

CAHIER DE ReCHERCHE

SEPTEMBRE 1997



N° 101

TRAITEMENT DES QUESTIONS OUVERTES :
COMPARAISON D'UNE POSTCODIFICATION
ET DE MÉTHODES LEXICOMÉTRIQUE
ET D'ANALYSE DU DISCOURS

PATRICK BABAYOU

Département " Prospective de la Consommation "

Crédoc - Cahier de recherche. N°
0101. Septembre 1997.

CREDOC•Bibliothèque



CRÉDOC

 L'ENTREPRISE DE RECHERCHE



**Traitement des questions ouvertes :
Comparaison d'une postcodification et
de méthodes lexicométrique
et d'analyse du discours**

Patrick Babayou

CRÉDOC - Département Prospective de la Consommation

août 1997

142, rue du Chevaleret
7 5 0 1 3 - P A R I S

*« Le français : idiome idéal pour
exprimer délicatement des sentiments
équivoques. »*

E.M. Cioran, *Aveux et anathèmes*,
Gallimard 1987

SOMMAIRE

La recherche francophone en analyse textuelle : état de l'art	7
<i>Alceste</i> versus postcodification ou la genèse d'une complémentarité des approches	11
Plan de post codification	13
Traitement des réponses par l'analyse lexicale d' <i>Alceste</i>	18
<i>Étude de la lemmatisation</i>	18
<i>Le vocabulaire utilisé</i>	19
<i>Analyse de la classification des réponses</i>	23
<i>Les limites d'une description des attitudes</i>	30
Croisement de la classification d' <i>Alceste</i> et de la postcodification	32
<i>Contexte d'étude</i>	32
<i>Étude des correspondances entre les classes</i>	33
Conclusion	45
L'analyse textuelle et les approches par graphes.....	48
Présentation des possibilités du logiciel <i>Tropes</i>	49
<i>Définitions et concepts utilisés</i>	49
<i>Principales étapes de l'analyse d'un corpus</i>	51
<i>Mode d'entrée dans le logiciel</i>	53
Trois exemples d'applications du logiciel <i>Tropes</i>	54
<i>Rappels sur les trois corpus utilisés</i>	54
<i>Résultats pour le corpus « Température idéale »</i>	55
<i>Résultats pour le corpus « bonheur »</i>	61
<i>Résultats pour le corpus « prospective »</i>	66
Synthèse des résultats obtenus par le logiciel <i>Tropes</i>	71
Perspectives	73
Bibliographie	75

SOMMAIRE DES TABLEAUX

Tableau 1. Catégories de réponses post codées	13
Tableau 2. Codage des « sans opinion »	14
Tableau 3. Codage des « caution scientifique »	15
Tableau 4. Codage des « attitudes pragmatiques »	16
Tableau 5. Codage des « certitudes absolues »	16
Tableau 6. Codage des « erreurs »	17
Tableau 7. Formes réduites les plus fréquentes dans le corpus	19
Tableau 8. Classification des réponses par Alceste	23
Tableau 9. Description du premier contexte	25
Tableau 10. Description du deuxième contexte	26
Tableau 11. Description du troisième contexte	27
Tableau 12. Description du quatrième contexte	28
Tableau 13. Ventilation des classes de réponses construites par Alceste	32
Tableau 14. Profil des catégories du plan de codage en fonction des classes construites par Alceste	34
Tableau 15. Classement par Alceste de la catégorie « sans opinion »	35
Tableau 16. Classement par Alceste de la catégorie « emballage du produit »	36
Tableau 17. Classement par Alceste de la catégorie « Température de présentation en magasin »	37
Tableau 18. Classement par Alceste de la catégorie « Confiance absolue dans le réfrigérateur »	38
Tableau 19. Classement par Alceste de la catégorie « Expérience acquise »	39
Tableau 20. Classement par Alceste de la catégorie « Température idéale »	40
Tableau 21. Classement par Alceste de la catégorie « Place des produits »	41
Tableau 22. Classement par Alceste de la catégorie « Mode d'emploi du réfrigérateur, recours à un spécialiste »	42
Tableau 23. Classement par Alceste de la catégorie « Dates limites »	43
Tableau 24. Classement par Alceste de la catégorie « Aspect du produit »	44
Tableau 25. Exemple de catégories et références utilisées par Tropes	50
Tableau 26. Styles et mises en scène selon Tropes	51
Tableau 27. Les trois corpus testés avec Tropes	54
Tableau 28. Qualification du corpus « Température idéale »	55
Tableau 29. Phrases illustrant le style du corpus « Température idéale »	56
Tableau 30. Phrases définissant la mise en scène du corpus « Température idéale »	57

Tableau 31. Univers de référence du corpus « Température idéale »	58
Tableau 32. Mises en relations significatives du corpus « Température idéale »	58
Tableau 33. Les vocabulaires utilisés en 1992 et 1996 pour répondre à la question « Qu'est-ce qu'être heureux ? »	62
Tableau 34. Univers de référence des deux textes composant le corpus « bonheur »	63
Tableau 35. Thèmes abordés dans les entretiens du corpus « prospective »	67
Tableau 36. Principaux univers de référence du corpus « prospective »	67
Tableau 37. Qualification globale des neuf entretiens du corpus « prospective »	68
Tableau 38. Principaux univers de référence des entretiens du corpus « prospective »	69

SOMMAIRE DES ILLUSTRATIONS

Figure 1. Distribution des formes les plus fréquentes du corpus.....	22
Figure 2. Dendrogramme de la classification descendante.....	24
Figure 3. Graphe du corpus « Température idéale » centré sur l'univers « alimentation »	59
Figure 4. Graphe du corpus « Température idéale » centré sur l'univers « confiance » et phrases types.....	60
Figure 5. Évolution des thématiques.....	64
Figure 6. Graphes des principaux univers de référence du corpus « bonheur » en 1992 et en 1996	65
Figure 7. Graphe du corpus « prospective ».....	70

La recherche francophone en analyse textuelle : état de l'art

Depuis plusieurs années¹, le CRÉDOC a entrepris des travaux sur l'analyse automatique des textes, privilégiant et soutenant en particulier le développement du logiciel *Alceste* de Max Reinert. De nombreuses publications de recherche et d'études des équipes du CRÉDOC témoignent de la vitalité de ces applications dans des contextes de traitements variés : textes littéraires², données biographiques³, structures narratives⁴, réponses ouvertes ou entretiens lors d'enquêtes d'évaluation, d'attitudes ou de comportements⁵, ou encore d'analyse de corpus de réponses en français et en anglais⁶. Dans l'ensemble, l'utilisation d'*Alceste* a été privilégiée et n'a en fait été confrontée à d'autres méthodes que dans les travaux les plus anciens. Il s'agissait alors de valoriser les possibilités offertes par ce logiciel et de les comparer à celles d'autres offres, tel SPAD-T.

L'utilisation et la participation à l'amélioration du logiciel *Alceste* continuent de faire partie des objectifs de recherche en analyse textuelle au CRÉDOC. Les méthodes statistiques descriptives utilisées par ce logiciel et par d'autres n'ont cependant pas évolué de manière importante depuis plusieurs années. Nous nous plaçons en effet là du point de vue de l'école française d'analyse des données, inspirée en particulier par Benzécri. L'apport considérable d'*Alceste* reste, vis-à-vis des travaux du CRÉDOC, d'offrir l'avantage de relier l'analyse des questions

¹ Cf. en particulier Hauesler, 1987 ; Yvon, 1990

² Beaudouin, Hébel, 1994 ; Beaudouin, 1993

³ Bauer, Maresca, 1992 ; Beaudouin, Aucouturier, 1994

⁴ Beaudouin, 1995

⁵ Gilles, 1995 ; Collerie de Borely, 1996

⁶ Beaudouin, Brochet, 1996

ouvertes par post codification et une automatisation de cette tâche coûteuse, par le biais d'une classification (descendante) des réponses.

Le contexte de ces méthodes ne doit toutefois pas faire oublier l'étendue des recherches qui sont menées, en particulier dans le monde francophone, sur l'analyse textuelle. Les techniques mises en œuvre se diversifient et se diffusent dans des disciplines variées, auxquelles le CRÉDOC s'intéresse : sciences humaines et sociales, psychologie cognitive, sémiologie, littérature comparée en sont des domaines d'application traditionnels. De nouveaux champs d'application émergent aussi, comme par exemple le développement des « moteurs de recherche » sur le réseau Internet.

Il nous a semblé raisonnable, aujourd'hui, de décrire l'état de l'art de l'analyse textuelle, au sens le plus varié du terme.

*

* *

De manière paradoxale, nous pouvons commencer cet « état de l'art » par quelques rappels en forme de justification de la démarche d'automatisation de l'analyse du discours. La littérature sur les justifications du recours ou du non recours à des techniques d'analyse textuelle est en effet abondante. De fait, nous supposons depuis plusieurs années au CRÉDOC qu'une question sous-jacente est d'emblée résolue : ces techniques ont une utilité, au moins dans le cas du traitement des questions ouvertes qui est celui auquel les équipes d'études et de recherches sont le plus usuellement confrontées.

Il subsiste toutefois un grand nombre de questions non résolues par ce seul postulat. Un deuxième niveau de remise en cause des techniques repose non plus sur la pertinence globale de leur utilisation mais sur les pertinences respectives des différents types d'approches.

En tout état de cause, nous n'apportons pas, ici ou ailleurs, de réponse définitive. Nos objectifs peuvent être mesurés d'une manière purement économique et, à ce titre, converger avec les objectifs et intérêts d'autres utilisateurs de l'analyse textuelle. Nous sommes confrontés en effet à la recherche d'une utilisation optimale des possibilités de questionnements dans un contexte d'enquête sociologique. A ce titre, la variété des modes d'interrogations doit être

préservée, même si les questions fermées passent pour les plus usuelles, voire les plus pratiques à traiter par des individus fortement influencés par leur culture statistique. Les questions ouvertes gardent une utilité décisive malgré les réserves qui peuvent être émises à leur égard¹, en permettant de s'assurer de ne pas enfermer les personnes interrogées dans une logique de réponses trop orientées par les suggestions faites.

La spontanéité donne ainsi un sens à des questions qui, si elles étaient fermées, n'en auraient guère. Par exemple, on ne pourrait interpréter une question comme celle que nous utilisons, « *pour vous, qu'est-ce qu'être heureux ?* », alors qu'une des réponses suggérées serait : « *avoir plus d'argent* ». L'argent étant indépendant du bonheur dans les proverbes seuls, il semble évident que dans une telle configuration de questionnement les personnes interrogées auraient tendance à se rallier massivement à cette réponse, quelle que soit leur situation financière personnelle ou la vision qu'elles ont de la société. La citation spontanée serait plus rare et, par conséquent, plus interprétable.

Dans une logique différente, il peut être souhaitable de poser une question sous une forme ouverte dans le cas où l'on mesure un phénomène dont on a une faible connaissance *a priori*. C'est par exemple le cas de la question que nous utiliserons : « *D'après vous, comment peut-on faire pour connaître la bonne température de conservation d'un produit frais ?* ».

Il faut donc orienter notre réflexion et nos efforts non vers la justification de la méthode d'interrogation ouverte mais bien sur la réduction du coût que représente le traitement analytique de ces questions. L'orientation de notre recherche actuelle et de l'état des lieux des connaissances que nous tentons de faire ici a donc pour objectif majeur de considérer le recours à l'analyse textuelle d'un point de vue strictement économique et d'établir par quelles méthodes il est possible de réduire le coût de traitement d'une question ouverte et selon quelle automatisation du processus de dépouillement.

¹ Cf. en particulier la critique générale que fait Bourdieu (1980) sur la possibilité de faire une analyse de l'opinion publique. Il n'existe pas nécessairement une comparabilité des réponses entre plusieurs individus lorsqu'on pose une question ouverte, chacun répondant en se situant dans un espace de représentations individuelles. Même le statisticien peut être troublé par cette limite épistémologique.

Enfin, et c'est le sens de notre exergue, il ne faut pas négliger la difficulté que représente l'interprétation automatisée d'un discours. Les ambiguïtés qui troublent souvent les analyses doivent être considérées comme un bruit statistique qu'il faut parvenir à contrôler.

*

* *

Ce cahier de recherche a pour objectif de passer en revue différents outils disponibles, en comparant leurs approches et en illustrant leurs différences à l'aide d'exemples concrets d'applications extraits d'enquêtes rendues publiques par le CRÉDOC au cours des trois dernières années. Nous décrivons de cette façon les grandes familles de techniques d'analyse des questions ouvertes actuellement appliquées, tant en France que dans d'autres pays francophones (Belgique, Québec, etc.).

Le premier exemple que nous proposons a trait aux différences qui existent entre les résultats d'une post codification et d'un traitement par l'analyse d'*Alceste* de réponses à une question ouverte décrivant un comportement de consommation, qui avait été posée en 1995 lors de l'enquête du CRÉDOC sur les comportements alimentaires¹. Cherchant d'abord à opposer ces deux approches, nous montrons en quoi elles peuvent être considérées en termes de complémentarité.

Le deuxième exemple exploite les possibilités d'une méthode d'analyse cognitivo-discursive, qui s'apparente à la théorie des graphes. Le logiciel *Tropes* a été utilisé dans cette phase. Le même corpus de réponses à une question ouverte que dans le premier exemple est utilisé dans un premier temps. Les réponses à une question ouverte posée en 1992 et 1996 sont ensuite traitées, cette question appartenant à l'enquête annuelle du CRÉDOC sur la consommation : « *Pour vous, qu'est-ce qu'être heureux ?* ». Les possibilités de cette approche sont illustrées aussi par le traitement d'un corpus composé d'entretiens semi-directifs auprès d'experts sur la prospective de la mobilité locale des personnes âgées².

¹ Babayou, 1995

² Babayou, Volatier, 1997

***Alceste* versus post codification ou la genèse d'une complémentarité des approches**

Nous entreprenons dans cette partie de comparer les résultats obtenus par la post codification et l'analyse par *Alceste* d'une question ouverte posée dans une enquête de comportements¹.

Sans donner lieu ici à une querelle d'écoles entre tenants de la post codification et héros de la l'analyse textuelle, nous nous interrogeons froidement sur leurs applications au domaine de la sociologie comportementale. Il est évident que les résultats peuvent être congruents sans pour autant coller strictement l'un à l'autre, ne serait ce par exemple que du fait de la structure hiérarchique de la typologie d'*Alceste* alors qu'une post codification autorise les codages multiples des réponses.

La question qui nous servira de support pour décortiquer les différences entre les deux méthodes est extraite de notre enquête sur les comportements alimentaires des Français. A la fin de l'année 1994, à la demande de la DGCCRF², le CRÉDOC a interrogé en face à face 1603 ménagères sur l'utilisation qu'elles font de leur réfrigérateur d'une part, et d'autre part sur leurs habitudes de consommation et leurs croyances en matière de nutrition³. Cette enquête comprenait la question ouverte suivante :

« D'après vous, comment peut-on faire pour connaître la bonne température de conservation d'un produit frais (comme les volailles ou les yaourts) ? »

¹ Ce type de comparaison a déjà été tenté par ailleurs (cf. par exemple Blot, 1993) mais restait inédit dans les réflexions du Crédoc alors même que, nous l'avons souligné, elle concerne directement l'une des principales évolutions de nos travaux d'études et de recherche depuis une dizaine d'années.

² Direction Générale de la Concurrence, de la Consommation et de la Répression des Fraudes

³ Babayou, 1995

Il est important de noter immédiatement que, outre une formulation qui inspirait à la fois des réponses en termes d'opinions et de comportements, la question traitée a donné lieu à un nombre très important de non réponses, près de la moitié des ménagères déclarant explicitement ne pas savoir comment déterminer une température idéale de stockage d'un produit frais (sans pour autant exprimer d'une seule manière cette ignorance). Nous avons de plus montré lors de l'exploitation de l'enquête que les réponses à cette question, loin de donner une idée précise des usages des ménagères, avaient plutôt pour résultat de permettre de dessiner une image de l'espace des représentations mentales des ménagères vis-à-vis de l'utilisation des appareils producteurs de froid et, plus généralement, de la conservation des aliments.

Les développements qui suivent ont pour but de rendre compte tout d'abord du plan de codification du corpus des réponses, puis de traiter avec *Alceste* ce même corpus, en réalisant l'analyse la plus simple qu'il soit possible de réaliser avec ce logiciel.

