CAHIER DE RECHERCHE

SEPTEMBRE 1996 ■ N° 95

ANALYSE LEXICALE DE CORPUS EN ANGLAIS

Valérie BEAUDOUIN
Frédéric BROCHET

Département "Prospective de la consommation"

CRÉDOC
ENTREPRISE DE RECHERCHE

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

CRÉDOC

ANALYSE LEXICALE DE CORPUS EN ANGLAIS

Valérie BEAUDOUIN

Frédéric BROCHET

Avec la collaboration de

Aude COLLERIE de BORELY

Claire EVANS

Ségolène EVEN

Chantal RENAULT

SEPTEMBRE 1996

142, rue du Chevaleret
75013 - PARIS

Sommaire

INTRODUCTION[1]

[1] Nous tenons à remercier Patrick Mac Leod pour la pertinence de ses commentaires et la relecture de ce rapport.

Nous présentons dans ce cahier de recherche l'état d'avancement d'un chantier en cours. Depuis le début de 1995, nous sommes confrontés à des corpus de textes en anglais. A cette époque, nous ne disposions pas de méthode d'analyse de statistique textuelle raffinée comme pour le français : le logiciel Alceste, conçu et réalisé par Max Reinert [1983, 1990]. Nous avons déjà eu l'occasion de présenter son fonctionnement et des exemples d'utilisation dans divers cahiers de recherche. On pourra se reporter à [Beaudouin et Lahlou, 1993]. Il faut rappeler que les méthodes d'analyse des données, et plus particulièrement les méthodes de statistique textuelle, ont vu le jour et se sont principalement développées en France, sous l'égide de J.P. Benzécri [1981]. Elles se sont diffusées tout le long du pourtour méditerranéen, mais n'ont pas atteint les pays anglo-saxons, ce qui explique le « vide » en statistique textuelle anglaise. L'examen de l'origine géographique des participants aux JADT (Journées internationales d'Analyse des Données Textuelles [S. J Anastex, 1993] et [Bolasco, Lebart et Salem, 1995]) en est une preuve parmi d'autres.

La question était : comment faire de l'Analyse des Données Textuelles (ADT) sur des textes anglais ?

Dans le but de rendre les fichiers de sortie (classes de la typologie) indépendantes de la langue des fichiers d'entrée (corpus), trois voies étaient possibles.

- La première voie, *l'autoroute*, aurait consisté à faire traduire les textes de l'anglais vers le français, soit par des traducteurs professionnels, soit à l'aide d'outils de TAO (Traduction Assistée par Ordinateur). L'analyse aurait ensuite porté sur des textes en français, un domaine bien balisé. Cette voie est en cours de test. Pour cela, il fallait disposer d'un traducteur automatique mais aussi de la version en anglais d'Alceste pour comparer les analyses sur textes en anglais et sur textes traduits.

- La seconde voie, ou *voie rapide,* mise en place pour répondre aux exigences de résultats rapides de l'enquête bilingue « Recherche et Environnement » a conduit à indexer les textes par mots-clefs, à procéder à une traduction sommaire de ces mots-clefs puis à effectuer des analyses sur les corpus de réponses réduites à leurs mots-clefs. Pour cela, il a quand même fallu tester la cohérence des analyses effectuées sur des textes indexés par

rapport aux analyses sur texte brut. Cette comparaison, qui fait l'objet d'une partie de ce rapport, a été effectuée sur les réponses en français. Nous ne disposions pas alors de la version anglaise d'Alceste.

- La troisième voie, ou *voie lente*, avait un double objectif : établir quels étaient les éléments indispensables pour construire une version d'Alceste adaptée à l'anglais puis mettre en place ces éléments. En vue du premier objectif, de nombreux tests comparatifs ont été effectués pour voir s'il était utile ou non de construire un dictionnaire de mots-outils (la question se pose déjà pour l'analyse d'autres langues) et de lemmatiser... Les tests n'ont pas tous abouti à des conclusions concordantes. Des difficultés liées à la nature des corpus analysés ont été rencontrées qui nous ont amenés à faire des détours par rapport à notre objectif initial. En effet, la mise en place d'une version en anglais a eu pour effet indirect de proposer des améliorations pour la version française d'Alceste, dont certaines sont déjà prises en compte dans la nouvelle version qui devrait être installée au CRÉDOC d'ici la fin de l'année 1996.

Cette interrogation est née de la nécessité d'analyser les résultats de l'enquête « Recherche et Environnement » qui est essentiellement constituée de questions ouvertes en anglais ou en français. Les résultats de cette étude figurent dans [Laredo, Volatier et Collerie de Borely , 1996] et devraient donner lieu à la publication d'un rapport d'ici la fin de l'année 1996.

Nous n'étions pas les seuls à vouloir traiter des textes en anglais. Hormis la *voie rapide* qui ne concernait que le CRÉDOC, tout le travail a été réalisé en collaboration avec Frédéric Brochet, doctorant de l'EPHE sous la direction de Patrick Mac Leod, qui avait lui aussi de gros corpus en anglais à analyser.

Concrètement, il existe aujourd'hui une version en anglais d'Alceste avec reconnaissance des mots-outils et lemmatisation : les dictionnaires utilisés ont été constitués par nous-mêmes puis adaptés et intégrés au logiciel par Max Reinert. La reconnaissance des mots-outils semble satisfaisante au regard des différents tests effectués, mais la qualité de la lemmatisation n'a pas encore été testée.

L'organisation de ce rapport ne correspond pas à la progression chronologique de notre travail : dans un souci de clarté et de concision, nous avons préféré suivre un enchaînement logique dicté par les aspects algorithmiques de notre travail. Nous commençons par revenir sur les problèmes de limitations en taille des tableaux gérés par le logiciel Alceste, que nous avons à nouveau rencontrés en dépit d'une très grande amélioration par rapport à la version précédente. Nous attendons avec impatience l'installation de la nouvelle version qui devrait enfin permettre de traiter correctement nos corpus les plus volumineux. Ensuite nous abordons la « voie lente », c'est-à-dire les différentes étapes de la constitution de la version anglaise d'Alceste. Nous en venons alors à la « voie rapide », l'indexation par mots-clefs pour tester la pertinence et la solidité de cette approche. La voie de la traduction est actuellement en cours de test et donnera lieu prochainement à une présentation des résultats.

PRÉSENTATION DES CORPUS

La mise au point de la version anglaise d'Alceste a nécessité de nombreux tests sur des corpus en anglais mais aussi en français. Les corpus sont les suivants :

- 10443 *fiches de dégustation de vins*, soit environ 3,5 MO, rédigées en **anglais** par un seul auteur, Robert Parker, critique de vin américain. Une fiche se présente de la manière suivante (la description comprend en moyenne 50 mots) :

  **** *pr_Chateau Mont Redon Chateauneuf Du Pape *m_1990 *a_Rhone *n_85 *t_Red
  Mont Redon has been utilizing some small oak barrels for aging a percentage of their red wine, and this has resulted in an almost Bordeaux-like structure and austerity in certain vintages. The 1990 displays deep ruby color, a ripe, medium to full-bodied, highly structured feel on the palate, plenty of tannin, and a closed, firm style. After bottling, there is significantly less to the wine than prior to its filtration. Drink it over the next 4-8 years before it dries out.

  Dans les fiches de dégustation rédigées par Parker, la première ligne (« étoilée ») consigne les informations figurant sur l'étiquette de la bouteille (propriété, appellation, millésime, grande région, couleur) et le texte qui suit est la description du produit.

  Ce corpus est utilisé dans le cadre d'un travail de doctorat (Frédéric Brochet) en neurosciences cognitives qui traite des aspects cognitifs de la dégustation, conduit à l'EPHE, sous l'égide de Patrick Mac Leod. Dans la problématique, la ligne étoilée est considérée comme l'objet, et le texte associé comme sa représentation. On cherchera donc à extraire du corpus les « mondes lexicaux » [Reinert, 1993] que l'on interprète en terme de « noyaux de représentation » [Lahlou, 1995]. Concrètement, on construit une typologie (où chaque classe est censée suggérer un noyau) à partir des cooccurrences de mots à l'intérieur d'une fiche.

- un corpus de *fiches de dégustation* rédigées en **français**. Ce corpus (utilisé par F. Brochet) est constitué des 32 000 fiches de dégustation des cinq derniers *Guide Hachette des vins de France,* aimablement communiqués sous forme numérique par les éditions Hachette. Le volume de ce corpus s'élève à 11 Méga-Octets. L'analyse a porté sur une sélection au hasard de 4 000 fiches, soit 1,2 MO.

- les 4 962 descriptions en **anglais** de *problèmes d'environnement* (0,7 MO) en réponse à la question ouverte : « Please provide maximum details when identifying the eight major environmental problems of the future, stressing possible areas of interdependence. ».

- les 2 805 descriptions en **français** de *problèmes d'environnement* (0,4 MO) en réponse à cette même question : « Veuillez préciser avec le maximum de détails quels seront dans le futur les huit problèmes d'environnement prioritaires, en insistant sur les interdépendances possibles » rédigées par les chercheurs francophones.

Cette question qui admet au maximum huit réponses a été posée dans le cadre de l'enquête « Recherche et Environnement ». Le CRÉDOC est un des partenaires principaux de cette enquête européenne auprès de chercheurs du monde entier réalisée en France sous la coordination de J. Theys et C. Courtet. Le questionnaire a été envoyé à près de dix mille chercheurs. Il était rédigé en français pour les pays francophones (donc rempli en français) et en anglais pour tous les autres pays. Un millier (1030 exactement) de chercheurs du monde entier ont eux-mêmes rédigé les réponses.

La disparité dans les réponses est très grande comme le montre la Figure 1 : longueur très variable, construction, articulation logique présentes ou absentes...

Questionnaire peu rempli	n° pb	Questionnaire très rempli
Demographic explosion and famine.	1	Much to my surprise what I deem to be most important doesn't figure in your list : improved energy efficiency. It is the most rational "source" of energy, a classical cross-issue task, and probably the most realistic way to implement reduction of energy needs (and related pollution), since Western publics don't seem to be keen to reduce their standard of living.
Overuse of fossil fuels. Involves depletion and pollution.	2	Since increased efficiency won't be sufficient, development of renewable -and hence sustainable- energy sources (solar would be best) is of utmost importance.
Terrorism.	3	Next, and equally basic, preservation of high quality drinking water resources seems a problem of both global and especially regional concern that not yet receive the amount of attention it deserves.
Uncontrolled disease.	4	We need to keep an eye on climate change. The -potential- scope and the irreversibility of its consequences make this an issue worth every attention. In the light of research results, no. 1 and 2 above may even increase in importance.
Impact on short-term priority (e.g. CAP) on long term environmental quality.	5	The option of nuclear power should be kept open, through it won't solve all energy problem and in itself it is highly problematic, especially the problem of nuclear waste disposal (long-term storage, save against links and terrorism and abuse for military purposes) deserve high attention.
Nuclear accidents/explosion.	6	Next, the bundle of problems you list under "Development : population growth" becomes an ecological problem and "eats up" economic development. Yet, there is no ethical way of population control except through development, especially for women. Only save prospects for a decent living will lead people to reduce their number of offspring.
Climate change.	7	The value problematic is next in importance: democracy gets under stain where it has taken roots long ago (in the West, that is) through a lack in civic responsibility and public contempt for politics; the rules of the democratic game have yet to be internationalized in many Eastern and Southern countries.
Biodiversity loss.	8	Soil conservation, like water preservation, gets too little attention so far. Even post-industrial societies will come to understand that they still depend on high quality soil, not concrete.

Figure 1 : Exemples de questionnaires (peu ou très rempli)

Ce sont les problèmes prioritaires qui sont ici retenus comme individus statistiques et non les chercheurs eux-mêmes.

Les textes des réponses ont été saisis *intégralement* au CRÉDOC dans la langue de rédaction. Ensuite, chaque problème a été indexé par mots-clefs (au maximum six). Ce travail a été effectué par Ségolène Even, Mathieu Fortineau et Zohra Amara.

Cette codification était nécessaire pour réaliser des analyses fines par mots-clefs. Par exemple, le mot-clef « sustainable development » a été affecté chaque fois qu'une réponse évoquait la problématique de « développement durable », « préférence pour le long terme », etc.

Cette option méthodologique présente en outre l'avantage de contourner le délicat problème de la traduction. Dans certains cas, l'existence de nuances dans les réponses francophones ou d'expressions-clefs non traduisibles —comme « parcellisation de l'espace »— a conduit à garder des mots-clefs français ou anglais non traduits.

Pour l'enquête « Recherche et Environnement », les noms suivants seront parfois utilisés pour désigner les corpus :

	Textes indexés par mots-clefs	**Textes bruts**
anglais	TIAN	TBAN
français	TIFR	TBFR

1. PROBLÈMES POSÉS PAR LA TAILLE DES CORPUS

La dernière version du logiciel Alceste sous UNIX permet en théorie de traiter des corpus de toutes tailles (jusqu'à 10 Méga-Octets). L'analyse construit un tableau qui contient en ligne les unités de contexte (réponse à une question ouverte, fiche de dégustation...), en colonne les mots du vocabulaire. A l'intersection d'une ligne et d'une colonne, on a « 1 » si le mot est présent dans l'unité de contexte, « 0 » sinon.

Dans la pratique, la taille du tableau analysé est soumise à trois contraintes :

1. Le nombre d'unités de contexte, c'est-à-dire le nombre de lignes dans le tableau analysé, (dans notre cas les unités analysées sont les problèmes d'environnement ou les fiches de dégustation), qui est limité à 10 000 ;
2. Le nombre de mots analysés (les colonnes du tableau) limité à 1 400 ;
3. Le nombre de « uns », c'est-à-dire d'occurrences dans le tableau, fixé à 90 000 (le tableau ne doit pas être rempli à plus de 6%).

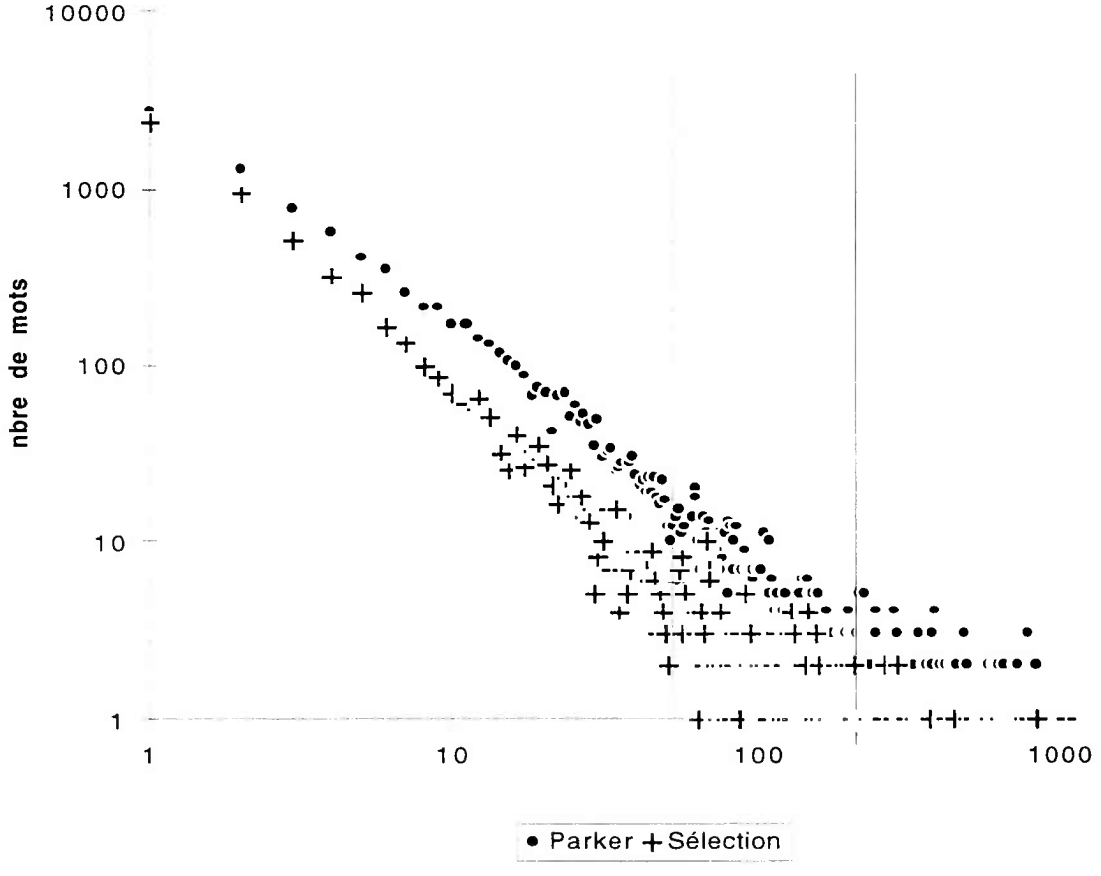
Dans le traitement des questions ouvertes, nous n'atteignons pas la première limite. En général, dans les enquêtes du CRÉDOC, les échantillons sont constitués de 1 000 à 2 000 personnes. Le corpus le plus vaste provient de l'enquête « Recherche et Environnement » à laquelle ont répondu près d'un millier de chercheurs : chacun a décrit en moyenne 7,5 problèmes prioritaires en matière d'environnement, ce qui nous amène à un corpus de 7767 textes.

