbonara
9	Derrick
9	endives
9	érable
9	Jaurès
9	Lara
9	Man
9	Marseille
9	Michael
9	Morgan
9	Périgord
9	Seychelles
9	Sud-Ouest-France
9	Télé-Loisirs
9	tiramisu
9	volley
8	décoration
8	gironde
8	gris
8	marron
8	plat
8	pleureur
8	suisse
8	beige
8	dimanche
8	amour
8	années

« Les préférences des Français » - Fréquence des formes réduites

Avec lemmatisation	
Fréquence	**Forme réduite**
9	Télé-Loisirs
9	tiramisu
9	volley
8	girond+
8	gris+
8	peint+
8	plein+
8	pleureur
8	beige+
8	dimanche+
8	arbre+
8	beurre+
8	bifteck+
8	bronze+
8	cadeau+
8	chaîne+
8	classe+
8	fer+
8	hêtre+
8	mark+
8	objet+
8	revue+
8	soup+e
8	super
8	travaux
8	tribun+
8	écout+er
8	éta+yer
8	jou+er
8	Caroll
8	Claude
8	Jacques
8	Michel
8	Victor
8	Guignol<
8	baba
8	Beatle+
8	bordelais+
8	Bouches-du-Rhône
8	Cacharel
8	Cardin
8	Chrysler
8	crème-anglaise
8	éclair-au-chocolat
8	Eddy-Mitchell
8	frambois+

Sans lemmatisation	
Fréquence	**Forme réduite**
8	beurre
8	bijoux
8	broderie
8	bronze
8	chaîne
8	classe
8	cuisine
8	dauphin
8	fer
8	hêtre
8	parents
8	poche
8	poire
8	raisin
8	super
8	travaux
8	tribune
8	Corsa
8	Etam
8	jouer
8	préférée
8	voir
8	Caroll
8	Claude
8	Jacques
8	Michel
8	Victor
8	Guignols
8	baba
8	Bouches-du-Rhône
8	Cacharel
8	Cardin
8	Chrysler
8	crème-anglaise
8	crêpes
8	éclair-au-chocolat
8	Eddy-Mitchell
8	je-ne-sais-pas
8	judo
8	La-chance-aux-chans
8	mimosa
8	Montand
8	Nougaro
8	parmentier
8	Pc
8	Poitou

« Les préférences des Français » - Fréquence des formes réduites

Avec lemmatisation	
Fréquence	Forme réduite
8	je-ne-sais-pas
8	judo
8	La-chance-aux-chans
8	liégeois+
8	Nougaro
8	parmentier
8	pc
8	picard+
8	Poitou
8	rhum
8	rosbif
8	Rover
8	Science-et-Vie
8	Souchon
8	Télé-Star
8	Tf1
8	Turquie
8	vidéo
7	frai+c„
7	français+
7	particulier+
7	pédestre+
7	quotidien+
7	mère+
7	Ardennes
7	Brésil
7	Vosges
7	Limousin
7	M6
7	Macias
7	Modes-et-Travaux
7	Monospace
7	Nul-part-ailleurs
7	pivoine
7	Rodier
7	spencer
7	tarte-aux-fruits
7	Volvo
7	Vsd
7	yaourts
7	Zone-Interdite
7	million+
7	six
6	dernier+
6	gros+
6	précis+

Sans lemmatisation	
Fréquence	Forme réduite
8	rhum
8	rosbif
8	Rover
8	Science-et-Vie
8	Souchon
8	spaghettis
8	Télé-Star
8	Tf1
8	Turquie
8	vidéo
7	grands
7	peint
7	policiers
7	quotidien
7	noire
7	mère
7	Ardennes
7	Brésil
7	Vosges
7	ami
7	bifteck
7	canal
7	cartes
7	chats
7	coton
7	feux
7	files
7	jambon
7	jeu
7	lot
7	moulin
7	mousse
7	objet
7	photo
7	piano
7	planche
7	poires
7	poterie
7	revue
7	sciences
7	sortie
7	soupe
7	tapisserie
7	toile
7	vanille
7	variété

« Les préférences des Français » - Fréquence des formes réduites

Avec lemmatisation	
Fréquence	Forme réduite
6	spécia+l
6	uni+
6	Allemagne
6	Angleterre
6	Montpellier
6	actualité+
6	an+
6	argent
6	berger+
6	boite+
6	bord+
6	canne+
6	chiffre+
6	club+
6	couleur+
6	dépêche+
6	dessin+
6	dire+
6	échec+
6	grille+
6	lampe+
6	lettre+
6	loup+
6	méta+l
6	paquet+
6	presse+
6	repas
6	vase+
6	achet+er
6	sortir.
6	Barbara
6	Daniel
6	Yves
6	Avantag+e
6	complic+e
6	muscul<
6	religi<
6	A4
6	albisia
6	Bach
6	Bob
6	Bx
6	Calédonie
6	Charlie-Hebdo
6	Che-Guevara
6	Dauphine