Nos objectifs sont donc les suivants :

- établir une grille de lecture d'une classification descendante de données textuelles dans un cas de présence de nombreuses réponses manquantes ;
- observer comment le lexique se structure automatiquement par l'algorithme de classification d'*Alceste*, et en quoi cette structure peut converger vers celle d'un plan de post codification ;
- observer le classement par *Alceste* des individus qui sont post codés en réponses multiples.

Plan de post codification

L'examen attentif du corpus des réponses à la question sur la connaissance d'une bonne température de conservation des produits frais a permis de dégager cinq types d'attitudes, elles-mêmes décomposables en deux ou trois sous thèmes. Le plan de post codification autorise la reconstitution de réponses multiples.

Tableau 1. Catégories de réponses post codées

Catégorie	Part	Sous Catégorie	Part
Sans opinion	46,6%		
Caution scientifique	30,8%	référence à l'emballage du produit	25,8%
		référence à la température de présentation du produit en magasin	5,0%
Attitudes pragmatiques	17,1%	confiance absolue dans le réfrigérateur	13,8%
		l'expérience acquise	3,3%
Certitudes absolues	12,1%	croyance en une température idéale précise	4,6%
		place des produits dans le réfrigérateur	2,6%
		mode d'emploi du réfrigérateur, recours à un spécialiste	4,9%
Erreurs	4,8%	référence aux dates limites	2,2%
		aspect du produit	2,6%

Source : Enquête Comportements et attitudes des ménagères à l'égard du réfrigérateur et du froid, CRÉDOC, DGCCRF 1995

Ce plan de codage cherchait à rendre compte des différentes attitudes face à des réponses d'opinion et des réponses de comportements, ce qui n'était pas toujours facile, d'autant moins que la question posée n'appelle pas une mise en relation nécessaire entre les réponses et les pratiques domestiques des ménagères. Néanmoins, les classes de réponses « pragmatiques » et

« erreurs » font indéniablement référence à un cadre pratique, ce qui n'est pas forcément le cas des autres.

Les tableaux suivants donnent un détail plus fin du plan de codage.

Tableau 2. Codage des « sans opinion »

Sous catégories	Interprétation
« sans opinion »46,6%	ménagères déclarant explicitement ne pas savoir comment déterminer une température idéale de stockage d'un produit frais
Illustration « Non, je ne vois pas » « Aucune idée » « On ne peut pas savoir » « Il faut être au courant » « Peuchère, là, vous me posez quelque chose ! » « C'est le genre de question que je ne me suis jamais posée » « Je ne m'en occupe pas » « Alors là, je n'en sais rien, c'est la grosse question ! »	

Source : Enquête Comportements et attitudes des ménagères à l'égard du réfrigérateur et du froid, CRÉDOC, DGCCRF 1995

Ce premier type de réponses n'est pas aussi simpliste que l'on pourrait le penser *a priori*. Si les réponses du genre « *je ne sais pas* », « *je ne vois pas* » sont effectivement très fréquentes, il existe une foule de réponses plus originales qui ont été rattachées à la classe des « sans opinion ».

Il existe aussi des nuances dans le mode de réponse, visibles y compris sur l'échantillon que nous donnons dans le tableau précédent, selon par exemple la présence ou non d'une idée de généralisation de l'absence d'opinion (je n'ai pas d'opinion, et il n'y a donc pas de raison pour que d'autres ménagères en aient une...), ou aussi selon l'absence totale d'intérêt pour la question (« *Je ne m'en occupe pas* ») ou au contraire une sorte de frustration de ne pas connaître la réponse (« ... *c'est la grosse question !* »).

On notera enfin que le codage « sans opinion » s'accompagne, par le jeu des réponses multiples recomposées, d'un autre codage dans environ 10% des cas. Non seulement les enquêteurs avaient pour consigne de relancer la question en cas de non réponse mais en plus de

nombreuses personnes ont manifestement répondu spontanément qu'elles ne savaient pas avant de se raviser (« ... *mais ça doit être écrit quelque part* »).

Tableau 3. Codage des « caution scientifique »

Sous catégories	Interprétation
« emballage du produit »25,8% « reproduire la température »..... 5,0%	recherche d'une information objective en particulier sur l'emballage des produits, de manière à trouver une réponse en forme de justification scientifique d'une opinion ou d'une pratique personnelle
Illustration « C'est généralement inscrit sur l'emballage, du moins cela devrait » « C'est écrit dessus » « C'est marqué dessus » « Il faut lire » « Sur l'emballage » « En regardant l'étiquette » « Je pense qu'elle est parfois indiquée sur l'emballage » « Je prends dans les rayons frais, donc je mets au frais » « En regardant le thermomètre du présentoir du magasin »	

Source : Enquête Comportements et attitudes des ménagères à l'égard du réfrigérateur et du froid, CRÉDOC, DGCCRF 1995

Les ménagères qui se situent dans cette classe de réponses ont tendance à rechercher une solution au problème posé à travers une attitude visant à privilégier la mise en application d'un jugement de valeur sous couvert de scientificité. Ainsi, elles vont faire confiance au fabricant (pour l'emballage), ou encore au distributeur (pour le présentoir ; il s'agit de chercher à reproduire chez soi la température du rayon froid du magasin ou l'on s'approvisionne en produits frais), réputés tous deux raisonnables et efficaces pour tout conseil concernant la conservation des produits.

Les réponses sont par conséquent très typées, voire sibyllines : la référence à l'emballage ou à l'étiquette est très fréquente, mais elle peut être sous entendue (« *C'est écrit dessus* », « *Il faut lire* »). Quant à la référence au distributeur, il n'est pas toujours simple de la repérer par des phrases typées, sinon par une évocation du rayon, de la notion de froid, ou encore du magasin ou d'un commerçant.

Tableau 4. Codage des « attitudes pragmatiques »

Sous catégories « référence au réfrigérateur » 13,8% « l'expérience acquise » 3,3%	Interprétation référence à des comportements surtout fondés sur une habitude dont l'origine reste souvent confuse pour la ménagère elle-même, et plus généralement sur la mise en pratique des comportements qui ont fait leurs preuves
Illustration « C'est le frigo qui décide » « Je fais confiance à mon réfrigérateur » « Les températures sont standard, les températures du frigo sont adéquates » « S'acheter un frigo qui est aux normes homologuées, et déjà ça va mieux » « Garder au frais à une certaine température » « Au début on essaie, puis lorsqu'on a trouvé la bonne position on laisse »	

Source : Enquête Comportements et attitudes des ménagères à l'égard du réfrigérateur et du froid, CRÉDOC, DGCCRF 1995

Les réponses de cette classe s'orientent vers la consolidation de pratiques qui ont fait peu ou prou leurs preuves du point de vue des ménagères qui les donnent.

Le codage a, en conséquence, été attribué en fonction d'une référence au réfrigérateur —et, sans distinction, à ses différentes déclinaisons plus ou moins familières : frigidaire, frigo...—, à l'idée de confiance, d'essais et de réussite. La référence à l'existence, au moins supposée, d'une échelle standardisée des températures de conservation a aussi été déterminante pour le codage de ce type de réponse.

Tableau 5. Codage des « certitudes absolues »

Sous catégories « croyance en une température » 4,6% « position des produits » 2,6% « mode d'emploi, recours à un spécialiste » ... 4,9%	Interprétation réponses données par des ménagères qui sont sûres de connaître la solution au problème posé, en faisant toutefois toujours référence, explicitement ou non, à l'utilisation d'un réfrigérateur
Illustration « Le froid, c'est 4 à 8 degrés. On le sait depuis toujours » « 4 degrés : j'ai travaillé dans l'alimentation » « On le sait : la viande c'est au bas du frigo, le yaourt c'est en haut » « Suivant l'étagé de mon réfrigérateur : c'est fait pour » « Il n'y a qu'à avoir le bouquin vendu avec le congélateur » « Je fais confiance au technicien qui m'a installé le réfrigérateur et je n'y touche plus » « C'est mon fils qui l'a réglé »	

Source : Enquête Comportements et attitudes des ménagères à l'égard du réfrigérateur et du froid, CRÉDOC, DGCCRF 1995

Les réponses qui ont été codées ici sont assez clairement identifiables par la référence à des notions très précises : référence à une température donnée (et utilisation du terme « degré »), référence à des positions (référentiel « haut-bas », utilisation du mot « étage »), ou encore référence à un « livre », ou à un spécialiste de la question : l'installateur du réfrigérateur, qui agit forcément *ès qualités*, qu'il soit ou non professionnel.

On remarquera, au niveau du sens à donner aux réponses qui ont été codées ici, que les références à une certitude absolue ne s'accompagnent pas nécessairement d'une attitude propre à engendrer un comportement sûr d'un point de vue d'hygiène alimentaire. Cette classe de réponses, comme la suivante, permet donc de visualiser, dans le plan de codage, des attitudes particulières à la frontière entre opinion et comportement.

Tableau 6. Codage des « erreurs »

Sous catégories	Interprétation
« référence aux dates limites » 2,2%	réponses caractéristiques de ménagères qui ont contourné la question et répondent plutôt à une interrogation du type : « pendant combien de temps peut-on conserver sans risque un aliment ? »
« aspect du produit » 2,6%	
Illustration	
« La date, la première chose c'est la date »	
« Si c'est frais, ça se voit »	
« Il faut mirer les oeufs »	
« A l'odeur »	

Source : Enquête Comportements et attitudes des ménagères à l'égard du réfrigérateur et du froid, CRÉDOC, DGCCRF 1995

Cette dernière classe de réponses étudiées lors de la post codification est particulièrement intéressante puisqu'elle est caractéristique d'individus qui ne répondent absolument pas à la question posée, mais se situent en dehors du champ d'interrogation. On ne fait jamais référence ici à une quelconque température, puisque le problème se situe ailleurs : tout aliment conservé l'est forcément à une température satisfaisante, le point crucial n'est donc pas *comment* conserver les produits, mais *combien de temps* voire *dans quel état*.

Cette idée apparaît clairement lorsqu'il s'agit de faire référence aux dates limites de consommation, et indirectement lorsque les enquêtés font appel aux qualités organoleptiques du produit (aspect général, couleur, odeur...).

Traitement des réponses par l'analyse lexicale d'*Alceste*

Étude de la lemmatisation

La première étape qu'effectue *Alceste* dans le traitement du corpus d'une question ouverte consiste à lemmatiser les formes ou expressions employées par les enquêtés. Il en découle une liste des mots réduits, que nous donnons par la suite en ordre de fréquence décroissante. L'exécution de cette étape, plutôt bien menée dans notre contexte d'étude par *Alceste*, permet toutefois de relever quelques premiers ajustements nécessaires avant toute analyse statistique :

- regroupement sous la même racine « réfrigérateur+ » des mots « réfrigérateur », « frigidaire » et « frigo », employés indifféremment par les enquêtés. Il n'y a donc pas lieu d'introduire une dispersion des réponses selon l'emploi d'un des trois mots, cette séparation semblant artificielle.
- rattachement au verbe « mettre » des formes « mets » qui apparaissent 100 fois dans le corpus, et jamais dans le sens nominal qui pourrait être possible, et « mise » qui n'est pas ici de la famille du verbe « miser ».
- rattachement au verbe « lire » des formes « lire » (prise par défaut au sens nominal de monnaie), « lis » et « lit » (prise elle aussi au sens nominal).
- rattachement au verbe « marquer » de la forme « marque », qui n'est jamais employée par les enquêtés à un sens nominal, mais toujours dans des expressions du type : « *c'est marqué dessus* ».
- création d'une locution « aucune idée », afin de distinguer les différents usages du mot « idée », employé par différents individus.

- « livre », « livret », « notice », « bouquin », « brochure », « manuel » se rapportent indifféremment au mode d'emploi du réfrigérateur — nous créons aussi l'expression « mode d'emploi » pour la faire apparaître et la rattacher à ce sens—, et nous allons donc les réduire à la même racine « notice+ » afin à nouveau de ne pas provoquer une dispersion artéfactuelle des réponses.

Ces corrections ne sont pas négligeables puisqu'elles affectent plus de la moitié des réponses¹.

Les ajustements effectués permettent de clarifier la description des réponses en particulier en évitant de prendre en compte des niveaux de langue qui n'ont pas d'intérêt direct pour une description des comportements (« frigo » vs. « réfrigérateur », « bouquin » vs. « notice »...)

Le vocabulaire utilisé

Après l'étape de lemmatisation, le corpus comprend principalement les formes indiquées dans le tableau suivant.

Tableau 7. Formes réduites les plus fréquentes dans le corpus

Forme	Fréquence	Forme	Fréquence	Forme	Fréquence
réfrigérateur+	338	notice+	38	paquet+	18
marqu+er	210	date+	35	bonne+	18
mettre.	197	frais	34	conserv+er	17
emballage+	166	lire.	31	froid+	17
regard+er	134	congélateur	28	habitu<	17
température+	87	voir.	27	demand+er	16
thermomètre+	78	question+	25	attenti+f	15
produit+	74	thermostat	24	principe+	15
étiquette+	71	pos+er	23	aliment+	13
degré+	68	regardant+	23	jour+	13
aucune-idée	59	achet+er	22	confi+ant	13
écrit<	58	yaourt+	22	fait	13
indiqu+er	42	consomm+er	19	normalement	13
faire.	39	viande+	19		

¹ Notons que les imprécisions relevées dans ce traitement indiquent les limites d'A/ceste en tant qu'analyseur syntaxique, ce qui ne nous intéresse pas au premier chef ici, puisque nous sommes essentiellement intéressés aux aspects statistiques du logiciel.

Remarquons d'emblée qu'en dépit du libellé de la question qui mentionnait deux exemples de produits —les volailles et le yaourt—, il n'y a pas un fort effet d'écho dans les réponses données par les enquêtés. Il n'a donc pas semblé nécessaire de donner de statut illustratif à ces deux mots. On s'était d'ailleurs aperçu au niveau de la post codification que les individus mentionnant l'un ou l'autre se rattachaient tous à des comportements précis, en particulier la position des aliments ou la mention d'une température de conservation. L'utilisation de ces termes n'intervenait par conséquent qu'à titre d'exemple.

La distribution des mots employés par les enquêtés permet de distinguer dans une première approche du mode de réponse à la question la prédominance de l'utilisation de cinq formes réduites :

1. **réfrigérateur+**, qui est de loin la notion la plus souvent citée par les enquêtés
2. **marqu+er**, qui est presque exclusivement employé à une voie pronominale (« *C'est marqué dessus* »)
3. **mettre.**, employé souvent en association avec des formes telles « réfrigérateur+ » ou « frais » (« *mettre au réfrigérateur* », « *mettre au frais* »), ou encore dans une expression comme : « *mettre le thermostat à...* »
4. **emballage+**, employé souvent comme complément d'objet du verbe suivant :
5. **regard+er**

On ne s'étonnera pas ici de l'absence de formes évoquant l'absence d'opinion (qui constituaient tout de même 46,6% des réponses, comme on l'a vu lors de la post codification du corpus), celle-ci entraînant la plupart du temps l'utilisation de formules lapidaires constituées le plus souvent de mots outils (« ne », « sait », « pas ») qui ne participent pas à

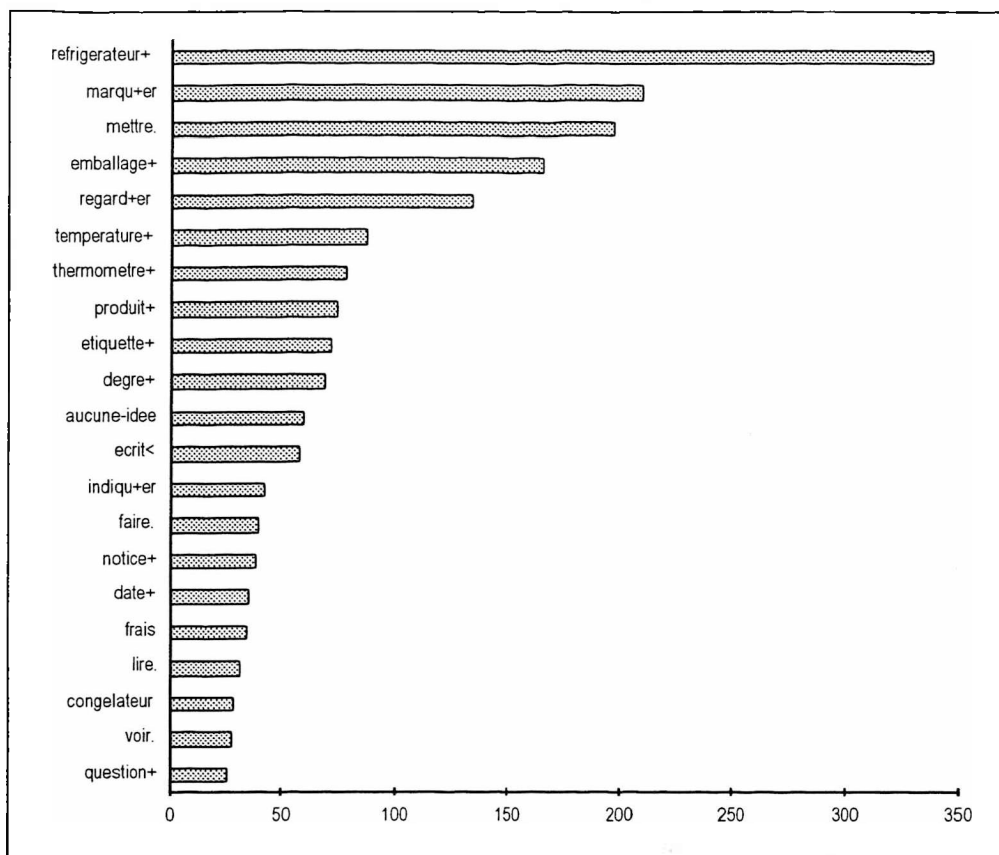
l'analyse dans *Alceste*¹. Nous reviendrons par la suite plus précisément sur le problème posé par la masse importante de non réponses spécifiques à notre corpus.