Nous avions déjà rencontré la troisième limitation sur les pièces de Corneille et Racine avec une version plus ancienne d'Alceste [Beaudouin, 1994]. Suite à cette expérience et à d'autres, Max Reinert avait fortement augmenté les capacités de calcul de son logiciel : le nombre de « uns » avait doublé par rapport à la version d'alors. Mais ce n'est toujours pas suffisant.

De fait, lorsque le corpus est de grande taille, et que le maximum de « uns » est dépassé, les seuils de fréquence sont automatiquement modifiés : les mots de plus hautes fréquences sont supprimés ainsi que ceux de plus basse fréquence. En général, dans les traitements de statistique textuelle, sont conservés tous les mots de fréquence supérieure à trois. Quand les corpus sont de taille très élevée, le vocabulaire analysé et —par conséquent— les occurrences

sont réduits comme peau de chagrin. Par exemple, sur le corpus de Parker, les contraintes de taille font que seules sont conservées les fréquences comprises entre 52 et 204 (cf. Figure 2).

Cet intervalle de fréquence correspond à 7,3 % du vocabulaire et 12,6 % des occurrences. L'analyse ne porte plus que sur une très faible tranche du contenu du corpus. Notamment, des mots qui, en cas d'échantillonnage du corpus, contribuent fortement à la constitution de classes, sont écartés de l'analyse de sorte que le contenu des classes s'en trouve fortement modifié. En revanche, en traitant un échantillon de Parker (sélection d'une fiche sur quatre, corpus « Sélection »), on traite toutes les fréquences supérieures à 3 et inférieures à 3000, ce qui correspond à 35,6 % du vocabulaire et 82,4 % des occurrences.



Clef de lecture : Dans le corpus « Sélection », près de 1000 mots ont une fréquence de 2. Les deux droites verticales délimitent la gamme de fréquence, très réduite, utilisée par Alceste sur le corpus « Parker ».

Figure 2 : Courbes de Zipf (corpus Parker intégral et sélection)

Le précepte qui consiste à travailler sur des corpus de grande taille reste valable, mais pour que les analyses restent crédibles, elles doivent être conduites sur des échantillonnages de l'ordre du méga-octet.

Max Reinert nous annonce que dans la prochaine version la limite de « 1 » est passée de 90 000 à 300 000. Différents tests effectués sur les sept Mo que constituent les oeuvres complètes de Claude Simon, objet de recherche de Pascal Mougin [1995], l'ont convaincu de la nécessité d'augmenter les capacités de calcul. Dernière évolution du logiciel : le nombre de classes maximum est passé de 12 à 15.

2. CONSTRUCTION D'UNE VERSION D'ALCESTE EN ANGLAIS

Le logiciel Alceste s'est montré adapté à l'analyse des textes en français. Qu'en est-il pour les textes rédigés en anglais ? Comment établir des comparaisons entre textes français et anglais autour d'une même thématique ?

Est-il nécessaire de constituer une version anglaise aussi élaborée que la version française ? Plus précisément, est-il nécessaire de constituer des dictionnaires de mots-outils, des index de suffixes et de désinences... ? Ou bien les analyses peuvent-elles fournir des résultats comparables, au prix de raccourcis commodes comme l'élimination des mots-outils et la lemmatisation ?

2.1. FAUT-IL OU NON UN DICTIONNAIRE DE MOTS-OUTILS ?

L'objectif est d'étudier l'incidence des mots-outils sur la constitution des champs lexicaux (construits par l'analyse des cooccurrences). Nous appelons mots-outils, les mots grammaticaux, non porteurs de sens qui constituent le ciment syntaxique du texte : prépositions, pronoms, conjonctions, déterminants... Ils sont définis par opposition aux mots lexicaux, « chargés d'une fonction sémantique » [Tesnière, 1959, p. 53], qui renvoient à un référent : noms, verbes, adjectifs, adverbes.

Comment la syntaxe modifie-t-elle la constitution des classes, et à travers elles les noyaux de la représentation ?

A la suite de différentes analyses affectant différents rôles aux mots-outils, on effectue des comparaisons sur la base de :

- l'architecture des classes : proximités, agrégation... ;
- le contenu sémantique et la taille des classes ;
- la concordance entre les classements des textes et de leurs variables (mots-étoilés).

CORPUS DE FICHES DE DÉGUSTATION

Dans un premier temps, nous avons cherché à comparer les analyses sur le corpus de Parker (fiches de dégustation rédigées en anglais). Pour des raisons de taille (cf. première partie), nous n'avons pas pu travailler sur le corpus complet. Les analyses sur le sous-échantillon « Sélection » ont conduit à des résultats décevants : aucune partition ne semblait satisfaisante au-delà de trois classes, elles semblaient toutes difficiles à interpréter. Par conséquent, les comparaisons d'analyses avec et sans mots-outils étaient peu pertinentes. Nous ne sommes pas encore en mesure de savoir si la médiocrité des analyses est liée à des biais du corpus (par exemple la pratique intense du « copier—coller ») ou au contenu même de la description qui se prête mal à une partition fine sémantiquement pertinente. Nous nous sommes donc rabattus sur le corpus de fiches de dégustation en français.

Sur le corpus du *Guide Hachette*, on compare deux typologies issues de deux analyses. L'une intègre les mots-outils dans le calcul des classes (les mots-outils sont actifs), l'autre non (les mots-outils sont alors des variables illustratives). L'intégration ou non des mots-outils dans la constitution statistique des classes est la seule différence entre ces deux analyses (le mode opératoire est strictement identique dans les deux cas).

Les résultats obtenus présentent peu de différences : la constitution des classes est peu ou pas influencée par la présence des mots-outils. D'ailleurs, lorsque des mots-outils sont spécifiques d'une classe, ils sont également rattachés à cette même classe, lorsqu'ils sont considérés comme des variables illustratives. Prenons par exemple la classe, qui a été désignée par l'expression « Coeur de Bourgogne » (pour l'interprétation cf. [Brochet, à paraître]) . Elle apparaît dans les deux classifications. Dans la première, avec mots-outils actifs, les mots typiques sont : an, bourgogne, car, celui, ci, coeur, comme, corps, coup, dernier, dit, il, millésime, ne, pas, quand, que, pourtant, davantage, viticulteur, ... Dans la seconde, avec mots-outils illustratifs, les mots pleins sont : bourguignon, clos, dernier, climat, coeur, corps, coup, dire, nuit, voir, millésime, viticulteur, tempérament, ... Les mots-outils spécifiques sont : croire, dire, jamais, ne, pas, rien, sans doute, volontiers, car, comme,

davantage, celle... La classe est suffisamment robuste pour se constituer en présence ou non des mots-outils.

Dans le cas présent, le rôle joué dans l'analyse par les mots-outils n'a pas d'incidence sur les résultats. Il n'empêche que pour la cohérence de la démarche (on recherche avant tout des éléments porteurs de sens qui renvoient à un référent : autant ne pas analyser les mots vides de sens) et pour le travail d'interprétation, il est plus pertinent de ne pas donner le même statut aux mots pleins et aux mots vides.

LES PROBLÈMES PRIORITAIRES EN MATIÈRE D'ENVIRONNEMENT

Les tests sont effectués dans cette partie sur les problèmes prioritaires en matière d'environnement, tels qu'ils ont été décrits par les chercheurs interrogés. Le corpus est constitué de 4962 problèmes décrits en anglais. Trois analyses sont ici effectuées et comparées, les variations dans les paramètres concernent comme précédemment le rôle des mots-outils :

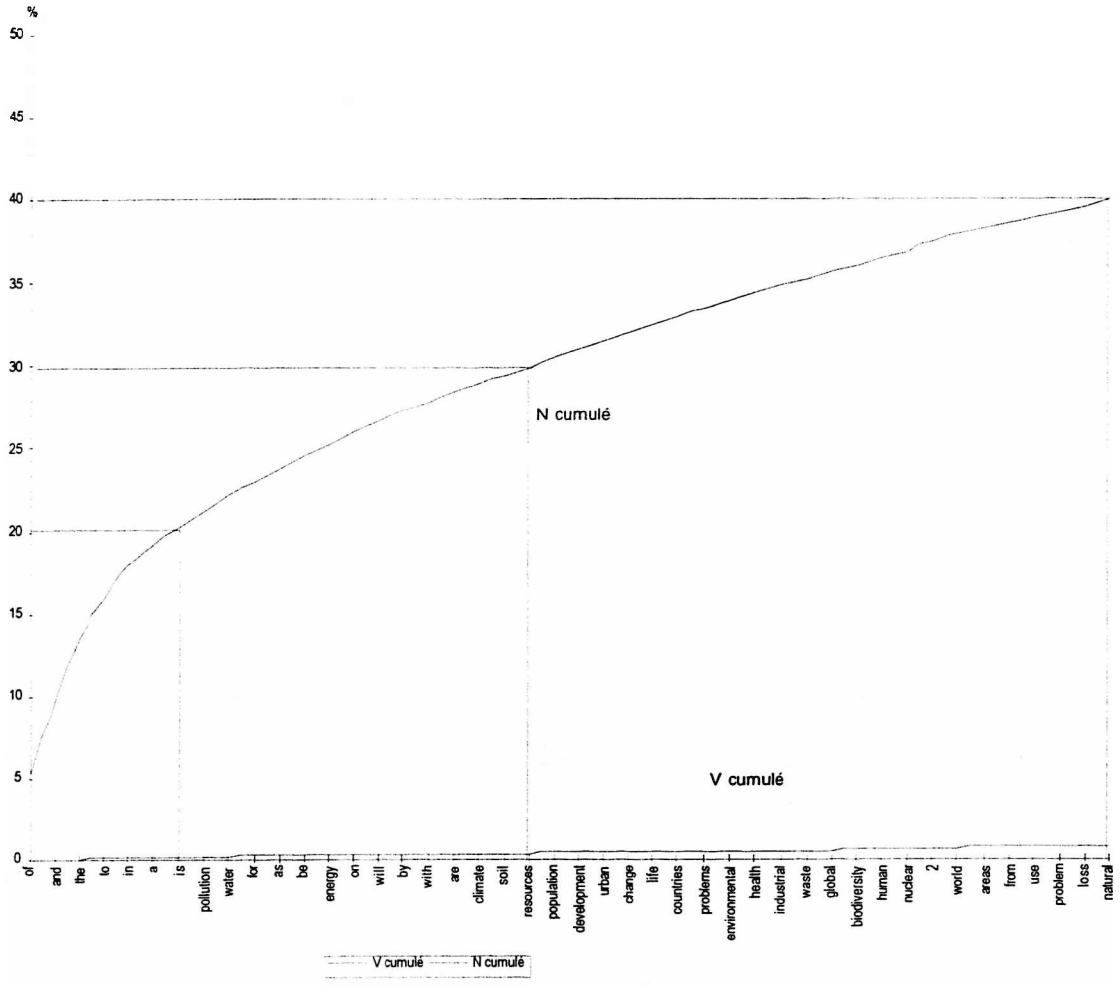
- Analyse 1 : les mots-outils sont éliminés, à l'aide d'un dictionnaire de mots-outils ;
- Analyse 2 : un seuil de fréquence supérieur est fixé pour exclure tous les mots au-dessus de cette fréquence, ceci permet d'éliminer une grande partie des mots-outils ;
- Analyse 3 : tous les mots du corpus interviennent dans l'analyse.

Il est nécessaire d'expliciter l'hypothèse qui se cache derrière la deuxième option. Le moyen le plus sûr d'éliminer les mots-outils est de disposer d'un dictionnaire de mots-outils et de rechercher dans le texte tous les mots qui y sont présents afin de les extraire. Cela suppose bien entendu la construction préalable de la liste des mots-outils. Supposons que ce dictionnaire n'existe pas. N'y-a-t-il pas un autre moyen d'extraire les mots vides ? L'étude de la répartition du vocabulaire apporte des résultats tout à fait intéressants. Zipf [1934] a montré

que le vocabulaire était organisé d'une manière systématique dans un corpus : il y a beaucoup de mots très peu fréquents, très peu de mots très fréquents ; le nombre de mots ayant une fréquence F, diminue quand F augmente selon une loi bien définie. Or une des particularités de cette courbe, très tôt mise en évidence [Encyclopedia Universalis,1968, p. 1056] est que les mots très fréquents sont en général des mots-outils.

En fixant un seuil de fréquence élevé, on a donc la possibilité d'éliminer une grande partie des mots-outils.

Dans Alceste, cela peut être fait non pas directement en fixant un seuil de fréquence, mais en éliminant un certain pourcentage des occurrences par suppression des mots les plus fréquents. La courbe des occurrences cumulées (cf. Figure 3) nous incite à penser qu'en supprimant 30% des occurrences, on élimine essentiellement des mots-outils, les exceptions étant : *pollution*, *water*, *energy*, *climate* et *soil*.



N = occurrence V = vocabulaire

Clef de lecture : les mots *of, than, the, to, in, a, is* représentent 20 % des occurrences et moins de 1 % du vocabulaire.

Figure 3 : Courbes cumulées des occurrences et du vocabulaire

Dans la pratique, les mots exclus de l'analyse avec un seuil de 30 % apparaissent dans le tableau ci-après.

Mots exclus de l'analyse	Fréquence
a	889
and	3674
are	447
as	624
be	608
by	459
climate	446
energy	601
for	670
in	1672
is	883

Mots exclus de l'analyse	Fréquence
of	4233
on	582
pollution	872
population	433
resources	437
soil	445
the	3108
to	1998
water	831
will	581
with	454

Figure 4 : Mots exclus avec seuil de fréquence supérieur

Nous pouvons donc comparer les trois analyses (à chaque fois, une typologie en douze classes a été demandée ; seules les classes d'effectif supérieur à 50 ont été conservées ce qui explique les variations dans le nombre de classes).

L'analyse sans mots-outils est incontestablement la plus satisfaisante, car elle est conforme aux hypothèses de départ : on recherche des noyaux de signification, qui renvoient à une référence. Dans ce contexte, les mots-outils ne nous intéressent pas directement, ils risquent même de brouiller la lisibilité des classes.

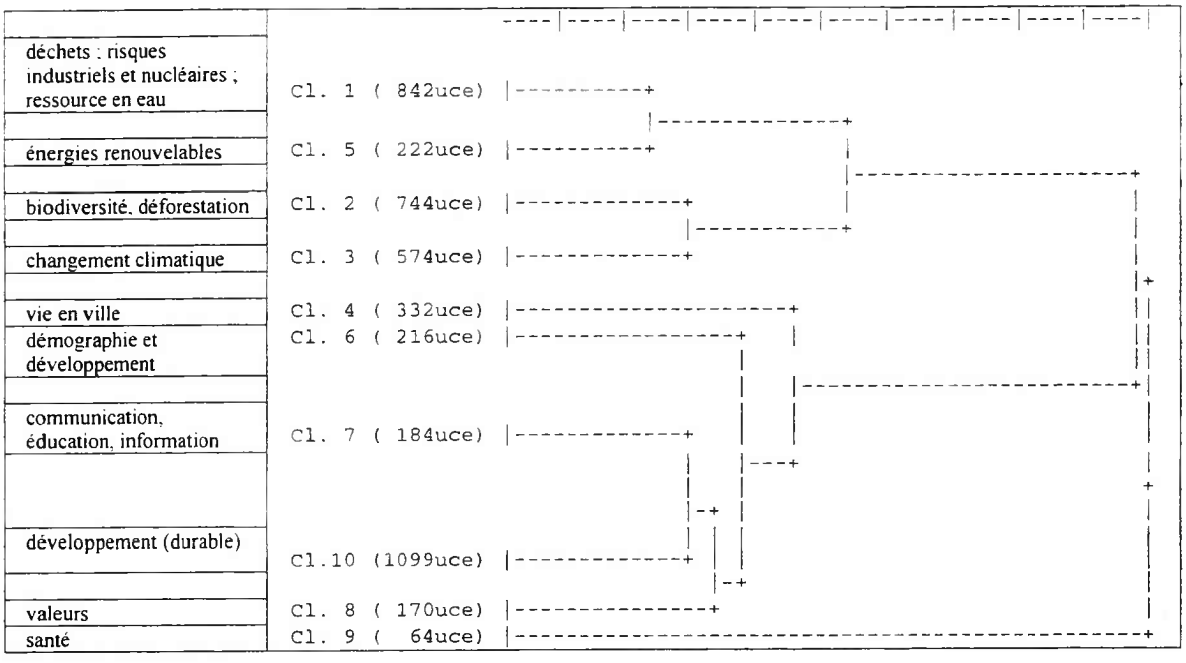


Figure 5 : Analyse 1 (sans mots-outils)

Cependant, peut-on parvenir à des résultats satisfaisants en faisant l'économie de la reconnaissance et de l'élimination des mots-outils ?