Sans lemmatisation	
Fréquence	Forme réduite
7	zéro
7	Balavoine
7	écouter
7	Honore
7	Christine
7	Maurice
7	Astra
7	Beatles
7	Bonne-Soirée
7	Boss
7	Coluche
7	crudités
7	Damart
7	Dassin
7	daube
7	Décathlon
7	Devernois
7	devred
7	Diesel
7	Enrico
7	Escort
7	fraisier
7	géranium
7	glaïeul
7	hachis
7	hibiscus
7	Kennedy
7	liégeois
7	Limousin
7	M6
7	Macias
7	Modes-et-Travaux
7	monospace
7	Nul-part-ailleurs
7	pivoine
7	Rodier
7	spencer
7	tarte-aux-fruits
7	Volvo
7	Vsd
7	yaourts
7	Zone-Interdite
6	hautes
6	pédestre
6	précise
6	unis

« Les préférences des Français » - Fréquence des formes réduites

Avec lemmatisation	
Fréquence	**Forme réduite**
6	Dior
6	Elvis
6	J-Clerc
6	Jimmy
6	jt
6	Kenya
6	Lagaf
6	Madagascar
6	Manoukian
6	M-Griffon
6	Mitterrand
6	Piaf
6	Picasso
6	Queen
6	raclette
6	scrabble
6	SergeLama
6	sorbet+
6	stones
6	tout-ce-qui-est
6	Trenet
6	U2
5	argent<
5	ferre+
5	manuel+
5	matin+
5	naturel+
5	régiona+l
5	Garonne
5	acier+
5	ail
5	avenir+
5	bouquin+
5	breton+
5	cèdre+
5	charcut+1
5	collection<
5	comte+
5	courrier+
5	débat+
5	eau+
5	écho+
5	entrée+
5	forge+
5	fureur+
5	haricot+

Sans lemmatisation	
Fréquence	**Forme réduite**
6	blanche
6	verte
6	Allemagne
6	Angleterre
6	Montpellier
6	argent
6	arts
6	berger
6	bord
6	cadeau
6	chiffres
6	club
6	dépêche
6	desserts
6	Dire
6	éclair
6	lampe
6	lettres
6	livres
6	loup
6	marks
6	match
6	métal
6	pâte
6	plante
6	presse
6	repas
6	rosiers
6	sorties
6	vase
6	brûlée
6	sortir
6	Barbara
6	Daniel
6	Yves
6	classique
6	complice
6	musculation
6	A4
6	albisia
6	Bach
6	Bob
6	Bx
6	Calédonie
6	Charlie-Hebdo
6	Che-Guevara

« Les préférences des Français » - Fréquence des formes réduites

Avec lemmatisation	
Fréquence	Forme réduite
5	huître+
5	jardin+
5	langue+
5	micro+
5	mode+
5	moule+
5	mouton+
5	parasol+
5	peuplier+
5	piscine+
5	plateau+
5	séjour+
5	semaine+
5	soeur+
5	tir+
5	vache+
5	velours
5	visite+
5	week-end+
5	dans+er
5	donn+er
5	fondre.
5	grill+er
5	jur+er
5	liber+er
5	port+er
5	Alain
5	André
5	Elne
5	Renaud
5	Roméo
5	prés+ent
5	René
5	telephon+3
5	urg+ent
5	Aveyron
5	Biarritz
5	cèpes
5	châtaign+
5	Chirac
5	Chopin
5	chrysanthème
5	City
5	civet+
5	Clinton
5	Curie

Sans lemmatisation	
Fréquence	Forme réduite
6	dauphine
6	Dior
6	Elvis
6	JClerc
6	Jimmy
6	jt
6	Kenya
6	Lagaf
6	Madagascar
6	Manoukian
6	Mgriffon
6	Mitterrand
6	Piaf
6	Picasso
6	Queen
6	raclette
6	scrabble
6	SergeLama
6	spaghetti
6	Stones
6	tout-ce-qui-est
6	Trenet
6	U2
5	farcies
5	farcis
5	ferre
5	matin
5	particulière
5	pleine
5	Garonne
5	acier
5	ail
5	ans
5	art
5	avenir
5	boite
5	bouquin
5	canne
5	charcuterie
5	comte
5	courrier
5	eau
5	échecs
5	forge
5	Fureur
5	grille