Le graphique qui suit (Figure 1) permet de mieux visualiser les poids relatifs des différentes formes utilisées par les enquêtés.

On voit en particulier que, outre les cinq formes déjà citées, sept autres se détachent nettement : **température+**, **thermomètre+**, **produit+**, **étiquette+**, **degré+**, **aucune-idée**, **écrit<**.

¹ Il aurait été possible de leur donner un statut « actif » dans l'analyse mais alors une perturbation supplémentaire était introduite par le statut particulier donné à la présence d'une négation en général dans les réponses des personnes interrogées.

Figure 1. Distribution des formes les plus fréquentes du corpus



Guide de lecture : en abscisse figurent les fréquences d'apparition des formes indiquées en ordonnées. Les barres sont donc proportionnelles au nombre d'occurrences des formes dans le corpus. Cet histogramme permet par conséquent de se rendre compte des disparités d'utilisation des différentes formes, à travers l'apparition de « coudes » dans la distribution.

Le simple examen du tri à plat des fréquences d'apparition des formes réduites dans le corpus permet déjà de retrouver les directions prises lors de la post codification, en particulier avec l'apparition des termes « emballage » et « étiquette » (réponses rattachées à la notion de caution scientifique relevée précédemment), « température » et « réfrigérateur » (termes apparaissant dans les réponses « pragmatiques-référence au réfrigérateur »), ou encore « livre » et « notice » (codage « certitude absolue-mode d'emploi »). Cependant, il ne s'agit là que d'une impression, et elle n'a qu'une valeur indicative, ne présupant pas de la suite des événements.

Analyse de la classification des réponses

La classification que fait *Alceste* des réponses permet de mettre en évidence quatre classes de réponses, auxquelles il faut bien entendu ajouter les réponses non classées par l'algorithme de constitution des classes. Le tableau suivant donne les parts respectives de ces classes dans l'ensemble des réponses :

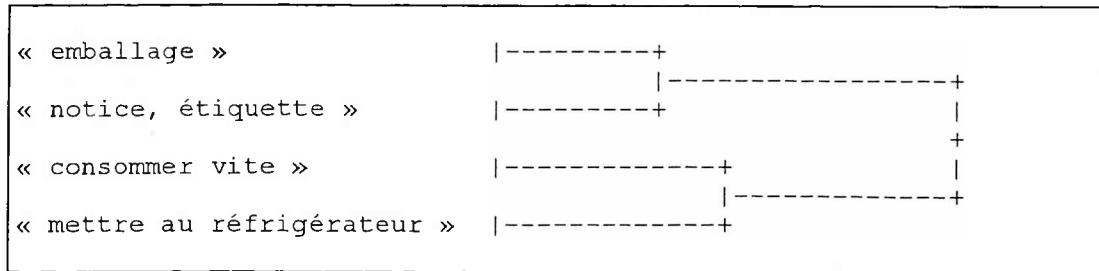
Tableau 8. *Classification des réponses par Alceste*

Classe de réponses	Part	
	Classées	Ensemble
réponses non classées		41,2%
« regarder l'emballage »	30,4%	17,9%
« lire la notice ou l'étiquette »	15,9%	9,3%
« consommer vite »	9,0%	5,3%
« mettre au réfrigérateur »	44,7%	26,3%
	100,0%	100,0%

Si l'on considère, à ce niveau d'analyse, que les réponses non classées correspondent majoritairement à des enquêtés qui n'émettent pas d'opinion, on voit que la classification automatique d'*Alceste* a permis de mettre en évidence cinq grands types de réponses, c'est-à-dire autant que le plan de codage en prévoyait. Si les groupes du tableau précédent peuvent, sans formuler d'hypothèse forte, se rapprocher des catégories de la postcodification, on voit néanmoins qu'il n'apparaît pas clairement une classe de réponses « erreurs ». En revanche, une nouvelle classe de réponses se forme avec l'utilisation des formes qui évoquent la rapidité de consommation des produits.

La hiérarchie des classes de réponses permet de montrer qu'il existe une dichotomie primordiale, selon que l'enquêté fait principalement référence au **réfrigérateur** ou à l'**emballage**. Cette dichotomie permet de mettre l'accent sur deux types de modes de réponses qui n'apparaissaient pas aussi clairement lors de la post codification.

Figure 2. Dendrogramme de la classification descendante



Guide de lecture : cet arbre, qui se lit de droite à gauche –puisque’il s’agit d’une classification descendante–, permet de représenter les trois partitions successives du corpus des réponses par l’algorithme de classification d’*Alceste* et la hiérarchie qui en découle.

Les tableaux qui suivent rendent compte du détail de la formation des classes construites par *Alceste*, ce qui donne la possibilité d’arriver à un niveau de détail plus fin dans la comparaison des deux méthodes.

Tableau 9. Description du premier contexte

« Référence à l'emballage »	195 réponses classées (30,4% du corpus) nombre moyen de formes par réponse : 2,48
Formes caractéristiques emballage+ (129) boîte+ (8) marqu+er (87) indicat+ion (8) écrit< (33) principe+ (11) indiqu+er (27) normalement (10) produit+ (36) regard+er (44) inscrire (9)	Interprétation la solution indiquée ici consiste à trouver l'information au sujet de la bonne température de conservation grâce à l'emballage des produits
Phrases caractéristiques « normalement c'est marqué sur l'emballage, autrement je ne sais pas » « des fois c'est écrit sur l'emballage » « c'est écrit sur l'emballage en principe » « c'est marqué en principe sur l'emballage » « c'est marqué sur les emballages quelques fois » « normalement c'est marqué sur les emballages » « c'est marqué sur le produit en principe » « ça doit être marqué sur l'emballage normalement » « en regardant sur l'emballage du produit » « c'est marqué sur l'emballage du produit » « c'est peut-être écrit sur l'emballage » « sur l'emballage c'est écrit dessus » « ne sait pas souvent c'est indiqué sur le produit » « car c'est marqué sur l'emballage » « en principe c'est marqué dessus » « ça doit être indiqué sur l'emballage » « écrit sur l'emballage » « l'inscrire sur l'emballage » « en l'inscrivant sur l'emballage »	

Tableau 10. Description du deuxième contexte

« Lire la notice ou l'étiquette »	102 réponses classées (15,9% du corpus) nombre moyen de formes par réponse : 2,71
Formes caractéristiques étiquette+ (56) limite+ (6) lire (28) regard+er (32) paquet+ (16) notice+ (12) date+ (15) marqu+er (32)	Interprétation le mode d'emploi de l'appareil de conservation est utilisé, en concurrence avec l'étiquette du produit en tout état de cause, il faut lire une information
Phrases caractéristiques « je lis les étiquettes et je regarde bien les dates de péremption » « lire la notice sur le paquet » « regarder la date limite » « on lit la notice sur l'étiquette » « je regarde uniquement la date limite, c'est tout » « je regarde la date, je suis et je goûte » « consulter un livre spécialisé ou alors lire les étiquettes » « je regarde la date sur le paquet je n'en sais pas plus » « on regarde sur la date limite » « déjà on regarde la date » « lire l'étiquette » « je le lis sur le paquet » « je lis sur l'étiquette » « je regarde sur les étiquettes » « je lis l'étiquette » « lire les étiquettes » « regarder sur l'étiquette » « lire la notice explicative » « lire la notice »	

Tableau 11. Description du troisième contexte

« Consommer vite »	58 réponses classées (9,0% du corpus) nombre moyen de formes par réponse : 3,69
Formes caractéristiques frais.....(17) volaille+.....(4) jour+.....(11) achet+er.....(9) mang+er.....(8) gard+er.....(4) conserve+.....(7) aller.....(5) possi+ble.....(6) poulet+.....(3) consomm+er.....(9) froid+.....(5) aucune-idée.....(6) colle+.....(2) voir.....(7) moyen+.....(2) rapidement.....(4) yaourt+.....(4)	Interprétation l'idée est ici de consommer au plus vite les aliments qui auront de toute façon été conservés au frais
Phrases caractéristiques « j'achète chez le boucher ma viande, c'est pour la consommer de suite » « en principe les laitages il faut les tenir au froid, le beurre aussi et la viande, mais on va tous les jours chercher du frais » « il faut le garder au plus froid possible » « ne sait pas, je ne le conserve pas, je le mange tous les jours » « un produit frais je le colle dans le frigidaire et je le mange le plus vite possible » « les yaourts au frigidaire et le poulet au congélateur car on le mange rapidement après » « je ne conserve pas de produit frais comme les volailles » « ne sait pas, voir s'ils sont frais » « 10 jours après l'abattage pour les volailles » « on l'achète à la banque réfrigérée et après on le mange assez vite » « si c'est frais ça se voit » « ne sait pas, aucune idée, j'achète au fur et à mesure » « quand il est frais je mets au congélateur le plus froid possible puis 3 jours dans le frigo » « ça j'en sais rien, vous me posez une colle, je le mets au frigo et c'est tout, mon boucher me dit vous n'allez pas au delà de tel jour » « je le mets tout de suite au réfrigérateur et je le consomme rapidement le jour ou le lendemain » « pour les yaourts température dessus 6 degrés, pour les volailles je les consomme en marmite immédiatement » « quand j'achète, j'achète du jour pour le lendemain donc ne conserve pas longtemps c'est marqué sur les étiquettes » « absolument aucune idée je le vois au produit lui même : si le produit cristallise du froid par exemple je baisse le thermostat » « c'est le frigidaire et je les consomme dans les 2 ou 3 jours »	

Tableau 12. Description du quatrième contexte

« Mettre au réfrigérateur »		287 réponses classées (44,7% du corpus) nombre moyen de formes par réponse : 2,89
Formes caractéristiques réfrigérateur+..... (233) degré+.....(34) mettre..... (156) haut+.....(10) thermomètre+..... (58) laiss+er.....(9) bonne+..... (15) thermostat.....(15) température+..... (50) aliment+.....(10) congélateur..... (20) régie+.....(8)		Interprétation réponses formulées par des individus pour lesquels le réfrigérateur semble être la voie nécessaire du salut face au problème qui leur est posé dans l'enquête, puisque c'est le réfrigérateur qui permettra de trouver une température idéale
Phrases caractéristiques « quand on est dans un supermarché on peut vérifier la température du thermostat et l'on met tout de suite dans le frigidaire » « dans le frigidaire y'a le haut et y'a le bas pour les différentes températures » « on les met dans le réfrigérateur sans contrôler la température, pas de thermomètre comme dans le congélateur » « y'a que le thermomètre qui puisse le dire, y'a qu'à le mettre plus bas et je mets les aliments frais dans le haut du frigo » « c'est entre 3 et 5 degrés, faut mettre un thermomètre dans le frigo » « mon frigidaire est réglé a la bonne température » « je me fie au magasin avec un thermomètre que j'ai dans le frigo » « il faut mettre un thermomètre et vérifier à 3-4 degrés » « eh bien il faut mettre la température enfin le thermostat du frigidaire au moyen degré entre les deux 5 ou 4 »		

La classe de « référence à l'emballage » semble très homogène, et correspond à ce titre à la définition de la catégorie équivalente qui existait déjà dans le plan de codage.

En revanche, la catégorie « emballage » du plan de codage se retrouve aussi dans la classe suivante construite par *Alceste*, « lire la notice ou l'étiquette ». En effet, on trouve dans cette classe des réponses qui se rattachent aussi à la catégorie « mode d'emploi du réfrigérateur » du plan de codage. Ceci semble s'expliquer principalement par le fait qu'on **lit** aussi bien une étiquette qu'une notice, alors qu'on **regarde** un emballage. De même, si l'information est **marquée** sur l'emballage, elle est surtout **écrite** sur l'étiquette. Il y a des différenciations dans l'utilisation d'expressions idiomatiques qui correspondent peut-être à une distinction sociale, au moins dans le discours, mais qui introduit une distinction artificielle entre deux comportements identiques. Parallèlement, on n'a plus de distinction du recours à la notice.

La classe « consommer vite » est plutôt hétérogène, puisque l'on y trouve différents styles de réponses comparables à ceux du plan de codage : recours aux techniques de production du froid (« *Il faut le garder au plus froid possible* »), notion de date de conservation (« *Un produit frais, je le colle au frigidaire et je le mange le plus vite possible* »), absence d'opinion ou d'intérêt pour la question (« *Ne sait pas, aucune idée. J'achète au fur et à mesure* »), voire un somptueux mélange de plusieurs attitudes relevées dans le plan de codage : « *Ça, je n'en sais rien, vous me posez une colle. Je le mets au frigo et c'est tout. Mon boucher me dit vous n'allez pas au delà de telle date* ».

Toutefois, cette classe peut aussi s'interpréter comme une généralisation de la catégorie du plan de codage « référence aux dates limites », et en tous cas relève d'un type d'attitudes qui n'apparaissent pas de manière aussi homogène dans la post codification.

La classe des individus qui utilisent principalement les formes « réfrigérateur » et « mettre » est peut-être moins homogène qu'il n'y paraît. Nous nous situons là à un point difficile de l'analyse de la question ouverte, qui touche autant le choix d'une méthode de classification des réponses que l'interprétation même du sens des classes mises en évidence. Il paraît clair que l'on fait ici de l'usage du réfrigérateur un préalable nécessaire à toute recherche de la température idéale. Pourtant, l'examen des réponses caractéristiques données par *Alceste* montre que les enquêtés vont aussi faire référence, en vrac, à une température précise et à

l'idée de mesure de la température dans le réfrigérateur¹ (« *C'est entre 3 et 5 degrés : faut mettre un thermomètre dans le frigo* »), à l'idée d'une confiance aveugle dans le réfrigérateur (« *Mon frigidaire est réglé à la bonne température* »), voire à une référence au thermomètre du présentoir du magasin (« *Quand on est dans un supermarché, on peut vérifier la température du thermostat et l'on met tout de suite dans le frigidaire* »).

En tout état de cause, cette classe regroupe les différents thèmes évoqués dans le plan de codage.

Les limites d'une description des attitudes...

Au total, cette analyse textuelle du corpus des réponses donne bien le même type de résultats que le plan de codage mais de manière nécessairement moins fine². Nous obtenons *stricto sensu* un résultat statistique, c'est-à-dire une information moins précise mais plus lisible.

Au chapitre des « portés disparus », insistons toutefois sur le fait que le logiciel *Alceste* n'a pas permis de distinguer dans cette première analyse les éléments de réponses suivants du plan de codage :

- les « *sans opinion* », que l'on ne peut mettre en bijection avec l'ensemble des « non classés »³.
- les « *expériences acquises* », noyées sans que l'on puisse les retrouver par l'utilisation d'un vocabulaire spécifique.
- les « *place des produits dans le réfrigérateur* », dont quelques spécimens sont apparus caractéristiques d'opinions plus générales sur l'utilisation du réfrigérateur.

¹ Cette idée de mettre un thermomètre dans le réfrigérateur apparaissait déjà lors de la phase de post-codage, et avait été rattachée, peut-être de manière imprécise, à la notion de "température idéale". Pourtant, le fait que l'on déclare utile de rechercher quelle est la température qui règne dans le réfrigérateur n'implique pas que l'on connaisse forcément le chiffre, ou tout au moins un intervalle, qui permettra de valider la mesure. En particulier, on ne sait si la validation viendra d'une recherche d'information —emballage, notice du réfrigérateur...—, ou bien d'une pré-connaissance ou d'une expérience acquise...