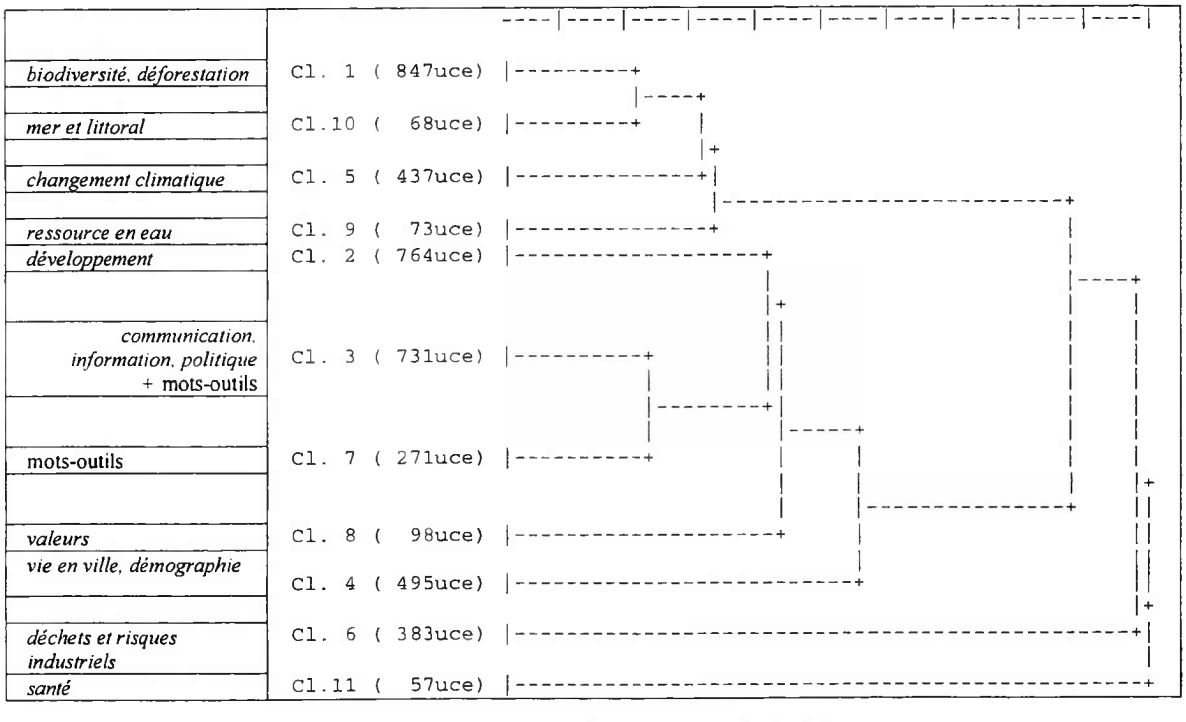


Figure 6 : Analyse 2 (avec seuil de fréquence)

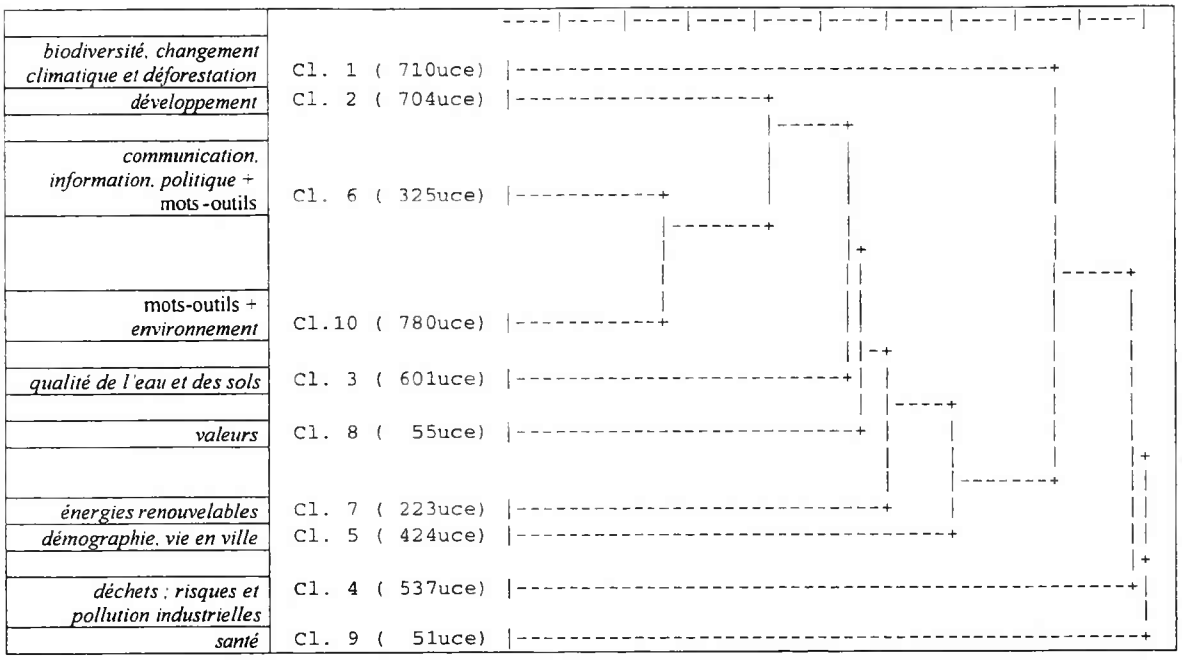


Figure 7: Analyse 3 (avec mots-outils)

Dans la colonne de gauche des figures précédentes, apparaissent des noms de classes qui résultent d'une interprétation grossière des classes obtenues (l'objectif n'étant pas ici l'étude du contenu). Dans les analyses 2 et 3, il existe deux classes qui ne se prêtent pas à la construction d'une interprétation sémantique car elles sont presque exclusivement décrites par des mots-outils. Dans la seconde analyse (en fixant un seuil de fréquence supérieur), on a certes éliminé une partie des mots-outils (les plus fréquents) mais il en reste encore une grande quantité qui ont contribué à la constitution de classes. Par ailleurs en fixant ce seuil, on a aussi éliminé une partie des mots pleins (*climate*, *energy*, *pollution*, *population*, *resources*, *soil*, *water*) qui dans d'autres analyses jouent un rôle central. On ne peut pas retenir le seuil de fréquence supérieur comme un outil acceptable pour la reconnaissance des mots-outils. On pourrait imaginer de filtrer les mots-outils par un double seuil : fréquence élevée et faible nombre de lettres (cet indice qui permet de repérer les mots courts est utilisé par SPAD.T). Pour travailler sur une langue étrangère, la meilleure option consiste à construire le dictionnaire des mots-outils.

Le fait qu'il existe des classes de mots-outils dans les analyse 2 et 3 met en évidence des différences stylistiques significatives dans la manière dont les chercheurs ont répondu à l'enquête. Celles-ci pourraient se prêter à des études spécifiques où les mots analysés seraient exclusivement les mots-outils : on pourrait repérer les différences de construction syntaxique, la position de l'énonciateur face à son énoncé...

CONCLUSION

Quand le locuteur est unique, ou quand les règles de rédaction sont très rigoureusement définies, les mots-outils se répartissent de manière homogène dans le corpus de sorte qu'ils n'ont pas d'incidence sur la constitution des classes, qui sont justement définies par la répartition hétérogène des mots. En revanche, quand les locuteurs emploient un style différent, les mots-outils ne se répartissent pas au hasard dans les unités de textes. Par exemple, une réponse rédigée en style télégraphique contiendra très peu de mots-outils, tandis qu'en style soutenu, l'articulation grammaticale sera très riche. La présence des mots-outils amène à la constitution de classes qui reflètent des différences de style plutôt que des différences de « monde ».

Lorsque l'on cherche à identifier des styles ou des modes d'énonciation, il est plus approprié de mener les analyses en conservant les mots-outils, voire en ne conservant qu'eux. Au contraire, à la recherche des noyaux de représentation dans une population, il convient de mener l'analyse sur les mots pleins uniquement.

2.2. FAUT-IL OU NON LEMMATISER ?

Si le français et l'anglais étaient, à l'instar du chinois [Tesnière, 1959, p. 54], des langues dans lesquelles la distinction entre mots pleins et mots vides était rigoureuse, nos tests se seraient arrêtés là. Mais il y a, en français comme en anglais beaucoup de mots composites qui associent des éléments pleins et des éléments vides. Ainsi *mangerons* contient un élément plein *mang-* et des éléments vides (*-er-* marque du futur et *-ons* marque de la première personne du pluriel). En français et en anglais, les éléments constitutifs de la seconde articulation du langage ne sont pas les mots mais ce que Martinet [1970, p. 16] désigne par le terme de monème. Traditionnellement, on oppose les monèmes lexicaux (ou léxèmes) aux monèmes grammaticaux (ou morphèmes), ce qui recouvre l'opposition entre éléments pleins et éléments vides.

Pour rester cohérent avec la démarche qui consiste à éliminer les mots-outils, il faut éliminer tous les monèmes grammaticaux pour que l'analyse porte exclusivement sur les léxèmes. Telle est l'ambition de la **lemmatisation** : elle consiste à remplacer par la vedette ou entrée de dictionnaire, toutes les formes qui en dérivent, toutes les variantes morphologiques. C'est l'opération que l'on fait naturellement lorsque l'on cherche un mot dans le dictionnaire (on ramène les formes conjuguées à l'infinitif, les pluriels au singulier...). Par exemple, *vont*, *va*, *iront*, *allâmes* sont ramenés à *aller*.

Depuis 1992, nous disposons au CRÉDOC des outils **Sylex,** boîte à outils linguistiques conçue par Patrick Constant [1991] et Frédéric Pigamo [1990] pour le français. Nous avions à l'origine une application SIT spécialement adaptée à nos besoins en analyse lexicale : elle produit à partir du texte brut un texte lemmatisé directement analysable par Alceste. Depuis fin 1994, nous disposons de la boîte à outils complète qui permet de développer des applications en fonction de nos besoins : en plus de l'analyse lexicale, de nombreuses fonctionnalités sont utilisées dans le cadre de l'Observatoire des Consommations Alimentaires. Cette boîte à outils comprend un **lex**ique et un analyseur **sy**ntaxique par

couches. Le lexique associe à chaque forme lexicale l'ensemble des catégories syntaxiques qu'elle peut avoir. Ainsi la forme « couvent » est définie comme pouvant potentiellement être un nom ou un verbe à la troisième personne du pluriel. Avant l'application des règles de syntaxe, on est dans une situation d'ambiguïté maximale. Les couches de règles syntaxiques successivement appliquées aux phrases permettront de lever pas à pas les ambiguïtés.

Certains logiciels, c'est le cas d'Alceste, proposent des outils de lemmatisation plus drastiques, qui visent à accéder au radical en supprimant les suffixes. Pour simplifier on parlera dans ce cas-là de **stemmatisation**. Par exemple, *rive*, *rives*, *rivage*, *rivages* seront ramenés à la racine riv+. En revanche, *rivière* et *rivières* seront réduits à *rivière+*. En fait, il y a dans cette approche une confusion entre les suffixes et les désinences morphologiques qui sont traités de la même manière. Au fil du temps, la stemmatisation d'Alceste a évolué vers une lemmatisation telle que définie précédemment grâce à l'introduction de dictionnaires.

Pour le moment, Alceste ne dispose pas de fonction de lemmatisation en anglais. Nous avons donc utilisé un module de lemmatisation du premier type (intégrant la forme syntaxique dans la réduction). Comme nous ne disposons pas encore de la version de **Sylex** pour l'anglais, Patrick Constant a eu l'amabilité d'effectuer pour nous le traitement de lemmatisation sur deux corpus anglais :

- les réponses à la question des problèmes prioritaires dans le cadre de l'enquête « Recherche et Environnement » ; seules les réponses parvenues avant la relance de l'enquête ont été lemmatisées, la comparaison sera donc faite sur un corpus incomplet d'un tiers par rapport à la comparaison des analyses avec ou sans mots-outils. Comme nos préoccupations sont plutôt méthodologiques, on se n'attardera pas non plus sur la variation des résultats ;
- les fiches de dégustation de vin de Parker (corpus « Sélection »).

Nous avons comparé les résultats d'analyses sur les corpus bruts et lemmatisés, et cherché à évaluer l'incidence de la lemmatisation. Rappelons que la lemmatisation de Sylex est légère, syntaxiquement cohérente : pas de réduction de mots appartenant à des catégories différentes, en tout cas elle est plus légère que la lemmatisation d'Alceste qui parfois encore regroupe des mots appartenant à des catégories différentes.

Précédemment, nous avions montré sur **un** corpus que le type de lemmatisation (par Alceste ou Langage Naturel) n'avait pas une incidence forte sur les typologies obtenues [Beaudouin et Lahlou, 1993, p. 43-44]. Les seules variations concernaient les effectifs des classes : l'écart maximum n'était cependant que de trois points. Nous avons ici l'occasion de montrer l'incidence de la lemmatisation Sylex par rapport à aucune lemmatisation.

LES PROBLÈMES D'ENVIRONNEMENT

Les indicateurs statistiques sur la taille du vocabulaire (V) et le nombre d'occurrences (N) (cf. Figure 8), pour reprendre les définitions de Charles Muller [1977], nous montrent l'ampleur des variations apportées par la lemmatisation et leurs effets sur le tableau analysé par Alceste.

Vocabulaire	Corpus brut		Corpus lemmatisé	
	V	N	V	N
Mots pleins	5025	38663	3914	39066
Mots analysés (fréq>3)	1239	32735	1288	35281
Pourcentage	24,6	84,6	32,8	90,3

Figure 8 : Tableau du vocabulaire

En gros, la lemmatisation permet de travailler sur une partie plus ample du vocabulaire et des occurrences, et donc de limiter les « pertes ». Ceci confirme un résultat général déjà présenté dans [Beaudouin, 1993].

Les classifications obtenues sont exactement les mêmes que l'on garde quatre classes terminales ou onze. Encore une fois, seule varie la taille de chacune des classes.

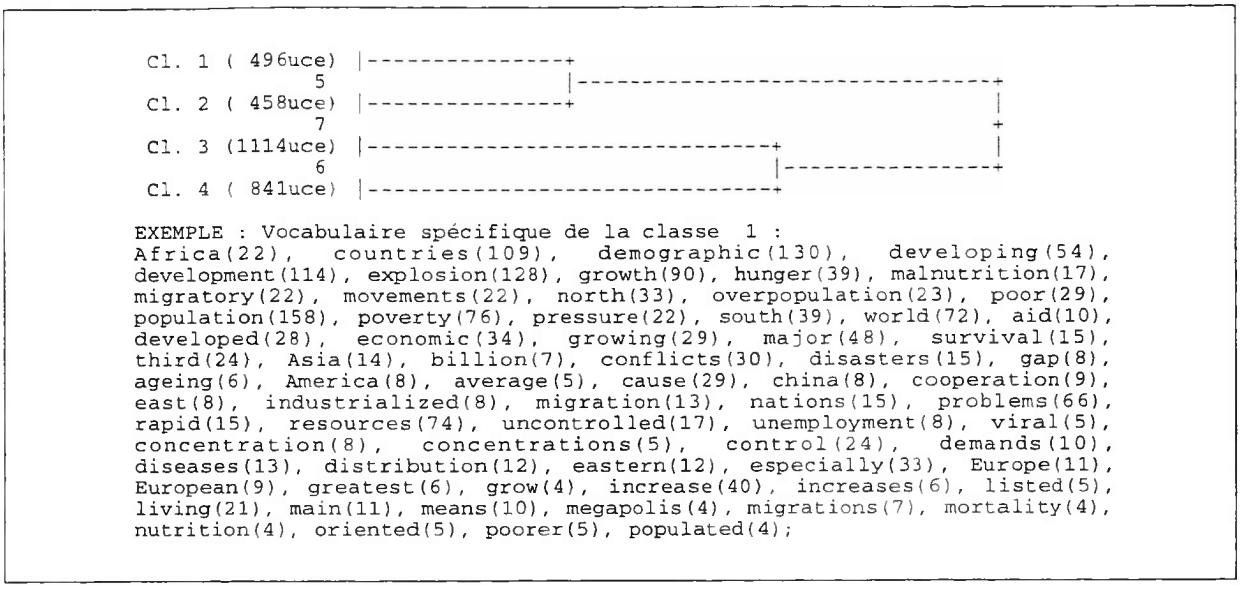


Figure 9 : Analyse sans lemmatisation

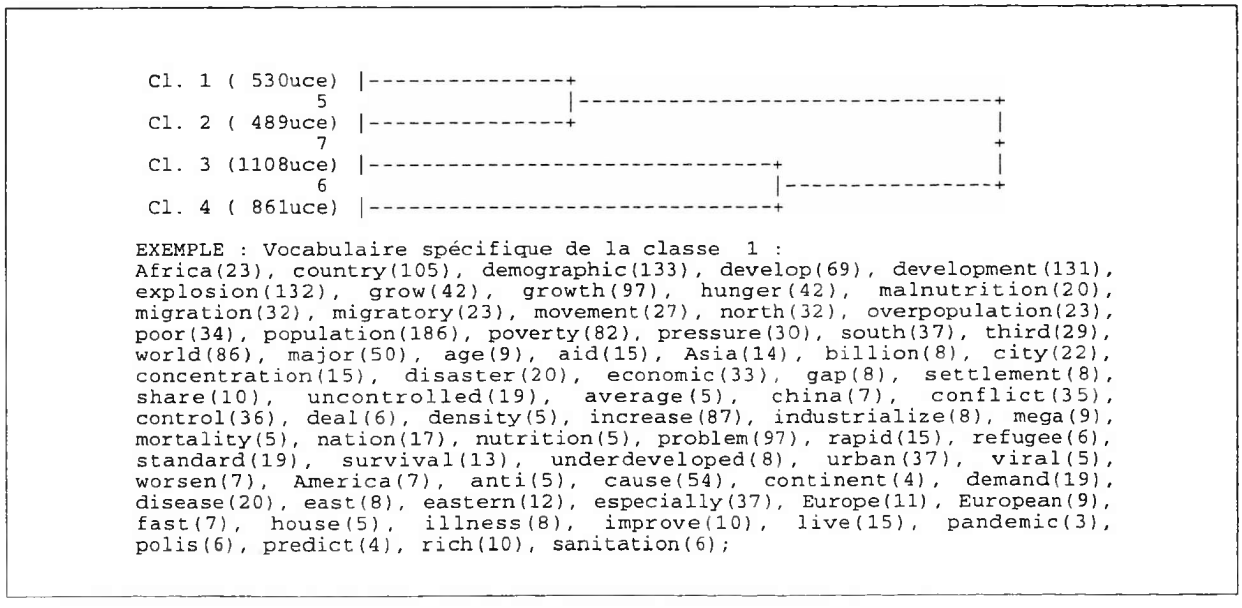


Figure 10 : Analyse avec lemmatisation

Les problèmes sont grosso modo classés de la même manière par les deux analyses comme le montre la Figure 11 : les déviations par rapport à une équivalence parfaite entre les deux typologies ne dépassent pas les 15 points. Dans le pire des cas, on a quant même 83,5 % des problèmes rgroupés dans une classe par l'analyse sur texte lemmatisé qui se retrouvent dans la classe correspondante avec l'anlyse sur texte brut.