« Les préférences des Français » - Fréquence des formes réduites

Avec lemmatisation	
Fréquence	**Forme réduite**
5	décapotable
5	documentaire
5	Elton-John
5	footing
5	Foucault
5	Friends
5	George
5	glace-à-la-vanille
5	Gti
5	gym
5	handball
5	hebdo
5	Hendrix
5	Hérault
5	hi-fi
5	Indonésie
5	info
5	Institt
5	jeans
5	je-suis
5	King
5	Lamborghini
5	laser
5	lorrain
5	Luther
5	magnétoscope
5	Mandela
5	Marley
5	Massif-Central
5	Méditerranée
5	muguet
5	ne-sais-pas
5	nougat
5	osier
5	Pink
5	pistache
5	pomme-de-terre
5	R5
5	Rolling
5	Schubert
5	Strait+
5	Touraine
5	Voix-du-Nord
5	Zara
5	Zx-Citroën
-	

Sans lemmatisation	
Fréquence	**Forme réduite**
5	huîtres
5	langue
5	meuble
5	micro
5	mode
5	moules
5	mouton
5	parasol
5	pieds
5	piscine
5	plateau
5	salades
5	séjour
5	soeur
5	tir
5	tomate
5	velours
5	vêtements
5	week-end
5	Jura
5	Passat
5	portable
5	André
5	Hélène
5	Renaud
5	Roméo
5	patin
5	présence
5	religieuse
5	rene
5	Téléphone
5	Aveyron
5	Biarritz
5	cèpes
5	Chirac
5	Chopin
5	chrysanthème
5	City
5	Clinton
5	crêpe
5	Curie
5	dahlia
5	décapotable
5	documentaire
5	Elton-John
5	footing

« Les préférences des Français » - Fréquence des formes réduites

Avec lemmatisation		Sans lemmatisation	
Fréquence	Forme réduite	Fréquence	Forme réduite
-		5	Foucault
-		5	framboisier
-		5	Friends
-		5	George
-		5	glace-à-la-vanille
-		5	Gti
-		5	gym
-		5	handball
-		5	hebdo
-		5	Hendrix
-		5	Hérault
-		5	hi-fi
-		5	hortensia
-		5	Indonésie
-		5	info
-		5	Instit
-		5	jeans
-		5	King
-		5	Lamborghini
-		5	laser
-		5	lorrain
-		5	Luther
-		5	magnétoscope
-		5	Mandela
-		5	Bob Marley
-		5	Massif-Central
-		5	Méditerranée
-		5	muguet
-		5	nougat
-		5	osier
-		5	Picardie
-		5	Pink
-		5	pistache
-		5	pomme-de-terre
-		5	R5
-		5	Rolling
-		5	Schubert
-		5	Touraine
-		5	Voix-du-Nord
-		5	Zara
-		5	Zx-Citroën

Dépôt légal : Septembre 1999

ISSN : 1257-9807

ISBN : 2-84104-141-7

CAHIER DE RECHERCHE

Récemment parus :

Utilisation de la modélisation statistique à des fins interprétatives
Bruno MARESCA, Pascale HÉBEL - **n°123 (1998)**

La dynamique interne du récit
Pierre LE QUÉAU, Mathieu BRUGIDOU - **n°124 (1998)**

Hétérogénéité des attitudes et comportements de consommation
Jean-Luc VOLATIER - **n°125 (1998)**

Les comportements des consommateurs européens
Ariane DUFOUR, Jean-Pierre LOISEL, Emmanuelle MAINCENT, Laurent POUQUET, Jean-Luc VOLATIER - **n°126 (1999)**

Eléments de méthode pour l'analyse du tissu économique local
Philippe MOATI, Stéphane LOIRE - **n°127 (1999)**

La construction sociale de la perception de la santé
Christine OLM, Pierre LE QUÉAU - **n°128 (1999)**

L'évolution des opinions et des comportements des séniors depuis vingt ans, en France
Franck BERTHUIT, Bertrand CHOKRANE, Georges HATCHUEL - **n°129 (1999)**

Le consommateur Français en 1998
Anne-Delphine BROUSSEAU, Jean-Luc VOLATIER - **n°130 (1999)**

Président : Bernard SCHAEFER Directeur Général : Robert ROCHEFORT
142, rue du Chevaleret, 75013 PARIS - Tél. : 01 40 77 85 01

ISBN : 2-84104-141-7

CRÉDOC

Centre de recherche pour l'Étude et l'Observation des Conditions de Vie