² Notons d'ailleurs qu'une limite est introduite par *Alceste* puisque le logiciel arrête son analyse dès lors qu'une classe comprend moins de 2% des u.c.e.

³ On regrettera au passage que la version du logiciel que nous avons utilisée ne donne aucune information sur les unités de contexte "non classées", alors qu'elles font ici partie intégrante de l'analyse des réponses à la question.

- les « *erreurs* », diluées dans les autres catégories de réponses.

Il convient d'ajouter à cette liste de disparus l'absence de distinction claire entre certains comportements différents : c'est ainsi le cas que nous avons relevé des catégories « *lecture de la notice de l'appareil* » ou « *emballages des produits* », qui n'ont pas été différenciées alors même qu'elles ne participent pas de la même approche de la question —elles devraient même se situer, en toute logique, de part et d'autre de la première dichotomie induite par l'algorithme de partitionnement, et il s'agit là d'une limite du logiciel délicate à appréhender.

Nous ne pouvons être clairement convaincus en l'état de notre analyse qu'il y a impossibilité d'obtenir un classement fin des individus. Cependant, nous pouvons établir plus sûrement que l'automatisation espérée par *Alceste* n'est guère possible sans intervenir directement sur l'interprétation des mots utilisés par les enquêtés, ce qui suppose en fait une pré-connaissance des réponses données par les enquêtés, avant d'entreprendre l'analyse, et cela à un niveau tel que l'analyse textuelle deviendrait alors peu utile.

Cette première analyse permet aussi et principalement de montrer l'existence de trois grands types de réponses à la question que le plan de codification ne mettait pas aussi bien en valeur : l'absence de réponse (les non classés), la référence externe (l'emballage), l'appel du réfrigérateur (réglage, position, ...).

Croisement de la classification d'*Alceste* et de la post codification

Contexte d'étude

Après la phase de construction d'une partition des types de réponses à la question ouverte, nous allons maintenant comparer plus finement les résultats obtenus par *Alceste* avec l'application du plan de post codage.

Cette comparaison s'effectue sur l'ensemble des réponses classées ou reclassées par *Alceste*¹. Le poids des différentes classes est donné dans le tableau suivant.

Tableau 13. Ventilation des classes de réponses construites par *Alceste*

Classe de réponses	Part
réponses non classées	38,7%
« référence à l'emballage »	17,6%
« lire la notice ou l'étiquette »	11,9%
« consommer vite »	12,4%
« mettre au réfrigérateur »	19,5%
	100,0%

¹Le Tableau 8, page 23, donne la ventilation des réponses selon la classe donnée par l'algorithme de classification descendante d'*Alceste*. Toutefois, on ne prenait là que les réponses effectivement prises en compte dans le processus de partitionnement. Les réponses "non classées" lors de cette première étape sont réexaminées par un algorithme s'apparentant à l'attribution d'un score et donc reclassées dans l'une ou l'autre des classes construites précédemment. Certaines restent toutefois non classées. Nous étudions désormais l'ensemble des réponses classées ou reclassées, et les marges données dans la partie précédente doivent donc être corrigées.

Cette distribution diffère sensiblement de la distribution donnée précédemment, en particulier du fait du gonflement de la classe « *consommer vite* » qui passe de 5,3% des réponses à 12,4% de l'ensemble du corpus. Nous verrons par la suite que cet écart important s'explique surtout par le reclassement douteux de réponses de la catégorie « sans opinion » du plan de codage.

Étude des correspondances entre les classes

L'examen des caractéristiques contextuelles des classes construites par *Alceste* montrait que l'on retrouvait certaines grandes directions du plan de codage appliqué précédemment. On ne s'étonnera donc pas de voir certaines classes de réponses se recouper fort logiquement avec les catégories du plan appliqué lors de la post codification.

Le tableau suivant permet de voir dans quelles classes d'*Alceste* se retrouvent les catégories du plan de codage.

Tableau 14. Profil des catégories du plan de codage en fonction des classes construites par Alceste

classes construites par <i>Alceste</i> catégories du plan de codage	Non classées	Regarder l'embal- lage	Lire la notice ou l'étiquette	Consom- mer vite	Mettre au réfrigéra- teur	Total
Sans opinion	72,4	2,3	4,5	14,3	6,5	100,0
Emballage du produit	2,6	63,1	27,4	5,2	1,6	100,0
Température de présentation en magasin	5,1	3,5	4,7	3,6	83,1	100,0
Confiance absolue dans le réfrigérateur	8,8	3,3	3,0	14,8	70,1	100,0
Expérience acquise	39,1	8,7	3,8	14,1	34,3	100,0
Température idéale	8,2	7,0	2,4	10,0	72,5	100,0
Place des produits	9,6	4,3	0,0	39,6	46,5	100,0
Mode d'emploi du réfrigérateur,... spécialiste	34,3	12,7	36,9	11,0	5,1	100,0
Dates limites	0,0	8,6	76,3	15,1	0,0	100,0
Aspect du produit	20,0	7,1	20,2	52,8	0,0	100,0

Guide de lecture : parmi les réponses des enquêtés affectées à la catégorie « sans opinion » du plan de post codification, 72,4% ont été « non classées » par *Alceste*, 14,3% ont été classées dans la classe « consommer vite ».

Une grande partie des différences de sens entre les deux classements s'explique par la présence d'une réponse multiple lors de la post codification, *Alceste* attribuant alors une seule des catégories possibles. C'est en particulier le cas des réponses de la catégorie « sans opinion », qui sont classées souvent par affinité avec l'autre réponse la plus proche.

Les tableaux des pages qui suivent donnent plus de précisions sur le classement par *Alceste* des réponses en fonction des catégories construites lors de la post codification. Nous rappelons en premier lieu la classe dans laquelle la majorité des réponses sont affectées, puis donnons des réponses permettant d'illustrer les différences de sens observées.

Tableau 15. Classement par Alceste de la catégorie « Sans opinion »

Classement principal : Non classées (72,4%)	
- ce classement principal était le plus logique compte tenu des caractéristiques ultra courtes des réponses données par les enquêtés de cette catégorie - les exemples qui suivent montrent qu'il y des classements différents suite à une ambivalence de la réponse, ou encore à cause d'une classe de réponse mal définie - la classe « consommer vite » semble avoir attiré certaines réponses lors de l'étape de classement sans raison apparente	
Classe	Illustration
regarder l'emballage	« Ne sait pas. Souvent, c'est indiqué sur le produit » « Je ne sais pas mais ça doit être marqué dessus » « ça doit être marqué sur le produit mais je n'ai jamais fait attention »
lire la notice ou l'étiquette	« Je ne sais pas. C'est marqué dessus : on regarde » « Je ne sais pas. C'est marqué sur les étiquettes » « Je me suis jamais posé la question »
consommer vite	« Aucune idée » « Je ne vois pas trop ce que vous voulez qu'on réponde, là » « En principe je mets à congeler et je sors au jour le jour » « Ne sait pas. Je consomme tout de suite » « Je n'ai jamais regardé mais je consomme au fur et à mesure » « On en conserve pas trop » « Pour les yaourts, je pense que c'est marqué dessus sinon j'en sais rien » « Je ne sais pas : ça se voit, non ? » « Bon j'en sais rien, moi, je fais confiance à celui chez qui j'achète »
mettre au réfrigérateur	« Je ne sais pas, je mets au frigo » « Je ne sais pas, je le laisse toujours pareil » « Je ne sais pas, on les met au frigo comme ça et on ne regarde pas là » « A-t-on la possibilité de vérifier ? Mise au congélateur » « Aucune idée. On le met au frigo et c'est terminé » « Je mets dans le réfrigérateur sans réfléchir »

Tableau 16. Classement par Alceste de la catégorie « Emballage du produit »

Classements principaux : « Regarder l'emballage » (63,1%) « Lire la notice ou l'étiquette » (27,4%)	
- le classement se fait principalement dans deux classes d'<i>Alceste</i>, à cause de la séparation de sens déjà notée entre ces deux contextes lexicaux (on regarde un emballage, on lit une étiquette) - le classement « non classées » reste douteux, et semble provoqué par la présence de mots isolés, ou encore employés dans un contexte que l'algorithme d'<i>Alceste</i> ne comprend plus : « c'est marqué sur les étiquettes » - les autres classements font apparaître des erreurs, dont en particulier celles qui touchent la classe « consommer vite »	
Classe	Illustration
non classées	« Au pif, quoi, j'ai mis au 2 le frigo, cela a l'air d'être froid, ou alors quand c'est marqué sur l'étiquette » « Mettre au frigidaire : c'est marqué sur l'étiquette » « La température de conservation doit être notée sur l'emballage » « Cela est marqué dessus, j'y fais attention »
lire la notice ou l'étiquette	« Je sais pas, c'est marqué dessus, on regarde » « Sur une étiquette » « Imprimer la température de conservation sur une étiquette » « En les mettant au frigidaire en faisant attention à la date indiquée sur les paquets » « Il y a une étiquette, non ? » « En lisant l'emballage et le livre du frigo » « Parfois c'est indiqué sur les paquets » « Je lis sur le produit »
consommer vite	« C'est marqué tenir au frais ou à l'abri de la chaleur » « Sur l'emballage mais ceci dit je ne sais pas ce qu'il fait dans mon frigo » « C'est marqué dessus à conserver entre tel et tel degré » « Aucune idée. Il faut s'acheter un frigo qui est aux normes homologuées et déjà ça va mieux » « C'est marqué sur la viande, peut-être pas sur les yaourts »
mettre au réfrigérateur	« Y'a la température dessus » « En mettant déjà un thermomètre dans le frigo et en regardant l'étiquette » « mettre le frigo à 6 degrés et c'est marqué sur l'emballage » « on lit sur l'emballage la température et j'ai un thermomètre dans mon frigo »

Tableau 17. Classement par Alceste de la catégorie « Température de présentation en magasin »

Classement principal : « Mettre au réfrigérateur » (83,1%)	
• on voit que la classe « mettre au réfrigérateur » est assez floue dans sa définition : reproduire la température du rayon frais implique certes l'utilisation d'un réfrigérateur, mais <i>Alceste</i> ne distingue pas ce comportement réellement • les classements douteux relèvent d'une imprécision de l'algorithme de classification ou bien d'une ambivalence qui n'est pas rendue	
Classe	Illustration
non classées	« Par rapport à la conservation en magasin » « C'est marqué sur le truc, il faut regarder le thermomètre dans les rayons »
regarder l'emballage	« Normalement, il y a un thermomètre dans le lieu de présentation » « En regardant en vitrine déjà la température qu'elle est » « C'est marqué sur l'emballage et il faut normalement vérifier la température du frigo »
lire la notice ou l'étiquette	« Je regarde où ils se trouvent en supermarché » « Regarder le thermomètre et il y a la date sur l'emballage » « On regarde le truc qu'ils ont... un thermomètre »
consommer vite	« Regarder s'il y a un thermomètre dans les étalages et voir la température » « Il faudrait un ruban dans le thermomètre frontal. Ne sait pas. ça presse de manger plus vite » « Quand on les achète ils sont déjà au rayon frais donc je les remets au frais »

Tableau 18. Classement par Alceste de la catégorie « Confiance absolue dans le réfrigérateur »

Classement principal : « Mettre au réfrigérateur » (70,1%)	
- le classement principal est ici naturel, et il est concurrencé directement par la classe « <i>consommer vite</i> » qui, on l'a vu, implique elle-même une référence au moins implicite au réfrigérateur - à côté des ambivalences, les autres classements correspondent en fait à l'utilisation de termes peu fréquents dans le corpus des réponses	
Classe	Illustration
non classées	« Faire confiance au réfrigérateur » « Je fais confiance à mon réfrigérateur » « Au frigo, c'est l'habitude » « Regarder sur l'emballage. Je ne regarde pas régulièrement, je me fie au réfrigérateur »
regarder l'emballage	« ça doit être écrit normalement et je le mets au frigidaire » « C'est marqué dans le frigo » « En regardant le frigo : c'est marqué, il y a un barème » « Les températures sont standards, les températures du frigo sont adéquates »
lire la notice ou l'étiquette	« En les mettant au frigidaire en faisant attention à la date limite indiquée sur l'emballage » « Je regarde la date, j'enlève le plastique et je mets au frigo » « Il y a des normes, il suffit de regarder le mode d'emploi au frigo » « On regarde dans le frigo s'il y a pas trop d'humidité »
consommer vite	« Faire le plus froid possible au frigo » « On sait bien que tout corps mort à l'air libre va se décomposer alors il faut mettre les aliments dans un chambre froide pour les conserver »

Tableau 19. Classement par Alceste de la catégorie « Expérience acquise »

Classement principal : « Non classées » (39,1%) « Mettre au réfrigérateur » (34,3%)	
- Alceste n'ayant pas créé de classe qui évoque spécifiquement la notion d'habitude, on ne s'étonnera pas des classements principaux qui montrent la limite atteinte ici par l'algorithme - cette limite semble s'expliquer par un vocabulaire varié et dont les fréquences d'utilisation restent faibles ou très particulières : les termes « <i>habitude</i> » ou « <i>expérience</i> » ne sont pas rattachés à un contexte précis par Alceste - les autres classements évoquent plutôt la présence d'ambivalences que de véritables erreurs	
Classe	Illustration
non classées	« C'est l'habitude, on a un frigidaire » « C'est un peu à l'oeil d'habitude » « On le sait par expérience » « Au frigo, c'est l'habitude » « Je les connais depuis longtemps » « Je le fais par habitude. Ne sait pas »
regarder l'emballage	« C'est mis sur le produit » « L'indiquer sur l'emballage. Moi, je sais à peu près » « Le bon sens et l'habitude, les indications de l'emballage » « Quand les produits s'abîment vite, on le remarque soi-même : on se trompe une fois, pas deux »
lire la notice ou l'étiquette	« Question d'habitude. Les professionnels l'indiquent souvent sur les brochures » « Question de feeling »
consommer vite	« ça se voit, on est habitué » « On l'achète à la banque réfrigérée et après on le mange assez vite » « On a toujours la même température, on n'en change pas »
mettre au réfrigérateur	« Par habitude je mets dans le réfrigérateur » « Je ne sais pas, je laisse toujours pareil » « Au début on essaie puis lorsqu'on trouve la bonne position on la laisse » « Je mets le thermostat à 2, c'est bon pour tout »

Tableau 20. Classement par Alceste de la catégorie « Température idéale »

Classement principal : « Mettre au réfrigérateur » (72,5%)	
- le classement principal s'explique, en l'absence d'une classe identique à la catégorie, par la référence quasi obligée au réfrigérateur - les autres classements, pour ceux qui ne correspondent pas à une ambivalence, s'expliquent par le caractère non explicite de la référence au réfrigérateur	
Classe	Illustration
non classées	« Conserver le produit entre 0 et 5 degrés » « Je la connais : entre 5 et 8 degrés, on le sait »
regarder l'emballage	« Il y a une graduation sur le produit » « 6 à 7 degrés : c'est marqué dessus » « On a l'habitude, c'est 4 à 5 degrés, c'est marqué dessus » « C'est marqué sur les produits : à 3, ça suffit »
lire la notice ou l'étiquette	« Je ne sais pas, on sait que c'est entre 0 et 6 degrés, c'est écrit sur les étiquettes » « C'est à 2-3 degrés à peu près. J'avoue que me suis pas posé la question »
consommer vite	« Le frais, c'est 4 à 8 degrés, on le sait depuis toujours » « Entre 0 et 3 degrés pour la volaille, le yaourt 6 » « On règle pas trop fort, moyen » « Il faut peut-être 3 ou 4 degrés en bas, et moins 18 en haut, faut voir »

Tableau 21. Classement par Alceste de la catégorie « Place des produits »

Classement principal : « Mettre au réfrigérateur » (46,5%)	
- comme précédemment, c'est ici la référence sous-jacente à l'utilisation du réfrigérateur qui explique pour une bonne part le classement principal - les autres classements montrent en particulier que l'algorithme d'<i>Alceste</i> n'a pas permis d'intégrer un vocabulaire hétérogène, à base de prototypes alimentaires en partie suggérés par le libellé de la question, et où le référentiel « haut »-« bas » n'est pas réellement distingué alors même qu'il est l'élément clef de la discrimination	
Classe	Illustration
non classées	« Charcuterie en bas » « En plaçant les produits à différents endroits du frigo : bac à légumes, etc... » « Fromages et fruits-légumes en bas » « Suivant les étages, le refroidisseur est moindre ou plus élevé »
regarder l'emballage	« En regardant sur l'emballage, il y a des produits que je mets en haut, d'autre en bas selon la température » « C'est marqué sur la porte du frigo par rapport au niveau du réfrigérateur »
consommer vite	« En bas du frigo, les produits frais se gardent bien » « Il faut savoir se servir d'un réfrigérateur : pas mettre la salade en haut et les yaourts en bas » « Suivant l'étage de mon frigo, c'est fait pour » « On le sait, la viande c'est en bas du frigo, le yaourt c'est en haut »