		Analyse sur texte lemmatisé					
		Non classés	classe 1	classe 2	classe 3	classe 4	Total
Analyse sur texte brut	Non classés	35,8 55,3	16,0 4,9	12,5 4,3	16,0 3,5	19,8 4,9	232 6,5
	classe 1	2,3 11,3	**87,3** **85,4**	5,6 6,1	2,0 1,4	2,9 2,2	738 20,6
	classe 2	3,2 14,0	6,6 5,7	**86,8** **83,5**	1,5 1,0	1,8 1,3	651 18,2
	classe 3	1,6 11,3	1,5 2,1	1,2 1,9	**93,8** **93,4**	1,8 2,0	1049 29,3
	classe 4	1,3 8,0	1,5 1,9	3,2 4,3	0,9 0,8	**93,1** **89,7**	912 25,5
	Total	150 4,2	754 21,1	677 18,9	1054 29,4	947 26,4	3582

Clef de lecture : 85.4 % des problèmes classés en classe 1 avec analyse sur texte lemmatisé sont classés en 1 avec l'analyse sur texte brut.

Figure 11 : Croisement des typologies avec et sans lemmatisation

Au niveau du vocabulaire, les quatre classes sont pratiquement les mêmes. La seule différence est que les mots au pluriel figurent au singulier dans la seconde analyse. On est dans un contexte particulier où la lemmatisation n'a pas une incidence forte dans la mesure où les verbes conjugués sont assez peu nombreux et les formes peu variées. Beaucoup de réponses se présentent sous la forme de groupes nominaux.

LES FICHES DE DÉGUSTATION DE VINS

Pour les fiches de dégustation de vin, les tests ne sont pas encore achevés. Un premier test a été effectué sur le corpus français Hachette. Sur les fiches de dégustation Hachette, deux analyses ont été effectuées : l'une sur les formes brutes et l'autre sur les formes lemmatisées (avec Alceste). Dans les deux cas, les mots-outils ont été exclus.

Les principales classes obtenues sont les suivantes : ROUGE (Rouge, Bordeaux), BLANC (Blanc, Lignée), et (Temps, Coeur de Bourgogne).

Les arbres de la classification sont rigoureusement identiques, même si les tailles des classes sont différentes.

Fiches	Sans lemmatisation	Avec lemmatisation
Temps, Coeur de Bourgogne	32,9 %	38,8 %
Rouge	45,4 %	24,6 %
Blanc (blanc)	12,6 %	18,9 %
Blanc (lignée)	9,1 %	17,7 %
ENSEMBLE	100,0 %	100,0 %

Figure 12 : Répartition des fiches(en pourcentage)

A nouveau, les conditions d'énonciation sont cruciales. En français, sur le corpus Hachette, toutes les formes d'un lemme sont regroupées dans la même classe et peu de mots significatifs sont écartés.

Un autre test est en cours sur le corpus anglais. Les résultats sont déconcertants car de nombreuses formes d'un même mot se trouvent dans des classes différentes rendant l'interprétation desdites classes problématique. On s'interroge cependant sur les particularités de ce corpus. Après l'élimination des « copiés-collés », de nouveaux tests seront effectués.

CONCLUSION

La lemmatisation est-elle donc une opération superflue ? Non, dans la mesure où la démarche est cohérente avec celle qui consiste à supprimer les mots-outils. Mais ces différents tests nous ont permis de voir que dans une situation d'urgence (l'analyse rapide de textes dans une langue étrangère), on pouvait éventuellement se passer de cette étape. En effet, on obtient globalement les mêmes résultats avec et sans lemmatisation : même arbre de classification, les mots typiques des classes sont grossièrement les mêmes et les problèmes sont à peu près classés de la même manière...

Finalement tout le débat sur la lemmatisation qui a fait couler beaucoup d'encre perd d'un coup de sa pertinence dans le cas de réponses rédigées en anglais. Sur deux corpus de nature très différente tous les deux en anglais, on parvient grosso modo aux mêmes résultats. Sans doute trouverait-on des corpus où ce n'est pas le cas.

En tout état de cause, il paraît pertinent pour chaque corpus à analyser de faire des tests avec et sans lemmatisation. C'est sans doute un moyen assez efficace pour tester la stabilité des typologies obtenues.

On admettra cependant que pour l'analyse du sens, le travail sur le texte lemmatisé est plus satisfaisant —ne serait-ce que pour l'interprétation des classes—. Elle est facilitée du fait du nombre plus faible de mots typiques.

2.3. ÉTAT D'AVANCEMENT DE LA NOUVELLE VERSION DU LOGICIEL ALCESTE EN ANGLAIS

Les différents tests comparatifs effectués ont permis de montrer que dans les problématiques qui nous animent (recherche de la référence au monde plutôt qu'analyse de la position de l'énonciateur face à son énoncé), il est fortement conseillé d'avoir une liste des mots-outils afin de pouvoir les exclure de l'analyse.

Nous avons donc construit un dictionnaire de mots-outils anglais. Pour ce faire nous avons utilisé le travail fait par Patrick Constant sur un échantillon du corpus Parker. Il avait fourni le corpus lemmatisé et indexé par catégories syntaxiques :

> Château Montelena's wines possess gorgeous potential when young, and are even better when revisited later. The 1984 Cabernet Sauvignon exhibits the forward, jammy, cassis, and other black fruit character of the vintage, marvelously rich, full-bodied, concentrated flavors, high extract, gobs of glycerin, and moderate tannin in the finish. Its virtues - purity, richness, and opulence - along with a firm underpinning of tannin, should serve it well for another 10-15 years.$

est ainsi devenu :

> chateau_N Montelena's_I wine_N possess_V gorgeous_A potential_N when_G young_A,- and_G be_G even_G better when_G revisit_A late_A.- the_G 1984 Cabernet_I Sauvignon_I exhibit_V the forward_N,- jammy_I,- cassis_I,- and_G other_G black_A fruit_N character_A of_G the_G vintage_A,- marvelously rich_A,- full-ody_A,- concentrate_A flavor_N,- high extract_N,- gob_N of_G glycerin_N,- and_G moderate_A tannin_N in the finish_N.- its virtue_N -- purity_N,- richness_N,- and_G opulence_N -- along_G with_G a_G firm_A underpinning_N of_G tannin_N,- shall_V serve_V it_G well_V for_G another_N 10-15 year_N.-$

(De nombreux points étranges apparaissent dans cette lemmatisation qui date d'il y a un an ; dans la dernière version utilisée pour les tests comparatifs, la qualité s'est sensiblement améliorée). Il a donc suffi d'extraire tous les mots suffixés en _G pour avoir déjà une première liste de mots-outils. Celle-ci a été enrichie à la main en examinant les dictionnaires complets des différents corpus en anglais.

Par ailleurs, nous avons partiellement résolu le problème des formes contractées (exemples : don't, I'll...) en les introduisant à la fois dans le dictionnaire des locutions et dans celui des mots-outils.

La liste des verbes irréguliers a été saisie. Chaque fois que le logiciel rencontrera une forme dérivée, il la rattachera à la forme infinitive : *am*, *is* et *are* seront rattachés à *be*. Une partie d'entre eux ont de plus été classés parmi les mots-outils. C'est le cas de *to be*, *to have*, *can*, *must*...). Ce dictionnaire a été testé sur nos corpus et se révèle satisfaisant, sauf quelques cas d'homonymie comme *left* (gauche) et *left* (dérivé de *to leave*). Des tests sur d'autres corpus sont bien entendu indispensables pour le compléter. L'avancement de cette version est soumis à la bonne collaboration entre les utilisateurs de la version anglaise.

Même si nous n'avons pas établi la nécessité impérieuse de la lemmatisation, une liste de suffixes a été constituée en vue de celle-ci. Sa qualité n'a pas été testée.

Ces différents dictionnaires ont été intégrés par Max Reinert dans une version spéciale d'Alceste. Il a également amélioré certaines de nos options. Un seul point est encore insatisfaisant : la gestion des apostrophes, mais devrait être résolu très prochainement.

3. COMPARAISON TEXTE BRUT / TEXTE INDEXÉ

Nous venons d'envisager la « voie lente » avec l'élaboration d'une version en anglais d'Alceste. Nous en venons maintenant à la « voie rapide », celle qui a été adoptée pour la production rapide de résultats et qui a consisté à indexer les textes par mots-clefs, à traduire partiellement ces derniers et à effectuer des analyses sur les réponses réduites à leurs mots-clefs.

Dans cette partie, la comparaison porte sur les réponses en français à la question ouverte sur les problèmes prioritaires en matière d'environnement. L'objectif est de comparer l'analyse lexicale sur les réponses brutes et sur les réponses indexées par mots-clefs. Cette comparaison n'a pas été effectuée sur les textes anglais puisque l'objectif était de valider l'approche par mots-clefs, à une époque où la version en anglais d'Alceste n'était pas même ébauchée.

Pourquoi ? L'enquête « Recherche et Environnement » a été réalisée auprès de chercheurs du monde entier. Elle est principalement constituée de questions ouvertes, ce qui marque un refus d'enfermer a priori les chercheurs dans un cadre de réflexion trop étroit. Ce choix paraît judicieux à la lumière des observations que les chercheurs ont fait sur les questionnaires (cf. Annexe 2 sur les observations) : les seules critiques de fond portent sur la partie des « scénarios », essentiellement constituée de questions fermées.

Etant donnée la liberté de parole (en fait d'écriture) accordée aux chercheurs, il fallait des méthodes adaptées pour analyser cette parole dans toute sa richesse. On a donc refusé d'emblée l'idée d'un postcodage simple, de même que l'on avait refusé l'idée de poser des questions fermées. La méthode d'analyse devait être cohérente avec les choix faits lors de la conception du questionnaire.

Mais pour la première fois dans une étude du CRÉDOC, on est confronté au problème de la langue. En effet, les questionnaires ont été rédigés soit en anglais, soit en français. Alors que pour le français, nous disposons de tous les outils nécessaires pour l'analyse des textes, tel n'était pas le cas pour l'anglais. Cela nous a amené à tenter de mettre au point une version en anglais d'Alceste (cf. Partie I)

Mais l'existence d'une version anglaise ne résout pas tous les problèmes. Comment analyser simultanément les réponses des anglophones et celles des francophones ?

La solution la plus naturelle aurait été de faire traduire toutes les réponses de l'anglais vers le français. Par des traducteurs professionnels, le budget aurait été d'environ 60 000 Francs, coût relativement élevé. Il aurait également été possible d'utiliser des outils de traduction automatique. N'ayant pu évaluer à temps la qualité de ces traducteurs automatiques, nous ne nous y sommes pas intéressés tout de suite. Outre les aspects financiers et techniques, on rencontre ici encore « les problèmes théoriques de la traduction » [Mounin, 1963] : la représentation de l'environnement n'est sans doute pas le même en français et en anglais. Les problèmes sont en partie différents, ainsi que leur organisation et la manière dont ils sont décrits. L'option a été prise de reporter le plus tard possible la phase de la traduction.

Par ailleurs, l'enquête « Recherche et Environnement » a été conçue pour pouvoir être réexploitée par d'autres laboratoires ou centres de recherche sur l'environnement. Il fallait donc pouvoir présenter les résultats sous une forme plus homogène et normée.

La solution adoptée a consisté à indexer par mots-clefs les réponses. Au maximum six mots-clefs pouvaient être attribués à chaque réponse. Les mots-clefs sont des mots ou des expressions comme on peut le voir dans la Figure 13 :

Texte brut	Texte indexé par mots-clefs
Le problème de la population globale, dont l'accroissement non contrôlé, empêche le développement de bien des pays, épuise des ressources non renouvelables, etc.	accroissement_démographique, épuisement_ressources, développement_économique
Ces pollutions diffuses sont et seront le problème majeur, car elles entraînent la pollution des sols et des problèmes difficilement maîtrisables du climat.	pollution_sol, changement_climatique, pollution_diffuse
Conséquence directe, une urbanisation anarchique avec des mégalopoles où la vie est pénible (transports, pollution, délinquance, ghettos et bidonville, ...)	maîtrise_urbanisme, mégalopoles, cadre_de_vie, transport, pollution, insécurité
Menaces sur la biodiversité : perte (irréversible) d'espèces liée à la surpopulation et à la mauvaise exploitation des ressources	appauvrissement_bio_diversité, disparition_des_espèces, surpopulation, gestion_ressources, irréversibilité.

Figure 13 : Exemples de textes bruts et de textes indexés par mots-clefs

L'indexation des textes s'est faite dans la langue du questionnaire : mots-clefs français pour les questionnaires en français, anglais pour les questionnaires en anglais. Il existe donc une nomenclature de 307 mots-clefs en français et de 305 mots-clefs en anglais (Cf. Annexe 1).

Pour travailler simultanément dans les deux langues, on a cherché à apparier les nomenclatures (Cf. Annexe 1). La correspondance entre mots-clefs français et anglais était parfois évidente (water/eau ; waste/déchets...), parfois moins et quelquefois inexistante. En dépit d'un effort d'harmonisation des nomenclatures au cours de l'indexation, il n'y a que 169 mots-clefs communs aux deux. Ceci justifie en partie le fait de n'avoir pas eu recours à la traduction : l'indexation montre déjà que le découpage des problèmes d'environnement n'est pas le même en anglais et en français. Prenons un exemple assez « simple », le « nucléaire », car il n'y a qu'un terme pour désigner ce type d'énergie. La Figure 14 montre que les items utilisés pour l'indexation ne coïncident pas ; globalement la manière d'aborder le nucléaire n'a pas été la même.

En français	Fréq
déchets nucléaires	41
risques nucléaires	40
énergie nucléaire	17
nucléaire	11
pollution nucléaire	9
centrale nucléaire	5
armes nucléaires	4
catastrophe nucléaire	1

En anglais	Fréq
nuclear wastes	82
nuclear-related risks	67
nuclear energy	42
nuclear powerplants	22
nuclear weapons testing	17
nuclear material transport	1

Figure 14 : Deux langues, deux manières d'indexer

Le corpus final analysé est constitué des réponses des anglophones indexées par les mots-clefs anglais et les réponses des francophones indexées par les mots-clefs traduits en anglais (quand il existait une correspondance) ou par des mots-clefs en français quand ils n'avaient pas d'équivalents en anglais. Pour l'analyse détaillée des réponses, le lecteur pourra se

reporter à l'article présenté au colloque de Fontevraud [Laredo, Volatier et Collerie de Borely, 1996].

Pour une analyse globale de l'enquête, l'option de la traduction n'ayant pas été adoptée, on est obligé de travailler sur les corpus indexés par mots-clefs au lieu de travailler sur le texte brut. Peut-on éviter cet effort supplémentaire ?

Nous faisons les hypothèses suivantes :

- l'indexation par mots-clefs permet d'harmoniser les réponses en évacuant d'emblée les différences de styles (suppression de tout l'appareillage syntaxique) et en procédant à une pré-interprétation des réponses ;
- l'indexation par mots-clefs appauvrit beaucoup les réponses en les réduisant à leur plus simple expression.

On a donc deux tendances contraires : l'une tend à rendre les réponses plus homogènes et par conséquent la typologie plus stable ; l'autre à appauvrir les réponses ce qui devrait déstabiliser la typologie. Les deux effets se compensent-ils ?

Pour donner des éléments de réponse à cette question, nous comparons deux typologies. L'une est obtenue sur le corpus des réponses brutes en français, l'autre sur le corpus indexé par mots-clefs en français. La typologie des réponses brutes sera dorénavant désignée par TBFR, la typologie sur les réponses indexées par TIFR.

	Texte brut (TBFR)	Texte indexé (TIFR)
Population	ensemble des problèmes (2810)	
Individu	problème en texte libre	problème indexé
Variables	mots pleins	mots-clefs

Figure 15 : Bases de la comparaison

L'individu statistique est ici le problème et les variables sont tantôt les mots pleins, tantôt les mots clefs.

Nous cherchons à voir s'il y a une relation de bijectivité entre TBFR et TIFR. Dans l'idéal, cela implique qu'on obtienne la même typologie et que les réponses brutes ou indexées soient classées de la même manière. Si c'est le cas, on prouverait qu'il est suffisant de travailler sur TBFR pour mettre à jour la représentation des problèmes d'environnement. Une identité parfaite est à exclure d'emblée : quel est le seuil de similitude qui nous paraît acceptable ? Peut-il être défini statistiquement ?

Pour comparer deux typologies, nous avons une gamme assez étendue de moyens. Doivent être comparés : la structure globale des classes (l'arbre de la classification), la nature et le contenu des classes, leur taille et le croisement statistique des deux typologies.

Si les écarts ne sont pas trop grands, l'analyse lexicale sur les mots-clefs peut être considérée comme équivalente à l'analyse sur texte brut.

On retrouve dans les deux traitements, trois grands axes communs : les problèmes d'environnement liés à l'action de l'homme sur la planète, à celle de l'homme sur la société et enfin les problèmes liés à l'énergie et aux déchets (Figure 16). On a donc une opposition entre un point de vue global et planétaire, « l'écosphère » et un point de vue social, la « sociosphère ». La seule variation est le rattachement différent du troisième pôle. Pour le texte brut, il est rattaché à la « sociosphère » et pour le corpus indexé à « l'écosphère ». Cela s'explique dans la mesure où les déchets et les problèmes d'énergie ont autant d'incidence sur la planète (décharges, pollution atmosphérique) que sur l'homme (propreté-hygiène, pollution atmosphérique), d'où leur position instable, ce que nous ont également prouvé d'autres tests.

La Figure 16 montre comment les deux typologies peuvent être mises en correspondance : en dehors des deux points que nous allons examiner, les correspondances semblent satisfaisantes.

Texte brut		Texte indexé
Changement climatique Ressources en eau Mer et littoral Désertification, déforestation Biodiversité	L'homme sur la nature	Changement climatique Ressources en eau Mer et littoral Désertification, déforestation Biodiversité Désertification des campagnes
Démographie, développement Répartition inégale Vie en ville Valeurs Santé	L'homme sur la société	Démographie, développement Vie en ville Valeurs Santé
Energie, transport Déchets, nucléaire	Energie, déchets	Energie, transport Déchets, nucléaire

Figure 16 : Typologies des problèmes d'environnement

Nous retrouvons globalement les mêmes classes. Revenons tout d'abord sur la démarche d'interprétation qui nous permet de conclure que deux classes sont identiques. Une classe est caractérisée par ses mots spécifiques. Pour dire que deux classes se ressemblent ou sont identiques (il faudrait mieux définir ce qu'est la similarité), un travail d'interprétation sémantique est indispensable.

Dans les analyses sur TBFR et TIFR, on obtient à chaque fois une classe qui peut être résumée par le concept global de « changement climatique ». Cette interprétation s'est faite à la suite de l'examen des mots typiques.

Mots typiques sur TBFR de la classe « changement climatique »

climatique+(72), globa+l(23), changement+(71), climat+(68), couche+(20), effet+(18), gaz(12), modification+(16), trou+(7), réchauff+er(14), co2(19), effet_de_serre(65), ozone(45), conséqu+ent(26), anthropique+(7), stratosphérique+(4), serr+er(3), évoluti+f(11), océanique+(3), incertitude+(3), perturb<(4), act+ion(16), inconnu+(2), augmentation+(10), réalité+(3), régime+(2), taux(3), température+(2), attendre.(2), éventu+el(4), proba+ble(4), carbone+(2), méthane(2), moyen_terme(3), affecte+(2), dramatique+(2), élévation+(2), programme+(3), réaction+(2), sécheresse+(2), asséch+er(2), détermin+er(2), influenc+er(3), souhait<(2), atmosphériques(2), o2(2), régulation+(2), végétation+(2), biomasse(2), indirect+(2), planétaire+(4), méconnaissance+(1), énumér+er(1), inadéquat+(1).

Mots typiques sur TIFR de la classe « changement climatique »

changement_climatiq(62), changement_global(12), climat(37), co2(26), couche_ozone(47), effet_de_serre(67), évaluation_scientif(15), gaz_à_effet_de_serr(6), réchauffement(15), niveau_des_mers(5), pluviosité(3), rejet_des_polluants(5), acidification_de_l_(2);

Sur TIFR, le nombre de mots spécifiques de la classe est beaucoup plus élevé que sur TBFR (19 contre 76). D'autre part, les mots-clefs sont plus homogènes par construction. L'interprétation est donc plus facile à construire sur TIFR, mais elle est aussi moins fine, car une partie des nuances a disparu dans l'indexation. Par exemple, le fait que les chercheurs font part de quelques incertitudes sur les mécanismes à l'origine d'un éventuel changement climatique (*incertitude, inconnu, éventuel, probable*) n'a pas été pris en considération lors de l'indexation. Le passage du texte brut aux mots-clefs entraîne la disparition de certaines nuances, mais permet un accès plus direct au coeur de la problématique du « changement climatique ».

Prenons à titre d'exemple, le terme *effet de serre.* A priori, toutes les occurrences (près d'une centaine) devraient être classées dans la classe « changement climatique ». Sur TBFR, 11 réponses contenant le terme *effet de serre* ont été classées ailleurs ; sur TIFR, 5 réponses. Le tableau ci-dessous présente les réponses contenant *effet de serre* non classées dans « changement climatique ». Globalement, le fait qu'elles soient classées dans une autre classe est assez cohérent, et les disparités entre TIFR et TBFR ne sont pas très grandes.

Réponses contenant le mot-clef *effet de serre* non classées dans la classe « changement climatique »	
(TIFR)	(TBFR)
	Energie, transport énergie: passage à des énergies moins polluantes, énergies renouvelables. problème du co2, effet-de-serre.
Energie, transport #transport_routier, #énergie, effet-de-serre, #économie_énergie, #choix_rail_route, #pétrole.	**Energie, transport** #transports #routiers, #énergie et effet-de-serre: les #prix bas du #pétrole et du dollar conduisent à un relâchement des économies d'#énergie. le développement du #transport routier au détriment du #rail ne peut qu'accroître l'effet-de-serre et rendre impossible le respect de la convention de #Rio.
	Désertification, déforestation la conquête de nouvelles terres sur les savanes pour le développement d'une agriculture minière conduit à une #rapide minéralisation de la matière organique des sols, le CO2 produit participe à l'effet-de-serre. il en résulte un effondrement de la structure des sols qui #provoque de #fortes augmentations du ruissellement et l' #érosion. le pedoclimax se sahélise, les rendements des #cultures #chutent.
Energie, transport gaz_à_effet-de-serre, #énergie_non_polluante, #transport,	**Energie, transport** #réduire les #émissions de gaz #polluants ou à effet-de-serre: #choix des #sources d' #énergie plus appropriées, lien avec la priorité 2, #revoir les #modes de #transports, trouver des #alternatives aux #transports #automobiles ou par #camions.
Energie, transport #énergie_non_polluante, gaz_à_effet-de-serre, #transport,	**Energie, transport** #énergie: recherche de #sources #énergetiques à faible #potentiel de pollution: faible #émission de gaz à effet-de-serre, pas de produits de combustion ou de transformation #polluants, pas de radiation, d' #émissions de #particules, d'ions. développement des filières poles à combustion à CH_4 ou H_2 par-exemple, lié aux problèmes de #transports.
Désertification, déforestation #pénurie_en_eau, #consommation_d_eau, accroissement_démographique, déforestation, désertification, effet-de-serre.	**Répartition inégale** #raréfaction des #ressources en #eau, à-cause de la consommation croissante, industries, urbanisation, style de vie, et à-cause de la croissance démographique, la déforestation, la désertification et l'effet-de-serre pourraient produire des disettes régionales.
	Répartition inégale les problèmes d'effet-de-serre, co2. liés au développement des ressources de l'homme auront entre autre pour effet, et on en #parle bien rarement, de modifier les équilibres #économiques agricoles, c'-est-à-dire à une #remontée de trois cent kilomètres vers le nord des #productions. les #différents pays #concernés devraient en tenir compte des maintenant.

Réponses contenant le mot-clef *effet de serre* non classées dans la classe « changement climatique »	
(TIFR)	(TBFR)
	Répartition inégale démographie et risques naturels et industriels. les grands risques potentiels que #font courir les activités humaines, de la reproduction non contrôlée à la #production effrénée sur des filières technologiques peu sûres: #catastrophes industrielles, nucléaires, marée noire, urbanisation, #densité de #population et tremblement de #terre ou éruption volcanique ou grandes inondations. ce-que l' on peut appeler les #facteurs synergétiques ou multiplicatifs des #catastrophes naturelles et industrielles et l'amplification des conséquences, déforestation, désertification, pollution des nappes phréatiques. je veux distinguer ici l'évolution lente, effet-de-serre, ozone. des conséquences brutales qui interviendraient à-cause de #facteurs cumulatifs de risques. en principe, une politique préventive peut éviter les #catastrophes et surtout leur #extension si l'on tient compte des #phénomènes de synergie de risques cumules, risques en chaîne d'une #guerre bactériologique, chimique.
	Energie, transport il est important de #réduire de-manière drastique la pollution de l'air due aux #émissions de CO_2 et de nox des #automobiles. #favoriser les autres moyens de #transports et réglementation stricte. ceci est lié à l'évolution du climat et à l' effet-de-serre. doivent s'y ajouter la maîtrise des pollutions industrielles.
Désertification, déforestation pollution_air, maladie, #effet-de-serre, pluies_acides, déforestation, acidification_sols.	**Ressources en eau** #pollution de l'#air: gaz polluants, cov, poussières, maladies respiratoires et de peaux. effet-de-serre, surélévation des #eaux, nouvelles maladies, déplacements de population, cf. 7, effet de retard. pluies-acides, déforestation, cf. 4 et 6, #acidification des #sols, cf. 3, #pollution de l'#eau, cf. 2.
	Valeurs climat: effet-de-serre, ozone. #classe en-tête non pas à-cause de son acuité #actuelle, mais parce-que rien ne #sert de résoudre les autres problèmes si ce problème cadre a #long-terme n'est pas #pris énergiquement en #compte. on peut d'-ailleurs, à l'#heure #actuelle, émettre l'avis que les actions #clefs, bioénergie, ferroutage, #restent trop timides.

Pour cette classe, il semble y avoir une bonne correspondance entre les deux méthodes.

En ce qui concerne les problèmes de développement et de démographie, sur TBFR on obtient deux classes, l'une centrée sur l'accroissement démographique et le développement dans le Tiers Monde, l'autre sur les inégalités économiques au niveau mondial. Cette distinction n'apparaît pas sur TIFR (cf. Figure 17). Cette distinction témoigne davantage d'une différence de vocabulaire utilisé que d'idées distinctes. On regroupe donc dans la suite les deux classes de TBFR pour pouvoir établir des comparaisons.

sur TBFR	sur TIFR
nord+(53), Afrique(21), Asie(7), conflit+(22), déséquilibre+(18), développement+(97), écart+(8), faim+(22), famine+(13), flux(29), pays(88), relation+(28), sud+(59), migratoires(24), tiers_monde(22), démograph<(129), explos+ion(89), pauvre+(73), riche+(30), accroissement+(18), intensifi+er(12), mégalopoles(16), pathogène+(5), sous-developpe+(7), aid+er(4), environnementaux(6), migrat+ion(7), coopérati+f(4), milliard+(4), misère+(4), amen+er(4), creus+er(3), engendr+er(10), surpopulation(10), survie(7), problem<(68), déstabilis+(4), expansi+f(4), aide+(4), course+(4), place+(4), tension+(4), vision+(3), détérior+er(4), gouvern<(2), grand+ir(3), poursuivre.(2), repartir.(2), résoudre.(6), risqu+er(5), chom+...(4), impossi+ble(6), religi<(4), aigu+(2), cohérent+(2), immense+(2), Amérique<(2), commerce+(3), démun+ir(2), développ+er(13), nourr+ir(2), pvd(6), intérieur+(2), puissance+(2), vue+(3), about+ir(2), concentr+er(2), craindre.(2), désastr+e(2), haut+(3);	accroissement_démog(60), afrique(17), conflits(17), coopération_nord_su(11), développement_écono(51), équilibre_nord_sud(40), explosion_démograph(62), famine(31), flux_migratoire_hum(33), pauvreté(40), pvd(56), paupérisation(8), surpopulation(12), inégalités_sociales(10), concentration_humai(10), intolérance(3), risques_conflits(6), vieillissement_popu(6),
inéga+l(19), mondia+l(23), partie+(12), croissance+(18), globe+(8), population+(51), atteint<(9), distribut+ion(8), econom+3(25), ethn+3(6), répartit+ion(9), sanitaire+(6), spatia+l(4), système+(11), terre+(12), faible+(9), fait(15), petit+(6), vieill<(10), enjeu+(5), continent+(5), producti+f(20), état+(10), pression+(9), région+(10), seuil+(5), exist+er(8), nécessit+er(5), résult+er(8), évid+ent(4), extensi<(6), crise+(5), facteur+(7), niveau+(17), nourriture+(3), oeuvre+(4), expliqu+er(3), faire.(13), écologiques(7), dens+e(3), différ+ent(9), gener+(5), age+(2), exclu+(2), nombr+eux(7), régiona+l(6), sauvage+(4), amont(2), associe+(5), cout+(9), fonction+(4), guerre+(4), ordre+(4), orient(2), retour+(4), situation+(5), tendance+(3), ag+ir(5), aller.(10), amplifi+er(4), approch+er(2), concern+er(6), cré+er(5), immigr+er(2), parl+er(2), remont+er(2), arid+e(2), catastroph<(8), cause+(6), chroniqu<(2), énorm+e(4), fréqu+ent(2), sever+e(2), détriment(4), impacts(3);	conflits_d_usage_de(4), conflits_militaires(3), religion(2), répartition_humaine(5), Asie(2), transfert_de_techno(2);

Figure 17 : Autour du développement

Une classe autour de la désertification des campagnes n'apparaît que dans l'analyse sur mots-clefs. Cette classe n'apparaît pas dans l'analyse sur mots-clefs des réponses en anglais. On ne peut donc pas supposer qu'il s'agisse d'un biais de codification. Il se peut de plus qu'elle apparaisse à un niveau de découpage plus fin.

Il reste une question en suspens : les problèmes ont-ils été classés de la même manière ? Le chi-deux calculé sur le tableau croisant les deux typologies est très élevé ; il nous conduit à rejeter l'hypothèse d'indépendance. La Figure 18 montre pour chaque classe le nombre de problèmes communs aux deux typologies et les parties complémentaires. En dehors de la classe « désertification, déforestation», la partie commune regroupe la majorité des individus

de chaque classe. La classe la plus homogène est de toute évidence la classe « changement climatique ».

On peut donc considérer que les résultats de l'analyse sur texte brut sont comparables à ceux de l'analyse par mots-clefs. Cette dernière offre une classe interprétable supplémentaire (désertification des campagnes) alors que l'analyse sur texte brut sépare deux types de réponses sur le développement économique difficiles à distinguer sur le plan sémantique. L'amélioration de l'interprétabilité apportée par la codification par mots-clefs justifie-t-elle le surcoût corrélatif ? Il est impossible de conclure scientifiquement. Si nous n'avions eu que des réponses en français, nous n'aurions sans doute pas choisi cette option de codification par mots-clefs.

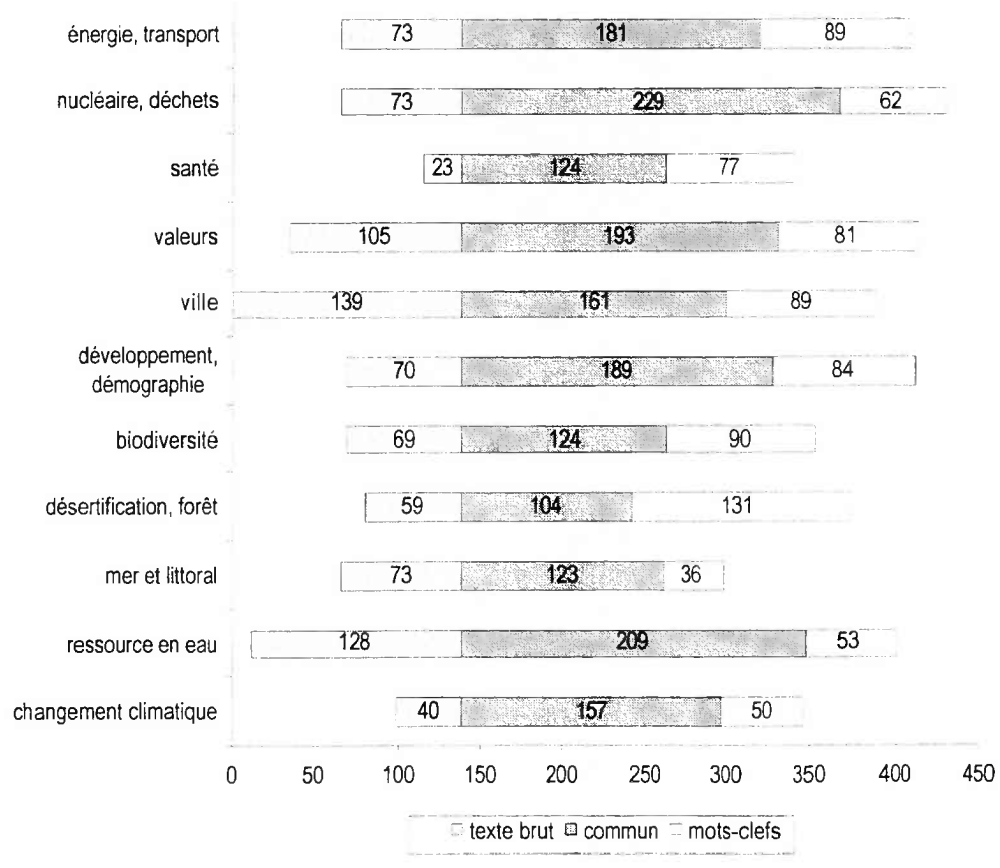


Figure 18 : Croisement des deux typologies

CONCLUSION

Où en sommes-nous ? Il existe actuellement une version en anglais d'Alceste avec reconnaissance des mots-outils et lemmatisation des verbes irréguliers. La possibilité d'une lemmatisation plus complète existe, mais les résultats n'ont pas encore été testés. Nous avons montré d'une manière qui nous semble définitive que, dans notre logique, les mots-outils ne devaient pas intervenir dans l'analyse. En revanche, pour la lemmatisation, d'autres tests comparatifs avec et sans lemmatisation seront nécessaires pour déterminer le caractère indispensable ou non de cette étape sur des textes anglais.