Tableau 22. Classement par Alceste de la catégorie « Mode d'emploi du réfrigérateur, recours à un spécialiste »

Classement principal : « Lire la notice ou l'étiquette » (36,9%)	
- le classement principal est ici en quelque sorte « naturel » puisqu'il correspond à une convergence déjà relevée entre les deux types de traitements - ce classement ne recouvre toutefois que les références précises à la notice, et non le recours au spécialiste - ceci explique en partie les autres classements, qui relèvent le plus souvent d'une incapacité de l'algorithme d'<i>Alceste</i> à identifier le vocabulaire spécifique autour de la notion élargie de « spécialiste »	
Classe	Illustration
non classées	« On demande au spécialiste » « Je fais confiance aux commerçants » « En se fiant à la notice du frigo, et l'habitude » « Je demande au vendeur si je ne sais pas » « Je fais confiance au technicien qui m'a installé le réfrigérateur et je n'y touche plus »
regarder l'emballage	« J'ai un manuel, tout est écrit dessus » « C'est noté sur le produit ou sur le réfrigérateur » « C'est indiqué sur la notice du réfrigérateur ou sur le produit lui-même »
consommer vite	« Acheter un livre là-dessus » « Sur une notice allant avec le frigidaire » « Déjà il faut se renseigner de la personne auprès de qui on l'a acheté » « ça je n'en sais rien, vous me posez une colle, je le mets au frigo et c'est tout »
mettre au réfrigérateur	« Je les mets au milieu du frigo et c'est bon pour conserver car il y a des dessins qui l'indiquent » « C'est bien marqué quelque part, il suffit d'avoir un thermomètre dans le réfrigérateur pour les mettre à la hauteur adéquate »

Tableau 23. Classement par Alceste de la catégorie « Dates limites »

Classement principal : « Lire la notice ou l'étiquette » (76,3%)	
• on se trouve ici en présence d'une catégorie de réponses très typées au niveau du plan de codage de la question ouverte : toutes les réponses font ici explicitement référence à la notion de « date » • le classement principal s'explique par une cooccurrence fréquente des termes « lire » ou « regarder » et « date », qui attirent en quelque sorte les réponses concernées vers une classe où l'on déclare déjà lire ou regarder quelque chose • les deux autres classements s'expliquent principalement par une ambivalence dans la réponse : recours à l'emballage, ou notion de « frais »	
Classe	Illustration
regarder l'emballage	« Il faut regarder les dates et regarder le degré sur l'emballage, normalement cela doit être indiqué » « La date sur l'emballage, et on sent le produit »
consommer vite	« Aucune idée. Je regarde la date, je la mange » « Si l'on s'en sert à la date avant, sinon ce n'est pas frais » « Je regarde la date, je mange peu de viande, je fais attention » « A l'aspect pour les volailles, à la date limite pour les yaourts »

Tableau 24. Classement par Alceste de la catégorie « Aspect du produit »

Classement principal : « Consommer vite » (52,8%)	
- le classement s'explique par l'utilisation de termes dont le contexte s'apparente à l'idée de vitesse de consommation, et qui sont déjà utilisés hors d'une référence à l'aspect du produit dans le reste du corpus des réponses - ce rattachement à la classe « <i>consommer vite</i> » s'explique aussi par la notion de fraîcheur, comme le montre une réponse du type : « <i>si c'est frais, ça se voit</i> » - le classement « <i>non classées</i> » s'explique par l'utilisation de termes rares dans le corpus, et qui ne peuvent être rattachés à un contexte connu par l'algorithme	
Classe	Illustration
non classées	« A l'odeur et à l'aspect » « A l'odeur » « Si il garde un bon aspect, si il ne sèche pas » « C'est un peu à l'oeil habitué » « Je sens, c'est au nez » « Ne sait pas. L'aspect ? »
regarder l'emballage	« C'est peut-être marqué dessus. Sinon c'est au feeling et à la tête du produit » « La date sur l'emballage et on sent le produit »
lire la notice ou l'étiquette	« Par la couleur » « Je ne sais pas quoi vous répondre : je regarde la couleur, à l'odeur, au toucher » « Le goûter » « Je regarde la date, je la suis, et je goûte »

Conclusion

L'analyse fine des croisements entre la post codification des réponses et leur classification par *Alceste* permet de confirmer certaines intuitions que l'on pouvait avoir au sujet des principales différences entre les deux traitements des questions ouvertes. On a ainsi la confirmation qu'au delà de la réelle difficulté d'*Alceste* à parvenir à un niveau de détail aussi fin que le permet une post codification, le niveau encore très synthétique auquel on est soumis par l'algorithme du logiciel n'assure pas pour autant l'homogénéité des classes de réponses en termes d'attitudes et de comportements. De grandes lignes sont mises en évidence et permettent donc de comprendre les axes d'orientation des discours prononcés par les enquêtés, mais dans le détail nous observons des divergences souvent injustifiées qui ne permettent pas de considérer au même plan la méthode de classification descendante et la post codification pour la description des comportements.

Si l'on veut bien détailler les principales défaillances des algorithmes utilisés par *Alceste* pour effectuer une partition du corpus, nous pouvons relever de manière synthétique les points suivants :

- la présence d'idées multiples et d'ambivalences dans le discours ne conduit pas sur notre corpus à la constitution des classes correspondantes de manière homogène : par exemple, la notion de recours à l'emballage en complément d'une absence d'opinion se trouve à la fois parmi les réponses « regarder l'emballage » et les réponses « non classées » (cf. Tableau 16 et Tableau 17) ;

- le traitement des « non réponses » qui font intervenir principalement des mots outils reste insatisfaisant, sans qu'il soit vraiment possible d'envisager un remède au niveau du pré-traitement du corpus ou au niveau des algorithmes d'*Alceste* ;
- la classe résiduelle des réponses « non classées » par *Alceste* ne retranscrit d'ailleurs pas pleinement l'idée d'absence d'opinion : si on y retrouve bien les trois quarts des « sans opinion » du plan de codage, on y trouve aussi plus du tiers des « expérience acquise » ou encore des « mode d'emploi du réfrigérateur » ;
- des classes « fourre-tout » apparaissent pour décrire notre corpus qui correspondent à des regroupements de plusieurs catégories du plan de codage, catégories qui ne sont pas forcément associées à des comportements comparables : c'est par exemple le cas de la classe « mettre au réfrigérateur » qui regroupe à la fois des comportements du type « caution scientifique » et d'autres du type « certitudes absolues » ;
- la classe « consommer vite » qui semblait plutôt claire lors de la phase de construction est beaucoup moins homogène après l'étape de reclassement et le croisement systématique avec les catégories de la post codification ;
- des comportements très spécifiques mais partagés par un faible nombre de personnes interrogées (« place des produits », « aspect du produit ») se diluent dans les classes d'*Alceste*, alors qu'elles appellent pourtant l'utilisation d'un vocabulaire très précis, se rattachant à une classe par présence d'une autre idée, éventuellement sous-entendue ou par un phénomène plus délicat à interpréter ;
- il n'est pas inutile d'insister enfin sur le fait que l'on n'a pas trouvé grâce à l'algorithme de classification d'*Alceste* les grands axes de la post codification, et en particulier l'absence d'opinion, la présence d'opinion et l'opinion erronée.

Les résultats des mises en œuvre des deux techniques sont différents et il n'y a pas lieu de s'en étonner ici ou ailleurs. Si cela n'avait pas été le cas, ce n'est certainement pas dans cette recherche que l'on aurait mis au jour ce résultat.

Les différences tiennent en fait aux fondements mêmes des deux démarches considérées : tandis que la post codification a pour but d'affecter une classe prédéfinie aux individus en fonction des réponses qu'ils ont données, l'algorithme de classification descendante du logiciel *Alceste* traite, lui, le lexique, en recherchant les convergences entre les différents contextes des mots utilisés. La comparaison entre post codification et classification descendante hiérarchique n'était par conséquent justifiée que dans la mesure où l'on acceptait de faire sien le postulat selon lequel les individus emploient un lexique dans un contexte propre à être comparé d'un individu à l'autre.

Pour parler comme un statisticien, nous dirons que la post codification effectue une typologie des individus, et la classification descendante les typologies conjointes du lexique et des individus. Autrement dit, la construction des variables de distinction des individus n'est pas prise en charge dans le processus de post codification, alors qu'elle est partie intégrante de l'algorithme de classification d'*Alceste*. Certaines limites interprétatives d'*Alceste* tiennent d'ailleurs aux ambiguïtés dans les définitions de la population étudiée et des variables.

Ainsi, la reconstitution des réponses multiples reste problématique avec *Alceste*. Lors de la post codification des réponses, on s'était ainsi aperçu du fait que certaines modalités de réponses constituaient une sorte de recours à une absence d'opinion formulée *a priori* par les ménagères interrogées. Ce type d'approche est en revanche plus délicat avec une classification qui reste hiérarchique : si l'on peut bien interpréter la hiérarchie elle-même comme une indication de la proximité entre plusieurs types de réponses, cette interprétation reste incertaine et ne permet pas la reconstitution de réponses multiples. On voit ici le rôle fondamental de la définition de l'unité de contexte élémentaire choisie par *Alceste*. Le découpage d'une réponse d'un individu en plusieurs unités de contexte reste critique.

Les deux méthodes considérées dans cette partie apparaissent donc complémentaires car elles ne donnent pas la même information. Dans cette approche très détaillée qui a été la nôtre, il est clair que de nombreuses limites sont mises en évidence et que la tâche d'automatisation de la codification des réponses à une question ouverte décrivant un comportement n'est pas entièrement maîtrisée.

L'analyse textuelle et les approches par graphes

Partant du principe aujourd'hui acquis qu'il existe un nombre croissant d'outils d'analyse textuelle qui s'appuient sur la théorie des graphes, nous en avons testé un dans le but d'en illustrer les capacités.

Le logiciel *Tropes*¹ a été testé à ce titre.

Trois applications sont proposées sur trois corpus de natures différentes, typiques de ceux auxquels les chargés d'études du CRÉDOC sont habituellement confrontés : la même question ouverte que celle qui a été utilisée pour comparer *Alceste* et une post codification ; les réponses à une même question ouverte posée, « *Pour vous, qu'est-ce qu'être heureux ?* », dans une enquête à quatre ans d'intervalle ; la retranscription d'entretiens semi-directifs sur la prospective de la mobilité locale des personnes âgées.

Les résultats établis sont ensuite complétés par une réflexion théorique sur les possibilités offertes par la théorie des graphes dans l'analyse de corpus. En effet, *Tropes* est un logiciel qui permet de produire une description « intelligente » des notions évoquées dans un corpus de texte(s). Toutefois, même si cette description s'appuie sur des tests statistiques, nous ne pouvons obtenir qu'une description, et non un modèle qui pourrait, par exemple, faciliter la post codification d'un corpus de réponses ouvertes. Dans cette perspective, c'est en particulier la voie représentée par les analyses par réseaux de neurones qui est explorée.

¹ Développé par Pierre Molette et commercialisé par la société Acétic.

Présentation des possibilités du logiciel *Tropes*

Définitions et concepts utilisés

Le logiciel *Tropes* a été développé par une équipe de chercheurs sémiologues et informaticiens dans la lignée de travaux sur l'analyse cognitivo-discursive¹ du Groupe de Recherche sur la Parole (Université Paris VIII).

Par le jeu d'une chaîne de traitement du corpus que nous détaillons par la suite, *Tropes* a pour objectif de construire un réseau de liaisons entre les concepts utilisés dans ce corpus, sous la forme d'un graphe permettant une représentation et une interprétation du sens du texte. Le corpus peut être composé d'un ou plusieurs textes.

Le logiciel utilise la notion de « classes d'équivalents » et analyse les cooccurrences entre ces classes, et non les cooccurrences entre les mots du corpus. A ce titre, l'étape d'identification des classes d'équivalence est l'analogue de la lemmatisation que produit le logiciel *Alceste*, c'est-à-dire qu'il s'agit de réduire intelligemment la quantité d'information disponible dans le texte de manière à privilégier la significativité des résultats statistiques. D'une manière générale, ce type de démarche correspond à l'arbitrage traditionnel qui doit être fait dans toute analyse statistique descriptive entre quantité d'information traitée et interprétabilité des résultats.

¹ Ghiglione et alii, 1995

Les concepts utilisés à cet effet par *Tropes* sont illustrés par l'exemple suivant, que nous extrayons du manuel de présentation du logiciel.

Tableau 25. Exemple de catégories et références utilisées par Tropes

Univers 1	Univers 2	Classes	Mots
politique	doctrine politique	communisme	communisme
politique	doctrine politique	communisme	marxisme
politique	doctrine politique	libéralisme	capitalisme
politique	doctrine politique	libéralisme	libéralisme
politique	homme politique	chef d'état	chef d'état
politique	homme politique	chef d'état	président de la République
politique	homme politique	ministre	ministre
politique	homme politique	ministre	garde des Sceaux

L'ensemble de ces concepts est utilisé par le logiciel et l'utilisateur final a la possibilité de contrôler les catégories ou références à partir desquelles il souhaite travailler. Il faut noter toutefois que ces emboîtements sont fixés par le logiciel, ce qui, dans le cas d'une synonymie présente dans un même texte, peut mener à des confusions incontrôlables.

Le logiciel *Tropes* utilise aussi deux concepts qui permettent de décrire globalement un texte et qui sont le **style général** et la **mise en scène**. Le logiciel donne pour chaque texte analysé la qualification selon ces deux critères (les résultats possibles sont définis par les concepteurs du logiciel comme il est indiqué dans le tableau suivant), et il fournit aussi quelques éléments complémentaires.

Tableau 26. *Styles et mises en scène selon Tropes*

Styles	
argumentatif.....	le sujet s'engage, argumente, explique ou critique pour essayer de persuader l'interlocuteur
narratif.....	un narrateur expose une succession d'événements, qui se déroulent à un moment donné, en un certain lieu
énonciatif.....	le locuteur et l'interlocuteur établissent un rapport d'influence, révèlent leurs points de vue
descriptif.....	un narrateur décrit, identifie ou classe quelque chose ou quelqu'un

Mises en scène	
dynamique, action.....	verbes d'action prédominants
ancrée dans le réel.....	verbes de l'être et de l'avoir
prise en charge par le narrateur.....	verbes qui permettent de réaliser une déclaration sur un état, une action...
prise en charge à l'aide du « je »...	nombreux pronoms à la première personne du singulier

Ces qualifications globales des corpus qui sont soumis à l'analyse du logiciel ont pour fondement la comparaison entre la structure du vocabulaire d'un corpus et la structure d'une base de référence interne au logiciel (*cf. infra*).

Principales étapes de l'analyse d'un corpus

Les analyses que produit *Tropes* reposent sur quatre phases : découpage des phrases du corpus en propositions, levée d'ambiguïtés des mots du corpus, tests statistiques, identification de classes d'équivalents et des relations entre ces classes. Avant d'être regroupés en classes, le logiciel regroupe les mots en grandes catégories : les verbes, les joncteurs (qui sont les conjonctions et locutions conjonctives de coordination et de subordination), les modalisations (adverbes et locutions adverbiales), les adjectifs qualificatifs, les pronoms personnels et les « classes d'équivalents » (noms communs et noms propres). Ce sont ces caractéristiques qui sont prises en compte pour qualifier globalement un texte.

Le **découpage du corpus en propositions** a pour but de transformer le texte à étudier en unités simples qui correspondent à une seule idée. Cette étape est l'analogue de ce que produit

Alceste lors du découpage du corpus en « unités de contexte élémentaire » (u.c.e.). Toutefois, alors qu'*Alceste* fait ce découpage en cherchant à produire des u.c.e. de tailles similaires (même si ce processus peut être contrôlé, ce qui sera par exemple utile dans le traitement de questions ouvertes où l'u.c.e. est, naturellement, la réponse donnée par un individu), *Tropes* fonde le découpage en propositions sur à la fois un examen de la ponctuation et une analyse de la syntaxe du texte. Par conséquent, même si les auteurs du logiciel annoncent que ce découpage comportera nécessairement des erreurs, il est un fait qu'une même idée aura peu de chances d'être séparée selon plusieurs unités de texte.

La **levée d'ambiguïtés** des termes rencontrés dans le texte a pour but de s'assurer que chaque mot est bien interprété conformément à sa fonction grammaticale et à son sens. Cette étape est nécessaire à l'identification des catégories des mots du corpus. A cette fin, *Tropes* a la possibilité de résoudre deux types d'ambiguïtés classiques en pareil cas, qui sont liées à la structure grammaticale et syntaxique, et aux aspects sémantiques du texte analysé. La résolution grammaticale permettra par exemple d'identifier une homonymie entre un verbe et un nom commun (par exemple, le mot « livre » qui peut désigner à la fois un objet et être une forme conjuguée du verbe « livrer », et qui ne remplissent donc pas la même fonction dans le texte). De la même manière, une analyse sémantique permettra de distinguer deux termes dans une homonymie pour deux mots qui ont la même nature et donc la même fonction dans le texte (par exemple, le mot « fils »). Ces levées d'ambiguïtés sont réalisées par *Tropes* par des algorithmes d'intelligence artificielle, avec un taux d'erreur réputé faible. En tout état de cause, notre utilisation de *Tropes*, telle qu'elle est illustrée par les exemples qui suivent, nous a permis d'identifier quelques-unes de ces erreurs, comme l'association du terme « scénario » à un contexte uniquement cinématographique, alors que c'est un terme couramment utilisé dans le vocabulaire de la prospective.

Les deux étapes qui précèdent sont réalisées conjointement avec l'identification des **catégories de mots**. Certaines catégories comme les pronoms, les articles, les prépositions ou encore certains adjectifs permettent d'identifier par leur seule présence certaines ambiguïtés du texte. Par exemple, dans le cas d'un mot qui peut s'interpréter à la fois comme un verbe ou un nom commun, la présence d'un pronom avant ce mot permettra de lever l'ambiguïté.

Chaque mot sera de plus identifié selon qu'il est **actant** ou **acté**. Ces notions seront représentées à l'aide de graphes orientés, les flèches du graphe allant des actants vers les actés.

Les **tests statistiques** qu'effectue *Tropes* sont fondés sur l'analyse des fréquences d'apparition des mots et des catégories de mots, ainsi que sur l'analyse des cooccurrences et des taux de liaisons entre les classes d'équivalents. Ces tests sont en principe fondés sur la mesure du chi-2. Ils ont pour but de permettre la construction d'un graphe représentant les liaisons entre classes d'équivalents du corpus, c'est-à-dire la structure des idées qui sont présentes dans le texte. Ils sont aussi utilisés pour identifier les catégories de mots sur-représentées dans le corpus en référence à une utilisation moyenne de ces catégories dans la langue française (le secret de fabrication de la dite référence étant bien entendu précieusement préservé), et donc aussi pour qualifier le style du texte.

Mode d'entrée dans le logiciel

Contrairement aux versions d'*Alceste* que le CRÉDOC utilise traditionnellement, le logiciel *Tropes* possède un mode d'entrée par interface interactive et graphique, sur le modèle des logiciels aisément pilotables par un utilisateur non chevronné en informatique. Cette caractéristique en fait un outil particulièrement adapté à toute forme d'analyse purement exploratoire d'un corpus inconnu. En indiquant au logiciel un certain nombre de textes qui formeront une « base documentaire », l'utilisateur a ensuite la possibilité de naviguer en interactif dans le corpus ainsi constitué et d'en visualiser la structure des idées. Pour des raisons de recherche d'une certaine comparaison entre les différentes approches de l'analyse textuelle aujourd'hui, nous n'avons pas pu, dans notre approche, intégrer cette dimension purement exploratoire de l'analyse que *Tropes* rend possible.

Trois exemples d'applications du logiciel *Tropes*

Rappels sur les trois corpus utilisés

Dans cette partie visant à illustrer les résultats que l'on peut obtenir avec l'utilisation d'un logiciel d'analyse par graphes tel que *Tropes*, nous utiliserons trois corpus recueillis à des dates différentes par le CRÉDOC. Ces corpus sont de natures différentes, ce qui permet d'illustrer les possibilités de *Tropes* face à ces contextes d'utilisations les plus fréquentes au CRÉDOC.

Tableau 27. Les trois corpus testés avec *Tropes*

Intitulé	Source	Nature	Remarques
<i>Température idéale</i>	Enquête « Comportements Alimentaires des Français », 1995	question ouverte entretien en face à face	1 600 réponses référence : Babayou, 1995
<i>Bonheur</i>	Enquêtes « Consommation », 1992 et 1996	question ouverte entretien téléphonique	deux fois 1 000 réponses références : Lahlou, 1993 et Babayou, 1997
<i>Prospective</i>	Recherche sur la prospective de la mobilité locale des personnes âgées, 1997	entretiens semi-directifs	neuf entretiens référence : Babayou, Volatier, 1997

Pour chacun de ces corpus, nous donnons dans un premier temps les résultats de la qualification globale que fournit *Tropes*.

Des graphes sont ensuite présentés, permettant en particulier de mettre l'accent sur les capacités exploratoires du logiciel.

Résultats pour le corpus « Température idéale »

Nous avons soumis à l'analyse de *Tropes* le corpus constitué des 1 603 réponses à la question sur la connaissance d'une température idéale de conservation des aliments frais. Cette analyse a pour avantage de rendre directement comparables les résultats de *Tropes* avec les deux analyses exposées dans la partie précédente de notre recherche.

Le premier résultat obtenu par l'analyse de *Tropes* concerne la qualification générale du texte. Le tableau suivant montre quels ont été les points relevés par le logiciel.

Tableau 28. Qualification du corpus « Température idéale »

Style	« plutôt argumentatif »
Mise en scène	prise en charge par le narrateur prise en charge à l'aide du « je »
Autres notions	syntaxe irrégulière détectée

Si le classement du style du texte en « argumentatif » n'est pas trop surprenante, puisqu'il s'agit de la retranscription de formulations d'attitudes et de comportements que les personnes qui les énoncent ont forcément tendance à justifier, en revanche les phrases types sur lesquelles *Tropes* fonde ce classement sont surtout des réponses qui expriment l'absence d'opinion. Le logiciel repère en effet la présence de négations (« ne... pas » comme dans « *je ne sais pas* », « aucune » comme dans « *aucune idée* », etc.) que l'on peut trouver dans la moitié du corpus étudié. Le tableau suivant donne les phrases qui ont été prises en compte pour établir ce classement.

Tableau 29. Phrases illustrant le style du corpus « Température idéale »

« ne sait pas, je ne vois pas »
« non je ne vois pas, c'est marqué dessus. »
« je ne sais pas du tout »
« ne sait pas »
« on doit marquer sur le produit la température pour bien le garder »
« je pense que c'est bien il faudrait le marquer quelque part »
« ah peut être 6 degrés mais je ne sais pas comment »
« je met la position 2 mais je ne sais pas si c'est bien »
« ne sait pas »
« par la vue si ca ne change pas de couleurs »
« ou alors plus si on le consomme plus tard. »
« je vois le bien si c'est très froid c'est à la couleur du produit »
« je ne conserve pas de produit frais comme les volailles »
« j'en sais rien »
« je n'en ai aucune idée »
« en principe c'est je ne sais pas quoi répondre »
« oh je ne sais pas je ne peux pas me rendre compte »
« je ne sais pas je mets au frigo »
« je ne crois pas savoir non vraiment »
« je ne sais pas on les met au frigo comme ca. »
« mais on ne regarde pas la température »
« il faut le garder au plus froid possible »
« aucune idée »
« ne sait pas je le mange de suite »
« aucune idée je regarde la date »
« et à la maison on applique les mêmes températures »
« je ne sais même pas je mets toujours le fromage et les laitages dans le bas et la viande en haut dans le plus froid »
« je ne vois pas trop ce que vous voulez qu'on réponde la »
« et c'est tout »
« j'en sais rien on met au frigo »
« au bas du frigo les produits frais se gardent bien »
« un thermomètre dans les appareils congélateur très propre »
« tout est au frigo »
« par rapport à la conservation en magasin »
« alors la je ne connais pas la température ni du yaourt ni du poulet »
« je le mets dans le frigo en haut ca ne me vient pas à l'idée »
« je le dis au toucher avec le dos de ma main il faut que ce soit bien froid »
« la couleur un yaourt gonfle s il n est plus bon c'est bon encore 1 semaine après »
« c'est dans le frigidaire bien sur non c'est tout »
« quand les produits s abîment vite on le remarque soi même il faut observer on se trompe une fois pas 2.faut les conserver au frigo »
« c'est marqué sur la viande peut être pas sur les yaourts »
« il faut le mettre aux environs de 6 degrés mais de les préserver enveloppés »
« aucune idée je demande si c'est un produit frais si oui congélation après achat »
« je ne regarde jamais de produits périssables plus de trois jours la question ne me concerne pas »
« je ne comprends pas votre question. »

Note de lecture : les réponses identiques, telles que « je ne sais pas » ne sont mentionnées qu'une seule fois.

Les mises en scènes sont quant à elles justifiées par les phrases présentées dans le tableau suivant.

Tableau 30. Phrases définissant la mise en scène du corpus « Température idéale »

« il le faudrait sur l'emballage »
 « je ne sais pas du tout »
 « c'est peut être écrit sur l'emballage »
 « on doit marquer sur le produit la température pour bien le garder »
 « je ne sais pas. »
 « je pense que c'est bien il faudrait le marquer quelque part »
 « ah peut être 6 degré mais je ne sais pas comment »
 « il faut être instruit on devrait faire nos éducations avant l'âge de 16 ans »
 « je regarde où ils se trouvent en hypermarché »
 « il faut regarder sur le petit livre du frigo »
 « il faut avoir un appareil spécial »
 « je juge »
 « je ne sais pas je laisse le congélateur sur 4. »
 « oh je ne sais pas je ne peux pas me rendre compte »
 « comment on peut faire je ne crois pas savoir non vraiment »
 « le poulet par exemple ça se voit à l'odeur »
 « il faut le garder au plus froid possible »
 « je ne vois pas trop ce que vous voulez qu'on réponde la »
 « je ne sais pas moi, je le laisse au frigidaire et c'est tout »
 « eh bien il faut mettre la température enfin le thermostat du frigidaire au moyen degré entre les deux 5 ou 4. »
 « il faut être au courant »
 « au niveau du frigo il faut avoir un thermomètre dedans »
 « il faut aller chez le volailler. »
 « je ne le sais plus je l'ai écouté à la télévision »
 « on sait disposer les aliments dans le frigo »
 « ne sait pas je ne vais pas commencer thermostat y a sûrement un truc mais quoi? »
 « je ne sais pas ma pauvre »
 « il faudrait qu'il y ait des thermomètres visibles et clairs dans les rayons »
 « si le frigidaire est réglé bien ça doit rester frais »
 « je ne sais pas, je fais au hasard »
 « ils le disent au poste on en parle sur les emballages »
 « le commerçant le sait. »

Note de lecture : les réponses identiques ne sont mentionnées qu'une seule fois.

On notera que les irrégularités syntaxiques relevées par *Tropes* sont justifiées en particulier par le fait qu'il s'agit ici d'un langage parlé, qui a conduit les personnes interrogées à de nombreuses hésitations.

L'analyse des univers de référence donnée par *Tropes* permet d'apporter une information supplémentaire par rapport aux analyses faites précédemment. Nous avons montré en particulier, par la lemmatisation d'*Alceste*, qu'il n'y avait pas d'effet d'écho massif et, de fait, la classification n'a pas été perturbée par un tel phénomène. Pourtant, *Tropes* montre que l'alimentation est la quatrième référence utilisée dans le corpus des réponses.

Tableau 31. Univers de référence du corpus « Température idéale »

Désignation	Effectif	Mots regroupés
matériel ménager..	370	réfrigérateur, frigidaire, frigo, congélateur
marchandise	166	emballage
outil	102	thermomètre, thermostat
aliment	85	yaourt, viande, volaille, fromage, poulet, légumes...
production	73	produit
pensée	62	idée, être au courant
commerce	27	commerçant, hypermarché, magasin, vendeur...
texte	21	notice, brochure, manuel
littérature	16	livre, bouquin
odorat.....	15	odeur, nez, pif
confiance.....	13	confiance

Les univers de référence qui sont mis en évidence par *Tropes* sont à la fois puissants comme lorsqu'ils regroupent sous une même bannière « odeur », « nez » ou « pif » ou encore lorsqu'ils forment le regroupement « commerce »¹, et insatisfaisants quand ils séparent les termes « texte » et « littérature ». Le logiciel permet certes de remédier à ce défaut en donnant à l'utilisateur la possibilité de construire un « scénario sémantique » qui sont des regroupements systématiques imposés par l'utilisateur.

Le logiciel *Tropes* propose une analyse des « mises en relations »² dont nous donnons ci-après un aperçu des plus significatives.

Tableau 32. Mises en relations significatives du corpus « Température idéale »

Relations	Eff.	Exemple de réponse
thermomètre → matériel ménager	(38)	« on met un thermomètre dans le frigo »
température → matériel ménager	(26)	« je me fie à la température du frigo »
matériel ménager → degré	(13)	« mettre le frigo à 6 degrés »
manuel → matériel ménager	(12)	« c'est marqué sur la notice du réfrigérateur »
matériel ménager → température	(10)	« mon frigo est toujours à la bonne température »

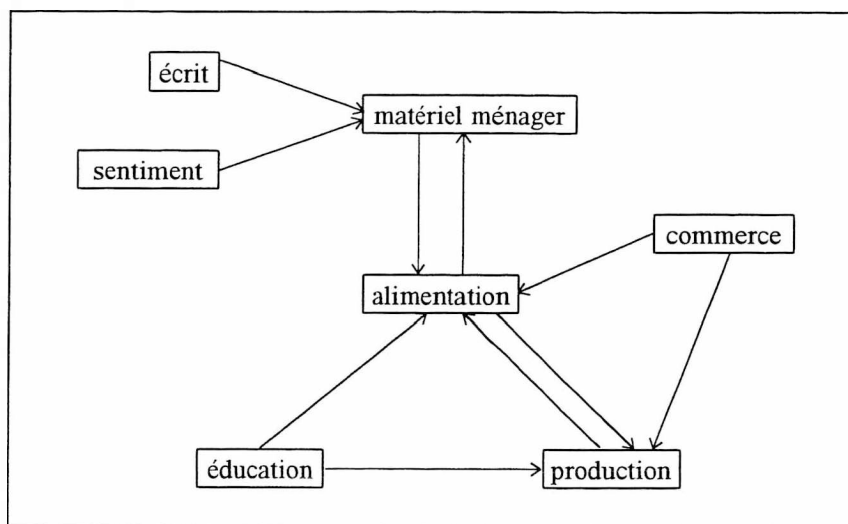
¹ à la limite près que notre contexte nous autorise à ne pas tenir compte de différences entre petits commerçants et grands distributeurs.

² ce qu'*Alceste* appelle segments répétés, sans toutefois les ordonner, comme c'est le cas ici.

La distribution des univers de référence montre qu'en fait il existe un effet d'écho non négligeable sur les aliments cités. L'absence de perturbations dans l'analyse d'*Alceste* s'explique donc surtout par cette variété des références alimentaires présentes dans le corpus, sans limitation au poulet et aux yaourts.

Les aliments cités le sont surtout par rapport à une utilisation du réfrigérateur, mais aussi par rapport à l'idée de goûter les produits, ou encore à la référence aux emballages et à la lecture des informations. Le graphe suivant, produit par *Tropes*, montre quelles sont les relations qui se nouent autour de l'univers de référence « alimentation ».