L'existence d'une version pour l'anglais, ne résout pas le problème de l'analyse de textes bilingues. Nous avons vu que l'indexation des réponses avec mots-clefs et la traduction partielle de ceux-ci était une solution envisageable. Nous sommes actuellement en train de tester la qualité d'un traducteur automatique, dont dispose l'EPHE. Si celle-ci se révèle acceptable, on comparera les analyses sur les textes en anglais, puis sur les textes traduits. Si les résultats sont convergents, toutes les réponses (françaises et anglaises traduites en français) pourront être analysées simultanément.

Les très nombreux tests que nous avons été amenés à effectuer nous ont conduit à certaines réflexions sur la démarche d'interprétation. La qualité de la structure lexicale et sémantique d'un corpus (le fait qu'il soit organisé autour d'un certain nombre de noyaux) est difficile à définir a priori. Pour dire les choses trivialement, comment peut-on savoir a priori qu'un corpus de textes se prêtera bien à une analyse lexicale (construction d'une typologie des discours) ?

Nous avons souvent dit que l'idéal était d'avoir des sources nombreuses parlant d'objets similaires, comme c'est le cas dans les réponses à des questions ouvertes, dans des récits de parcours, dans un corpus de fiches de dégustation de vin. Mais il vaut mieux aussi que le canevas rédactionnel soit similaire et bien respecté. Une fois ces contraintes de départ respectées, est-on sûr d'obtenir des résultats intéressants, ce qui revient à dire que l'organisation sémantico-lexicale est interprétable ? Pour avoir une réponse, il n'y a pas d'autre moyen que de tenter des analyses.

L'identification de la structure textuelle n'est pas aisée. Benzécri annonçait avec enthousiasme que l'analyse des données permettait de faire émerger le modèle contenu de manière cachée dans les données. Mais il n'insistait pas sur le fait que cette « mise à jour » peut être longue et difficile. Pour un praticien, de nombreux, parfois innombrables, essais sont nécessaires pour faire surgir cette structure, autrement dit pour sélectionner le meilleur découpage. Il faut bien reconnaître que parfois les efforts sont vains : que faire de résultats triviaux ou ininterprétables ?

Conscient de ces difficultés, Max Reinert a conçu au coeur même de son logiciel un test de stabilité extrêmement robuste : à chaque analyse, deux types de découpages sont effectués et les deux typologies sont croisées afin de ne conserver que les classes les plus stables. Il convient en sus de ce test préliminaire d'effectuer de nombreux tests en faisant varier le découpage, la taille des unités de contexte, le nombre de locutions, le rôle de certains termes, en lemmatisant ou non... En examinant les résultats, petit à petit se forge dans l'esprit la représentation globale sous-jacente. On peut dire que la mise à jour est faite quand on reconnaît presque immédiatement les noyaux recherchés dans une nouvelle analyse.

Ensuite, pour le confort du lecteur, on ne conservera qu'une analyse qui sera le prototype de toutes les autres et qui mettra le mieux en évidence le modèle. Ce que nous disons ici pour l'analyse statistique des données textuelles est tout aussi vrai pour les tableaux de données numériques, pour le traitement des enquêtes par exemple. Construire une typologie est d'une simplicité enfantine, en revanche, construire une typologie qui soit interprétable par celui qui la construit et par celui qui la lit est une tâche plus ardue. Là encore il faut présenter une structure des données où chaque noyau ait sa propre cohérence interne et participe à la construction de la structure globale. Mais ne passe-t-on pas alors d'une approche descriptive à une forme de « modélisation » ?

Nous avons par exemple fait des essais sur deux corpus : les fiches Parker et les fiches Hachette de dégustation de vin. A priori, les corpus respectent les règles de cohérence présentées ci-dessus. De nombreuses analyses ont été faites, en conservant ou en excluant les mots-outils, en introduisant ou non certaines locutions, en lemmatisant ou non. Sur l'un des

corpus, Hachette, nous sommes parvenus assez aisément à identifier la structure en découpant les fiches en sous-unités :

ROUGE (Rouge, Bordeaux), BLANC (Blanc, Lignée), (Temps, Coeur de Bourgogne).

Faire une analyse sur des fragments de fiches ou sur les fiches entières conduit à des résultats assez différents. Dans le premier cas, on obtient à la fois des aspects liés à la mise en mots de la dégustation, mais aussi des éléments de la structure interne des fiches. Dans le second, on a bien une typologie des objets.

En revanche, sur le corpus Parker, aucune analyse n'a semblé satisfaisante, parce que les discordances étaient trop fortes entre les différentes classifications.

BIBLIOGRAPHIE

ANASTEX S. J. (eds.), (1993). *JADT 1993 : Actes des Secondes journées internationales d'analyse statistique de données textuelles*. Paris : TELECOM.

BEAUDOUIN Valérie, (1994). « Avancées en analyse lexicale : structure lexicale, registres et thèmes », *Cahiers de recherche CRÉDOC*, n°61. Paris :, 1994, pp. 1-78.

BEAUDOUIN Valérie, LAHLOU Saadi, (1993). « L'analyse lexicale : outil d'exploration des représentations ». *Cahiers de recherche CRÉDOC*, n°48. Paris.

BENZÉCRI, J.P., coll. (1981). *Pratique de l'analyse des données, Linguistique et lexicologie*. Paris : Dunod.

BOLASCO Sergio, LEBART Ludovic, SALEM André (eds) (1995). *JADT 1995 : III Giornate internazionali di Analisi Statistica dei Dati Testuali.* Roma : CISU, vol. I et II.

CONSTANT Patrick, (1991). *Analyse syntaxique par couche*. Paris : Doctorat ENST.

Encyclopedia Universalis (1968). « Linguistique et statistique », vol 9, 1968.

LAHLOU Saadi, (1995). *Penser Manger : les représentations sociales de l'alimentation*, Thèse EHESS sous la direction de Serge Moscovici.

MARTINET André (1970, 1991). *Éléments de linguistique générale*, Paris, Armand Colin.

MOUGIN Pascal (1995). « Mondes lexicaux et univers sémantiques : le logiciel ALCESTE au service de l'étude de l'imaginaire simonien, à partir du traitement de *La Route des Flandres* », Literary and Linguistic Computing, Vol. 10, n°1 pp. 59-68.

MOUNIN Georges (1963). *Les problèmes théoriques de la traduction*, Paris, Gallimard.

MULLER Charles, (1977 réed 1992).- *Principes et méthodes de statistique lexicale*, Larousse, réimpression Slatkine-Champion, Genève-Paris.- 211p.

PIGAMO Frédéric, (1990). *Outils de traitement sémantique du langage naturel.* Paris : Doctorat ENST.

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

REINERT Max, (1990). ALCESTE : "Une méthodologie d'analyse des données textuelles et une application : Aurélia de Gérard de Nerval". *Bulletin de méthodologie sociologique*, n°26, pp. 24-54.

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

TESNIÈRE Lucien, (1959, 1982). *Eléments de syntaxe structurale.* Paris : Klincksieck, 1959, 674 p.

ZIPF G. K., (1974 (1ère édition 1936)).- *La psychobiologie du langage : une introduction à la philologie dynamique*, RETZ-CEPL, Paris.

ANNEXES

ANNEXE 1 : LE THESAURUS DES PROBLÈMES PRIORITAIRES DE L'ENQUÊTE RECHERCHE ET ENVIRONNEMENT

MOTS-CLÉS ET FRÉQUENCES DE CITATION

N.B. : Les mots français sont repérés par le préfixe f_

Mot-clé	Fréquence d'apparition
health	502
climatic change	468
air pollution	459
biodiversity	401
demographic growth	387
pollution	382
deforestation	359
developing countries	338
transportation	323
energy	321
demographic explosion	302
industry	283
greenhouse effect	252
water pollution	252
values	240
urban areas	237
development	226
water	225
soil pollution	217
climate	212
ozone	204
soil erosion	204
agriculture	198
urban life	196
desertification	193
fossil fuels	188
over exploitation of resources	187

Mot-clé	Fréquence d'apparition
migration	182
water scarcity	182
water resources	181
forest	180
quality of life	180
poverty	177
nuclear wastes	167
marine sites	163
nuclear related risks	162
environmental policy	159
environmental degradation	157
environmental consciousness	156
overfishing	151
extinction of species	150
CO2	148
food	148
marine pollution	143
rural areas	141
industrial pollution	137
soil	134
communication	131
natural ecosystems	129
education	128
social cohesion	127
water quality	124
traffic	122
renewable resources	121

Mot-clé	Fréquence d'apparition
consumption consumers	118
hunger	117
industrial risks	115
sustainable development	111
ethics	109
global change	109
industrial wastes	109
demands on energy	108
global warming	107
depletion of resources	102
wastes	102
accidents	101
overpopulation	100
f_ appauvrissement biodiversité	98
drinking water	97
waste management	96
megacities megapolis	95
f_ développement économique	92
destruction of ecosystems	91
soil degradation	91
scientific assessment	90
technology	90
chemical pollution	87
interdependency of problems	86
protection of resources	84
acid rain	83
hygiene	83
noise	82
waste recycling	82
epidemics	79
productive land	79

Mot-clé	Fréquence d'apparition
war	79
scarcity of resources	78
conflicts for resources	76
eastern European countries	76
social unrest	75
aids	74
f_ équilibre Nord Sud	74
unemployment	73
younger generations	73
domestic wastes	72
malnutrition	72
uncontrolled illnesses	72
population pressure	70
alternative energies	69
solidarity	69
rising sea level	68
anthropogenic action	66
north south relation	66
groundwater resources	65
moral outlooks	65
pesticides	64
inefficient environmental policy	63
landscapes	63
northern western countries	63
nuclear energy	63
wasting of resources	63
distribution of world resources	62
waste storage	61
water supply	60
Africa	59
world economy	59

Mot-clé	Fréquence d'apparition
f_ appauvrissement du sol	58
natural disasters	58
population control	58
chaotic urbanization	56
democracy	55
diffuse pollution	55
social inequalities	55
urban rural areas relations	55
clean processes	54
soil quality	54
chemical compounds	53
demands on resources	51
f_ gestion ressource	51
natural resources	51
tropical rain forest	51
cars motor vehicles	50
supremacy of economy	50
flora	49
genetically modified organisms	49
public health	48
urban ecology	48
non renewable resources	47
salinization	47
ageing population	46
congestion of space	46
tourism	46
f_ action de l'homme	44
f_ pollution urbaine	44
individualism / selfishness	44
f_ désertification des campagnes	43
f_ production agricole	43

Mot-clé	Fréquence d'apparition
public awareness	43
viruses	43
fisheries resources	42
fresh water	42
northern western way of life	42
f_ nucléaire	41
threatened species	41
spread of diseases	40
cleansing	39
genetic research	39
biotechnology	38
f_ préservation espaces	38
fauna	37
f_ concentration urbaine	37
f_ surexploitation	37
overconsumption	37
acidification	36
f_ maladie	36
genetic related risks	36
predictability	36
toxic wastes dumps	36
Asia	35
floods	35
nuclear powerplants	35
clean energies	34
international institutions	34
new viruses	34
oil spills	34
environmental disasters	33
f_ aménagement du territoire	33
f_ cadre de vie	33
f_ consommation alimentaire	33
f_ financement coût	33

Mot-clé	Fréquence d'apparition
micro organisms	33
international agreement	32
political systems	32
new diseases	31
social explosion	31
fertilizers	30
natural regulation	30
undeliberated development	30
coastal ecosystems	29
coastal pollution	29
energy saving	29
nitrate	29
oil	29
survival	29
contaminants	28
destruction threat on habitat	28
environmental sciences	28
f_ pays industrialisés	28
irreversible destruction	28
manipulation disinformation	28
solar energy	28
working conditions	28
f_ maximisation du profit	27
f_ transport routier	27
new species	27
f_ littoral	26
nuclear weapons testing	26
toxic materials	26
agricultural industrialization	25
degradation of resources	25
droughts	25
drug addiction	25
former USSR Russia	25

Mot-clé	Fréquence d'apparition
f_ devenir des infrastructures	25
f_ maîtrise des procédés	25
UV effects	25
f_ villes	24
pauperization	24
terrorism	24
weapons	24
biomass	23
demands on water	23
f_ développement industriel	23
f_ irréversibilité	23
greed	23
organic pollution	23
overcrowding	23
rainfalls	23
culture	22
delayed effects	22
exclusion homeless	22
law	22
raw materials	22
environmental cost	21
f_ atmosphère	21
f_ pollution agricole	21
f_ prévention des risques	21
f_ traitement des déchets	21
political instability	21
suburbs	21
cancer	20
ecotoxicology	20
f_ économie libérale	20
f_ pollution des nappes	20
f_ rejet des polluants	20
f_ stockage	20

Mot-clé	Fréquence d'apparition
man made environment	20
timber industry	20
f_ choix rail route	19
f_ transport collectif	19
threatened cultural activities	19
fundamentalism religious conflict	18
f_ risques naturels	18
local travels	18
overproduction	18
arid regions	17
computers / data works	17
dumps	17
ethnic conflicts	17
f_ mobilité	17
heavy metals	17
newly industrialized countries	17
contempt for politics	16
financial tools	16
f_ insécurité	16
interdisciplinary approaches	16
long distance travels	16
military activity	16
stratosphere damage	16
biosphere	15
f_ gestion environnement	15
f_ maîtrise urbanisme	15
f_ moyens de communication	15
f_ répartition humaine	15
f_ surexploitation agricole	15
human / animal adaptation	15
marine plankton	15
north south conflicts	15

Mot-clé	Fréquence d'apparition
antibiotics	14
CFC production	14
civism	14
electricity production	14
f_ coopération nord sud	14
f_ modes de vie	14
sewage	14
aid to development	13
biotopes	13
f_ aménagement espaces	13
f_ gestion écosystèmes	13
f_ parcellisation de l'espace	13
f_ prise de conscience	13
f_ terre arable	13
illegal activities	13
media	13
productivity	13
waste exportation	13
bacteria	12
death	12
f_ génétique	12
f_ pollution nucléaire	12
f_ rejets industriels	12
f_ relation au travail	12
f_ risques de conflits	12
f_ surmédiatisation	12
replanting	12
research funds	12
resistant species	12
volcanic seismic related risks	12
credibility of science	11
dependence on energies	11
fires	11

Mot-clé	Fréquence d'apparition
f_ approvisionnement en eau	11
f_ concurrence compétition	11
f_ disparition patrimoine génétique	11
f_ respect de la personne	11
global entropy	11
aerosols	10
civil wars	10
f_ acculturation	10
f_ pénurie d'énergie	10
f_ qualité aliments	10
f_ responsabilité individuelle	10
transboundary pollution	10
f_ problème de société	9
f_ uniformisation des idées	9
indoor environment	9
nationalism	9
pests	9
railroad	9
spread of species	9
biological chemical	8
f_ embouteillage	8
f_ gaz à effet de serre	8
f_ irrigation	8
f_ mer poubelle	8
f_ production alimentaire	8
respect to human life	8
rural lifestyle	8
agricultural monocultures	7
authoritarian regimes	7
economical crisis	7
f_ commerce international	7
f_ conflits eau	7

Mot-clé	Fréquence d'apparition
f_ course aux armements	7
f_ élimination déchets	7
f_ emballages	7
f_ législation pollution	7
human fertility sterility	7
innovation imagination	7
international crisis	7
product life cycle	7
uniformity of behavior	7
women	7
f_ concentration du pouvoir	6
f_ délocalisation des industries	6
f_ environnement	6
f_ hydrocarbures	6
f_ loisir	6
f_ non respect du droit	6
f_ perte des valeurs	6
f_ transport aérien	6
grazing of cattle	6
leaders awareness	6
wood fuel	6
Amazonia	5
asthma	5
atmospheric pollution	5
availability of resources	5
ecologists	5
f_ agriculture propre	5
f_ agriculture traditionnelle	5
f_ climat régional	5
f_ déchets urbains	5
f_ désengagement de l'état	5
f_ France	5

Mot-clé	Fréquence d'apparition
f_ habitation décente	5
f_ limiter les déchets	5
f_ océans	5
f_ pays poubelles	5
f_ prix de l'eau	5
f_ produit phytosanitaire	5
f_ religion	5
f_ risques technologiques	5
geological timescale	5
middle east	5
polar regions	5
algae	4
carcinogens	4
f_ acidification de l'eau	4
f_ acidification sol	4
f_ chine	4
f_ gaz	4
f_ intégrisme	4
f_ intolérance	4
f_ inventaire du patrimoine	4
f_ lac rivière	4
f_ pénurie du bois	4
f_ produits polluants	4
f_ relation humaine	4
f_ ressource renouvellement	4
f_ risques biologiques	4
f_ transfert de technologies	4
f_ transport électricité	4
genetic illnesses	4
space matters	4
taxation	4
abortion	3
diesel	3

Mot-clé	Fréquence d'apparition
family planning	3
f_ allergies	3
f_ assèchement	3
f_ choix fleuve route	3
f_ comportement écologique	3
f_ développement aquaculture	3
f_ énergie éolienne	3
f_ jeunesse	3
f_ matériaux non dégradables	3
f_ valorisation forêt	3
property rights	3
self destruction	3
biotic exchanges	2
corruption	2
f_ analphabétisme	2
f_ désaccord scientifique	2
f_ diversité génétique	2
f_ eau de surface	2
f_ emploi de proximité	2
f_ maîtrise développement	2
f_ militantisme écologique	2
f_ PME PMI	2
f_ résistance des mutants	2
f_ sécurité routière	2
f_ travail des enfants	2
insects	2
non recyclable materials	2
availability of energy	1
concrete	1
f_ aire de jeux	1
f_ approximation scientifique	1

Mot-clé	Fréquence d'apparition
f_ droit d'ingérence	1
f_ écologie des transports	1
f_ économie des transports	1
f_ Inde	1
f_ investissement dans PVD	1
f_ méthane	1
f_ MST	1
f_ vaccination	1
individual liberty	1
mutagens	1
nuclear materials transport	1
phosphate	1
self sufficiency	1

APPARIEMENT MOTS-CLÉS ANGLAIS ET FRANÇAIS

Français	Anglais	Français	Anglais
accident	Accidents	générations futures	Younger generations
accroissement démographique	Demographic growth	gestion déchets	Waste management
aérosol	Aerosols	hygiène	Hygiene
Afrique	Africa	individualisme	Individualism/selfishness
agriculture s.a.i.	Agriculture	industrie s.a.i.	Industry
appauvrissement ressources	Scarcity of resources	inégalités sociales	Social inequalities
arme chimique/biologique	Biological/chemical weapons	inondation	Floods
arme nucléaire	Nuclear weapons testing	insectes	Insects
Asie	Asia	institutions internationales	International institutions
banlieue/ghetto	Suburbs	interdépendance des problèmes	Interdependency of problems
biodiversité	Biodiversity	interdisciplinarité	Interdisciplinary approaches
biomasse s.a.i.	Biomass	maladie nouvelle	New diseases
biotechnologie s.a.i.	Biotechnology	malnutrition	Malnutrition
biotopes	Biotopes	manipulation génétique	Genetic research
bruit	Noise	matières premières	Raw materials
cancers	Cancer	mégalopoles	Megacities/megapolis
catastrophe écologique	Environmental disasters	micro-organismes	Micro-organisms
centrale nucléaire	Nuclear powerplants	milieux marins	Marine sites
changement climatique	Climatic change	milieux ruraux	Rural areas
changement global	Global change	mondialisation de l'économie	World economy
chômage / emploi	Unemployment	morale civique	Moral outlooks
circulation automobile	Traffic	morbidité / mortalité	Death
climat s.a.i.	Climate	mutation génétique	Genetically-modified organisms
CO2	CO2	nappes phréatiques	Groundwater resources
cohésion sociale	Social cohesion	nitrates	Nitrate

Français	Anglais	Français	Anglais
combustible fossile	Fossil fuels	niveau des mers	Rising sea level
communication/information	Communication	nouvelles espèces	New species
concentration humaine	Population pressure	NPI	Newly Industrialized Countries
conflits d'usage des ressource	Conflicts for resources	paupérisation	Pauperization
conflits ethniques	Ethnic conflicts	pauvreté	Poverty
conflits militaires	Military activity	pays de l'est	Eastern-European countries
conflits s.a.i.	War	paysage	Landscapes
conflits sociaux	Social unrest	pénurie en eau	Water scarcity
consommation d'eau	Demands on water	pesticides	Pesticides
consommation d'énergie	Demands on energy	pétrole	Oil
consommation s.a.i.	Consumption/consumers	pluies acides	Acid rain
couche ozone	Ozone	pluviosité	Rainfalls
culture	Culture	politique environnementale	Environmental policy
décharges	Dumps	pollution air	Air pollution
décharges/déchets toxiques	Toxic wastes/dumps	pollution chimique	Chemical pollution
déchets industriels	Industrial wastes	pollution des côtes	Coastal pollution
déchets ménagers	Domestic wastes	pollution diffuse	Diffuse pollution
gaspillage des ressources	Wasting of resources		
déchets nucléaires	Nuclear wastes	pollution eau	Water pollution
déchets s.a.i.	Wastes	pollution industrielle	Industrial pollution
déforestation	Deforestation	pollution mer	Marine pollution
dégradation du sol	Soil degradation	pollution s.a.i.	Pollution
dégradation ressources	Degradation of resources	pollution sol	Soil pollution
démocratie	Democracy	procédés propres	Clean processes
désertification	Desertification	PVD s.a.i.	Developing countries
développement durable	Sustainable development	qualité eau	Water quality
diesel	Diesel	qualité vie	Quality of life
disparition des espèces	Extinction of species	reboisement	Replanting
drogues	Drug addiction	réchauffement	Global warming
droit s.a.i.	Law	recyclage déchets	Waste recycling

Français	Anglais	Français	Anglais
eau douce	Fresh water	région aride	Arid regions
eau potable	Drinking water	réglementation internationale	International agreement
eau s.a.i.	Water	régulation naturelle	Natural regulation
écologie urbaine	Urban ecology	relation ville/campagne	Urban/rural areas relations
économie énergie	Energy saving	répartition des ressources	Distribution of world resource
écosystèmes naturels	Natural ecosystems	respect de l'environnement	Environmental consciousness
éducation	Education	ressource eau	Water resources
effet de retard	Delayed effects	ressource marine	Fisheries resources
effet de serre	Greenhouse effect	ressource non renouvelable	Non-renewable resources
énergie de substitution	Alternative energies	risques génétiques	Genetic related risks
énergie non polluante	Clean energies	risques industriels	Industrial risks
énergie nucléaire	Nuclear energy	risques nucléaires	Nuclear-related risks
énergie renouvelable	Renewable resources	risques sismiques/volcaniques	Volcanic/seismic related risks
énergie s.a.i.	Energy	salinisation	Salinization
énergie solaire	Solar energy	santé	Health
engrais	Fertilizers	santé publique	Public health
épidémie	Epidemics	sécheresse	Droughts
épuisement ressources	Depletion of resources	SIDA	AIDS
érosion	Soil erosion	solidarité	Solidarity
éthique	Ethics	sols s.a.i.	Soil
évaluation scientifique	Scientific assessment	surexploitation halieutique	Overfishing
ex-URSS	Former USSR/ Russia	surexploitation ressources naturelles	Over exploitation of resources
exclusion	Exclusion/homeless people	surpopulation	Overpopulation
explosion démographique	Demographic explosion	terrorisme	Terrorism
exportation déchets toxiques	Waste exportation	tourisme	Tourism
famine	Hunger	transport s.a.i.	Transportation
faune	Fauna	transport urbain	Local travels
fertilité du sol	Soil quality	urbanisation	Urban areas
flore	Flora	valeurs s.a.i.	Values

Français	Anglais	Français	Anglais
flux migratoire humain	Migration	vie en ville	Urban life
forêt amazonienne	Amazonia	vieillissement population	Ageing population
forêt s.a.i.	Forest	virus	Viruses
forêt tropicale humide	Tropical/rain forest	virus nouveaux	New viruses

ANNEXE 2 : ANALYSE DES OBSERVATIONS DES QUESTIONNAIRES EN FRANÇAIS DE L'ENQUÊTE RECHERCHE ET ENVIRONNEMENT

Ségolène EVEN

Comme de nombreux questionnaires, l'enquête *« Recherche et Environnement : problèmes prioritaires et problèmes émergents »* propose aux répondants de noter, en fin de l'étude, leurs remarques et leurs perceptions : *« Quelles sont vos observations sur le contenu de cette enquête ? »*. Il nous a semblé intéressant d'effectuer un travail d'analyse sur cette question précise, par ailleurs très rarement traitée. Le répondant, libre de réagir dans le sens qui lui convient, nous informe sur la lisibilité et la compréhension du questionnement. Cette analyse permettrait en outre d'enrichir les résultats finaux, du moins de les relativiser..

Pour traiter cette question, nous n'avons pris en compte que les questionnaires en français, car l'analyse simultanée des deux langues anglais - français était trop complexe. Nous avons utilisé le logiciel ALCESTE afin de dégager au travers d'une analyse des cooccurrences les profils de réponses. Notons que globalement, le nombre de réponses est peu élevé : sur les 373 questionnaires français, 102 chercheurs n'ont rien répondu à cette question. Nous travaillons donc sur une base de 201 observations classées.

Les observations se répartissent en deux ensembles principaux et distincts, d'égale importance (102 observations contre 99). Chaque ensemble se subdivise jusqu'aux sept sous-ensembles finaux que nous nommerons classes.

Ces deux groupes se distinguent par la nature des critiques formulées par les chercheurs. En l'occurrence, il ne s'agit pas d'une distinction entre critique positive et critique négative, mais entre un constat de pure forme et une analyse des fondements sur lesquels repose l'enquête.

Le premier groupe est constitué d'observations qui évaluent essentiellement la forme du questionnaire. La réaction est instinctive, affective, elle exprime le sentiment qui se dégage immédiatement, les chercheurs se racontent. Le questionnaire est défini par sa complexité et sa longueur ; quoiqu'intéressant et suscitant la réflexion, le temps manque aux scientifiques pour mieux y répondre. Il serait intéressant d'y comparer le taux de questions ouvertes laissées sans réponse. Cet ensemble constate, et donc ouvre peu le débat, bien que des jugements positifs émergent selon les classes qui composent ce groupe.

Le second groupe formule une critique de fond en s'interrogeant sur les choix de l'enquête. Les remarques s'articulent autour du sens des questions, de la compétence des scientifiques, de la formulation des enjeux et du rôle des différents acteurs de l'environnement. Le débat est ouvert sur les postulats formulés par l'enquête, sur la manière dont les questions furent posées, sur l'opportunité même d'une enquête s'adressant aux scientifiques. La critique est plus profonde, bien qu'ici encore il faille se garder d'analyser le contenu des réponses dans une optique uniquement négative.

Car, et cela est valable pour la totalité des observations, il transparaît une note de satisfaction au travers de toutes les réponses que nous avons eues : l'enquête, si elle est jugée parfois mal orientée, ambiguë ou longue, a le mérite d'être ambitieuse.

L'ENSEMBLE A : PERCEPTION ET CONSTAT

Classe 1 - Enquête complexe

Cette classe représente 17% des réponses classées, soit 34 observations. La remarque type du chercheur porte ici sur la complexité de l'enquête. Les mots significatifs qui forment cette classe sont l'adjectif « difficile+ » et le verbe « rempl+ir ». C'est en considérant le manque d'aisance qu'il a éprouvé pour répondre aux nombreuses questions de l'enquête que le chercheur formule ce constat : répondre au questionnaire est compliqué. Cette réaction est d'ordre affectif, il nous est fait part d'un certain ressentiment.

Les phrases qui suivent illustrent bien ce point :

« Beaucoup de mal à le remplir [...]. En les relisant, je ne suis pas fier, mais c'est fait. »
« Certaines rubriques sont très difficiles à remplir. »

Des justifications supplémentaires enrichissent cette première remarque au travers des mots « ambigu<», « complet+ », « ouvert+ », « heure+ », et l'explicitent plus en détail.

- L'ambiguïté de certaines questions justifie la qualification initiale sur la complexité et l'étaye précisément. Elle renvoie même le plus souvent à des exemples particuliers.

  « Questions 2 à 8 ambiguës ».

- L'exhaustivité est une considération positive sur l'extrême richesse de l'enquête, mais elle rend compte également de sa complexité : faire le choix d'une réponse adéquate et intelligente est compliqué par la multitude des solutions proposées.

  « Difficile à remplir car très vaste ».
  « Utile, complet et cohérent ».

- Le mode de questionnement par le biais des nombreuses questions ouvertes est apprécié.

  « Bon compromis entre le questionnaire ouvert et fermé ».

Mais, par cette allusion implicite à la rédaction, le chercheur renchérit de nouveau sur la complexité de cette enquête, où l'abondance des questions ouvertes oblige à l'écriture et à l'argumentation. La référence au temps au travers du mot « heure+ » transparaît ici.

Classe 2 - Enquête trop longue mais bien faite

Cette classe est formée de 30 observations, soit 15% des réponses classées.

Deux éléments de réponses composent cette classe : un constat d'ordre général à connotation parfois négative, nuancé par des termes laudatifs.

L'essentiel de la remarque porte ici sur la durée excessive du questionnaire. Les mots les plus représentatifs sont le verbe « répondre. » et le mot « long+ ». Tels que, ces termes sont strictement informatifs, comme l'illustre cette phrase type :

« C'est long ! ».

Ce constat s'enrichit toutefois avec la présence du mot « difficulté+ »,

« souvent difficile à répondre simplement »,

et des adverbes « trop » et « peu » (à comprendre dans le sens de « quelque peu »), qui sont des segments répétés fréquemment.

« Trop long, trop complexe »
« Un peu long. Difficulté pour répondre. »

Les mots suivants, également typiques de la classe, renvoient à une qualification plus positive du questionnaire : « simple+ », « idée+ », cohérence », et le segment répété « bien faite ».

La nuance est donc de mise d'autant que, le mot outil le plus représentatif est « mais ». Il y a donc bien rupture dans la formulation de la critique :

« A nécessité un temps assez long, mais a permis de synthétiser des idées, voire à déboucher sur une certaine cohérence. »

A ce titre, les deux autres verbes représentatifs de cette classe, « espér+er » et consacr+er », exprimant l'engagement, révèlent une réelle volonté de participation. Du moins le chercheur ne reste pas indifférent à l'enquête, en dépit des propos parfois négatifs qu'il peut tenir :

« Un peu long, mais très bien faite. Espérons que ça serve aussi à notre environnement ».

Cette classe se caractérise d'ailleurs par l'emploi fréquent de la première personne du singulier dans les segments répétés, marque que les chercheurs se sont sentis concernés par l'enquête : « je ai », « je espér+er qu+ », « je ne », qu+ je ai ».

Classe 4 - Enquête intéressante qui oblige à la réflexion

Cette classe comprend 21 observations, soit 10% des réponses classées. Le questionnaire est d'emblée qualifié par cette classe sur le mode de l'estime. « Intéressant+ » et « réflex+ion » sont en effet les deux mots les plus caractéristiques. Ils sont relayés par d'autres termes : « demand+er », « travail<», « aller. », « nécessite+ », « permis », soit autant de termes qui font référence au mot « réflexion+ » et qui construisent autour de ce qualificatif un réseau de remarques, qui éclaire sur la façon dont l'enquête interpelle le chercheur :

demande - réflexion,

travail - réflexion,

nécessite - réflexion

permis - réflexion.

Au travers de ces mots, le chercheur décline les modalités de sa participation ; il démontre qu'il a joué le jeu, peut-être même le revendique. Il indique en tout cas deux axes récurrents dans les autres classes de ce groupe de remarques :

- le questionnaire a demandé de la réflexion, donc l'obligation de travail, ce qui n'est pas sans laisser poindre un léger reproche quant à l'effort qu'exige l'enquête. Cet effort renvoie à la qualité complexe du questionnaire déjà évoquée précédemment ;
- le questionnaire oblige à la réflexion, c'est-à-dire qu'en la suscitant, il permet de clarifier les idées, de rebondir, d'ouvrir de nouvelles perspectives d'analyse.

C'est d'ailleurs sous cet angle de la réflexion suscitée que l'enquête acquiert de l'intérêt. La phrase qui suit illustre bien cet aspect :

> « Questionnaire qui a nécessité un très important effort de réflexion pour essayer d'apporter des réponses appropriées et qui pourraient être utilisées par les décisionnaires ».

Classe 5 - Enquête qui mérite du temps difficile à se ménager

Cette classe comprend 17 observations, soit 8% des réponses classées. Cette classe traite en fait de l'emploi du temps des chercheurs. L'organisation de leur temps de travail est au centre de leurs propos et c'est sous cet aspect que le questionnaire est considéré.

Le mot le plus important est donc « temps ». Deux verbes caractérisent le temps : « trouv+er » et « mérit+er ».

Il est donc question du temps nécessaire que le chercheur doit se ménager dans son emploi du temps pour répondre au questionnaire, mais également de temps qu'il aurait souhaité avoir pour mieux y répondre.

- La contrainte du temps

Le premier verbe « trouv+er », en impliquant la durée excessive du questionnaire, fait référence à la contrainte de temps qui lie le chercheur à ses activités. La longueur de l'enquête, critique latente pour l'ensemble de ce groupe, est soulevée ici parce que les chercheurs luttent pour aménager leur planning.