Figure 3. Graphe du corpus « Température idéale » centré sur l'univers « alimentation »

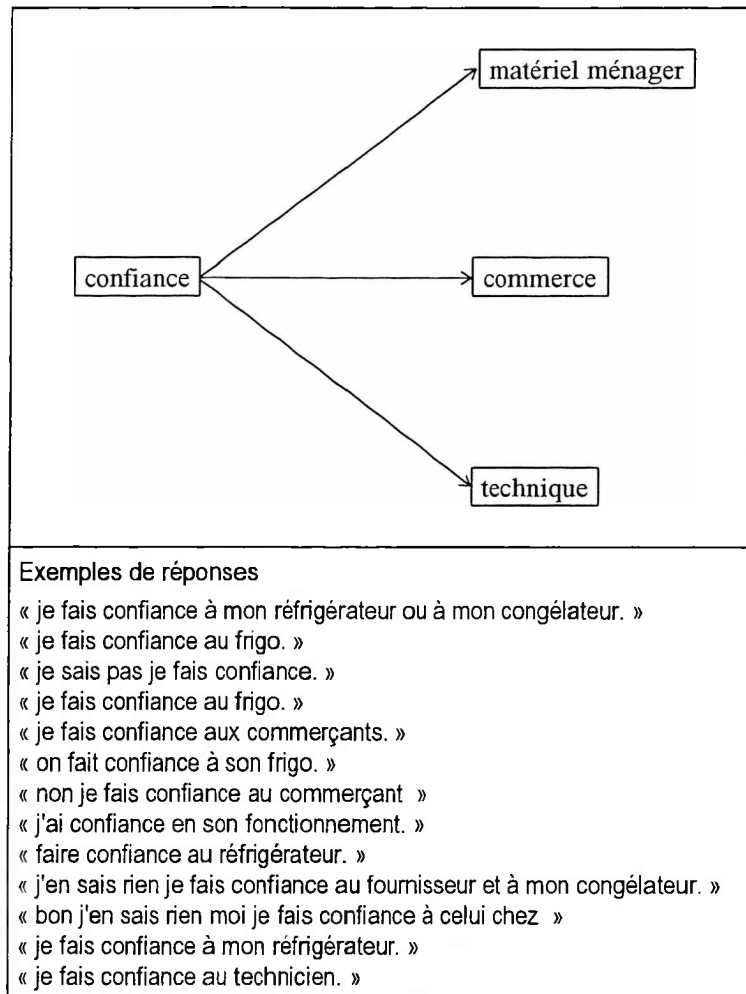


La valeur ajoutée de *Tropes* apparaît à travers ce style de représentations par graphe des univers de référence présents dans un corpus, ce que ne permet pas nécessairement une classification des unités de contexte par *Alceste*. En effet, pour que de telles relations apparaissent dans ce logiciel, il faudrait être assuré que l'algorithme de classification ait à chaque étape un nombre d'u.c.e. suffisant pour relier les notions entre elles et que cette liaison apparaisse à travers la hiérarchie produite. Notre expérimentation au sein de cette recherche montre que ce résultat ne peut être obtenu avec *Alceste* de manière simple.

De la même manière, rien ne permet de penser qu'une post codification permettra d'obtenir le même type d'information que celle présente dans les graphes d'un logiciel du type de *Tropes*, en particulier du fait qu'une codification ne permet pas de restituer simplement les notions d'actants et d'actés qui sont utilisées dans un graphe orienté.

Nous pouvons illustrer cette limite de la codification par un exemple. Nous avons pu montrer dans les traitements précédents de la question sur la température idéale qu'il existe une référence à la confiance, et que celle-ci peut se reporter sur différentes « personnes » : un technicien, un membre de la famille, le réfrigérateur lui-même. *Tropes* restitue cette information à travers le graphe suivant, que nous complétons à l'aide des phrases types qui sont prises en compte dans la notion de « confiance ».

Figure 4. Graphe du corpus « Température idéale » centré sur l'univers « confiance » et phrases types



Il est clair que la confiance dont il est question est uniquement utilisée en tant qu'actant. Cette propriété en fait un cas à part. Nous avons séparé, dans la grille de post codification, la confiance absolue dans le réfrigérateur du recours à un spécialiste. Ce dernier apparaît en effet à la fois en tant que personne vers laquelle se porte la confiance, donc acté, mais aussi en tant que personne qui a réglé un appareil auquel on ne touche plus, donc actant.

Pour montrer ce fait, il est nécessaire dans une phase préliminaire de construire un **scénario sémantique**, selon la terminologie employée par le logiciel *Tropes*. Il s'agit d'induire un traitement du corpus qui prenne en compte d'une manière unique les « spécialistes » auxquels les ménagères peuvent avoir fait appel, en particulier un technicien ou un membre de la famille. Ce type de manipulation des synonymes ne nous est pas inconnu : d'une certaine manière, la lemmatisation d'*Alceste* peut revenir à construire un groupe de termes n'ayant pas la même racine mais ayant dans un contexte donné la même valeur de sens. On peut par exemple penser au regroupement que nous avons effectué des termes « frigo », « réfrigérateur » et « frigidaire », mais aussi des termes « livre », « livret », « notice », « bouquin », « brochure », « manuel » sous le même concept de « mode d'emploi ».

Les ajustements que nous faisons à l'aide des scénarios sémantiques correspondent donc à une extension du travail effectué en standard par *Tropes* dans le but d'adapter le logiciel au contexte particulier du corpus. Dans notre cas, le spécialiste est, nous l'avons montré, un technicien ou encore un membre de la famille de la ménagère. On notera que cette mise en équivalence permet de préciser à *Tropes* des relations qui sont absolument dépendantes du contexte d'analyse du corpus. En effet, association hasardeuse en d'autres contextes, notre scénario sémantique conduit à considérer comme équivalents un « mari » et un « spécialiste de la réfrigération »...

Finalement, l'utilisation de *Tropes* sur le corpus « Température idéale » permet de mettre en évidence des informations qui n'avaient pas pu être détaillées à l'aide des méthodes traditionnelles déjà employées.

Résultats pour le corpus « bonheur »

Nous avons soumis, pour cette analyse, les deux corpus constitués par les réponses de 1 000 personnes interrogées en 1992 et en 1996 sur leur conception de ce que signifie « être heureux ». En 1992, l'analyse des réponses¹ avait été effectuée grâce à *Alceste*. En 1996, les

¹ Lahlou et alij, 1993

réponses ont été post codées, et le corpus de 1992 a été repris pour être soumis à la même grille de post codification, afin de mieux représenter les évolutions des représentations¹.

Le tableau suivant, extrait du traitement publié en 1997 de cette question, donne les évolutions dans les notions employées par les personnes interrogées aux deux dates d'enquête.

Tableau 33. Les vocabulaires utilisés en 1992 et 1996 pour répondre à la question « Qu'est-ce qu'être heureux ? »

Mot racine	1992	1996		Mot racine	1992	1996	
santé	317	397	↗	souci	-	65	
famille.....	359	308	↘	monde.....	-	62	
travail.....	176	226	↗	libre.....	-	61	
argent	208	211	=	ami.....	45	55	↗
enfant	165	203	↗	chômage.....	-	55	
bien	184	149	↘	voyage	37	52	↗
vie	209	143	↘	maison.....	-	48	
pouvoir.....	48	136	↗	bonheur.....	72	47	↘
vivre.....	122	109	=	problème.....	56	36	↗
temps	-	77		femme.....	39	33	=
amour, aimer.....	147	72	↘	peau.....	52	27	↘
heureux.....	103	72	↘	réussir.....	39	26	↘
vacances	-	69		mari+.....	41	23	↘

Source : CRÉDOC, Enquêtes consommation, fin 1992 et fin 1996

Guide de lecture : on indique pour chaque forme réduite (les mots sont ramenés à leur racine) le nombre d'occurrences apparaissant dans les deux corpus. Les réponses se portent respectivement en 1992 et 1996 à 998 et 1 005. Le signe '-' est utilisé pour désigner des effectifs qui étaient, en 1992, inférieurs à 40.

Ces analyses permettent de montrer que le discours des Français sur le bonheur est en particulier de plus en plus lié à une référence aux notions de « pouvoir » (par exemple au sens d'avoir le pouvoir de faire ce que l'on veut) et aussi de « temps » (avoir plus de temps) ou encore de liberté. Il s'agit là d'une approche globale que nous pouvons compléter, avec *Tropes*, avec la qualification générale que le logiciel peut donner du style et de la mise en scène du corpus des réponses lors des deux années d'enquête. Le résultat en est que les deux textes

¹ Babayou, 1997

sont qualifiés par *Tropes* de la même manière, c'est-à-dire un style « plutôt énonciatif » et une mise en scène à la fois « ancrée dans le réel » et « prise en charge à l'aide du 'je' ».

L'analyse des univers de référence selon *Tropes* permet d'obtenir une autre image des thèmes abordés par les personnes interrogées, qui diffère toutefois par des regroupements de termes dans un sens beaucoup plus étendu que la lemmatisation d'*Alceste*.

Tableau 34. Univers de référence des deux textes composant le corpus « bonheur »

1992		1996	
famille.....	379	famille.....	340
santé.....	242	emploi.....	288
emploi.....	219	santé.....	270
finance.....	217	finance.....	227
sentiment.....	208	sentiment.....	224
vie.....	204	enfant.....	208
enfant.....	190	loisir.....	188
gens.....	64	vie.....	133
environnement.....	52	habitat.....	70
loisir.....	52	gens.....	68
habitat.....	50	paix.....	38
végétal.....	41	femme.....	35
femme.....	41	environnement.....	31
paix.....	37	sécurité.....	27
liberté.....	24	voyage.....	25
voyage.....	18	pauvreté.....	24
climat.....	17	liberté.....	22
conflit.....	12	climat.....	19
mer.....	11	économie.....	18

Les regroupements en univers effectués par *Tropes* permettent de montrer que les quatre références principales citées sont : la famille, la santé, le travail et l'argent. On retrouve les mêmes résultats qu'avec *Alceste*, mais l'ordre est différent : famille-santé-emploi-finance pour *Tropes* et santé-famille-travail-argent pour *Alceste*. Les évolutions des principaux thèmes sont très concordantes entre les deux logiciels, pour les notions identifiées des deux côtés.

Certaines notions sont en revanche identifiées différemment par les deux logiciels, ce qui ne facilite pas les comparaisons. Par exemple, le fort accroissement des lemmes « pouvoir » et

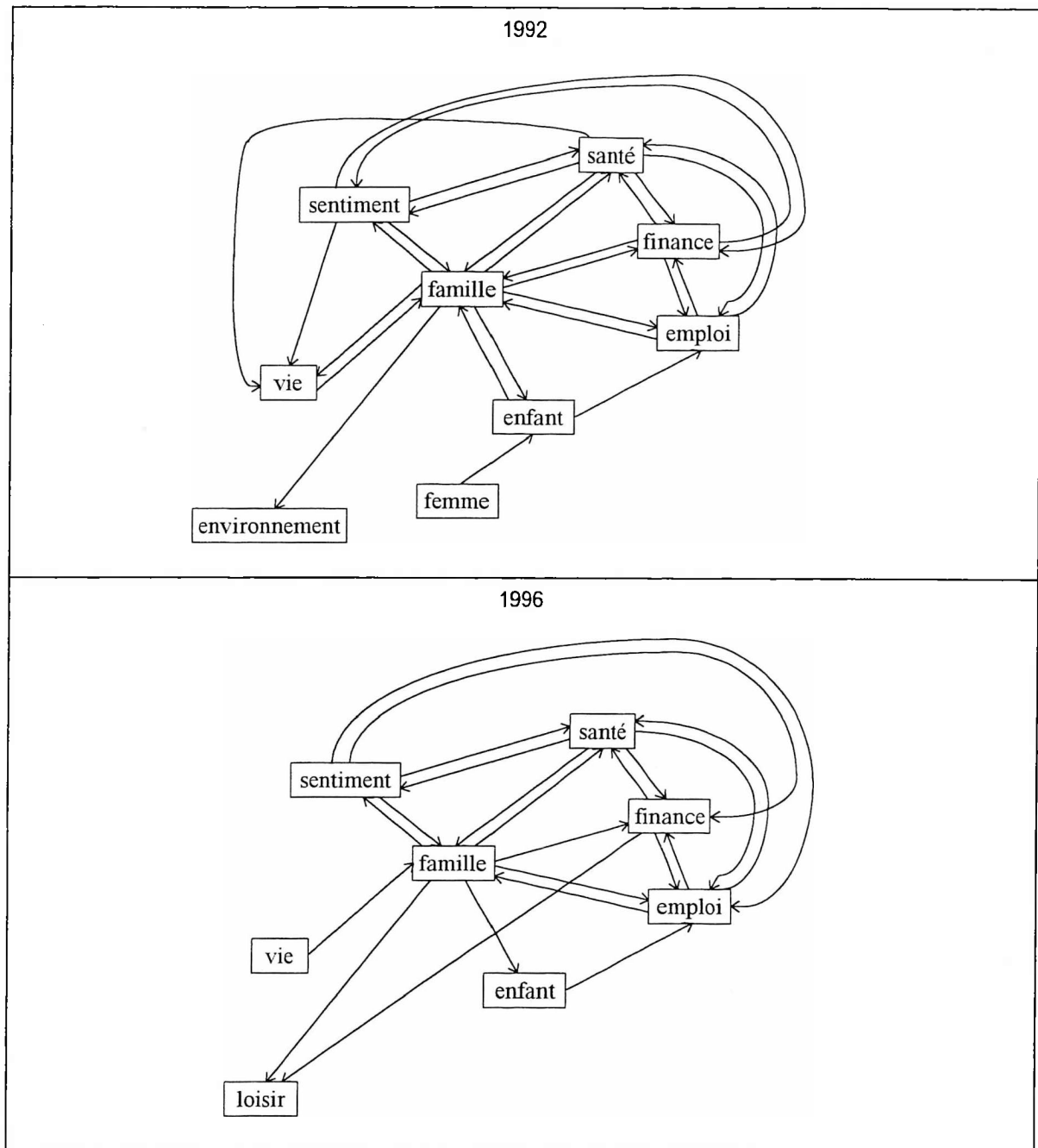
« temps » dans *Alceste* ne peut faire état dans *Tropes* qu'au développement de l'univers de référence « loisirs ». La stabilité de l'univers de référence « sentiment » apparaît contradictoire avec la décroissance des lemmes « amour, aimer » et « heureux » mesurée par *Alceste*.

Figure 5. Évolution des thématiques

		Tropes		
		en hausse	stable	en baisse
Alceste	en hausse	emploi, santé, enfant	-	-
	stable	-	argent	-
	en baisse	-	-	famille

Ceci nous amène à nous intéresser au graphe des univers de référence centré autour de cette notion de famille.

Figure 6. Graphes des principaux univers de référence du corpus « bonheur » en 1992 et en 1996



Si un noyau dur subsiste entre emploi, famille, santé, sentiments (les (petit(e)s) ami(e)s, l'amour...) et finance (c'est-à-dire l'argent), on voit aussi que d'autres relations évoluent, telle la relation à l'enfant qui mène systématiquement au travail et aux préoccupations liées au chômage, la conception de la vie qui est de plus en plus une vie de famille, et surtout

l'apparition d'un désir de disposer de loisirs, ce qui n'est pas incompatible avec la volonté de disposer de pouvoir et de temps que nous avons déjà montrée.

L'intérêt principal du travail sur le corpus « bonheur » est de donner la possibilité de qualifier l'évolution dans le temps d'un discours représentatif des Français sur une attitude.

Résultats pour le corpus « prospective »

Cette dernière analyse porte sur un corpus composé de neuf textes qui sont les retranscriptions d'entretiens semi directifs réalisés par le CRÉDOC auprès d'experts de différentes disciplines sur la prospective de la mobilité locale des personnes âgées, dans le cadre d'une recherche financée par le PREDIT et le CNRS PIR Villes. Seule une approche par thèmes abordés a déjà été proposée à partir de ces entretiens.

Le tableau suivant résume les thèmes qui avaient été abordés par les experts lors de ces entretiens.

Tableau 35. Thèmes abordés dans les entretiens du corpus « prospective »

Thèmes	Sous thèmes
Définition de la population âgée	étapes dans le cycle de vie vie seule ou en couple analyse des effets de la dépendance
Revenu futur des retraites	historique de l'évolution du niveau de vie hypothèses pour les évolutions futures
Enjeux de la population âgée pour la consommation	caractéristiques de la demande des personnes âgées arbitrages consommation – épargne potentiels de l'approvisionnement à distance
Environnement et cadre de vie	localisation géographique influence du milieu de vie sur la mobilité adaptation de la ville aux personnes âgées mobilité et sociabilité
Services à domicile	état de la demande influence des technologies analyse de la mobilité générée par les personnes âgées

L'analyse des univers de référence produits par *Tropes* permet de montrer comment le logiciel restitue un classement analogue des thèmes abordés dans l'ensemble des neuf retranscriptions qui constituent le corpus.

Tableau 36. Principaux univers de référence du corpus « prospective »

Univers	Eff.	Univers	Eff.
vieillesse.....	66	transport	18
gens	64	problème.....	14
retraite.....	34	nation	13
ville.....	25	femme	13
famille.....	25	environnement.....	12
transport terrestre....	25	transport par rail.....	10
logement.....	19	paiement.....	9
vie	19	peur	9

Cette première approche ne fait bien sûr que confirmer l'analyse « à la main » qui avait été faite de ce corpus.

L'utilisation du logiciel *Tropes* permet de tirer parti de quelques-unes de ses caractéristiques originales, par exemple en différenciant nettement les entretiens entre eux. Deux points de vue semblent intéressants : la qualification globale des neuf textes les uns par rapport aux autres d'une part, l'examen des différences dans les approches d'une même référence par chacun des neuf experts d'autre part.

Tableau 37. Qualification globale des neuf entretiens du corpus « prospective »

Entretien	Style	Mise en scène	Autres
n°1	plutôt argumentatif	prise en charge par le narrateur	éléments de doute détectés
n°2	plutôt argumentatif	ancrée dans le réel	
n°3	plutôt argumentatif	ancrée dans le réel	éléments de doute détectés
n°4	plutôt argumentatif	ancrée dans le réel	éléments de doute détectés
n°5	plutôt argumentatif	prise en charge par le narrateur	
n°6	plutôt argumentatif	ancrée dans le réel	éléments de doute détectés
n°7	plutôt argumentatif	ancrée dans le réel	
n°8	plutôt argumentatif	ancrée dans le réel	
n°9	plutôt argumentatif	dynamique, action	

A partir de ce tableau, il est clair que, parlant du même sujet dans un même contexte d'entretien, le discours de chaque expert peut s'écarter d'une certaine norme de présentation.

Chaque entretien réalisé rend compte en fait de quelques notions précises. Ceci s'explique par le fait que les domaines de spécialité des experts ne les amenaient pas forcément à s'exprimer avec précision sur chacun des thèmes abordés dans la recherche sur la mobilité locale des personnes âgées. *Tropes* permet ainsi de différencier les différents entretiens selon la dispersion des principaux univers de référence qui ont été abordés par chacun.

Tableau 38. Principaux univers de référence des entretiens du corpus « prospective »

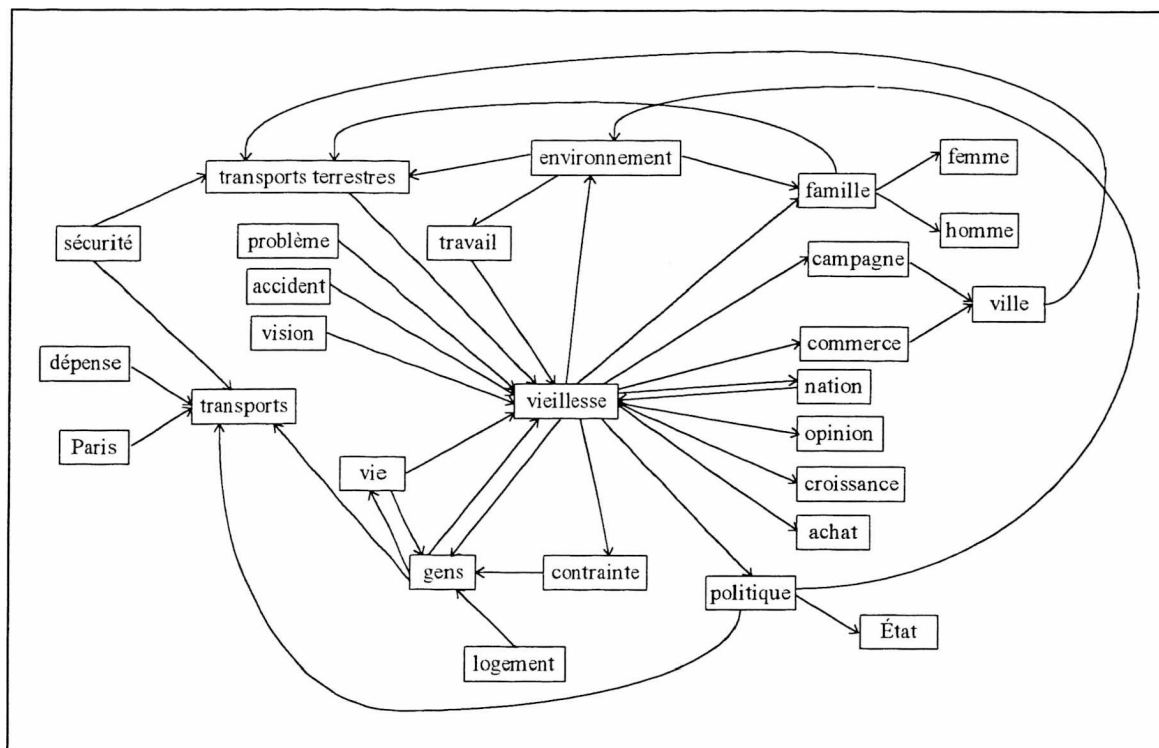
Entretien n°1	vieillesse (17)	gens (11)	vie (6)	famille (5)	pensée (5)
Entretien n°2	famille (6)	gens (6)	retraite (5)	logement (4)	
Entretien n°3	gens (15)	vieillesse (12)	peur (7)	famille (7)	femme (5)
Entretien n°4	retraite (36)	économie (20)	vieillesse (14)	paiement (12)	environnement (5)
Entretien n°5	ville (7)	retraite (7)	gens (7)	vie (6)	femme (5)
Entretien n°6	vieillesse (14)	gens (6)	travail (6)	retraite (4)	vie (3)
Entretien n°7	gens (9)	transports (7)	vieillesse (7)	transport terrestre (5)	logement (5)
Entretien n°8	retraite (15)	vieillesse (10)	ville (8)	consommation (3)	gens (3)
Entretien n°9	transport par rail (7)	gens (7)	ville (6)	transport terrestre (5)	

Au delà des différences dans les thèmes abordés (information qu'une lecture diagonale des neuf retranscriptions aurait de toute façon donnée plus rapidement), c'est surtout l'homogénéité de chaque sous-corpus qui est intéressante. On constate ainsi que certains experts ont eu tendance à ne s'exprimer principalement que sur un nombre limité de sujets, tandis qu'au contraire d'autres ont été beaucoup plus prolixes sur plusieurs sujets.

A partir de cette analyse, il est bien entendu possible de produire un graphe pour un univers de référence commun à plusieurs entretiens. Toutefois, cette analyse n'apporte pas pléthore d'informations car il s'avère que les différents sujets abordés ont toujours eu tendance à l'être selon une même logique et un même ordre. Sur le point des implications et imbrications entre problèmes, les neufs experts convergent vers un formalisme que l'on peut donc représenter par une vision globale du corpus composé de l'ensemble des entretiens.

Nous illustrons cette vision globale à travers un graphe des univers de référence tels que *Tropes* les organise autour du terme « vieillesse », commun à l'ensemble des entretiens. Nous pouvons ainsi montrer que le logiciel identifie l'ensemble des thèmes qui ont été évoqués par les experts.

Figure 7. Graphe du corpus « prospective »



Dans l'ensemble, les analyses que l'on a pu produire sur le corpus « prospective » ont été plutôt des approches globales des neuf textes de retranscription des entretiens. On voit clairement que ce type de résultats a pour avantage majeur de permettre à un chercheur d'appréhender globalement un corpus d'entretiens, en différenciant rapidement les différents modes de discours qui ont été adoptés par les personnes interrogées. A l'usage, il est toutefois visible que *Tropes* s'avérerait être un outil assez lourd pour tout ce qui a trait aux différences d'approches entretien par entretien. On pourrait ainsi lui reprocher l'absence d'une forme d'automatisation de l'analyse comparative des entretiens qui lui sont soumis.

Synthèse des résultats obtenus par le logiciel *Tropes*

Nous avons volontairement mis l'accent sur la première caractéristique de *Tropes* qui est de proposer une qualification globale du style d'un texte. Certaines différences entre les textes étudiés ont ainsi pu être mises en évidence. Dans le dernier cas que nous avons examiné, le corpus « prospective », des résultats rapides peuvent ainsi être obtenus sur des critères de distinctions entre les styles des textes qui composent le corpus.

Toutefois, le travail sur le corpus « température idéale » conduit à douter de l'efficacité et du caractère signifiant d'un tel critère de style général d'un texte, ce que nous avons pu montrer en particulier à travers une polarisation sur les réponses de la classe « sans opinion ». Au passage, on remarquera que *Tropes* ne traite guère mieux qu'*Alceste* les réponses indiquant une absence de connaissance ou d'opinion de la part des personnes qui répondent à une question ouverte d'attitude ou de comportement. Cela est dû dans les deux cas au fait que ce sont finalement les substantifs qui priment dans l'interprétation que ces logiciels font d'un corpus. Cette démarche est aussi symptomatique de l'erreur qui a été commise par *Tropes* -comme par *Alceste*- de considérer le mot « lire » comme le substantif désignant une monnaie (avec dans *Tropes* un univers de référence explicitement financier) et non comme un verbe, alors même que ce mot était suivi d'un complément d'objet direct.

Le travail sur les univers de référence produits par *Tropes* permet d'une certaine manière de représenter ce logiciel sous l'aspect d'une aide à la post codification. En effet, les corpus de questions ouvertes qui ont été analysés nous permettent de repérer des thèmes saillants, ce qui peut conduire à la construction d'une grille de codification des réponses, par repérage des principales thématiques évoquées.

Enfin, *Tropes* apporte un éclairage original sur les enchaînements entre les idées qui sont présentes dans le discours analysé, ce que rend possible l'approche par l'intermédiaire d'un graphe orienté. C'est surtout visible dans les analyses que l'on peut mener en termes d'actants et d'actés.

Au delà de toutes ces caractéristiques et de ces aspects novateurs par rapport aux travaux qui avaient déjà pu être effectués au CRÉDOC, il est certain que la version que nous avons pu tester de *Tropes* reste un outil puissant d'exploration d'un corpus, peu orienté vers la restitution de classes types de réponses adaptées à la présentation de la structuration des réponses à une question ouverte.

Perspectives

Ce rapport montre clairement que le potentiel d'amélioration des outils lexicométriques existant reste grand. Le nombre des méthodes que nous avons envisagées et testées, le nombre de questions que nous avons laissées en suspens, le nombre d'incertitudes qui ressortent des analyses, les champs des complémentarités entre les méthodes qui ne sont certainement pas encore tous explorés... tous ces éléments fondent aujourd'hui la richesse des perspectives pour les recherches en analyse textuelle.

L'approche par la classification automatique, même soutenue par de puissants outils de réduction synonymique et d'analyse grammaticale, ne doit pas être considérée comme la seule voie possible dans l'exploration d'une question ouverte.

Les approches par réseaux sont séduisantes mais il faut rester vigilant sur les équilibres à atteindre entre les possibilités quasi infinies d'exploration du corpus qui peuvent longuement occuper l'analyste et sa nécessaire description.

On doit par ailleurs souligner la complémentarité entre les approches par graphes et celles par classification automatique, les premières pouvant venir d'outil exploratoire pour préparer les secondes.

Il y a également une bonne complémentarité entre l'analyse des réponses par réseau et la post codification. Afin d'utiliser la statistique et les classifications automatiques pour améliorer une post codification, il faudrait davantage penser à une utilisation de type analyse discriminante,

c'est-à-dire à l'affectation de réponses à des classes déjà constituées. On aurait ainsi une chaîne de traitements cohérente :

1. analyse exploratoire par analyse du discours au moyen d'un outil de type « graphes ».
2. constitution contrôlée manuellement de classes de réponses .
3. affectation automatique des répondants par analyse discriminante.

Cette constitution d'une chaîne de traitement semi-automatique rendrait nécessaire le développement informatique de passerelles entre les différents outils existants. Nous travaillerons dans les prochains mois sur l'étude de la faisabilité d'une telle chaîne de traitements.

Les développements actuellement réalisés par les différentes équipes auteurs des logiciels nous laissent présager la possibilité de tester de nouvelles fonctionnalités dès l'année prochaine.

Bibliographie

- [1.] BABAYOU P., 1995 - *Nutrition et hygiène alimentaire : attitudes et croyances des ménagères en 1995*, Collection des Rapports n°164, CRÉDOC
- [2.] BABAYOU P., 1997 - *La consommation en 1997, vers le cyber consommateur ?* Cahier de recherche n°99, CRÉDOC
- [3.] BABAYOU P., VOLATIER J.-L., 1997 - *Prospective de la mobilité locale des personnes âgées*, Prédit, PIR Villes, CRÉDOC
- [4.] BAUER D., MARESCA M., 1992 - *Lignes de vie : méthodologie de recueil et de traitement des données biographiques*, Cahier de recherche n°37, CRÉDOC
- [5.] BEAUDOUIN V., 1993 - *Analyse lexicale et stylistique : Gravitations de Jules Supervielle*, Cahier de recherche n°49, CRÉDOC
- [6.] BEAUDOUIN V., HÉBEL P., 1994 - *Avancées en analyse lexicale*, Cahier de recherche n°61, CRÉDOC
- [7.] BEAUDOUIN V., AUCOUTURIER A.-L., 1994 - *Parcours d'insertion de jeunes en difficulté*, Cahier de recherche n°66, CRÉDOC
- [8.] BEAUDOUIN V., 1995 - *Analyse textuelle et structures narratives de récits*, Cahier de recherche n°82, CRÉDOC
- [9.] BEAUDOUIN V., BROCHET F., 1996 - *Analyse lexicale de corpus en anglais*, Cahier de recherche n°95, CRÉDOC

- [10.] BLOT I., 1993 - *L'apport d'une classification automatique pour étudier une question ouverte : les réponses orales de commerçants*, Actes des secondes journées internationales d'analyse de données textuelles, Telecom Paris ; S.J.Anastex
- [11.] BOURDIEU P., 1980 - « *L'opinion publique n'existe pas* », in *Questions de sociologie*, Éditions de Minuit
- [12.] BRUGIDOU M., LE QUÉAU P., 1995 - *L'analyse des entretiens non-directifs par la méthode des rafales*, JADT
- [13.] COLLIERIE DE BORELY A., 1996 - *Consommateur et préférences de consommation en 1996*, Cahier de recherche n°88
- [14.] GHIGLIONE R., KEKENBOSCH C., LANDRÉ A., 1995 - *L'analyse cognitivo-discursive*, Presses Universitaire de Grenoble
- [15.] GILLES M.-O., 1995 - *Les spécificités des enquêtes quantitatives auprès de populations socialement marginales*, Cahier de recherche n°68, CRÉDOC
- [16.] HAUESLER L., 1987 - *Analyse lexicale de réponses libres, le coût de l'électricité*, Collection des rapports du CRÉDOC, n°14
- [17.] JENNY J., 1997 - *Méthodes et pratiques formalisées d'analyse de contenu et de discours dans la recherche sociologique française contemporaine. État des lieux et essai de classification*. Bulletin de Méthodologie Sociologique, n°54
- [18.] LAHLOU S., BEAUDOUIN V., COLLIERIE DE BORELY A., 1993 - *Où en est la consommation aujourd'hui ?* CRÉDOC, Collection des cahiers de recherche, n°46
- [19.] LE QUÉAU P., 1997 - *Les fleurs mystiques de Babylone*, Thèse de doctorat, Paris V
- [20.] YVON F., 1990 - *L'analyse lexicale appliquée à des données d'enquête : état des lieux*, Cahier de recherche n°5, CRÉDOC

Dépôt légal : Septembre 1997

ISSN : 1257-9807

ISBN : 2-84104-087-9

CAHIER DE Recherche

Récemment parus :

Méthode d'étude sectorielle - Vol. 2

Philippe MOATI - n°93 (1996)

Modélisation des choix alimentaires des ménages

Patrick BABAYOU, Aude COLLIERIE DE BORELY - n°94 (1996)

Analyse lexicale de corpus en anglais

Valérie BEAUDOUIN, Frédéric BROCHET - n°95 (1996)

Évaluer ou l'esprit des méthodes

Michel LEGROS, Guy POQUET - n°96 (1996)

Approche de la structure du paysage associatif dans le domaine de l'environnement

Bruno MARESCA - n°97 (1996)

**Les nouvelles logiques productives dans les PMI :
déterminants et impact sur les performances**

Philippe MOATI, Laurent POUQUET - n°98 (1997)

La consommation en 1997 : vers le cyber consommateur ?

Patrick BABAYOU - n°99 (1997)

**L'impensé rebelle : l'identification des facteurs
d'incertitude dans les enquêtes sur fichiers**

Patrick DUBÉCHOT, Marie-Odile SIMON - n°100 (1997)

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

ISBN : 2-84104-087-9

CRÉDOC

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