Voici quelques phrases typiques construites autour de trouver du temps :

« Répondre sérieusement à ce type d'enquête requiert un temps énorme ! ».
« Un peu trop long ! On se lasse rapidement. En fait, on n'a pas forcément le temps ».
« Enquête bien construite, demande bien évidemment du temps difficile à se ménager ».

- L'insatisfaction.

Le second verbe « mérit+er » introduit une valeur positive, qui va dans le sens du souhait de participation et de la reconnaissance, mais révèle sans doute dans le même temps l'insatisfaction, le regret.

« Il donne envie d'écrire des pages et des pages sur le sujet et en même temps suscite le regret de n'avoir pas le temps nécessaire ».
« Beaucoup de questions auraient mérité de longues réponses pour répondre complètement ».

En mentionnant la charge de travail très lourde qui lui incombe, en évoquant un emploi du temps chargé, le chercheur valorise implicitement l'enquête. Puisque l'organisation de son temps de travail est au centre de sa préoccupation, l'enquête à laquelle il a cependant bien voulu répondre, acquiert en conséquence une certaine valeur. Elle fut jugée suffisamment

intéressante pour prendre lieu et place d'autres travaux. Mieux, le regret de ne pouvoir faire plus en terme de participation est sincèrement exprimé.

Les chercheurs de cet ensemble qualifient pour l'essentiel la forme même du questionnaire. Délicat à traiter, celui-ci se définit quasi strictement par sa longueur et sa complexité. Ce jugement est essentiellement descriptif. De fait, ce sont les sentiments qui guident l'essentiel des remarques ; on note ainsi l'importance de l'utilisation de la première personne du singulier. Car les chercheurs parlent d'eux mêmes et de l'aisance avec laquelle ils ont répondu aux nombreuses questions ouvertes et fermées.

Surtout, par delà une description critique de l'enquête, les chercheurs avouent implicitement leur curiosité, si ce n'est leur intérêt. Certes, l'enquête est excessive : trop longue, trop riche ; elle est exigeante : trop compliquée, trop envahissante pourrait-on dire au regard des contraintes de temps énoncées par les chercheurs. Mais elle intéresse. Car, en dépit de ces contraintes qui limitent de manière si draconiennes son emploi du temps, le chercheur a pu répondre à l'enquête, car il l'a jugée nécessaire et utile. Elle intéresse sans doute aussi parce qu'elle ouvre une perspective : on peut supposer que ce questionnement, qui a mobilisé effort et réflexion, se poursuivra au-delà de ces quelques heures passées à répondre à l'enquête. En proposant des idées, des scénarios, en donnant à discuter des points de vue, peut-être l'enquête se prolongera-t-elle dans le temps, dans cet avenir qu'elle aura contribué à décrypter.

On ne peut cependant faire l'économie de cette dernière remarque : si l'enquête se trouve valorisée d'avoir été intégrée à l'emploi du temps des chercheurs parce qu'elle était intéressante et bien faite, il n'en demeure pas moins vrai que c'est en considérant le Temps, comme contrainte et comme exigence, que peut être compris le nombre infime des réponses par rapport à la quantité de questionnaires envoyés.

ENSEMBLE B : ANALYSE ET PROSPECTIVE

Classe 6 - Pertinence des évaluations

Cette classe est formée de 16 observations, soit 8% des réponses classées. Si aucun mot particulier n'émerge réellement de cette classification, on notera que les mots « colonne+ », « impact », « tendance+ », « positi+f », « négati+f », renvoient au chapitre sur *« Vos hypothèses sur l'avenir de la planète »*, pages 12 à 15. La lecture des phrases types le confirme :

« La partie scénario est très peu claire[...] »
« [...] Pour certaines tendances, les colonnes impact de la tendance n'a pas toujours un sens clair. »
« [...] Pages 12, 13, 14, impact sur l'environnement faible à fort : on est gêné, car il n'y a pas de distinction entre impact positif ou négatif [...] ».
« Il n'est pas facile de répondre à la question : vos hypothèses[...] ».

Un chapitre entier de l'enquête est donc soumis à l'observation critique des chercheurs. Propositions et systèmes d'évaluations sont dénoncés pour leur manque de pertinence ou de clarté. Ces questions sont trop ambiguës.

On constatera que cette remise en cause porte sur l'unique partie du questionnaire essentiellement composée de questions fermés. L'enquête proposait d'y évaluer l'importance de tels ou tels facteurs pour l'avenir (facteurs politique, économique, social, technologique, juridique). Dans ce cadre fermé, rigide et directif, le chercheur n'a pas trouvé de propositions de réponse (impacts faible à fort, positif ou négatif, probabilité...) qui le satisfasse.

Classe 3 - Compétence scientifique et sens du questionnaire

Cette classe est formée de 29 observations, soit 14% des réponses classées. Cette classe se constitue dans une opposition entre « compét+ent » et « opinion+ » face à un double problème de « sens ».

« Sens » est le mot le plus caractéristique de cette classe mais il a deux acceptations : il désigne le champs de l'enquête et de l'autre, la signification des questions. « Compét+ent » renvoie au préambule du questionnaire *« Enquête prospective internationale auprès de la communauté scientifique [...] Votre métier vous confère une place particulière [...]. »,* donc au choix des enquêteurs d'interroger les chercheurs parce que ce sont des scientifiques. « Opinion+ » est entendu dans l'acceptation opposée du terme « compét+ent », surtout si la compétence est scientifique ; il se comprend comme « formuler un avis » différent de la « vérité scientifique », qui est prouvée, argumentée, obéissant à des lois, vérifiée par la méthodologie expérimentale.

- Questionnaire généraliste et expertise

Le problème posé par « sens » s'enrichit ici des mots « général+ » et « large+ » qui rendent compte de l'ampleur, de la diversité et de la richesse de l'enquête. « Sens » est compris ici comme étendue ; il est question du « sens large », du « sens général » de l'enquête.

C'est à cette multitude de sujets à traiter que s'oppose « compét+ent ».

L'abondance des situations envisagées par l'enquête, qui font référence à des domaines scientifiques très variés, place le chercheur dans une situation paradoxale : il n'est pas en mesure de traiter toutes les questions en tant que spécialiste, car il ne connaît pas toutes les spécificités des problèmes d'environnement. En place et lieu de l'expertise demandée par l'enquête, il répond de son ignorance et postule donc pour une réponse d'ordre général, qui est loin du jugement scientifique.

Voici quelques phrases types de cette classe qui illustrent parfaitement ce paradoxe :

> « Le fait d'être spécialiste d'un problème ne rend pas obligatoirement compétent pour traiter d'un autre et je n'ai pas répondu à plusieurs questions que dans cet esprit. J'ignore l'état de la question par exemple sur l'effet de serre, l'ozone. Beaucoup de questions posées sont des questions d'opinion et doivent être traitées comme telles sans donner un caractère plus certain. »

« J'ai peu de compétence sur beaucoup de questions. Beaucoup de mes réponses sont instinctives plus que scientifiques. »
« Questionnaire large, qui dépasse forcément les compétences de chacun. »
« Dans son principe, excellente idée ; dans la pratique, très marquée par une idéologie utilitariste. Par ailleurs, les questions sont si générales que les réponses relèvent souvent plus de l'opinion que de la réelle expertise. »

Ainsi, l'importante variété des sujets abordés par le questionnaire ne permet pas aux scientifiques de jouer leur rôle d'expert.

- Opinion et expertise

La formulation des questions soulève un autre problème de sens. En effet, les termes « champ+ » et « échelle+ », renvoyant aux questions elles-mêmes, font directement référence à la manière dont sont abordés les problèmes d'environnement. Non seulement, la signification et la clarté des questions sont remis en cause —elles sont parfois considérées comme ambiguës—, mais leur pertinence est également sujette à caution : l'enquête appellerait plus des opinions que des certitudes. En conséquence, le chercheur s'inquiète de nouveau sur la validité de son interrogation en tant qu'expert.

« Les questions sont quelques fois formulées en fonction de la réponse que vous souhaitez, cf. pages 12 et 15 : cela rend les réponses confuses et peu claires. Manque de précision des intitulés des questions : on ne sait pas à quelle échelle vous parlez, ou là où vous voulez en venir. »
« Beaucoup de questions posées sont des questions d'opinion et doivent être traitées comme telles sans donner un caractère plus certain ».

Cette classe est sous le signe du doute. On note d'ailleurs deux autres termes « risque+ » et « réserve+ », qui mettent en avant le souci de prudence et de restriction. De même, les mots outils caractéristiques de cette classe expriment majoritairement la négation, « ne », « pas », et l'alternative par « ou ».

Les chercheurs font part de leur perplexité face à ce qu'ils considèrent comme deux approches contradictoires du questionnaire : d'une part, une volonté affichée d'interroger des experts ; de l'autre un questionnaire généraliste aux interrogations ambiguës, dans lequel l'expert ne se retrouve pas.

Classe 7 - Définition des responsabilités et des modes d'action

Cette classe représente 54 observations, soit 27% des réponses classées. Il s'agit de la classe la plus importante. Elle se distingue par le nombre et la variété de son vocabulaire. On peut d'ailleurs noter les nombreux mots outils dont le plus important est le mot de liaison « et », qui marque la longueur de l'argumentation.

- Les acteurs de l'environnement

Une articulation essentielle apparaît au travers de deux ensembles de mots :

- « politit+3 », « économ+3 » et « industri »,

- « recherche+ », « scientifi » et « cherch+eur ».

Ces deux ensembles peuvent être compris comme opposés ou complémentaires. Ils stipulent en tout cas que chacun de ces acteurs joue un rôle face aux « problèmes d'environnement » (segment répété le plus caractéristique de cette classe). Les chercheurs sont loin d'être les seuls acteurs concernés ; il faut élargir le débat et interpeller les mondes politique et économique, responsables face aux détériorations de l'environnement, et impliqués la résolution des problèmes. Les chercheurs de cette classe souhaitent redéfinir les obligations respectives de chaque acteur de l'environnement, car, chacun d'eux doit prendre en charge ce qui lui incombe réellement : les termes « act+ion » et « choix », ainsi que les verbes « concern+er » et « consid+er » y réfèrent. Sans doute pourrait-on considérer trois grands acteurs :

- Industrie et économie

La responsabilité, en terme de détérioration, incomberait en majorité à l'industrie et l'économie qui, si l'on s'en réfère aux phrases types caractérisant cette classe, proposent et imposent un modèle de développement économique néfaste pour l'environnement.

> « [L'environnement] ne sera en aucun cas résolu par les lois du marché, par une croissance accrue, y compris par une croissance accrue au niveau mondiale. Tant que polluer coûtera moins cher aux pollueurs que

dépolluer, tant que les règlements internationaux les protégeront, même involontairement, par leur complexité, alors il y aura toujours plus de problèmes d'environnement ».
« [...] Le monde des affaires, qui raisonnent eux à court terme et à travers des profits croissants, [font] abstraction de données capitales pour l'équilibre de la planète ».
« En dernière analyse, les responsables des problèmes retrouvés sont les négociants, capitalisme sauvage [...]. Arrêter la production pour faire monter les prix est la cause la plus sévère d'une nouvelle pression contre les ressources naturelles ».

- Monde politique

Transition entre la responsabilité et l'action, le monde politique est critiqué pour son inertie face aux problèmes d'environnement, mais il est aussi défini comme le plus habilité à agir pour le respect de l'environnement.

« [...] L'environnement est avant tout un choix politique. »

Au regard de son pouvoir d'intervention, les chercheurs reprochent au politique son absence ou son peu d'efficacité.

- La recherche

Enfin, en se confrontant au politique, la recherche délimite la place qu'elle pense raisonnable d'occuper, dans des modalités d'interventions à la fois totalement autonomes sur certains points, mais aussi tributaires de l'action politique pour d'autres.

« [...] La solution des problèmes soulevées sont souvent d'ordre politique et la recherche sert d'alibi pour retarder les prises de position plus courageuses ».
« L'environnement est un problème de politique mondiale. Les scientifiques n'ont pas de solutions à proposer dans ce domaine, sauf des idées individuelles, des dangers à signaler ».
« [...] Un souhait : bien distinguer ce qui est connu et attend des solutions politiques sans nécessiter de recherche nouvelle, ce qui est incertain et nécessite des recherches de techniques, de politiques, d'actions le plus souvent sociales ».
« [...] Les thèmes de recherche devraient être confrontés et discutés avec des responsables des grands groupes industriels et de commerce international ».

Cette distinction que la recherche opère entre des projets qu'elle mène à terme, des solutions qu'elle apporte, et leurs mises en application souvent tardives par les décideurs politiques, la rend sereine. L'enquête interroge la recherche ; celle-ci lui répond en s'interrogeant sur l'efficacité de l'action politique.

D'une certaine manière, la recherche suggère ici que la pérennité et l'aggravation des problèmes d'environnement n'est pas tant de sa responsabilité que celle d'autres acteurs,

politique ou économique. Il est donc souhaitable que son propre pouvoir d'action et de décision soit replacé dans cette problématique qui intègre l'économie de marché et la réglementation.

- Globalité et interactions des problèmes d'environnement

Une seconde tendance d'importance se dégage de cette classe ; elle complète la définition des responsabilités en considérant les niveaux d'interventions souhaitables pour une prise en charge effective des « problèmes d'environnement ».

Cette tendance se définit au travers des mots suivants : « globa+l », « mondia+l », « planète+ », que l'on suppose en opposition aux termes « loca+l » ou « nationa+l ».

L'action à mener par les acteurs répertoriés précédemment nécessite que soit au préalable défini à quel niveau l'intervention est souhaitée pour qu'elle soit efficace. L'enjeu est ici l'échelle ou les échelles dont il est nécessaire de débattre, car la résolution des problèmes d'environnement dépend tout particulièrement de l'analyse que l'on fait de ces mêmes problèmes en termes d'intensité, d'étendue, de durée et de connexion. C'est en s'accordant sur la manière d'aborder les problèmes, puis de les traiter, que l'on sera réellement efficace.

Or, l'efficacité passe, pour les chercheurs de cette classe, par la conscience de la globalité.

> « [...] Les solutions doivent aussi devenir mondiales. L'écologie montre que tout est lié. On ne peut traiter la question de l'effet de serre, par exemple, qu'en réduisant la combustion de produits fossiles. Il faudrait pour compenser développer le nucléaire. De même, ces questions sont liées à des problèmes de développement économique. On ne résoudra rien en s'attaquant aux problèmes les uns après les autres ; seule une réflexion globale peut s'avérer payante à long terme, cf. la devise bien connue « think globally, act globally », si elle est suivie de décision au plan local ».
>
> « La gestion des problèmes d'environnement s'inscrit davantage dans une meilleure prise en compte globale de développement durable et de l'équilibre politique et économique ».
>
> « Il est extrêmement encourageant qu'une initiative telle celle-ci serve à la prise de décision globale ».

Résoudre les problèmes d'environnement demande que toute la planète soit considérée. Une prise en charge parcellaire ne fera que retarder la résolution des problèmes.

L'ensemble de ces observations est à placer sous le signe de la critique constructive. En partant d'une critique du sens, les chercheurs enrichissent le débat d'une réflexion sur la

responsabilité et la prise en charge. De ce point de vue, la remise en cause du questionnaire est progressive.

Il est tout d'abord question d'exemples précis, aux propositions trop rigides pour permettre aux chercheurs d'exprimer avec justesse leurs certitudes. Le propos s'élargit ensuite au caractère généraliste de l'enquête, qui oblige l'expert à relativiser ses certitudes et à ne prononcer que des avis. Sur ce point, les chercheurs avouent leur perplexité face à ce qui leur semble comme le paradoxe de cette enquête.

De là, ils ouvrent le débat sur la responsabilité et l'action, non sans noter qu'une fois encore, les problèmes d'environnement sont abordés du point de vue de la connaissance scientifique, alors qu'il s'avère urgent de les appréhender aussi du point de vue de l'action politique. Une telle démarche tiendrait compte de cette notion de globalité que les chercheurs définissent comme essentielle.

On perçoit enfin au travers de cette analyse une certaine ambiguité : Si la globalité et l'interaction sont des notions fondamentales sans lequelles aucune solution ne peut être engagée, pourquoi une enquête généraliste, qui en beaucoup de points adopte cette perspective, désarme les chercheurs.

Dépôt légal : Septembre 1996

ISSN : 1257-9807

ISBN : 2-84104-072-0

CAHIER DE RECHERCHE

Récemment parus :

Consommateurs et préférences de consommation en 1996
Aude COLLERIE DE BORELY - **n°88 (1996)**

Deux articles sur la localisation industrielle
Philippe MOATI - **n°89 (1996)**

Les inégalités en France : les différentes façons de "penser" en haut et en bas de l'échelle sociale
Georges HATCHUEL, Anne-Delphine KOWALSKI et Jean-Pierre LOISEL - **n°90 (1996)**

Estimations de lois de consommation alimentaire sur un pseudo panel d'enquêtes de l'INSEE
Nilton CARDOSO, François GARDES - **n°91 (1996)**

L'évolution de l'emploi dans l'industrie manufacturière française
Philippe MOATI, Laurent POUQUET - **n°92 (1996)**

Méthode d'étude sectorielle - Vol. 2
Philippe MOATI - **n°93 (1996)**

Modélisation des choix alimentaires des ménages
Patrick BABAYOU, Aude COLLERIE DE BORELY - **n°94 (1996)**

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

ISBN : 2-84104-072-0

CRÉDOC

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